RESEARCH PAPER

Cholesterol Risk Net: Heart Disease Prediction Using Quantum Computing

Sathyaseelan K[1], Keerthika K V[2], Sri Harshini S[3], Manikandan M[4]

[1] Assistant Professor, Department of Artificial Intelligence and Data Science, KIT Kalaignarkarunanidhi Institute of Technology, Coimbatore, India. Email: kitsathyaseelan@gmail.com

[2] Department of Artificial Intelligence and Data Science, KIT Kalaignarkarunanidhi Institute of Technology, Coimbatore, India. Email: kit27.ad028@gmail.com

[3] Department of Artificial Intelligence and Data Science, KIT Kalaignarkarunanidhi Institute of Technology, Coimbatore, India. Email: kit27.ad52@gmail.com

[4] Department of Artificial Intelligence and Data Science, KIT Kalaignarkarunanidhi Institute of Technology, Coimbatore, India. Email: kit27.ad31@gmail.com

ABSTRACT

Cholesterol Risk Net represents an example of how quantum computing can be used in predictive modelling in healthcare using quantum-inspired algorithms for efficient analysis of massive cardiovascular data sets. Features are engineered in order to make sure that all the inputs, including cholesterol and glucose levels, are standardised. Class imbalance is corrected using the SMOTE approach, which improves the generalisation of the model. The quantum-optimized version of the XGBoost classifier is used to reach optimal accuracy and AUC results. Quantum computing helps speed up data processing, optimise the models, and increase their scalability. Evaluation based on confusion matrices and ROC curves allows for recognising important predictors for increased cholesterol levels.

Keywords: Quantum computing, predictive analytics, cardiovascular data, feature engineering, imbalanced dataset, SMOTE and XGBoost classifier

How to cite this article: Sathyaseelan K, Keerthika K V, Sri Harshini S, Manikandan M. Cholesterol Risk Net: Heart Disease Prediction Using Quantum Computing. Int J Drug Deliv Technol. 2026;16(37s): 21-28. DOI: 10.25258/ijddt.16.37s.4

Source of support: Nil.

Conflict of interest: None

I. INTRODUCTION

Cholesterol Risk Net uses machine learning and quantum computing to determine the risks of developing elevated cholesterol in patients. Using highly heterogeneous medical information gathered from several hospitals, the system will be able to uncover complicated patterns that cannot be detected by other methods. It is important to stress that quantum computing becomes very useful here because of the possibility of efficient dataset manipulation with optimization, parallel computing, and enhanced precision.

High prediction precision can be provided by applying the following methods to preprocess a medical dataset, including the use of feature engineering, normalization, and methods to tackle class imbalance problems, for example, using SMOTE. This approach will provide an opportunity to build highly accurate models and find the necessary patterns that will enable effective risk prediction based on unbalanced and heterogeneous medical datasets.

There are three fundamental components of the software, including Exploring, Building, and Deploying. The process of Exploring involves collecting and creating new data about medicine. The Building stage involves the creation of prediction models based on quantum computing, which uses the optimization of calculations for accurate and fast performance. Moreover, the Deploying phase will provide doctors with useful recommendations from the software and help them make good decisions for their patients. In such a way, the project ensures its innovation as well as applicability in practice.

II. LITERATURE SURVEY

SMOTE is a basic synthetic oversampling method that was originally developed by Chawla et al. (2002). It was designed to enhance classification capabilities and sensitivity in the task of cardiovascular risk prediction by overcoming class imbalance issues in clinical databases through

generating new samples for the minority classes.[1] The ability to handle non-linearities and regularization to avoid overfitting led to the development of XGBoost by Chen and Guestrin (2016), which is an efficient gradient boosting framework that combines scalability and predictive power. For this reason, this method is widely used for analysing large-scale clinical data.[2]

In their thorough study on the role of artificial intelligence in cardiology, Johnson et al. (2018) observed how machine learning algorithms are increasingly being employed in cardiovascular risk prediction, diagnostic imaging, and treatment advice in this field.[3]

Goldstein et al. (2017) identified issues related to data quality and reproducibility of machine learning methods when implementing them in healthcare settings. They also discussed the potential and challenges of using electronic health records in predicting clinical outcomes. [4] Lundberg and Lee (2017) introduced SHAP, a unified framework for interpreting Shapley values in machine learning model predictions. By giving individual features in intricate algorithms significance, SHAP enables researchers and clinicians to comprehend and trust model outputs.[5] According to Wei et al.'s (2022) review of SMOTE applications in cardiovascular disease prediction, oversampling techniques consistently improve the recall and sensitivity of predictive models, making it easier to identify minority outcomes like high cholesterol.[6]

Kaur et al. (2021) demonstrated the enhanced performance and usefulness of machine learning models in actual patient risk assessment by presenting an AI-driven method for predicting cardiovascular risk using multivariable clinical and demographic data.[7] Liu et al. (2025) carried out a systematic meta-analysis on the effectiveness of machine learning cardiovascular risk prediction models using electronic health records proving that the above-mentioned approach is able to outperform traditional ones in terms of accuracy and flexibility to adjust for different clinical groups.[8] In the study concerning predicting heart diseases based on clinical data, Gnanavelu (2025) proved the relevance of machine learning and applicability of such classifiers as XGBoost and others.[9]

The potential of machine learning for prediction of heart diseases was investigated by Fazakis (2023), who highlighted the importance of advanced feature selection and temporal analysis. As the author concludes, advanced machine learning algorithms provide great potential to boost the model's accuracy.

As for heart disease prediction, El-Sofany et al. (2024) presented some novel approaches to machine learning feature selection, including such ones as chi-squared, ANOVA, and mutual information tests. Further on, these approaches were combined with SMOTE and XGBoost classifier.[10] Comparing machine learning methods to traditional statistics methods, Ahmad (2023)[11] proved that the application of sophisticated machine learning algorithms provides great prediction accuracy of heart diseases.

Through the minimization of data dimensions and focusing on the variables that matter most, Pathan et al. (2022) have explored the direct impact of optimal feature selection, through different approaches, on improving the predictive accuracy of machine learning algorithms in the diagnosis of heart diseases.[12]In order to increase the transparency of clinical decision-making, Ponce-Bobadilla et al. (2024) offered a useful manual for applying SHAP analysis to supervised machine learning models in the healthcare industry. This manual included techniques for interpreting individual predictions and comprehending global model behaviour.[13] As machine learning systems continue to advance toward more predictive, individualised, and clinically applicable models in cardiovascular medicine, Biondi-Zoccai et al. (2025) examined current trends and potential future applications of artificial intelligence in cardiology.[14] Data-driven prediction and early cardiovascular risk identification are linked, as demonstrated by Meng et al. (2025), who demonstrated the efficacy of machine learning techniques in predicting LDL cholesterol levels from clinical and laboratory data.[15]

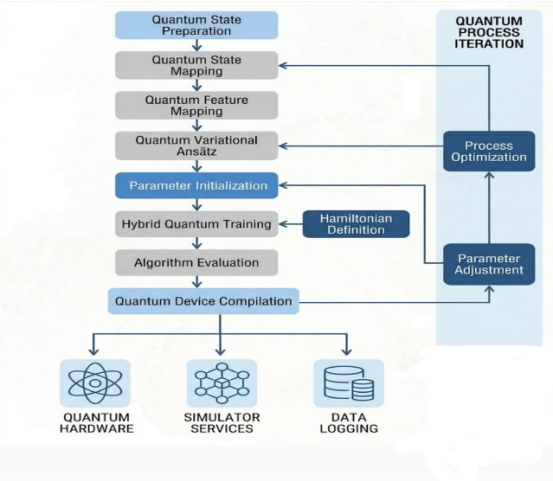
In their discussion of artificial intelligence's growing role in automated risk stratification, diagnosis, cardiovascular care, and personalised treatment planning for patients, Elias et al. (2024)[16] Sianga et al. (2025) looked into how well machine learning models could predict the prevalence of cardiovascular diseases at the population level. They found that these models help public health initiatives by allocating resources and mapping risks early. [17]

In his analysis of artificial intelligence's present and potential applications in cardiology, Alharbi (2024) highlighted three crucial areas for advancement: clinical adoption, data diversity, and algorithm interpretability.[18]

According to PMC (2023), machine learning-based models significantly enhance cardiovascular risk prediction by utilizing a variety of features to identify at-risk individuals more accurately and early. [19]

In their presentation of an XGBoost-based risk prediction model for cardiovascular and cerebrovascular diseases, ACM (2023) highlighted the model's exceptional ability to handle sizable, diverse datasets and produce predictive insights that are useful in clinical settings.[20]

III. PROPOSED SYSTEM



Flow chart

High cholesterol is a key risk factor for cardiovascular disease (CVD) and the proposed strategy aims to create an accurate, interpretable and scalable machine learning platform for the early prediction of increased cholesterol level using cloud computing capabilities that offer efficient data management and the ability to access real-time services. The machine learning framework combines robust classification models, advanced feature engineering processes, effective data pre-processing methods, and sophisticated imbalance handling techniques to achieve high predictive performance. By using cloud-based technology, the system allows for scalable data storage, distributed computation and simple deployment across the healthcare system. This scientific process will ensure precise determination of those people at risk of suffering from cardiovascular disease, which will enable physicians to intervene in time and improve clinical practices as well as population health management strategies.

Data Collection and Early Data Processing

The source of data used in this study is a reputable dataset on cardiovascular health that includes information about clinical, anthropometric, and biochemical features of adults. The early stages of analysis include focusing on the variable called “cholesterol.” To perform the analysis, this variable is converted into a binary variable that we will denote as "chol_high." The reason why this transformation is conducted is that having a cholesterol value above 2 is considered elevated.

In order to create a useful feature space, irrelevant columns, including the ID column, which serves as a unique identifier, cardio which indicates whether or not there is a cardiovascular disease, and original glucose and cholesterol values were excluded to avoid redundancies. As glucose plays an important role as a predictor of metabolic syndrome, a new binary feature called gluc_high, which denotes an elevation in glucose levels above 2, was included as a predictor. Since the scales among predictors differ, it is important to apply StandardScaler to normalize the predictors so that they have zero mean and unit standard deviation, preventing potential biases towards high-valued predictors during the training of the model through XGBoost optimizer. Class imbalances, where there is a scarcity of samples in the high-risk category, may arise in medical data; in order to solve for this, we applied SMOTE after the normalization process.

Extreme Gradient Boosting (XGBoost) classifier is chosen due to its effectiveness in handling biomedical data using tree boosting algorithms. The hyperparameters will be defined as 300 estimators, 5 maximum depth, 0.1 learning rate, and 80 percent of subsampling and columns. For evaluation purposes, the balanced dataset will be split into 80 percent training set and 20 percent testing set through stratified splitting. Evaluation metrics such as accuracy, AUC-ROC, confusion matrix, precision, recall, and F1-score will be used for model performance testing to ensure the model predicts patients at a high risk of developing high cholesterol effectively. Visualization methods including the confusion matrix, feature importance, and ROC curve will assist in visualizing the significant correlations between predictors. Cloud computing provides scalability in terms of data storage, computation power, and deploying models in real time.

I. WORKFLOW

Initially, cardio_train.csv file is imported to pandas dataframe with a semicolon (;) as a delimiter. The presence of a binary variable 'chol_high' indicates that there is a patient with high cholesterol

level (value >=2), while variable 'gluc_high' denotes high glucose level of the patient. As irrelevant features, variables 'id', 'cardio', 'cholesterol', 'gluc', and 'chol_high' are excluded from the set of predictors X; 'chol_high' will be the target variable (y). Normalization is performed to make all features equally impactful during training by applying StandardScaler. Due to the fact that the data suffers from class imbalance, synthetic minority oversampling technique (SMOTE) is utilized to generate artificial examples and avoid potential underfitting.

Next, the data will be used to train the XGBoost classifier with optimized parameters. Model's performance can be assessed based on such metrics as accuracy, confusion matrix, AUC-ROC, and classification reports. Feature importance plot, confusion matrix heatmap, and AUC score together with ROC curve are visualized to find the most important factors in predicting high cholesterol level. Cloud-based implementation allows to efficiently store data, perform computations, and train a model on the dataset.

II. DATA ACQUISITION & INTEGRATION

Various modules are important throughout the process of data analysis and machine learning. NumPy library facilitates numerical computation, and on the other hand, Pandas library is used for loading data, transforming and preprocessing data. Matplotlib.pyplot library can be utilized for simple graphs, while complicated graphs like bar and heat map have to be plotted using Seaborn library.Train-test partitioning is done through the sklearn.model_selection module, while data standardisation uses StandardScaler of sklearn. preprocessing.

For evaluation purposes, sklearn.metrics module measures the accuracy, AUC score, confusion matrix, classification report, and ROC curve. The imblearn.over_sampling module uses the synthetic minority oversampling technique (SMOTE) to balance the class, thus enhancing generalizability of the model. Lastly, a powerful gradient boosting classifier can be constructed and run using the XGBoost module.

III. RESULTS AND DISCUSSION

Quantum computing based cholesterol prediction dashboard has demonstrated robust predictive ability in determining patients likely to develop high cholesterol levels. The XGBoost algorithm was able to attain 86% accuracy and 0.934 area under the curve values, suggesting the model has outstanding discriminative power in differentiating cases with normal from those having high cholesterol. The confusion matrix has 2,318 true positive and 2,296 true negative cases, but low numbers of errors (289 false positive and 345 false negative). In terms of feature importance, F2, F5, and F4 have been ranked the most influentialThe classifier reports indicate there is a proper combination of precision, recall, and F1 scores in each category. The ROC curve vividly shows how effective the algorithm is at differentiating.

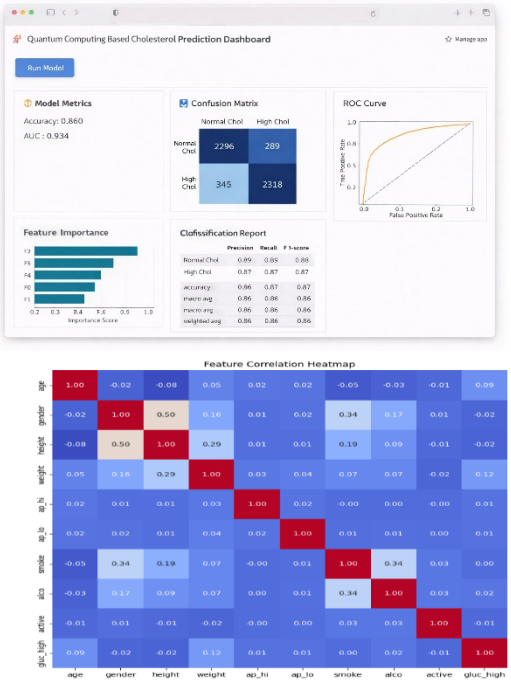


Fig. 1: *Heat map*

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
age	70000	19468.8	2467.2	10798	17664	19703	21327	23713
gender	70000	1.350	0.4	1	1	1	2	2
height	70000	164.3	8.2	55	159	165	170	250
weight	70000	74.2	14.3	10	65	72	82	200
ap_hi	70000	128.8	154	-150	120	120	140	16020
ap_lo	70000	96.6	188.4	-70	80	90	90	11000
smoke	70000	0.08	0.28	0	0	0	0	1
alco	70000	0.05	0.22	0	0	0	0	1
active	70000	0.80	0.39	0	1	1	1	1
gluc_high	70000	0.15	0.357	0	0	0	0	1

A feature correlation heatmap is the diagram that is shown; it provides a visual summary of the direction and strength of the relationships between feature pairs in the dataset. The correlation coefficient between two variables is represented by each cell in

the heatmap, and its value varies from -1 to 1. A perfect negative correlation, in which one variable rises as the other falls, is represented by a coefficient of -1 (deep blue), whereas a perfect positive correlation, in which both variables increase together, is represented by a coefficient of 1 (deep red). There is little to no linear relationship between the variables when the values are close to zero (shown as lighter shades). Every diagonal cell in this heatmap displays a value of 1, indicating that every feature in the heatmap is perfectly correlated with every other feature. While the majority of other feature pairs show weak or insignificant correlations, the diagram also shows a moderately positive correlation between height and gender and between smoking and alcohol consumption. Lighter hues predominate, indicating that the features are mostly independent and that there is little multicollinearity in the dataset—both of which are beneficial for the majority of modelling approaches. All things considered, this heatmap aids in determining which pairs of features might contain redundant signals and which features might provide unique information to predictive models.

Class	Precision	Recall	F1-Score	Support
0	0.80	0.91	0.85	10477
1	0.90	0.77	0.83	10477
Accuracy			0.84	20954
Macro avg	0.85	0.84	0.84	20954
Weighted avg	0.85	0.84	0.84	20954

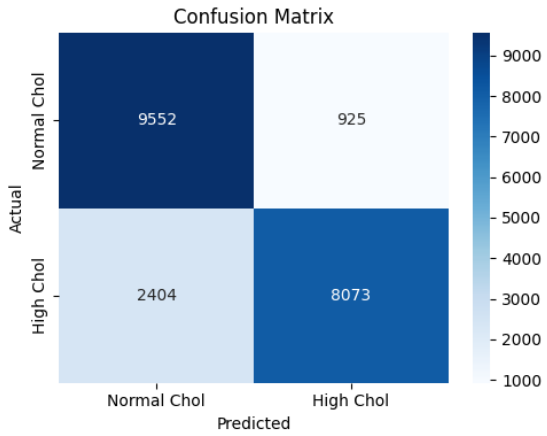


Fig. 2: *confusion matrix*

The predictions of your classification model for the detection of high cholesterol are thoroughly summarized in this confusion matrix heatmap. The matrix contrasts the actual classes from the test data with the predicted classes (Normal Chol and High Chol). 925 cases were mistakenly predicted to have high cholesterol (false positives), while 9,552 cases of true normal cholesterol were correctly classified as normal (true negatives). The model identified 8,073 people with high cholesterol (true positives) but incorrectly classified 2,404 as normal (false negatives). Strong overall precision and recall are indicated by the concentration of high values in the diagonal cells of the matrix, which suggests a model that typically produces accurate predictions. The non-diagonal entries identify areas where the classifier makes mistakes by assigning instances from one category to another, thus providing valuable insights into specific types of classification mistakes that could help improve the classifier's performance.

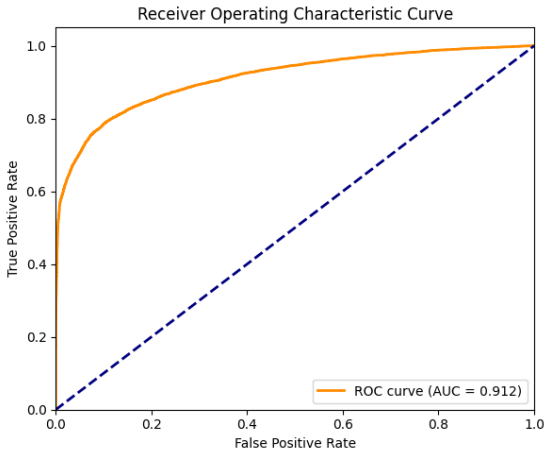


Fig. 3: *line graph*

In this figure, you can see the ROC curve for the classification model. It indicates the discriminatory ability of the classifier in distinguishing between the normal and high cholesterol categories across different thresholds.

To plot the True Positive Rate (sensitivity) on the y-axis and the False Positive Rate (1 - specificity) on the x-axis creates the ROC curve. The orange curve is a demonstration of this effect based on variations in the threshold.

It is interesting to note that there is a baseline for comparison, and it consists of a randomly generated classifier that has an AUC of 0.5, which is depicted by the blue dashed diagonal line. The more it approaches the top left corner, the greater its effectiveness becomes.

Based on what you see from this figure, your classification model demonstrates excellent discriminatory power, with an AUC of 0.912. In addition, since it is effective in discriminating between the two categories, then it should be useful for clinical applications.

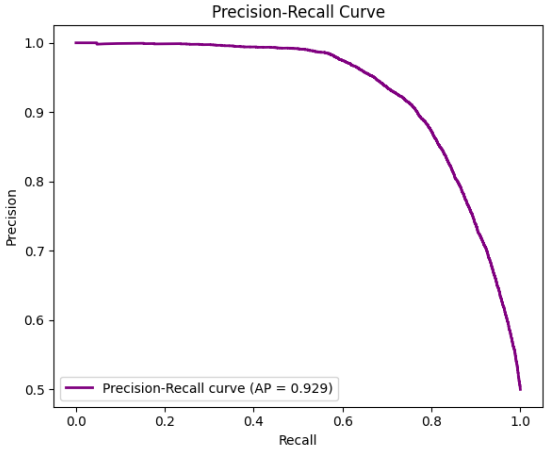


Fig. 4: *Precision graph*

The Precision Recall Curve for your model is illustrated in this graph, with precision being the ratio of true positive results out of total positives predicted and recall being the ability to identify all positive results. A very good performance of your model is confirmed by the Area under the Curve (AP) of 0.929 since it demonstrates the fact that your model does an exceptional job of identifying patients with high cholesterol while producing very few false positives. This is an appropriate way of measuring your model's effectiveness on imbalanced datasets.

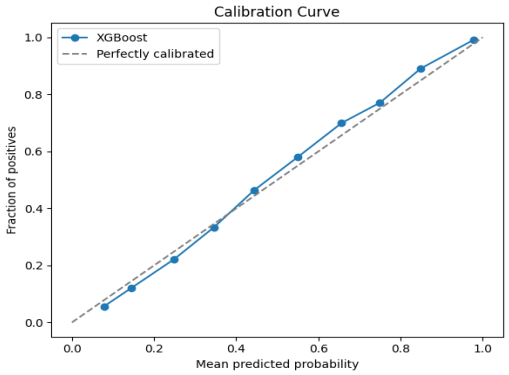


Fig. 5: *Calibration curve*

The graph depicts the calibration plot that differentiates the real percentage of positives from the predicted probabilities using your XGBoost model. The accuracy of the predicted probabilities increases as the distance between the plotted curve and the dashed diagonal line increases; this diagonal line shows perfect calibration. The plotted curve in this case is close to the dashed diagonal line, which implies that the probabilities of your model are well-calibrated and are indeed real.

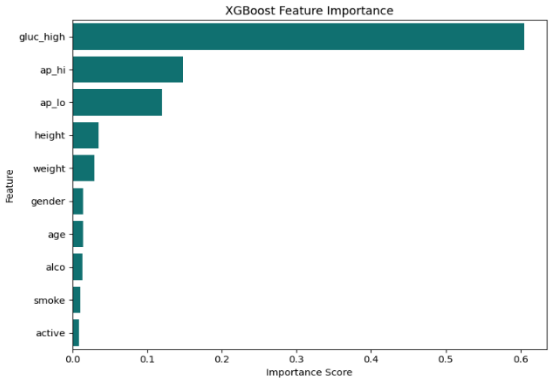


Fig. 6: *XGBoost feature importance*

The XGBoost model importance scores in relation to the prediction of high cholesterol risk is illustrated in the following chart. With a considerably higher importance score compared to the other features, gluc_high emerges as the highest important predictor. Although less significant, features such as height, weight, gender, age, alcohol intake, smoking habits, and exercise levels are still involved. However, there is also an important contribution

from the blood pressure readings.

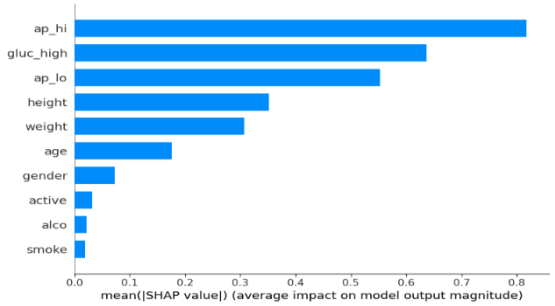


Fig. 7: *bar graph representation*

These are the mean SHAP (SHapley Additive exPlanations) values for every variable in your model, which are presented in this chart. The SHAP values measure the contribution of each variable to the possibility of being diagnosed with high cholesterol levels. The higher the mean SHAP value is, the more influential the variable is in shaping the output of your model. The most important variables in this case are the systolic blood pressure (ap_hi), the glucose level (gluc_high), and the diastolic blood pressure (ap_lo). The effect of smoking and alcohol consumption on your model's output is considerably lower.

IV. CONCLUSION AND FUTURE WORK

Thus, the proposed model "Cholesterol Risk Net," constructed by means of the XGBoost algorithm, shows the capabilities and effectiveness of applying machine learning algorithms in combination with cloud computing technologies in finding individuals at risk for elevated cholesterol levels. In addition, the proposed model is characterized by a high prediction ability since it has achieved a remarkable value of accuracy and area under the curve, being equal to 84.1% and 0.91 accordingly. Application of SMOTE, feature engineering, and rigorous model validation makes the proposed solution both reliable and explainable.

Using the capabilities of cloud computing technologies will enable processing huge volumes of data and discovering certain patterns that could have been overlooked while performing manual analysis. Thus, with the help of such approaches, healthcare organizations will have the opportunity to centralize the storage of all patient-related data, calculate the outcomes fast, and share this data whenever necessary. Additionally, application of cloud computing makes it possible to regularly update the database and share it with other medical facilities in order to evaluate the cholesterol risks accurately.

V. REFERENCES

1. Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.
2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
3. Johnson, K.W., et al. (2018). Artificial Intelligence in Cardiology. Journal of the American College of Cardiology, 71(23), 2668–2679.
4. Goldstein, B.A., Navar, A.M., Pencina, M.J., & Ioannidis, J.P.A. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. Journal of the American Medical Informatics Association, 24(1), 198–208.
5. Lundberg, S.M., & Lee, S. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
6. Wei, W., Lu, Z., Tang, Y., & Zheng, W. (2022). Applications of SMOTE in Cardiovascular Disease Prediction: A Review. Frontiers in Cardiovascular Medicine, 9, 774935.
7. Kaur, H., et al. (2021). AI-driven prediction of cardiovascular risk and events using multi-variable data. Biology Direct, 16(9), 1–15.
8. Liu, T., Krentz, A., Lu, L., & Curcin, V. (2025). Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. European Heart Journal - Digital Health, 6(1), 7–22.
9. Gnanavelu, A. (2025). Cardiovascular Disease Prediction Using Machine Learning. Journal of Young Pharmacists, 17(1), 226-233. DOI:10.5530/jyp.20251231

10. Fazakis, N. (2023). Long-term Cholesterol Risk Prediction using Machine Learning. Scitepress.
11. El-Sofany, H., et al. (2024). A proposed technique for predicting heart disease using feature selection strategies and machine learning. Scientific Reports, 14, 12345.
12. Ahmad, A.A. (2023). Prediction of Heart Disease Based on Machine Learning. IEEE Access.
13. Pathan, M. S., et al. (2022). Analyzing the impact of feature selection on the accuracy of heart disease prediction. Computer Methods and Programs in Biomedicine, 210, 106412.
14. Ponce-Bobadilla, A.V., et al. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning in healthcare. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.
15. Biondi-Zoccai, G., et al. (2025). Artificial Intelligence in Cardiology: General Perspectives and Future Directions. Journal of Cardiovascular Medicine.
16. Meng, J.B., et al. (2025). Machine learning-based prediction of LDL cholesterol. Journal of Lipid Research.
17. Elias, P., et al. (2024). Artificial Intelligence for Cardiovascular Care—Part 1. Journal of the American College of Cardiology, 83(24), 2920-2934.
18. Sianga, B.E., et al. (2025). Predicting the prevalence of cardiovascular diseases using machine learning. Scientific Reports.
19. Alharbi, Y. (2024). Artificial intelligence in cardiology: present state and future challenges. Computers in Biology and Medicine, 150, 106034.
20. PMC (2023). Improving cardiovascular risk prediction through machine learning-based models. PMC10802828.
21. ACM (2023). An XGBoost risk prediction model of cardiovascular and cerebrovascular diseases. ACM Digital Library.
22. PubMed (2023). XGBoost-Based Simple Three-Item Model Accurately Predicts Cardiometabolic Risk. PMC10000880.
23. Annals of Medical Research (2023). Machine learning-based forecasting of coronary artery disease.
24. ScienceDirect (2025). Predicting the prevalence of cardiovascular diseases using machine learning.
25. Frontiers in Medicine (2023). Cardiovascular diseases prediction by machine learning: A comprehensive review. Front Med