Computers in Biology and Medicine 195 (2025) 110599

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

# Optimizing treatment to control LDL cholesterol using machine learning

Deiby Boneu Yepez [a,b,c,d,*], David Sierra Porta [b], Liz Morales Aguas [d], Fernando Manzur Jattin [c]

[a] Ciencias básicas, programa de medicina, Corporación Universitaria Rafael Núñez, Cartagena, Bolívar, Colombia
[b] Ciencias básicas, Maestría en estadística aplicada y ciencia de datos, Universidad Tecnológica de Bolívar, Cartagena, Bolívar, Colombia
[c] Centro de Diagnostico Cardiológico SAS Cartagena, Bolívar, Colombia
[d] TimeMed-IA Cartagena, Bolívar, Colombia

ARTICLE INFO

*Keywords:*
Machine learning
LDL cholesterol
Hypolipidemic agents
Risk factors

ABSTRACT

*Introduction:* Increased LDL cholesterol is one of the main risk factors for cardiovascular diseases; therefore, adequate therapy reduces the risk of developing cardiovascular disease. Artificial intelligence (AI) is a tool that can significantly help doctors select the optimal treatment aimed at each patient individually. Based on the above, the question arises: Which artificial intelligence model allows us to recommend the best treatment to control LDL levels according to the patient's cardiovascular risk factor?

*Methodology:* The performance of various machine learning models was compared to assess their ability to predict the best-individualized therapy for each patient according to their cardiovascular risk from a registry of patients at a clinic specializing in cardiovascular diseases. The Machine learning models used included: RandomForestClassifier (RFC), GradientBoostClassifier (GBC), AdaBoostClassifier (ABC), ExtraTreeClassifier (ETC), Decision Tree Classifier (DTC), KNN, Support Vector Machine (SVM), Multinomial Logistic Regression (MLR), and Naive Bayes Classifier (NBC).

*Population and sample:* The records of 166 patients with any cardiovascular risk who had LDL alterations and used some therapeutic interventions were obtained. However, four medical records did not have creatinine levels; therefore, they were excluded, counting at the end with 162 observations. Of these, a sample of 115 patients who achieved the LDL therapeutic goal was obtained.

*Results:* The Random Forest Classifier (RFC) and Gradient Boosting Classifier (GBC) demonstrated superior performance in classifying optimal LDL-lowering therapy. In contrast, Naïve Bayes Classifier (NBC) overestimated outcomes and was deemed unsuitable.

*Conclusions:* Machine learning models, particularly Random Forest Classifier, provide valuable tools for optimizing LDL control in high-risk cardiovascular patients. These models enhance clinical decision-making by enabling personalized therapy selection based on patient-specific risk factors.

## 1. Introduction

Cardiovascular diseases (CVDs) are the leading cause of mortality worldwide [1], accounting for an estimated 38 % of global deaths, according to the World Health Organization [2].

A combination of risk factors influences the formation of cardiovascular diseases where some factors are changeable and others aren't. The factors of genetics and age together with sex establish non-modifiable risk factors because they exist inherently and cannot be controlled. In contrast, modifiable risk factors encompass blood pressure, glycemic control, diet, physical activity, renal function, and body weight, among others [3]. The medical concern about dyslipidemia has elevated low-density lipoprotein (LDL) levels as its top priority among these issues. The development of cardiovascular disease depends heavily on elevated LDL levels because high-density small cholesterol particles tend to become trapped inside blood vessel walls leading to atheroma plaque development [4].

The formation of atheromatous plaques blocks arterial blood flow to various organs which results in ischemic events. The affected tissues receive less oxygen and essential nutrients when blood flow reduces. The disease evolution produces acute ischemic events through this process which affects vulnerable organs specifically heart and brain [3]. Lipid

* Corresponding author. Ciencias básicas, programa de medicina, Corporación Universitaria Rafael Núñez, Cartagena, Bolívar, Colombia.
*E-mail address:* deiby.boneu@campusuninunez.edu.co (D.B. Yepez).

https://doi.org/10.1016/j.compbiomed.2025.110599
Received 20 October 2024; Received in revised form 13 June 2025; Accepted 16 June 2025
Available online 20 June 2025

control must reach its optimal level through a risk-based analysis of each patient. Cardiovascular protection occurs when cholesterol management is both timely and effective because it prevents lipoprotein build-up on arteries which prevents atheromatous plaque development and associated ischemic events [5].

Currently, several therapeutic strategies are available to lower LDL cholesterol levels include lifestyle modifications, statins, fibrates, bile acid sequestrants (resins), ezetimibe, PCSK9 inhibitors, and plasmapheresis. However, Despite the existence of the several treatment options exist but patients still demonstrate inadequate LDL cholesterol levels [6]. Two main factors contribute to poor lipid management, therapeutic inertia and inappropriate treatment selection among the most prominent contributing factors [7].

In many cases, assessments of cardiovascular risk profiles and LDL target levels are often insufficient which causes inadequate individual treatment approaches. Additionally, therapeutic inertia remains a significant challenge, as many patients continue receiving the same dose and type of lipid-lowering therapy, regardless of their specific clinical needs. This results in some patients being underdosed, failing to achieve adequate LDL reduction, while others may be overdosed, increasing the risk of adverse effects [8]. This inertia persists even when follow-up laboratory tests indicate that LDL targets have not been met, yet treatment adjustments are not made accordingly [7].

The combined existence of cardiovascular diseases with additional health problems like diabetes or hypertension increases the likelihood of subpar LDL control. When hypertension or diabetes are coexisting conditions the medical staff may direct their attention toward achieving stricter blood pressure or glycemic control, leading to the underemphasis of cholesterol management [9].

Additionally, uncertainty about what LDL cholesterol levels is medically safe contributes significantly to the inability to reach target cholesterol levels. Current research shows that minimizing LDL levels decreases cardiovascular risks but does not cause harm to patient health or create physiological problems [10].

Another significant barrier to achieving optimal cholesterol control is the cost and accessibility of lipid-lowering therapies. Each medication exhibits different potency in LDL reduction, which directly impacts treatment costs for healthcare systems. Consequently, prescribing high-cost therapies without aligning them with individualized cholesterol-lowering targets may limit resource allocation and restrict access to optimal treatments for other patients in need.

## 2. Materials and methods

### 2.1. Study population

The predictive model was evaluated through medical data from a Cartagena-Colombia, based specialized cardiovascular clinic that treats patients across different cardiovascular conditions. A total of 166 patient records from the Latin Caribbean population in northern Colombia were reviewed. These patients belonged to diverse socioeconomic backgrounds, although specific socioeconomic classifications were not documented in the clinical records. All patients included in the study had some degree of cardiovascular risk, which was necessary for individualizing LDL targets. Their blood LDL levels needed therapeutic treatment for LDL control because healthcare professionals used individual patient-specific cardiovascular risk assessments. Four recorded medical cases did not include serum creatinine levels which impeded glomerular filtration rate (GFR) computation for cardiovascular risk classification. A total of 162 patient observations made up the final dataset because researchers excluded the incomplete medical records from the analysis. Strategic data within medical records presented all necessary patient information starting from their medical background alongside vital signs and lab results before and after therapy completion (Table 1).

Medical record was through reviewed secondary data sources that received retrospective analysis during December 2022 up to May 2023. The selection of this six-month time frame was based on the clinical observation that patients typically returned for follow-up consultations within this period, resulting in repeated assessments of the same cohort rather than the inclusion of new patients. As a result, the dataset size was inherently limited, as the patient population remained consistent without the addition of new cases.

**Table 1**
Demographic data.

| Factors | Media (DS) | Min | Max | Skewness | Kurtosis |
|---|---|---|---|---|---|
| Total | 162 | – | – | – | – |
| **Men** | n = 60 | – | – | – | – |
| | 37.04 % | – | – | – | – |
| **Women** | n = 102 | – | – | – | – |
| | 62.96 % | – | – | – | – |
| Age (years) | 65.45 (12.28) | 26 | 93 | −0.35 | 0.02 |
| Height(cm) | 163.28 (8.39) | 142 | 187 | 0.42 | −0.01 |
| Weight (Kg) | 66.6 (16.16) | 38.5 | 138.8 | 0.8 | 1.53 |
| BMI | 26.58 (5.19) | 16.8 | 46.4 | 0.61 | 0.86 |
| SBP | 124.94 mmHg (18.04) | 90 | 180 | 0.57 | 0.39 |
| DBP | 75.75 mmHg (10.59) | 50 | 110 | 0.28 | −0.13 |
| Glucose | 105.38 mg/dL (30.14) | 53 | 317 | 3.59 | 17.87 |
| LDL | 137.98 mg/dL (30.77) | 75 | 220 | 0.29 | −0.26 |
| HDL | 44.59 mg/dL (10.74) | 18 | 81 | 0.51 | 0.26 |
| Triglycerides | 139.85 mg/dL (63.36) | 47 | 540 | 0.91 | 0.94 |
| Total cholesterol | 210.73 mg/dL (36.48) | 134 | 306 | 0.18 | −0.40 |
| Creatinine | 1.11 mg/dL (0.78) | 0.5 | 10 | 9.58 | 106.15 |
| GFR | 84.16 (23.84) | 4.8 | 158.9 | −0.06 | 0.59 |

Data were collected on regarding primary cardiovascular risk evaluation variables which included established cardiovascular diseases together with diabetes status, arterial hypertension status, chronic kidney disease status, cigarette use patterns, vital signs measurements and laboratory results. Risk assessment features in this study were selected because they proved to correspond directly with cardiovascular threat and they serve extensively in clinical practice for predicting cardiovascular events. This approach supports the development of personalized treatment strategies, ensuring that therapeutic interventions are optimized according to each patient's unique risk profile.

Distribution patterns of crucial medical parameters required analysis through skewness and kurtosis values. The distribution of most measured variables was approximately normal based on their observed Skewness values. Glucose and creatinine showed positive skewness which indicates formulation of extreme data points to the right side of the distribution. The elevated kurtosis score indicates that outliers exist in these variables and could potentially affect both treatment responses and treatment classification processes. The statistical properties of these variables received consideration before data analysis to select appropriate methodology for training and evaluation and minimizing potential biases in model training and evaluation.

Many patients were female, comprising 62.96 % (102 individuals), while males accounted for 37.04 % (60 individuals). Patient's ages ranged from 26 to 93 years, with a mean age of 65.54 years (SD ± 12.28). Regarding cardiovascular history, hypertension was the most prevalent condition, affecting 90.12 % (146 individuals). Among these, 88 women (60.27 %) and 58 men (39.72 %) were diagnosed with hypertension. The mean LDL cholesterol level in the study population was 144.2 mg/dL (SD ± 29.64).

The cardiovascular risk assessment method followed the 2019 ESC/EAS guidelines [11] that integrated diabetes results alongside hypertension and chronic kidney disease findings and history of cardiovascular events together with information about smoking habits. Personal

and demographic characteristics such as age along with sex distribution and Body Mass Index (BMI) were processed together with lipid profiles because they directly influence lipid-lowering treatment results. The measurement of BMI consisted of dividing weight in kilograms by the square of height in meters. The research divided patients into four BMI groups based on these criteria: Underweight had BMI below 18.49, Normal weight existed between 18.5 and 24.99, Overweight ranged between 25 and 29.99, and Obesity involved a BMI of 30 or higher.

To evaluated chronic kidney disease (CKD) stage and assess its contribution to cardiovascular risk, the glomerular filtration rate (GFR) was calculated using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation. This equation provides a more accurate cardiovascular risk assessment compared to serum creatinine alone. The Spanish Society of Nephrology (Senefro) filtration calculator was used to estimate eGFR based on the CKD-EPI algorithm. The CKD-EPI equation provides improved eGFR calculation for individuals with normal kidney function since it does better than the MDRD equation at measuring eGFR at higher filtration rates.

This study utilized the 2012 KDIGO guidelines to create chronic kidney disease (CKD) classification divisions through glomerular filtration rate (GFR) measurement combined with presence of albuminuria. However, since albuminuria results were not consistently available in the database, stratification was performed solely based on GFR levels. CKD stages were defined as follows: Stage 1: GFR >90 mL/min/1.73 $m^2$, Stage 2: GFR 60–89 mL/min/1.73 $m^2$, Stage 3a: GFR 45–59 mL/min/1.73 $m^2$, Stage 3b: GFR 30–44 mL/min/1.73 $m^2$, Stage 4: GFR 15–29 mL/min/1.73 $m^2$ and Stage 5: GFR <15 mL/min/1.73 $m^2$. Patients who experienced a major cardiovascular event, including acute coronary syndrome, peripheral arterial stenosis, or ischemic stroke, were classified within the same category as these represent the most clinically significant cardiovascular outcomes.

The studied population revealed a 24.22 percent (39 patients) rate of cardiovascular disease occurrence. Data showed that arterial recanalization received treatment by 36 patients which amounted to 22.36 % among the cardiovascular disease patient population (Table 2).

Type 2 diabetes mellitus was the second most prevalent condition after hypertension, affecting 37.03 % (60 patients). Among diabetic patients, women were more affected (65 %) than men (35 %), making diabetes more prevalent in the female population. The mean baseline LDL level in diabetic patients was 129.15 mg/dL, while the mean blood glucose level was 121.33 mg/dL (SD ± 42.20), indicating that most patients were within their glycemic targets.

Regarding smoking status, none of the patients reported being current smokers. Among former smokers, 67 patients were identified, consisting of 31 men (46.27 %) and 36 women (53.73 %). Chronic kidney disease (CKD) is classified as a high cardiovascular risk factor when renal function falls below 59 mL/min/1.73 $m^2$, according to ESC/EAS guidelines [11]. Based on this criterion, 18 patients were identified with moderate stage 3 CKD, with a GFR between 30 and 59 mL/min/1.73 $m^2$. Among them, 7 were men (38.89 %) and 11 were women (61.11 %). The mean LDL level in this group was 125.89 mg/dL (SD ± 28.41).

Severe chronic kidney disease (CKD) is defined as renal function below 30 mL/min/1.73 $m^2$. In this study, two patients were identified with stage 4 CKD, one male and one female, with a mean LDL level of 155.5 mg/dL (SD ± 67.18). Regarding end-stage CKD (stage 5), only one female patient was observed. The lipid-lowering therapy adjusted to cardiovascular risks led to LDL target in 67.9 % (110) of the total patient population. Patients with middle or low grades of cardiovascular risk showed the best therapeutic outcome with LDL control reaching 80.71 % and 90 % among this population segment. The successful achievement of LDL targets reached 50 % in patients with high cardiovascular risk (Table 3, Fig. 1).

The reduction of LDL cholesterol values differed based on individual drug therapies when used separately or in combination with statins. The addition of PCSK9 inhibitors to medical treatment produced the most significant LDL cholesterol reduction among all therapies due to their multi-faceted cholesterol-lowering capacity per evidence from several clinical trials [12]. This therapeutic approach showed the minimum response variability because its clinical outcomes spanned between 60.86 % and 82.84 %. Statin monotherapy generated the highest variability when assessing LDL response levels. Statistical analysis from Table 4 shows Atorvastatin 40 mg caused LDL reduction from 10.18 % to 73.75 % with an average of 46.36 % (SD ± 14.24).

**Table 2**
Cardiovascular risk factors.

| History | Total | Men | Women | LDL Mean (SD) |
|---|---|---|---|---|
| **Cardiovascular** | n = 39 | 27 | 12 | 134.16 mg/dL (36.14) |
| **disease** | 24.22 % | 69.23 % | 30.76 % | |
| **Coronary** | n = 36 | 24 | 12 | 132.75 mg/dL (36.7) |
| **Revascularization** | 22.36 % | 66.66 % | 33.33 % | |
| **Low weight** | n = 8 | 1 | 7 | 135.25 mg/dL (32.86) |
| | 4.93 % | 12.5 % | 87.5 % | |
| **Normal weight** | n = 61 | 22 | 39 | 144.42 mg/dL (31.10) |
| | 37.65 % | 36.07 % | 63.93 % | |
| **Overweight** | n = 56 | 25 | 31 | 134.93 mg/dL (28.99) |
| | 34.56 % | 44.64 % | 55.36 % | |
| **Obesity** | n = 40 | 12 | 28 | 133.2 mg/dL (31.78) |
| | 24.69 % | 30 % | 70 % | |
| **Diabetes** | n = 60 | 21 | 39 | 129.15 mg/dL (27.54) |
| | 37.03 % | 35 % | 65 % | |
| **Hypertension** | n = 146 | 58 | 88 | 136.86 mg/dL (29.64) |
| | 90.12 % | 39.72 % | 60.27 % | |
| **Smoker** | Never = 95 | 29 | 66 | 143.44 mg/dL (31.27) |
| | Former = 67 | 31 | 36 | 130.22 mg/dL (28.50) |
| | Active = 0 | 0 | 0 | |
| **Chronic Kidney Disease** | Stage 5 = 1 | 0 | 1 | 169.0 mg/dL (0) |
| | Stage 4 = 2 | 1 | 1 | 155.50 mg/dL (67.18) |
| | Stage 3b = 4 | 0 | 4 | 125.5 mg/dL (41.1) |
| | Stage 3a = 15 | 8 | 7 | 122.2 mg/dL (28.89) |
| | Stage 2 = 76 | 27 | 49 | 140.39 mg/dL (32.34) |
| | Stage 1 = 64 | 24 | 40 | 138.08 mg/dL (28.04) |

**Table 3**
Cardiovascular risk stage.

| Risk | # Patients | Initial LDL (SD) | LDL control (SD) | Controlled |
|---|---|---|---|---|
| **Very high risk** | n = 40 24.69 % | 135.13 mg/dL (38.35) | 56.6 mg/dL (27.68) | 20 50 % |
| **High risk** | n = 48 29.63 % | 134.52 mg/dL (34.52) | 68.58 mg/dL (23.47) | 26 54.17 % |
| **Moderate risk** | n = 64 39.51 % | 141.97 mg/dL (24.69) | 75.55 mg/dL (20.88) | 55 85.94 % |
| **Low risk** | n = 10 6.17 % | 137.4 mg/dL (17.61) | 82.1 mg/dL (23.09) | 9 90 % |

### 2.2. Statistical analysis and modeling

The dataset was studied with descriptive statistics which included both measures of central tendency and measures of dispersion. The

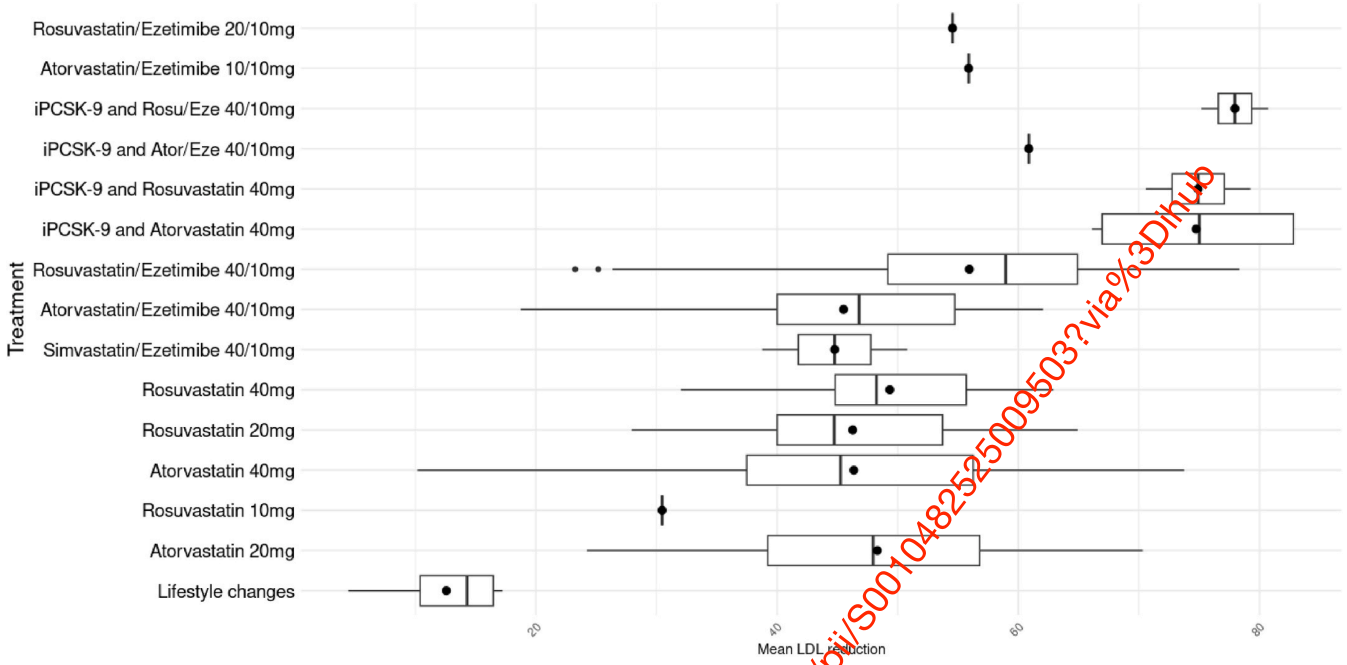


**Fig. 1.** Difference in LDL reduction between medications used.

**Table 4**
Mean LDL reduction.

| Treatments | Doses | n° | LDL reduction | Min | Max |
|---|---|---|---|---|---|
| Lifestyle Changes | Daily | 4 | 12.57(5.79 %) | 4,46 | 17.21 |
| Atorvastatin | 20 mg | 12 | 48.31 (14.68 %) | 24,22 | 70,31 |
| | 40 mg | 63 | 46.36 (14.24 %) | 10,18 | 73,73 |
| | 10 mg | 1 | 30.46 (0 %) | – | |
| Rosuvastatin | 20 mg | 5 | 46.26 (13.97 %) | 27,94 | 64,93 |
| | 40 mg | 14 | 49.34 (9.68 %) | 32,03 | 62,91 |
| Simvastatin/Ezetimibe | 40/10 mg | 2 | 44.76 (44.76 %) | 38,75 | 50,78 |
| Atorvastatin/Ezetimiba | 10/10 mg | 1 | 55.88 (0 %) | – | – |
| | 40/10 mg | 17 | 45.51 (12.39 %) | 18.75 | 62.04 |
| Rosuvastatin/Ezetimibe | 20/10 mg | 1 | 54.54 (0 %) | | |
| | 40/10 mg | 33 | 55.94 (14.39 %) | 23.26 | 65.53 |
| iPCSK9 + Statins | e/15 days | 1 | 60.86 (0 %) | – | – |
| Atorvastatin/iPCSK9 | 40 mg | 4 | 74.74 (9.34 %) | 66.10 | 82.84 |
| Rosuvastatin/iPCSK9 | 40 mg | 2 | 74.90 (6.11 %) | 70.59 | 79.23 |
| Rosuvastatin/Ezetimibe/ iPCSK9 | 40/10 mg | 2 | 77.94 (3.90 %) | 75.19 | 80.70 |

Shapiro-Wilk test determined normality conditions thus non-parametric statistical methods were used accordingly. The evaluation of variable relationships involved multiple statistical tests that incorporated Student's t-test, ANOVA and chi-square analysis together with regression approaches. The model required reduced predictors after we eliminated variables that presented high correlation to achieve superior performance and better interpretability. The variable selection process relied on clinical significance to choose variables which directly link to LDL control measures and cardiovascular health risk factors.

The multiclass classification technique let us find the best LDL-lowering therapy for individual patients in our predictive modeling analysis. A 5-fold cross-validation method was applied to training and validating models for preventing overfitting and enhancing the model's general use capabilities. Technical optimization through hyperparameter tuning of each model occurred to maximize performance. The goal to pick the best treatment from various possibilities led us to choose multiclass classification as our most fitting solution. The model determines suitable medical treatment by analyzing both LDL measurements with cardiovascular risk component tests.

### 2.3. Machine learning models

The selection of machine learning models occurred considering their tested success rates in classification work combined with their compatibility with structured clinical information. The choice of supervised learning algorithms fit the labeled dataset because it allowed direct correlations between patient attributes and best LDL-lowering treatments. The examined models consisted of Random Forest (RFC), Gradient Boosting (GBC), AdaBoost (ABC), Extra Trees (ETC), Decision Trees (DTC), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Multinomial Logistic Regression (MLR) together with Naïve Bayes (NBC) (Table 5).

The models demonstrated great ability to process complex non-linear connections between variables as they commonly applied to medical classification tasks. RFC together with GBC and ABC and ETC constitute ensemble methods since they use multiple decision trees for generating accurate predictions. The evaluation stage of this study included both Logistic Regression and Naïve Bayes models as baseline analyses. SVM and KNN received evaluation in high-dimensional feature spaces although their poor interpretability in clinical practice proved to be their main weakness. To identify precise results and medication selection models which complied with clinical needs the study examined various methods.

- RandomForestClassifier (RFC): Breiman developed this machine learning model in 2001 [13]. It is a predictive and classification method that uses binary divisions of predictor variables to determine results, achieving high-probability and low-probability results. The

**Table 5**
Comparison of machine learning models.

| Model | Type | Key Characteristics | Advantages | Disadvantages |
|---|---|---|---|---|
| RFC | Ensemble (Trees) | Uses multiple decision trees to improve accuracy | High accuracy, robust to noise, handles nonlinear data well | Computationally expensive, less interpretable |
| GBC | Ensemble (Boosting) | Builds trees sequentially, minimizing errors at each step | High predictive power, handles missing data well | Sensitive to overfitting if not tuned properly |
| ABC | Ensemble (Boosting) | Assigns more weight to misclassified instances | Improves weak learners, good generalization | Sensitive to noisy data and outliers |
| ETC | Ensemble (Trees) | Like to Random Forest but uses more randomization | Faster than RFC, reduces variance | Less accurate than RFC in some cases |
| DTC | Tree-Based | Simple structure for decision-making | Easy to interpret, fast training | Prone to overfitting, less robust |
| SVM | Kernel-Based | Finds the best hyperplane for classification | Works well with high-dimensional data | Computationally expensive with large datasets |
| KNN | Instance-Based | Classifies based on nearest neighbors | Simple, non-parametric, adaptable | Slow with large datasets, sensitive to irrelevant features |
| MLR | Regression | Generalizes logistic regression for multiple classes | Interpretable, efficient for linear relationships | Assumes linearity, struggles with complex relationships |
| NBC | Probabilistic | Uses Bayes' Theorem for classification | Fast, works well with small datasets | Assumes independence between features, often unrealistic |

model is trained with increasingly complex splits forming the branches of each tree, each tree continuing to split until a forest is formed [14].

- AdaBoostClassifier (ABC): A Random Forest-derived meta estimator that adjusts its prediction by training the model on an original training data set, then adjusts additional copies of the regression on the same data set but where the weights of the instances are adjusted according to the error of the current prediction, repeating until the learning process is complete [15].
- GradientBoostClassifier (GBC): It's a machine learning system which utilizes regression along with classification issues according to the algorithm. The reinforcement algorithm receives universal acclaim because it converts several bilinear learning models into strong performance models. The process uses each step to build new models which execute loss function optimization based on the previous model. The loss function gradient calculation for current ensemble predictions helps train a new bilinear model which reduces the gradient value. The concept of gradient emerges because the algorithm depends on gradient descent to reduce loss. The predictive function of gradient boosting suits regression tasks better than linear regression which determines only continuous values like house prices. This algorithm creates its structure through multiple stages so it optimizes any differentiable loss function [16].
- ExtraTreesClassifier (ETC): is an automatic learning algorithm designed to address regeneration and classification tasks. It belongs to the category of ensemble learning techniques, which combines the results of multiple independent decision trees. Its operation is like that of the well-known Random Forest Classifier. The ETC stands out as a powerful option compared to the RFC since it also performs a selection of significant features. The algorithm is constructed by creating several decision trees using the original training sample. At each decision point, each of these trees receives a random selection of 'k' features from the total set of available features. Each decision tree, based on certain mathematical criteria, generally the Gini index, chooses the most relevant characteristic to divide the data. This process of random selection of characteristics results in the generation of multiple unrelated decision trees. One of the key advantages of the ETC model is its ability to deal with noisy data and characteristics that may not be relevant. This robustness makes it less susceptible to the influence of data that do not add substantial value to the model, which is particularly beneficial in realistic and challenging data environments [17].
- Decision Trees Classifier (DTC): is a supervised learning method exists as an algorithm which finds frequent use during classification and regression tasks. The model appears as a tree with each evaluation of an attribute displayed at internal nodes and the result of evaluations identified through branches before class labels or numeric values appear at leaf nodes. Decision trees attain both popular ease-of-understanding attributes as well as the capability to handle both numeric and categorical data. Multiple decision tree variations exist alongside features that allow them to work with other machine learning models like ABC and GBC for enhanced effectiveness. Decision trees serve medical applications by analyzing medical pathologies and performing image interpretation and diagnosis of COVID-19 [18].
- Support Vector Machines (SVM): Is a powerful classification method in machine learning. The central idea behind SVM is to find the optimal hyperplane that best separates different classes in a feature space, maximizing the margin between classes. SVM seeks to find a hyperplane in a high-dimensional feature space that can optimally separate instances of different classes. A hyperplane is a generalization of the concept of a plane in higher dimensions.
- K-Nearest Neighbors (KNN): It's a supervised learning algorithm. Although it's working principle doesn't involve an explicit training process like some supervised algorithms, such as logistic regression or neural networks, it's still supervised because it requires labeled input data to classify new instances. In KNN, to classify a new instance, the algorithm finds the "k" closest data points in the training set and classifies the new instance into the most common category among these "k" neighbors. The key here is that the training set must be pre-labeled so that the algorithm can calculate the distance and determine the class of the near neighbors. Therefore, although KNN does not follow an explicit training phase, it is still a supervised learning algorithm because it relies on labeled training data to make predictions [19].
- Multinomial Logistic Regression (MLR) is a type of supervised learning algorithm used to predict a target variable's probability. The nature of the target variable is dichotomous, meaning there would only be two possible classes. The probability that an observation belongs to one of the two classes is calculated. The logistic classification algorithm predicts the probability that the target variable belongs to a particular class. If the probability of membership in a class is greater than the 50 % threshold, it is classified as belonging to that class, and if it is less than 50 %, it is classified as belonging to the other class. This is done using the sigmoid function, which converts any value to a value between 0 and 1 [20].
- Naive Bayes Classifier (NBC): It's works as a supervised method that uses Bayes' theorem to categorize data by making independence assumptions about features relative to the class variables. NBC remains efficient even though it is simple to understand and operates effectively in practical applications including spam filtering and duplicate document detection and text classification among others. NBC operates by determining the probability that a particular instance belongs to a class when analyzing its attributes. It relies on

features showing conditional independence in practice which enables the model to operate effectively even when processing big data sets. The selection process for classification takes place when an instance obtains its highest posterior probability classification [21].

### 2.4. Validation of machine learning models

The assessment of machine learning models' performance and generalization occurred through k-fold cross-validation procedures. The chosen cross-validation method included 5-fold splitting of the dataset into equal parts called folds which were randomly selected. The model training incorporated four subsets which served alongside the validation set in its cycle to build the connection between input and output parameters. The validation method applied each set of data five times to obtain performance metrics during validation of each subset despite its repeated use as validation data precisely once before averaging the obtained values. The chosen method prevents model overfitting and generates more reliable predictive performance estimates specifically for cases involving small datasets like the present research. The use of 5-fold cross-validation provides dependable results because it ensures each observation gets employed for both training and validation phases independently.

The hyperparameters of the machine learning algorithms were optimized using a grid search approach combined with 5-fold cross-validation. For each model, a specific range of hyperparameters was defined based on commonly used values in the literature and practical considerations to balance model complexity and performance. Both RFC and ETC received structure optimization through adjusting their number of estimators between 100 and 1000 and maximum tree depth spanning 10 to 30 levels inclusive of no limit. Hybrid GBC and ABC went through parameter optimization by examining different estimator counts from 200 to 1000 and learning rate variations between 0.01 and 0.5. SVM needed parameter optimization where C values between 0.01 and 10.0 were tested together with three different kernels ('linear', 'rbf', 'poly') under two gamma settings ('scale', 'auto'). The optimization of DTC involved maximizing depth between 10 and 50 including unlimited depth. The KNN algorithm was optimized through adjustments to both its number of neighbors from 3 to 15 and its weighting mechanisms between 'distance' and 'uniform'. MLR received three different optimizer C values of 1.0, 2.0, 3.0 accompanied by using the 'lbfgs' solver for tuning. NBC didn't need hyperparameter optimization since it conducts operations through probabilistic methods. Cross-validation allowed identification of the select best hyperparameters through minimization of Root Mean Squared Error (RMSE) which produced both optimized generalization and accuracy in model performance (Fig. 2).

### 2.5. Performance metrics and statistical programs

In this study, recall wasn't included because accuracy proved comparable in evaluation of this setting. We didn't use Area Under the Curve (AUC) because its main application involves binary classification problems. The evaluation of machine learning models used: Accuracy, Precision and F1-score for performance assessment.

- Accuracy: In the context of multi-class classification, this metric calculates the exact number of instances that received proper classifications in multi-classification systems. A prediction achieves accuracy only when every single predicted label matches exactly with the true labels within the dataset.
- Precision: Precision evaluates the percentage of correct positive predictions which represent a value percentage of both valid positive and incorrect positive results. When the model exhibits high precision scores it proves to be reliable in its positive classifications because it indicates fewer incorrect predictions.
- F1-score: Defined as the harmonic mean of precision and recall, the F1-score penalizes models that exhibit a significant imbalance between precision and recall. The main utilization of this score focuses on establishing balance between precision and recall metrics yet it maintains value as an evaluation metric for the model's success in positive instance detection alongside error reduction.

The metric of accuracy serves commonly as a model evaluation tool yet lacks sufficient capacity to measure performance when class distribution becomes unbalanced. Some categories with few patient instances required additional evaluation metrics such as precision and F1-score to achieve complete assessment effectiveness. The evaluation of the model's accuracy needed precision because it assessed the model's capability to identify appropriate patients for lipid-lowering medication correctly and reduce wrong test results. The F1-score served as the primary assessment metric for dealing with class distribution skewness because it balances precision and recall measurement. The study delivers a more comprehensive performance evaluation through its



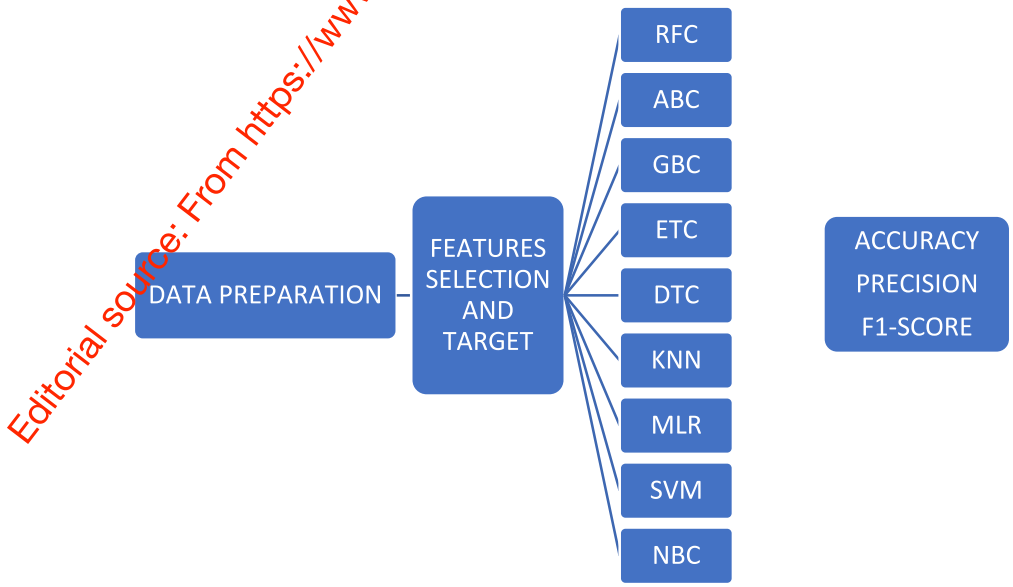


**Fig. 2.** Schematic representation of the machine learning model analysis process.

combined reporting of accuracy with precision and the F1-score to assess model behavior in authentic clinical scenarios.

The clinical and paraclinical patient data entered Microsoft Excel (Microsoft Office 2023) database for storage. Statgraphics Centurion software version 19.1.2 (64-bit) carried out the statistical examinations. A two-factor factorial design within Statgraphics analyzed different machine learning models' capacity to determine lipid-lowering therapies for specific cholesterol levels in single-interaction conditions. The visualization and image creation tasks took place through RStudio version 2023.06.2. A collection of LDL cholesterol reduction data came from both scientific literature and clinical records after which various machine learning models ran to forecast the best medication for reaching specific LDL reduction targets. The use of Python (version 3.10) with Scikit-learn (version 1.5.2) enabled machine learning implementation because these tools provided both free statistical and machine learning libraries that simplfied data processing and model evaluation and performance analysis. The performance assessment of different machine learning models used an analysis of variance (ANOVA) which analyzed results at a 95 % confidence interval, considering a P-value $<0.05$ as statistically significant.

### 2.6. Ethical considerations

This study qualifies as minimal-risk research under the provisions of Article 11 from Resolution 8430 of 1993 that governs Colombian law. The research design followed retrospective observational methods while avoiding any contact with either patients or their family members. Since this study analyzed secondary data there was no need for acquiring informed consent. Additionally, the study received approval from the Research Ethics Committee of the Centro de diagnostic cardiologico for Biomedical Research, as documented in Minutes No. 102, during a meeting held on April 12, 2024.

The scientific director of the institution authorized data collection procedures. The study maintained patient confidentiality through complete hiding of identifiable information as it adhered to ethical research standards throughout its duration.

## 3. Results

### 3.1. Factors influencing failure to achieve the LDL target

The dataset contained many outliers which meant we could not remove them because such an approach would harm sample size significantly and possibly include selection bias. The researchers evaluated two alternative techniques for extreme value treatment (winsorization) and robust statistical transformations. However, after clinical validation, these outliers represented real patient variability instead of entry mistakes thus providing evidence to keep them in the model for accurately representing the diversity of treatment responses. The analysis purpose required keeping the original data untouched since any normalization methods would affect its natural composition leading to invalid results. A logistic regression analysis determined the principal elements that caused LDL control failure while permitting outlier data inclusion to maintain data authenticity. The study data demonstrated that past cardiovascular incidents together with baseline LDL measurements and LDL treatment effectiveness and general cardiovascular risk level were the primary contributors to nonachievement of LDL goals based on the findings in Table 6. The data analysis confirmed statistical relationships between these factors and meeting the LDL control targets ($p < 0.05$).

Table 7 shows the LDL values from baseline assessments between patients whose treatment was unsuccessful and those achieving their prescribed LDL targets. Data showed a significant difference in LDL cholesterol between the two groups ($p = 0.0212$) while setting risk stratification at the primary focus. The study results demonstrate why specific treatment approaches should follow risk assessment to

**Table 6**
Logistic regression analysis of factors influencing failure to achieve the LDL target.

| | Estimate | Error Est. | z value | Pr(>(z)) |
|---|---|---|---|---|
| (Intercept) | 62.764630 | 19.278327 | 3.256 | 0.00113 ** |
| Age | 0.032088 | 0.065827 | 0.487 | 0.62594 |
| Sex | 0.805869 | 1.327709 | 0.607 | 0.54388 |
| IMC | 0.172751 | 0.133904 | 1.290 | 0.19701 |
| HTA | −0.186663 | 2.432350 | −0.077 | 0.93883 |
| Diabetes | −0.369109 | 2.173694 | −0.170 | 0.86516 |
| Initial LDL | −0.374389 | 0.119709 | −3.127 | 0.00176 ** |
| LDL Reduction | 0.760813 | 0.252929 | 3.008 | 0.00263 ** |
| CV Disease | 6.880410 | 3.283952 | 2.095 | 0.03616 * |
| Treatments | −0.062324 | 0.236138 | −0.264 | 0.79183 |
| TFG | −0.007585 | 0.035714 | −0.212 | 0.83182 |
| CV Risk | −16.660644 | 5.313780 | −3.135 | 0.00172 ** |
| Glucose | −0.018259 | 0.026658 | −0.685 | 0.49339 |

**Table 7**
Difference in Expected LDL Percentage Reduction: Observed vs. Clinical Trial Data.

| Treatments | Dose | Patient | Expected | Observed | Clinical trials |
|---|---|---|---|---|---|
| Lifestyle Changes | Daily | 4 | 5.69 | 12.57 | 10 |
| Atorvastatin | 20 mg | 12 | 23.72 | 48.31 | 41.4 |
| | 40 mg | 63 | 36.39 | 46.36 | 46.2 |
| Rosuvastatin | 10 mg | 1 | 21.88 | 30.46 | 44.1 |
| | 20 mg | 5 | 27.91 | 46.26 | 49.5 |
| | 40 mg | 14 | 39.20 | 49.34 | 54.7 |
| Ezetimibe + Statine | 10 mg | 2 | 44.44 | 30.30 | 20 |
| Simvastatin/ Ezetimibe | 40/10 mg | 15 | 44.90 | 44.76 | 65.3 |
| Atorvastatin/ Ezetimibe | 10/10 mg | 1 | 31.37 | 55.84 | 54.8 |
| | 40/10 mg | 2 | 43.84 | 51.12 | 64.5 |
| Rosuvastatin/ Ezetimibe | 20/10 mg | 1 | 58.33 | 54.54 | 64 |
| | 40/10 mg | 33 | 45.65 | 55.94 | 71.6 |
| iPCSK9 + Statins | Every 15 days | 1 | 23.91 | 60.86 | 60 |
| Atorvastatin/ iPCSK9 | 40 mg | 4 | 63.68 | 82.82 | 79 |
| Rosuvastatin/ iPCSK9 | 40 mg | 2 | 48.77 | 74.91 | 83 |
| Rosuvastatin/ Ezetimibe/ iPCSK9 | 40/10 mg | 2 | 58.86 | 77.94 | 87 |

**Table 8**
Performance of machine learning models applied to lipid-lowering treatment selection.

| Metrics | Performance | | |
|---|---|---|---|
| Models | Accuracy (SD) | Precision (SD) | F1-score (SD) |
| RFC | 0.8182 (0.0426) | 0.8439 (0.0343) | 0.8247 (0.0397) |
| GBC | **0.8218 (0.0516)** | **0.844 (0.0413)** | 0.8277 (0.0495) |
| ABC | 0.52 (0.1403) | 0.8025 (0.0238) | 0.612 (0.1029) |
| ETC | 0.8182 (0.0453) | 0.8483 (0.03) | **0.8249 (0.0413)** |
| DTC | 0.8018 (0.0436) | 0.8306 (0.0343) | 0.8073 (0.0399) |
| SVM Classifier | 0.4291 (0.0106) | 0.8303 (0.0532) | 0.5524 (0.0061) |
| KNN Classifier | 0.4618 (0.1559) | 0.6633 (0.0843) | 0.5354 (0.1269) |
| MLR | 0.4146 (0.0123) | 0.6827 (0.0207) | 0.5105 (0.0139) |
| Naive Bayes | 0.3163 (0.0503) | 0.5318 (0.0443) | 0.3395 (0.0363) |

maximize the effectiveness of LDL cholesterol management when improving patient outcomes (Table 8).

### 3.2. Comparison of expected, observed, and reported responses in clinical trials

Table 7 presents the distribution of patients across different lipid-lowering therapies, specifying the number of individuals receiving each treatment. Additionally, the table differentiates between observed LDL reductions in real-world patients and expected LDL reductions based on clinical trial data. The expected LDL reductions were estimated using the percentage decrease in baseline LDL levels relative to target LDL levels, as defined by cardiovascular risk categories. This methodology aligns with established clinical guidelines, which recommend more intensive LDL-lowering targets for high-risk patients.

Based on this comparative analysis, a multifactorial experimental design was conducted using the percentage of LDL reduction as the response variable, comparing observed reductions in real-world patients with those reported in clinical trials. ANOVA analysis indicated significant variations in LDL-lowering efficacy across treatments ($p < 0.05$), aligning with prior clinical research. However, treatment effectiveness was notably higher in clinical trials than in real-world studies, with a statistically significant difference ($p = 0.0083$) (Fig. 3).

### 3.3. Data preparation

#### 3.3.1. Removal of variables with high collinearity

A high correlation was identified among several variables, leading to suboptimal model performance due to the dilution of feature importance among correlated predictors (Fig. 4). To fix this and improve predictions of machine learning models, highly correlated variables were identified and removed where necessary. The most significant correlations observed were: Cardiovascular risk and LDL target ($r = 0.98$), Total cholesterol and baseline LDL levels ($r = 0.93$), BMI and weight ($r = 0.89$), as well as BMI and BMI category ($r = 0.91$), Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) ($r = 0.62$), Sex and height ($r = 0.62$) and cardiovascular disease and diabetes with cardiovascular risk ($r = 0.65$ and $r = 0.58$, respectively). Addressing these correlations was essential to prevent redundancy and improve model interpretability and performance.

To mitigate the issue of high collinearity, a regression-based approach was applied, restricting the selection of variables to those with a correlation coefficient greater than 0.4 with the target variable (prescribed medication), as illustrated in Fig. 5. Based on this criterion, the following 15 variables were retained for model training: Sex, Age, Hypertension, Glomerular Filtration Rate (GFR), Smoking, Cardiovascular Risk, High-Density Lipoprotein (HDL), Triglycerides, Total Cholesterol, Glucose, Creatinine, Systolic Blood Pressure (SBP), LDL Reduction, LDL Target and BMI Category. Despite reducing the number of predictors to 15 features to enhance machine learning model training, classification performance remained suboptimal. Furthermore, some variables included in the model were determined to have limited clinical relevance in therapy selection or LDL reduction. As a result, smoking status was excluded from the analysis, as none of the patients in the dataset were current smokers.

Additionally, since creatinine values are used to calculate the glomerular filtration rate (GFR) through the CKD-EPI equation, both creatinine and GFR were consolidated under the broader category of kidney disease. Given that the primary objective of lipid-lowering therapy is to target LDL levels, the variables total cholesterol, HDL, and triglycerides were excluded from the analysis. Furthermore, in clinical practice, factors such as sex and blood pressure are not considered key determinants in the selection of lipid-lowering treatment.

Following the removal of highly collinear variables and the elimination of features deemed clinically insignificant in therapy selection, the final set of predictive variables included: **LDL Reduction, Age, Cardiovascular Risk, Stages of Chronic Kidney Disease, Presence of Arterial Hypertension.** The target variable for prediction was lipid-lowering treatment selection.

#### 3.3.2. Application of machine learning models to different treatments

The machine learning models performance exhibited variable metrics during classification stage of all treatment options whereby predictive accuracy, precision and F1-score differed substantially. Predictive excellence came from ensemble methods as the models showed outstanding results including RFC (accuracy: $0.8182 \pm 0.0426$, F1-score: $0.8247 \pm 0.0397$), GBC (accuracy: $0.8218 \pm 0.0516$, F1-score: $0.8277 \pm 0.0495$) and ETC (accuracy: $0.8182 \pm 0.0453$, F1-score: $0.8249 \pm 0.0413$). These results highlight the robustness of ensemble models in their ability to find the best possible treatment approaches. The predictive power of NBC and MLR models was noticeably lower because of multi-class complexity and clinical overlapping characteristics between treatment groups (Table 5). These models demonstrated accuracy rates of $0.3163 \pm 0.0503$ and F1-score of $0.3395 \pm 0.0363$ and accuracy rates of $0.4146 \pm 0.0123$ and F1-score of $0.5105 \pm 0.0139$ respectively. These findings underscore a critical challenge in predicting optimal lipid-lowering therapy when all treatment options are considered as distinct categories. The substantial variability in model performance suggests that grouping treatments based on therapeutic similarities or established clinical guidelines may enhance classification accuracy and improve model interpretability.

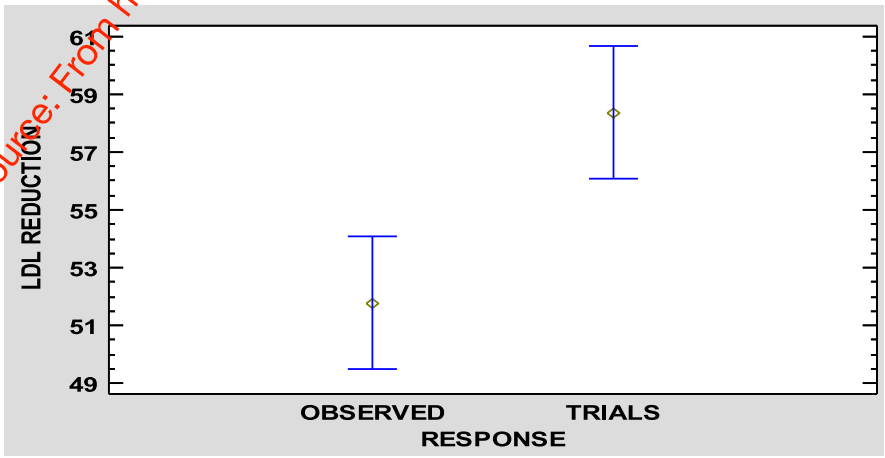


**Fig. 3.** Comparison of LDL Reduction: Clinical Trial Data vs. Observed Patient.

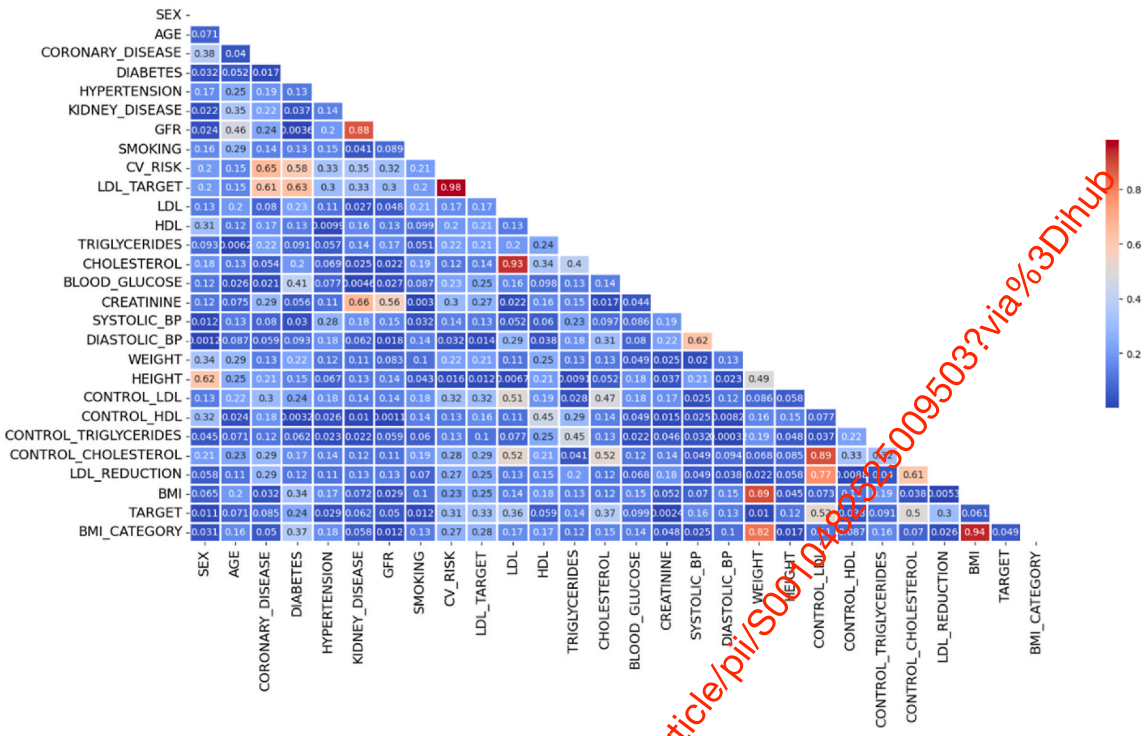


**Fig. 4.** Correlation of feature variables with the target variable.

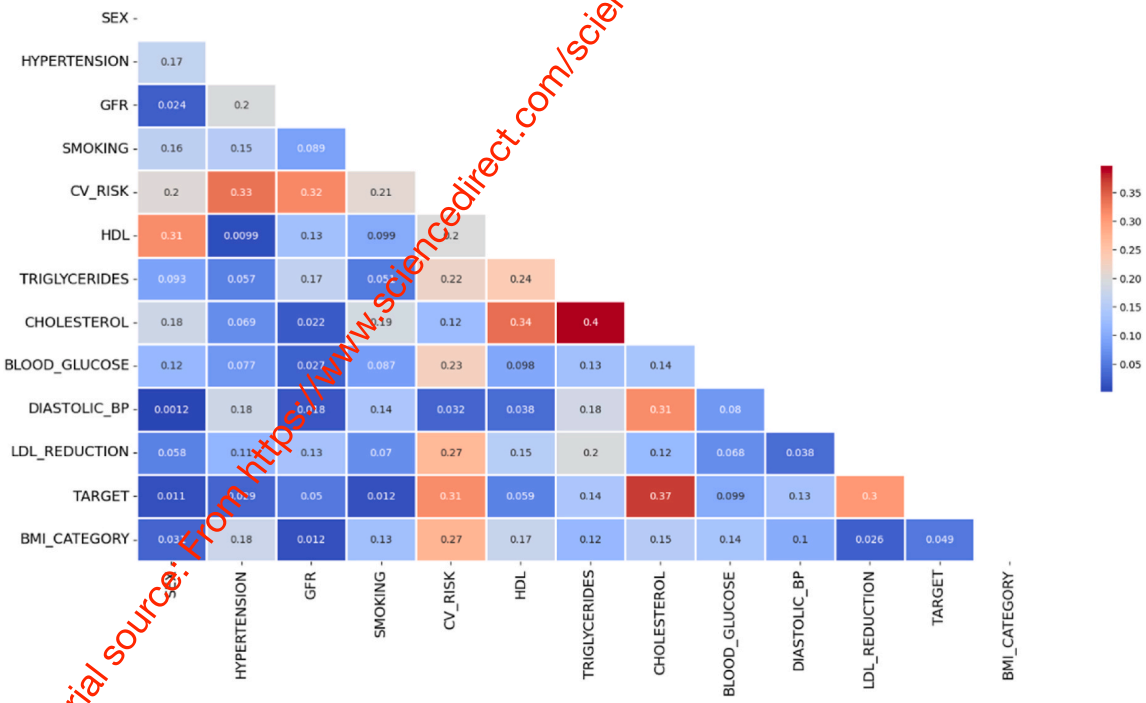


**Fig. 5.** Variables with a correlation greater than 0.4 with the target variable.



### 3.3.3. *Drug categorization*

Given the suboptimal performance of machine learning models in predicting the optimal individual therapy for each patient, along with the overlapping LDL reduction outcomes among different treatments (Fig. 1) and the standard classification of therapies based on cardiovascular risk, a new categorization approach was adopted. To improve predictive accuracy and model interpretability, lipid-lowering therapies were grouped into five distinct categories, where statistically significant differences in mean LDL reduction were observed (Fig. 6).

Category 1 was assigned to lifestyle modifications, decreased LDL by 12.57 %. Category 2 included low-intensity statins in monotherapy (e.g., Lovastatin, Simvastatin); due to limitations in patient registry data, only one instance of single-drug Rosuvastatin 10 mg was recorded, achieving an LDL reduction of 30.47 %. Category 3 encompassed moderate-

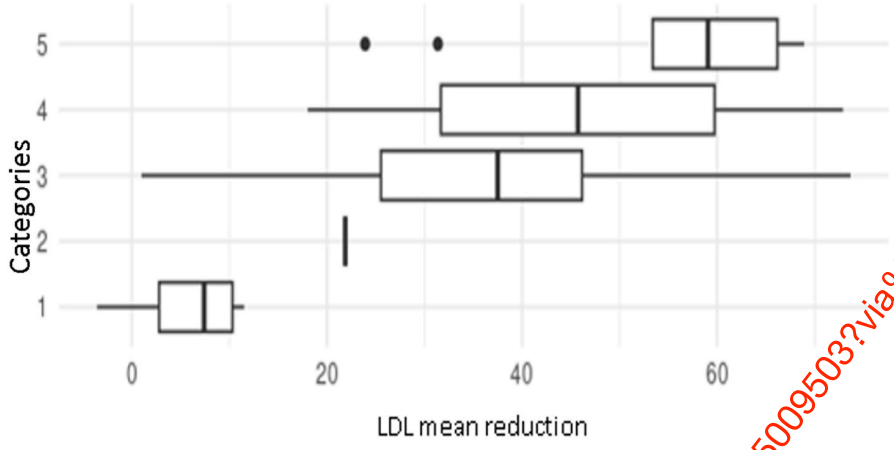


**Fig. 6.** Drug categorization.

intensity statins, including: Atorvastatin (20 mg, 40 mg), Rosuvastatin (20 mg, 40 mg), Simvastatin and Atorvastatin/Ezetimibe (40/10 mg). Category 4 included combination therapies with ezetimibe, which achieved a mean LDL reduction exceeding 50 %, including: Atorvastatin/Ezetimibe (10/10 mg), Rosuvastatin/Ezetimibe (20/10 mg, 40/10 mg). Category 5 was assigned to combination therapies incorporating PCSK9 inhibitors (iPCSK9), which achieved a mean LDL reduction greater than 70 %. These included: Statin + iPCSK9, Atorvastatin 40 mg + iPCSK9, Rosuvastatin 40 mg + iPCSK9 and Rosuvastatin/Ezetimibe 40/10 mg + iPCSK9. These therapy categories were established to better differentiate LDL-lowering effectiveness across treatment options (Table 9).

Given the statistically significant difference ($p < 0.05$) between the LDL reductions observed in clinical trials and those reported in real-world patient data; therefore, the decision was made to train the machine learning models exclusively on the studied population. Specifically, the research sample focused on patients who reached their LDL cholesterol targets according to ESC/EAS 2019 cardiovascular risk guidance. Amongst the 162 patients under evaluation, 110 patients (67.9 %) reached their individually specified LDL goal (Table 3) to become the foundation for modeling.

Additionally, after removing highly collinear variables and eliminating features with minimal clinical relevance in the selection of lipid-lowering therapy, the following variables were included: **LDL reduction, Age, Cardiovascular risk, Stages of chronic kidney disease and Presence of hypertension.** These variables were selected based on their established clinical significance in determining optimal lipid-lowering treatment strategies.

**Table 9**
Mean reduction in LDL levels observed in the treatment groups.

| Categories | Treatment | Mean reduction LDL | #Patients |
|---|---|---|---|
| 1 | Lifestyle changes | 12.57 % | 4 |
| 2 | Rosuvastatin 10 mg | 30.47 % | 1 |
| 3 | Atorvastatin 20 mg<br>Atorvastatin 40 mg Simvastatin/Ezetimibe 40/10 mg<br>Atorvastatin/Ezetimibe 40/10 mg<br>Rosuvastatin 20 mg<br>Rosuvastatin 40 mg | 46.78 % | 113 |
| 4 | Atorvastatin/Ezetimibe 10/10 mg<br>Rosuvastatin/Ezetimibe 20/10 mg<br>Rosuvastatin/Ezetimibe 40/10 mg | 55.89 % | 35 |
| 5 | iPCSK9 + Statins<br>Atorvastatin/iPCSK9 40 mg<br>Rosuvastatin/iPCSK9 40 mg<br>Rosuvastatin/Ezetimibe/iPCSK9 40/10 mg | 73.95 % | 9 |

Finally, to enhance the interpretability of the machine learning models, SHapley Additive exPlanations (SHAP) analysis was conducted using a Random Forest classifier (Fig. 7). The SHAP interaction analysis provided key insights into how specific clinical variables influence the model's decision-making process for optimizing lipid-lowering therapy. A strong positive interaction was observed between LDL reduction and cardiovascular risk, indicating that patients with greater LDL reductions and higher cardiovascular risk were more likely to receive intensive therapy recommendations. Additionally, the interaction between chronic kidney disease (CKD) stage and age revealed that younger patients with advanced CKD tend to be prioritized for more aggressive treatment strategies. Furthermore, the analysis demonstrated that hypertension amplifies the impact of cardiovascular risk on therapy selection, suggesting that hypertensive patients with elevated cardiovascular risk are more likely to require intensive lipid-lowering interventions.

### 3.3.4. *Application of machine learning models to drug categorization*

During model training, various hyperparameter configurations were explored to identify the optimal settings that maximized predictive performance and accuracy. The results revealed specific tuned configurations that enhanced each model's effectiveness. GBC achieved optimal performance with a learning rate of 0.075, a maximum depth of 10, and 200 estimators, leveraging sequential reinforcement to enhance predictive capacity. ABC performed best with a learning rate of 0.5 and 200 estimators, improving generalization capability. RFC benefited from 100 estimators with no restriction on maximum tree depth, enabling it to capture complex variable relationships effectively. ETC demonstrated robustness with a maximum depth of 20 and 400 estimators, providing high precision in predicting therapeutic responses. SVM was configured with a penalty parameter (C) of 10.0 and a polynomial kernel with a moderate gamma scale, proving effective in high-dimensional classification tasks. KNN performed optimally with 10 neighbors and a distance-based weighting scheme, ensuring a context-sensitive prediction strategy. RLC excelled with a regularization parameter (C) of 1.0 using the 'lbfgs' solver, ensuring optimal model convergence. NBC lacked hyperparameters for optimization, limiting the ability to improve its performance. DTC was employed as the base model for AdaBoost, with no restriction on maximum tree depth, achieving the best results for training this ensemble model. These refined hyperparameter configurations allowed for the optimization of predictive accuracy and model robustness across the different machine learning approaches.

To assess model performance, 5-fold cross-validation ($k = 5$) was applied to the entire dataset due to its limited sample size. While this approach maximized data utilization for both training and validation, it

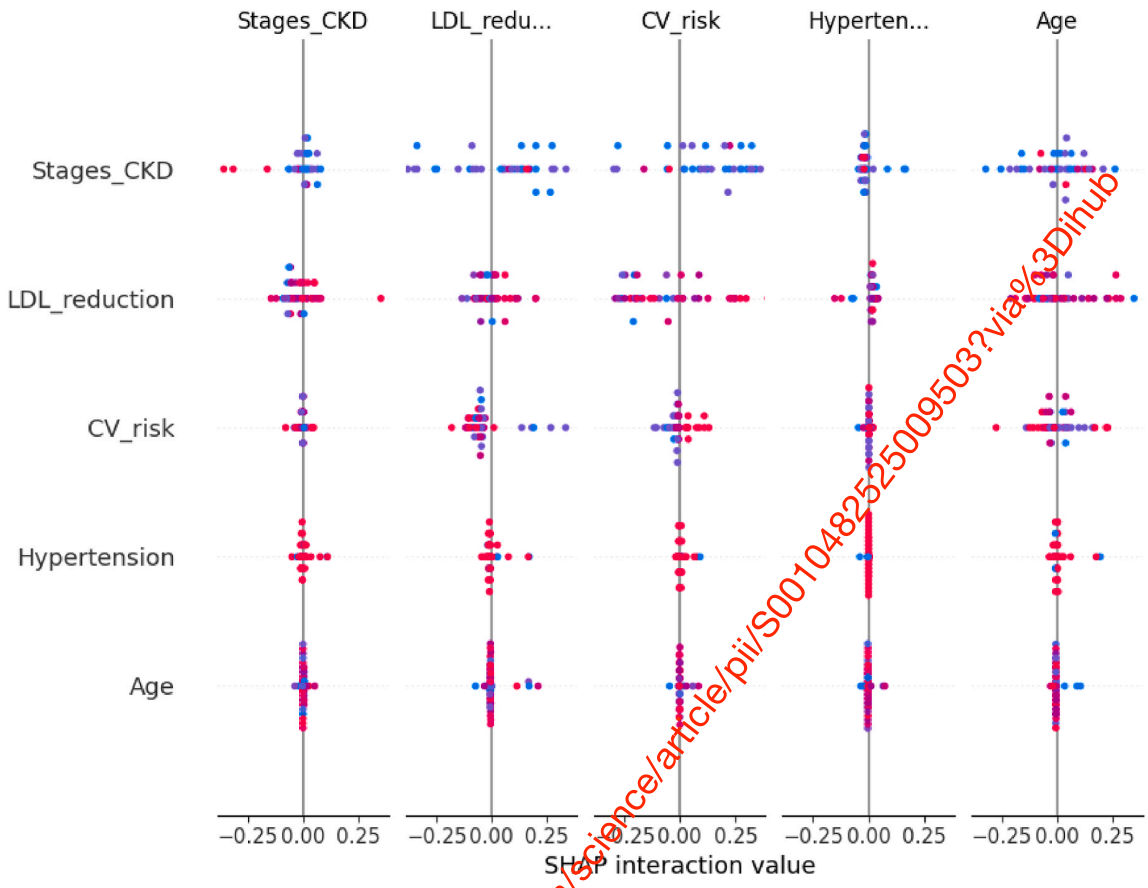


**Fig. 7.** SHAP interaction value powered by RFC.



is acknowledged that the inclusion of an independent test set would have provided a more robust evaluation of model generalization.

Therefore, the performance of various machine learning models was evaluated based on accuracy, precision, and F1-score, with standard deviations calculated from repeated experiments (Table 10). Among the models, RFC demonstrated superior performance, particularly due to its effective integration of cardiovascular risk variables into treatment selection. RFC consistently achieved the highest accuracy (0.9391 ± 0.029) and F1-score (0.9395 ± 0.0286), highlighting its robustness and reliability in predicting optimal LDL-lowering therapies. However, its performance remained comparable to models such as GBC and ETC when considering overall metrics. In contrast, DTC exhibited greater variability in accuracy (0.8906 ± 0.1003), suggesting lower stability compared to ensemble-based methods.

Among the non-ensemble methods, SVM demonstrated high precision (0.95 ± 0.0188) but lower accuracy (0.801 ± 0.056), suggesting potential overfitting or sensitivity to data imbalance. Similarly, KNN achieved moderate performance (accuracy: 0.8232 ± 0.0874) but exhibited high variability, indicating potential instability in classification performance across different datasets.

MLR and ABC demonstrated moderate performance, with MLR maintaining consistent accuracy (0.7836 ± 0.0176) but achieving lower F1-scores. In contrast, NBC exhibited significantly lower performance (accuracy: 0.5691 ± 0.1525), reflecting its limitations in handling complex, non-linearly separable data.

As shown in Fig. 8, the updated confusion matrices provide detailed insights into the classification performance of each machine learning model. Notably, RFC, GBC, ETC, and KNN demonstrated consistently high predictive accuracy across most treatment categories, with minimal misclassifications, particularly centered in class 3, which contained the largest proportion of correctly classified cases. Despite the overall strong performance, slight misclassification patterns were observed at the boundaries between adjacent classes, reflecting subtle overlaps in clinical features.

In contrast, models such as SVM, Decision Trees, and Logistic Regression maintained moderate performance, with some dispersion of predictions across neighboring categories, indicating reduced discrimination power in borderline cases. Most strikingly, Naïve Bayes exhibited pronounced misclassification errors, especially in class 4, where the number of correct predictions was notably lower compared to other models. This observation aligns with its overall poorer performance metrics and underscores its limitations in handling the multi-class nature of the dataset and the clinical complexity of the task.

Conversely, ANOVA analysis demonstrated that the examined models produced different results statistically significant at $p < 0.001$. The Naïve Bayes model achieved lower performance than other classifiers that demonstrated its restrictions in processing cardiovascular risk

**Table 10**
Machine learning models applied to drug categorization.

| Metrics | Performance | | |
|---|---|---|---|
| Models | Accuracy (SD) | Precision (SD) | F1-score (SD) |
| RFC | **0.9391 (0.029)** | 0.9438 (0.0263) | **0.9395 (0.0286)** |
| GBC | 0.9282 (0.0324) | 0.9446 (0.0338) | 0.9326 (0.0328) |
| ABC | 0.8101 (0.0525) | 0.8513 (0.0316) | 0.8275 (0.0436) |
| ETC | 0.9251 (0.0364) | 0.9377 (0.0234) | 0.9299 (0.0303) |
| DTC | 0.8906 (0.1003) | 0.9462 (0.0246) | 0.9064 (0.0652) |
| SVM Classifier | 0.801 (0.056) | **0.95 (0.0188)** | 0.865 (0.0245) |
| KNN Classifier | 0.8232 (0.0874) | 0.896 (0.0485) | 0.8526 (0.0705) |
| MLR | 0.7836 (0.0176) | 0.8907 (0.0189) | 0.8258 (0.0067) |
| Naive Bayes | 0.5691 (0.1525) | 0.7343 (0.0948) | 0.5556 (0.1619) |

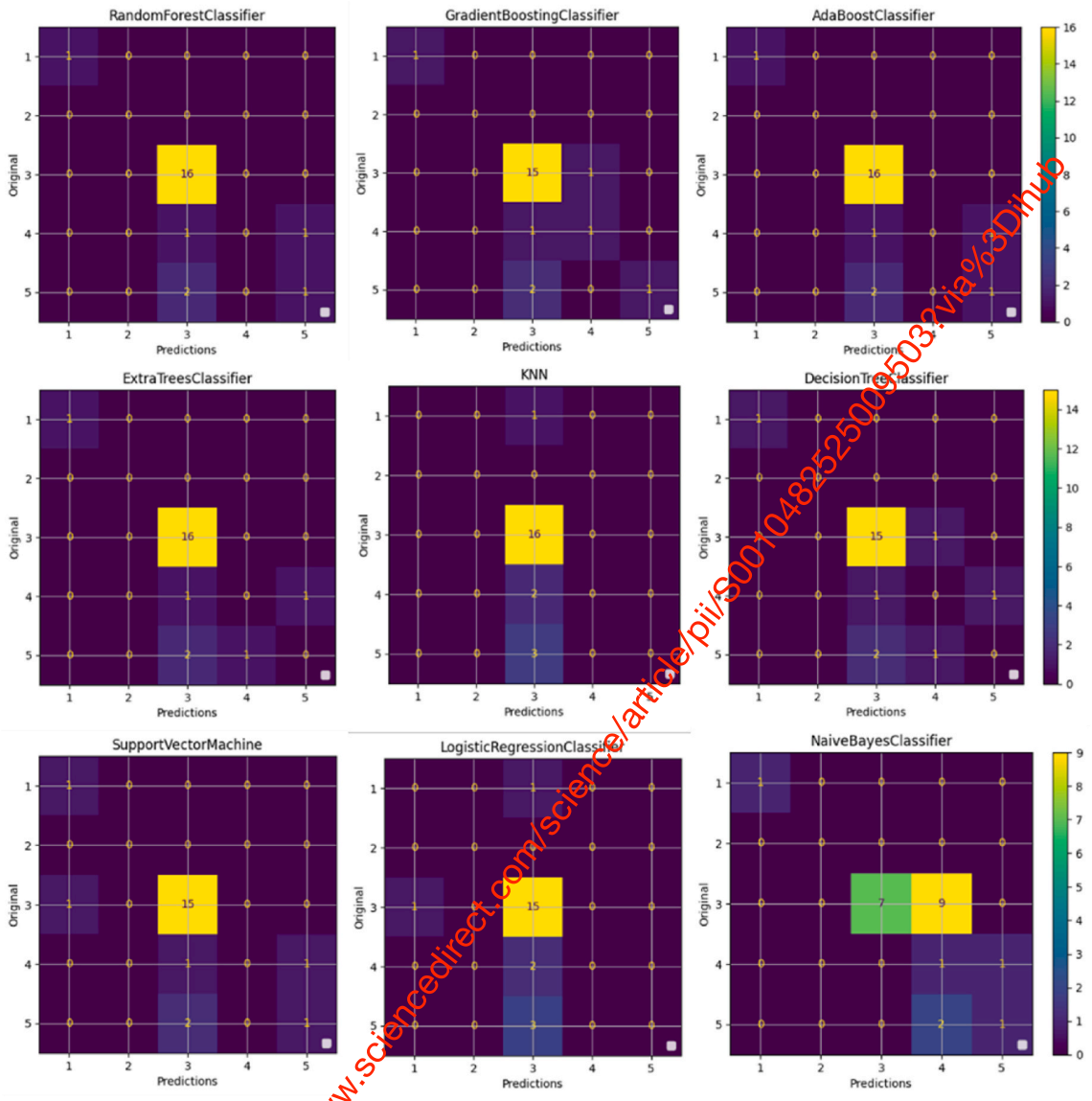


**Fig. 8.** Confusion matrices of machine learning models.

factor complexities. RFC demonstrates superior effectiveness in predicting optimal lipid-lowering therapies due to its ability to handle cardiovascular risk-based LDL cholesterol control yet other models fail to achieve similar results.

## 4. Discussion

AI implementation in medicine field has shown major growth which has resulted in more studies focused on making clinical decisions for individual treatment selection. Researchers Song and Park [22] created "RCC-Supporter" which stands as an AI-powered clinical decision-support system to assist medical practitioners with renal cell carcinoma (RCC) treatment selection decisions. Applying Gradient Boost algorithms in 4 real-world data from 1867 RCC patients and, the system achieved an accuracy of 95 % in predicting the most suitable treatment options. These findings underscore the growing potential of AI in personalized medical care, enhancing clinical decision-making and ultimately improving patient outcomes. Similarly, another study focused on prostate cancer treatment leveraged AI to refine clinical decision-making [23]. The examination employed explainable machine learning methods on data collected from 255,000 localized prostate cancer patients in the SEER Prostate Watchful Waiting database. The model used multiple variables based on oncologic data along with demographic data and socioeconomic information and geographical variables to forecast among three therapeutic options including: Active surveillance/watchful waiting (AS/WW), radical prostatectomy (RP), and radiation therapy (RT). The predictive model displayed reliable performance through its achieved multiclass Area Under Curve (AUC) score of 0.77. Inverse correlations existed between clinical prostate cancer characteristics and patients chose RP as their treatment option, but geographic variables proved important in the selection of RT. The research indicates that AI technology is escalating its impact on oncology which enables improved personalized therapies for cancer patients.

This study reveals different elements that influence the response to lipid-lowering treatments when compared to the findings of a 2021 Spanish research [24]. That research established that inadequate treatment response mainly depended on patient age combined with sex and measurement LDL cholesterol levels; indicating, young males with elevated LDL cholesterol levels failed to meet their target LDL goals. The

primary elements affecting treatment response according to our research included initial LDL and desired LDL reductions combined with cardiovascular disease status and overall cardiovascular risk levels. This discrepancy likely stems from differences in study populations and treatment methodologies. The analysis in the Spanish research drew data from the Dyslipidemia Registry of the Spanish Society of Atherosclerosis that established therapy based on clinical needs without establishing standardized procedures. In contrast, the present research employed LDL control data which adjusted treatments based on cardiovascular risk standards following clinical guidelines; thus, superior control to that observed in other studies is achieved [25].

Experiencing a cardiovascular event or having a high cardiovascular risk of developing one has been identified as a statistically significant factor contributing to poor response to lipid-lowering treatment. This finding suggests a potential innate resistance to therapy in these patients. Although no conclusive studies have yet confirmed this hypothesis, it represents a promising avenue for future research, warranting further investigation into the underlying mechanisms influencing treatment resistance.

The major discrepancy between authentic treatment outcomes and the data presented in clinical trials matches what scientific research demonstrates [26]. Clinical trials implement controlled environments to maintain strict patient protocol adherence which maximizes therapeutic outcomes. Real world studies face many external factors such as patient treatment non-compliance and medical conditions and diverse clinical practices which decrease medication effectiveness. A European patient registry showed results that differed from those obtained through clinical trials testing different statins for lowering LDL levels. The research findings showed substantial differences between clinical trial and real-world groups thus demonstrating trial-based lipid lowering effects cannot be directly used for real-world patient management [27].

The integration of machine learning as an artificial intelligence tool offers significant potential for enhancing clinical outcomes in patients with high cardiovascular risk. Algorithms such as Random Forest and Gradient Boost enable the analysis of large datasets, allowing for the identification of complex patterns that may not be readily detectable through traditional methods. By leveraging these capabilities, machine learning models provide valuable insights into patient-specific clinical behaviors, facilitating the optimization of treatment strategies based on individual patient characteristics.

Ensemble methods proved their expertise, particularly RFC (F1-score: 0.8247 ± 0.0397) and GBC (F1-score: 0.8277 ± 0.0495), consistently outperformed other classifiers, confirming their robustness in predicting optimal treatment strategies. Conversely, models such as NBC (F1-score: 0.3395 ± 0.0363) and MLR (F1-score: 0.5105 ± 0.0139) exhibited limited generalization capabilities, likely due to the complexity of multi-class classification in real-world clinical data. These findings reinforce the importance of evaluating multiple performance metrics beyond accuracy, particularly when working with datasets containing underrepresented treatment categories.

Despite the limited dataset size (162 instances), classifier complexity was carefully considered to mitigate the risk of overfitting. A combination of simple models and more complex ensemble methods was evaluated to assess performance across different levels of complexity. Additionally, 5-fold cross-validation was applied to enhance model generalization and reduce variance. In this research, the ensemble methods utilized optimized hyperparameters to regulate their complexity levels and reduce exorbitant overfitting conditions. The research outcomes demonstrate that properly adjusted ensemble methods generate accurate predictions although the dataset is smaller because they maintain their clarity for clinical decision-based interpretations.

The application of machine learning models reaches above basic treatment guidelines through integration of various factors which affect response including medical conditions of patients and their genetic profile and their behaviors and other relevant clinical aspects. Random Forest identifies patient subgroups who experience low success rates with statins for reaching lipid control targets. The acquired data allows clinicians to foresee and modify treatment strategies or employ combination therapy approaches thus decreasing the risk of cardiovascular issues appearing.

Through machine learning integration into the clinical practices healthcare in a real world, providers create specific and highly effective medical strategies for treating patients with high cardiovascular risk. The implementation of this strategy both enhances patient recovery and benefits healthcare institutions to reduce unnecessary treatments and preventing avoidable complications.

This study establishes crucial importance because it adopts pioneering approaches to use machine learning models that optimize LDL-lowering therapy according to varied individual risk variables. Traditional lipid-lowering approaches involving dietary and lifestyle modifications, statins, ezetimibe and iPCSK-9 have been studied by previous research but these studies did not completely use machine learning to personalize the treatment.

## 5. Conclusions

Machine learning models is a powerful tool to be applied in clinical practice, with the potential to enhance patient care processes by enabling personalized treatment strategies based on individual risk factors.

Clinical trials deliver different LDL reductions than real world studies because trial conditions maximize effectiveness while patients in a real-world experience hurdle buying their prescribed medicines.

Patients fail to reach LDL targets for various causes when they maintain insufficient LDL levels or begin with elevated LDL values in the presence of continuing cardiovascular disease and global vascular risk factors.

The application of machine learning models serves clinical practice well because they enable customized treatment plans through risk-based evaluation of individual patients.

Future studies should include their patient large number scope across various medication treatments to study machine learning model effectiveness for individualized optimal care planning.

## CRediT authorship contribution statement

**Deiby Boneu Yepez:** Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **David Sierra Porta:** Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Data curation. **Liz Morales Aguas:** Writing – review & editing, Supervision. **Fernando Manzur Jattin:** Writing – review & editing, Supervision.

## Ethical Statement for Solid State Ionics

Hereby, I Deiby Boneu Yepez consciously assure that for the manuscript Optimizing treatment to control LDL cholesterol using artificial intelligence models the following is fulfilled:

1) This material is the authors' own original work, which has not been previously published elsewhere.
2) The paper is not currently being considered for publication elsewhere.
3) The paper reflects the authors' own research and analysis in a truthful and complete manner.
4) The paper properly credits the meaningful contributions of co-authors and co-researchers.
5) The results are appropriately placed in the context of prior and existing research.

6) All sources used are properly disclosed (correct citation). Literally copying of text must be indicated as such by using quotation marks and giving proper reference.
7) All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

The violation of the Ethical Statement rules may result in severe consequences.

To verify originality, your article may be checked by the originality detection software iThenticate. See also http://www.elsevier.com/editors/plagdetect.

I agree with the above statements and declare that this submission follows the policies of Solid State Ionics as outlined in the Guide for Authors and in the Ethical Statement.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Deiby Boneu Yepez has patent Cardiolipid IA pending to 1-2024-101098. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledge

Special Acknowledge to Camila Boneu for help in the translation.

## References

[1] N. Townsend, L. Wilson, P. Bhatnagar, K. Wickramasinghe, M. Rayner, M. Nichols, Cardiovascular disease in Europe: epidemiological update 2016, Eur. Heart J. 37 (42) (2016) 3232–3245, https://doi.org/10.1093/eurheartj/ehw334.
[2] WHO, World health organization (WHO)-The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, 2021.
[3] S. Yusuf, S. Hawken, S. Ôunpuu, T. Dans, A. Avezum, F. Lanas, L. Lisheng, Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study, Lancet 364 (9438) (2004) 937–952, https://doi.org/10.1016/S0140-6736(04)17018-9.
[4] B.A. Ference, H.N. Ginsberg, I. Graham, K.K. Ray, C.J. Packard, E. Bruckert, A. L. Catapano, Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the european atherosclerosis society consensus panel, Eur. Heart J. 38 (32) (2017) 2459–2472, https://doi.org/10.1093/eurheartj/ehx144.
[5] H. Tada, M.A. Kawashiri, M. Yamagishi, Comprehensive genotyping in dyslipidemia: mendelian dyslipidemias caused by rare variants and Mendelian randomization studies using common variants, J. Hum. Genet. 62 (4) (2017) 453–458, https://doi.org/10.1038/jhg.2016.159.
[6] E. Galve, A. Cordero, A. Cequier, E. Ruiz, J.R. González-Juanatey, Grado de control lipídico en pacientes coronarios y medidas adoptadas por los médicos. Estudio REPAR, Rev. Española Cardiol. 69 (10) (2016) 931–938, https://doi.org/10.1016/j.recesp.2016.02.013.
[7] E.G. Díaz, D.R. Medina, Ó.M.M. Porras, J.L.C. Mateos, Determinants of inertia with lipid-lowering treatment in patients with type 2 diabetes mellitus, Endocrinología, Diabetes y Nutrición (English ed.) 66 (4) (2019) 223–231, https://doi.org/10.1016/j.endinu.2018.08.014.
[8] F. López-Simarro, I. Moral, A. Aguado-Jodar, C. Cols-Sagarra, J. Mancera-Romero, M. Alonso-Fernández, C. Brotons, The impact of therapeutic inertia and the degree of medication adherence on the control goals for patients with diabetes, Semergen 44 (8) (2017) 579–585, https://doi.org/10.1016/j.semerg.2017.10.002.
[9] F.Y. Man, C.X. Chen, Y.C. Lau, C. King, Therapeutic inertia in the management of hyperlipidaemia in type 2 diabetic patients: a cross-sectional study in the primary care setting, Hong Kong Med. J. 22 (4) (2016) 356, https://doi.org/10.12809/hkmj154667.
[10] J. Pedro-Botet, X. Pintó, Colesterol LDL, cuanto más bajo mejor, Clín. Invest. Arterioscler. 31 (2019) 16–27, https://doi.org/10.1016/j.arteri.2019.10.003.
[11] F. Mach, C. Baigent, A.L. Catapano, K.C. Koskinas, M. Casula, L. Badimon, O. Wiklund, 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), Eur. Heart J. 41 (1) (2020) 111–188, https://doi.org/10.1093/eurheartj/ehz455.
[12] F.J. Raal, E.A. Stein, R. Dufour, T. Turner, F. Civeira, L. Burgess, D. Gaudet, PCSK9 inhibition with evolocumab (AMG 145) in heterozygous familial hypercholesterolaemia (RUTHERFORD-2): a randomised, double-blind, placebo-controlled trial, Lancet 385 (9965) (2015) 331–340, https://doi.org/10.1016/S0140-6736(14)61399-4.
[13] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010950718922.
[14] J.L. Speiser, M.E. Miller, J. Tooze, E. Ip, A comparison of random forest variable selection methods for classification prediction modeling, Expert Syst. Appl. 134 (2019) 93–101, https://doi.org/10.1016/j.eswa.2019.05.028.
[15] Y. Zhang, M. Ni, C. Zhang, S. Liang, S. Fang, R. Li, Z. Tan, Research and application of AdaBoost algorithm based on SVM, in: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), IEEE, 2019, May, pp. 662–666, https://doi.org/10.1109/ITAIC.2019.8785556.
[16] I.H. Sarker, Machine learning: Algorithms, real-world applications and research directions, SN Computer Sci. 2 (3) (2021) 160, https://doi.org/10.1007/s42979-021-00592-x.
[17] A.S. Parvin, B. Saleena, An ensemble classifier model to predict credit scoring-comparative analysis, in: 2020 IEEE International Symposium on Smart Electronic Systems (iSES)(Formerly iNiS), IEEE, 2020, December, pp. 27–30, https://doi.org/10.1109/iSES50453.2020.00017, 10.1109/iSES50453.2020.00017.
[18] B. Charbuty, A. Abdulazeez, Classification based on decision tree algorithm for machine learning, J. Appl. Sci. Technol. Trends 2 (1) (2021) 20–28, https://doi.org/10.38094/jastt20165.
[19] P. Thanh Noi, M. Kappas, Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery, Sensors 18 (1) (2017) 18, https://doi.org/10.3390/s18010018.
[20] F. Abramovich, V. Grinshtein, T. Levy, Multiclass classification by sparse multinomial logistic regression, IEEE Trans. Inf. Theor. 67 (7) (2021) 4637–4646, https://doi.org/10.48550/arXiv.2003.01951.
[21] D. Prabha, J. Aswini, B. Maheswari, R.S. Subramanian, R. Nithyanandhan, P. Girija, A survey on alleviating the naive bayes conditional Independence assumption, in: 2022 International Conference on Augmented Intelligence and Sustainable Systems, (ICAISS), Trichy, India, 2022, pp. 654–657, https://doi.org/10.1109/ICAISS55157.2022.10011103.
[22] W.H. Song, M. Park, RCC-Supporter: supporting renal cell carcinoma treatment decision-making using machine learning, BMC Med. Inf. Decis. Making 24 (Suppl 2) (2024) 259, https://doi.org/10.1186/s12911-024-02660-7.
[23] J.H. Han, S. Lee, B. Lee, O.K. Baek, S.L. Washington 3rd, A. Herlemann, P. E. Lonergan, P.R. Carroll, C.W. Jeong, M.R. Cooperberg, Explainable ML models for a deeper insight on treatment decision for localized prostate cancer, Sci. Rep. 13 (1) (2023) 11532, https://doi.org/10.1038/s41598-023-38162-1.
[24] E. Climent, A.M. Bea, D. Benaiges, Á. Brea-Hernando, X. Pintó, M. Suárez-Tembra, Dyslipidaemia Registry of the Spanish Atherosclerosis Society, LDL cholesterol reduction variability with different types and doses of statins in monotherapy or combined with ezetimibe. Results from the Spanish arteriosclerosis society dyslipidaemia registry, Cardiovasc. Drugs Ther. (2022) 1–8, https://doi.org/10.1007/s10557-020-07137-z.
[25] T. Liu, D. Zhao, Y. Qi, Global Trends in the Epidemiology and Management of Dyslipidemia, 2022, https://doi.org/10.3390/jcm11216377.
[26] A. Attar, Response to statin therapy in the real world, Eur. J. Prev. Cardiol. 28 (14) (2021) e25–e26, https://doi.org/10.1177/2047487320905718.
[27] D.D. Bacquer, D.D. Smedt, Ž. Reiner, L. Tokgözoğlu, E. Clays, K. Kotseva, G. D. Backer, Percentage low-density lipoprotein-cholesterol response to a given statin dose is not fixed across the pre-treatment range: real world evidence from clinical practice: data from the ESC-EORP EUROASPIRE V Study, Eur. J. Prev. Cardiol. 27 (15) (2020) 1630–1636, https://doi.org/10.1177/2047487319874898.

**Deiby Boneu Yepez**: Physician from the Universidad de Cartagena, expert in lipids Universidad del Bosque, Master's degree in applied statistics and data science Universidad Tecnológica de Bolívar. Clinical trials subinvestigator of Centro de Diagnóstico Cardiológico, Professor of programa de medicina Corporacion Universitaria Rafael Núñez, Cartagena, Bolivar, Colombia. Co-founder of TimeMed-IA.

**David Sierra Porta:** Graduate in Mathematics and Physics from the University of Zulia, Master Scientiarum in Fundamental Physics and PhD in Fundamental Physics from the University of the Andes (Venezuela), Professor of Ciencias básicas, Maestría en estadistica aplicada y ciencia de datos, Universidad Tecnológica de Bolívar. Researcher in the area of muography and high energy particles and Universidad de los Andes as a researcher in Computational Astrophysics.

**Liz Morales Aguas**: Bacteriologist from the Universidad San Buenaventura, specialist's degree in epidemiology Universidad Juan N de Corpas, Clinical writer. Co-founder of TimeMed-IA.

**Fernando Manzur Jattin:** Physician from the Universidad de Cartagena, Cardiologist Universidad Complutense Madrid España, master's degree in clinical trials in human beings at the Universidad de Sevilla, Principal investigator of centro de diagnóstico cardiológico, Professor of Universidad de Cartagena.