Clinica Chimica Acta 489 (2019) 35–40

Contents lists available at ScienceDirect

Clinica Chimica Acta

journal homepage: www.elsevier.com/locate/cca

Deep neural network for estimating low density lipoprotein cholesterol

Taesic Lee[a], Juwon Kim[b], Young Uh[b,**], Hyunju Lee[a,c,*]

[a] Department of Biomedical Science and Engineering, Gwangju Institute of Science and Technology, Gwangju, South Korea.
[b] Department of Laboratory Medicine, Yonsei University Wonju College of Medicine, Wonju, South Korea.
[c] School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, South Korea.

ARTICLE INFO

Keywords:
Deep neural network
LDL cholesterol

ABSTRACT

Background: LDL cholesterol (LDL-C) has been mainly estimated using the Friedewald equation, and other equations have recently been developed to complement the Friedewald equation. The present study aims to employ a deep neural network (DNN) to improve LDL-C estimation.

Methods: We used two independent datasets obtained from the Korean National Health and Nutrition Examination Survey and the Wonju Severance Christian Hospital as training and test datasets, respectively. We used the training dataset to construct the DNN architecture, which takes three input values of total cholesterol, HDL cholesterol, and triglyceride, and estimates LDL-C as the output. The model consists of six hidden layers, and each hidden layer has 30 nodes. The performance of the DNN model constructed by the training dataset was measured using the test dataset.

Results: In fivefold cross-validation using the training dataset, the DNN model showed the lowest mean and median squared errors compared to the Friedewald equation and Novel method. For the independent test dataset, our DNN model outperformed other existing methods on the basis of mean and median squared errors.

Conclusions: The DNN model provided the most accurate estimation of LDL-C compared to other existing methods including the Friedewald and Novel methods.

1. Introduction

LDL cholesterol (LDL-C) is an important modifiable risk factor for cardiovascular disease (CVD) and a primary target in national and international clinical practice guidelines [1]. It is primarily estimated via the Friedewald equation [2]. However, several studies have suggested that the Friedewald method misclassifies the risk of CVD in many cases, particularly those with hypertriglyceridemia [3]. This misclassification is mainly attributed to the triglyceride:5 ratio, representing the triglyceride:VLDL cholesterol (VLDL-C) ratio, used in the Friedewald equation [3]. To overcome this limitation, the Novel method for estimating LDL-C was advocated, where a hyperparameter X in triglyceride:X is empirically determined on the basis of a subject's triglyceride and non-HDL cholesterol (non-HDL-C) [4].

In one study, an artificial neural network (ANN) with one hidden layer was used to estimate LDL-C on patients with diabetes mellitus [5]. The ANN, a machine learning technique consisting of many linear and non-linear units [6], uses back-propagation to determine its internal parameters by reducing differences between ground truth and estimation [6]. Recently, deep neural network (DNNs) have been highlighted because a DNN can represent highly-complex data [7]. In this study, we use a DNN for estimating LDL-C from two independent datasets involving the Korean population.

2. Methods

For a training dataset, we obtained a dataset of 14,812 anonymized laboratory test records from the Korean National Health and Nutritional Examination Survey (KNHANES) from 2009 to 2015. Data with any missing results of total cholesterol, triglyceride, HDL−C, and LDL-C were excluded. Data involving dyslipidemia medication or higher triglyceride concentration of > 400 mg/dL [4] were not excluded. All lipid profiles were tested after > 12 h of fasting.

In the training dataset, total cholesterol, triglyceride, HDL−C, and LDL-C levels were measured by the Hitachi analyzer 7600 (Hitachi, Tokyo, Japan) using commercially available kits from Sekisui

[*] Correspondence to: Hyunju Lee, School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, 123 Chemdangwagi-ro, Gwangju 61005, Korea

[**] Corresponding author at: Department of Laboratory Medicine, Yonsei University Wonju College of Medicine, 20 Ilsan-ro, Wonju-si, Gangwon-do, 26426, South Korea

E-mail addresses: u931018@yonsei.ac.kr (Y. Uh), hyunjulee@gist.ac.kr (H. Lee).

https://doi.org/10.1016/j.cca.2018.11.022
Received 31 July 2018; Received in revised form 13 November 2018; Accepted 13 November 2018
Available online 15 November 2018

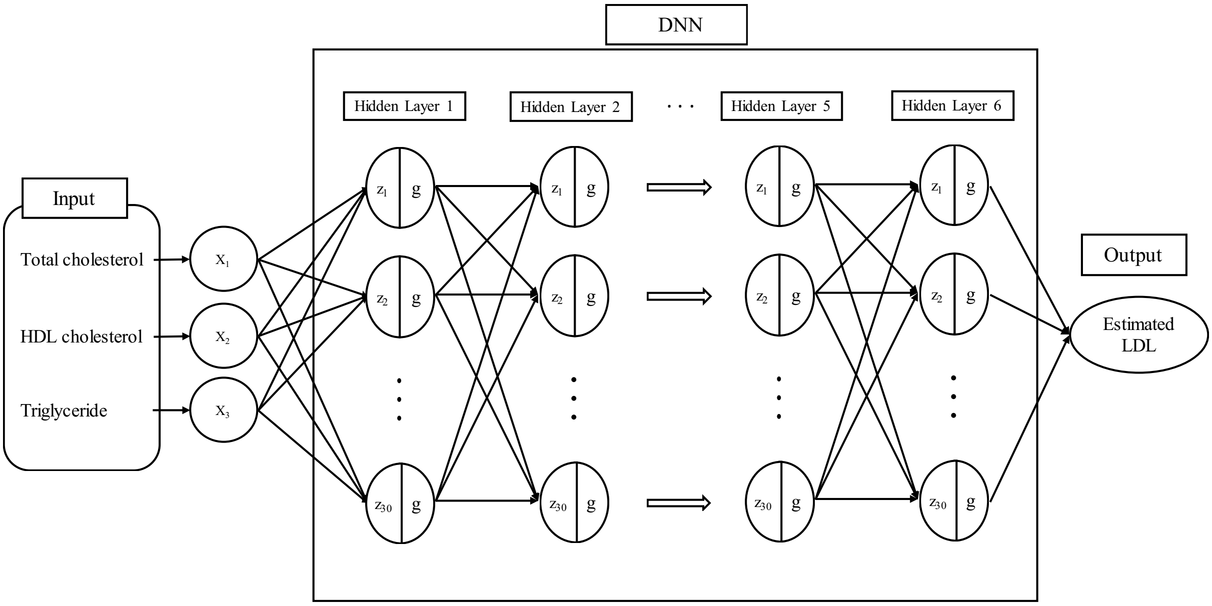

Fig. 1. DNN model for estimating LDL-C.
Input takes three values: total cholesterol, HDL−C, and triglyceride, and the output is the LDL-C. The DNN model consists of six hidden layers, and each hidden layer has 30 nodes. Each node's response is "z" based on activation function "g," where z = $bias + \Sigma_{i=1}{}^{n}X_iw_i$, and g(z) = max (0, z) (Rectified Linear Unit).

Diagnostics (Tokyo, Japan).

For a test dataset, lipoprotein subfraction test results were selected from September 2008 to March 2013 recorded at Wonju Severance Christian Hospital (WSCH). A total of 4520 samples were finally enrolled after excluding samples with missing data for lipid profiles. All subjects fasted overnight for 12 h before blood collection, and the levels of total cholesterol, triglyceride, HDL−C, and LDL-C in serum samples were measured on the day of blood collection. If a patient record contained multiple test results of lipoprotein subfraction tests during the study period, only the first result was chosen in the dataset. The total cholesterol, triglyceride, HDL-C and LDL-C results were extracted only when LDL-C subfraction test results were collected within the same day. The total cholesterol, triglyceride, HDL−C, and LDL-C for the testing dataset were analyzed using the modular DPE system (Roche Diagnostics, Basel, Switzerland). The lipoprotein subfraction test was performed using the Lipoprint LDL System (Quantimetrix, Redondo Beach, CA, USA). The Lipoprint system uses polyacrylamide gel electrophoresis to separate the various lipoprotein subfractions on the basis of particle size as VLDL band, three mid-bands [MID-C = VLDL remnants, MID-B = large intermediate density lipoprotein (IDL), and MID-A = small IDL], and seven LDL bands and an HDL band. LDL-1 and -2 are defined as large LDL subfractions whereas LDL-3, 4, 5, 6, and − 7 are defined as small LDL subfractions [8,9].

All data were accessed in compliance with the Helsinki Declaration, and training data was in the compliance with Institutional Review Board of WSCH (approval no. CR317314).

For the performance comparison of our DNN model for estimating LDL-C, the Friedewald eq. (LDL-C_F) [2] and the Novel method (LDL-C_N) developed by Martin et al. [4] were used, which are defined as follows:

$$LDL - C_F = (\text{Total cholesterol–HDL} - C)–(\text{Triglyceride}/5)$$

$$LDL - C_N = (\text{Total cholesterol–HDL} - C)–(\text{Triglyceride}/X)$$

In LDL-C_N, X is an adjustable factor based on the 180-cell method.

Our DNN model takes three input values of total cholesterol, HDL−C, and triglyceride, and estimates LDL-C as the output. The model consists of six hidden layers, and each hidden layer has 30 hidden nodes. In each node, when the input is represented by X_i, output is obtained through the activation function g(z), where z = f (X_i) = $bias + \Sigma_{i=1}{}^{n}X_iw_i$ and g(z) = max (0, z) (Rectified Linear Unit, ReLU). Here, we chose ReLU as the activation function to apply non-linearity in each hidden layer. Weight parameters w_i were learned using the training dataset. The architecture of the DNN model is shown in Fig. 1.

We performed fivefold cross-validation using the training dataset (Fig. 2). Cross-validation is a popular strategy for model selection [10]. The algorithm for this technique is described in the following steps:

1. Randomly split the training samples into five equal parts.
2. Select one part, and merge the other four parts into a subset for training the DNN.
3. Use the single part to calculate the prediction errors of the DNN.
4. Repeat the above steps from Step 2 five times, replacing the part used for error calculation with the next part.

This procedure will yield five different DNN models with the same basic structure, but with different weights and different prediction errors. In our study, fivefold cross-validation was selected to determine the structure of the DNN and to internally and objectively check its performance. During cross validation, we increased the hidden layers and nodes of each DNN, and the performance change was observed. After this process, a DNN consisting of six hidden layers with 30 nodes in each layer was selected (Fig. 1).

Applying the KNHANES data consisting of 14,812 samples to the determined DNN structure, we constructed the final DNN model for estimating LDL-C. To externally validate the DNN model, the testing dataset obtained from 2008 to 2013 at WSCH, which was not included in the training, was used to check the prediction error of the DNN.

The performance of each method including LDL-C_F, LDL-C_N, and DNN was checked through the squared error as follows:

$$\text{Squared error} = (\text{Measured LDL} - \text{C–Estimated LDL} - C)^2$$

where "Measured LDL-C" was considered as the ground truth and "Estimated LDL-C" as the approximate value for "Measured LDL-C" calculated by the three methods.

Statistical analyses and density plots of the squared error were performed in R (http://r-project.org), version 3.4.1. Python (Version 3.6) based on Tensorflow-GPU (Version 1.5) was used for training and testing the DNN model.

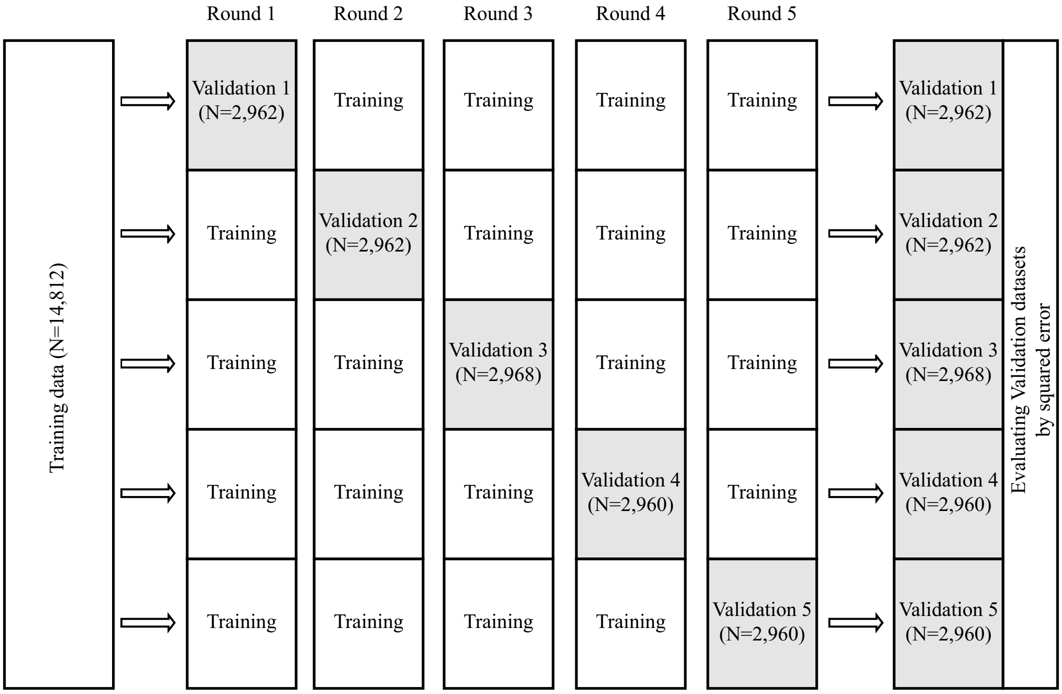

Fig. 2. Fivefold cross-validation.
Training data in the KNHANES dataset are split into five subsets. A validation subset is used for measuring performance of the DNN model constructed using the merged data of the remaining four subsets. This process takes place five times using each subset once for validation purposes, and performance is checked by squared error.

Table 1
Baseline characteristics of two independent datasets in Korean population.

Characteristics	Training dataset		Testing dataset	
Total number	14,812		4520	
Age, yr, median (IQR)	46	(32–59)	67	(56–74)
Age group, yr, n (%)				
< 20	1515	(10.2)	16	(0.4)
20–64	10,826	(73.1)	1936	(42.8)
≥65	2471	(16.7)	2568	(56.8)
Gender, n (%)				
Male	7375	(49.8)	2526	(55.9)
Female	7437	(50.2)	1994	(44.1)
Cholesterol, mg/dL, median (IQR)				
Total	186	(162–212)	165	(140–192)
HDL-C	47	(40–56)	37	(30–45)
Measured LDL-C	110	(89–132)	100	(79–123)
VLDL-C	–		22.3	(16.5–30.5)
Triglyceride, mg/dL, median (IQR)	120	(76–211)	103	(75–148)
Triglyceride/VLDL-C, median (IQR)	–		4.71	(3.75–5.97)

3. Results

The general characteristics and lipid profiles of participants are presented in Table 1. The median age of patients in the test dataset was older than the training dataset. Among the 14,812 participants (training dataset), 2471 were elderly patients aged ≥65 (17%). In the testing dataset including 4520 subjects, 56.8% were elderly patients (≥ 65 years of age). The proportions of men in the training dataset and testing dataset were 49.8% and 55.9%, respectively.

Lipid profiles and triglyceride:VLDL ratios were presented as median and interquartile range (IQR). Total cholesterol and LDL-C were higher in the training dataset than the testing dataset. Only presented in the training dataset, the median of VLDL-C and triglyceride:VLDL-C ratio were 22.3 and 4.71, respectively (Table 1).

We conducted fivefold cross-validation for establishing the deep learning model, and five validation datasets ranging from 2960 to 2968 subjects were constructed (Fig. 2). Mean and median squared errors of LDL-C_F, LDL-C_N, and DNN are presented in Table 2. Mean of squared errors ranged from 167.5 to 269.8 in LDL-C_F, from 69.0 to 123.6 in LDL-C_N, and from 59.6 to 69.4 in DNN (Table 2). Standard deviations of squared errors ranged from 771 to 2364 in LDL-C_F, 171 to 1267 in LDL-C_N, and 139 to 601 in DNN (Table 2). For all mean and median squared errors of five validation sets (n = 14,812), the DNN outperformed LDL-C_F and LDL-C_N (Table 2) in the estimation of LDL-C.

The final DNN model trained from 14,812 samples in the training dataset was tested with 4520 different samples in the testing dataset (Table 3). Both mean and median squared errors were the lowest in the DNN among the three methods (Table 3). Standard deviations were 537.7 in LDL-C_F, 174.0 in LDL-C_N, and 166.4 in the DNN model. After selecting samples with TG < 400 mg/dL, the DNN model also provided the best accurate values of estimated LDL-C compared to the Friedewald and Novel methods (Table 3).

Three plots of the density distribution of squared errors from the three methods are shown in Fig. 3. All three distributions peaked at squared errors between 0 and 25, and were left-skewed (Fig. 3). The DNN exhibits the highest peak among three methods, indicating that DNN submitted estimates with lower error than LDL-C_F or LDL-C_N (Fig. 3).

4. Discussion

In considering the implications to patient care, LDL-C estimation should be carried out with the approach which provides the most accurate and closest value to the measured LDL-C. In this study, we developed and validated a DNN model for estimating LDL-C from the standard lipid profile. Different from those of the Friedewald method that applies a fixed factor of 5 [2] and the Novel method proposed by Martin et al. [4], which uses an experimentally obtained adjustable factor for the triglyceride:VLDL-C ratio, 180 (6 × 30) perceptrons/nodes in the DNN model are activated for estimating LDL. Similar to other previous studies [4,11], our data also showed that the Novel

Table 2
Square error between measured LDL-C and estimated LDL-C of fivefold cross validation in the training dataset.

		All samples (n = 14,812)									
Methods for LDL-C estimation	Estimator	Cross validation 1 (n = 2962)[a]		Cross validation 2 (n = 2962)		Cross validation 3 (n = 2968)		Cross validation 4 (n = 2960)		Cross validation 5 (n = 2960)	
LDL-C_F (Friedewald equation)	Mean (SD)	236.7	(1370)	269.8	(2364)	180.1	(771)	167.5	(932)	198.3	(1033)
	Median (IQR)	33.64	(6.76–121.0)	36.00	(7.84–130.0)	36.00	(6.76–125.4)	31.36	(6.76–121.0)	33.64	(7.84–125.4)
LDL-C_N	Mean (SD)	89.7	(536)	123.6	(1267)	69.0	(171)	71.9	(225)	75.30	(265)
	Median (IQR)	19.14	(4.00–63.3)	18.78	(4.00–61.1)	19.60	(4.00–60.1)	18.24	(4.00–62.6)	19.14	(4.48–60.8)
DNN	Mean (SD)	69.4	(393)	82.5	(601)	60.6	(139)	59.6	(158)	63.1	(164)
	Median (IQR)	17.72	(3.64–57.7)	17.32	(3.70–57.6)	17.64	(4.03–60.0)	17.29	(3.88–56.7)	18.76	(3.73–59.9)

[a] Number of validation dataset

method [4] outperformed the Friedewald method. Compared to these two methods, the DNN estimated the LDL-C with the lowest error. In addition to application of deep learning in estimating LDL-C, a major strength of this study is to internally validate the DNN model on the basis of cross-validation, after which the independent dataset is used to test the DNN model.

The ratio of triglyceride:VLDL is the most important factor regarding the accuracy of estimated LDL-C [4,11]. The median triglyceride:VLDL-C ratio of 4.7 derived from the test dataset in our study is closer to the factor of 5.2 proposed by Martin et al. *(4)* than the 5.7 ratio reported by Meeusen et al. [11]. The lower triglyceride:VLDL ratio may indicate the inclusion of lower dyslipidemic subjects in our study compared to the previous studies [4,11].

The first study for estimating LDL-C by neural network was conducted on 500 individuals with Type II diabetes mellitus in 2010 [5]. Ten inputs including age, height, and lipid profile, one hidden layer including 15 nodes, and a sigmoid activation function were used in that study [5]. In the present study, three input features, namely the total cholesterol, HDL—C, and triglyceride were used to compare the model with other previously proposed methods. Furthermore, to improve the performance of the DNN model, we selected six hidden layers and ReLU as the nodes' activation function. During the fivefold cross-validation process, we observed improved performance with higher numbers of hidden layers and nodes. Thus, we constructed the DNN using six hidden layers and 30 nodes per layer similar with the Novel method clustering all data to 180 cases [4].

There have been many studies for estimating LDL-C using linear regression [12,13], including the equations "0.9 × non-HDL-C – triglyceride × 0.1" [12] and "0.75 × (Total cholesterol – HDL-C)" [13]. However, these studies used only one layer for estimating LDL-C, which is incapable of implementing more complex interrelations between inputs *(6)*. In one layer-based regression, or any other "shallow regression," the numbers of optimized parameters range from one to four. However, the DNN model in our study requires training of 4620 parameters ('$3\times30 + 30\times30\times5 + 30$' parameters for input to hidden, hidden to hidden, and hidden to output layers, respectively), which entails significant computation. Nonetheless, using backpropagation [6] and Tensorflow-GPU [14], only 1 min was required for training the 4620 parameters. Taken together, our DNN model yields the most accurate LDL-C by performing 4620 calculations in 180 chambers (6 hidden layers by 30 hidden nodes).

We showed that most values of squared errors were concentrated between 0 and 25, indicating that all three methods could well estimate LDL-C (Fig. 3). Our DNN exhibited a higher peak than LDL-C_F and LDL-C_N, where LDL-C_N displayed better performance than LDL-C_F, indicating that the DNN can be an alternative method to estimate LDL-C.

We also applied different architectures and learning approaches when constructing the DNN model, such as drop-out [15], early stopping [15], increasing hidden layers, and changing the method of selecting initial weights [16], to evaluate the performance of the DNN. Through several experiments, we found that early stopping, increasing hidden layers, and Xavier initialization [16] to randomly select initial weights were appropriate, but drop-out was not an appropriate approach to improve performance for LDL-C estimation. Furthermore, as suggested in another study [5], additional information, such as age, weight, blood pressure, and liver function could improve the DNN model.

5. Conclusions

From a large sample of lipid profiles, we constructed a DNN model to estimate LDL-C, and tested the DNN model with an independent dataset provided by a tertiary care hospital. The DNN model achieved the most accurate estimates compared to the Friedewald method [2] and the method proposed by Martin et al. [4]. Application of the DNN could be a simple task applied in the modern laboratory that is efficient in terms of technology and cost. Our DNN model, which we continue to develop, can be a suitable candidate.

Author contributions

All authors confirm that they have contributed to the intellectual content of this paper and have met the following three requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors' disclosures or potential conflicts of interest

Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:

Employment or leadership

None declared.

Consultant or advisory role

None declared.

Table 3
Square error between measured LDL-C and estimated LDL-C in the testing dataset.

	All samples (n = 4,520)[a]	
Methods for LDL-C estimation	Mean (SD)	Median (IQR)
LDL-C_F (Friedewald equation)	115.9 (537.7)	27.04 (5.9-86.5)
LDL-C_N	68.9 (174.0)	23.86 (5.5-74.2)
DNN	65.6 (166.4)	21.64 (5.1-66.3)

	Samples with triglyceride < 400mg/dL (n = 4,446)[a]	
Methods for LDL-C estimation	Mean (SD)	Median (IQR)
LDL-C_F (Friedewald equation)	73.9 (152.3)	26.01 (5.8-82.8)
LDL-C_N	62.1 (111.0)	23.11 (5.4-71.9)
DNN	58.8 (109.7)	21.66 (5.4-67.1)

[a] This was collected from Wonju Severance Christian Hospital for testing dataset.

Stock ownership

None declared.

Honoraria

None declared.

Research funding

This research was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (NRF-2016R1A2B2013855) and (NRF-2018M3C7A1054932).

Expert testimony

None declared.

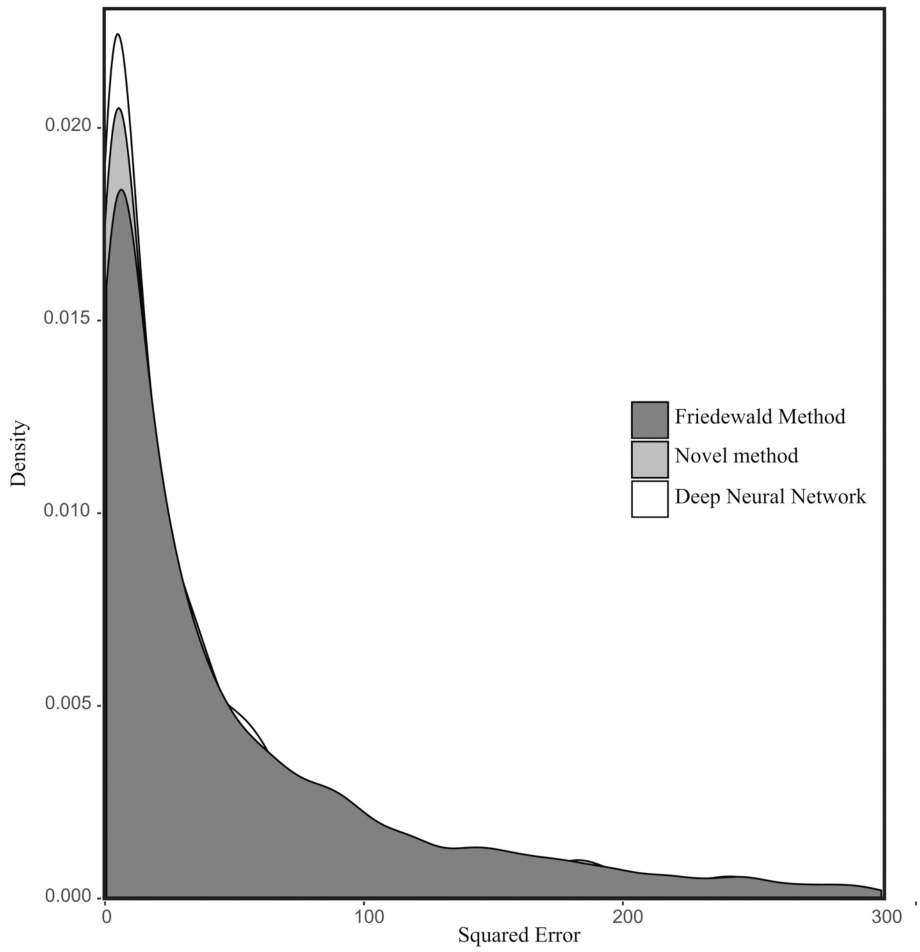

Fig. 3. Distributions of squared errors of LDL-C_F, LDL-C_N, and DNN.
Squared errors of testing data in WSCH dataset < 300 are illustrated in the density plot. Areas of each distribution equal one.

Patents

None declared.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cca.2018.11.022.

References

[1] Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults, Executive summary of the third report of the national cholesterol education program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III), JAMA 285 (2001) 2486–2497, https://doi.org/10.1001/jama.285.19.2486.

[2] W.T. Friedewald, R.I. Levy, D.S. Fredrickson, Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge, Clin. Chem. 18 (1972) 499–502.

[3] M. Nauck, G.R. Warnick, N. Rifai, Methods for measurement of LDL-cholesterol: a critical assessment of direct measurement by homogeneous assays versus calculation, Clin. Chem. 48 (2002) 236–254, https://doi.org/10.1373/clinchem.2014.227710.

[4] S.S. Martin, M.J. Blaha, M.B. Elshazly, P.P. Toth, P.O. Kwiterovich, R.S. Blumenthal, et al., Comparison of a novel method vs the Friedewald equation for estimating low-density lipoprotein cholesterol levels from the standard lipid profile, JAMA 310 (2013) 2061–2068, https://doi.org/10.1001/jama.2013.280532.

[5] M. Chakraborty, B. Tudu, Comparison of ANN models to predict LDL level in Diabetes Mellitus type 2, International Conference on Systems in Medicine and Biology (ICSMB), IEEE, 2010, pp. 392–396, , https://doi.org/10.1109/ICSMB.2010.5735410.

[6] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, https://doi.org/10.1038/nature14539.

[7] D. Cireşan, U. Meier, J. Masci, J. Schmidhuber, Multi-column deep neural network for traffic sign classification, Neural Netw. 32 (2012) 333–338, https://doi.org/10.1016/j.neunet.2012.02.023.

[8] H. Arai, Y. Kokubo, M. Watanabe, T. Sawamura, Y. Ito, A. Minagawa, et al., Small dense low-density lipoproteins cholesterol can predict incident cardiovascular disease in an urban Japanese cohort: the Suita study, J. Atheroscler. Thromb. 20 (2013) 195–203, https://doi.org/10.5551/jat.14936.

[9] R.C. Hoogeveen, J.W. Gaubatz, W. Sun, R.C. Dodge, J.R. Crosby, J. Jiang, et al., Small dense low-density lipoprotein-cholesterol concentrations predict risk for coronary heart disease: the Atherosclerosis Risk in Communities (ARIC) study, Arterioscler. Thromb. Vasc. Biol. 34 (2014) 1069–1077, https://doi.org/10.1161/ATVBAHA.114.303284.

[10] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Stat. Surv. 4 (2010) 40–79, https://doi.org/10.1214/09-SS054.

[11] J.W. Meeusen, A.J. Lueke, A.S. Jaffe, A.K. Saenger, Validation of a proposed novel equation for estimating LDL cholesterol, Clin. Chem. 60 (2014) 1519–1523, https://doi.org/10.1373/clinchem.2014.227710.

[12] Y. Chen, X. Zhang, B. Pan, X. Jin, H. Yao, et al., A modified formula for calculating low-density lipoprotein cholesterol values, Lipids Health Dis. 9 (2010) 52 https://doi.org https://doi.org/10.1186/1476-511X-9-52.

[13] C.M. de Cordova, M.M. de Cordova, A new accurate, simple formula for LDL-cholesterol estimation based on directly measured blood lipids from a large cohort, Ann. Clin. Biochem. 50 (2013) 13–19, https://doi.org/10.1258/acb.2012.011259.

[14] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al., Tensorflow: A system for large-scale machine learning, Preprint at https://arxiv.org/pdf/1605.08695 (2016).

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.

[16] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.