ORIGINAL ARTICLE

Computational statistical modelling for parameters optimization of LDL-cholesterol levels in patients with type 2 diabetes

Wan Muhamad Amir W Ahmad[1*], Farah Muna Mohamad Ghazali[2], Hazik Bin Shahzad[3], Mohamad Nasarudin Adnan[4], Nor Azlida Aleng[5]

ABSTRACT

Introduction

LDL is the acronym used to denote low-density lipoproteins. When the level of LDL is high, it leads to the accumulation of cholesterol in the arteries, and as a result, it is commonly known as "bad" cholesterol. Recognizing the significance, a study is conducted on computational and statistical modelling for parameter optimization of LDL-cholesterol levels. In recent years, there has been an increase in the application of precise statistical analysis methodologies. Consequently, scientists are more focused on producing reliable results and are more determined to provide credible findings.

Objective

This study aims to develop and validate a proposed hybrid method that combines Multilayer Feedforward Neural Network (MLFFNN), and Multiple Linear Model (MLR) and to present the R-syntax applications of the proposed hybrid method with clinical study data.

Material and Methods

The proposed hybrid model will be built using the generalized mixture model, which is formulated using an MLFFNN framework and Linear Model (LM). The performance of the hybrid model will be assessed utilizing the predicted mean square error neural network (MSE.net) and the predicted mean square error (P.MSE), which will serve as a benchmark for precision and efficacy. ***Results:*** The result indicates that hybrid method modelling is superior, with the highest R-squared value and the lowest MSE.net value. The hybrid model technique was found to produce a more precise forecast of the outcome when the data is separated into training and testing datasets. The R-square score in this report demonstrates that the LM model is a good fit (84.97%), and the MSE.net value of 0.00529 indicates that the model is both accurate and predictive.

Conclusion

The research concludes that the hybrid model method proposed is preferable. This critical conclusion helps us comprehend the hybrid method's proportional contribution to this illustration's result.

Keywords

Multiple linear regression; Multilayer feedforward neural network (MLFFNN); Hybrid methodology; Bootstrapping

INTRODUCTION

Cholesterol is a lipid resembling wax, found in every cell of our body. Endogenously synthesised by the liver and exogenously obtained from animal-derived sources such as meat and dairy, this substance is naturally occurring. Although cholesterol is necessary for the body's proper functioning, elevated levels of cholesterol in the bloodstream can raise the likelihood of developing coronary heart disease [1, 2]. In cross-sectional or prospective studies of young or middle-aged people, total and LDL cholesterol levels tend to rise with age. However,

1. Wan Muhamad Amir W Ahmad, School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia. wmamir@usm.my
2. Farah Muna Mohamad Ghazali, School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia. farahmuna@student.usm.my
3. Hazik Bin Shahzad, School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia. hazikshahzad@hotmail.com
4. Mohamad Nasarudin Adnan, School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia. nasarudinadnan@student.usm.my
5. Nor Azlida Aleng, Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu (UMT), 21030 Kuala Nerus, Terengganu, Malaysia. azlida_aleng@umt.edu.my

Correspondence

Wan Muhamad Amir W Ahmad, School of Dental Sciences, Health Campus, Universiti Sains Malaysia, 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia
Email: wmamir@usm.my

Bangladesh Journal of Medical Science Vol. 23 No. 02 April'24

DOI: https://doi.org/10.3329/bjms.v23i2.72191

cross-sectional and prospective studies of patients older than 65 years have shown that total and LDL cholesterol levels decrease with age [3, 4, 5, 6]. According to Borel et al. [7], there is an association between the waist circumference ratio, total cholesterol, LDL, and triglyceride levels. According to Hernández-Reyes' study in 2020 [8], individuals with a waist circumference of 100 cm or more had notably higher levels of TC, LDL-C, non-HDL-C, and triglycerides and lower levels of HDL-C. The increases in total cholesterol were due to increases in LDL cholesterol [9]. In cross-sectional studies, the concentration of plasma triglycerides is inversely related to the size of low-density lipoprotein (LDL) particles. Changes in LDL size were significantly ($p < 0.0001$) correlated with reductions in triglyceride levels [10]. Postmenopausal women often have increased LDL levels and decreased HDL values. At the current levels, estrogen medication seems to prevent this drop in HDL [11]. Postmenopausal estrogen's effects on heart disease, cancer, and osteoporosis are still being discussed [12, 13]. In middle-aged women Total cholesterol and low-density lipoprotein cholesterol (LDL-C) concentrations are associated with cardiovascular events and total mortality in middle-aged women and men [4, 14, 15]. As per the statement, the percentage of Canadians who had unhealthy levels of LDL (bad) cholesterol showed a considerable rise with increasing age. It was less than 6% in the age group of 6 to 19 years, but it increased to 12% for individuals aged 20 to 39 years and 40% for those aged 40 to 59 years. Researchers are now developing hybrid models to increase the precision of data forecasts and minimize the shortcomings of single models. As a technique for prediction models, hybrid approaches are gaining popularity in the present day, and currently, the use of statistical forecasting tools is on the rise. To overcome the disadvantages of standalone models, hybrid models are required since they provide more accurate and predictable outcomes. According to Adnan's research in 2021 [16] and Li et al's study in 2022, the hybrid prediction model was discovered to be more precise than conventional prediction models when compared to the standalone model. The application of multiple linear regression (MLR) in research permits the evaluation of the impact of several independent variables on a dependent variable. Linear regression is employed to formulate a model function (consisting of response and predictors) to represent the linear relationship between two or more variables. The objective of linear regression analysis is to discover whether the relationship between independent and dependent variables is increasing or decreasing. In statistical linear modeling, response variables (dependent variables) and predictor variables are used (independent variables). Another usefulness of this study is to create a hybrid methodology by combining specific statistical tools using R-syntax. This hybrid methodology includes the bootstrapping, splitting, regression modeling, and validation of derived models utilizing MLFFN. The derived model should be accurate and validated. MLFFNN is a known neural network for providing validity to models [17]. The combined methodology will increase the accuracy of the results and the precision of the models. Hence, the study aims to develop a hybrid methodology using qualitative predictors in regression analysis.

METHODOLOGY:

Figure 1 depicts the conceptual structure of the suggested technique. Sections data sources, study design, computational biostatistics modelling, bootstrapping (case resampling technique), multiple linear regression and multilayer feedforward neural network provide a comprehensive description.

Data Sources

This investigation analyzed information from patients who attended the ambulatory clinic at the Hospital Universiti Sains Malaysia (USM). The Universiti Sains Malaysia Research Ethics and Human Research Committee (USM/JEPeM/20090462) approved the study, and patient confidentiality and medical status were maintained. The study included 97 patients who had type 2 diabetes mellitus, and it had the advantage of exploring a model that takes into account clinically significant factors. A bootstrapping method will be created after the data is processed. Table 1 presents a synopsis of the selected variable's data in the study.

Table 1. Data description of the selected blood profile

Variable	Description
LDL	Low-density lipoprotein
Age	Age in Year
Wc	Waist reading
Tc	Total of cholesterol
Tg	Triglycerides reading
HDL	high-density lipoprotein

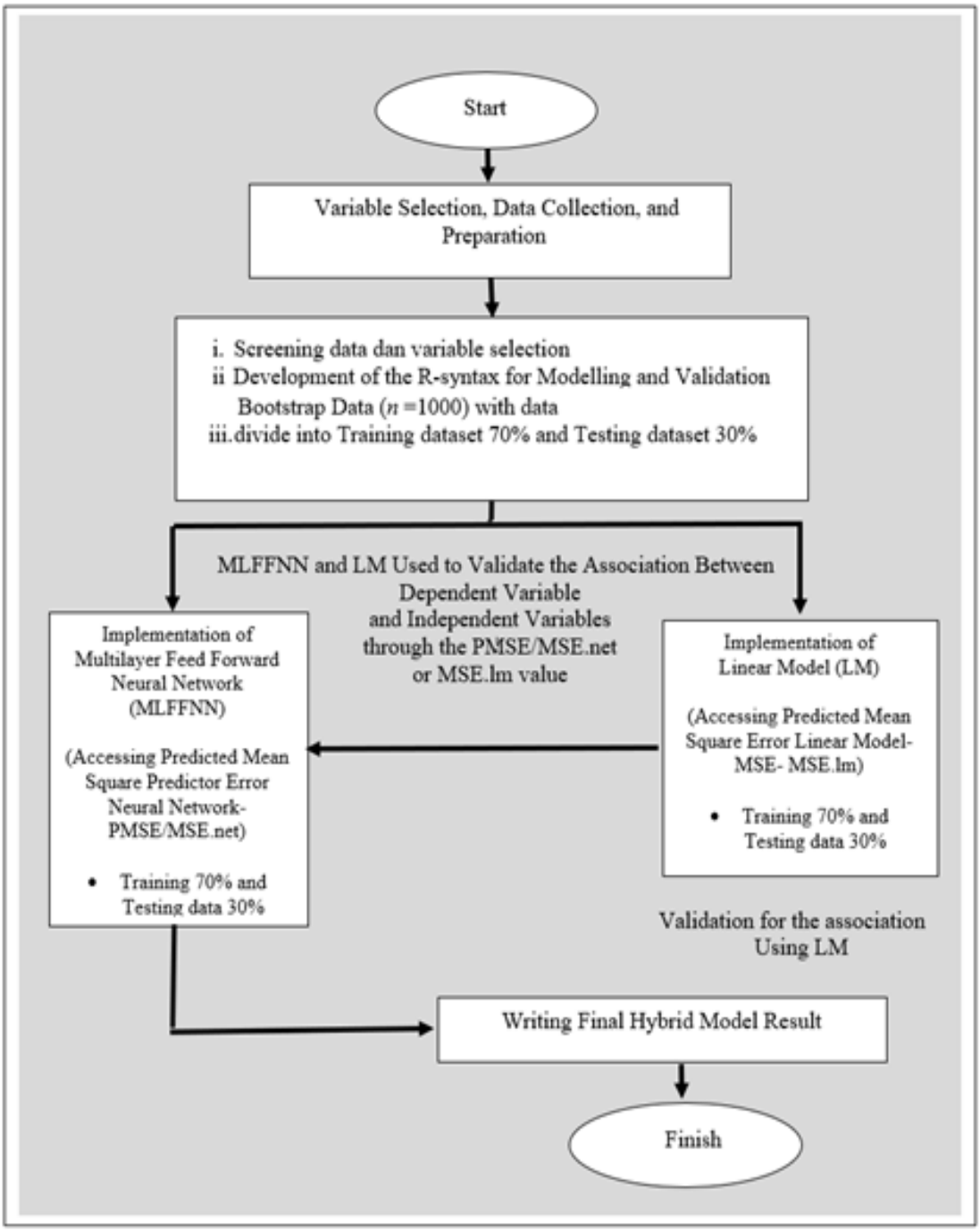

Figure 1. The conceptual framework of the proposed methodology

The Study Design

This study employs a computational biostatistical study design. This study design involves the (i) bootstrap technique to improve the parameter estimates (ii) Multilayer Feed-Forword Neural Network (MLFFNN) and (iii) Multiple Linear Regression. Figure 1 shows the conceptual framework of the development method. R-software is employed as a research tool in this research. While the validation of derived models is done using MLFFNN.

Computational Biostatistics Modeling

The R software is utilized for the purpose of composing syntax and conducting data analysis. The R-syntax was developed by integrating various statistical techniques such as the bootstrap method, multiple linear regression, multilayer feedforward neural network, and fuzzy linear regression. The utilization of this hybrid approach has the potential to aid researchers in the enhancement of their models and outcomes with greater accuracy and precision, as noted by Sharkawy (2020) [18] and Wynants (2015) [19]. The present investigation involves the execution of modelling and validation of the model through the process of partitioning the data set into two distinct sets, namely the training dataset (70%) and the testing dataset (30%). The statistical model shall be constructed through the utilization of training data, while the evaluation of the model's efficacy shall be conducted through the application of testing data.

Bootstrapping (Case Resampling Technique)

Bootstrapping is a basic statistical interference approach that involves repeatedly resampling a sample to build a statistical sample distribution. Efron suggested a potential alternative computer technology known as the bootstrapping methodology in 1979 [20]. The bootstrap does not generate a new sample; instead, it replaces the population's current data values and draws simulated samples from inside the sample. This is accomplished by taking a sample (through replacement) and building a larger sample "made of case resampling" also known as bootstrap samples. In this study, bootstrapping increases the regression model's predictability and precision.

Multiple Linear Regression

Linear regression is a statistical technique that models the linear association between two variables: the response variable and the predictor variable. It allows for an analysis of how the explanatory variables affect the response variable and provides predictions of the dependent variable's value as the independent variable's value changes. As per Adnan et al.'s research in 2021 [16] and Ahmad et al.'s study in 2016 [21], the linear regression method is utilized for this purpose. LDL represents the response variable (dependent variable) in the linear relationship with n explanatory variables (independent variables), which are defined $x_1, x_2, x_3, x_4, x_5, x_6$ through $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ by the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon$. Therefore, the model can be described as where LDL is the dependent variable; x_n is the independent variable; β_0 is the intercept, β_n is the regression coefficient, and ε the error term, also referred to as the statistical error.

The regression equation may be used to determine the direction and magnitude of the independent variable's x influence on the dependent variable y in addition to forecasting the value of the dependent variable y. Let's say that we are thinking about an experiment with k independent variables and a sample size n of observations. The R-syntax in R-Studio was created utilizing the methodologies of bootstrap, multiple linear regression, and multi-layer feed-forward neural networks. Training data (for modelling) and testing data (for validation) are the two categories into which the data is separated. To look into the relationship between the total number of instances and the selected explanatory variables, linear regression models are built.

$$LDL = \beta_0 + \beta_1(Age) + \beta_2(Wc) + \beta_3(Tc) + \beta_4(Tg) + \beta_5(Hdl) + \beta_6(Alp) \quad (1)$$

where,

$\beta_0, \ldots, \beta_6$ are the coefficient parameters,

Age, Wc Tc, Tg, Hdl, and *Alp* are the independent variables

LDL is the dependent variable of the reading of low-density lipoprotein

measurement.

The present study utilizes the Maximum Likelihood Estimator (MLE) approach to determine the parameters of a regression model. R. A. Fisher proposed MLE as a point estimation method in 1922, as indicated in Ghazali et al.'s research in 2020 [22] and Alexopoulos' study in 2010 [23]. MLE is a statistical method that identifies the function that is most likely to explain the observed data.

Multilayer Feedforward Neural Network (MLFFNN)

The present study will employ the Multi-Layer Feed-Forward Neural Network (MLFFNN) method, which is widely recognised as the most commonly utilised form of artificial neural network. As per Ahmad et al.'s (2021, 2022) research [17, 24], the MLFFNN generally consists of three fundamental layers, namely the input

layer, the hidden layer, and the output layer. The output node in this research is held constant at a value of one due to the presence of a solitary dependent variable. The MLFFNN with N input nodes, H hidden nodes, and one output node is described by Equation (1). The values $\hat{y}$ are presented in the following manner.

$$\hat{Y} = g_i\left(\sum_{j=1}^{H} w_j h_j + w_0\right) \quad (2)$$

where w_j an output weight from hidden node j to the output node, W_0 is the bias for the output node, and g is an activation function. The values of the hidden node h_j, j =1…H are given by;

$$h_j = g_i\left(\sum_{j=1}^{H} v_{ji} x_i + v_{j0}\right) \quad (3)$$

where v_{ji} the *weight* of the output from input node i to hidden node j, v_{j0} serves as the bias for hidden node j, where j ranges from 1 to H. The independent variables are represented by x_i, where i ranges from 1 to N. An activation function denoted by k is utilised in this context, as per the works of Ahmad et al. (2021; 2022) [17, 24]. Figure 2 depicts the overall structure of the MLP model.

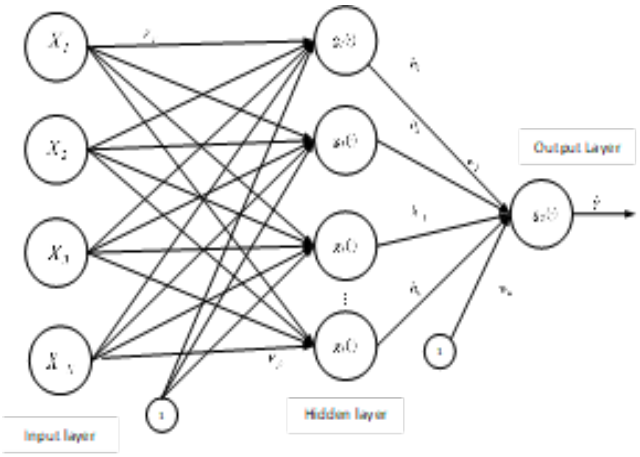

Figure 2. The general architecture of the MLFFNN with one hidden layer, N input nodes, H hidden nodes, and one output node.

RESULTS

The objective of this study is to assess the effectiveness of a Multilayer Layer Feed Forward Neural Network (MLFNN) utilizing a linear activation function. This research investigates both the training and testing datasets. The selection of the most optimal multiple linear regression model was based on the identification of variables that generated the smallest Predicted Mean Square Error, which was computed using the multiple linear regression algorithm.

Table 2. Result of Hybrid Linear Regression Model

Model	Unstandardized Coefficients		t	p-value
	B	Std. Error		
(Intercept)	0.782531	0.0987209	7.927	6.02e-15 ***
Age	-0.006304	0.0012019	-5.245	1.91e-07 ***
Wc	-0.002445	0.0005546	-4.409	1.15e-05 ***
Tc	0.777554	0.0109904	70.749	< 2e-16 ***
Tg	-0.284475	0.0208545	-13.641	< 2e-16 ***
Hdl	-0.471913	0.0422559	-11.168	< 2e-16 ***
Alp	-0.001250	0.0003892	-3.211	0.00136 **

Dependent Variable: LDL

R^2 : 0.8559, [F(df) = 983.2 (6, 993); p < 2.2e-16], MSE.lm =0.087

Multiple Linear Regression. Model Assumption is met.

*Sig. codes: 0 '***' 0.001 '**'*

Figure 3 depicts the architecture of the MLFFNN. The neural network in question comprises a single concealed layer consisting of four neurons, one output node, and six input nodes. The phase involved the utilization of the MLR technique for the purpose of variable selection. Six factors are Age (β_1= -0.006304; Std SE= 0.0012019; p < 0.01), Wc (β_2= -0.002445; Std SE= 0.0005546; p < 0.01), Tc (β_3= 0.777554; Std SE= 0.0109904; p < 0.01), Tg (β_4= -0.284475; Std SE= 0.0208545; p < 0.01), Hdl (β_5= -0.471913; Std SE= 0.0422559; p < 0.01), Alp (β_6= -0.001250; Std SE= 0.0003892; p < 0.01) reading have significantly influenced the LDL. This indicates that, the input variable,

The objective of this study is to examine the

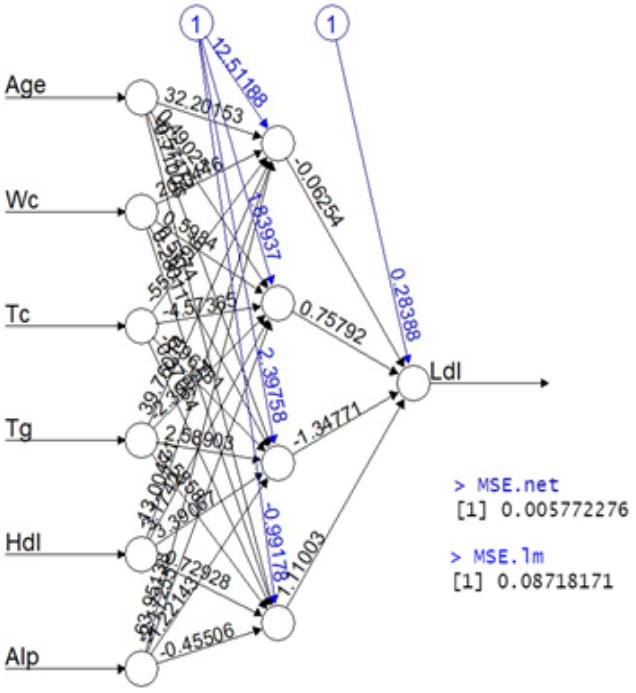

Figure 3. The architecture of the MLFFNN with six input nodes, one hidden layer, and one output node

effectiveness of Multi-Layer Feedforward Neural Networks (MLFFNN) in light of the outcomes obtained from Multiple Linear Regression (MLR). The MLR could be interpreted as evidence supporting the chosen element in this case. A smaller value indicates a greater impact on the study. For regression modeling, the choice of variables that yields the smallest MSE will be deemed optimal. In this study, we proposed the MLFFNN, which has six inputs, one hidden layer, and a single output. These six variables produce a small MSE.net, indicating that the six independent variables studied contribute significantly to LDL. The LDL level is shown by the output node being set to one (a dependent variable). For the test of the train, the ratio is 70:30. Seventy per cent of the available data is used to train the network, and thirty per cent is used to test the network [24, 25]. MLFFNN's performance was judged using the testing/out-of-sample MSE.net. The Predicted Mean Squared Error (MSE.net) shows how far off our predictions are from the actual data. Table 2 is a summary of the results of the multiple regression modeling (MLR). The models are illustrated below. Therefore the proposed linear model is given by

LDL =0.782531 -0.006304 (Age) -0.002445 (Wc)+ 0.777554(Tc)- 0.284475 (Tg) -0.471913(Hdl) -0.001250(Alp) (4)

Equation 4 presents multiple linear models of LDL levels. The results indicate a notable relationship between age, waist circumference, total cholesterol, triglycerides, and high-density lipoprotein level with LDL levels in a certain direction.

DISCUSSION

The discussion of this proposed model may be focused on the technique's development; and the discussion of the results, which contribute to the study's conclusions. The hybrid model and technique improve the precision of the analysis by providing extra information regarding the model fitting and the model prediction. Higher accuracy can be seen through the error obtained from the selected model. This method can be used as an alternative method to regression modeling [24, 25]. The second discussion that can be obtained from the analysis is on the result obtained. The variable utilized in this modeling approach is chosen based on its clinical significance and the opinion of a clinical expert. Because the high volume of LDL is related to several diseases, the LDL value can be a potential tool for the primary indicator of health [10]. For public health professionals, predictive modeling is essential for predicting the evolution of LDL as healthcare costs continue to rise as a result of the prevalence of non-communicable chronic diseases. This research will aid medical professionals in spotting patients who are at risk for developing chronic diseases. The suggested approach helps to achieve the best level of predicting accuracy [27]. Furthermore, this approach can help policymakers and health professionals improve existing preventive measures, which can in turn help establish a new policy. Diagnostic parameters that are both accurate and reliable can be generated by the developed predictive models using Bootstrap, MLFFNN, and multiple linear regression. Preventative measures can be derived from these models.

CONCLUSION

The primary objective of this research is to create a hybrid model using MLFFNN and MLR techniques. R's syntax provides a basis for coordination and consistency across an application. The training dataset was composed of 70% of the data, while only 30% was included in the testing dataset. The MLFFNN

model was implemented on a single hidden layer. The mean MSE.net was calculated for both MLFFNN and MLR in this study. MSE.net is a measure of network efficiency used in various statistical settings. To achieve the best possible MSE.net value, the output of the MLR model was fed into the MLFFNN model. This ensures that the resulting model is the most accurate prediction model possible. The main challenges include selecting appropriate input parameters, data preparation, and data standardization to enable integration into the ANN. By using these techniques, the study found that age, waist circumference, total cholesterol, triglycerides, and high-density lipoprotein have a significant impact on low-density lipoprotein. Existing literature suggests a linear correlation between LDL levels and age, waist circumference, total cholesterol, triglycerides, and high-density lipoprotein.

Acknowledgement

The authors express gratitude to Universiti Sains Malaysia (USM) for providing financial support for the research under grant number FRGS/1/2022/STG06/USM/02/10.

Conflicts of Interest:

The authors affirm that they do not have any conflicts of interest.

Ethical Approval:

Approval for the study was granted by the Human Research Ethics Committee USM (HREC) under the JEPeM Code: USM/JEPeM/20090462. Stringent protocols were observed to safeguard patient confidentiality and maintain their medical status.

Author's Contribution:

Data gathering and idea owner of this study: WMAWA, FMMG, HBS, MNA, NAA.

Study design: WMAWA, FMMG, HBS, MNA, NAA.

Data gathering: WMAWA, FMMG, HBS, MNA, NAA.

Writing and submitting a manuscript: WMAWA, FMMG, HBS, MNA, NAA.

Editing and approval of final draft: WMAWA, FMMG, HBS, MNA, NAA.

References

1. Abu Kholdun Al-Mahmood, Aziz Al-Safi Ismail, FA Rashid, Geoff Gill. (July 2007). Effect of Therapeutic Lifestyle Changes on Insulin Sensitivity of Non-obese Hyperlipidemic Subjects: Preliminary Report. *Journal of Atherosclerosis and Thrombosis*, 14(3):122-7. DOI: 10.5551/jat.14.122
2. Abu Kholdun Al-Mahmood, Aziz Al-Safi Ismail, FA Rashid, Wan Mohamed. (June 2006). Isolated Hypertriglyceridemia: An Insulin-Resistant State with or without Low HDL Cholesterol. *Journal of Atherosclerosis and Thrombosis*, 13(3):143-8.DOI: 10.5551/jat.13.143
3. Al-Mahmood, A. K. SF Afrin, N Hoque. (2013). Metabolic Syndrome and Insulin Resistance: Global Crisis. *Bangladesh Journal of Medical Biochemistry*, 4(1). DOI: 10.3329/bjmb.v4i1.13779
4. Genest J., McPherson R., Frohlich J.(2009). Canadian Cancer Society/Canadian Guidelines For The Diagnosis And Treatment Of Dyslipidemia And Prevention of Cardiovascular Disease In The Adult. *Canadian Journal of Cardiology*. 10:567-79
5. Goel, S., Garg, P. K., Malhotra, V., Madan, J., Mitra, S., & Grover, S. (2016). Dyslipidemia in Type II Diabetes Mellitus - An assessment of the main lipoprotein abnormalities. *Bangladesh Journal of Medical Science*, 15(1), 99–102. https://doi.org/10.3329/bjms.v15i1.21170
6. Ferrara, A., Barrett-Connor, E., & Shan, J. (1997). Total, LDL, and HDL cholesterol decrease with age in older men and women:

The Rancho Bernardo Study 1984–1994. *Circulation*, *96*(1), 37-43.

7. Borel A.L., Coumes S., Reche F, Ruckly S., Pépin J.L., Tamisier R., et al (2018). Waist, neck circumferences, waist-to-hip ratio: which is the best cardiometabolic risk marker in women with severe obesity? The SOON cohort. *PloS one*.13(11):e0206617.
8. Hernández-Reyes, A., Vidal, Á., Moreno-Ortega, A., Cámara-Martos, F., & Moreno-Rojas, R. (2020). Waist circumference as a preventive tool of atherogenic dyslipidemia and obesity-associated cardiovascular risk in young adults males: a cross-sectional pilot study. *Diagnostics*, 10(12), 1033.
9. Schonfeld, G., Patsch, W., Rudel, L. L., Nelson, C., Epstein, M., & Olson, R. E. (1982). Effects of dietary cholesterol and fatty acids on plasma lipoproteins. *The Journal of Clinical Investigation*, 69(5), 1072-1080.
10. McNamara, J. R., Jenner, J. L., Li, Z., Wilson, P. W., & Schaefer, E. J. (1992). Change in LDL particle size is associated with change in plasma triglyceride concentration. *Arteriosclerosis and Thrombosis: A Journal of Vascular Biology*, *12*(11), 1284-1290.
11. Afrin, S. F., Mahmood, A. K. A., Bari, K. F., Rahman, F., & Hassan, Z. (2017). Pattern of lipid levels of subjects seeking laboratory services in an established laboratory in the Dhaka city. *Bangladesh Journal of Medical Science*, 16(3), 375–379. https://doi.org/10.3329/bjms.v16i3.32849
12. Wilson, P. W. (1990). High-density lipoprotein, low-density lipoprotein and coronary artery disease. *The American Journal of Cardiology*, *66*(6), A7-A10.
13. Wilson, P. W., Abbott, R. D., & Castelli, W. P. (1988). High-density lipoprotein cholesterol and mortality. The Framingham Heart Study. *Arteriosclerosis: An Official Journal of the American Heart Association, Inc.*, *8*(6), 737-741.
14. Al-Mahmood, A. K., Ismail, A. A., R, F. A., Wan Bebakar, W. M., & Tai, E. S. (2016). The metabolic syndrome in normal weight Malay subjects. *Bangladesh Journal of Medical Science*, 15(1), 123–128. https://doi.org/10.3329/bjms.v15i1.27149
15. Lewington S., Whitlock G., Clarke R., Sherliker P., Emberson J., Halsey J., Qizilbash N., Peto R., Collins R. (2007). "Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths," *The Lancet*, vol. 370, no. 9602, pp. 1829–1839.
16. Adnan, R. M., Mostafa, R. R., Kisi, O., Yaseen, Z. M., Shahid, S., & Zounemat-Kermani, M. (2021). Improving streamflow prediction using a new hybrid ELM model combined with hybrid particle swarm optimization and grey wolf optimization. *Knowledge-Based Systems*, *230*, 107379.
17. Ahmad, W. M. A. W., Ahmed, F., Noor, N. F. M., Aleng, N. A., Ghazali, F. M. M., & Alam, M. K. (2022). Prediction and Elucidation of Triglycerides Levels Using a Machine Learning and Linear Fuzzy Modelling Approach. *BioMed Research International 2022*, 1-7.
18. Sharkawy, A. N. (2020). Principle of neural network and its main types. *Journal of Advances in Applied & Computational Mathematics*, *7*, 8-19.
19. Wynants, L., Bouwmeester, W., Moons, K. G. M., Moerbeek, M., Timmerman, D., Van Huffel, S., ... & Vergouwe, Y. (2015). A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. *Journal of clinical epidemiology*, *68*(12), 1406-1414.
20. LaFontaine, D. (2021). The history of bootstrapping: Tracing the development of resampling with replacement. *The Mathematics Enthusiast*, *18*(1), 78-99.
21. Ahmad, W. M. A. W., Azmi, N. A. N., Aleng, N. A., Mohd, M. S. B., Ibrahim, R. H., Ali, Z., Rosdi W.M.L., & Harun, M. (2016). Modified Bayesian regression modeling involving qualitative predictor variables: A tumor size study. *Journal of Scientific Research and Development.* 3 (7), 14-19.
22. Ghazali, F.M.M., Ahmad, W.M.A.W., Nawi, M.A.A., Noor, N.F.M., Ghazali, N.F., Aleng, N.A., Ibrahim, M.S.M., & Halim, N.A. (2020). Ordered Logistic Regression with Artificial Neural Network Models for Variable Selection for Prediction of Hypertension Patient Outcomes. *Sapporo Medical Journal*, 54(10).
23. Alexopoulos, E. C. (2010). Introduction to multivariate regression analysis. *Hippokratia*, *14*(Suppl 1), 23.
24. Ahmad, W. M. A. W., Ghazali, F. M. M., Noor, N. F. M., Alam, M. K., & Aleng, N. A. (2021). Using Two-Layered Feed-Forward Neural Networks To Model Blood Uric Acid Among Diabetic Patients. *Bangladesh Journal of Medical Science*, 20(4), 741-747.
25. Ahmad, W. M. (2012). Forecasting short-term load demand using multilayer feed-forward (MLFF) neural network model. *Applied Mathematical Sciences*, 6(108), 5359-5368.
26. Al-Mahmood, A. K. SF Afrin, N Hoque. (2014). Dyslipidemia in Insulin Resistance: Cause or Effect. *Bangladesh Journal of Medical Biochemistry*, 7(1). DOI: 10.3329/bjmb.v7i1.18576.
27. Ghazali, F. M. M., W Ahmad, W. M. A., Srivastava, K. C., Shrivastava, D., Noor, N. F. M., Akbar, N. A. N., Aleng, N. A., & Alam, M. K. (2021). A Study of Creatinine Level among Patients with Dyslipidemia and Type 2 Diabetes Mellitus using Multilayer Perceptron and Multiple Linear Regression. *Journal of pharmacy & bioallied sciences*, 13(Suppl 1), S795–S800. https://doi.org/10.4103/jpbs.JPBS_778_20

Appendix

```
#/Dataset for Biometry: Biometry Modeling Study 2003 #
Input =("
Age Wc Tc Tg Ldl Hdl Alp
39 101 4.09 1.11 2.59 1.00 54
39 101 4.09 1.11 2.59 1.00 54
53 32 4.74 2.74 2.70 .79 84
68 89 4.64 .70 2.84 1.45 89
⋮ ⋮ ⋮
46 34 4.93 1.53 3.30 .93 96
56 36 5.79 2.74 3.55 .99 84
58 35 3.40 1.40 3.02 1.48 89
71 33 3.58 .91 2.36 .81 143
65 30 6.89 3.48 3.97 1.34 107
47 32 6.04 1.41 4.43 .97 51
66 33 3.20 1.33 1.64 .96 93
68 33 4.46 1.36 2.84 1.00 111
47 33 1.96 1.04 1.96 1.51 40
")
data = read.table(textConnection(Input),header=TRUE)

#/Performing bootstrap for 1000 : case resampling
    procedure /#
mydata <- rbind.data.frame(data, stringsAsFactors =
    FALSE)
iboot <- sample(1:nrow(mydata),size=1000, replace = TRUE)
bootdata <- mydata[iboot,]

#/Install the neuralnet package/#
if(!require(neuralnet)){install.packages("neuralnet")}
library("neuralnet")

#/Checking for the missing values/#
apply(bootdata, 2, function(x) sum(is.na(x)))

#/Scaling the data for normalization/#
#/Method (usually called feature scaling) to get all the
    scaled data/#
#/In the range [0,1]/#
max_data <- apply(bootdata, 2, max)
min_data <- apply(bootdata, 2, min)
data_scaled <- scale(bootdata, center = min_data, scale =
    max_data - min_data)

#/Randomly split the data into 70:30/#
#/70 Percent of the data at our disposal to train the
    network/#
#/30 Percent to test the network/#
index =
    sample(1:nrow(bootdata),round(0.70*nrow(bootdata)))
train_data <- as.data.frame(data_scaled[index,])
test_data <- as.data.frame(data_scaled[-index,])

#/Build the network/#
#/There are 3 hidden layers have 3 and 2 neurons
    respectfully/#
#/Input  = 6/#
#/Output  = 1/#
n = names(bootdata)
f = as.formula(paste("Ldl ~", paste(n[!n %in% "Ldl"],
    collapse = " + ")))
nn = neuralnet(f,data=train_data,hidden=c(4),linear.
    output=T)
plot(nn)
options(warn=-1)

#/30 Percent of the available data to do this:
# Using only the first 2 columns representing the input
    variables
# of The network and 1 is the output for NN/
predicted <- compute(nn,test_data[,1:6])

#/Use the Mean Squared Error NN (MSE.net-forecasts the
    network) as a measure of how far away our predictions
    are from the real data/#
MSE.net <- sum((test_data$Ldl-predicted$net.result)^2)/
    nrow(test_data)
MSE.net
Model <- lm(Ldl~Age+Wc+Tc+Tg+Hdl+Alp, data=bootdata)  #
    build the model
summary(Model)

data$PredictedLdl <- predict(Model,data)
distPred <- predict(Model, data)
preds <- predict(Model, data)
modelEval <- cbind(data$Ldl, preds)
colnames(modelEval) <- c('Actual','Predicted')
modelEval <- as.data.frame(modelEval)
print (modelEval)

#/Calulate mean square error /#
test <- data[-index,]
predict_lm <- predict(Model,test)
MSE.lm <- sum((predict_lm - test$Ldl)^2)/nrow(test)
MSE.lm

#/Printing the Value of MSE for Linear Model and Neural
    Network/
print(paste(MSE.lm, MSE.net))
# Finished/#
```