INFORMATICS JOURNALS

77

SHRI-LEKH (Multi-disciplinary Research Journal) Vol 1(2), December 2025, p. 77-82

Online ISSN: 3139-5694
DOI:10.18311/SHRI-LEKH/2025/v1i2/53072

An Efficient LightGBM Approach for Cardiovascular Disease Detection Highlighting Cholesterol and Body Mass Index

Swati S. Khandalkar* and Shweta M. Barhate

Department of Electronics and Computer Science, Rashtrasant Tukadoji Maharaj (RTM) Nagpur University, Nagpur – 440033, Maharashtra, India

Abstract

Obesity is a health condition that directly impacts an individual's quality of life and overall well-being. Its occurrence is directly associated with lifestyle, ultimately leading to other health conditions like diabetes, hypertension and Cardiovascular Diseases (CVD). In addition to obesity, which is measured by the Body Mass Index (BMI), CVD is also influenced by other factors like age, gender, diabetes, and cholesterol levels. Total cholesterol is another key manageable risk factor for CVD, heart attack or stroke. Heart-related disorders can be managed if detection, risk evaluation and prediction are done at an early stage. This calls for an urgent need for AI-based technologies that could reliably predict the forthcoming consequences to individuals having obesity, high cholesterol or CVD. Feature importance analysis suggests that the BMI and total cholesterol are the most influential markers, followed by pulse pressure, heart rate and glucose level. The findings indicate that the proposed Light Gradient Boosting Machine (GBM) performance model achieved an accuracy of 90%, a recall of 86%, a precision of 95%, an F1 score of 90% and a ROC-AUC of 95%, respectively. These outcomes highlight the crucial role of obesity and cholesterol management in reducing cardiovascular risk. The proposed LightGBM-based framework offers a reliable and interpretable solution for early detection of cardiovascular disease, thereby supporting preventive healthcare and informed clinical decision-making.

Keywords: Cardiovascular Disease, Cholesterol, LightGBM, Machine Learning, Obesity, Risk Detection

INTRODUCTION

Cardiovascular Disease (CVD) is the main reason for a significant number of mortalities worldwide. For the reduction in the occurrence of CVD, information on risk factors to prevent and/or manage them well becomes important (Youk *et al.*, 2020). Detection of people at high risk is very important for timely intervention and prevention.

Numerous diseases like cholesterol, heart issues and diabetes are directly associated with obesity. The most common fat quantification mechanism is BMI. Other methods include skinfold thickness, bioimpedance and waist circumference. BMI analyses give vital information concerning increased fat in the body and the classification of body mass/weight (Alsareii *et al.*, 2023).

A global study suggests that amongst all variable risk factors of heart disease, abnormal cholesterol level is associated with the incidence of CVD, particularly ischemic heart disease. High level of Total Cholesterol (TC) is considered the foremost reason for coronary atherosclerosis and increased risk of CVD (Jung *et al.*, 2022).

Machine learning (ML) algorithms, particularly ensemble-based models like LightGBM, provide advanced capabilities in detecting subtle risk patterns from large datasets. This study focuses on applying

Email: *swatiskhandalkar@gmail.com
shwetab73@yahoo.com

LightGBM to detect cardiovascular disease based primarily on obesity (BMI) and cholesterol level, which are recognised as major modifiable risk factors.

Related Work

A wide research scope has already been done on this research. Some of these are the following works:

A research paper presented by Jindal *et al.* (2021) proposed ML techniques like Logistic Regression (LR) and K-Nearest Neighbour (KNN) for predicting and classifying those with heart disease. The K-NN algorithm shows higher performance with 88.52% accuracy.

Reddy *et al.* (2022) presented the Heart Failure Dataset (HFD) prediction models, which use decision trees and naive Bayes algorithms. These two algorithms resulted in accuracies of 86% and 82% respectively, on HFD.

Ghosh *et al.* (2021) utilised supervised algorithms that included GBM, Decision Tree (DT), Random Forest (RF), KNN and Ada Boost (AB) classifier, in combination with composite classifiers and were applied to the Cleveland dataset using the bagging and boosting strategies. The comparison was done using runtime, computational complexity, and hyperparameter adjustment that gave an accuracy of 99.05%.

Ashraf *et al.* (2021) utilised individual learning algorithms as well as ensemble approaches for prediction, such as J48, multilayer perceptron, Bayes Net, KNN, random tree, Naïve Bayes, and RF. Amongst all, J48 scored 70.77% accuracy. Applying the new-fangled technique, KERAS scored an accuracy of 80%.

Nalluri *et al.* (2020) studied two predictive models created on the Extreme Gradient Boost (XGBoost) and LR algorithms. This study used the Framingham Heart Study (FHS) dataset and mean imputation to handle missing values. A 10-fold cross-validation was carried out on the train set, in which XGBoost achieved an accuracy of 84.46 % and LR models achieved an accuracy of 85.86%.

Hasan and Hasan (2020) identified and analysed the probable medical risk factors of diabetes and proposed a system which could classify diabetic individuals as per their diverse attributes using LR and ANOVA methods. Results revealed age, cholesterol and BMI as the utmost potential risk factors related to diabetes. Several classification algorithms were applied in the proposed method, in which the RF algorithm reached the highest accuracy of 90% and an Area Under Curve (AUC) score of 98%.

Kavitha *et al.* (2021) applied a unique machine-learning method in data processing with regression and classification to the Cleveland cardiopathy dataset to predict cardiopathy. DT, RF and Hybrid Model methods were utilised for ML to get the accuracy of 88% in predicting coronary heart disease.

Latifah *et al.* (2020) studied the performance of the LR and RF for detecting coronary heart disease. All rows holding missing values were removed using the FHS dataset. An accuracy of 85.04 % was attained by the LR model, and 84.4 % accuracy was attained by the RF model.

Uddin *et al.* (2023) combined three different datasets to train and test the proposed system for predicting the chances of heart disease. Different ML algorithm techniques were used to classify the risk level of the patient, like DT, Extra Trees Classifier (ET), KNN, MLP, GB, Support Vector Machine (SVM), AB Classifier, RF, XGB and LR. When equated to other ML algorithms, the highest accuracy of 99.16% was shown by the DT algorithm.

METHODOLOGY

Dataset

The study uses structured cardiovascular dataset patient records sourced from open-access medical repositories such as Kaggle datasets, which provide standardised heart disease datasets. The Framingham dataset consists 3390 records, including medical features (totChol, BPMeds, prevalentStroke, diabetes, prevalentHyp, diaBP, sysBP, BMI, glucose, heartRate),

Table 1. Features information of the FHS dataset with a description

Sr. No.	Features	Description	Categorical / Continuous
1	sex	Patient's gender	Categorical
2	age	Patient's age	Continuous
3	education	Educational level of the individual	Categorical
4	is_smoking	Smoking habit of the patient	Categorical
5	cigsPerDay	One day average cigarette consumption	Continuous
6	BPMeds	If patient consumes blood pressure medicine or not	Categorical
7	PrevalentStroke	If patient has previous stroke history or not	Categorical
8	prevalentHyp	If patient has hypertension or not	Categorical
9	diabetes	If patient has diabetes or not	Categorical
10	totChol	Serum total cholesterol level	Continuous
11	sysBP	Systolic blood pressure	Continuous
12	diaBP	Diastolic blood pressure	Continuous
13	BMI	Body Mass Index	Continuous
14	heartRate	Heart or pulse rate	Continuous
15	glucose	Blood Glucose level	Continuous
16	TenYearCVD	10-year risk of cardiovascular disease (CVD event)	Categorical

behavioural features (is_smoking, cigsPerDay), demographic information (sex, age, education) and target variable (TenYearCVD), indicating whether the patient developed CVD within 10 years (1 = Yes, 0 = No). Table 1 gives the list of attributes of FHS dataset and its description.

Data Preprocessing

The dataset was cleared by removing missing values, duplicates, and outliers. Raw data in feature engineering was converted into meaningful features that boost machine learning models. After the dataset was prepared, the data was divided into a training set and a test set. Class Balancing Synthetic Minority Over-sampling Technique (SMOTE) (Albattah & Khan, 2025) was applied to balance the distribution of CVD and non-CVD samples. The data was normalised so that all features have the same range and distribution.

Model

The Light Gradient Boosting Machine (LightGBM) Algorithm

The LightGBM gradient boosting (Omotehinwa *et al.*, 2024) is a tree-based learning framework which can detect heart diseases by using numerous machine learning problems that require features and target variables like cholesterol levels, sex, blood pressure, age, family history of heart disease, smoking status, and a few other related information. After preparation of the dataset, LightGBM uses gradient boosting to make a model which determines the target variable, i.e,

whether an individual has heart disease or not, on the basis of provided factors (Ramalingamsakthivelan *et al.*, 2022).

RESULTS AND DISCUSSION

In this section, the model's performance is highlighted, using several evaluation metrics and a confusion matrix.

The LightGBM algorithm was implemented with optimised parameters:
n_estimators =1000
max_depth = 6
learning_rate = 0.05

The system is easy to use and efficiently handles missing variables. For training and testing, the established system used 15 variable inputs from the clinical database of Kaggle clinic patients.

This database was divided into 80% training and 20% testing subsets. Performance of the model was assessed with a classification report, accuracy and feature importance ranking. The confusion matrix of the LightGBM Classifier model is displayed in Figure 1.

Using the testing dataset, the risk assessment and monitoring system for predicting the occurrence of heart disease attained 90% accuracy. The system displayed 95% for ROC-AUC, 95% for precision, 86% for recall, and 90% for f1-score (as shown in Table 2). Because of higher accuracy, efficient memory usage, and fast training times, the LightGBM technique was preferred for this project.

With a low false positive rate and high true positive rate, the ROC curve demonstrates outstanding model performance, indicating strong CVD detection accuracy (as shown in Figure 2).

Figure 3 shows the top 10 features which are important for predicting cardiovascular disease detection using the LightGBM model. totChol was the most important marker, while other indicators of body composition like BMI and pulse_pressure were also ranked in the top 3 features in predicting cardiovascular disease detection, followed by heartRate, glucose levels and age. The feature importance ranking is presented below (Figure 3).

Table 2. Light GBM performance

Metrics	Performance
Accuracy	90%
Precision	95%
Recall	86%
F1-score	90%
ROC-AUC	95%



Figure 1. Confusion matrix.

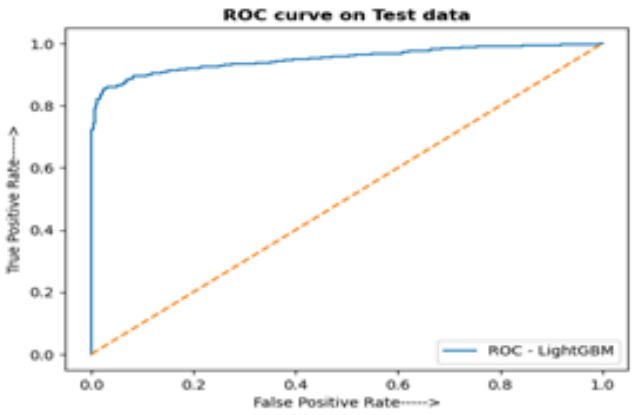


Figure 2. ROC-AUC.

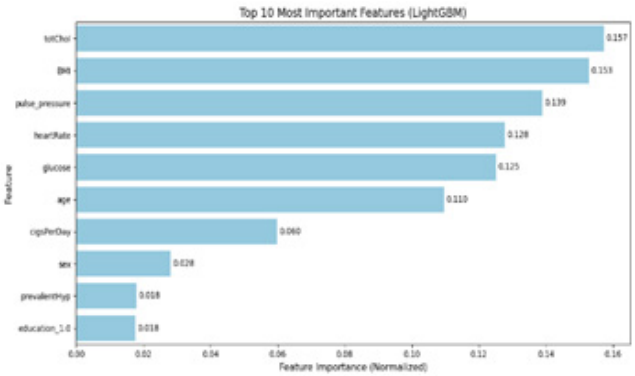


Figure 3. The top-10 ranked variables for lightGBM.

CONCLUSION

This study presented a machine learning-based framework for detecting cardiovascular disease with an organised healthcare database. This work successfully reflects the complex character of CVD causation by combining a wide range of demographic, clinical, and biochemical factors, such as cholesterol levels, BMI and Pulse pressure.

LightGBM model trained on balanced data achieves promising results, with an accuracy 90%, a recall 86%, a precision 95%, an F1-score of 90%, and ROC-AUC 95%. It demonstrates that the LightGBM algorithm can effectively detect cardiovascular disease risk, with cholesterol and BMI emerging as dominant factors.

Future Work

Future work can extend this research by integrating more diverse datasets, performing hyperparameter optimisation, and developing explainable AI models to enhance medical interpretability.

REFERENCES

Albattah, W., & Khan, R. U. (2025). Impact of imbalanced features on large datasets. *Frontiers in Big Data, 8*(1), 1-13. https://doi.org/10.3389/fdata.2025.1455442 PMid:40151465 PMCid: PMC11948280.

Alsareii, S. A., Awais, M., Alamri, A. M., AlAsmari, M. Y., Irfan, M., Raza, M., & Manzoor, U. (2023). Machine-learning-enabled obesity level prediction through electronic health records. *Computer Systems Science and Engineering, 46*(3), 3715-3728. https://doi.org/10.32604/csse.2023.035687

Ashraf, M., Ahmad, S. M., Ganai, N. A., Shah, R. A., Zaman, M., Khan, S. A., & Shah, A. A. (2021). Prediction of cardiovascular disease through cutting-edge deep learning technologies: An empirical study based on TensorFlow, PyTorch, and Keras. *In Advances in Intelligent Systems and Computing, 1*(1), 239-255. https://doi.org/10.1007/978-981-15-5113-0_18

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. M. J. M., Ignatious, E., Shultana, S., Beeravolu, A. R., & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with Relief and LASSO feature selection techniques. *Sensors, 9*(1), 19304-19326. https://doi.org/10.1109/ACCESS.2021.3053759

Hasan, K. A., & Hasan, M. A. M. (2020). Prediction of clinical risk factors of diabetes using multiple machine learning techniques resolving class imbalance. *In Proceedings of the 23rd International Conference on Computer and Information Technology (ICCIT)*, 1-6. https://doi.org/10.1109/ICCIT51783.2020.9392694

Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. *IOP Conference Series: Materials Science and Engineering, 1022*(1),1-11. https://doi.org/10.1088/1757-899X/1022/1/012072

Jung, E., Kong, S. Y., Ro, Y. S., Ryu, H. H., & Shin, S. D. (2022). Serum cholesterol levels and risk of cardiovascular death: A systematic review and a dose-response meta-analysis of prospective cohort studies. *International Journal of Environmental Research and Public Health, 19*(8272), 1-12. https://doi.org/10.3390/ijerph19148272 PMid:35886124 PMCid:PMC9316578.

Kavitha, M., Gnaneswar, G., Dinesh, R., Rohith Sai, Y., & Sai Suraj, R. (2021). Heart disease prediction using hybrid machine learning model. *In Proceedings of the Sixth International Conference on Inventive Computation Technologies (ICICT 2021)*, 1329-1333. https://doi.org/10.1109/ICICT50816.2021.9358597

Latifah, F. A., Slamet, I., & Sugiyanto. (2020). Comparison of heart disease classification with logistic regression algorithm and random forest algorithm. *AIP Conference Proceedings, 2296*(1), 1-9. https://doi.org/10.1063/5.0030579

Nalluri, S., Vijaya Saraswathi, R., Ramasubbareddy, S., Govinda, K., & Swetha, E. (2020). Chronic heart disease prediction using data mining techniques. *In Advances*

in Intelligent Systems and Computing,1079(1), 903-912. https://doi.org/10.1007/978-981-15-1097-7_76

Omotehinwa, T. O., Oyewola, D. O., & Moung, E. G. (2024). Optimizing the light gradient-boosting machine algorithm for an efficient early detection of coronary heart disease. *Informatics and Health, 1*(2), 70-81. https://doi.org/10.1016/j.infoh.2024.06.001

Ramalingamsakthivelan, N. M. K., Silambarasan, V., Thavasi, S., & Vijaya Shankar, P. (2022). Heart disease risk assessment by using LightGBM technique. *International Journal for Multidisciplinary Research (IJFMR),5*(2), 1-11. https://doi.org/10.36948/ijfmr.2023.v05i02.2620

Reddy, V. S. K., Meghana, P., Subba Reddy, N. V., & Ashwath Rao, B. (2022). Prediction on cardiovascular disease using decision tree and naïve Bayes classifiers. *Journal of Physics: Conference Series, 2161*(1), 1-8. https://doi.org/10.1088/1742-6596/2161/1/012015

Uddin, K. M. M., Ripa, R., Yeasmin, N., Biswas, N., & Ku, S. (2023). Machine learning-based approach to the diagnosis of cardiovascular disease using a combined dataset. *Intelligence-Based Medicine, 7*(1), 1-15. https://doi.org/10.1016/j.ibmed.2023.100100

Youk, T. M., Kang, M. J., Song, S. O., & Park, E.-C. (2020). Effects of BMI and LDL-cholesterol change pattern on cardiovascular disease in normal adults and diabetics. *BMJ Open Diabetes Research and Care, 8*(1), 1-7. https://doi.org/10.1136/bmjdrc-2020-001340 PMid:33355207 PMCid:PMC7757466.