Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/68/20241487

Cardiovascular disease prediction and a visualization platform building through machine learning

Qian Li[1,*], Ruiyang Liu[2], Qianxue Guo[3]

[1]Central South University, ChangSha, 410083, China
[2]Harbin Engineering University, Heilongjiang, 150001, China
[3]Central South University, ChangSha, 410083, China,

*8214210206@csu.edu.en

Abstract. In the increasing trend of population aging, medical security has become an important issue in social life that cannot be ignored. The elderly population generally faces the threat of a variety of cardiovascular diseases, which not only bring potential hazards to their physical health, but also pose more serious challenges to the medical system. Therefore, there is an urgent need to monitor the health indicators of the elderly and provide timely medical care. This article shows the corresponding machine learning model, which can provide relatively accurate predictions based on a given data set. By comparing and studying three Bayesian-optimized algorithms and baseline models (including Decision Tree, Support Vector Machine, Logical Regression, Random Forest, XGB, LGBM), a relatively better algorithm was selected. In this paper, it is believed that the model performance given by LGBM after Bayesian optimization is relatively good. It has a sound framework, high accuracy in predicting cardiovascular disease, and good performance when processing large-scale data, making it feasible for application in the field of medical security. At the same time, this research builds a visualization platform to afford assistance in the early detection of heart disease, is more friendly to non-professionals, and has made contributions in the fields of medical security and public health.

Keywords: Artificial Intelligence, machine learning, Visualization platform for digital healthcare.

1. Introduction

According to the World Health Organization, cardiovascular disease (CVD) is the leading cause of death worldwide. It is estimated that 17.9 million people have died from CVD in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attacks and strokes [1]. Meanwhile, heart failure is the leading cause of death overall in China [2].With age, the cardiovascular system of the elderly population gradually shows a physiological aging trend, which makes heart and blood vessel function gradually decline and easily triggers potential health risks. Therefore, timely medical coverage and effective health management are particularly important for the elderly. Addressing the challenges of cardiovascular disease in the elderly requires a comprehensive and systematic healthcare protection system to ensure that the elderly have access to timely and effective treatment and management. Traditional medical tools are limited by time and space and do not allow for timely or early detection. Existing AI technologies propose a variety of prediction models that make excellent contributions to the

DOI: 10.54254/2755-2721/68/20241487

measurement and evaluation of remote human physiological parameters, which can be instantly acquired by various types of sensors and transmitted via the network to be evaluated by AI models, eliminating the limitations imposed by geographic distances, which greatly increases the scope of medical insurance coverage, and further facilitates the popularization of healthcare. Cardiovascular disease risk prediction is an area of significance where improved risk prediction algorithms can benefit the most at-risk populations. Using machine learning techniques, researchers have developed various models to predict the risk of heart disease. These models can analyse large amounts of data, including electronic health records , vital signs, to provide personalized risk assessments. While AI has made significant progress in the field of heart disease prediction and management, there are still some gaps and challenges that need to be solved by future research and technological developments. AI models are often viewed as "black boxes", making it difficult to understand their decision-making processes. Improving the interpretability of models is essential for trust and adoption in medical practice.

This research reviews recent work on heart disease prediction, researches two algorithms, Bayesian-optimised LGBM, XGBoost, and evaluates how much they improve performance compared to the baseline model. The superiority of the Bayesian-optimised models is verified by discriminating, calibrating and analysing several metrics, and a corresponding visualisation platform is produced. This provides visualization data for healthcare professionals as well as helps laypersons understand their health status and encourage them to take appropriate preventive measures. Based on the research, the idea of using MMOE for joint detection of multiple diseases is further proposed.

2. Related Work

The research by Cho S.Y. compared the new model with the PCE model by testing several ML models (Logistic Regression , Treebag, Random Forest, and AdaBoost, etc.), verifying the advantages of the new model by several metrics [3].

Nikkila Prakash used a Logistic Regression Model for the prediction of cardiovascular diseases using Kaggle dataset for training the model [4] and briefly explored other models such as support vector machines. The research provided some inspiration for the selection of the topic for this paper.

P. Gupta and D. D. Seth proposed the idea of an application based on machine learning in which one can search for nearby heart central facilities and their details [5]. On this basis, this research proposes to build a visualization platform, where by analyzing and visualizing large-scale clinical data, researchers can better understand the pathogenesis and influencing factors of heart disease for subsequent clinical diagnosis. The public can also understand the risk factors and pathological features of heart disease, enabling patients to take timely preventive measures to reduce the risk of heart attack.

3. Methodology

3.1. Dataset

In this research, a dataset provided by the Kaggle platform was selected for the research of predictors of heart disease [6]. This dataset combines key biomarkers in cardiology and the patient's basic health profile, including age, gender, type of chest pain presentation (cp), resting blood pressure (trestbps), serum total cholesterol (chol), fasting blood sugar (fbs), resting electrocardiogram (restecg), maximal heart rate (thalach), exercise-induced angina pectoris (restecg), post-exercise ST-segment depression (oldpeak), slope of the ST-segment at peak exercise (slope), number of major vessels observed (ca), and thalassemia status (thal).

3.2. Data preprocessing

Data preprocessing cannot be ignored in data analysis, aiming to improve the quality of the data and the accuracy and reliability of the analysis results. In this research, a dataset provided on the Kaggle website is used, which contains 14 attributes (13 features + one label) without any duplicate, missing, or outlier values. Out of 1025 data points, there are 499 data points without heart disease (shown as 0) with a percentage of 48.68% and 526 data points with heart disease (shown as 1) with a proportion of 51.32%.

DOI: 10.54254/2755-2721/68/20241487

The dataset has shown a balanced distribution, with equal sample sizes for each category and no significant bias. This state ensures the accuracy and reliability of data analysis and model training, and effectively avoids bias caused by skewed data, so no further adjustment is needed.

3.3. Data explotation

Exploring the relationship between the variables is an integral part of how to conduct this research, which considers the correlation between the attributes and the target attributes and draws a Pearson heat map. Unlike traditional research methods, this research uses visualization to research the data. This approach makes it easy for non-specialists to understand trends, patterns, and relationships in the data without the need for in-depth knowledge of statistics or data analysis methods. A graphic presentation of data can quickly convey information and is a visible representation of data, which not only enhances patients' awareness and engagement with their own health status, but also helps to increase the probability of early diagnosis and provides effective support for the development of digital health care.

3.4. Feature selection

In order to improve the interpretability of the model, reduce the computational cost, decrease the risk of overfitting and increase the accuracy of the model, the features are screened in this research. A subset of risk factors have been identified through medical and clinical studies as influencing the risk of heart disease: age, gender, smoking, high cholesterol, high blood glucose, and high blood pressure [6]. A positive correlation between heart rate and CVD has been demonstrated in Tverdal's research [7]. The trestbps is strongly associated with hypertension, which is one of the major risk factors for CVD. The clinical manifestation of chest pain, as a direct indicator of myocardial ischemia, is an important factor in assessing the risk of heart disease. Chol and fbs levels reflect the state of lipid and glucose metabolism, and their abnormal elevation is associated with an increased risk of cardiovascular disease. The number of ca provides a quantitative indicator for assessing the extent of coronary artery disease, which is essential for the diagnosis and management of cardiovascular disease. By comprehensively analyzing these parameters, this research strives to improve the prediction accuracy of heart disease and thus provide scientific evidence and data support for the prevention, diagnosis and treatment of cardiovascular disease. The final characteristics screened in this research are age, gender, fbs, cp, thalach, trestbps, ca, and target.

3.5. Model selection

Decision Tree(DT): DT aims to create a model that matches the characteristics of the dataset and has excellent generalization effects.

Logistic Regression (LR): LR has been optimized in terms of accuracy, but this also leads to a decrease in computational speed.

Random Forest Regression Algorithm: Random Forest is particularly suitable for dealing with data containing high-dimensional features and large datasets, with excellent generalization capabilities, while effectively preventing overfitting.

Support Vector Machines (SVM): Support Vector Machines can predict well even when the data volume is small and can improve the fitting accuracy.

In this research, the above four classes of models are used as baseline models, focusing on Random Forest Regression, XGBoost and LGBM and their Bayesian optimization for research and comparison.

XGBoost: XGBoost performs well with large-scale datasets, and is particularly well suited to heart disease datasets, which can contain a large number of samples and features. In the task of classifying heart disease datasets, XGBoost is able to effectively identify key features related to heart disease, thus improving the accuracy of prediction.

LightGBM Algorithm: Combining the advantages of both SVM and Logistic.

Bayesian Optimization: Bayesian optimization searches for the optimal solution through iterative updating, effectively balancing the relationship between exploration and utilization, and efficiently improving the transformation ratio between computational cost and data results.

DOI: 10.54254/2755-2721/68/20241487

In this research, the parameters of the XGBoost and LGBM algorithms are optimized by the Bayesian optimization method and their performance is compared with other models. This strategy not only optimizes the performance of the models, but also improves the adaptability and efficiency of the algorithms in specific application scenarios.

3.6. Evaluation criteria

Commonly used evaluation criteria in evaluating machine learning are precision rate, recall, F1 score and ROC curve. Precision: the proportion of correctly predicted positive samples to all positively predicted samples.

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{1}$$

Recall: also known as sensitivity and hit rate, is the proportion of positive samples that are correctly predicted.

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{2}$$

F1 Score: The F1 score is the harmonic mean of precision and recall, it takes into account both precision and recall, maximizing both at the same time, and striking a balance between the two.

$$\mathrm{F1} = \frac{2\times\mathrm{PRE}\times\mathrm{REC}}{\mathrm{PRE+REC}} \tag{3}$$

ROC curve: Area Under Curve (AUC) is an important indicator of the ROC curve used to quantify the diagnostic or predictive ability of a model. The closer the value of AUC is to 1, the better the performance of the model.

3.7. Visualization platform

This research builds a large-scale visualization platform, that analyzes and models the existing data and selects three models with relatively better F1 scores. The platform is intended for both the public and healthcare organizations, providing heart disease prediction services as well as a large amount of research data for healthcare organizations to assist in clinical diagnosis. The visualization platform is superior in that it graphically displays a wide range of data, including various physiological indicators and risk assessments of existing patients in the database. This helps medical professionals quickly recognize abnormal values of physiological indicators, more accurately assess the patients' risk levels and disease development, and also provides a powerful aid for early medical diagnosis.

4. Results and discussions

Distribution map of features: For the diagnostic model of heart disease, patients can be divided into two categories: "sick" and "not sick". In this study, all the features and their associations with the disease were visualized, partially as shown in Figure 1.

DOI: 10.54254/2755-2721/68/20241487

Figure 1. Feature analysis

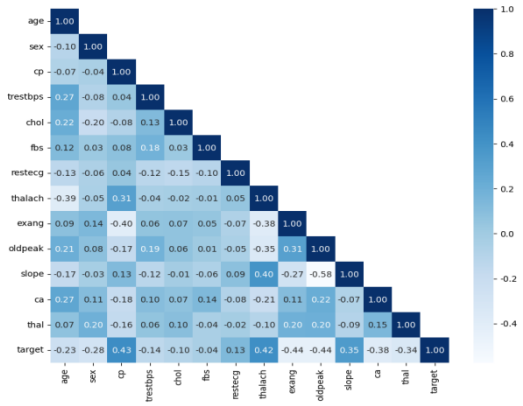


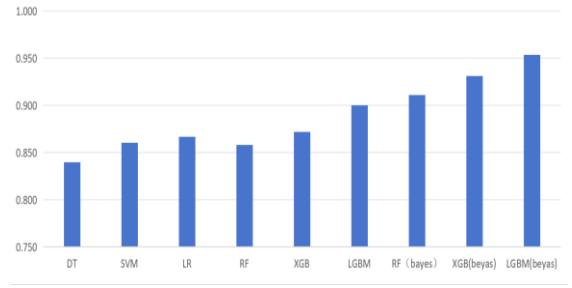
Figure 2. Pearson heat map

Feature analysis: Logical analysis is based on the correlation between data. Therefore, the Pearson heat map (Figure 2) is used to show the correlation between the data.

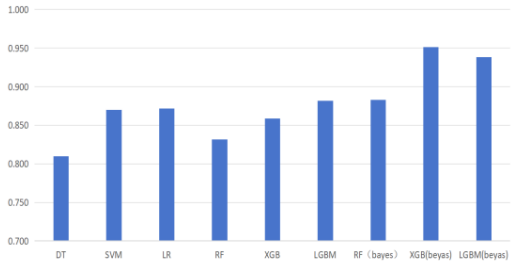
In this study, the indicators used by commonly used machine learning algorithms for heart disease prediction are compared. Summarized in Figure 3 and Figure 4.

As can be seen from Figure 4, the performance indicators of the enhanced model are relatively better. After optimizing XGB, and LGBM all improved in accuracy and recall compared to when they were not optimized. It can be seen from the chart that the indicators of the optimized algorithm are significantly better than the pre-optimized algorithm. Among them, the accuracy and F1 value of the LGBM after Bayesian optimization have reached more than 95%. The optimized XGB and LGBM have improved compared with the accuracy and recall rates of XGB and LGBM obtained in [9].

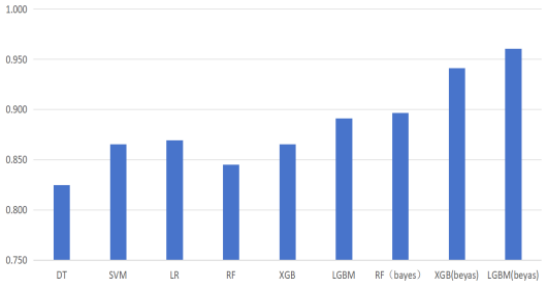
DOI: 10.54254/2755-2721/68/20241487



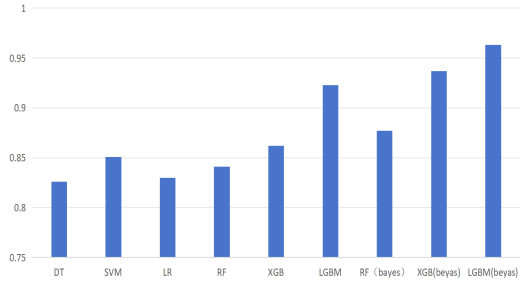
(a) Precision



(b) Recall

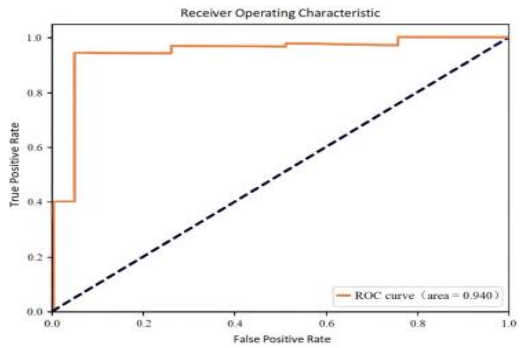


(c) F1 score

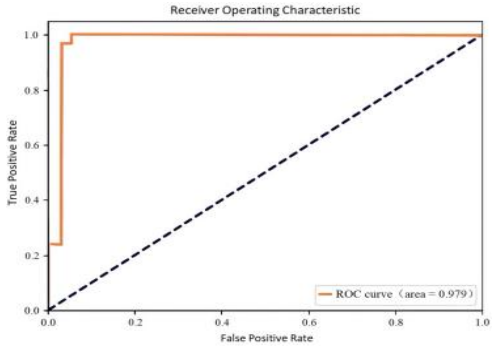


(d) Accuracy

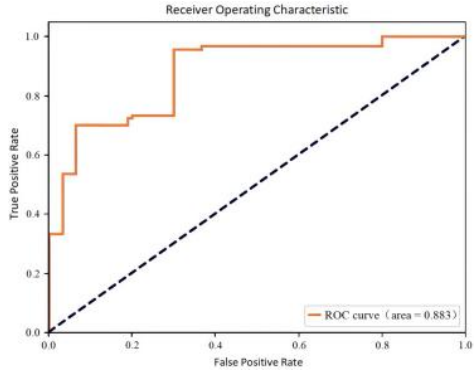
Figure 3. Model parameter comparison



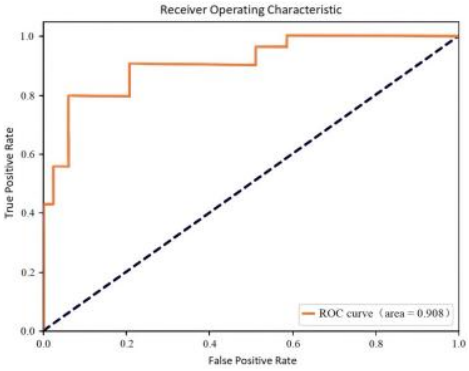
(a) LGBM



(b) LGBM (Bayesian optimization)



(c) XGB



(d) XGB(Bayesian optimization)

Figure 4. ROC curve

DOI: 10.54254/2755-2721/68/20241487

Visualization platform: This research has developed a basic data visualization platform based on HTML technology. The platform conducts in-depth analysis and mining of large-scale clinical data, processes the data in a refined manner and displays the chart. Medical staff can use the visualization platform to access and analyse data anytime, anywhere, and make accurate diagnoses and treatment plans. At the same time, visualization tools provide patients with an intuitive way to understand their own condition, so that they can more clearly understand their health and treatment options, thereby enhancing their participation and confidence in the treatment process. The transparency and visualization of this kind of information help to promote communication and trust between doctors and patients, and improve patients' treatment satisfaction and rehabilitation effects.

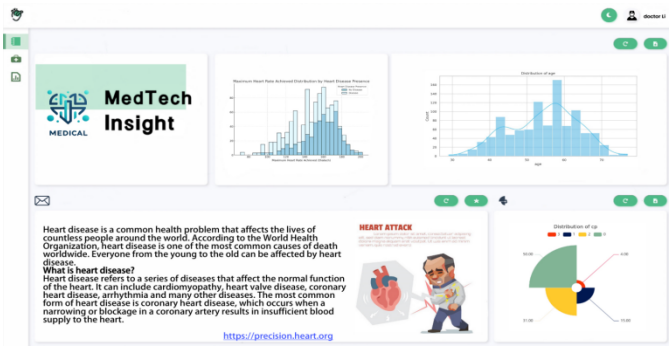


Figure 5. Visualization platform

5. Conclusion

In this paper, a comprehensive comparative study was conducted to explore and evaluate the performance and efficiency of a variety of advanced machine learning algorithms in the field of heart disease prediction. In this study, DT, SVM,LF, and RF are used as the base line models. The performance gap between XGB, LGBM and their Bayesian-optimized algorithms is studied and compared in detail. Through rigorous experimental design and data analysis, it not only verifies the superiority of machine learning technology in heart disease prediction, but also proves the significant effect of Bayesian tuning technology in optimizing complex models such as XGB and LGBM. In order to better demonstrate the practicality and intuitiveness of the research results, this research has developed a basic visualization platform. This platform not only achieves a high degree of user-friendliness in technology, including but not limited to excellent cross-platform and cross-device access capabilities, but its design concept and functional implementation are designed to lay a solid foundation for future technological upgrades and functional expansion. Through this platform, users can intuitively understand the performance of various machine learning models in heart disease prediction, as well as the performance comparison between them. It is believed that with the continuous progress of technology and the continuous expansion of application fields, this visualization platform will be able to bring more innovation and value to the field of medical and health data analysis.

Future outlook: Zhang Ran [10] proposed that stroke, heart disease, diabetes or elevated blood sugar have a high risk of co-morbidity and should attract attention in the fields of clinical diagnosis and treatment. From this, this paper proposes the idea of applying MMOE to scenarios where stroke, diabetes and heart disease are predicted at the same time. The full name of the MMOE model is Multi-gate Mixture-of-Experts. The model was proposed by Google in the article Modeling Task Relationships in Multi-task Learning with a Multi-gate Mixture of Experts [11] published in KDD in 2018. The high co-morbidity rate of stroke, heart disease, diabetes, or elevated blood sugar among these diseases means that they tend to occur simultaneously or affect each other in the same patient, which makes their prediction and treatment more complicated. MMOE, as an advanced MTL framework, can handle multiple tasks at the same time, and minimize the noise interference between multiple tasks. It is very suitable for the target needs of joint detection of multiple diseases at risk of co-morbidity. The "expert network" in MMOE can be individually designed for each disease, focusing on capturing characteristics

DOI: 10.54254/2755-2721/68/20241487

related to a particular disease. The "gating mechanism" can dynamically assign weights to different experts based on the characteristics of the input data to optimize the prediction of each task. This means that the model can be flexibly adjusted to meet the specific needs of different diseases. By simultaneously predicting multiple related diseases in the same model, MMOE can be more efficient than training independent models for each disease separately. In addition, it can also improve the accuracy of predictions, because multitasking learning usually leads to better generalization. There are complex interactions and co-morbid relationships between stroke, heart disease, and diabetes. By learning the common characteristics and differences of these diseases, MMOE helps to gain an in-depth understanding of the association between them, which is of great significance for the prevention and treatment of diseases.In short, it is not only feasible to apply MMOE to the simultaneous prediction of stroke, heart disease, and diabetes, but also in view of the high co-morbidity and interaction between these diseases. It provides a comprehensive solution to these common and serious health problems, and has the potential to significantly improve the accuracy and efficiency of forecasting. This preliminary research shows that the predictive scenario based on the MMOE model is not only theoretically feasible, but also shows great potential in practical applications.

References

[1] World health organization. Cardiovascular diseases (CVDs), 11 June 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] J. He, D. Gu, X. Wu, K. Reynolds, X. Duan, C. Yao, et al., Major causes of death among men and women in China, New England, J. Med., vol. 353, no. 11, pp. 1124-1134, Sep. 2005.

[3] Cho, SY., Kim, SH., Kang, SH. et al. Pre-existing and machine learning-based models for cardiovascular risk prediction. Sci Rep 11, 8886 (2021). https://doi.org/10.1038/s41598-021-88257-w.

[4] N. Prakash, M. Mahesh and P. Gouthaman. Cardiovascular Disease Risk Assessment using Machine Learning, 2023 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2023, pp. 249-256, doi: 10.1109/ICICT57646.2023.10133957.

[5] P. Gupta and D. D. Seth. Comparative Analysis of Machine Learning Classifiers for Accurate and Early Detection of Heart Disease, 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2022, pp. 1-5, doi: 10.1109/ICRITO56286.2022.9964882.

[8] R. Das, I. Turkoglu and A. Sengur. Effective diagnosis of heart disease through neural networks ensembles, Expert Syst. Appl., vol. 36, pp. 7675-7680, 2009.Tverdal, Aage, Vidar Hjellvik, and Randi Selmer. Heart rate and mortality from cardiovascular causes: a 12 year follow-up study of 379 843 men and women aged 40–45 years." European heart journal 29.22 (2008): 2772-2781.

[9] Das, SS Gourab Kumar, et al. Heart Disease Prediction Using Different Boosting Models. 2023 International Conference on Advanced & Global Engineering Challenges (AGEC). IEEE, 2023.

[10] Ran Zhang, Lu Yun, Shan-shan Zhang, et al. Prevalence pattern and component correlation of chronic disease comorbidity among the elderly in China. Chinese Journal of Public Health, 2019, 35(8): 1003-1005.

[11] Ma, Jiaqi, et al. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2018.