DOI: 10.4274/anatoljmed.2025.79663
Anatol J Gen Med Res 2025;35(3):307-320

CLINICAL RESEARCH / KLİNİK ARAŞTIRMA

ANATOLIAN JOURNAL of GENERAL MEDICAL RESEARCH

Validation and Improvement of a Machine-learning-based LDL Prediction Model Using Retrospective Lipid Profile Data

Retrospektif Lipid Profili Verileri Kullanılarak Makine Öğrenmesi Tabanlı LDL Tahmin Modelinin Doğrulanması ve İyileştirilmesi

Ferhat Demirci[1,2], Murat Emeç[3], Mehmet Hilal Özcanhan[4], Özlem Gürsoy Doruk[2,5], Pınar Akan[2,5]

[1]University of Health Sciences Türkiye, İzmir Tepecik Education and Research Hospital, Department of Medical Biochemistry, İzmir, Türkiye
[2]Dokuz Eylül University Institute of Health Sciences, Department of Neurosciences, İzmir, Türkiye
[3]İstanbul University Faculty of Computer and Informatics, Department of Computer Engineering, İstanbul, Türkiye
[4]Dokuz Eylül University Faculty of Engineering, Department of Computer Engineering, İzmir, Türkiye
[5]Dokuz Eylül University Faculty of Medicine, Department of Medical Biochemistry, İzmir, Türkiye

Cite as: Demirci F, Emeç M, Özcanhan MH, Gürsoy Doruk Ö, Akan P. Validation and improvement of a machine-learning-based LDL prediction model using retrospective lipid profile data. Anatol J Gen Med Res. 2025;35(3):307-320

Abstract

Objective: Direct measurement of low-density lipoprotein cholesterol (LDL-C) is time-consuming and expensive when triglycerides (TG) exceed 400 mg/dL. We sought to validate and refine a machine-learning (ML) model for rapid estimation of LDL-C in hypertriglyceridemic sera.

Methods: We extracted 25.991 lipid profiles (TG: 400-800 mg/dL) collected between 2010 and 2022 from two Turkish university hospitals. After an 80/20 split, seven ML algorithms were trained; the top two (random forest and XGBoost) were stacked with a decision tree meta-learner (model-3). Performance on the external test set (n=1.279) was compared with that of direct homogeneous LDL-C assays and the Sampson's formula (NIH-Equ-2) using balanced accuracy, precision, recall, F1 score, specificity, Pearson correlation coefficient, and Bland-Altman analysis, following International Federation of Clinical Chemistry and Laboratory Medicine analytical performance specifications.

Results: Model-3 yielded balanced accuracy =99.3%, precision =98.9%, recall =98.9%, and specificity =99.8%. Predicted LDL-C correlated strongly with direct measurement (r=0.996, p<0.001) and reduced the mean absolute error by 54% compared with NIH-Equ-2. Only 0.39% of cases were underclassified relative to the European Society of Cardiology/European Atherosclerosis Society LDL-C risk categories. Bland-Altman plots demonstrated no significant proportional bias across the LDL-C range (mean bias =-0.2 mg/dL; 95% limits of agreement: -7.8 to+7.4 mg/dL).

Conclusion: A stacked ensemble ML model delivers near-assay accuracy for LDL-C prediction in high-TG samples and markedly outperforms current formula. Implementation could enable same day, low-cost LDL-C reporting without extra laboratory procedures, supporting faster dyslipidaemia management.

Keywords: LDL, lipid profile, machine-learning, artificial intelligence

Öz

Amaç: Trigliserid (TG) düzeyi 400 mg/dL'nin üzerine çıktığında düşük yoğunluklu lipoprotein kolesterol (LDL-K) ölçümü zaman alıcı ve maliyetli hale gelmektedir. Bu çalışmada, hipertrigliseridemik serumlarda hızlı LDL-K tahmini için geliştirilen makine öğrenmesi (ML) modelinin doğrulanması ve iyileştirilmesi amaçlandı.

Address for Correspondence/Yazışma Adresi: Murat Emeç, PhD, İstanbul University Faculty of Computer and Informatics, Department of Computer Engineering, İstanbul, Türkiye
E-mail: murat.emec@istanbul.edu.tr
ORCID ID: orcid.org/0000-0002-9407-1728

Received/Geliş tarihi: 20.09.2025
Accepted/Kabul tarihi: 03.11.2025
Published date/Yayınlanma tarihi: 30.12.2025

Öz

Yöntem: 2010-2022 yılları arasında iki üniversite hastanesinden elde edilen 25,991 lipid profili (TG: 400-800 mg/dL) retrospektif olarak incelendi. Veriler %80/20 oranında ayrıldıktan sonra yedi ML algoritması eğitildi; en iyi iki algoritma (random forest, XGBoost), karar ağacı tabanlı bir meta-öğrenici ile birleştirilerek (model-3) istiflendi. Dış test setinde (n=1.279) modelin performansı doğrudan homojen LDL-K testleri ve Sampson formülüyle (NIH-Equ-2) karşılaştırıldı. Değerlendirmede dengeli doğruluk, kesinlik, duyarlılık, F1 skoru, özgüllük, Pearson korelasyonu ve Bland-Altman analizi kullanıldı; Uluslararası Klinik Kimya ve Laboratuvar Tıbbı Federasyonu analitik performans kriterleri dikkate alındı.

Bulgular: Model-3, dengeli doğruluk %99,3; kesinlik %98,9; duyarlılık %98,9 ve özgüllük %99,8 elde etti. Tahmin edilen LDL-K ile doğrudan ölçüm arasında güçlü korelasyon saptandı (r=0,996, p<0,001). Model, NIH-Equ-2 formülüne göre ortalama mutlak hatayı %54 azalttı. Avrupa Kardiyoloji Derneği/Avrupa Ateroskleroz Derneği LDL-K risk kategorilerine göre yanlış sınıflandırma oranı yalnızca %0,39 idi. Bland-Altman analizinde anlamlı orantısal yanlılık gözlenmedi (ortalama fark =-0,2 mg/dL; %95 güven aralığı, -7,8 ile +7,4 mg/dL).

Sonuç: Yığılmış topluluk ML modeli, yüksek TG düzeylerinde LDL-K tahmininde doğrudan testlere yakın doğruluk sağlamış ve mevcut formüllerden belirgin olarak üstün bulunmuştur. Modelin uygulanması, ek laboratuvar işlemleri olmadan aynı gün, düşük maliyetli LDL-K raporlamasına olanak tanıyabilir ve dislipidemi yönetiminde hız kazandırabilir.

Anahtar Kelimeler: LDL, lipid profili, makine öğrenmesi, yapay zeka

Introduction

Artificial intelligence (AI) applications in medicine have become increasingly widespread. Machine-learning (ML) developments have been widely adopted in medical AI (MAI) applications. The growth of MAI is due to the ever-increasing abundance of health data, the primary input for ML. Interest in MAI stems from its ability to generate diagnostic predictions from complex datasets. The MAI prediction and visualization applications have produced fast and accurate results in solving many medical problems[1].

However, dataset size is not the only driving force behind MAI. The number, variety, and accuracy of input data that are directly related to the output significantly influence the success of results produced by the designed ML models. Therefore, the data are expected to include all information related to the research output. In computer science, supervised ML (i.e., controlled ML) is currently the most widely used tool in MAI.

Computer algorithms such as computer-aided diagnosis or clinical decision support systems used for supporting diagnosis, decision-making, and prediction are classified as diagnostic devices[2]. The methods for clinical validation and development are similar to those for standard diagnostic tests. Therefore, medical devices used in MAI applications must undergo rigorous clinical and experimental validation before use in patients to ensure patient safety and the efficacy of the method. Additionally, the reproducibility of the ML prediction results is a major concern of the International Federation for Clinical Chemistry and Laboratory Medicine (IFFC)[3].

The MAI applications used for clinical validation vary according to their form, model, and function. Our work aims to rapidly predict patients' low-density lipoprotein (LDL) levels prior to the costly and delayed direct measurement. The present work aims to validate and improve our previous "LDL predictor model" (p-LDL-M {2}) designed for LDL prediction in patients with 400≤ triglyceride (TG) ≤ 800 mg/dL (abbreviated as high-TG for the rest of the article)[4]. Our ultimate goal is to recommend an improved MAI application for the research community. Although different models were tested in this work, our present and previous data were obtained from similar models of testing devices. In other words, there is no data discrepancy.

The design and validation of generalized, reproducible, and improved p-LDL-M models is a five-step iterative process. The steps involve formulating the problem, collecting and preparing the data, validating and selecting a model, and interpreting and finally implementing the model. The improved target model is obtained after identifying the model with the best performance. Finally, optimization and feature selection techniques are applied to further enhance the performance of the developed model. However, the performance results of interim models have not been included in the results section to save space and avoid repeating noncritical results.

LDL-C concentration is the principal target for lipid-lowering therapy and a key determinant of atherosclerotic cardiovascular disease (ASCVD) risk, as emphasized by recent European Society of Cardiology (ESC)/European Atherosclerosis Society (EAS) and American College of Cardiology/American Heart Association Guidelines. However, the reliability of conventional LDL-C estimation equations is limited under hypertriglyceridemic conditions (TG >400 mg/dL). The Friedewald formula becomes invalid, the Martin-Hopkins method underestimates LDL-C in low LDL-C ranges, and even the National Institutes of Health (NIH)-Equ-2 method may introduce bias at very high-TG levels. Therefore, developing a robust ML-based estimation method is essential for precise LDL-C assessment in these patients.

The main aim of this study was to develop and validate a reliable ML-based model capable of predicting LDL-C levels in patients with hypertriglyceridemia (TG >400 mg/dL) before the costly and delayed laboratory measurements.

Materials and Methods

Study Design

Before commencing the study, the necessary approval was obtained from the Non-Interventional Ethics Committee of University of Health Sciences Türkiye, İzmir Tepecik Education and Research Hospital, (approval no: 2023/13-23, date: 12.04.2023) and Non-Interventional Ethics Committee of Dokuz Eylul University Faculty of Medicine (approval no: 2023/20-04, date: 14.06.2023). This study was first conducted at the University of Healt Sciences Türkiye, Dr. Suat Seren Chest Diseases and Chest Surgery Training and Research Hospital (hospital 1) and at the Dokuz Eylül University Research and Application Hospital (hospital 2) as the validation and improvement phase of the first phase. All experiments on humans were conducted according to relevant ethical guidelines and regulations. The experiments followed protocols approved by the Ethics Committees of hospital 1 and hospital 2. All experimental protocols used in this study have been reviewed and approved by the relevant institutional and/or licensing committee. The study's participants are three healthcare scientists and two engineering scientists from four institutions. The participants comply with the first recommendation to involve diverse stakeholders in developing clinically useful, practical, and ethical models. A total of 6.404 patient records with high-TG levels were presented in the hospital 1 biochemistry laboratory[4]. The hospital 2 biochemistry laboratory maintains records for 20.690 high-TG patients. During data analysis, records were omitted if they had missing results for total cholesterol (TC), TG, high-density lipoprotein (HDL), or LDL; if results exceeded the linear limits of specific analysis methods; if they contained zero or negative values; if they were from patients younger than 18 years of age; or if they lacked numerical data.

Of 27.094 patient records across the two hospitals, only 25.991 high-TG patient records (6.392 from hospital 1 and 19.599 from hospital 2) were processed using Python® software (Wilmington, Delaware, USA). As a rule of thumb in ML design and testing, the dataset was split into three training subsets (80% of all TG records, all-TG) and three testing subsets (20% of each dataset). Training and testing were conducted using nine combinations of datasets and three ML models. The study results are valid only for patients with high-TG levels. The TC, HDL, LDL, and TG were analyzed using Roche Cobas c702 (Mannheim, Germany) and Beckman Coulter AU5800 (California, USA) automated analyzers at hospital 1 and hospital 2, respectively. Our training and test sets were completely independent, meaning no test data was used in training the models. In addition, no data that would be unavailable during actual use were used; i.e., there was no data leakage in our analysis. With no data leakage and an 80:20 independent training-test split, our sample sets comply with the IFFC recommendations.

Study Population/Subjects

Our study population consisted of 6.392 lipid profile results obtained between January 2010 and December 2022 at hospital 1 and 19.559 results obtained between August 2011 and July 2022 at hospital 2. Standardized lipid profile data collected from the laboratory database included TC, TG, HDL, and LDL levels that were measured on the same day. Table 1 shows the main characteristics of the two high-TG study populations.

At hospital 1, 3.431 cases were male and 2.961 were female. The mean age of men was 49.72 years, while the mean age of women was 54.07 years. The mean directly measured LDL was 149.76±45.28 mg/dL. At hospital 2, 16.638 cases were male 2.961 were female. The mean ages were 56.81 years for men and 54.06 years for women. The mean measured direct LDL level was 151.10±46.44 mg/dL. Figure 1 displays the standard diagram used for reporting diagnostic accuracy, illustrating the progression of subjects throughout the study. Participants were divided into two datasets for statistical evaluation and ML analysis. The first dataset, typically comprising 80% of the participants, was used as the training set, while the remaining 20% formed the test set. In ML, the training set is utilized to build predictive models, and the test set is used to assess the prediction accuracy of those models.

Table 2 shows that the training set was divided into three groups. The first group (n=5113) contains direct LDL data from hospital 1 with TG levels >400 mg/dL. The second group (n=19.599) contains hospital 2 direct LDL data with TG >400 mg/dL. The third group (n=24.712) comprises the combined direct LDL data from hospital 1 and hospital 2 for cases with TG levels >400 mg/dL. In the designed ML models, the training set of model-1 (the model most similar to our previous p-LDL-M {2} model) included only the first group of data, while model-2 used only the second data group. Model-3 was trained using the sum of the training sets[4]. It should be pointed out that the ML models also differ in their AI architectures.

To ensure unbiased comparability, the test set is the same for all three models: 20% of the hospital 1 data. The test set had to be from hospital 1 because testing newly designed models with a new training and test set from hospital 2 could have been misleading by eliminating cross-hospital prediction

The LDL level distribution of the 1279 test subjects is shown in Figure 1. The classification is based on the 2019 ESC/EAS Guidelines for managing dyslipidemia[5]. The most undesirable error in LDL level classification is assigning a patient to an LDL level below the actual classification (under-classification). Therefore, preventing under-classification was one of the primary objectives of the new model designs. The above properties of the study population indicate full compatibility with the sample size, race, gender, data diversity, and train-test set partitioning recommendations of IFFC.

Table 1. Characteristics of the study population

Characteristics	Units	Hospital 1: n=6.392 value ± SD	Hospital 2: n=19.599 value ± SD
Age		51.73±11.61	56.39±13.75
Male	years	49.72±11.20	56.81±14.05
Female	years	54.07±11.61	54.06±11.65
Sex			
Male	N/A	3431 (%53.7)	16638 (%84.9)
Female	N/A	2961 (%46.3)	2961 (%15.1)
Total cholesterol	mg/dL	243.16±52.79	248.46±57.82
	mmol/L	6.29±1.37	6.45±1.50
Triglycerides	mg/dL	510.98±96.71	509.33±97.13
	mmol/L	5.77±1.09	5.79±1.10
HDL	mg/dL	37.57±8.97	40.51±11.17
	mmol/L	0.97±0.23	0.92±0.21
Non-HDL cholesterol	mg/dL	205.62±48.12	207.96±51.30
	mmol/L	5.32±1.24	5.32±1.24
Direct LDL	mg/dL	149.76±45.28	151.10±46.45
	mmol/L	3.87±1.17	3.87±1.17

SD: Standart deviation, N/A: Not applicable, HDL: High-density lipoprotein, LDL: Low-density lipoprotein

Table 2. Description of model abbreviations according to data sets

Model	Training set	Test Set
Model-1	80% records of high-TG subjects only (group 1: 5.113)	20%
Model-2	100% records of high-TG subjects only (group 2: 19.599)	20%
Model-3	80% records of high-TG subjects in hospital 1 and 100% records of high-TG subjects in hostpital 2 (group 3: 27.712)	20%
NIH-Equ-2	The number of n in the groups for each model (group 1: 5.113, group 2: 19.599, group 3: 27.712)	20%

TG: Triglycerides, NIH-Equ-2: National Institutes of Health-Equ-2

Lipid Profile Testing

All lipid profile parameters were analyzed using automated chemistry analyzers: the Roche Cobas c702 in the biochemistry laboratory of hospital 1 and the Beckman Coulter AU5800 in the biochemistry laboratory of hospital 2. Only the initial test results of each patient were considered in the study; repeated measurements were excluded. TC and TG were determined using the enzymatic cholesterol esterase/oxidase and glycerol phosphate oxidase methods, respectively.

HDL levels were measured using a direct homogeneous assay that did not involve precipitation. LDL was quantified using a direct homogeneous assay that employs a selective protective agent to isolate LDL from chylomicrons, HDL, and very LDL, with measurement by the cholesterol esterase/oxidase method. The maximum allowable total error for LDL based on these methodologies was 11.9%. The actual total error rates recorded by the Roche c702 and Beckman AU5800 analyzers were 9.48% and 8.67%, respectively. Since both error rates were below the acceptable limit, the lipid profile data were deemed reliable and suitable for the study.

ML Analysis

Python 3.9 was used as the primary programming language. Data manipulation and analysis were performed using the Pandas Library (version 1.4.4) in Python. NumPy (version 1.21.5), which supports the handling of large, multidimensional arrays and provides advanced mathematical functions for array operations, was also used. ML models were developed using the Scikit-learn (Sklearn) library, version 1.0.2. To evaluate the contribution of individual features to model predictions, SHapley Additive exPlanations (SHAP) analysis was conducted. The SHAP library (version 0.42.1) was used to measure feature importance, and the corresponding

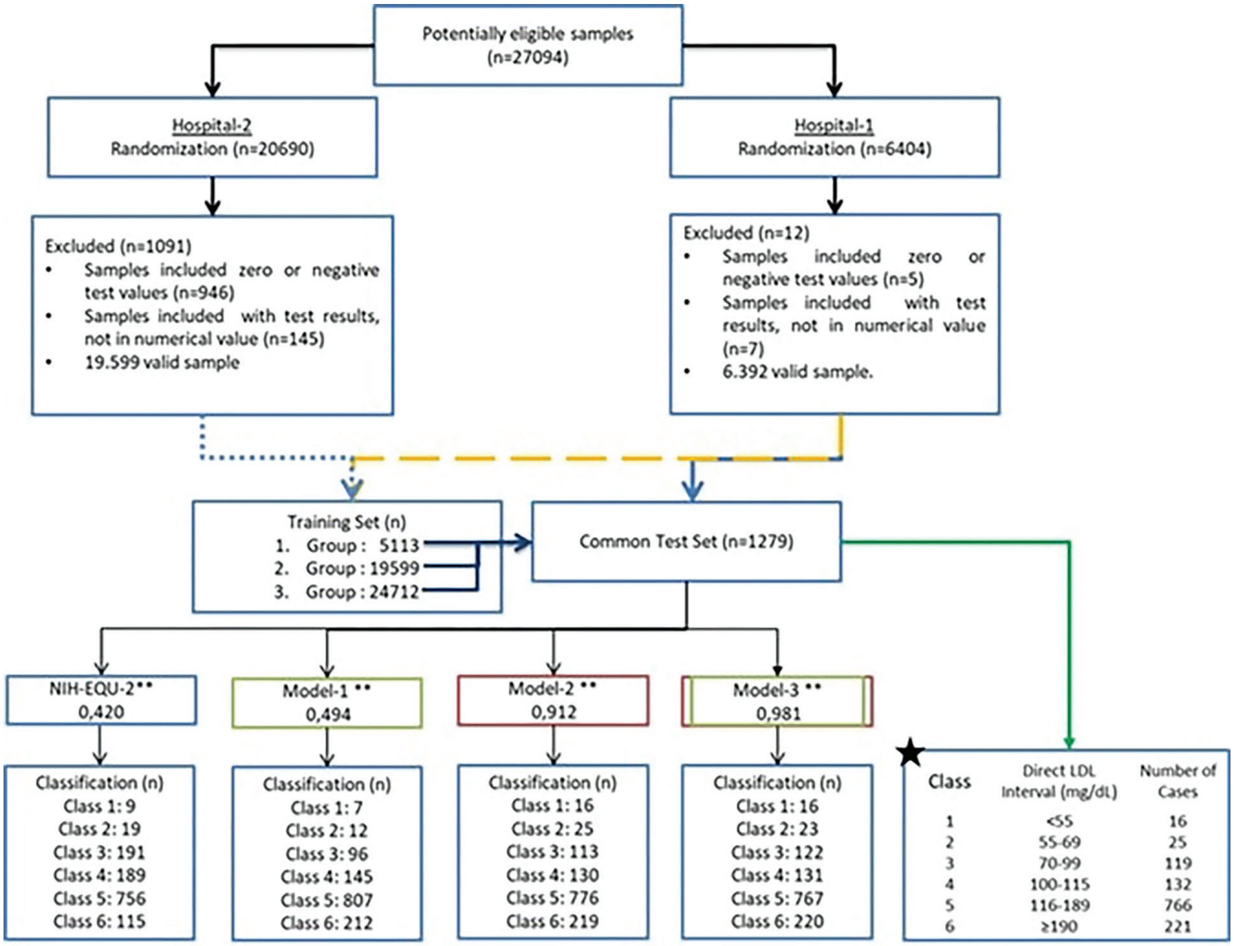

Figure 1. The flow of the subjects through the study shown in standards for reporting diagnostic accuracy diagram

NIH-Equ-2: National Institutes of Health-Equ-2, LDL: Low-density lipoprotein

SHAP graph for the LDL dataset is presented in Figure 2. Additionally, the GradientBoost library (version 1.5.0) was used to implement gradient boosting algorithms during model training and testing. A detailed literature search identified previous LDL estimation algorithms[6-10]. Our selection of analysis methods is based on various reviews of the use of the best combined ML algorithms that can detect linear or non-linear relationships between independent and dependent variables. Three ML algorithms (decision tree, random forest, and gradient boosting) were used to test linearity in the preprocessed data.

The linearity of the new dataset all-TG, obtained by combining the hospital 1 and hospital 2 subsets, was also verified. After verifying the linearity of the high-TG data set, the high-TG analysis was considered a regression analysis, and the prediction scores for LDL values from seven individual ML algorithms were determined separately for the three data sets (hospital 1, hospital 2, and all-TG). Next, LDL values for patients at both hospitals were predicted using a combination of the three algorithms described above. The new models were constructed by stacking the highest-performing algorithms: random forest, XGBoost, and decision tree. Stacking is an ensemble ML technique that combines multiple high-performing ML algorithms to produce the highest-performing predictive model. Early results indicated that the ensemble ML method using all-TG improved predictions of LDL values and LDL-level classification. Accordingly, the highest-performing random forest and XGBoost models were used as base learners, and the decision tree algorithm was stacked as the meta-learner to produce a meta-model.

Hence, a new stacked ensemble model was designed in our study. The untuned model was tested on the all-TG dataset and was later tuned. Hyperparameter tuning is a technique in ML model design used to achieve the highest final test scores across all performance parameters. After hyperparameter tuning, the highest-performing stacked ensemble ML model (model-3) shown in Figure 3 was obtained. The start-up model was model 1. Model 2 was obtained using only the hospital 2 dataset. Model-1 is our previous p-LDL-M {2} model in work[4]. The performances of all three models were tested. Model-3's LDL prediction was tested on the all-TG dataset to evaluate the effects of a larger dataset and model improvements on predictive performance. The predicted LDL values were placed into LDL-level classes in the final step, as shown in Figure 3. During the above design and selection processes, all key steps and recommendations of the IFFC for developing a medical ML application were followed. Figure 3, supported by the above-detailed explanations of our design's architecture, meets the reproducibility recommendation of IFFC.

Statistical Analysis

The measured direct LDL was accepted as the actual value. The predicted and calculated LDL values were compared with the actual LDL values. Statistical analyses were performed using IBM® SPSS® Statistics 26 for Windows®. A paired t-test was used to compare the means. Pearson's and Spearman's correlation tests were performed to assess the association between direct LDL and the predictions of the designed ML models and the Sampson-NIH equation (hereafter referred to as the NIH-Equ-2 method). Sometimes, the two correlations

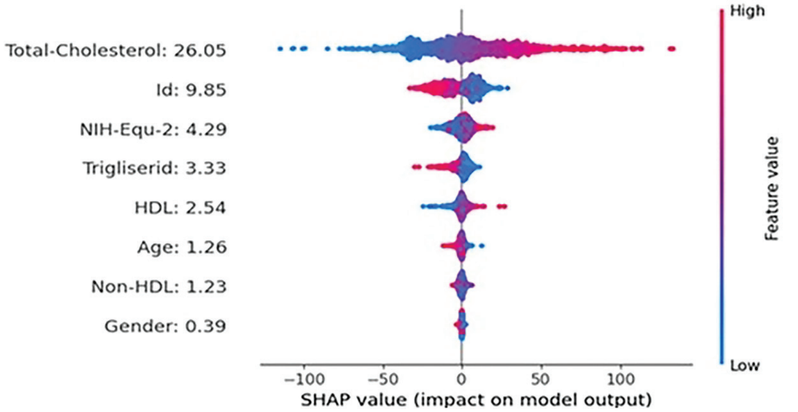

Figure 2. SHAP graph of direct LDL dataset parametersfeatures

SHAP: SHapley Additive exPlanations, LDL: Low-density lipoprotein

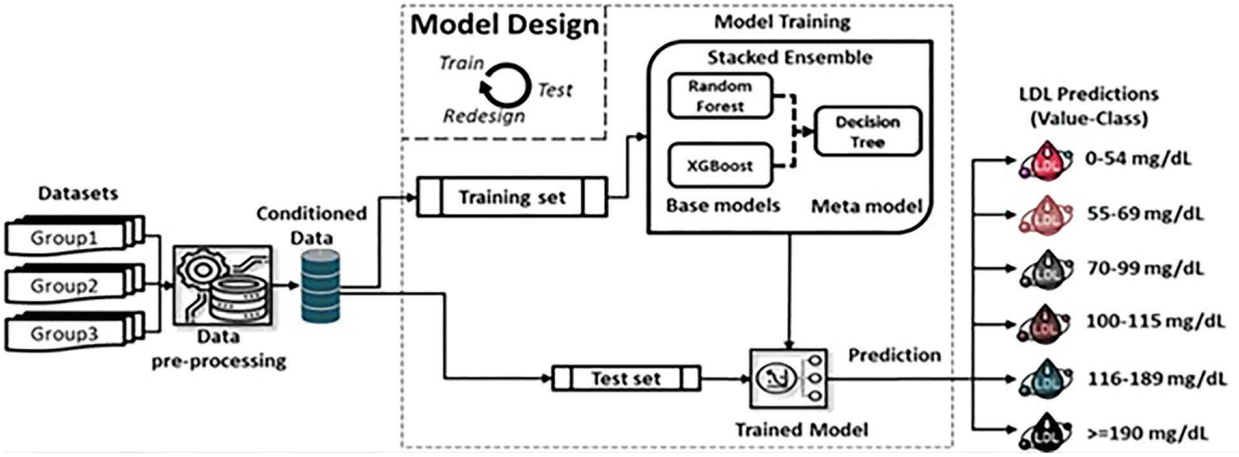

Figure 3. The proposed highest-performing LDL prediction ML model-3 architecture

LDL: Low-density lipoprotein, ML: Machine-learning

disagree on the strength of the correlation between an independent and dependent variable because of outliers[6]. Therefore, we included both in our work to determine whether such a disagreement existed. The present study found no discrepancy between the Pearson and Spearman correlation coefficient matrices; only minor differences, all less than 1.000, were observed.

Statistical significance was set at p<0.05, and a Passing-Bablok regression was conducted to determine the agreement between the prediction models and the current measurements. Bland-Altman plots were used to assess systematic bias across different direct LDL concentrations. In the Bland-Altman plots, differences between methods were plotted against direct LDL measurements.

The LDL levels classification performance of the designed models and the NIH-Equ-2 in classifying them was also assessed in accordance with the 2019 ESC/EAS Guidelines. Each subject's predicted or calculated LDL level class was compared with the subject's actual LDL level class. One way to measure and compare ML model performance is to report precision, recall, balanced accuracy, F1 score, and specificity for the model's predictions. The parameters used in the calculations for the equations in our study are defined as follows:

- **True positive:** The number of cases when the subject's LDL class was correctly identified.

- **False positive:** The number of cases when the subject's LDL class is incorrectly identified.

- **True negative:** The number of cases when the subjects out of an LDL class are correctly identified (not applicable in our study).

- **False negative:** The number of cases that subjects out of an LDL class is incorrectly identified Cohen's Kappa statistic was used to assess agreement between the designed models and NIH-Equ-2 classifications. The Kappa result can be interpreted as follows: values ≤0 indicating no agreement, 0.01-0.20 as none to slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1.00 as almost perfect agreement[6].

Results

Basic Statistics Results

A correlation matrix of designed models, NIH-Equ-2, and the actual direct LDL is given in Figure 4. The Figure shows that all models and NIH-Equ-2 results are strongly correlated with the exact values. However, the correlation for model-3 is exceptionally high (r=0.996). NIH-Equ-2's correlation with the actual values is the lowest, at 0.862.

The analysis of the scatter correlation plots of the compared methods (Figure 5) showed that the NIH-Equ-2 results were scattered, and the R2 value was low (R2=0.7443). Model-3 produced the best results, with R^2 close to 1 (R^2=0.9923) and low scatter. Interim model-2 exhibited a slightly high degree of scatter, with an R^2 value of 0.9494.

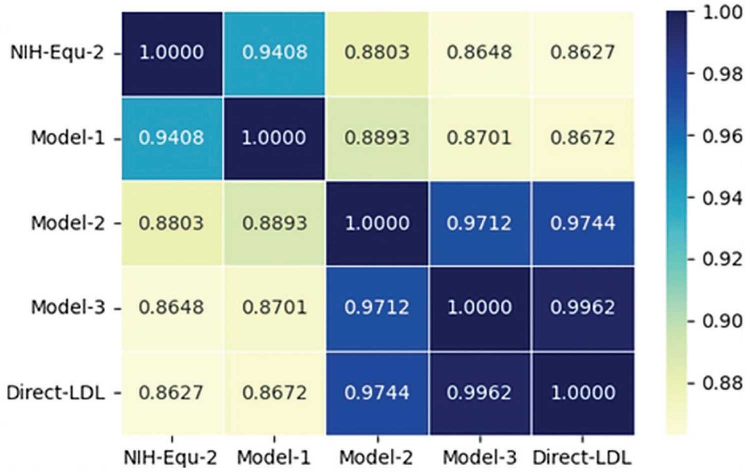

Figure 4. Correlation matrix of NIH-Equ-2, designed ML models and Direct LDL

NIH-Equ-2: National Institutes of Health-Equ-2, LDL: Low-density lipoprotein, ML: Machine-learning

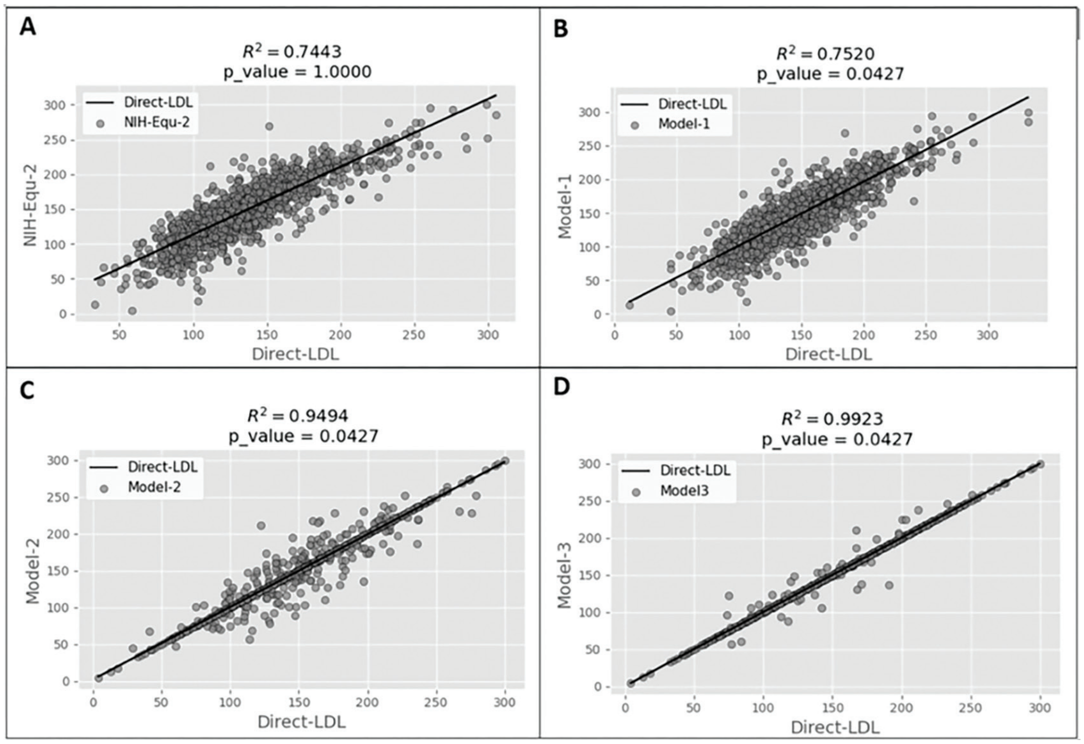

Figure 5. Scatter plots of correlations between predicted and direct LDL

A: NIH-Equ-2 vs. direct-LDL scatter graph, B: Model-1 vs. direct-LDL scatter graph, C: Model-2 vs. direct-LDL scatter graph, D: Model-3 vs. direct-LDL scatter graph, NIH-Equ-2: National Institutes of Health-Equ-2, LDL: Low-density lipoprotein

The receiver operating characteristic (ROC) curves for the six classes predicted by model-3 are shown in Figure 6. The area under the ROC curve (AUC) indicates the performance of a model across all possible classification thresholds. A value greater than 0.9 is considered outstanding. Our ROC curves showed a micro-averaged AUC of 97% across five classes.

The AUC for the non-critical class 3 was the lowest [89%; 95% confidence interval (CI), 8493%]. Therefore, the average AUC

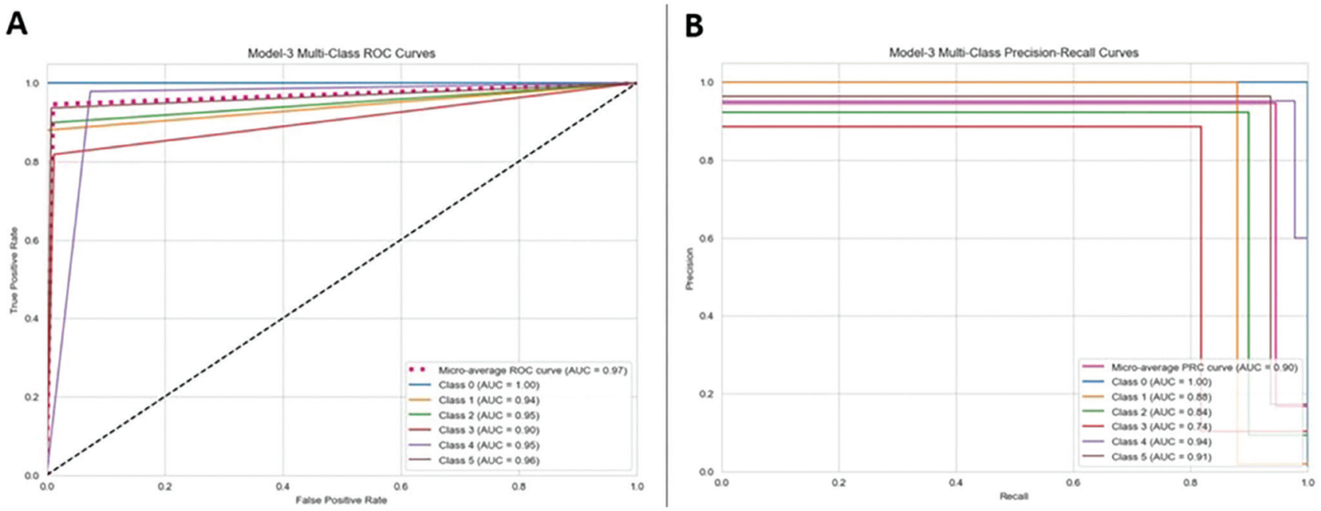

Figure 6. ROC curve and PRC of model-3 class predictions

A: Model-3 multi-class ROC curves, B: Model-3 multi-class precision-recall curves, ROC: Receiver operating characteristic, PRC: Precision-recall curve, AUC: Area under the curve

indicates that our proposed model achieves good predictive accuracy and precision across all classes. The average AUC of the precision-recall curve is also satisfactory at 0.90 (95% CI, 0.86-0.93). However, the AUC of class 3 is the lowest at 0.74 (95% CI, 0.71-0.77). Fortunately, class 3 is not the level at which LDL underestimation is critical. The AUC for the class 4 critical LDL level is high at 0.94 (95% CI, 0.90-0.98).

The Bland-Altman plot of direct LDL and model-3 is given in Figure 7. As observed, most measurements are above the mean and fall within the 95% CI. However, some values do not fall within the 95% CI, which requires an explanation. The reason for the high number of outliers in the CI is presented in the next section.

Kappa scores (Table 3) were obtained by comparing the actual LDL levels with the predicted and calculated LDL levels. The lowest Kappa score was 0.420 for NIH-Equ-2's LDL-level classification, while the highest was 0.981 for model-3. These results indicate that the performance of model 3 was best when data from hospital 1 and hospital 2 were combined.

Discussion

Our study aims to predict LDL levels in high-TG subjects. Based on our literature review, this study is the second ML-based study on high-TG subjects in Türkiye.

The categorical classification of patients' LDL levels is as important as the quantitative LDL value for guiding lipid-lowering therapy. Clinicians apply various treatments, from dietary changes and exercise to multidrug therapies, depending on the patient's LDL level. Therefore, the LDL values under study were categorized into classes according to the 2019 ESC/EAS Guideline(5). The data preparation, model selection, design, and validation steps were completed in accordance with the IFFC recommendations. The most important findings of our research are discussed below.

As illustrated in Figure 1, class 1 (0-54 mg/dL) had the lowest number of cases, whereas class 5 (116-189 mg/dL) had the highest (766 cases). The mean values of the datasets play a crucial role in representing the characteristics of the studied population. Upon examining the lipid profiles of the individuals included in our research, it was observed that the average TC, TG, and LDL levels were elevated, whereas average HDL levels were comparatively low compared with similar ML studies(7,9,11). These discrepancies may be attributed to the dietary patterns prevalent in our country. Nevertheless, with the exception of TG levels, the lipid values reported by the NIH in the multicenter study by Sampson et al. (12) were largely consistent with ours. In contrast, the other four centers reported lower lipid values than those in our study.

In prior studies focused on low-TG LDL prediction, random forest has been the most commonly used ML algorithm. However, alternative approaches such as XGBoost, deep neural networks, support vector machines, linear regression,

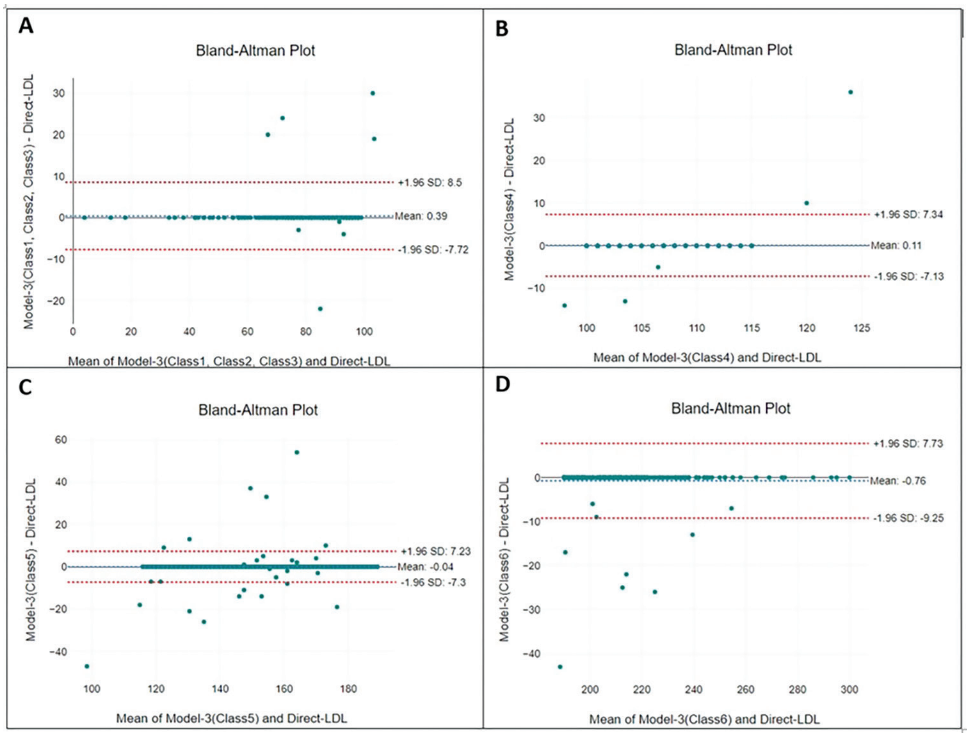

Figure 7. Bland-Altman plot between direct LDL and model-3

LDL: Low-density lipoprotein, SD: Standard deviation

Table 3. Model-1, model-2, model-3, and NIH-Equ-2 predicted Kappa scores of direct

Model/formula	Cohen's Kappa score
NIH-Equ-2	0.420[a]
Model-1	0.494[a]
Model-2	0.912[b]
Model-3	0.981[b]

[a]: Moderate aggrement, [b]: Almost perfect aggrement, NIH-Equ-2: National Institutes of Health-Equ-2

and k-nearest neighbors have also been employed[8,9,11]. In the present study, the stacked ensemble ML model-3 demonstrated superior performance and yielded the most accurate predictions.

Model 3 yielded several noteworthy findings. When its predictions were compared with direct LDL measurements, model-3 demonstrated the highest accuracy and correlation coefficients and the lowest error rates (mean absolute error, mean squared error, mean absolute percentage error). Notably, the best performance was achieved using the full all-TG dataset rather than the high-TG subset. When the entire dataset, including calculated LDL values, was utilized, model-3's prediction performance improved by 12.90% compared with predictions based solely on the high-TG group. Additionally, classification of LDL levels exhibited superior accuracy and minimal variability when using the all-TG dataset. Model-3 also outperformed the well-established NIH-Equ-2 method, showing a 13.45% improvement margin. These findings highlight the advantage of combining data from hospital 1 and hospital 2 to enhance LDL estimation. Moreover, the study confirms the following:

- The previously known strong correlations between TC, non-HDL, and direct LDL.
- The strong performance of NIH-Equ-2 in calculating LDL.
- The success of ML in estimating LDL values.
- The linear relationship between TG and LDL.

Model-3-predicted results and direct LDL measurements were significantly correlated (r=0.996). The algorithm results of Anudeep et al.[7] and Singh et al.[9] were also significantly associated with the direct-LDL measurements (0.98 and 0.982). The more robust correlation results in independent studies indicate that ML algorithms and ensemble techniques can predict LDL values better than previously developed formulae[13].

Another interesting result was the difference between NIH-Equ-2 mean and the direct LDL mean (136.12±39.38 and 148.80±44.42 mg/dL, respectively). The relatively large difference was disappointing. In contrast, the mean value obtained in model 3 did not differ significantly. Our model-3's superior statistical performance was further supported by higher precision, recall, balanced accuracy, F1 score, and specificity.

The resulting SHAP graph is shown in Figure 2. TC was the most impactful feature in our SHAP graph (Figure 2). The impact of TC was validated by the highest Pearson correlation value, 0.844, in Figure 8. There was also a high correlation between TC and direct-LDL in the study by Chen et al.[14].

Beyond TC, the SHAP summary plot revealed that TG and non-HDL cholesterol made substantial positive contributions to the prediction of LDL, reflecting the well-known metabolic coupling between TG-rich lipoproteins and LDL particles. HDL-C exerted a mild inverse effect, consistent with its protective role in reverse cholesterol transport. Age showed a modest positive effect, whereas sex contributed minimally, likely because lipid distributions were similar between sexes in the dataset. These findings support the biological plausibility of the model outputs.

The scatter plot of model-3 results in Figure 5 shows that our model results are almost linear, in contrast to the scattered results of NIH-Equ-2. Model-3's p-value (Figure 5 4.27%) is lower than the %5.46 of desirable biological variation database specifications for the LDL cholesterol[15]. Our scatter performance is also consistent with Anudeep et al.[7] low-scatter study. Our study also agreed with Anudeep et al.[7] find that different formulae can produce negative results, even though their correlation values (r) vary between 0.89 and 0.94.

The R^2 value of model-3 in our study was comparable to, yet slightly higher than, the R^2 reported in the study by Chen et al.[14] In the research conducted by Kim et al.[16] where the XGBoost method-also employed in our study-was applied consecutively, the R^2 value was relatively high but still moderately lower than that achieved by our model.

Extensive evidence from epidemiological studies, Mendelian randomization analyses, and randomized controlled trials has established a log-linear association between LDL levels and ASCVD. Consequently, clinical guidelines consistently emphasize lipid-lowering therapies as essential for improving ASCVD-related outcomes. The effectiveness of these interventions is supported by foundational scientific research, clinical data, genetic studies, randomized trials, and population-based analyses[17,18]. Furthermore, LDL concentrations directly inform the selection and dosage of cholesterol-lowering treatments. One study, for instance, reported that each 1 mmol/L reduction in LDL was associated with a 20% decrease in major cardiovascular events[19].

Previous studies have shown that traditional predictive models and formulae exhibit greater error rates at lower LDL concentrations (<70 mg/dL). In contrast, our model-3 achieved a classification error rate of just 4.8% (2 out of 41) in this range (Table 4), outperforming the Weill-Cornell model, which had an error rate of 7.5%[9]. Similar performance was observed in the study by Çubukçu and Topcu[11] although it is important to note that their cohort consisted of patients with TG levels between 177 and 399 mg/dL[9]. Based on the 2019 ESC/EAS Guidelines, our model exhibited a 3.4% (4/119) classification error across the first three LDL categories (LDL <100 mg/dL). For comparison, error rates were 43.75% (77/176) for the NIH-Equ-2 formula, 11.4% (143/1254) for the Weill-Cornell model, and 3.47% (53/1528) for the model by Anudeep et al.[7] In the study by Barakett-Hamade et al.[8] LDL values were categorized into three groups, with the lowest group defined as <80 mg/dL. The misclassification rate for this category was 12.5% (793/6327).

It is well established that elevated LDL levels contribute significantly to morbidity and mortality among patients with cardiovascular disease; intensive hyperlipidemia management has been shown to improve quality of life, particularly in patients who are over-classified[20,21]. A correct or slightly higher classification of LDL levels (over-classification) ensures that patients receive appropriate or more aggressive treatment regimens. In our study, model 3 occasionally overclassified LDL levels by one category (Table 4). However, we argue that this slight overestimation poses minimal clinical risk, as it would lead to intensified treatment, which is generally safer than the risk of undertreatment[22]. Notably, therapies such as ezetimibe and monoclonal antibodies, when added to statin-based treatment, effectively reduce LDL levels and improve cardiovascular outcomes and overall survival. Studies have shown that inclusion of ezetimibe lowers the risk of cardiovascular events without increasing adverse effects or toxicity, even in patients with acute coronary syndromes or with already optimal LDL

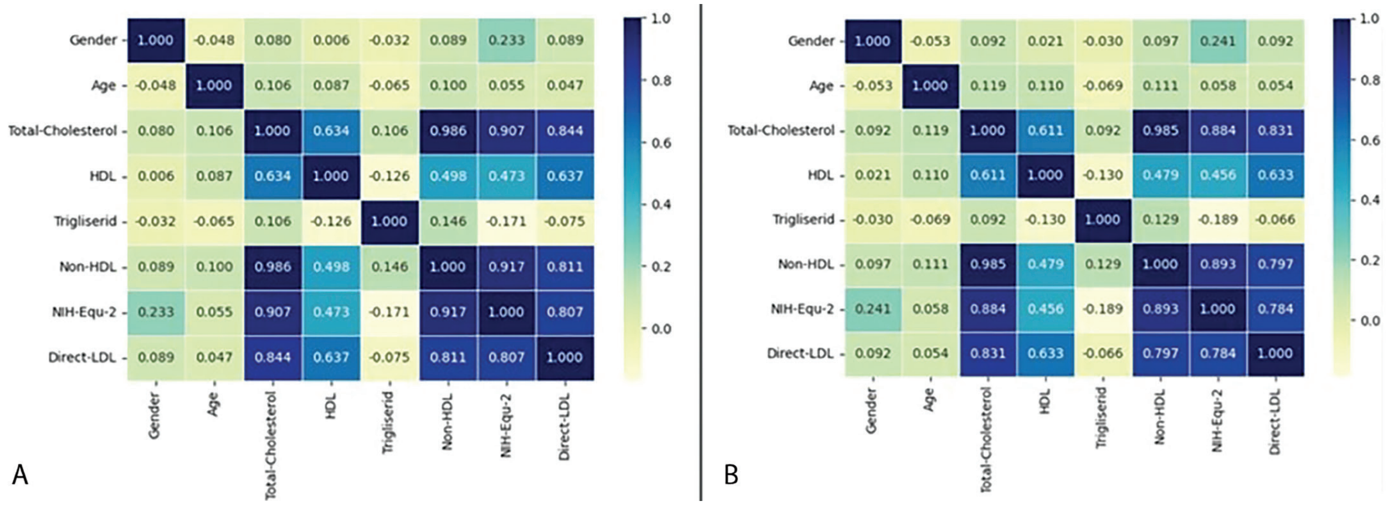

Figure 8. Correlation matrices of direct LDL parameters

A: Pearson correlation matrix, B: Spearman correlation matrix, NIH-Equ-2: National Institutes of Health-Equ-2, LDL: Low-density lipoprotein, HDL: High-density lipoprotein

Table 4. Results of classifications by model-3 (n=1279)

Model-3 class	**Direct LDL class**						
	1	**2**	**3**	**4**	**5**	**6**	**Total**
1	16	0	0	0	0	0	16
2	0	23	0	0	0	0	23
3	0	2	117	2	1	0	122
4	0	0	1	128	2	0	131
5	0	0	1	2	762	2	767
6	0	0	0	0	1	219	220
Total direct LDL	16	25	119	132	766	221	1279

LDL: Low-density lipoprotein

levels. Moreover, there is currently no evidence suggesting harm from adding ezetimibe in cases with moderate LDL concentrations[23,24]. In contrast, under-classification may have more serious consequences, such as misclassifying a high-risk patient from class 5 to class 4, resulting in inadequate treatment. Importantly, the under-classification rate of model 3 was remarkably low-only 0.39% (three of 766 cases).

Lipid-lowering therapies impose a considerable financial burden on healthcare systems globally[25]. Although beta-quantification is considered the reference standard for LDL measurement, it is not feasible for routine clinical use due to its high cost and labor intensity[26]. As a result, enzymatic and homogeneous immunoassay methods have largely replaced it. Following the establishment of correlations between LDL and other lipid fractions, direct LDL measurement techniques were gradually replaced by calculation-based methods such as the Friedewald, Martin-Hopkins, and NIH-Equ-2 formulae. Each of these approaches offers distinct advantages. For instance, whereas combined hospital datasets perform well when TG levels exceed 400 mg/dL, NIH-Equ-2 can extend its calculations up to 800 mg/dL. Notably, the Martin-Hopkins method utilizes variable adjustment factors tailored to TG levels, allowing for accurate estimations even in non-fasting individuals[9]. To overcome the limitations of traditional formulae, recent studies have shifted focus toward ML-based LDL prediction models. Based on our literature review, this study is the second to use ML algorithms to predict LDL levels in patients with TG >400 mg/dL, following our initial publication. This makes our work particularly relevant in clinical settings where precise LDL classification is crucial for initiating appropriate therapeutic interventions[5]. The effectiveness of our model is further validated by its Kappa score, a statistical measure of classification agreement. Model-3 achieved a Kappa score of 0.981, significantly outperforming the NIH-Equ-2 formula, which had a Kappa score of 0.420. These results indicate that model-3 is in "almost perfect agreement" with direct LDL measurements, whereas NIH-Equ-2 falls under the category of "moderate agreement". Although numerous studies endorse the reliability of NIH-Equ-2 for LDL estimation[12,27], it is important to note that NIH-Equ-2 was calibrated against LDL values derived from the reference beta-quantification technique. In contrast to many existing studies, our research utilized results from two distinct auto-analyzer platforms. Therefore, variations in accuracy may stem from differences in analytical methods, ML models, or the analyzers themselves[26].

The number of predictions deviating from the mean in the Bland-Altman plot comparing direct LDL and model-3 was high. The highest number deviation is observed in classes 5 and 6 (Figure 7). The large numbers in both classes are due to the very wide range of class 5 (116-189) and the open-ended range of class 6 (>190). The number of deviations from the acceptable region for the critical class 4 is four. The percentage of undesired predictions is negligible, with 4 instances out of 1279 total predictions, i.e., 0.31%.

The recommendations of the IFCC Working Group[3] have been followed carefully in our present work. In the last stage, the prediction results for the same high-TG patients were used to cross-validate the final model's reliability. Thus, we can now claim the model's applicability to data from different hospitals. We believe that our work will increase acceptance of ML in MAI applications and pave the way for future real-world applications.

Study Limitations

Our study has some limitations. First, because our data were retrospective and beta-quantification is not used in routine laboratory practice, the more commonly used direct homogeneous immunoassay was employed. In this study, LDL-C values were obtained using a commercial homogeneous direct assay. However, previous studies have shown that homogeneous LDL-C methods can overestimate LDL-C levels by 7-12% compared to the beta-quantification[28]. This discrepancy is particularly pronounced in hypertriglyceridemic individuals (TG ≥ 400 mg/dL), where cholesterol from very LDL and intermediate DL fractions may be mistakenly included in LDL-C measurements. As a result, the LDL-C values used for training our model may contain a systematic bias. Future studies should validate the model using beta-quantification-based reference data or incorporate an adjustment to account for potential measurement bias in homogeneous LDL-C assays. Secondly, the effects of diseases that may influence the lipid profiles could not be assessed independently because of the study's retrospective design. Third, sub-categorization of ethnic groups could not be performed. Since the target range was TG >400 mg/dL, TG values could not be subcategorized or analyzed in detail.

Conclusion

An MAI application that fully complied with the IFFC recommendations for predicting LDL using ensemble ML methods was presented. Performance results indicate that our newly designed ML estimation model, model-3, for predicting target high-TG values outperforms the NIH-Equ-2 formula and our previous model. Our work can be included in the routine lipid profile without changing the main principles and methods if similar work is planned as a multicenter study, enriched with data from different races, and expanded to include multiple autoanalyzers.

Ethics

Ethics Committee Approval: Before commencing the study, the necessary approval was obtained from the Non-Interventional Ethics Committee of University of Health Sciences Türkiye, İzmir Tepecik Education and Research Hospital, (approval no: 2023/13-23, date: 12.04.2023) and Non-Interventional Ethics Committee of Dokuz Eylul University Faculty of Medicine (approval no: 2023/20-04, date: 14.06.2023).

Informed Consent: Retrospective design.

Footnotes

Authorship Contributions

Surgical and Medical Practises: F.D., M.E., Ö.G.D., P.A., Concept: F.D., M.E., M.H.Ö., P.A., Design: F.D., M.E., M.H.Ö., Ö.G.D., P.A., Data Collection or Processing: F.D., M.E., Ö.G.D., Analysis or Interpretation: F.D., M.E., M.H.Ö., Ö.G.D., Literature Search: F.D., M.E., M.H.Ö., Ö.G.D., P.A., Writing: F.D., M.E.

Conflict of Interest: No conflict of interest was declared by the authors.

One of the authors of this article (F.D.) is a member of the Editorial Board of this journal. He was completely blinded to the peer review process of the article.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6:94-8.
2. Park SH, Choi J, Byeon JS. Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. Korean J Radiol. 2021;22:442-53.
3. Master SR, Badrick TC, Bietenbeck A, Haymond S. Machine learning in laboratory medicine: recommendations of the IFCC Working Group. Clin Chem. 2023;69:690-8.
4. Demirci F, Emec M, Gursoy Doruk O, Ormen M, Akan P, Hilal Ozcanhan M. Prediction of LDL in hypertriglyceridemic subjects using an innovative ensemble machine learning technique. Turkish Journal of Biochemistry. 2024;48:641-52.
5. Mach F, Baigent C, Catapano AL, et al. 2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk. Eur Heart J. 2020;41:111-88.
6. McHugh ML. Interrater reliability: the Kappa statistic. Biochem Med (Zagreb). 2012;22:276-82.
7. Anudeep PP, Kumari S, Rajasimman AS, Nayak S, Priyadarsini P. Machine learning predictive models of LDL-C in the population of eastern India and its comparison with directly measured and calculated LDL-C. Ann Clin Biochem. 2022;59:76-86.
8. Barakett-Hamade V, Ghayad JP, Mchantaf G, Sleilaty G. Is machine learning-derived low-density lipoprotein cholesterol estimation more reliable than standard closed form equations? Insights from a laboratory database by comparison with a direct homogeneous assay. Clin Chim Acta. 2021;519:220-6.
9. Singh G, Hussain Y, Xu Z, et al. Comparing a novel machine learning method to the Friedewald formula and Martin-Hopkins equation for low-density lipoprotein estimation. PLoS One. 2020;15:e0239934.
10. Hidekazu I, Nagasawa H, Yamamoto Y, et al. Dataset dependency of low-density lipoprotein-cholesterol estimation by machine learning. Ann Clin Biochem. 2023;60:396-405.
11. Çubukçu HC, Topcu Dİ. Estimation of low-density lipoprotein cholesterol concentration using machine learning. Lab Med. 2022;53:161-71.
12. Sampson M, Ling C, Sun Q, et al. A new equation for calculation of low-density lipoprotein cholesterol in patients with normolipidemia and/or hypertriglyceridemia. JAMA Cardiol. 2020;5:540-8.
13. Atabi F, Mohammadi R. Clinical validation of eleven formulas for calculating LDL-C in Iran. Iran J Pathol. 2020;15:261-7.
14. Chen L, Rong C, Ma P, Li Y, Deng X, Hua M. A new equation for estimating low-density lipoprotein cholesterol concentration based on machine learning. Medicine (Baltimore). 2024;103:e37766.
15. Westgard J. Desirable biological variation database specifications. Available from: https://www.westgard.com/clia-a-quality/quality-requirements/238-biodatabase1.html
16. Kim Y, Lee WK, Lee W. Prediction of low-density lipoprotein cholesterol levels using machine learning methods. Lab Med. 2024;55:471-84.
17. Writing Committee; Lloyd-Jones DM, Morris PB, et al. 2022 ACC expert consensus decision pathway on the role of nonstatin therapies for LDL-cholesterol lowering in the management of atherosclerotic cardiovascular disease risk: a report of the American College of Cardiology Solution Set Oversight Committee. J Am Coll Cardiol. 2022;80:1366-418. Erratum in: J Am Coll Cardiol. 2023;81:104.
18. Catapano AL, Graham I, De Backer G, et al. 2016 ESC/EAS Guidelines for the management of dyslipidaemias. Eur Heart J. 2016;37:2999-3058.
19. Cholesterol Treatment Trialists' (CTT) Collaboration; Baigent C, Blackwell L, et al. Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials. Lancet. 2010;376:1670-81.
20. Emerging Risk Factors Collaboration; Di Angelantonio E, Gao P, et al. Lipid-related markers and cardiovascular disease prediction. JAMA. 2012;307:2499-506.
21. Ference BA, Yoo W, Alesh I, et al. Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: a Mendelian randomization analysis. J Am Coll Cardiol. 2012;60:2631-9.
22. Sathiyakumar V, Blumenthal RS, Elshazly MB: New information on accuracy of LDL-C estimation. American College of Cardiology. March 20. 2020. (Accessed: March 27, 2024). Available from: https://www.acc.org/latest-in-cardiology/articles/2020/03/19/16/00/new-information-on-accuracy-of-ldl-c-estimation
23. Oyama K, Giugliano RP, Blazing MA, et al. Baseline low-density lipoprotein cholesterol and clinical outcomes of combining ezetimibe with statin therapy in IMPROVE-IT. J Am Coll Cardiol. 2021;78:1499-507.
24. Koskinas KC, Siontis GCM, Piccolo R, et al. Effect of statins and non-statin LDL-lowering medications on cardiovascular outcomes in secondary prevention: a meta-analysis of randomized trials. Eur Heart J. 2018;39:1172-80.
25. Michaeli DT, Michaeli JC, Boch T, Michaeli T. Cost-effectiveness of lipid-lowering therapies for cardiovascular prevention in Germany. Cardiovasc Drugs Ther. 2023;37:683-94.
26. Contois JH, Langlois MR, Cobbaert C, Sniderman AD. Standardization of apolipoprotein B, LDL-cholesterol, and Non-HDL-cholesterol. J Am Heart Assoc. 2023;12:e030405.
27. Sampson M, Wolska A, Meeusen JW, Otvos J, Remaley AT. The Sampson-NIH Equation is the preferred calculation method for LDL-C. Clin Chem. 2024;70:399-402.
28. Yano M, Matsunaga A, Harada S, et al. Comparison of two homogeneous LDL-cholesterol assays using fresh hypertriglyceridemic serum and quantitative ultracentrifugation fractions. J Atheroscler Thromb. 2019;26:979-88.