Hindawi
Journal of Engineering
Volume 2023, Article ID 1388200, 20 pages
https://doi.org/10.1155/2023/1388200

Research Article

Efficiency Comparison of Prediction Methods and Analysis of Factors Affecting Savings of People in the Central Region of Thailand

Achara Phaeobang and Saichon Sinsomboonthong

Department of Statistics, School of Science, King Mongkut's Institute of Technology Ladkrabang, Chalongkrung, Bangkok 10520, Thailand

Correspondence should be addressed to Saichon Sinsomboonthong; saichon.si@kmitl.ac.th

Received 9 June 2023; Revised 26 October 2023; Accepted 8 November 2023; Published 7 December 2023

Academic Editor: Assed Naked Haddad

Inarguably, saving is very important for the life of a senior citizen. Artificial neural network (ANN) and multiple linear regression (MLR) analyses have been successfully used to predict and analyze factors affecting the savings of people in several regions of the world. Many studies concluded that ANN is more efficient than MLR. However, some studies concluded that MLR is more efficient. To investigate this issue further, this study directly compared the efficiencies of unoptimized ANN and MLR in predicting and analyzing factors affecting the savings of people in the central region of Thailand in 2019, based on secondary data from a household socioeconomic survey, i.e., the National Statistical Staff Household Income Survey. The data were collected from January 2019 to December 2019 from questionnaires distributed to samples of households. The savings of people in the 25 provinces of Thailand were investigated with MLR and unoptimized ANN. Their prediction efficiencies were compared in terms of root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and processing time. The results showed that for all categories of savings—savings of low-, middle-, and high-income households—MLR was faster in processing time. It also provided a lower RMSE and a higher R^2 than the unoptimized ANN. Nevertheless, unoptimized ANN provided a lower MAE than MLR for the savings of low- and high-income household data. The most important factor affecting the savings of low-, middle-, and high-income households was the factor of deposit interest, bond, share dividends, and other types of investment.

1. Introduction

The meaning of savings in the Royal Institute dictionary is conserving, economizing, and preserving [1]. Saving means the careful use of property and money, which are two major factors of living. Savings money is important because it ensures security for the savers and contributes to economic stability. Savings are important to the economy because they support people's wellbeing, families, and communities. Management of finance and savings is essential in building stability and security, in terms of the livelihood of the people, families, communities, and nations. In addition to the economic benefits to the country and the livelihood of individuals, especially to elderlies, savings also have social, cultural, and educational benefits. In education on savings, member groups must learn the principles and practices as well as the potential outcomes of the savings to be motivated to save. Saving process is educational. People self-train to know how to save [1].

Saving is important for future needs because one may never know for certain whether the source of income would be as secure as it has been. Regular income may disappear, but living expenses still remain. The main objective of saving campaigns is to make people aware of the importance of saving and learn how to save money constantly to maintain life stability [2].

After Thailand's economic crisis in 1997, the situation of Thai people's savings slowly improved. Thailand was able to

gradually repay the loan to the International Monetary Fund and developed from just surviving to becoming sustainable [3].

Savings are closely related to the actual earned income and household consumption. Such income is considered household income and can be used for actual expenses. Households allocate this income for consumption. The remaining money is then kept as savings. Savings are closely related to the theory of consumption. It is believed that households with certain consumption expenditures and income levels find it difficult to reduce their consumption expenditures when their income levels are reduced. The amount of consumption expenditure over a period of time was based on the past, present, and future income expectations over a lifespan [4].

There exist several statistical procedures for finding the relationship between factors affecting the savings of people in some regions and the actual savings of those people. One of the popular methods is MLR analysis. There are also several machine learning procedures for predicting a quantitative variable. The ANN method is a popular one [5]. MLR and optimized ANN have been used to predict the outcomes of several fields of application, and many studies have reported that ANN performed better than MLR. Some studies, however, reported in contrary. Since an optimized ANN has to undergo many processing rounds, which takes time and effort, we thought of comparing unoptimized ANN and MLR, to see which one would be suitable for rapid screening purpose, i.e., to see the power of raw ANN and MLR.

Therefore, our research objectives were to compare the prediction efficiencies of MLR and unoptimized ANN and to analyze factors affecting the savings of people in the central region of Thailand. The comparison was in terms of RMSE, MAE, R^2, and processing time. The analysis used secondary data from the National Statistical Office on the Household Socio-Economic Survey 2019. The survey was conducted every two years. Data were collected from January 2019 to December 2019 with questionnaires and interviews of sampled households.

Our contributions to the field of statistics and economics include the following points: (1) MLR analysis was found to be superior (faster preparation and processing time, lower RMSE, and higher R^2) to the unoptimized ANN method in processing data on the savings of people in the central region of Thailand and (2) deposit interest, bond, share dividends, and other types of investments were the major factors affecting the savings of people in the central region of Thailand in 2019.

2. Literature Review

In 2019, factors affecting the savings behavior of people in Songkhla, Thailand, were studied. The aim of that study was to use MLR to determine factors affecting the savings behavior of people. The research results demonstrated that the factors affecting savings were personal factors, including gender and age. In addition, macroeconomic factors in monetary policy significantly affected the savings behavior of the people [6]. In this same year, economic factors affecting the household savings of people in Thailand were studied based on MLR analysis. The objective was to determine factors affecting household savings. The study showed that the economic factors affecting the household sector savings included inflation, long-term stock funds, and national saving fund [7].

Two years later in 2021, a study was conducted to identify and estimate the main determinants of household saving behavior in rural Ethiopia. The authors analyzed the data using MLR. Their findings suggested that household disposable income, education of household head, number of income earners in the family, and livestock ownership had a statistically significant positive effect on household savings. Similarly, family size, participation in off-farm activities, and distance from the data collection center had a statistically significant negative effect on household savings [8]. Another study based on MLR analysis aimed to investigate factors affecting Thailand's household savings and saving behavior, using data collected from the Household Socio-Economic Survey, 2016, of Thailand's National Statistical Office. The results indicated that household savings were affected by regional factors. Positive factors affecting the cumulative savings were the age of the head of the family, average monthly income, digital expenses, records of income and expenses, and retirement saving plans. Average monthly debt payment and number of family members negatively impacted household savings [9]. In the following year, another study used MLR to investigate the determinants of household savings in a model. Fixed-effect least squares and two-stage least squares estimation procedures in MLR were applied to data from 14 countries spanning the period 2000–2018. The analysis presented some evidence that social security affected the savings significantly but not the interest rate or old-age dependency ratio [10].

In the current year, 2023, household savings and negative interest rates in many countries were investigated. The objective of that study was to analyze the determinants of household savings in a model. An MLR's fixed-effect least squares estimation procedure was used to analyze a set of data from 20 countries in the period of 2000–2020. The analysis provided evidence that the negative interest rates led to a statistically and economically significant increase in savings. The positive effect of income uncertainty and lagged saving rates was smaller with negative interest rates [11].

ANN has been used in many fields of application. Examples of recent studies are prediction of the outbreak of coronavirus disease (2019), prediction of air quality, and prediction of air pollution. An optimized ANN model was developed to predict confirmed cases and deaths from COVID-19. The best prediction performance, in terms of RMSE, R, and MAE, was realized using past 7 days' cases as input variables in the training and testing dataset. The ANN model would be suitable for predicting confirmed cases and deaths of COVID-19 in the time afterward. The predicted confirmed cases and deaths of COVID-19 were very close to the actual confirmed cases and deaths [12]. Another example is the prediction of air quality. ANN was a significant method for protecting public health because it could provide

early warning of harmful air pollutants. The objective of that example was to use ANN and wavelet ANN (WANN) to identify the linear and nonlinear associations between the air pollution index (API) and meteorological variables. The research results demonstrated that WANN ($R = 0.8846$ for Xi'an and $R = 0.8906$ for Lanzhou) performed better than the ANN ($R = 0.8037$ for Xi'an and $R = 0.7742$ for Lanzhou) during the forecasting stage. WANN was effective in short-term API forecasting because it could recognize historic patterns and thereby identify nonlinear relationships between the input and output variables [13].

The final example is the prediction of particulate matter air pollution (PM2.5 and PM10). Fine particulate matter (PM2.5) affects climate change and human health. A study was conducted to use an optimized ANN to predict monthly PM2.5 concentration in Liaocheng, China, from 2014 to 2021. The ANN employed in the study contained a hidden layer with 6 neurons, an input layer with 11 parameters, and an output layer. The ANN achieved a high prediction performance in terms of R (0.9570), MAE (4.6 $\mu g/m^3$), and RMSE (6.6 $\mu g/m^3$) [14]. In the year after, two similar studies were conducted. ANN and WANN were used to predict daily PM2.5 concentration in Shanghai, China. The results show that the optimal input elements for daily PM2.5 concentration-predicting models were the PM2.5 from the previous 3 days and fourteen meteorological elements. It was emphasized that the WANN model obtained optimal prediction results in terms of R (0.9316) [15]. Finally, accurate prediction of air pollution is a difficult problem to be solved in the field of atmospheric research. ANN was exploited to predict hourly PM2.5 and PM10 concentrations in Chongqing, China. Thirteen kinds of training functions to obtain the optimal function were compared. The ANN model exhibited good performance in predicting hourly PM2.5 and PM10 concentrations. The forecast results would support fine management and help improve the ability to prevent and control air pollution in advance, accurately and scientifically [16].

Regarding papers comparing MLR and optimized ANN, in 2020, a paper reported that MLR and ANN models were applied to public spending execution in Peru. The aim of that research was to use MLR and an ANN model with multilayered perceptron to determine the influence of spending execution on the regional government's public budget. The determination coefficient R^2 was 95.9% for the MLR model, which was slightly better than 95.3% for the ANN model. ANN and MLR models obtained very similar results, achieving good models [17]. Another study, a comparative study of MLR analysis and the back propagation ANN method for predicting the financial strength of banks, was conducted in India. The main objective of that study was to forecast the performance of Indian banks. The two methods were compared of their prediction accuracy. Financial data spreading over 10 years from 2010 to 2019 were collected from 19 Indian public sector banks. The data consisted of 17 financial ratios collected from financial statements and other publications of the sampled organizations. Significant ratios that were determinants of the Capital Adequacy Ratio (CAR) were identified by MLR; then, these identified ratios were used as the input for the ANN model. MLR analysis identified 7 financial ratios that had a positive relationship with the dependent variable (CAR). These 7 independent variables were used to predict the financial strength of CAR of the banks. Then, a feed-forward back propagation ANN model was developed with these 7 independent variables to predict the CAR. Finally, the performances of these two methods were compared in terms of MSE, RMSE, and MAPE. The result was that the ANN model scored an improvement of 55.67% in MSE over the MLR model, 33.425% in RMSE over the MLR model, and 99.32% over the MLR model in MAPE [18].

Another study in 2021 compared MLR and ANN in bank performance prediction: a study of 11 Botswanan banks. Return on assets was used as the dependent variable, while management quality, credit risk, liquidity, financial leverage, and capital adequacy were the independent variables. When using MLR, the cost-to-income ratio and the loan loss provision to total loan ratio were found to be the two most significant drivers of bank performance. ANN achieved an R^2 value of 84.37% which was significantly higher than the R^2 value of 70.00% for MLR. ANN also showed a better predictive ability in terms of MAE and MSE [19].

In the last year, 2022, a study compared MLR and ANN predicting BYD stock price. BYD was a leading company in the new energy industry. That paper concluded that the backpropagation ANN had a better explanation ability than the MLR model [20]. Finally, a study investigating factors affecting savings of Gen Y in Bangkok, Thailand, was conducted. One of the study's objectives was to investigate the factors affecting the savings of Gen Y in Bangkok, using quantitative research and survey research methods with a questionnaire for 400 samples, which is selected by simple random sampling. The inferential statistic used was MLR. The results showed that Gen Y saved 4,378.03 baht per month with the major purpose of saving for spending in an emergency situation [21].

All of the abovementioned works that compared MLR and optimized ANN concluded that ANN was better than MLR in terms of MAE. However, for rapid deployment, ANN might not be better since it needed more time to run through a lot of processing rounds to optimize its parameters. Therefore, we wished to compare these two methods in terms of processing time in addition to various prediction error measures.

3. Materials and Methods

3.1. Data Collection. This research is based on secondary data from the National Statistical Office, Government Complex Commemorating Majesty the King's 80th Birthday Anniversary, Ratthaprasasanabhakdi Building, 2nd Floor, Chaeng Watthana Road, Lak Si, Bangkok, Thailand, on the Household Socio-Economic Survey 2019 (Household Income Survey), which was carried out every 2 years. The reason for using these secondary data is that we could not collect primary data nationwide in Thailand ourselves. This is because the research funding that we were able to procure was quite limited. Therefore, the research team had to rely on

secondary data from the public sector of the National Statistical Office, which were publicly available data. The previous research under this project was conducted according to current research standards, which had been updated every 2 years.

Data were collected from January 2019 to December 2019 from a questionnaire distributed to a sample of households. The National Statistical Office had already collected data every two years to track the trend of savings in the country, whether they were increasing or decreasing. The data collected before that were from January 2017 to December 2017, and the data collected after that were from January 2021 to December 2021, which were similar in nature.

The questionnaire had 40 questions, with many types of question, including quantitative and qualitative questions. Initially, more questions were considered for analysis about topics that were expected to affect savings. However, some questions were answered in a variety of ways and could not be categorized into smaller groups, such as questions about occupation. Some questions were not answered completely, causing a large amount of missing values, such as the number of rooms in the households. The research team, therefore, ignored the responses to those questions. The selected questions were suitable to eliminate bias because they do not have much influence on the respondents, such as the number of people gaining income from work, the number of household members, and the number of household members who were not working. The other questions were similar in nature.

The data were collected from January 2019 to December 2019 with a questionnaire to interview people in the central region of 25 provinces of Thailand selected as sample units. The National Statistical Office was sampling using a stratified two-stage sampling. The provinces in the central region were divided into 25 stratums, each of which was divided into 2 substratums according to the characteristics of municipal and nonmunicipal areas. Each substratum has an enumeration area (EA) as the sample unit for the first step and the private household as the sample unit for the second step. In that study, the authors were interested only in the central region comprising 1,902 EA, the population, divided into 900 EA municipal areas and 1,002 EA nonmunicipal areas.

The samples were 1,256 EA in the central region, where 616 EA were municipal areas, and 640 EA were nonmunicipal areas. The total number of private household samples was 15,640, including 9,240 municipal and 6,400 nonmunicipal households. The researchers selected only 11,586 sample households with complete information, divided into 2,191 low-income households savings (Y_1: the first quintile), 6,994 middle-income households savings (Y_2: the second quintile to the fourth quintile), and 2,401 high-income households savings (Y_3: the fifth quintile).

3.2. Variables. The method of selecting independent variables was based on a review of previous literature on variables that affected household savings. Most of the variables in the questionnaire were selected for analysis, except for some variables, such as occupation. This is because there were a lot more occupations in Thailand, exceeding 10. If occupation was converted to a dummy variable, it would have many levels, making it difficult to analyze and interpret the results. Therefore, the research team excluded the occupation variable. A limitation of multiple regression analysis is that the number of independent variables should not be too many, especially qualitative variables, as this can make the interpretation difficult. The research team therefore selected three important qualitative independent variables, excluding occupation. In addition, some of the independent variables were not fully answered by respondents, resulting in a large number of missing values, such as the number of rooms in the households. To conclude, 26 variables were chosen and analyzed based on significance and practicality in order to gain the most accurate information from responders.

The dependent variables in this research consist of three levels: savings of low-income households (the first quintile) in the range of 638–11,037 baht (Y_1); savings of middle-income households (the second quintile to the fourth quintile) in the range of 11,043–37,767 baht (Y_2); and savings of high-income households (the fifth quintile) in the range of 37,771–683,347 baht (Y_3). The twenty six independent variables were twenty three quantitative variables and three qualitative variables, as shown in Table 1.

3.3. Data Partition and Data Analysis. The authors used SPSS Statistical Package to divide the dataset into low-, middle-, and high-income levels. The complete dataset consisted of 11,586 households, divided into 2,191 low-income households, 6,994 middle-income households, and 2,401 high-income households. As a training dataset, 1,557 low-income households (70%) were selected for modeling and 634 households (30%) were used as the testing dataset for the prediction. Similarly, as a training dataset, 4,887 middle-income households (70%) were selected for modeling and 2,107 households (30%) were used as the testing dataset for the prediction. For high-income households, 1,695 households (70%) were selected as the training dataset for modeling and 706 households (30%) were used as the testing dataset for the prediction. The multiple linear regression analysis was carried out using SPSS Statistical Package, and the unoptimized artificial neural network method was carried out using the WEKA package. As a matter of fact, the authors had intended to use WEKA package for both multiple linear regression analysis and unoptimized artificial neural network method, but WEKA was not able to verify some assumptions, such as multicollinearity, autocorrelation, and homoscedasticity. Therefore, we choose to use SPSS to examine the assumptions and analyze the results of multiple linear regression analysis instead of WEKA.

3.3.1. Multiple Linear Regression Analysis. MLR is a relationship analysis of multiple variables. It consists of a quantitative dependent variable and k independent variables. The independent variables may all be quantitative

Table 1: Description of the twenty-six independent variables.

Variables	Description
X_1	Number of people gaining income from work
X_2	Number of household members
X_3	Number of household members who were not working
X_4	Household expenses
X_5	Household consumption expenditures
X_6	Expenses for food, beverage, and household tobacco
X_7	Total household income
X_8	Pension and allowance
X_9	Grants received from other people
X_{10}	Elderly allowance and disabled person, other grants from government and organizations
X_{11}	Income from renting rooms/land and other assets
X_{12}	Deposit interest, bond, share, dividends, and other types of investments
X_{13}	Other income such as lottery money, prize money, commission, and gambling money
X_{14}	Household debt
X_{15}	Sex: sex = 1 is female, sex = 0 is male
X_{16}	Age of household leader
X_{17}	Marital status: marital status = 1 is married, marital status = 0 is single
$X_{18\text{-}1}$	Secondary school and high school education level = 1, other = 0
$X_{18\text{-}2}$	Vocational, diploma, bachelor's and master's degree education level = 1, other = 0
X_{19}	Number of household members under the age of 15
X_{20}	Number of household members 60 years of age and over
X_{21}	Number of disabled persons in the household
X_{22}	Number of members entitled to claim medical expenses from government agencies/state enterprises
X_{23}	Number of members who received health insurance card
X_{24}	Number of members that have a card to certify the right of medical treatment
X_{25}	Number of members receiving other types of benefits
X_{26}	Number of members receiving subsistence allowances for the elderly

Note. The independent variables X_{15}, X_{17}, and X_{18} ($X_{18\text{-}1}$, $X_{18\text{-}2}$) are qualitative variables. The other variables are quantitative variables.

variables or may be both quantitative and qualitative variables. The relationship between the dependent variable and the independent variables is linear [22].

Operationally, MLR is a method for selecting independent variables that significantly affected a dependent variable. The method provides a smaller number of independent variables that significantly affect a dependent variable. A stepwise regression was used to select the independent variables because it was popular. A capability of MLR featured in this research is that it was able to find factors that affect people's savings in the central region of Thailand. The qualitative independent variables had to be converted to dummy variables before the data were imported into the software package. The quantitative independent variables and the dependent variable could be imported directly into it.

A flowchart of the overall MLR investigation and analysis steps is shown in Figure 1. The first step in the flowchart is the assumption checking before constructing the regression. The things that needed to be checked were the correlation among independent variables and the correlation between an independent variable and a dependent variable [22]. The dependent variable was checked for normal distribution using the Lilliefors significance correction [22]. In the case that the dependent variable did not have a normal distribution, it could be corrected by transforming the data using the Box–Cox transformation method [23]. The independent variables were checked for multicollinearity using the variance inflation factor (VIF). If VIF was greater than 10, then the independent variables in the model had multicollinearity [24]. In the case that the independent variables had high multicollinearity, this could be solved by a factor analysis method [22].

Second, data were used to construct a regression equation with a stepwise method [25]. Third, prediction efficiency was compared in terms of root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and processing time.

Finally, assumption was checked after constructing the regression. The residuals were checked for normal distribution using the Lilliefors significance correction [22]. If the residuals did not have a normal distribution, then they could be transformed by the Box–Cox transformation method [26]. The residuals were checked for independence using the Durbin–Watson method. If the residuals were in the range of 1.5–2.5, they were independent [22]. If the residuals were not independent, they could be corrected by taking ρ to adjust the regression model so that there would be no relationship between ε_i and ε_{i-1} [27]. The residuals were checked for equal variance using the graph between the standardized residual and predicted value [28]. In the case that the residuals did not have equal variance, this could be corrected by a weighted least square (WLS) method [22].

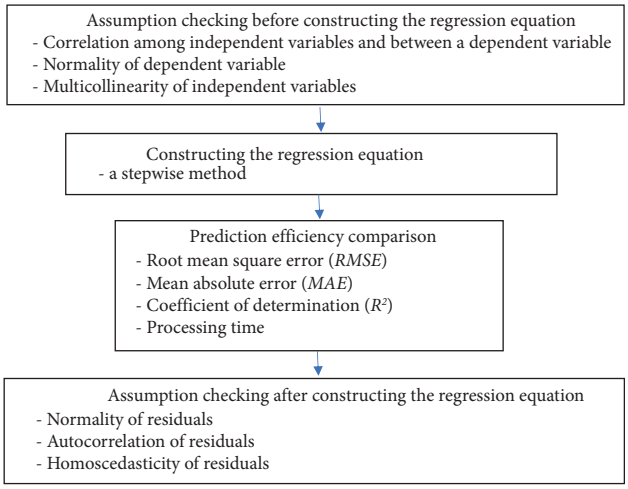

Figure 1: Flowchart of steps of multiple regression analysis.

3.3.2. Artificial Neural Network Method. ANN is a computer technique that simulates human brain with the ability to learn, recognize, and classify things. ANN's processing mimics the function of the brain and transmits information among neurons, with many neuron connections and parallel processing. A backpropagation algorithm is used to teach a multilayer perceptron ANN. Backpropagation algorithm is a popular ANN model because it can solve linear and nonlinear problems [5]. The ANN used in this study worked with a quantitative dependent variable. Many studies have compared ANN to multiple linear regression analysis (MLR) and found that it produced good results.

Operationally, ANN is a method of multiplying the input data by the weights of each input data path. The results from the nodes in each input data path will be passed on to a neuron to combine the data values using a combination function. The neuron will then send the output data to an activation function to adjust the values to the desired range. After that, the output data will be sent out as the input data for neurons connected in the next layer of the neural network. The output node contains the processed output. The importance of ANN in this research is that it had the capability to find factors that affect people's savings in the central region of Thailand. Both the qualitative and quantitative independent variables and the quantitative dependent variable could be imported directly into the software package.

MLR and ANN are generally used as a predictive technique in many fields of application, such as in engineering, science, economic, education, social science, and medicine. More detailed examples are such as the following various kinds of prediction: outbreak of coronavirus disease 2019, bank performance, financial strength of banks, stock price, air quality, air pollution, public spending execution, savings of Gen Y, and household savings behavior. The household savings behavior and factor affecting savings observed in this study would be cross-checked with those observed in several previous studies to find out whether they support this study or not.

A flowchart of the overall ANN steps is shown in Figure 2. First, the structure of ANN was established: the number of hidden layers, the number of nodes in the input layer, hidden layers, output layer, and the type of activation function [29]. Our unoptimized ANN had 27 input nodes according to the number of independent variables, 23 hidden nodes, and 1 output node for predicting the savings of low-income households. It had 27 input nodes, 25 hidden nodes, and 1 output node for predicting the savings of middle-income households and high-income households. Learning rate, momentum, and number of training iterations were set to 0.2, 0.8, and 10,000, respectively [30].

Second, the data were separated into 2 sets. The first dataset was the training dataset, which was 70 percent of the total dataset. It was used to train the given network. The second dataset was the testing dataset, which made up 30 percent of the total data set. It was used to predict the outcome and evaluate the performance of the network [5, 30]. Then, the data were imported into the nodes of the input layer. The data were pushed from the nodes of the input layer to the nodes of the hidden layer. The sum of all nodes was calculated in the hidden layers. The values of the nodes in the input layer were multiplied by the weight of each connection line [31]. The sum of all nodes in the hidden layers was adjusted with a sigmoid activation function [31, 32]. After that, the sum of the nodes in the output layer was computed using the sum function, which multiplies the values of the nodes in the hidden layer by the weight of each connection line [31]. The sum of all data in the output layer was then adjusted with a linear activation function. The errors in the output layer were calculated from the output values, compared to the target values [31]. The weights of the connection lines between the nodes in the output layer were then adjusted so that the training dataset has the minimum RMSE [30] and MAE [33]. Then, the process was repeated until the errors in the output layer reached the specified minimum threshold or the assigned number of iterations.

The RMSE and MAE were calculated for each iteration. Then, the validity of the neural network was verified by applying the weights obtained from the network training to validate the testing dataset. The RMSE and MAE obtained on the training dataset were compared with those obtained on the testing dataset. If the data values were very different, this can be corrected by trying a new weight setting or redesigning the neural network [30]. Prediction efficiency was compared in terms of RMSE, MAE, R^2, and processing time. Finally, the obtained network was used on the testing data to find the predicted value [30].

3.4. Prediction Efficiency Comparison. Multiple linear regression analysis and unoptimized artificial neural network methods were compared of their prediction efficiencies in terms of root mean square error, mean absolute error, coefficient of determination, and processing time.

3.4.1. Root Mean Square Error (RMSE). Root mean square error is based on the same principle as statistical variance. Measuring square root of error with this method yields

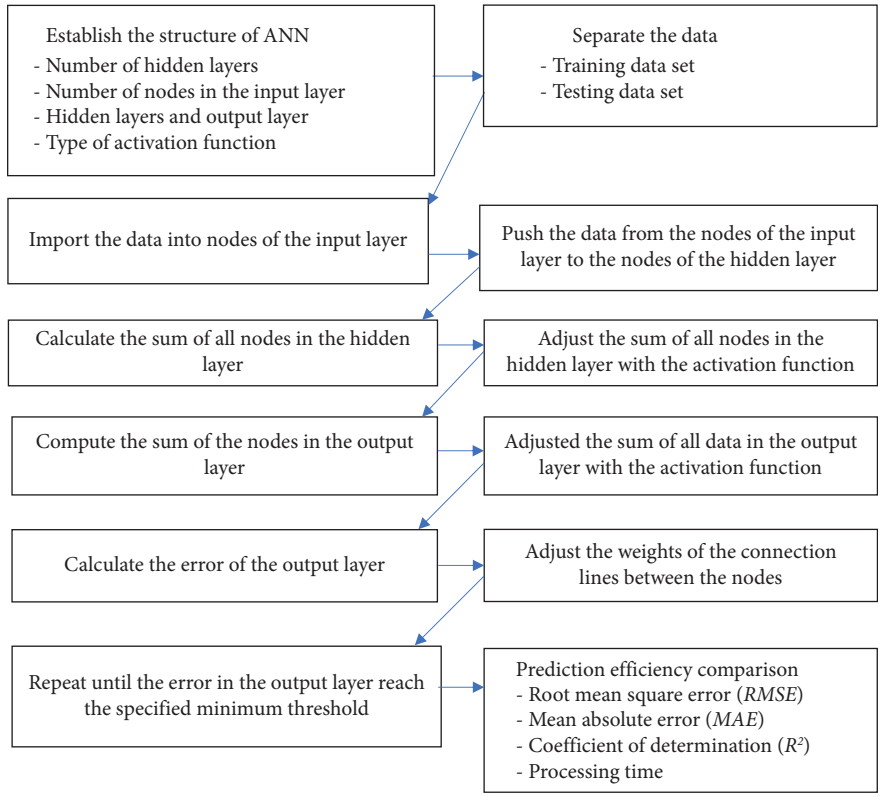

Figure 2: Flowchart of steps of ANN.

a relatively high error since the error is squared at any time before the sum is taken and the mean is calculated. At the end, the square root is removed. The smaller RMSE indicates a more accurate prediction. The formula is as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}, \quad (1)$$

where y_i is actual value and $\hat{y}_i$ is predicted values [30]. Many studies used RMSE for prediction efficiency comparison because the dependent variable was quantitative [34–36]. Therefore, the researchers chose to use RMSE. We also chose to use other measures such as mean square error (MSE). MSE is the same as RMSE, but MSE is obtained by taking the square of RMSE. The simplest method for estimating the accuracy of a model is using MSE. The lower the MSE, the better the model. MSE is a good measure because it consists of both bias and variance [37]. It is a function of estimation error and model complexity (i.e., degrees of freedom) [38].

3.4.2. Mean Absolute Error (MAE). Mean absolute error is the mean of the absolute error of $e_i = |y_i - \hat{y}_i|$. Mean absolute error measures how close the predicted value is to the actual value. The formula is as follows:

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}, \quad (2)$$

where y_i is actual value, $\hat{y}_i$ is predicted value, and n is sample size [33].

3.4.3. Coefficient of Determination (R^2). Coefficient of determination is the proportion or percentage that the independent variables explain the variation in the dependent variable. The formula is as follows:

$$R^2 = \frac{\text{SSR}}{\text{SST}}, \quad (3)$$

where SSR is the variation of Y due to the influence of X_1, X_2, . . ., X_k, and SST is the total variation [39].

3.4.4. Processing Time. Processing time is the time period (in seconds) from the start of processing to the end of processing either ANN or MLR. The simulation running process was manually timed with a digital watch (Solvil et Titus, Switzerland) by the author from start to finish and the measured time interval was taken as the processing time. A shorter processing time is considered more efficient in predicting the savings of people in the central region [40].

3.5. Investigational Steps. A flowchart of the overall investigation steps is shown in Figure 3: data collection, data partition, data analysis, and prediction efficiency comparison. Data collection collected and divided the data into 3 groups of low-, middle-, and high-income household savings, to differentiate among households of different incomes. The first dataset was the training dataset (70 percent of the entire dataset). It was used for training the given method. The second dataset was the testing dataset (30 percent), which was used for predicting the data to compare the performance of the

network. Data analysis consisted of applying MLR and unoptimized ANN to the data. Prediction efficiency comparison was in terms of lower root mean square error (RMSE), lower mean absolute error (MAE), higher coefficient of determination (R^2), and shorter processing time.

4. Results

4.1. Results of the Statistical Data Analysis of the Underlying Variables. The household saving data of people were collected from the secondary data of the central region of Thailand. Descriptive statistics for the savings of low-, middle-, and high-income household are shown, categorizing as statistics of various independent variables and dependent variable in Tables 2 and 3.

Table 2 shows the mean and standard deviation of low-, middle-, and high-income household savings classified by independent variables.

Table 3 shows that sex (X_{15}), marital status (X_{17}), secondary school and high school education levels ($X_{18\text{-}1}$), and vocational, diploma, bachelor's, and master's degree education levels ($X_{18\text{-}2}$) were the qualitative variables. Both X_{15} and X_{17} were nominal scale, while $X_{18\text{-}1}$ and $X_{18\text{-}2}$ were ordinal scale. The frequencies and percentages for low-, middle-, and high-income household savings of these four qualitative variables are shown in Table 3.

Descriptive statistics in the form of histogram for savings of low-, middle-, and high-income households are shown in Figures 4–6, respectively.

Figure 4 shows that the histogram of 2,191 low-income household savings is skewed to the right.

Figure 5 shows that the histogram of 6,994 middle-income household savings is skewed to the right.

Figure 6 shows that the histogram of 2,401 high-income household savings is only slightly skewed to the right.

4.2. Analysis and Results of Multiple Linear Regression

4.2.1. Savings of Low-Income Households. The results on savings of low-income households were obtained from the testing dataset for 634 low-income household savings (the first quintile, 30 percent).

(1) Check the multiple linear regression assumptions before constructing the regression equation.

(1.1) The dependent variable of savings of low-income households had a normal distribution. The savings of low-income household variable was tested for a normal distribution. It was found that the Lilliefors test statistic was 0.381 and p value was ≤ 0.001 ($<\alpha = 0.05$), so it was not a normal distribution (not shown). Therefore, the savings of low-income household variable were transformed using the Box–Cox transformation method. It was found that λ was 0. We, then, chose λ of 0 and used the natural logarithm transformation. After that, it was tested again for a normal distribution. It was found that the Lilliefors test statistic was 0.087 and p value was ≤ 0.001 ($<\alpha = 0.05$), so the savings variable in the natural logarithm was also not a normal distribution. Nevertheless, the central limit theorem stated that if a population did not a normal distribution, and if the random sample size was larger than or equal to 30, then the sample mean had an approximate normal distribution. Here, the sample size is 634, so it is assumed that the savings of low-income household variable had an approximate normal distribution.

(1.2) The independent variables had no multicollinearity.

For the savings of low-income households, the independent variables were checked whether the assumption of multicollinearity, based on tolerance and VIF, was satisfied, as shown in Table 4. Table 4 shows that the VIF of every independent variable was between 1.005 and 2.068, which was less than 10; therefore, every independent variable had no multicollinearity. Hence, the assumptions on the data were validated before the multiple linear regression analysis.

(2) Construct multiple linear regression analysis.

Due to t-test statistic of 8.317 and p value of ≤ 0.001 ($<\alpha = 0.05$), the independent variable X_{12} was correlated with the savings of low-income households when the other independent variables were constant. Similarly, the independent variables X_4, X_6, and $X_{18\text{-}1}$ were correlated to the savings of low-income households with an estimate regression equation as follows:

$$\widehat{Y} = 28{,}581.273 + 140.882X_{12} + 4.564X_4 - 10.163X_6 - 20{,}093.676X_{18\text{-}1}. \quad (4)$$

Factors affecting the savings of low-income households in the order of importance were determined by standardized coefficients beta. The factors were deposit interest, bond, share dividends, and other types of investment (X_{12}), household expenses (X_4), food expenses, beverages and household tobacco (X_6), and secondary school and high school education levels ($X_{18\text{-}1}$). The root of mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) were 91,397.9315, 243,314.6593, and 0.7152, respectively, as listed in Table 5.

The scatter plot shows the existence of a positive relationship between the savings of low-income households and the prediction of the multiple linear regression model in Figure 7.

(3) Check the multiple linear regression assumptions after constructing the regression equation.

(3.1) The residuals had a normal distribution (normality). The Lilliefors test statistic was 0.295, and the p value was ≤ 0.001 ($<\alpha = 0.05$),

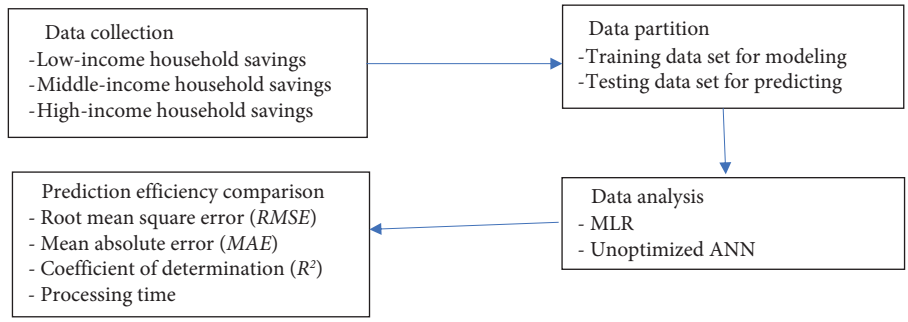

Figure 3: Flowchart of the investigational steps.

Table 2: Descriptive statistics of mean and standard deviation for the savings of low-, middle-, and high-income households.

Variable	Household savings					
	Low-income		Middle-income		High-income	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
X_1	0.85	0.796	1.58	0.884	2.17	1.224
X_2	1.75	0.969	2.63	1.403	3.64	1.752
X_3	0.90	0.884	1.05	1.153	1.47	1.283
X_4	9,042.35	5,117.727	18,925.57	7,772.952	44,175.57	22,537.834
X_5	8,354.60	4.501.227	16,541.55	7,010.194	37,583.34	19,925.162
X_6	3,906.02	1,677.865	6,870.89	2,885.858	11,709.33	5,242.726
X_7	7,747.00	2,340.018	21,790.58	7,377.310	68,005.65	50,449.169
X_8	10.77	177.644	685.60	3,831.239	5,926.65	16,007.368
X_9	1,456.56	2,051.221	1,276.91	3,500.191	1,203.85	5,597.407
X_{10}	618.82	618.856	456.40	1,040.264	458.39	1,343.689
X_{11}	39.56	307.733	64.22	818.332	636.48	9,696.787
X_{12}	19.35	214.954	33.53	218.413	329.93	1,480.741
X_{13}	35.54	252.633	99.81	993.601	169.55	1,085.259
X_{14}	682.46	2,927.860	2,126.11	6,076.712	9,515.80	17,568.616
X_{16}	59.35	15.622	51.37	15.197	54.89	13.066
X_{19}	0.20	0.521	0.39	0.709	0.54	0.812
X_{20}	0.76	0.760	0.53	0.766	0.69	0.841
X_{21}	0.11	0.354	0.06	0.304	0.09	0.532
X_{22}	0.06	0.324	0.17	0.522	0.86	1.237
X_{23}	1.58	1.061	1.87	1.599	1.84	1.849
X_{24}	0.06	0.252	0.55	0.732	0.93	1.089
X_{25}	0.01	0.112	0.01	0.119	0.01	0.092
X_{26}	0.71	0.754	0.45	0.711	0.44	0.686
Y	29,680.46	97,396.755	82,416.67	282,812.632	394,878.35	937,632.709

Table 3: Descriptive statistics of qualitative variables: frequencies and percentages for low-, middle-, and high-income household savings.

Variable	Household savings					
	Low-income		Middle-income		High-income	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
X_{15}						
Male	323	50.9	1,261	59.8	427	60.5
Female	311	49.1	846	40.2	279	39.5
X_{17}						
Single	113	17.8	316	15.0	51	7.2
Married	521	82.2	1,791	85.0	655	92.8
$X_{18\text{-}1}$						
Secondary school and high school	101	15.9	660	31.3	194	27.5
Other	533	84.1	1,447	68.7	512	72.5
$X_{18\text{-}2}$						
Vocational, diploma, bachelor's and master's degree	34	5.4	343	16.3	293	41.5
Other	600	94.6	1,764	83.7	413	58.5

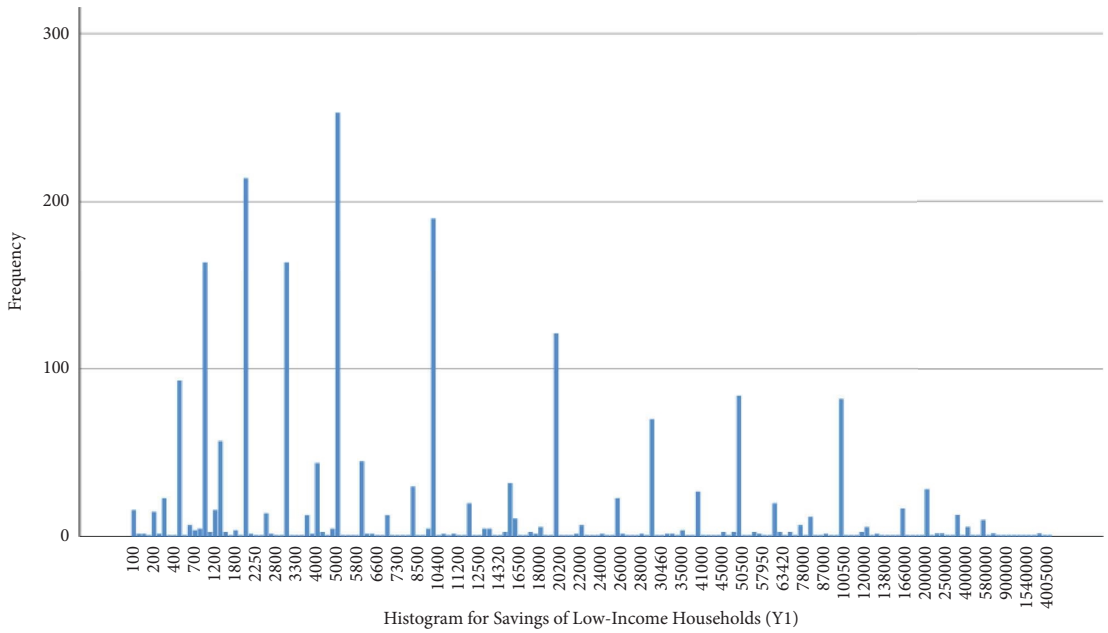

Figure 4: Histogram for savings of low-income households.

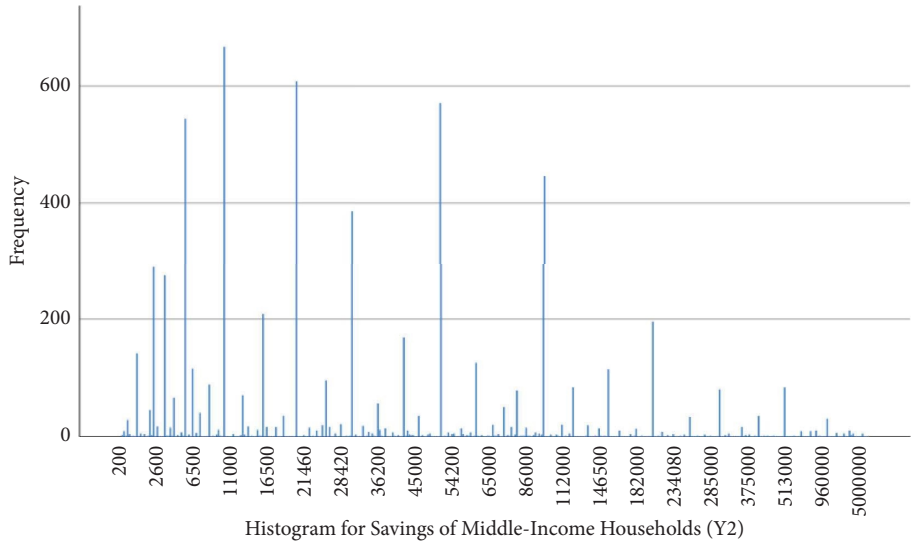

Figure 5: Histogram for savings of middle-income households.

which means the residuals were not normally distributed (not shown). Nevertheless, based on the central limit theorem (CLT) for a large sample size ($n > 30$), it can be assumed that the residuals had a normal distribution.

(3.2) The residuals were not correlated, or they were independent (no autocorrelation). The Durbin–Watson test statistic was 2.008, which was between 1.5 and 2.5. Therefore, the residuals had no correlation, or the residuals were independent.

(3.3) The residuals had equal variance (homoscedasticity). Figure 8 shows that the standard residuals between ± 2 were randomly distributed about zero and parallel to the horizontal axis. Therefore, the residuals had equal variance. Hence, the assumptions on the data were validated after multiple linear regression analysis.

4.2.2. Savings of Middle-Income Households. The results on savings of middle-income households were obtained from the testing dataset for 2,107 middle-income household savings (the second-fourth quintile, 30 percent).

(1) Check the multiple linear regression assumptions before constructing the regression equation.

(1.1) The dependent variable of savings of middle-income households had a normal distribution.

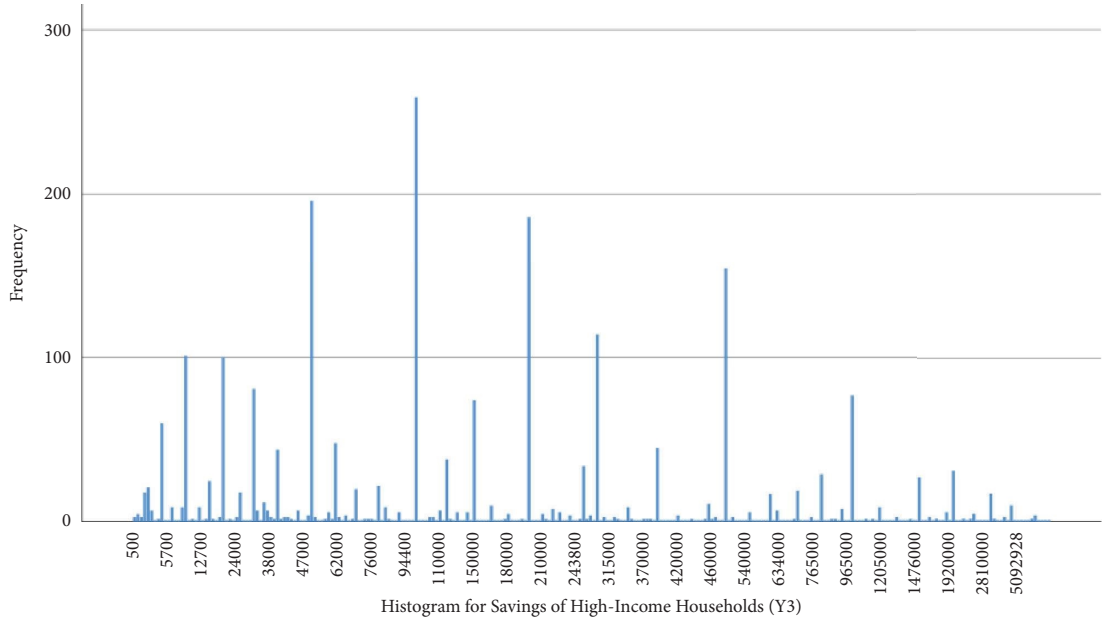

Figure 6: Histogram for savings of high-income households.

Table 4: Unstandardized coefficients B and standard error, standardized coefficients beta, t, p value, collinearity statistics of tolerance, and VIF for savings of low-income households.

Models	Unstandardized coefficients		Standardized coefficients	t	p value	Collinearity statistics	
	B	Std. error	Beta			Tolerance	VIF
Constant	28,581.273	9,342.901		3.059	0.002		
X_{12}	140.882	16.939	0.311	8.317	≤0.001	0.995	1.005
X_4	4.564	1.021	0.240	4.471	≤0.001	0.484	2.068
X_6	−10.163	3.109	−0.1 75	−3.269	0.001	0.485	2.061
X_{181}	−20,093.676	9,973.222	−0.076	−2.015	0.044	0.989	1.011

Table 5: Results of efficiency comparison in predicting savings of people in the central region by multiple linear regression analysis and optimized artificial neural network.

Household savings	*Root mean square error (RMSE)*		*Mean absolute error (MAE)*	
	MLR	Unoptimized ANN	MLR	Unoptimized ANN
Low-income	**91,397.9315**	127,910.8087	243,314.6593	**54,384.4958**
Middle-income	**259,150.5120**	326,001.4820	**58,097.1327**	170,355.9166
High-income	**851,552.7515**	2,779,219.6003	1,926,924.5832	**877,077.2266**
Household savings	Coefficient of determination (R^2)		Processing time (second)	
	MLR	Unoptimized ANN	MLR	Unoptimized ANN
Low-income	**0.7152**	0.6807	**3**	8
Middle-income	**0.7608**	0.6934	**5**	52
High-income	**0.7895**	0.6277	**4**	13

Bold values indicate the better method from the comparison of the two methods.

The savings of middle-income household variable was tested for normal distribution. It was found that the Lilliefors test statistic was 0.386 and the p value was ≤0.001 ($<\alpha = 0.05$), so it was not a normal distribution (not shown). Therefore, the savings of middle-income household variable were transformed using the Box–Cox transformation method. It was found that λ was −0.01, which was close to 0. We, then, chose λ of 0 and used the natural logarithm transformation. After that, it was tested again for a normal distribution. It was found that the Lilliefors test statistic was 0.052 and the p value was ≤0.001 ($<\alpha = 0.05$); so, the savings variable in the natural logarithm was also not a normal distribution. Nevertheless, the central limit theorem stated that if a population did not had a normal distribution, and if the random sample size was larger than or equal to 30, then the sample mean had an

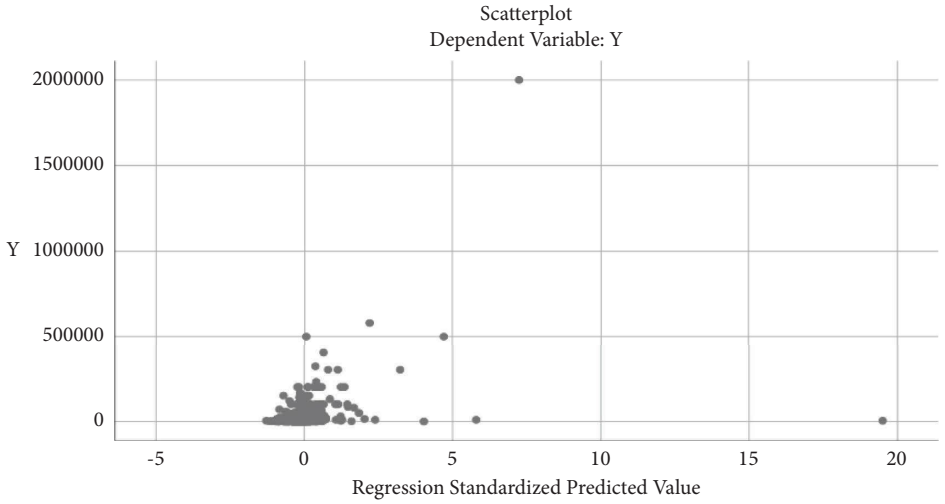

Figure 7: Scatter plot of savings of low-income households versus regression standardized predicted value.

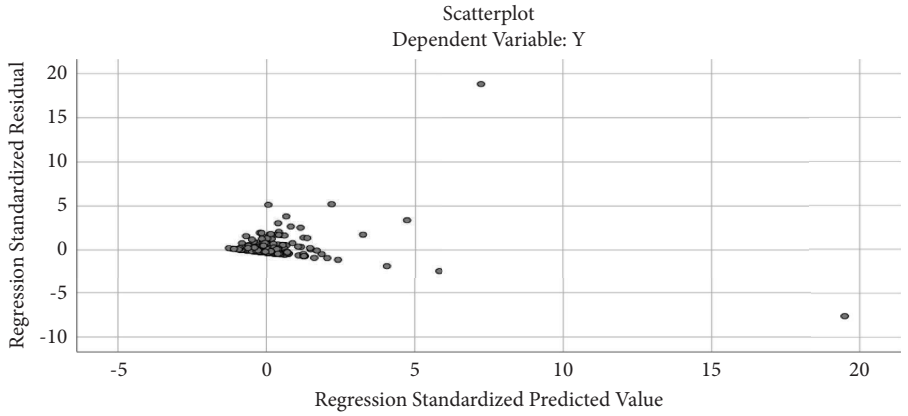

Figure 8: Scatter plot of regression standardized residual versus regression standardized predicted value for savings of low-income households.

approximate normal distribution. Here, the sample size was 2,107. Therefore, it was assumed that the savings of middle-income household variable had an approximate normal distribution.

(1.2) The independent variables had no multicollinearity.

For the savings of middle-income households, the independent variables were checked whether the assumption of multicollinearity, based on tolerance and VIF, was satisfied, as shown in Table 6.

Table 6 shows that the VIF of every independent variable was between 1.010 and 1.974, which was less than 10; therefore, every independent variable had no multicollinearity. Hence, the assumptions on the data were validated before the multiple linear regression analysis.

(2) Construct multiple linear regression analysis.

Due to t-test statistics of 11.93 and p value of ≤0.001 ($<\alpha = 0.05$), the independent variable X_{12} was correlated with the savings of middle-income households when the other independent variables were constant. Similarly, the independent variables X_5, X_8, X_2, X_{16}, X_{11}, X_{20}, X_{14}, and X_{24} were correlated to the savings of middle-income households with an estimate regression equation as follows:

$$\begin{aligned}\widehat{Y} = {} & -72,684.433 + 315.638X_{12} + 9.216X_5 - 6.663X_8 \\ & - 28,810.144X_2 + 1,225.560X_{16} + 19.532X_{11} \\ & + 26,182.367X_{20} - 2.218X_{14} - 18,337.871X_{24}.\end{aligned} \tag{5}$$

Factors affecting the savings of middle-income households in the order of importance were determined by standardized coefficients beta. The factors were deposit interest, bond, share dividends, and other types of investment (X_{12}), household consumption expenditures (X_5), number of household members (X_2), pension and allowance (X_8), number of household members ages 60 and over (X_{20}), age (X_{16}), income from renting rooms/land

and other assets (X_{11}), household debt (X_{14}), and number of members that had a card to certify the right for medical treatment (X_{24}). The root of mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) were 259,150.5120, 58,097.1327, and 0.7608, respectively, as listed in Table 5.

The scatter plot shows the existence of a positive relationship between the savings of middle-income households and the prediction of the multiple linear regression model in Figure 9.

(3) Check the multiple linear regression assumptions after constructing the regression equation.

(3.1) The residuals had a normal distribution. The Lilliefors test statistic was 0.282, and the p value was ≤ 0.001 ($<\alpha = 0.05$), which means the residuals were not normally distributed (not shown). Nevertheless, based on the central limit theorem (CLT) for a large sample size, it can be assumed that the residuals had a normal distribution.

(3.2) The residuals were not correlated, or they were independent. The Durbin–Watson test statistic was 2.001, which was between 1.5 and 2.5. Therefore, the residuals had no correlation, or the residuals were independent.

(3.3) The residuals had equal variance. Figure 10 shows that the standard residuals between ± 2 were randomly distributed about zero and parallel to the horizontal axis. Therefore, the residuals had equal variance. Hence, the assumptions on the data were validated after multiple linear regression analysis.

4.2.3. Savings of High-Income Households. The results on savings of high-income households were obtained from the testing dataset for 706 high-income household savings (the fifth quintile, 30 percent).

(1) Check the multiple linear regression assumptions before constructing the regression equation.

(1.1) The dependent variable of savings of high-income households had a normal distribution. The savings of high-income household variable was tested for normal distribution. It was found that the Lilliefors test statistic was 0.337 and the p value was ≤ 0.001 ($<\alpha = 0.05$), so it was not a normal distribution (not shown). Therefore, the savings of high-income household variable were transformed using the Box–Cox transformation method. It was found that λ was 0.04, which was close to 0. We, then, chose λ of 0 and used the natural logarithm transformation. After that, it was tested again for a normal distribution. It was found that the Lilliefors test statistic was 0.058 and the p value was ≤ 0.001 ($<\alpha = 0.05$), so the saving variable in the natural logarithm was also not a normal distribution. Nevertheless, the central limit theorem stated that if a population did not have a normal distribution, and if the random sample size was larger than or equal to 30, then the sample mean had an approximate normal distribution. Here, the sample size was 706, and it was assumed that the savings of high-income household variable had an approximate normal distribution.

(1.2) The independent variables had no multicollinearity.

For the savings of high-income households, the independent variables were checked whether the assumption of multicollinearity, based on tolerance and VIF, was satisfied, as shown in Table 7. Table 7 shows that the VIF of every independent variable was between 1.036 and 2.378, which was less than 10; therefore, every independent variable had no multicollinearity. Hence, the assumptions on the data were validated before the multiple linear regression analysis.

(2) Construct multiple linear regression analysis.

Due to t-test statistic of 7.215 and p value of ≤ 0.001 ($<\alpha = 0.05$), the independent variable X_{12} was correlated with the savings of high-income households when the other independent variables were constant. Similarly, the independent variables X_{24}, X_{20}, X_7, X_9, X_{14}, X_{26}, and X_3 were correlated to the savings of high-income households with an estimate regression equation as follows:

$$\begin{aligned}\widehat{Y} &= 254,740.696 + 163.838X_{12} - 83,640.060X_{24} \\ &\quad + 268,330.100X_{20} + 2.590X_7 + 20.483X_9 - 5.232X_{14} \\ &\quad - 172,209.800X_{26} - 66,259.688X_3.\end{aligned} \tag{6}$$

Factors affecting the savings of high-income households in the order of importance were determined by standardized coefficients beta. The factors were deposit interest, bond, share dividends, and other types investment (X_{12}), number of household members ages 60 years and over (X_{20}), total household income (X_7), number of members receiving subsistence allowances for the elderly (X_{26}), grants received from other people (X_9), household debt (X_{14}), and number of members that have a card to certify the right of medical treatment (X_{24}), and number of household members not working (X_3). The root of mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) were 851,552.7515, 1,926,924.5832, and 0.7895, respectively, as listed in Table 5.

The scatter plot shows the existence of a positive relationship between the savings of high-income households and the prediction of the multiple linear regression model in Figure 11.

Table 6: Unstandardized coefficients, B and standard error, standardized coefficients beta, t, p value, collinearity statistics of tolerance and VIF for savings of middle-income households.

Models	Unstandardized coefficients		Standardized coefficients	t	p value	Collinearity statistics	
	B	Std. error	Beta			Tolerance	VIF
Constant	−72,684.433	27,885.015		−2.607	0.009		
X_{12}	315.638	26.459	0.244	11.929	≤0.001	0.955	1.047
X_5	9.216	0.924	0.228	9.971	≤0.001	0.760	1.317
X_8	6.663	1.563	0.090	4.264	≤0.001	0.890	1.124
X_2	−28,810.144	4,670.825	−0.143	−6.168	≤0.001	0.742	1.347
X_{16}	1,225.560	519.435	0.066	2.359	0.018	0.512	1.954
X_{11}	19.532	6.934	0.057	2.817	0.005	0.990	1.010
X_{20}	26,182.367	10,359.627	0.071	2.527	0.012	0.507	1.974
X_{14}	−2.218	0.969	−0.048	−2.288	0.022	0.919	1.088
X_{24}	−18,337.871	8,073.654	−0.047	−2.271	0.023	0.913	1.096

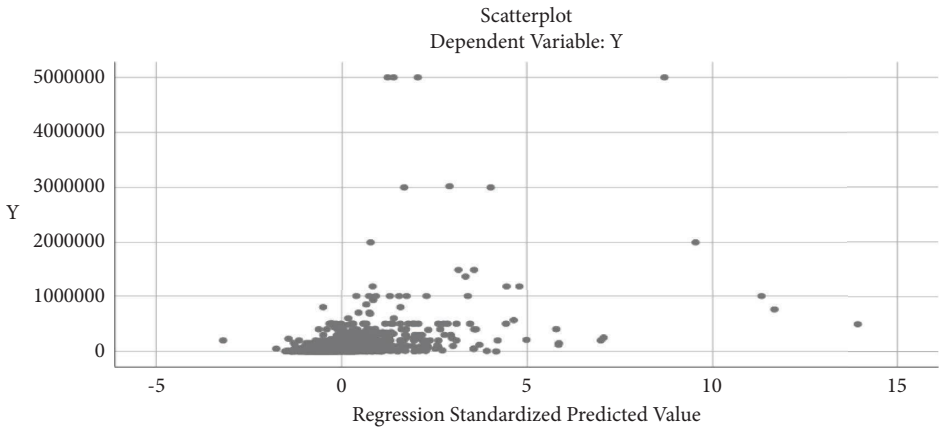

Figure 9: Scatter plot of savings of middle-income households versus regression standardized predicted value.

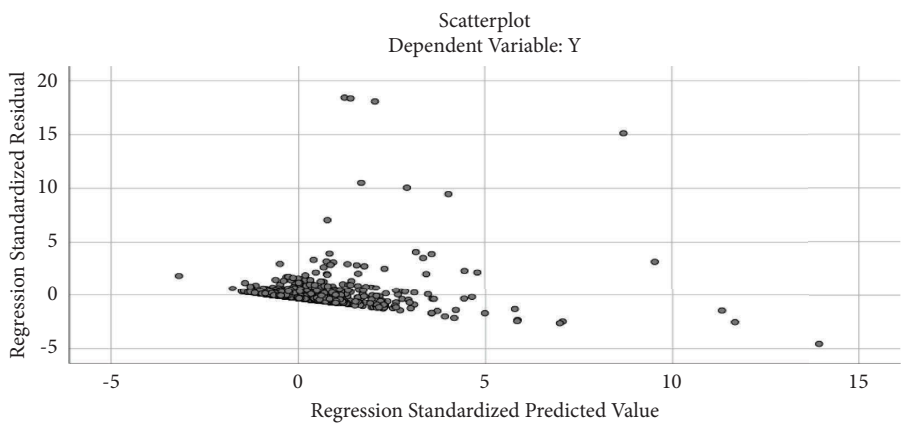

Figure 10: Scatter plot of regression standardized residual versus regression standardized predicted value for savings of middle-income households.

Table 7: Unstandardized coefficients B and standard error, standardized coefficients beta, t, p value, collinearity statistics of tolerance and VIF for savings of high-income households.

Models	Unstandardized coefficients		Standardized coefficients	t	p value	Collinearity statistics	
	B	Std. error	Beta			Tolerance	VIF
Constant	254,740.696	73,080.622		3.486	0.001		
X_{12}	163.838	22.707	0.259	7.215	≤0.001	0.910	1.099
X_{24}	−83,640.060	30,720.395	−0.097	−2.723	0.007	0.919	1.088
X_{20}	268,330.100	58,802.196	0.241	4.563	≤0.001	0.421	2.378
X_7	2.590	0.680	0.139	3.810	≤0.001	0.874	1.144
X_9	20.483	5.831	0.122	3.513	≤0.001	0.966	1.036
X_{14}	−5.232	1.898	−0.098	−2.756	0.006	0.925	1.081
X_{26}	−172,209.800	69,966.559	−0.126	−2.461	0.014	0.447	2.239
X_3	−66,259.688	27,093.433	−0.091	−2.446	0.015	0.852	1.174

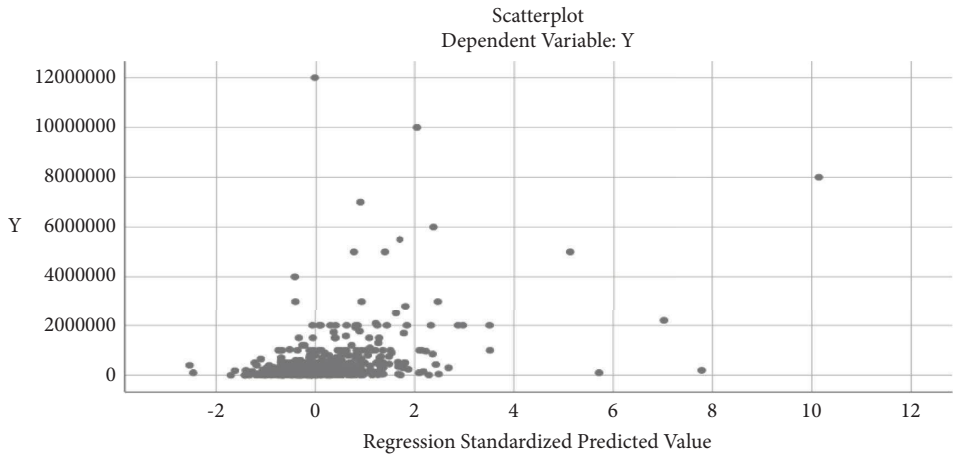

Figure 11: Scatter plot of savings of high-income households versus regression standardized predicted value.

(3) Check the multiple linear regression assumptions after constructing the regression equation.

(3.1) The residuals had a normal distribution. The Lilliefors test statistic was 0.247 and the p value was ≤0.001 ($<\alpha = 0.05$), which means the residuals were not normally distributed (data were not shown). Nevertheless, based on the central limit theorem (CLT) for a large sample size, it can be assumed that the residuals had normal distribution.

(3.2) The residuals were not correlated, or they were independent. The Durbin–Watson test statistic was 1.822, which was between 1.5 and 2.5. Therefore, the residuals had no correlation, or they were independent.

(3.3) The residuals had equal variance. Figure 12 shows that the standard residuals between ± 2 were randomly distributed about zero and parallel to the horizontal axis. Therefore, the residuals had equal variance. Hence, the assumptions on the data were validated after the multiple linear regression analysis.

4.3. Analysis and Results of Unoptimized Artificial Neural Network

4.3.1. Savings of Low-Income Households. The prediction efficiencies of savings of low-income households of unoptimized ANN on many runs are shown in Table 8.

Table 8 shows that the testing dataset for 634 low-income household savings (the first quintile) was 30 percent of the total dataset. The results of the analysis of the testing dataset for low-income household savings, with unoptimized artificial neural network using multilayer perceptron algorithm, were MAE, RMSE, and R^2 of 54,384.4958, 127,910.8087, and 0.6807, respectively.

4.3.2. Savings of Middle-Income Households. The prediction efficiencies of savings of middle-income households of unoptimized ANN on many runs are shown in Table 9.

Table 9 shows that the testing dataset for 2,107 middle-income household savings (the second quintile to the fourth quintile) was 30 percent of the total dataset. The results of the analysis of the testing dataset for middle-income household savings, with unoptimized artificial neural network using

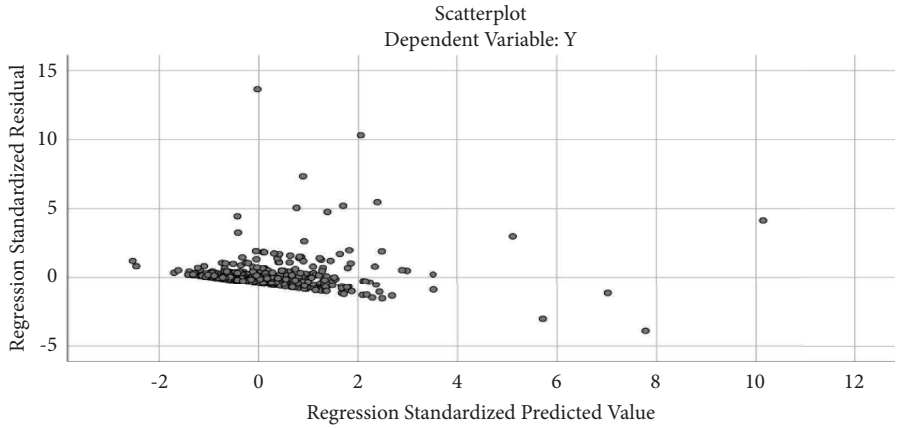

Figure 12: Scatter plot of regression standardized residual versus regression standardized predicted value for savings of high-income households.

Table 8: Mean absolute error, root mean square error, relative absolute error, root relative squared error, and coefficient of determination for savings of low-income households.

Mean absolute error	54,384.4958
Root mean square error	127,910.8087
Relative absolute error	131.7482%
Root relative squared error	130.8888%
Coefficient of determination	0.6807
Total number of instances	634

Table 9: Mean absolute error, root mean square error, relative absolute error, root relative squared error, and coefficient of determination for savings of middle-income households.

Mean absolute error	170,355.9166
Root mean square error	326,001.4820
Relative absolute error	177.1802%
Root relative squared error	115.2931%
Coefficient of determination	0.6934
Total number of instances	2,107

a multilayer perceptron algorithm, were MAE, RMSE, and R^2 of 170,355.9166, 326,001.4820, and 0.6934, respectively.

4.3.3. Savings of High-Income Households. The prediction efficiencies of savings of high-income households of unoptimized ANN on many runs are shown in Table 10.

Table 10 shows that the testing dataset for 706 high-income household savings (the fifth quintile) was 30 percent of the total dataset. The results of the analysis of the testing dataset for high-income household savings, with unoptimized artificial neural network using multilayer perceptron algorithm, were MAE, RMSE, and R^2 of 877,077.2266, 2,779,219.6003, and 0.6277, respectively.

The developed artificial neural network methods for savings of low-, middle-, and high-income households are shown in Figure 13.

Figure 13 shows an input layer and an output layer consisting of 27 input nodes and 1 output node. Figure 13 shows a hidden layer with 23 hidden nodes for low-income

Table 10: Mean absolute error, root mean square error, relative absolute error, root relative squared error, and coefficient of determination for savings of high-income households.

Mean absolute error	877,077.2266
Root mean square error	2,779,219.6003
Relative absolute error	191.8692%
Root relative squared error	296.2293%
Coefficient of determination	0.6277
Total number of instances	706

households. For middle- and high-income households, the number of hidden nodes was 25. The weights of the nodes were placed on the connection lines between input nodes, hidden nodes, and output node.

4.4. Results of Efficiency Comparison in Predicting Savings of People in the Central Region. The prediction efficiencies of MLR and unoptimized ANN on savings of low-, middle-, and high-income households were compared in terms of RMSE, MAE, R^2, and processing time and are as shown in Table 5.

Table 5 shows the efficiency comparison in predicting savings of low-, middle-, and high-income households of MLR and unoptimized ANN. The testing dataset was used to predict the outcomes. It was found that MLR had a lower RMSE, processing time, and a higher R^2 than unoptimized ANN for all savings of low-, middle-, and high-income households. Nevertheless, unoptimized ANN accomplished a lower MAE than MLR for the savings of low- and high-income households.

5. Discussion

In this study, an efficiency comparison of prediction methods of household savings of people in the central region of Thailand and analysis of factors affecting savings of people in the central region of Thailand were conducted, using secondary data on the 2019 Household Socio-Economic Survey, the National Statistical Staff's Household Income Survey. The investigation involved using MLR and

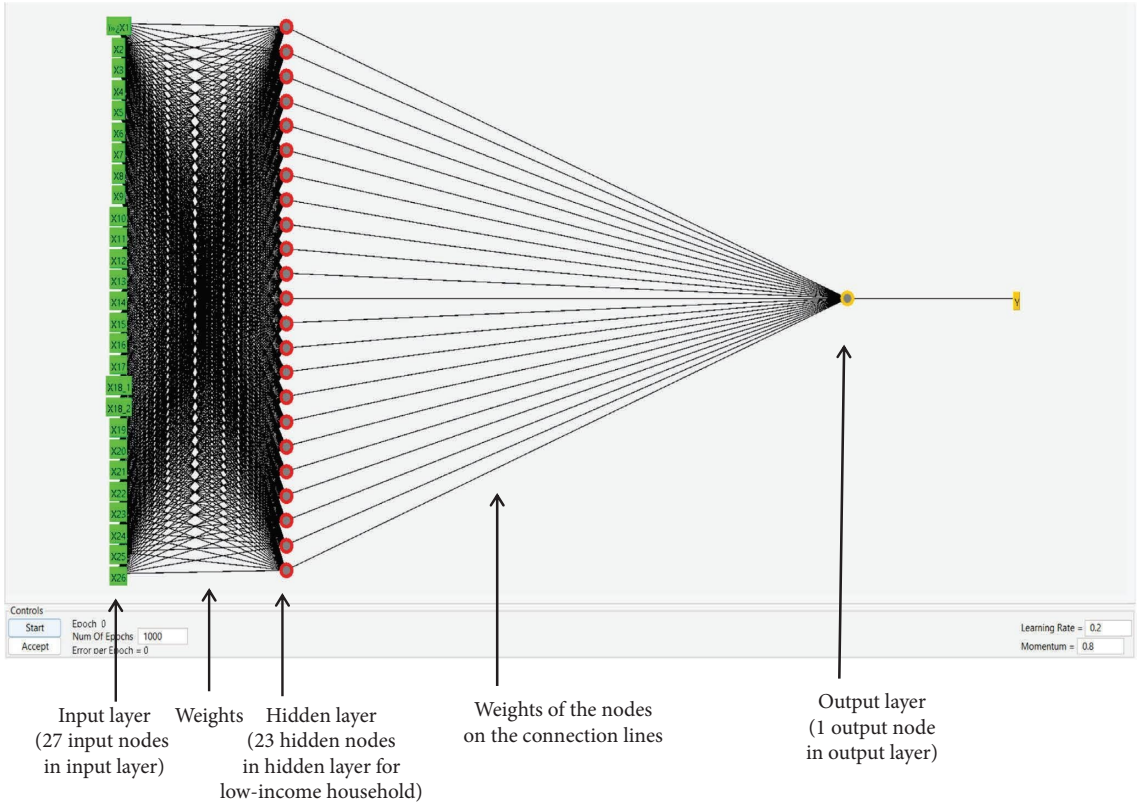

Figure 13: The developed artificial neural network model for low-income households.

unoptimized ANN. Their efficiency comparison was based on RMSE, MAE, R^2, and processing time. Three main topics are discussed as follows:

(1) The results of this study demonstrated that MLR provided a lower RMSE, a shorter processing time, and a higher R^2 than unoptimized ANN on all saving categories of low-, middle-, and high-income households. Nevertheless, the unoptimized ANN provided a lower MAE than MLR for the savings of low- and high-income households. Another study partly confirms the abovementioned conclusion. Using R^2 as the metric, Morales and Huanca [17] applied MLR and ANN to public spending execution in Peru and concluded that MLR was better than ANN. The determination coefficients R^2 achieved was 95.9% for the MLR model and 95.3% for the ANN model. Their conclusion is in agreement with ours but in contrast to several other studies [18–20].

(2) One of the strengths and weaknesses of the ANN model was that it relies on multiple internal parameters. The strength is that these parameters can be evaluated and adjusted to achieve the most accurate prediction. The weakness is that it takes a lot of processing rounds to optimize these parameters. Therefore, for a rapid use like for a screening purpose, an unoptimized ANN may not be as accurate as an optimized ANN or an MLR. In this research, the ANN's internal parameters—the learning rate and momentum—were set to default values, 0.2 and 0.8, respectively [30]. With those default values, the unoptimized ANN did not perform as well as the MLR.

(3) The most important factor affecting savings of low-, middle-, and high-income households was the factor of deposit interest, bond, share dividends, and other types of investment. The investigated factors in this study were similar to the investigated factors in another study in 2018 by [41], that investigated the saving behavior and the factors affecting saving behavior of people in Bangkok. The results showed that return, risk, and promotion affected the behavior, the amount of savings, savings objectives, and savings patterns, which are the same as our conclusion involving deposit interest. In 2019, factors affecting the saving behavior of people in Songkhla Province were investigated by [6]. The results of that study demonstrated that the macroeconomic factors in monetary policy had an effect on the saving behavior of people in Songkhla Province, which is the same as our conclusion involving the factor of bond, share dividends, and investments. Another study [7] investigated the economic factors affecting household saving of people in Thailand. The authors showed that the economic factors affecting the household sector saving included inflation, long-term stock funds, and the national saving fund,

which are the same as our conclusion. In 2021, the study of [9] based on MLR analysis aimed to study factors affecting Thai household savings and saving behavior. The results indicated that household savings were affected by retirement saving plans which are the same as our conclusion involving deposit interest. In the following year [10], another study used MLR to investigate the determinants of household savings in a model. The study concluded that savings were not affected by the interest rate which differs from our conclusion that savings were affected by deposit interest. Finally, in 2023, the study of [11] investigated household savings and negative interest rates in many countries. The result demonstrated that negative interest rate led to a statistically and economically significant increase in savings. This is an interesting apparent conflict with our conclusion that positive deposit rate increased savings.

6. Conclusions

In this paper, the authors were to compare the prediction efficiency of two predictive methods—multiple regression analysis (MLR) and unoptimized artificial neural network (unoptimized ANN)—and to investigate the factors affecting savings of people in central region of Thailand. The comparison of MLR and unoptimized ANN was in terms of RMSE, MAE, R^2, and processing time. In addition, factors affecting the savings of people in the central region of Thailand were analyzed using MLR analysis. The results can be summarized as follows:

(1) For all savings of income categories, low-, middle-, and high-income households, MLR achieved a lower RMSE and processing time as well as a higher R^2 than the unoptimized ANN. However, for the savings of income categories of low-, and high-income households, the unoptimized ANN provided a lower MAE than MLR did. Lower MSE, MAE, and processing time are good, but higher R^2 is good.

(2) The estimated multiple regression equation for savings of low-income households is as follows:

$$\widehat{Y} = 28,581.273 + 140.882X_{12} + 4.564X_4 - 10.163X_6 - 20,093.676X_{18-1}. \quad (7)$$

The estimated multiple regression equation for savings of middle-income households is as follows:

$$\widehat{Y} = -72,684.433 + 315.638X_{12} + 9.216X_5 - 6.663X_8 - 28,810.144X_2 + 1,225.560X_{16} + 19.532X_{11} + 26,182.367X_{20} - 2.218X_{14} - 18,337.871X_{24}. \quad (8)$$

The estimated multiple regression equation for savings of high-income households is as follows:

$$\widehat{Y} = 254,740.696 + 163.838X_{12} - 83,640.060X_{24} + 268,330.100X_{20} + 2.590X_7 + 20.483X_9 - 5.232X_{14} - 172,209.800X_{26} - 66,259.688X_3. \quad (9)$$

The most influential factor affecting savings of low-, middle-, and high-income households was deposit interest, bond, share dividends, and other types of investment (X_{12}).

As for the speed at which the two predictive methods executed, if the data were processed by an optimized ANN, several parameters such as momentum and learning rate had to be adjusted by several processing rounds. The total amount of time for optimizing then running the optimized ANN could be a matter of days. Therefore, we compared the processing time between the MLR and unoptimized ANN, and it was found that unoptimized ANN still took more processing time than MLR. Therefore, for rapid screening purpose, MLR may be better than ANN. The disadvantage of MLR is that the data must conform to the assumptions, whereas artificial neural network method does not require the data to conform to any assumptions [42, 43].

The reliability of MLR is unquestionable since it has been used in various fields of study for a very long time. ANN, on the other hand, is a newer method. Even though ANN has been proven to be reliable in many fields of study recently, it was not as well-established as MLR.

The most important factor affecting savings of low-, middle-, and high-income households was deposit interest, bond, share dividends, and other types of investment.

A limitation of this research was that in the survey of household incomes, some variables might have had missing values and some variables could not even be collected for analysis. In addition, since the survey of household incomes was carried out sporadically, the conclusion from the analysis may not reflect the current situation fully.

6.1. Future Recommendations

(1) Other default parameter values for the unoptimized ANN should be assigned. For example, learning rate, momentum, and number of training iterations should be set to 0.3, 0.7, and 20,000, respectively.

(2) There were various occupations that could not be collapsed into a small number of occupations. If they could be collapsed, then we would add an occupational variable to the data analysis as well.

(3) The research can be extended to cover the people from all regions of Thailand if there is sufficient funding for it.

(4) Other variables related to savings may need to be collected such as economic factors: inflation, long-term stock funds, and the national savings fund.

(5) New methods other than MLR and ANN should be further investigated.

(6) From the results of this study, government agencies should devise a plan to encourage savings for the people in order for them to live a better life in the future.

Data Availability

The data used to support the findings of this study are not restricted by the Ethics Board. The data were available freely from the National Statistical Office, Government Complex Commemorating Majesty the King's 80th Birthday Anniversary, Ratthaprasasanabhakdi Building, 2nd Floor, Chaeng Watthana Road, Lak Si, Bangkok, Thailand, 10210, email address: service@nso.go.th and sanonoi@nso.go.th, telephone: 02-141 7500–03, fax: 02-143 8132, website: https://www.nso.go.th, for researchers who meet the criteria for access to confidential data.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank the committee of the School of Science, King Mongkut's Institute of Technology Ladkrabang for funding this research project (Grant no. 2564-02-05-001). The authors would like to thank a master degree project student in the Department of Statistics for her help in coordinating data collection from her agency.

References

[1] The Royal Institute, *Dictionary*, Nanmebook, Bangkok, Thailand, 2011.

[2] K. Prachompan, T. Thaphiranrak, and S. Kheanamkham, "Factors affecting consumer's behaviour of future expenses of people in Bangkok," M.Sc. thesis in Business Administration, Suan Sunandha Rajabhat University, Bangkok, Thailand, 2018.

[3] P. Khusirisin, "Factors affecting household savings in Muang district, Chinagmai province," M.Sc. thesis in Economics, Chiangmai University, Chiang Mai, Thailand, 2018.

[4] P. Kaewbangwan, "Savings and the theory of consumption," 2008, http://www.fpo.go.th/S-I/Source/ECO/ECO6.htm.

[5] S. Sinsomboonthong, *Data Mining 1: Discovering Knowledge In Data*, Jamjuree Product, Bangkok, Thailand, 2017.

[6] N. Chaisiri, J. Noknoi, and V. Suvanvijit, "Factors affecting savings behaviour of people in Songkhla province," *Journal of Nakbut Periscope*, vol. 11, no. 3, pp. 121–132, 2019.

[7] M. Maytawee, "Economic factors affecting household saving of Thailand," M.Sc. thesis in Business Administration, Rajamangala University of Technology krungthep, Bangkok, Thailand, 2019.

[8] O. Urgessa and B. Alemayehu, "Determinants of household saving behavior in rural Ethiopia: evidence from southwest Shoa zone," *Innovations*, vol. 67, pp. 81–90, 2021.

[9] S. Daenghem and N. Charoenphan, "Factors affecting Thai household savings in the digital economy," *Journal of Demography*, vol. 37, no. 1, pp. 49–68, 2021.

[10] C. Fredriksson and K. Staal, "Determinants of household savings: a cross-country analysis," *International Advances in Economic Research*, vol. 27, no. 4, pp. 257–272, 2021.

[11] K. Staal, "Household savings and negative interest rates," *International Advances in Economic Research*, vol. 29, no. 1-2, pp. 1–13, 2023.

[12] Q. Guo and Z. He, "Prediction of the confirmed cases and deaths of global COVID-19 using artificial intelligence," *Environmental Science and Pollution Research*, vol. 28, no. 9, pp. 11672–11682, 2021.

[13] Q. Guo, Z. He, S. Li et al., "Air pollution forecasting using artificial and wavelet neural networks with meteorological conditions," *Aerosol and Air Quality Research*, vol. 20, no. 6, pp. 1429–1439, 2020.

[14] Z. He, Q. Guo, Z. Wang, and X. Li, "Prediction of monthly PM2.5 concentration in Liaocheng in China employing artificial neural network," *Atmosphere*, vol. 13, no. 8, pp. 1221–1316, 2022.

[15] Q. Guo, Z. He, and Z. Wang, "Predicting of daily PM2.5 concentration employing wavelet artificial neural networks based on meteorological elements in Shanghai, China," *Toxics*, vol. 11, no. 1, pp. 51–19, 2023.

[16] Q. Guo, Z. He, and Z. Wang, "Prediction of hourly PM2.5 and PM10 concentrations in Chongqing City in China based on artificial neural network," *Aerosol and Air Quality Research*, vol. 23, no. 6, pp. 220448–220511, 2023.

[17] J. Morales and J. Huanca, "Regression model and neural network applied to the public spending execution," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 195–200, 2020.

[18] Y. Alsaawy, A. Alkhodre, M. Benaida, and R. A. Khan, "A comparative study of multiple regression analysis and back propagation neural network approaches on predicting financial strength of banks: an Indian perspective," *Wseas Transactions on Business and Economics*, vol. 17, pp. 627–637, 2020.

[19] H. Kablay and V. Gumbo, "Comparison of multiple linear regression and neural network models in bank performance prediction in Botswana," *Journal of Mathematics and Statistics*, vol. 17, no. 1, pp. 88–95, 2021.

[20] Y. Zhao, "A comparative analysis of multiple linear regression models and neural networks for stock price prediction – take BYD as an example," *Advances in Economics and Management Research*, vol. 656, pp. 221–226, 2022.

[21] K. Kumnerdpetch, "Factors affecting savings of gen Y in Bangkok," *Journal of Community Development Research (Humanities and Social Sciences)*, vol. 15, no. 1, pp. 125–136, 2022.

[22] K. Wanichbancha, *Statistic for Research*, Chulongkorn University Press, Bangkok, Thailand, 2007.

[23] J. Net er, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied Linear Statistical Models*, Irwin, Chicago, IL, USA, 1996.

[24] D. C. Montgomery, E. A. Peck, and G. C. Vining, *Introduction to Linear Regression Analysis*, John Wiley and Sons, New York, NY, USA, 2012.

[25] S. Chanabon, *Inferential Data Analysis*, Khonkaen Provincial Public Health Office, Khonkaen, Thailand, 2017.

[26] R. Tansuchart, "Regression model," 2004, http://lms.mju.ac.th/courses/159/locker/Econometrics2/content10.htm.

[27] B. Chaiwichayachat, *Econometrics 1*, Protech, Bangkok, Thailand, 2010.
[28] S. Taesombut, *Regression Analysis*, Kasetsat University, Bangkok, Thailand, 2005.
[29] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Pearson Addison Wesley, Boston, MA, USA, 2006.
[30] S. Sripaaraya and S. Sinsomboonthong, "Efficiency comparison of classification for chronic kidney disease: case study of a hospital in India," *Journal of Science and Technology*, vol. 25, no. 5, pp. 839–853, 2017.
[31] M. T. Hagan, H. B. Demuth, M. H. Beale, and O. D. Jesus, *Neural Network Design*, Matin Hagan, New York, NY, USA, 2010.
[32] T. Prakobphol, "Artificial neural networks," *Journal of Huachiew Chalermprakiet University, Academic*, vol. 12, no. 24, pp. 73–87, 2009.
[33] S. Sinsomboonthong, *Data Mining 1: Discovering Knowledge In Data*, Jamjuree Product, Bangkok, Thailand, 2017.
[34] B. P. Swerpel and L. Paszke, "Application of neural networks to the prediction of significant wave height at selected locations on the Baltic sea," *Archives of Hydro-Engineering and Environmental Mechanics*, vol. 53, no. 3, pp. 183–201, 2006.
[35] H. Z. Abyaneh, "Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters," *Zare Abyaneh Journal of Environmental Health Science & Engineering*, vol. 12, no. 40, pp. 1–8, 2014.
[36] 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.
[37] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley and Sons, New Jersey, NJ, USA, 2005.
[38] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, MIT Press, Cambridge, MA, USA, 2001.
[39] K. Wanichbancha, *Using Windows SPSS to Analyze Data*, Samlada Press, Bangkok, Thailand, 2014.
[40] E. SathiyaPriya and S. M. Venila, "A study on classification algorithms and performance analysis of data mining using cancer data to predict lung cancer disease," *International Journal of New Technology and Research*, vol. 3, no. 11, pp. 88–93, 2017.
[41] L. Hakham, "Saving behaviours of people in Bangkok," M.Sc. thesis in Business Administration, Ramkhamhaeng University, Bangkok, Thailand, 2018.
[42] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann, Amsterdam, The Netherlands, 2006.
[43] D. T. Larose and C. D. Larose, *Data Mining and Predictive Analytics*, John Wiley and Sons, New Jersey, NJ, USA, 2015.