Received: 13 April 2025 | Revised: 30 June 2025 | Accepted: 18 August 2025 | Published online: 1 September 2025

RESEARCH ARTICLE

Medinformatics
2025, Vol. 00(00) 1–7
DOI: 10.47852/bonviewMEDIN52025919

BON VIEW PUBLISHING

An Artificial Intelligence and Biochemical Modelling Approach for LDL Cholesterol Prediction

Petros Paplomatas[1,*], Maria Kantartzi[2], Kostas Anagnostopoulos[3] and Aristidis Vrahatis[1]

[1]*Department of Informatics, Ionian University, Greece*

[2]*Independent Researcher, Germany*

[3]*Department of Medicine, Democritus University of Thrace, Greece*

Abstract: Estimating LDL cholesterol accurately remains a critical challenge in cardiovascular medicine. While ultracentrifugation delivers precise results, its high cost and complexity make it unsuitable for everyday clinical practice. Current estimation formulas like Friedewald and Martin-Hopkins are convenient but frequently miss the mark due to patient-specific factors and biological variations. Direct LDL testing, though reliable, puts a significant financial strain on healthcare systems. We developed a sophisticated machine learning approach that combines the strengths of 13 existing estimation equations with essential patient data—age, sex, total cholesterol, triglycerides, and high-density lipoprotein levels—alongside actual LDL measurements. The framework uses three different algorithms (k-Nearest Neighbors, Random Forest, and Support Vector Machine) as foundational learners, while XGBoost acts as the final decision-maker to identify intricate data relationships. What sets our method apart is its ability to classify LDL levels according to both NCEP III and European Society of Cardiology standards, offering clinicians a more complete risk assessment tool. When we tested our integrated model against simpler versions, the results were striking: our comprehensive approach achieved an R^2 value exceeding 0.98, significantly outperforming models that relied solely on basic clinical parameters (R^2 around 0.95). This advancement could transform how we assess cardiovascular risk, especially in settings where expensive laboratory methods aren't feasible.

Keywords: LDL cholesterol, cardiovascular risk, machine learning, clinical prediction, lipid profile estimation, NCEP III, European Society of Cardiology Guidelines

1. Introduction

Low-density lipoprotein (LDL) cholesterol is a pivotal biomarker in cardiovascular risk assessment, as its elevated levels are strongly correlated with the development of atherosclerosis [1], coronary artery disease [2, 3], and increased mortality rates [4, 5]. The primary method for precise LDL measurement, ultracentrifugation, is both expensive and labor-intensive, making it an impractical choice for routine clinical use [6]. As a result, clinicians often resort to estimation equations like the Friedewald [7] and Martin-Hopkins formulas [8]. However, these equations may yield inaccuracies due to the variability in lipid profiles influenced by triglycerides, age, metabolic conditions, and other demographic factors [9–11]. Direct LDL assays, while accurate, are cost-prohibitive due to high reagent costs, further limiting their use [12, 13].

The advent of machine learning has opened new avenues for improving diagnostic accuracy in medicine, particularly in lipid profile estimation [14]. This paper presents an innovative stacking machine learning framework that combines traditional LDL equations with clinical variables to maximize prediction accuracy. Our approach integrates 13 established LDL estimation equations as input features, combined with dual classification labels based on National Cholesterol Education Program Adult Treatment Panel III (NCEP III) and European Society of Cardiology (ESC) guidelines [15]. This design leverages existing biochemical knowledge while enhancing predictive accuracy through machine learning. Our results demonstrate the effectiveness of this approach, achieving state-of-the-art accuracy and showcasing the clinical applicability of the model.

Building on our previous work [13], we employed the R package LDLcalc for computing these equations, which has now been integrated into our new R package named AutoLDLStack. This package automatically handles the computation of equations and applies labels according to both NCEP III and ESC guidelines for dual-label stratification. **Our approach is novel in incorporating both traditional equation outputs and clinical features as model inputs, while providing dual classification according to NCEP III and ESC guidelines [16, 17] for comprehensive cardiovascular risk assessment.

By constructing a stacking ensemble learning model, we integrate 13 traditional LDL estimation equations and classification labels alongside clinical parameters such as sex, age, total cholesterol,

*Corresponding author: Petros Paplomatas, Department of Informatics, Ionian University, Greece. Email: p.paplomatas@ionio.gr

triglycerides, high-density lipoprotein (HDL), and directly measured LDL values. Our base models include K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM), with XGBoost as the meta-learner to optimize prediction accuracy [18]. This advanced integration overcomes the limitations of individual models, providing a robust and practical alternative for LDL estimation in clinical settings.

This research not only seeks to refine the accuracy of LDL cholesterol estimation but also aims to democratize access to precise cardiovascular risk assessment, particularly in resource-limited healthcare systems. By reducing dependency on costly direct measurement techniques, our model could revolutionize the management of cardiovascular risks on a global scale.

2. Materials and Methods

2.1. Data preprocessing

A dataset with 4,244 records was sourced from the biochemical laboratory at the "Sismanoglio" General Hospital in Komotini, Greece, for our study. The dataset included patients aged 1–103 years, with 1,794 females (42.3%) and 2,450 males (57.7%). Inclusion criteria comprised patients with complete lipid panel measurements (total cholesterol, triglycerides, HDL, and directly measured LDL), while exclusion criteria included incomplete laboratory data or missing demographic information. Our analysis employed the following parameters: SEX (Sex), AGE (Age), CHOL (Total Cholesterol), TG (Triglycerides), HDL, and LDLd (Directly Measured LDL).

2.2. Model selection and implementation

In this study, we employed three distinct base learning algorithms—KNN, RF, and SVM with a Radial Basis Function kernel (SVM-Radial)—selected for their complementary strengths in handling different aspects of the data:

1) K-Nearest Neighbors (KNN) [19] was chosen for its simplicity and effectiveness in capturing local structures within the data, particularly useful for modeling non-linear relationships. The number of neighbors (k) was tuned within the range of 6 to 26 (incremented by 2). Since KNN relies on distance metrics, preprocessing steps included centering and scaling to normalize feature impact.
2) Random Forest (RF) [20, 21] was incorporated due to its robustness against overfitting and its ability to handle high-dimensional data while providing feature importance insights. The number of variables randomly sampled as candidates at each split (mtry) was optimized over values of 11, 15, and 22. Preprocessing was not required, as RF is generally insensitive to variations in feature scale.
3) Support Vector Machines (SVMs) [22, 23] with a Radial Basis Function kernel (SVM-Radial) was selected for its strong performance in high-dimensional spaces, making it particularly suited for capturing complex patterns in the data. The tuning process involved optimizing the radial basis function parameter (σ) within the range [0.01, 0.02, 0.05] and the cost parameter (C) across values [6, 9, 10]. Given SVM's reliance on distance-based calculations, features were scaled to ensure uniform contribution to the model.

To integrate traditional LDL cholesterol estimation, we automatically computed 13 LDL equations using the R package LDLcalc, which derives LDL cholesterol levels from total cholesterol, HDL cholesterol, and triglycerides. These computed values were incorporated into the dataset as additional input features. Additionally, LDL cholesterol levels were classified according to both the NCEP III and the ESC guidelines, generating categorical labels that provide a more comprehensive risk assessment perspective. A custom function was implemented to systematically assign these classification labels within the dataset. All analyses were conducted using R version 4.3.0 with the following packages: caret (6.0–94), caretEnsemble (2.0.3), xgboost (1.7.5.1), and LDLcalc (2.0.0). Fixed random seeds (123 for data splitting, 1,987 for base learners, and 123 for meta-learner) were used throughout to ensure reproducibility.

2.3. Stacking model framework

To enhance predictive performance, a stacking ensemble approach was implemented. The dataset was partitioned into 80% training and 20% testing subsets, ensuring reproducibility through the use of a fixed random seed. Each base model—KNN, RF, and SVM-Radial—was trained using 10-fold cross-validation to obtain robust performance estimates. Model hyperparameters were fine-tuned using specific grid search strategies for each algorithm to optimize individual performance.

Extreme Gradient Boosting (XGBoost) [18] was employed as the meta-learner due to its ability to handle both regression and classification tasks with high predictive accuracy. XGBoost hyperparameters, including nrounds [50, 100, 150], max_depth [3, 5], eta [0.1, 0.3], gamma [0, 1], colsample_bytree [0.6, 0.8, 1.0], min_child_weight [1, 3], and subsample [0.75, 1.0], were systematically optimized to maximize model performance and generalization.

The final stacked model was validated on the test dataset, assessing predictive performance using multiple statistical metrics: root mean squared error (RMSE), R-squared (R2), mean absolute error (MAE), mean squared error (MSE), mean absolute percentage error (MAPE), and median absolute error (MedAE). These metrics provided a comprehensive evaluation of model accuracy and error distribution. Additionally, variable importance analysis was conducted to interpret the influence of individual features within the stacked model.

This methodological framework has been encapsulated in the AutoLDLStack package, enabling users to seamlessly train their own models and perform LDL cholesterol predictions. By simply providing an input dataset containing the parameters SEX, AGE, CHOL, TG, HDL, and LDLd, the package automates the entire process, from model training to inference, making it accessible for clinical and research applications.

3. Results

To evaluate the contribution of traditional LDL equations and dual classification labels, we compared two model configurations: (1) Clinical-Features-Only Model—trained exclusively on demographic and basic lipid parameters (SEX, AGE, CHOL, TG, HDL, LDLd), and (2) Enhanced Stacked Model—incorporating clinical features plus 13 traditional LDL equation outputs and dual classification labels based on NCEP III and ESC guidelines.

3.1. Baseline vs. full model comparison

Initially, we evaluated a baseline model, trained solely on clinical features, against the full model, which leverages additional predictive variables, including traditional LDL equations and dual classification labels. Table 1 presents a summary of key evaluation metrics.

Table 1. Clinical-features-only vs. Enhanced Stacked Model comparison

Metric	Stack_Model	Baseline_Model
RMSE(1)	5.630	8.359
R^2	0.981	0.959
MAE(2)	4.247	6.117
MSE(3)	31.707	69.877
MAPE(4)	4.489	6.328

Note: (1) Root mean squared error. (2) Mean absolute error. (3) Mean squared error. (4) Mean absolute percentage error.

The metrics reveal a substantial improvement with the Enhanced Stacked Model. The RMSE dropped from 8.36 to 5.63, and the coefficient of determination (R^2) rose from 0.96 to 0.98, underscoring the enhanced predictive capability. Moreover, the MAE was reduced by approximately 30%, confirming the advantage of incorporating traditional equations and classification features into the model.

3.2. Model performance analysis and risk classification

To comprehensively evaluate our Enhanced Stacked Model, we examined both the individual model contributions and the dual classification system that enhances predictive capability. To understand the contributions of each base model, we compared their performance metrics as shown in Table 2.

Interestingly, the SVM Radial model achieves the lowest RMSE and highest R^2, indicating that it is the best-performing base model in terms of error minimization. However, the Enhanced Stacked Model achieves the lowest MAE, suggesting more consistent predictions with fewer extreme deviations. While the differences in RMSE and R^2 between the stacked model and SVM Radial are small, the stacking approach ensures robustness and reliability across various test scenarios, providing stable predictions critical for real-world clinical applications.

The Enhanced Stacked Model incorporates dual classification features based on established cardiovascular risk guidelines. Table 3 presents the distribution of patients across both NCEP III and ESC categories.

The dual classification system provides complementary cardiovascular risk assessment perspectives, with ESC guidelines offering more granular stratification, particularly in lower LDL ranges. These categorical variables, combined with the 13 traditional LDL equations, contribute significantly to the Enhanced Stacked Model's superior performance demonstrated in Table 1.

3.3. Visualization of baseline models and stacking performance

To visually bridge the numerical results with their implications, we introduce Figure 1, which elucidates the performance characteristics of our stacking algorithm. In the learning curve analysis, the left plot presents the learning curve, plotting the RMSE against the increasing percentage of the training set size, using 10-fold cross-validation with confidence intervals. The convergence of the training RMSE, depicted by the blue line, and the test RMSE, shown by the red line, indicates effective generalization. The stacked model's lower error rates and the tighter confidence intervals, represented by the shaded regions at each data point, underscore its enhanced stability and efficiency in learning from data. At the point of performance stabilization, the final test RMSE is noted as 6.086, providing visual evidence that the model's learning capacity has been optimized without overfitting.

Turning to the actual versus predicted LDL values, the right plot displays a scatter of these values for the stacked model. Each blue point represents a test sample, with the red dashed line ($y = x$) serving as the ideal prediction line. The close alignment of points along this line illustrates the stacking algorithm's superior accuracy, corroborating the high R^2 and low RMSE values from our numerical analysis. This visualization not only confirms the predictive prowess of the stacking approach but also visually demonstrates the absence of systematic bias, enhancing the transition from quantitative data to a visual representation of performance enhancement.

Table 3. Dual LDL risk classification distribution in study population ($n = 4,244$)

CLASSIFICATION	*LDL RANGE (MG/DL)*	*N (%)*
NCEP III GUIDELINES		
OPTIMAL	<100	2005 (47.2%)
NEAR OPTIMAL–ABOVE OPTIMAL	100–129	1032 (24.3%)
BORDERLINE HIGH	130–159	727 (17.1%)
HIGH	160–189	336 (7.9%)
VERY HIGH	≥190	144 (3.4%)
ESC GUIDELINES		
LDL CAT 1	<55	375 (8.8%)
LDL CAT 2	55–69	398 (9.4%)
LDL CAT 3	70–99	1232 (29.0%)
LDL CAT 4	100–115	578 (13.6%)
LDL CAT 5	116–189	1517 (35.7%)
LDL CAT 6	≥190	144 (3.4%)

3.4. Analysis of model heterogeneity and justification for stacking

To evaluate model stability and justify the stacking approach, we examined performance variability across cross-validation folds. Beyond the high performance of the stacking algorithm as evidenced by the metrics in Table 1 and the visualizations in Figure 1, Figure 2 presents scatter plot matrices for RMSE and MAE across the base learners (KNN, RF, SVM Radial) and the Enhanced Stacked Model over different cross-validation instances.

Table 2. Performance comparison of machine learning models

MODEL	MAE(1)	MSE(2)	RMSE(3)	R^2	MAPE(4)	MEDAE(5)
KNN	4.465	34.209	5.848	0.9804	4.765	3.625
RANDOM FOREST	4.337	34.056	5.835	0.9803	4.563	3.432
SVM RADIAL	4.260	31.651	5.625	0.9817	4.448	3.357
STACKED	4.247	31.707	5.630	0.9817	4.489	3.366

Note: (1) Mean absolute error. (2) Mean squared error. (3) Root mean squared error. (4) Mean absolute percentage error. (5) Median absolute error.

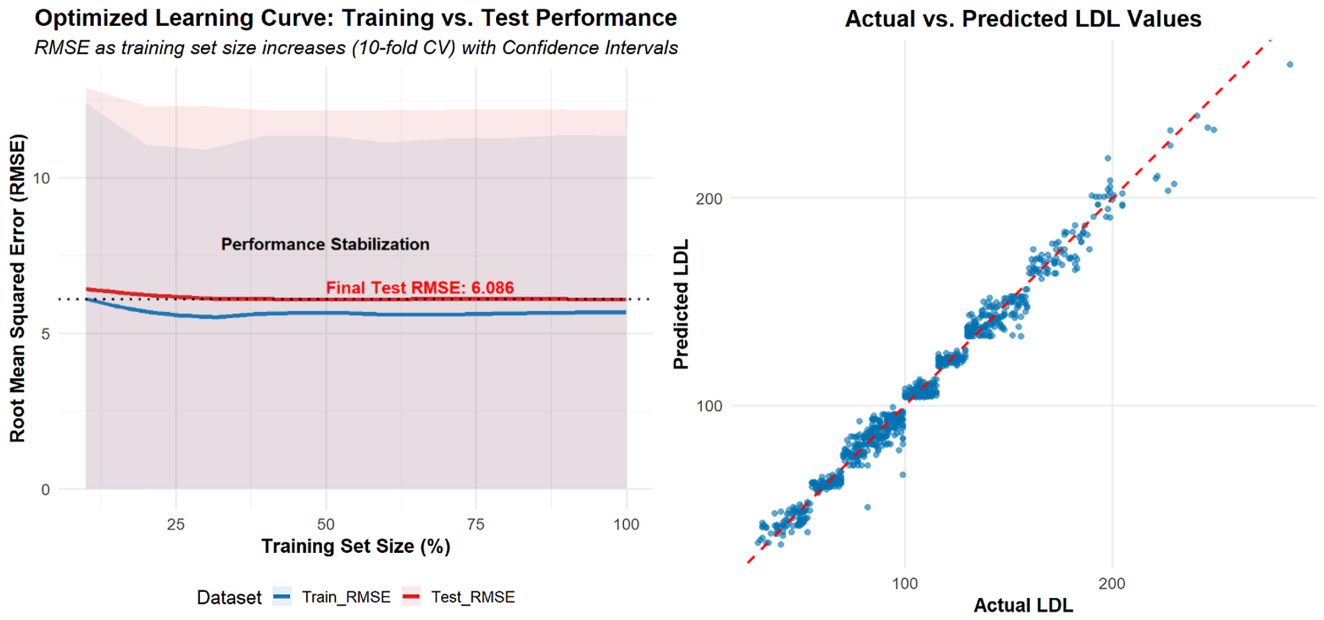

Figure 1. Enhanced Stacked Model performance analysis. (Left) Learning curve showing training and test RMSE convergence across increasing training set sizes with 10-fold cross-validation confidence intervals. (Right) Scatter plot of actual vs. predicted LDL values on test set ($n = 849$), with red dashed line representing perfect prediction ($y = x$). Blue points represent individual predictions from the Enhanced Stacked Model.

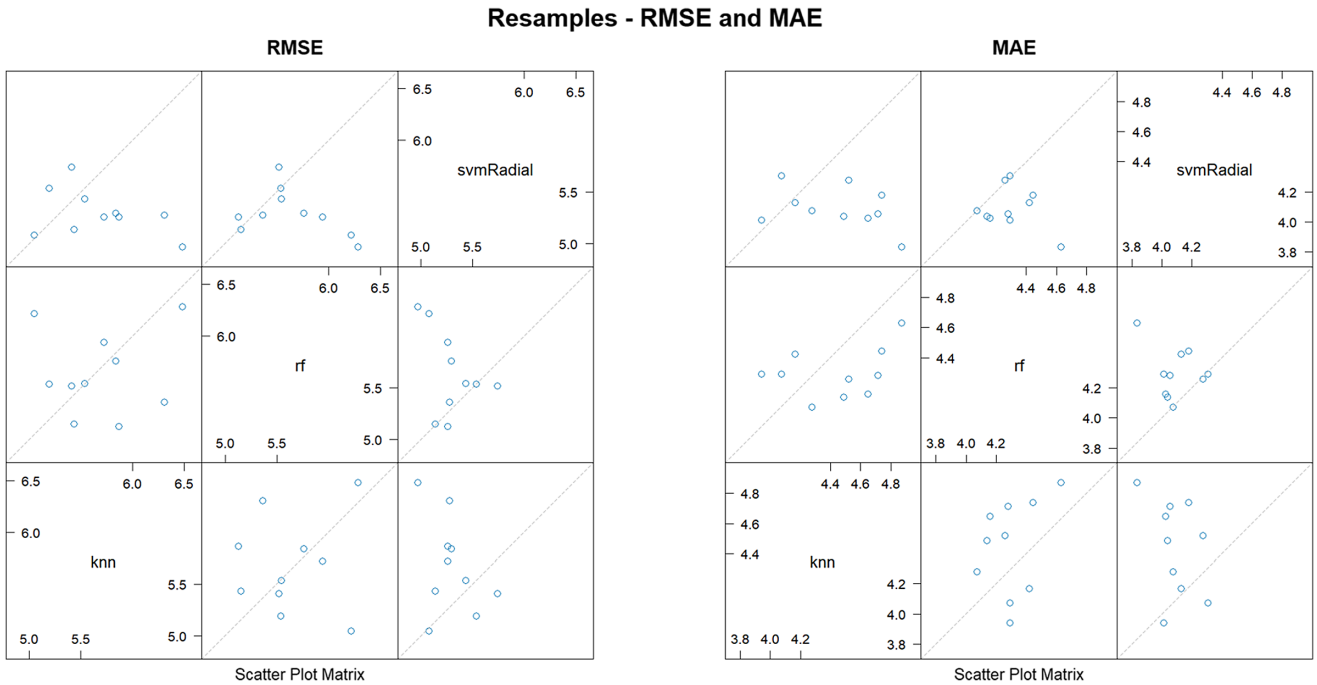

Figure 2. Performance variability analysis across cross-validation folds

These matrices illustrate how error metrics fluctuate, with each point representing a model's performance on a specific fold.

The deviations from the diagonal reference lines are notably more pronounced for the baseline models, indicating significant variability in performance. For instance, while SVM Radial generally shows lower RMSE and MAE values, suggesting strong predictive capability, KNN and RF exhibit more dispersion, revealing their vulnerabilities to different data subsets. This heterogeneity in model behavior supports the stacking approach, where the integration of diverse models leverages their individual strengths, creating a balanced system that enhances predictive stability and accuracy.

The stacked model, as shown in Figure 1, exhibits reduced variability and consistently lower error metrics across resamples, suggesting a more robust performance. By integrating different models, the stacking strategy reduces reliance on any single algorithm, thereby improving generalization and providing a more robust predictive framework. This method ensures consistent performance across varying conditions, minimizing error

variations for a more reliable prediction strategy, particularly when compared to individual baseline models.

Scatter plot matrices showing RMSE and MAE distributions for base learners (KNN, RF, SVM Radial) and Enhanced Stacked Model across 10-fold cross-validation resamples. Each point represents model performance on a specific fold, demonstrating reduced variability in the stacked ensemble.

4. Discussion and Conclusion

Accurate estimation of LDL cholesterol is critical for cardiovascular risk assessment, yet direct measurement remains a challenge in routine clinical practice. The gold standard method, ultracentrifugation, requires specialized laboratory equipment that is not widely available, making it costly and impractical for large-scale population screening. As a result, clinicians often rely on LDL estimation equations, such as the Friedewald and Martin-Hopkins formulas, which provide an alternative approach based on total cholesterol, HDL cholesterol, and triglycerides. However, these equations are not always effective due to their sensitivity to demographic factors, including age, metabolic status, and triglyceride levels. This variability can lead to misclassification of LDL levels, potentially affecting clinical decision-making, particularly in cases with high triglyceride levels or non-standard lipid profiles.

To address these challenges, we developed AutoLDLStack, an automated, machine-learning-driven approach that integrates traditional LDL estimation equations, clinical features, and machine learning algorithms into a single, cohesive system. AutoLDLStack offers a robust alternative to traditional methods by providing a fully automated solution. This system not only calculates LDL values but also incorporates classification according to both NCEP III and ESC guidelines, making it a truly plug-and-play approach. Users can train the stacking model on their dataset and then apply it to predict LDL values, ensuring estimations are tailored to their specific patient population with minimal setup required.

One of the key advantages of AutoLDLStack is its flexibility. Users can opt to utilize the full stacking ensemble model, which combines the predictive strengths of multiple machine learning algorithms, or alternatively, select a single model from the ensemble if they prefer a specific approach. This feature allows customization based on data availability, computational resources, and individual clinical needs.

Moreover, AutoLDLStack is designed for continuous adaptation. Given the evolving nature of lipid profiles and population characteristics, we recommend retraining the model periodically to keep predictions accurate and relevant. This adaptability ensures that the system evolves with changes in lifestyle trends, treatment guidelines, and demographic shifts, enhancing its long-term utility in clinical settings.

For ease of use and accessibility, we have provided a comprehensive tutorial on our GitHub repository, guiding users through the process of training, deploying, and updating the model. This resource makes AutoLDLStack accessible to researchers, clinicians, and healthcare institutions with minimal technical expertise, fostering its wide adoption.

Numerical comparisons in Table 2 highlight key performance differences between models, while Figure 1 visually confirms these findings. The learning curve shows effective generalization without overfitting, with train and test RMSE lines converging with minimal variance. The scatter plot of actual vs. predicted LDL values demonstrates the stacked model's strong predictive performance, with test points tightly clustered around the ideal prediction line, indicating minimal bias. Although SVM Radial achieved the lowest RMSE among base models, its performance isn't consistently optimal across all conditions, as shown by the variability in Figure 2. This figure illustrates the performance fluctuations across different resamples, reinforcing the need for a robust ensemble approach like AutoLDLStack.

The stacked model, as evidenced by reduced variability and lower error metrics in Figure 1, not only minimizes model bias but also maintains stable predictive performance under diverse conditions. This underscores the importance of stacking in balancing performance trade-offs and enhancing model reliability for real-world applications.

The superior accuracy and stability of our stacking model highlight its potential for clinical deployment, providing a cost-effective alternative for cardiovascular risk assessment, particularly in resource-limited settings. From a clinical perspective, AutoLDLStack bridges the gap between traditional biochemical estimation and AI-driven predictive modeling, offering accurate LDL assessments without the financial burden of direct measurement techniques. Given LDL cholesterol's critical role in guiding lipid-lowering therapies and cardiovascular risk management, this method has significant implications for improving patient outcomes.

Although AutoLDLStack is designed to be comprehensive and user-friendly, it does come with considerations. Our approach aligns with recent advances in ensemble learning for biomedical applications, where stacking methods have demonstrated consistent improvements over individual models across diverse clinical datasets [24, 25]. Our dataset was derived from a single laboratory, necessitating external validation across multiple, geographically diverse datasets to confirm its broad applicability. While stacking introduces some computational complexity, the increasing availability of machine learning frameworks in clinical research mitigates this concern.

In conclusion, our results confirm that AutoLDLStack significantly enhances LDL cholesterol estimation by integrating traditional equations, clinical features, and machine learning algorithms into an automated, adaptable, and user-friendly system. This fully automated, plug-and-play approach provides a highly accurate, scalable, and cost-effective solution for LDL prediction in clinical practice, empowering researchers and clinicians to improve cardiovascular risk assessment with ease.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support this work are available upon reasonable request to the corresponding author.

Author Contribution Statement

Petros Paplomatas: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – review & editing, Visualization, Project administration. **Maria Kantartzi:** Validation, Writing – original draft, Visualization. **Kostas Anagnostopoulos:** Conceptualization, Investigation, Resources, Data curation, Writing – original draft, Visualization, Supervision. **Aristidis Vrahatis:** Validation, Visualization, Supervision.

References

[1] Sampson, M., Ling, C., Sun, Q., Harb, R., Ashmaig, M., Warnick, R., . . . , & Remaley, A. T. (2020). A new equation for calculation of low-density lipoprotein cholesterol in patients with normolipidemia and/or hypertriglyceridemia. *JAMA Cardiology*, *5*(5), 540–548. https://doi.org/10.1001/jamacardio.2020.0013

[2] Ertürk Zararsız, G., Bolat, S., Cephe, A., Kochan, N., Yerlitaş, S. İ., Doğan, H. O., & Zararsız, G. (2022). Validation of Friedewald, Martin-Hopkins and Sampson low-density lipoprotein cholesterol equations. *PLoS ONE*, *17*(5), e0263860. https://doi.org/10.1371/journal.pone.0263860

[3] Andraschko, L. M., Gazi, G., Leucuta, D.-C., Popa, S.-L., Chis, B. A., & Ismaiel, A. (2025). Atherogenic index of plasma in metabolic syndrome—A systematic review and meta-analysis. *Medicina*, *61*(4), 611. https://doi.org/10.3390/medicina61040611

[4] Zhang, Y., Song, Y., Lu, Y., Liu, T., & Yin, P. (2025). Atherogenic index of plasma and cardiovascular disease risk in cardiovascular-kidney-metabolic syndrome stage 1 to 3: A longitudinal study. *Frontiers in Endocrinology*, *16*, 1517658. https://doi.org/10.3389/fendo.2025.1517658

[5] Fras, Z., Tršan, J., & Banach, M. (2020). On the present and future role of Lp-PLA$_2$ in atherosclerosis-related cardiovascular risk prediction and management. *Archives of Medical Science: AMS*, *17*(4), 954–964. https://doi.org/10.5114/aoms.2020.98195

[6] Won, K.-B., Han, D., Lee, J. H., Choi, S.-Y., Chun, E. J., Park, S. H., . . . , & Chang, H.-J. (2020). Atherogenic index of plasma and coronary artery calcification progression beyond traditional risk factors according to baseline coronary artery calcium score. *Scientific Reports*, *10*(1), 21324. https://doi.org/10.1038/s41598-020-78350-x

[7] Friedewald, W. T., Levy, R. I., & Fredrickson, D. S. (1972). Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical Chemistry*, *18*(6), 499–502. https://doi.org/10.1093/clinchem/18.6.499

[8] Nair, S. S., Kiran, R., Jisna, K. K., Prathima, M. B., Sushith, P., & D'sa, J. (2024). Comparison of ten formulae for calculating low-density lipoprotein cholesterol with direct low-density lipoprotein cholesterol measurement. *Current Medicine Research and Practice*, *14*(5), 192–199. https://doi.org/10.4103/cmrp.cmrp_98_24

[9] Singh, G., Hussain, Y., Xu, Z., Sholle, E., Michalak, K., Dolan, K., . . . , & Al'Aref, S. J. (2020). Comparing a novel machine learning method to the Friedewald formula and Martin-Hopkins equation for low-density lipoprotein estimation. *PLoS ONE*, *15*(9), e0239934. https://doi.org/10.1371/journal.pone.0239934

[10] Sajja, A., Park, J., Sathiyakumar, V., Varghese, B., Pallazola, V. A., Marvel, F. A., . . . , & Martin, S. S. (2021). Comparison of methods to estimate low-density lipoprotein cholesterol in patients with high triglyceride levels. *JAMA Network Open*, *4*(10), e2128817. https://doi.org/10.1001/jamanetworkopen.2021.28817

[11] Samuel, C., Park, J., Sajja, A., Michos, E. D., Blumenthal, R. S., Jones, S. R., & Martin, S. S. (2023). Accuracy of 23 equations for estimating LDL cholesterol in a clinical laboratory database of 5,051,467 patients. *Global Heart*, *18*(1), 36. https://doi.org/10.5334/gh.1214

[12] Bzdok, D., Krzywinski, M., & Altman, N. (2017). Machine learning: A primer. *Nature Methods*, *14*(12), 1119–1120. https://doi.org/10.1038/nmeth.4526

[13] Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. (2001). Executive summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). *Journal of the American Medical Association, 285*(19), 2486–2497. https://doi.org/10.1001/jama.285.19.2486

[14] Mach, F., Baigent, C., Catapano, A. L., Koskinas, K. C., Casula, M., Badimon, L., . . . , & Wiklund, O. (2020). 2019 ESC/EAS Guidelines for the management of dyslipidaemias: Lipid modification to reduce cardiovascular risk: The Task Force for the management of dyslipidaemias of the European Society of Cardiology (ESC) and European Atherosclerosis Society (EAS). *European Heart Journal*, *41*(1), 111–188. https://doi.org/10.1093/eurheartj/ehz455

[15] Paplomatas, P., Nikolidaki, M., Vrahatis, A., & Anagnostopoulos, K. (2024). Estimation of Low-Density Lipoprotein (LDL) values using equations and Machine Learning and variance calculation of LDL and Atherogenic Index of Plasma (AIP). *Academia Molecular Biology and Genomics*, *1*(1).

[16] Piepoli, M. F., Hoes, A. W., Agewall, S., Albus, C., Brotons, C., Catapano, A. L., . . . , & Verschuren, W. M. (2016). Guidelines: Editor's choice: 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts) Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). *European Heart Journal*, *37*(29), 2315.

[17] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[18] Hastie, T., Tibshirani, R., & Friedman, J. (2009). Prototype methods and nearest-neighbors. In T. Hastie, R. Tibshirani, & J. Friedman (Eds.), *The elements of statistical learning: Data mining, inference, and prediction* (pp. 459–483). Springer. https://doi.org/10.1007/978-0-387-84858-7_13

[19] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[20] Wallace, M. L., Mentch, L., Wheeler, B. J., Tapia, A. L., Richards, M., Zhou, S., . . . , & Buysse, D. J. (2023). Use and misuse of random forest variable importance metrics in medicine: Demonstrations through incident stroke prediction. *BMC Medical Research Methodology*, *23*(1), 144. https://doi.org/10.1186/s12874-023-01965-x

[21] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

[22] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R*. USA: Springer.

[23] Paplomatas, P., Krokidis, M. G., Vlamos, P., & Vrahatis, A. G. (2023). An ensemble feature selection approach for analysis and modeling of transcriptome data in Alzheimer's disease. *Applied Sciences*, *13*(4), 2353. https://doi.org/10.3390/app13042353

[24] Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine learning-based predictive models for detection of cardiovascular diseases. *Diagnostics*, *14*(2), 144. https://doi.org/10.3390/diagnostics14020144

[25] Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023). Ensemble learning for disease prediction: A review. *Healthcare*, *11*(12), 1808. https://doi.org/10.3390/healthcare11121808

How to Cite: Paplomatas, P., Kantartzi, M., Anagnostopoulos, K., & Vrahatis, A. (2025). An Artificial Intelligence and Biochemical Modelling Approach for LDL Cholesterol Prediction. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN52025919