Comparative Analysis of Machine Learning Models for LDL Cholesterol Estimation

Barthi Dasan

Universitas Nusa Mandiri, Indonesia

E-mail: barthi77@gmail.com

Abstrak. Estimasi low-density lipoprotein cholesterol (LDL-C) yang akurat sangat penting dalam penilaian risiko kardiovaskular dan pengambilan keputusan terapi. Metode estimasi LDL-C berbasis formula tradisional, seperti persamaan Friedewald, Sampson, dan Martin, menunjukkan penurunan akurasi pada kadar trigliserida (TG) yang tinggi. Studi ini membandingkan sembilan model machine learning (ML) dengan formula konvensional menggunakan dataset besar yang terdiri dari 120.174 subjek. Setelah prapemrosesan data dan seleksi fitur, empat prediktor utama (TC, TG, HDL-C, dan usia) digunakan untuk melatih model ML dengan validasi silang 5-fold. Di antara seluruh model, Light Gradient Boosting Machine (LightGBM) menunjukkan kinerja terbaik dengan R^2 = 0,8749, MSE = 204,53 mg^2/dL^2, dan PCC = 0,935 pada internal test set. Kinerja superior yang serupa juga diamati pada external validation cohort (n = 10.183), terutama pada kategori hypertriglyceridemia (TG ≥ 200 mg/dL), di mana formula konvensional mengalami penurunan performa yang signifikan. Model ML, khususnya pendekatan berbasis ensemble, mempertahankan akurasi prediksi yang stabil di seluruh rentang TG dan secara nyata mengurangi kesalahan prediksi pada ambang klinis LDL-C yang relevan (70, 100, dan 130 mg/dL). Temuan ini mendukung integrasi estimasi LDL-C berbasis ML ke dalam alur kerja laboratorium rutin dan menyoroti potensinya dalam mendukung pengambilan keputusan klinis.
Kata kunci: Lipid; Kolesterol LDL; Machine Learning; Trigliserida

Abstract. *Accurate estimation of low-density lipoprotein cholesterol (LDL-C) is essential for cardiovascular risk assessment and treatment decision-making. Traditional formula-based LDL-C estimations, such as Friedewald, Sampson, and Martin equations, show decreasing accuracy at higher triglyceride (TG) levels. This study compares nine machine learning (ML) models against conventional formulas using a large dataset of 120,174 subjects. After data preprocessing and feature selection, four predictors (TC, TG, HDL-C, and age) were used to train ML models with 5-fold cross-validation. Among all models, Light Gradient Boosting Machine (LightGBM) demonstrated the best performance, achieving R^2 = 0.8749, MSE = 204.53 mg^2/dL^2, and PCC = 0.935 on the internal test set. Similar superiority was observed in the external validation cohort (n = 10,183), particularly in hypertriglyceridemic ranges (TG ≥ 200 mg/dL), where classical equations showed substantial performance degradation. Machine learning models, especially ensemble-based approaches, maintain robust predictive ability across TG strata and significantly reduce error around clinically relevant LDL-C thresholds (70, 100, and 130 mg/dL). These findings support the integration of ML-assisted LDL-C estimation into routine laboratory workflows and highlight its potential contribution to clinical decision support.*

Keywords: *Lipids; LDL cholesterol; Machine learning; Triglyceride*

1. Introduction

Atherosclerotic cardiovascular disease (ASCVD) is a leading cause of global morbidity and mortality, with elevated low-density lipoprotein cholesterol (LDL-C) being a major and extensively validated risk factor. Consequently, LDL-C reduction has been established as a primary target for both primary and secondary cardiovascular prevention in clinical practice and guidelines. Traditionally, LDL-C has been estimated using equations such as the Friedewald formula, developed in 1972, which calculates LDL-C from total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), and triglycerides (TG). However, this formula has known limitations, including limited accuracy in scenarios with high triglycerides (TG >400 mg/dL) or very low LDL-C (LDL-C <70 mg/dL), and it typically requires a fasting sample. The Martin-Hopkins equation was developed to address some of these inaccuracies by using an adjustable factor for the TG:VLDL-C ratio, and while it generally outperforms Friedewald, it still presents inaccuracies, particularly at lower LDL-C estimates. In light of these challenges and the critical need for precise LDL-C estimates to inform individualized treatment plans and monitor aggressive LDL-C lowering therapies, there has been a significant trend toward developing novel methods. Machine learning (ML) approaches, which can model complex and non-linear relationships between variables, have emerged as a promising development for more accurate LDL-C estimation. Recent studies have shown that ML models often outperform traditional formulas, particularly in challenging subgroups such as those with elevated TG and very low LDL-C.

The traditional estimation of low-density lipoprotein cholesterol (LDL-C) using formulas such as Friedewald and Martin-Hopkins has posed a significant problem in clinical practice due to their limited accuracy, particularly in scenarios involving high triglyceride (TG) levels (e.g., TG >400 mg/dL) or very low LDL-C concentrations (e.g., LDL-C <70 mg/dL). The Friedewald formula, developed in 1972, notably requires a fasting sample and can underestimate LDL-C, especially with elevated TG, while the more recent Martin-Hopkins equation, despite being an improvement, still exhibits inaccuracies at lower LDL-C estimates. Given that LDL-C is a primary target for cardiovascular disease prevention, these inaccuracies can lead to inappropriate treatment decisions, highlighting the critical need for more precise and reliable estimation methods. To address this, multiple studies have adopted the method of developing and validating novel LDL-C prediction models using various machine learning (ML) algorithms, including random forests, Gradient Boosting, artificial neural networks (ANN), and K-nearest neighbors (KNN), often utilizing standard lipid profile components (total cholesterol, high-density lipoprotein cholesterol, and triglycerides) as primary inputs, and sometimes incorporating additional clinical and laboratory parameters. Results consistently indicate that these ML models generally outperform traditional formulas, demonstrating higher correlation coefficients with directly measured LDL-C and lower root mean squared errors (RMSE). For instance, the Weill Cornell model showed a correlation of 0.982 compared to 0.950 (Friedewald) and 0.962 (Martin-Hopkins), and ML models often provide better performance across challenging subgroups like those with TG >500 mg/dL or LDL-C <70 mg/dL, leading to improved reclassification of patients according to guideline-determined thresholds. Specific ML approaches, such as the 2-step prediction model, have shown the highest accuracy (RMSE 7.015) and concordance rates (85.1%). Critically, ML models exhibit greater robustness and maintain high predictive power across varying triglyceride levels, unlike traditional formulas which see a notable decline in accuracy (e.g., R^2 values dropping below zero for Friedewald/Martin at TG >300 mg/dL). However, a comprehensive analysis reveals several limitations: many ML models are validated against homogeneous direct assays rather than the gold-standard beta-quantification method, which is labor-intensive and costly but offers superior accuracy; the generalizability of these models is often limited due to development on single-center or specific population datasets, necessitating extensive external validation across diverse demographics and different analyzer platforms; and model performance can be affected by the size of the dataset for specific

subgroups (e.g., very high TG, very low LDL-C), or by their inherent computational complexity, which requires integration into electronic health records for practical application.

In this study, we propose using the Light Gradient Boosting Machine (LightGBM), eXtreme Gradient Boosting (XGBoost), and Adaptive Boosting (AdaBoost) algorithm to predict LDL-C based on routine lipid profile data. These ensemble machine learning methods are known for their strong predictive performance, scalability, and robustness on large datasets. LightGBM offers high computational speed and low memory usage, while XGBoost is recognized for its regularization capabilities and flexibility, and AdaBoost is valued for its simplicity and effectiveness in reducing bias. We hypothesize that these models will produce more accurate and consistent LDL-C predictions across all triglyceride ranges, including in cases with elevated triglyceride levels, and will outperform traditional formulas (Friedewald, Martin, Sampson) as well as other machine learning models. By providing reliable LDL-C estimates, these ensemble methods can enhance cardiovascular risk stratification and support precise treatment decisions in clinical practice.

2. Related Work

Accurate estimation of low-density lipoprotein cholesterol (LDL-C) is critical for cardiovascular risk stratification and treatment planning. Traditionally, formulas such as Friedewald, Martin, and Sampson have been used to estimate LDL-C from standard lipid panels (TC, HDL-C, and TG). However, these formulas exhibit limited accuracy, particularly in patients with hypertriglyceridemia (TG > 300 mg/dL), where they tend to either over- or underestimate LDL-C levels due to their reliance on fixed or semi-flexible ratios.

In recent years, machine learning (ML) approaches have emerged as powerful alternatives, offering data-driven, non-linear modeling capabilities that can accommodate the complex relationships among lipid parameters. Prior studies have demonstrated the superiority of models such as Random Forest, Gradient Boosting, and Multilayer Perceptron (MLP) in LDL-C estimation, outperforming traditional equations across a range of triglyceride levels.

Light Gradient Boosting Machine (LightGBM), eXtreme Gradient Boosting (XGBoost), and Adaptive Boosting (AdaBoost) are advanced ensemble learning methods that extend the capabilities of conventional tree-based models. LightGBM, developed by Microsoft, is designed for high efficiency, fast training speed, and lower memory usage compared to other tree-based algorithms, utilizing histogram-based algorithms and leaf-wise tree growth to enhance performance on large-scale, high-dimensional data. XGBoost, recognized for its scalability and effective regularization techniques, consistently achieves high predictive accuracy in clinical datasets. AdaBoost, while simpler, effectively reduces bias by sequentially focusing on misclassified samples, offering robust performance in various regression and classification tasks.

In the context of LDL-C prediction, LightGBM offers several advantages:

- Improved generalization on diverse patient populations.
- Robustness to outliers, which are common in lipid data.
- Better scalability for use in real-time clinical settings or integration into hospital information systems.

Although previous studies have explored Random Forest and Gradient Boosting extensively, the application of LightGBM, XGBoost and AdaBoost in lipidology remains underreported. Its inclusion in this study not only expands the range of ML models evaluated but also brings a state-of-the-art ensemble learning technique into the conversation around LDL-C estimation. The model's ability to retain high predictive accuracy even in extreme TG ranges (>400 mg/dL) reinforces its potential for deployment in clinical laboratories facing the limitations of formula-based calculations.

3. Method

3.1. Study Population

This retrospective study utilized anonymized data without any direct contact or intervention with the subjects and was approved by the Ethics Review Committee of Yanbian University Hospital (Ethics No. 2024665). The cohort included consecutive samples of standard lipid profiles—directly measured total cholesterol (TC), HDL-C, triglycerides (TG), and LDL-C—collected between January 1, 2020, and March 31, 2023, from inpatient and outpatient units for clinical purposes. Inclusion criteria required all lipid components to be measured on the same day to minimize daily variations. Data were extracted via the laboratory information system (LIS). All continuous variables in this study underwent Kolmogorov-Smirnov testing (Srimani et al., 2021), using a significance level of $P > 0.05$ to assess conformity to the assumption of normal distribution.

3.2. Lipid profile testing

Serum levels of triglycerides (TG), total cholesterol (TC), HDL-C, and LDL-C were measured in the clinical laboratory of Yanbian University using the Roche Cobas 702 chemistry analyzer, which is calibrated every 14 days and operated under quality control protocols in accordance with the regulations and certification standards of the Jilin Provincial Government. TC was measured using the cholesterol oxidase-peroxidase-aminoantipyrine phenol (CHOD-PAP) method, and TG was measured using the glycerol phosphate oxidase-peroxidase-aminoantipyrine phenol (GPO-PAP) enzymatic colorimetric method (Rifai, 2006). LDL-C was assessed using the surfactant-based LDL-C assay, while HDL-C was measured using the catalase-based HDL-C assay. All assays demonstrated linear measurement ranges: TG (44.3–1,000 mg/dL or 0.5–11.3 mmol/L), TC (19.3–500 mg/dL or 0.5–12.9 mmol/L), HDL-C (3.8–96.7 mg/dL or 0.1–2.5 mmol/L), and LDL-C (7.7–450 mg/dL or 0.2–11.6 mmol/L). Calibration of TC and TG is conducted every 15 days, while LDL-C and HDL-C are calibrated daily. Additionally, two levels of quality control materials (high and low) are tested daily to ensure measurement accuracy.

3.3. Data preprocessing

The initial screening process excluded patients with missing values in TC, TG, HDL-C, LDL-C, age, or gender, as well as those with measurements falling outside the assay's detection limits, due to relative deviations exceeding 10%, which classified them as outliers. Data collected from January 2020 to December 2022 was designated for internal training and validation, while data from January to March 2023 served as a secondary internal validation set. Duplicate entries—cases with identical feature values but differing outcomes—were removed from the internal dataset. To identify and eliminate low-importance features, we implemented a multi-model consensus approach using RandomForestRegressor, DecisionTreeRegressor, XGBRegressor, and LightGBMRegressor, with each model trained on the same dataset and their feature importance scores averaged (Pedregosa et al., 2011). Features deemed least important were discarded based on the averaged scores (Fig. 1). Feature scaling was then performed using the StandardScaler from the sklearn.preprocessing module: the scaler was fit on the training data, and the resulting parameters were reused to transform the validation data, ensuring consistent scaling across datasets.

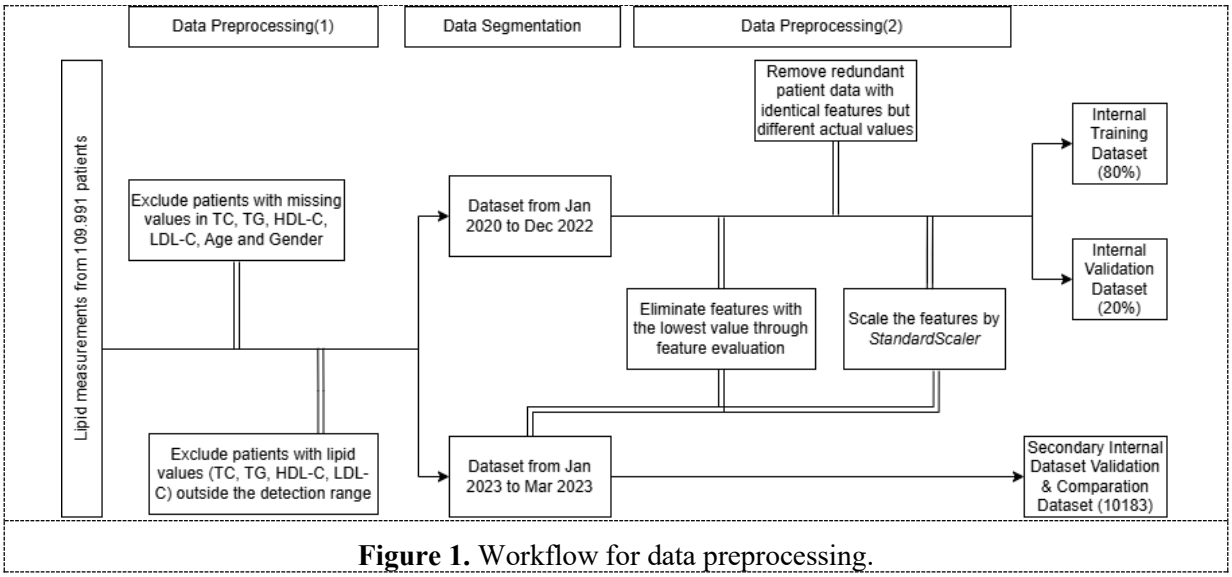

Figure 1. Workflow for data preprocessing.

3.4. Machine Learning algorithm and assessment methods

Using the Scikit-learn application programming interface (API) (Pedregosa et al., 2011), we conducted machine learning (ML) analyses to predict LDL-C values from actual measurements of total cholesterol (TC), triglycerides (TG), and HDL-C. Multiple regression models were developed, including linear regression, K-nearest neighbors (KNN), decision tree, random forest, Gradient Boosting, eXtreme Gradient Boosting (XGBoost), Adaboost, LightGBM, and multi-layer perceptron (MLP). The directly measured LDL-C values were used as ground truth labels to train and evaluate the models.

The dataset was randomly divided into training (80%) and test (20%) sets. To optimize model performance, hyperparameter tuning was performed using a combination of grid search and 5-fold cross-validation. For each algorithm, we explored a range of hyperparameters, such as learning rate, maximum tree depth, minimum samples per leaf, number of estimators for ensemble models, and the structure of hidden layers and neurons for neural networks. The goal was to identify the best hyperparameter configuration that minimized mean squared error (MSE) while maximizing both the R-squared (R^2) value and Pearson correlation coefficient (PCC) on the validation data, thereby enhancing the model's generalizability.

After selecting the optimal ML model based on internal validation results, we further evaluated it using a temporally separated internal validation set comprising data from January to March 2023. This time-based split ensured the model was tested on data collected later than the training set, although still originating from the same clinical source. In addition to evaluating model performance, we compared the ML predictions with traditional LDL-C estimation methods, including the Friedewald formula, Sampson formula, and Martin equation, using the same validation dataset.

Tabel 1. Equations for LDL-C estimation.

Friedewald Equations	LDL_C(mmol L) TC HDL_C TG5
Martin Equations	LDL_C(mg dl) TC HDL_C - TGX(adjustable coefficient)
Sampson Equations	LDL_C(mg dl) TC0948 HDL_C0971 TG859 + (TG Non_HDL_C) 2140 TG TG16100 944

Tabel 2. Conversions between mmol/L and mg/dL

mmol/L mg/dL
TG(mmol/L) * 88.57 TG(mg/dL)
TC(mmol/L) * 38.67 TC(mg/dL)
HDL-C(mmol/L) * 38.67 HDL-C(mg/dL)
LDL-C (mmol/L) * 38.67 LDL-C(mg/dL)

3.5. Hyperparameter tuning and model selection

All regression models were tuned using grid search with 5-fold cross-validation on the training set. Before model fitting, all continuous predictors (total cholesterol, triglycerides, HDL-C, and age) were standardized using a StandardScaler fitted only on the training data to avoid data leakage; the same scaler was then applied to the test set and the external validation set. For each algorithm, a predefined hyperparameter search space was constructed based on common recommendations for non-linear regression models (Table 3). The primary optimization metric during grid search was the cross-validated coefficient of determination (R^2), and the final configuration for each model was selected as the one achieving the highest mean R^2 on the validation folds. When two configurations exhibited similar performance, the one with lower model complexity (e.g., smaller depth or fewer trees) was preferred to reduce overfitting.

Feature importance was first inspected using a Random Forest regressor trained on the full set of predictors (TC, TG, HDL-C, age, and sex). Variables with a mean importance below 0.05 were considered to have negligible contribution. In this analysis, the sex variable (gender) consistently showed the lowest importance and was therefore excluded from the final models, leaving TC, TG, HDL-C, and age as the core predictors for LDL-C estimation.

All models were implemented in Python using scikit-learn, LightGBM, and XGBoost libraries. Grid search was parallelized (n_jobs = -1) to reduce computation time. The complete search spaces and the best hyperparameter configurations resulting from tuning are summarized in Table 3.

Tabel 3. Hyperparameter search space and optimal settings for each machine learning model.

Model	Hyperparameter search space	Best configuration (this study)
Linear Regression	fit_intercept ∈ {True, False}	fit_intercept = True
k-Nearest Neighbors (KNN)	n_neighbors ∈ {3,5,…,19}; weights ∈ {uniform, distance}; algorithm ∈ {auto, ball_tree}; leaf_size ∈ {20, 30}; metric ∈ {euclidean, manhattan}	n_neighbors = 19, weights = uniform, algorithm = ball_tree, leaf_size = 20, metric = euclidean
Decision Tree	min_samples_split ∈ [2–10]; min_samples_leaf ∈ [1–10]; max_features ∈ {None, sqrt, log2, 1.0}	(e.g.) min_samples_split = 5–6, min_samples_leaf = 10, max_features ≈ log2 / 1.0*
Random Forest	n_estimators ∈ {50, 100, 150}; max_depth ∈ {None, 20}; min_samples_split ∈ {2, 5}; min_samples_leaf ∈ {1, 2}; max_features ∈ {sqrt, log2}	n_estimators = 150, max_depth = 20, min_samples_split = 5, min_samples_leaf = 2, max_features = sqrt
Gradient Boosting	n_estimators ∈ {50, 100, 150}; max_depth ∈ {3, 5, 7}; learning_rate ∈ {0.01, 0.1, 0.3}; subsample ∈ {0.5, 0.7, 1.0}; max_features ∈ {sqrt, log2, None}	n_estimators = 100, max_depth = 5, learning_rate = 0.1, subsample = 1.0, max_features = None
MLP Regressor	hidden_layer_sizes ∈ {(50), (100), (50,50)}; alpha ∈ {0.0001, 0.001}; activation ∈ {relu, tanh}; solver = adam; learning_rate = adaptive; max_iter = 1000	hidden_layer_sizes = (100,), activation = relu, alpha = 0.001, solver = adam, learning_rate = adaptive, max_iter = 1000

LightGBM	n_estimators ∈ {50, 100, 150}; learning_rate ∈ {0.01, 0.1, 0.3}; max_depth ∈ {3, 5, −1}; num_leaves ∈ {15, 31, 63}; subsample ∈ {0.7, 1.0}; colsample_bytree ∈ {0.7, 1.0}	n_estimators = 150, learning_rate = 0.1, max_depth = 5, num_leaves = 15, subsample = 0.7, colsample_bytree = 1.0
XGBoost	n_estimators ∈ {50, 100, 150}; learning_rate ∈ {0.01, 0.1, 0.3}; max_depth ∈ {3, 5, 7}; subsample ∈ {0.7, 1.0}; colsample_bytree ∈ {0.7, 1.0}	n_estimators = 100, learning_rate = 0.1, max_depth = 5, subsample = 0.7, colsample_bytree = 1.0
AdaBoost	n_estimators ∈ {50, 100, 150}; learning_rate ∈ {0.01, 0.1, 0.3}	n_estimators = 50, learning_rate = 0.1

All experiments were executed on a workstation equipped with a 13th Gen Intel® Core™ i7-1365U processor (1.80 GHz), Intel Iris Xe Graphics, and 32 GB RAM running Windows 11 Pro (64-bit). Parallel processing was enabled during hyperparameter tuning using GridSearchCV with n_jobs = −1. Training time for individual models ranged from approximately 1–3 minutes depending on model complexity, with ensemble models (e.g., Gradient Boosting, Random Forest, LightGBM, XGBoost) requiring longer computation time than simpler models such as Linear Regression or k-Nearest Neighbors. The complete grid-search procedure for all nine machine learning models—including cross-validation, model fitting, and selection of optimal hyperparameters—was completed within several minutes, indicating that the workflow is computationally efficient and suitable for routine deployment in clinical laboratory environments..

3.6. LDL calculation formulas

LDL-C levels were estimated using the Friedewald, Martin, and Sampson formulas as outlined in Table 1. For the Martin method, LDL-C values were obtained using the online calculator available at http://www.LDLCalculator.com, while the Friedewald and Sampson calculations were carried out using Microsoft Excel 2021. Furthermore, any necessary unit conversions between mmol/L and mg/dL were performed based on the reference values provided in Table 2.

3.7. Statistics analysis

To assess the performance of both machine learning (ML) models and traditional LDL-C estimation formulas, we employed three widely used evaluation metrics: R-squared (R^2), mean squared error (MSE), and Pearson correlation coefficient (PCC). R^2 quantifies the proportion of variance in the target variable that can be explained by the input features. An R^2 value approaching 1 signifies strong predictive power, while a negative R^2 suggests that the model underperforms compared to simply predicting the mean (Chicco, Warrens & Jurman, 2021). Such outcomes often occur when the model fails to capture the underlying relationships in the data, especially in the presence of outliers or when the model structure is not well-suited to the data.

Lower MSE values represent better predictive accuracy, as this metric reflects the average squared difference between predicted and actual LDL-C values. The PCC was used to assess the strength of the linear relationship between predicted and observed values, with values near 1 indicating strong positive correlation. In this study, models and formulas demonstrating higher R^2 scores, lower MSE, and higher PCC were interpreted as having greater predictive accuracy. All statistical analyses and model evaluations were conducted using Python version 3.11.5.

Given the known variability in formula performance across different triglyceride (TG) concentrations, we further stratified the test data into six TG-based groups: TG <100 mg/dL, 100–149 mg/dL, 150–199 mg/dL, 200–299 mg/dL, 300–399 mg/dL, and ≥400 mg/dL. This stratification allowed for a more detailed and robust evaluation of model performance under varying TG conditions. By assessing accuracy within these distinct subgroups, we aimed to ensure that the conclusions drawn about model and formula effectiveness were consistent and applicable across a broad range of clinical lipid profiles.

Tabel 4. Descriptive Statistics.

Parameter	Q1 (25%)	Median	Q3 (75%)
TC	143.85	178.66	208.04
TG	93.88	135.51	200.17
HDL-C	35.58	43.31	52.20
LDL-C	85.07	116.78	141.53
Age	46	57	66

4. Results

4.1. Original clinical data

The study's original clinical data involved a comprehensive lipid profile analysis of 120,174 unique individuals, collected from Yanbian University Hospital between January 1, 2020, and March 31, 2023. Participants' ages spanned from 1 to 103 years, with males constituting the majority across all datasets, specifically 63,392 males (52.8% of the total cohort). It was observed that all continuous variables, including lipid levels and age, failed the normality test, indicating a non-normal distribution. The internal training and testing dataset comprised 109,991 cases from January 1, 2020, to December 31, 2022, where the median LDL-C was 116.78 mg/dL, median total cholesterol (TC) was 178.66 mg/dL, and median triglycerides (TG) were 135.51 mg/dL. For secondary internal validation, a distinct cohort of 10,183 cases from January 1, 2023, to March 31, 2023, was evaluated, showing a median LDL-C of 110.21 mg/dL, median TC of 184.46 mg/dL, and median TG of 138.17 mg/dL. Notably, the internal dataset had a higher median LDL-C but lower median TC and TG values compared to the secondary internal validation dataset.

4.2. Feature importance

Feature importance was evaluated using a comprehensive multi-regressor consensus approach to ensure robustness and mitigate model-specific biases in assessing the contribution of each variable toward LDL-C prediction. In this method, multiple regression algorithms including Linear Regression, K-Nearest Neighbors (KNN), DecisionTreeRegressor, RandomForestRegressor, GradientBoostingRegressor, Multilayer Perceptron (MLP), LightGBMRegressor, XGBRegressor, and AdaBoostRegressor were independently trained on the same dataset, and feature importance scores were extracted where applicable. For models that do not inherently provide feature importance scores, such as KNN and MLP, permutation importance was utilized to derive comparable estimates of feature relevance.

The feature importance scores from each model were then averaged to generate a unified measure reflecting the relative contribution of each predictor across diverse algorithmic perspectives. This multi-regressor consensus analysis revealed that Total Cholesterol (TC) consistently emerged as the most influential predictor of LDL-C across models, highlighting its central role in LDL-C estimation frameworks. In contrast, gender was consistently ranked with the lowest importance, indicating minimal direct predictive power relative to the lipid parameters within the dataset, even when evaluated across models with differing assumptions and architectures.

Tabel 5. Baseline characteristics of the study subjects.

	Inner dataset	**Secondary internal validation dataset**
	n = 109,249	**n = 10,183**
Age	57 (46,66)	58 (49,67)

Gender			
	Male	57,774 (52.88%)	5,208 (51.14%)
	Female	51,475 (47.12%)	4,975 (48.86%)
Lipid Profile			
	TC	178.66 (143.85, 208.04)	184.46 (148.88, 214.23)
	TG	135.51 (93.88, 200.17)	138.17 (97.43, 201.05)
	HDL-C	43.31 (35.58, 52.20)	41.74 (35.19, 49.50)
	LDL-C	116.78 (85.07, 141.53)	110.21 (80.05, 136.89)

Notes.

Data are expressed as medians (interquartile range) for continuous variables and frequencies (percentages) for categorical variables.

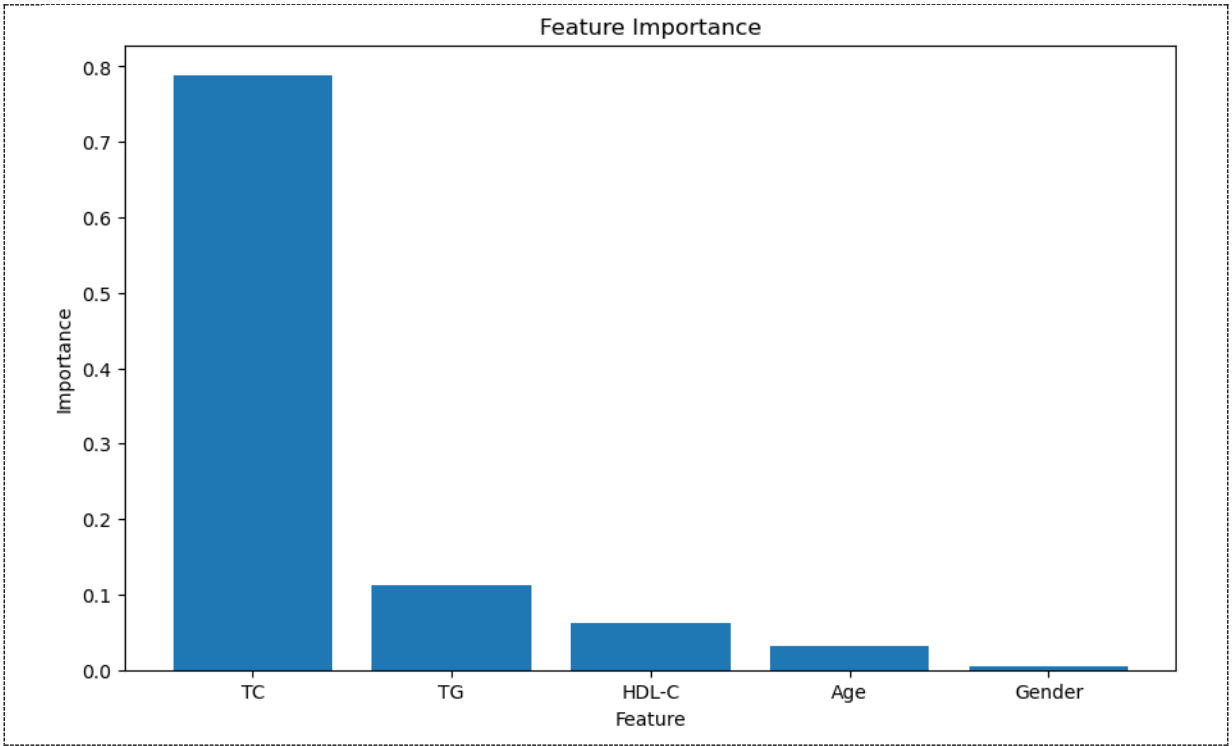

Figure 2. Feature importances.

4.3. Inner training and test

Learning curve analysis on the inner training and test data demonstrated that all nine models evaluated Linear Regression, KNN, Decision Tree, Random Forest, Gradient Boosting, MLP, LightGBM, XGBoost, and AdaBoost exhibited stable training without significant underfitting. Slight overfitting was observed in the Decision Tree and Random Forest models, reflected by higher training scores compared to test scores,

yet this did not compromise training stability. Ensemble models such as Gradient Boosting, LightGBM, and XGBoost, alongside the MLP neural network, displayed minimal train-test score gaps, indicating strong generalization and robust performance across increasing sample sizes.

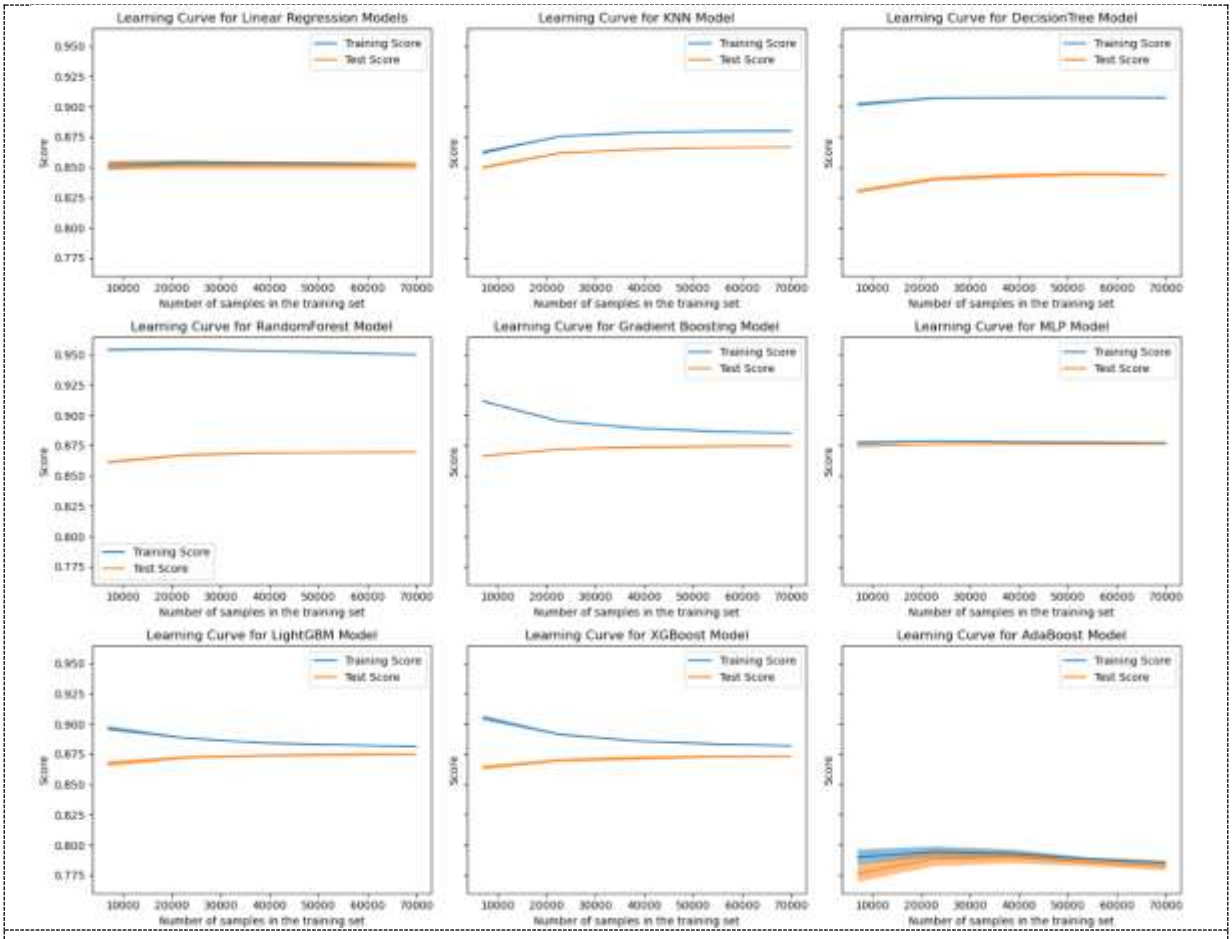

Figure 3. Learning curves for the nine machine learning models. Each curve shows the evolution of training and cross-validated performance (R^2) as the training sample size increases. Ensemble models (Gradient Boosting, LightGBM, XGBoost, Random Forest) display stable generalization with minimal train–test divergence, whereas simpler models such as Decision Tree exhibit larger gaps indicative of overfitting.

Scatter and residual plots further confirmed model fitting performance, with the Decision Tree model showing the lowest predictive scores (R^2 = 0.843, PCC = 0.918, MSE = 257.5), while ensemble methods (Random Forest, XGBoost) and MLP consistently achieved higher accuracy in LDL-C estimation. Feature importance analysis revealed Total Cholesterol (TC) as the most influential predictor across models, notably with importance scores of 0.8491 (Decision Tree) and 0.8698 (XGBoost), underscoring its critical role in LDL-C prediction within lipid profile-based machine learning frameworks.

4.4. Secondary internal validation

In the secondary internal validation, all machine learning models demonstrated strong predictive performance in estimating LDL-C, with R^2 values ranging from 0.843 (Decision Tree) to 0.876 (MLP), and Pearson correlation coefficients consistently above 0.91 across models. Notably, the ensemble methods,

including Random Forest (R^2 = 0.869, PCC = 0.932), Gradient Boosting (R^2 = 0.875, PCC = 0.935), LightGBM (R^2 = 0.875, PCC = 0.935), and XGBoost (R^2 = 0.872, PCC = 0.934), along with the MLP model (R^2 = 0.876, PCC = 0.936), achieved the highest accuracy with lower mean squared errors (MSE < 210), indicating robust fitting and consistency across the validation dataset.

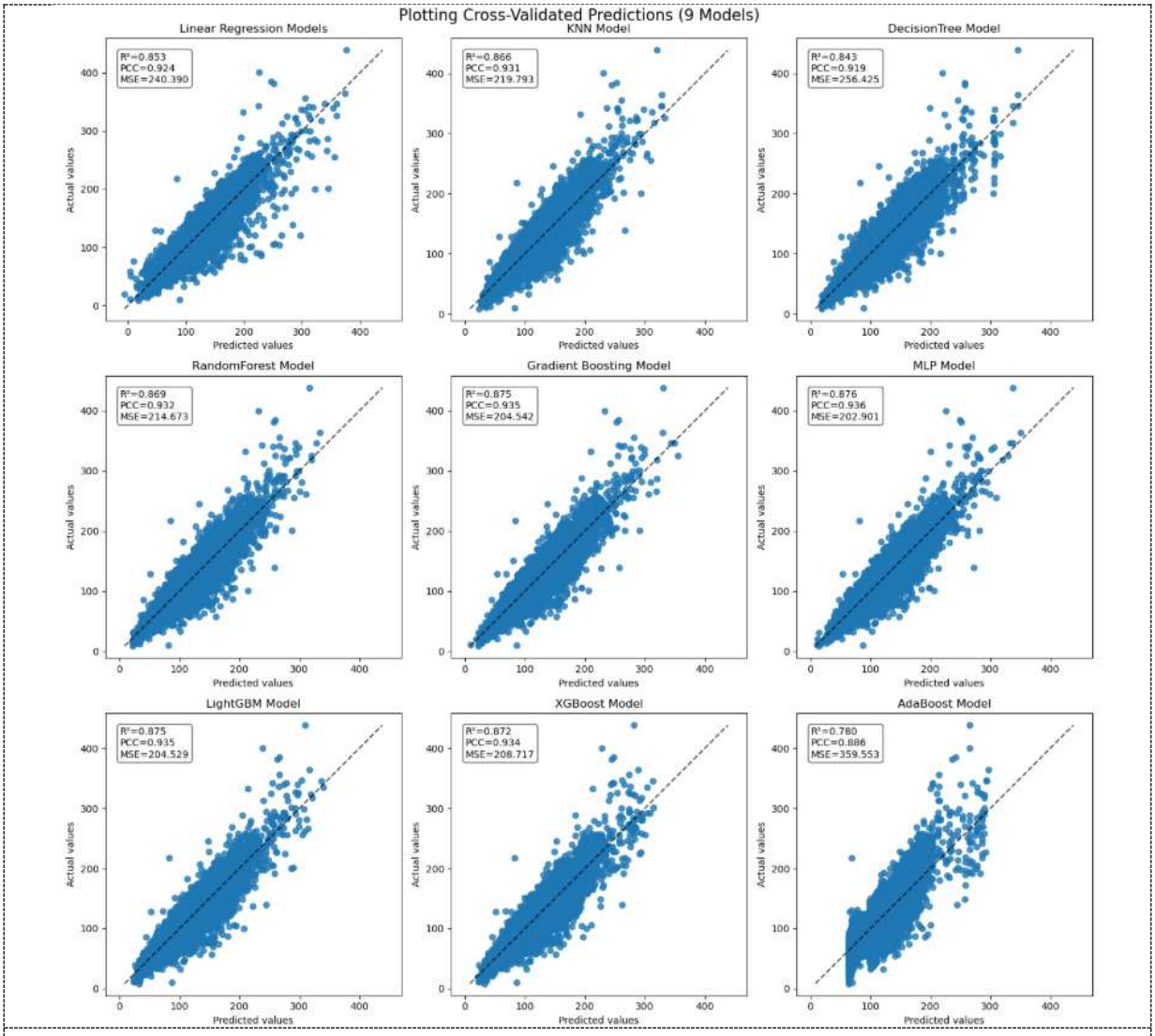

Figure 4. Actual versus predicted LDL-C values for nine machine learning models on the internal test set. LightGBM, XGBoost, Gradient Boosting, and MLP show the strongest alignment with the identity line, indicating high predictive agreement. Simpler models such as Linear Regression and Decision Tree show wider dispersion, particularly at higher LDL-C levels.

Compared to traditional formulas, these machine learning models exhibited superior predictive accuracy, particularly in higher LDL-C ranges and in samples with elevated triglyceride levels, where conventional formulas often fail. The results emphasize the capability of machine learning approaches to capture the complex, non-linear relationships within lipid profile data, providing reliable LDL-C estimates even in challenging contexts. This reinforces the potential of these models to enhance cardiovascular risk

assessment by addressing the limitations observed in formula-based estimations under conditions of hypertriglyceridemia.

5. Discussion

Accurate LDL-C measurement is essential for cardiovascular risk management, yet the high costs of direct measurement limit its widespread use, leading many laboratories to rely on estimation formulas such as Friedewald, Martin, and Sampson, which often show reduced accuracy in patients with elevated triglycerides. This study evaluated machine learning algorithms with systematic hyperparameter tuning to predict LDL-C reliably using routine lipid panel data while accounting for patient variability in gender, age, triglycerides, total cholesterol, and HDL-C. In Northeast China, where economic constraints and a high prevalence of dyslipidemia (62.1%) and metabolic syndrome (32.9%) present additional challenges, the implementation of accurate, cost-effective LDL-C prediction models is particularly critical. Our findings highlight that machine learning models can address the limitations of traditional formulas, offering robust and precise LDL-C estimation even in complex cases, thereby supporting improved cardiovascular risk stratification and treatment planning in clinical practice.

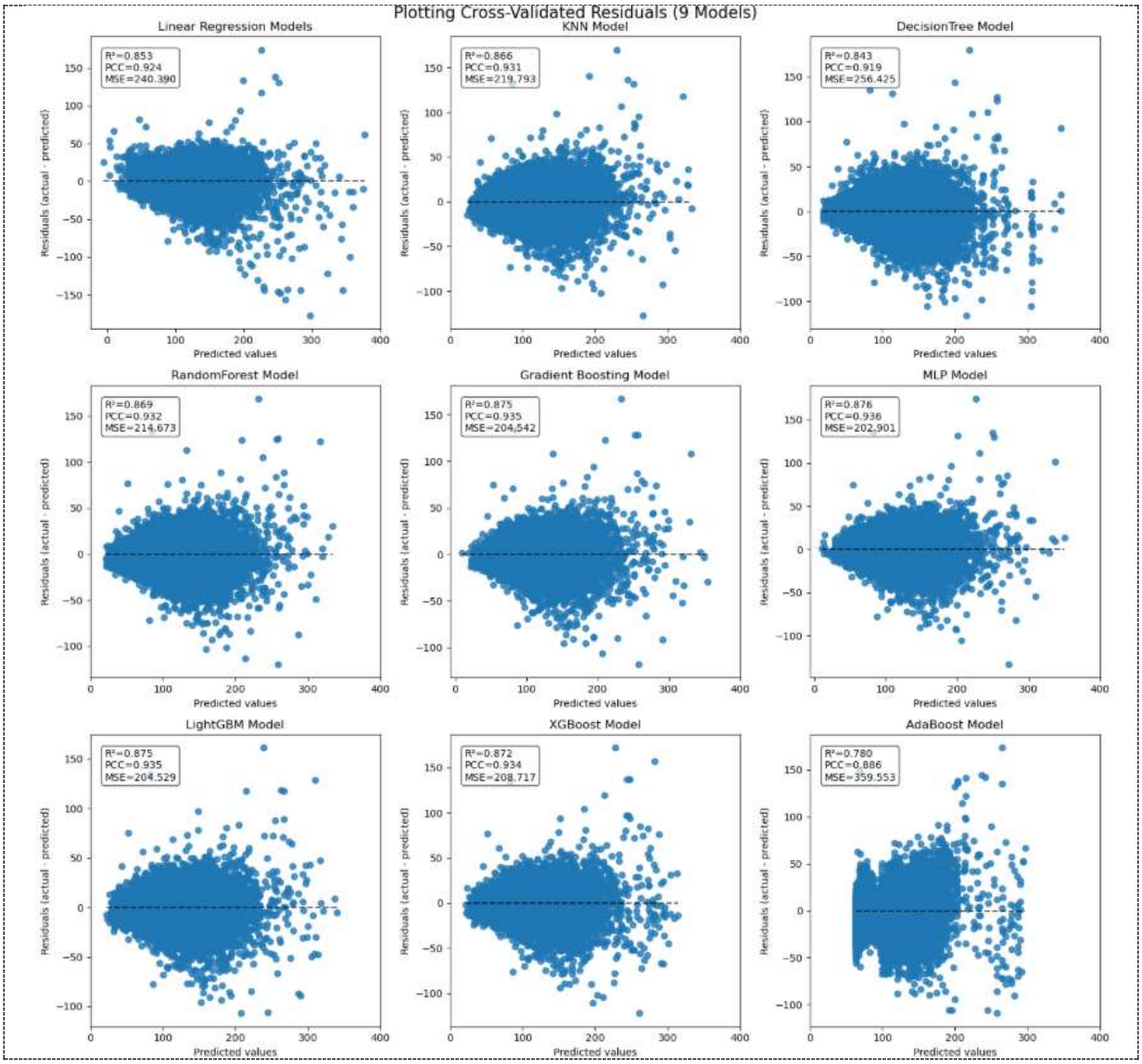

Figure 5. Residuals versus predicted LDL-C values for each of the nine machine learning models. Ensemble models demonstrate homoscedastic, randomly distributed residuals, while Decision Tree and AdaBoost show greater variance and patterns at the distribution tails, suggesting reduced stability.

On the internal training set, the linear, KNN, XGBoost, LightGBM, AdaBoost, and MLP regression models exhibited well-fitted learning curves, indicating stable model learning without significant underfitting. Although slight overfitting was noted in the Decision Tree and Random Forest models, it did not critically impact their predictive stability. Ensemble models, including Random Forest, Gradient Boosting, LightGBM, and XGBoost, along with the MLP neural network, consistently achieved higher R^2 values (up to 0.876) and lower MSE values, demonstrating superior predictive performance for LDL-C estimation across various sample sizes. These findings indicate that machine learning models can capture the complex, non-linear relationships within lipid profiles, providing reliable LDL-C predictions even under challenging conditions.

Additionally, our study confirmed that all machine learning models outperformed traditional formulas (Friedewald, Martin, Sampson) across different triglyceride ranges, particularly in high-TG scenarios (>300 mg/dL), where conventional formulas exhibited a marked decline in predictive accuracy. These results emphasize the robustness and flexibility of machine learning approaches in LDL-C estimation and align with previous research demonstrating the superior performance of ensemble and neural network models over closed-form equations for lipid parameter prediction. Collectively, these findings support the potential of machine learning integration into clinical laboratory workflows to enhance cardiovascular risk assessment and personalized treatment planning in diverse patient populations.

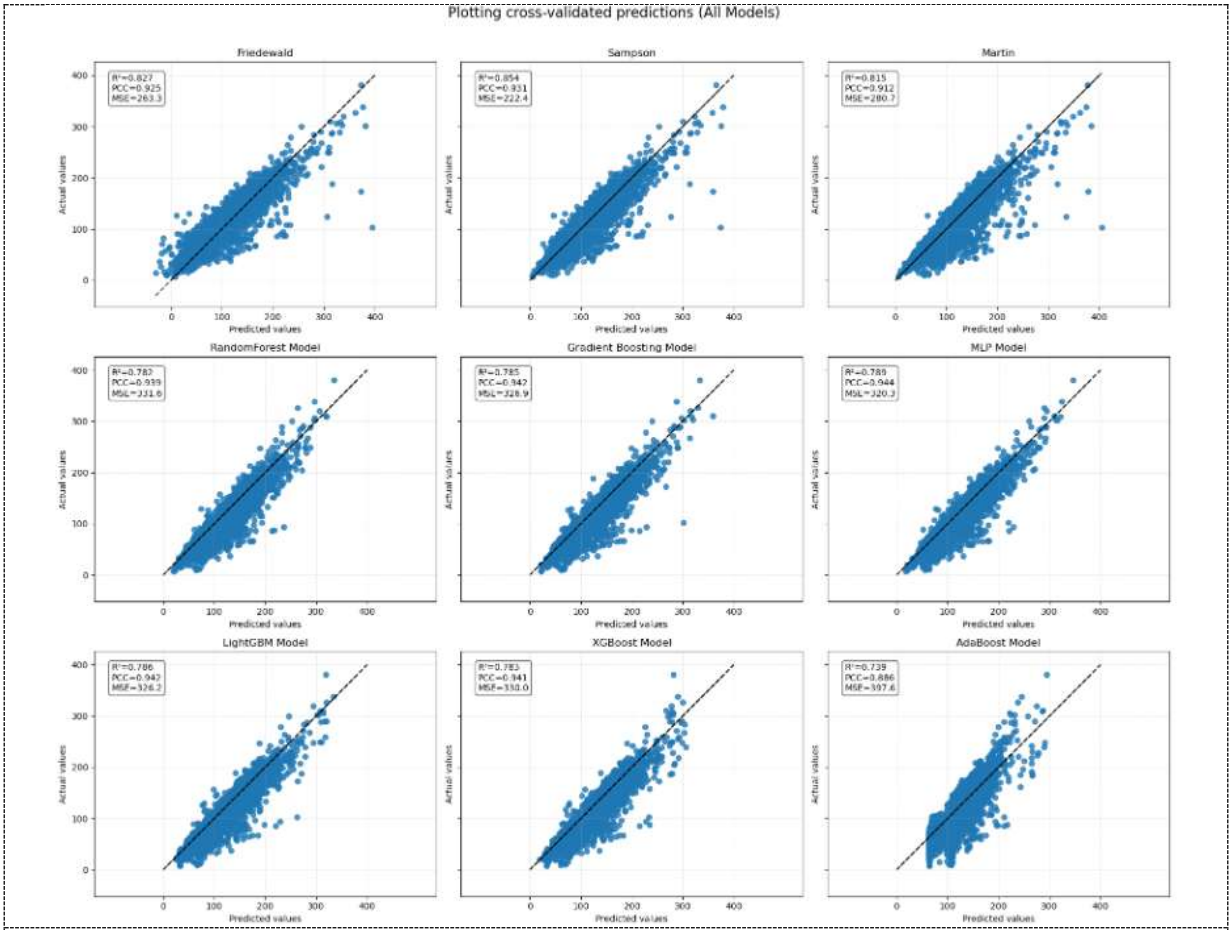

Figure 6. Comparison of actual versus predicted LDL-C for classical estimation formulas (Friedewald, Sampson, Martin) and machine learning models on the external validation dataset. Classical formulas exhibit increasing deviation from the identity line as triglyceride levels rise, whereas machine learning models maintain more stable accuracy across the full LDL-C range.

A notable limitation of the Friedewald and Martin formulas is their substantial decline in predictive accuracy under conditions of hypertriglyceridemia. In this study, the Friedewald formula demonstrated a consistent decrease in R^2 values across increasing TG categories, dropping from 0.93 in the <100 mg/dL group to -0.14 when TG exceeded 400 mg/dL, indicating that the formula's predictions were worse than a simple mean-based estimate in extreme TG conditions. The Martin formula exhibited a similar pattern, with its R^2 declining from 0.93 in the lowest TG category to -0.77 in the ≥400 mg/dL group, reflecting a tendency toward overestimation in high TG contexts while also failing to maintain predictive stability. These limitations align with previous reports highlighting the vulnerability of fixed-ratio estimation methods in capturing the complex, non-linear interactions between TG and LDL-C, particularly when chylomicrons and VLDL remnants dominate the lipid profile.

Tabel 4. Values of R2 and MSE for different models at various TG intervals.

Model	TG category (mg/dL)	R2	MSE	PCC
Friedewald	>0 and <100	0.93	96.83	0.97

	≥100 and <150	0.91	135.65	0.96
	≥150 and <200	0.85	213.94	0.94
	≥200 and <300	0.73	380.76	0.90
	≥300 and <400	0.52	711.97	0.82
	≥400	-0.14	1498.04	0.68
Sampson	>0 and <100	0.93	101.82	0.97
	≥100 and <150	0.91	129.9	0.96
	≥150 and <200	0.88	178.68	0.94
	≥200 and <300	0.79	305.62	0.90
	≥300 and <400	0.61	578.33	0.82
	≥400	0.15	1118.14	0.67
Martin	>0 and <100	0.93	94.32	0.97
	≥100 and <150	0.91	124.0	0.96
	≥150 and <200	0.88	172.72	0.94
	≥200 and <300	0.77	331.29	0.90
	≥300 and <400	0.46	805.7	0.82
	≥400	-0.77	2320.61	0.64
RandomForest	>0 and <100	0.87	189.70	0.97
	≥100 and <150	0.82	257.95	0.96
	≥150 and <200	0.79	308.12	0.94
	≥200 and <300	0.68	462.23	0.91
	≥300 and <400	0.51	732.57	0.86
	≥400	0.33	876.58	0.78
Gradient Boosting	>0 and <100	0.87	184.86	0.97
	≥100 and <150	0.83	250.63	0.96

	≥150 and <200	0.79	296.29	0.95
	≥200 and <300	0.68	460.77	0.91
	≥300 and <400	0.48	765.17	0.85
	≥400	0.34	859.43	0.79
MLP	>0 and <100	0.86	198.18	0.97
	≥100 and <150	0.83	252.41	0.96
	≥150 and <200	0.79	297.27	0.95
	≥200 and <300	0.69	441.09	0.91
	≥300 and <400	0.56	656.28	0.87
	≥400	**0.38**	**814.22**	**0.79**
LightGBM	>0 and <100	0.87	186.73	0.97
	≥100 and <150	0.83	251.26	0.96
	≥150 and <200	0.79	297.92	0.95
	≥200 and <300	0.68	463.46	0.91
	≥300 and <400	0.50	740.80	0.86
	≥400	**0.36**	**839.43**	**0.79**
XGBoost	>0 and <100	0.87	184.81	0.97
	≥100 and <150	0.83	252.49	0.96
	≥150 and <200	0.79	307.38	0.94
	≥200 and <300	0.67	466.89	0.91
	≥300 and <400	0.49	752.01	0.85
	≥400	0.34	865.33	0.78
AdaBoost	>0 and <100	0.81	264.67	0.91
	≥100 and <150	0.84	229.55	0.92
	≥150 and <200	0.81	273.70	0.92

≥200 and <300	0.64	517.09	0.88
≥300 and <400	0.31	1023.35	0.83
≥400	-0.38	1809.45	0.62

In contrast, the Sampson formula displayed improved robustness, maintaining positive R^2 values across all TG categories, including an R^2 of 0.15 in the ≥400 mg/dL group, indicating better handling of elevated TG scenarios than the Friedewald and Martin formulas. However, despite this improvement, the Sampson formula still underperformed relative to machine learning (ML) methods. Ensemble learning models such as Random Forest, Gradient Boosting, LightGBM, and XGBoost, along with the Multilayer Perceptron (MLP) neural network, consistently outperformed traditional formulas across all TG strata. Notably, in the challenging TG ≥400 mg/dL subgroup, LightGBM achieved an R^2 of 0.36 with a PCC of 0.79, while XGBoost and Gradient Boosting maintained R^2 values above 0.34, demonstrating their capacity to capture non-linear dependencies and interactions within the lipid profile that traditional equations fail to account for. Across lower TG strata, ML models sustained high predictive performance (R^2 >0.87, PCC >0.93), demonstrating stable generalization while avoiding the systematic biases observed in formula-based estimates.

These findings emphasize the superior adaptability and accuracy of ML models in LDL-C estimation across diverse TG conditions, highlighting their potential for integration into clinical laboratory workflows. By providing consistent and accurate LDL-C predictions, especially under high TG conditions where precise LDL-C estimation is critical for cardiovascular risk stratification, ML models offer a reliable alternative to conventional formulas. This is particularly relevant for laboratories managing populations with high hypertriglyceridemia prevalence, where reliance on traditional formulas may contribute to misclassification of cardiovascular risk and inappropriate therapeutic decisions.

LightGBM achieved slightly superior performance compared to the other machine learning models, including XGBoost and Random Forest. This improvement can be attributed to its leaf-wise tree growth strategy, which allows the model to prioritize splits that produce the greatest reduction in loss at each iteration. In contrast, XGBoost grows trees in a level-wise manner, which is computationally stable but often less expressive for capturing complex local interactions among lipid variables. LightGBM's histogram-based binning further improves efficiency and reduces memory usage, enabling it to model non-linear relationships between TC, TG, HDL-C, and LDL-C more effectively.

Meanwhile, Random Forest and Gradient Boosting models also performed well but were limited by their reliance on fully greedy tree construction without the fine-grained optimization present in LightGBM and XGBoost. The MLP model achieved competitive accuracy but required substantially longer training time and exhibited higher sensitivity to hyperparameter initialization. Overall, the combination of fast histogram-based learning, leaf-wise splitting, and aggressive loss minimization explains why LightGBM consistently achieved high R^2 and PCC values across triglyceride strata.

Another factor contributing to the performance advantage of LightGBM is its ability to handle feature interactions that are not strictly additive. In lipid metabolism, triglyceride-rich lipoproteins often influence the Friedewald estimation error in a non-linear fashion, especially when TG exceeds 200 mg/dL. LightGBM's leaf-wise strategy effectively captures these localized patterns, resulting in lower prediction error compared with XGBoost, which tends to generalize more smoothly due to its regularized level-wise tree expansion.

From a clinical perspective, reducing prediction error around therapeutic decision thresholds is highly important. Common LDL-C cutoffs such as 70 mg/dL (very-high-risk patients), 100 mg/dL (high-risk), and 130 mg/dL (moderate-risk) guide the initiation or intensification of lipid-lowering therapy. Even small prediction errors around these boundaries can lead to under-treatment (if LDL-C is underestimated)

or unnecessary therapy escalation (if overestimated). The superior performance of LightGBM—particularly its lower MSE and higher PCC in patients with elevated triglycerides—reduces the risk of such misclassification. This is clinically relevant because traditional formula-based estimates such as Friedewald, Sampson, or Martin show increasing error at higher triglyceride levels, whereas the machine learning models maintain more stable accuracy in these challenging ranges.

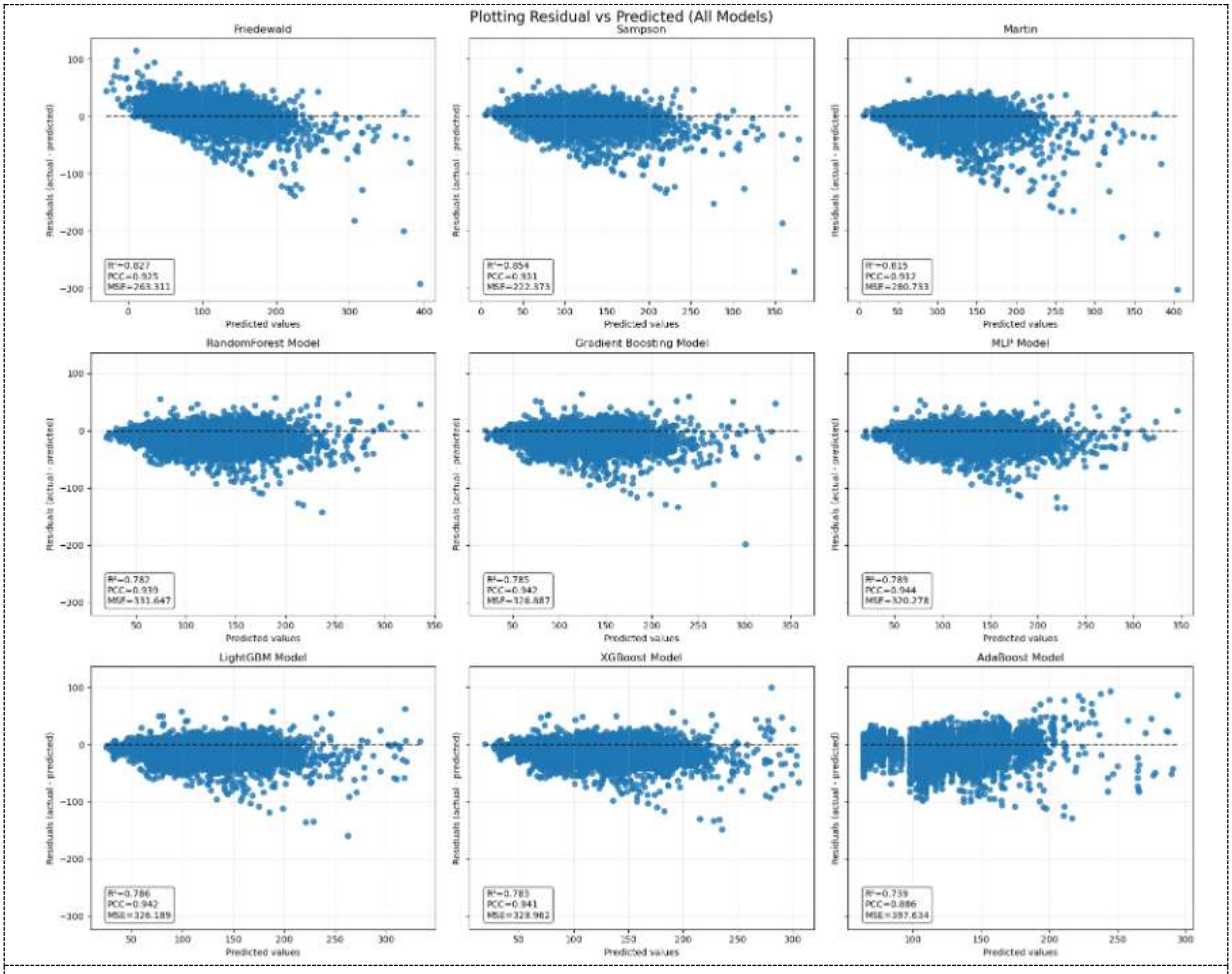

Figure 7. Residual plots for classical LDL-C formulas and machine learning models on the external validation dataset. Residuals from Friedewald, Sampson, and Martin formulas show increasing positive bias at higher LDL-C values, while ML models—especially LightGBM and Gradient Boosting—produce narrower and more symmetric residual distributions.

Collectively, the results from this study underscore the value of implementing machine learning methods in routine LDL-C estimation workflows to enhance personalized patient care in dyslipidemia management.

Our findings demonstrate that traditional LDL-C estimation formulas, including the Friedewald and Martin equations, exhibit a significant decline in predictive accuracy as triglyceride (TG) levels increase, with R^2 values dropping below zero in the TG ≥400 mg/dL interval. This decline is particularly concerning given the elevated cardiovascular risk in patients with high TG, underscoring the need for more reliable estimation methods in this subset. In contrast, machine learning models, including Random Forest, Gradient Boosting, LightGBM, XGBoost, and MLP, maintained high predictive accuracy across all TG levels, with

R^2 values remaining above 0.67 even in the TG ≥400 mg/dL range. Notably, LightGBM and XGBoost demonstrated robust performance in this challenging subgroup, achieving R^2 values of 0.36 and 0.34, respectively, with high correlation coefficients, highlighting their potential as reliable tools for LDL-C estimation in clinical practice where traditional formulas underperform.

Previous studies have shown that ML models such as XGBoost and Random Forest can outperform traditional formulas in LDL-C estimation, although concerns regarding hyperparameter tuning time and computational costs have been noted. In our study, we confirmed that ML models, including Random Forest, Gradient Boosting, LightGBM, XGBoost, and MLP, maintained superior predictive performance across all TG strata, particularly in the challenging TG ≥400 mg/dL interval where traditional formulas like Friedewald and Martin exhibited a sharp decline, with R^2 values dropping below zero, reflecting prediction instability. In contrast, the ML models consistently achieved R^2 values above 0.30 with lower MSE and higher Pearson correlation coefficients, demonstrating resilience in capturing nonlinear TG-LDL-C relationships under hypertriglyceridemic conditions. Although the MLP model demonstrated stable predictive accuracy, it required careful parameter tuning and showed similar but not superior performance compared to ensemble methods in our dataset.

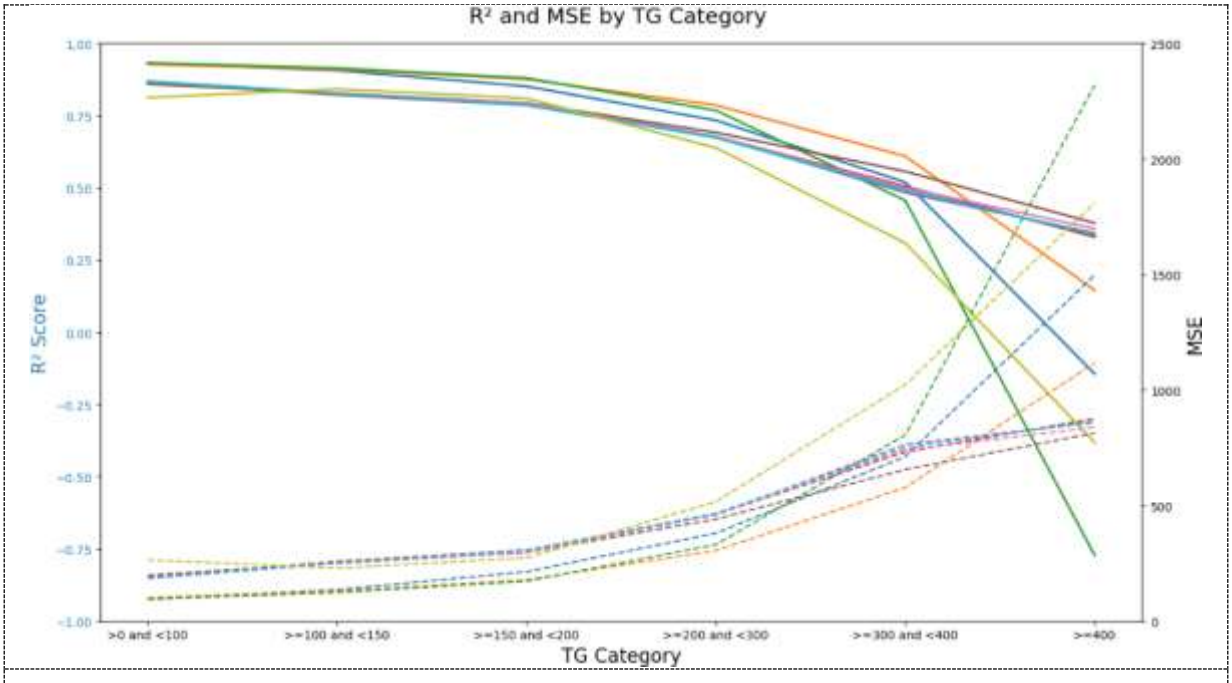

Figure 8. Model performance across triglyceride (TG) categories on the external validation dataset. Classical formulas show rapid degradation in R^2 and substantial increases in MSE beginning at TG ≥ 200 mg/dL. In contrast, machine learning models demonstrate more gradual performance decline, with LightGBM, Gradient Boosting, and MLP maintaining comparatively strong accuracy even at TG ≥ 300 mg/dL. Pearson correlation remains high for ML models across categories, indicating preserved monotonicity.

To address computational efficiency and improve interpretability, we applied the Mean Decrease in Impurity (MDI) method for feature selection prior to model fitting, which identified Total Cholesterol (TC) as the most influential predictor while reducing noise from less informative features such as gender. This approach facilitated faster convergence during hyperparameter tuning and reduced computational burden, supporting the practical integration of ML models for LDL-C estimation in clinical laboratories where accurate lipid profiling is critical for cardiovascular risk management

Our open-source machine learning models, including LightGBM, XGBoost, Random Forest, and MLP, provide a flexible and scalable solution for LDL-C estimation across diverse healthcare settings. These models can be easily retrained using local patient data to enhance predictive accuracy within specific populations, ensuring optimal performance while accounting for regional variations in lipid profiles. Additionally, the models' compatibility with integration into electronic health record systems supports seamless clinical implementation, facilitating real-time LDL-C estimation to inform cardiovascular risk assessment and treatment decisions. This adaptability and ease of integration position our ML models as practical tools for laboratories and healthcare providers aiming to improve precision in lipid management, particularly in settings where direct LDL-C measurement is limited.

Our findings indicate that LightGBM is not only computationally efficient but also clinically meaningful, as its improved predictions may better support risk stratification and lipid-lowering therapy decisions. Although the performance differences among advanced ML models were modest, LightGBM consistently demonstrated robustness across triglyceride categories, suggesting its suitability for real-world clinical deployment.

6. Conclusions

Accurate LDL-C determination, particularly in high TG ranges, remains challenging, with traditional formulas like Friedewald and Martin showing a marked decline in predictive performance when TG levels exceed 300 mg/dL, as evidenced by negative or near-zero R^2 values and elevated MSE. In contrast, our results demonstrate that machine learning models, including Random Forest, Gradient Boosting, LightGBM, XGBoost, and MLP, maintain higher predictive accuracy across all TG categories, including the challenging ≥400 mg/dL range, with R^2 values remaining positive (up to 0.36 for LightGBM) and PCC values consistently above 0.77. This improved performance indicates the potential of ML models to enhance cardiovascular risk assessment by providing more precise LDL-C estimates, supporting better-informed treatment decisions, especially in hypertriglyceridemic patients. However, while these models outperform formula-based estimates, successful integration into routine clinical workflows requires addressing data availability, computational resources, and prospective validation in diverse populations to ensure robust, reliable application in real-world healthcare settings.

References

[1] Meng, Jing Bi, Zai Jian An, and Chun Shan Jiang. 2025. “Machine Learning-Based Prediction of LDL Cholesterol: Performance Evaluation and Validation.” PeerJ 13(4). doi: 10.7717/peerj.19248.

[2] Anudeep, P. P., Suchitra Kumari, Aishvarya S. Rajasimman, Saurav Nayak, and Pooja Priyadarsini. 2022. “Machine Learning Predictive Models of LDL-C in the Population of Eastern India and Its Comparison with Directly Measured and Calculated LDL-C.” Annals of Clinical Biochemistry 59(1):76–86. doi: 10.1177/00045632211046805.

[3] Çubukçu, Hikmet Can, and Deniz İlhan Topcu. 2022. “Estimation of Low-Density Lipoprotein Cholesterol Concentration Using Machine Learning.” Lab Medicine 53(2):161–71. doi: 10.1093/labmed/lmab065.

[4] Ghayad, Jean Pierre, Vanda Barakett-Hamadé, and Ghassan Sleilaty. 2022. “Prospective Validation of a Machine Learning Model for Low-Density Lipoprotein Cholesterol Estimation.” Lab Medicine 53(6):629–35. doi: 10.1093/labmed/lmac049.

[5] Kim, Yoori, Won Kyung Lee, and Woojoo Lee. 2024. “Prediction of Low-Density Lipoprotein Cholesterol Levels Using Machine Learning Methods.” Lab Medicine 55(4):471–84. doi: 10.1093/labmed/lmad114.

[6] Singh, Gurpreet, Yasin Hussain, Zhuoran Xu, Evan Sholle, Kelly Michalak, Kristina Dolan, Benjamin C. Lee, Alexander R. van Rosendael, Zahra Fatima, Jessica M. Peña, Peter W. F. Wilson, Antonio M. Gotto, Leslee J.

[7] Shaw, Lohendran Baskaran, and Subhi J. Al'Aref. 2020. "Comparing a Novel Machine Learning Method to the Friedewald Formula and Martin-Hopkins Equation for Low-Density Lipoprotein Estimation." PLoS ONE 15(9 September). doi: 10.1371/journal.pone.0239934.

[8] Srimani S, Parai M, GhoshK,RahamanH. 2021. A statistical approach of analog circuit fault detection utilizing kolmogorov smirnov test method. Circuits, Systems, and Signal Processing 40:2091 2113 DOI 10.1007/s00034-020-01572-x.

[9] Rifai N. 2006. Lipids, lipoproteins, apolipoproteins, and other cardiovascular risk factors. In: Tietz textbook of clinical chemistry. St. Louis, Missouri: Elsevier Saunders, 903 968.

[10] PedregosaF,VaroquauxG, GramfortA, MichelV, ThirionB, GriselO,BlondelM, Prettenhofer P, Weiss R, DubourgV. 2011. Scikit-learn: machine learning in Python. The Journal of Machine Learning Research 12:2825 2830.

[11] Chicco D,WarrensMJ,JurmanG. 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science 7:e623 DOI 10.7717/peerj-cs.623.

[12] FanG,ZhangS,WuQ,SongY,JiaA,LiD,YueY,WangQ. 2022. A machine learning based approach for low-density lipoprotein cholesterol calculation using age, and lipid parameters. Clinica Chimica Acta 535:53 60 DOI 10.1016/j.cca.2022.08.007.