PREDICTING CORONARY HEART DISEASE USING AN IMPROVED LIGHTGBM MODEL

ABSTRACT

Coronary heart disease (CHD) presents a persistent threat without a definitive cure. Early detection is crucial for effective treatment. To address this, we introduce HY\_OptGBM, a predictive model leveraging an optimized LightGBM classifier. We fine-tuned LightGBM's hyperparameters and enhanced its loss function, utilizing the advanced hyperparameter optimization framework, OPTUNA. Training involved adjusting hyperparameters and applying the improved loss function. We evaluated our model using CHD data from the Framingham Heart Institute, employing various metrics like precision, recall, F-score, accuracy, MCC, sensitivity, specificity, and AUC. Notably, our model achieved an AUC of 97.9%, outperforming other comparative models. These findings suggest that our approach enhances early CHD identification within the general population, potentially reducing the economic burden associated with CHD treatment.

INTRODUCTION

CHD is a prevalent cardiovascular disorder resulting from the buildup of atherosclerotic plaques in the coronary arteries, leading to a reduction in blood flow to the heart muscle. This condition presents a range of symptoms, including chest pain or angina, shortness of breath, palpitations, and heart failure. In severe cases, CHD may lead to a heart attack, which can result in permanent damage to the heart muscle and have a profound impact on an individual's quality of life. Therefore, it is imperative to recognize and manage CHD through appropriate medical intervention and lifestyle modifications. [1]. Early detection of CHD can improve the cure probability and can decrease the cost of treatment. Numerous machine learning algorithms and data mining technologies have been widely used in the medical field [2]–[6] in recent years, owing to advancements in machine learning algorithms and a significant reduction in the cost of data storage. Data mining technology has become essential for healthcare data mining, such as disease diagnosis, auxiliary diagnosis, drug mining, and biomedicine. Through data mining technology, hidden knowledge about diseases can be extracted from large quantities of unstructured medical data, disease prediction models can be developed, and results can be analyzed. Health organizations face tremendous challenges in providing high-quality and affordable healthcare. A hospital provides quality healthcare services that require physicians to have comprehensive knowledge and a correct diagnosisfor the patient to avoid wasting healthcare resources due to inaccurate diagnoses. Data mining technology can perform efficiently and can play a crucial role in clinical cases. The optimal hyperparameters [7], [8] for any classification algorithm significantly affect its performance. The accuracy of the classification algorithm can be improved by selecting the optimal set of hyperparameters. In this study, a state-of-the-art hyperparameter optimization framework (OPTUNA) [9] was employed to obtain optimal hyperparameter values for the LightGBM model. Therefore, in this study, the most suitable set of hyperparameters was determined from the available hyperparameters. Hyperparametric optimization can be accomplished by different methods, such as random and grid searches. Another method is the OPTUNA hyperparametric search. Because the number of hyperparameters in the LightGBM significantly affects its performance, conventional random and grid search methods do not learn from the previous optimization, which wastes considerable time and is inefficient. The OPTUNA framework continuously learns from previous optimizations and adjusts the hyperparameters as necessary. Therefore, OPTUNA was chosen in this paper for hyperparameter optimization. The loss function also affects the model accuracy [10]. In this paper, the focal loss function was proposed based on the cross-entropy loss by adding the category weight α and the sample difficulty weight modulating factor γ. The aim of this study was to address the problem of unbalanced proportions of positive and negative samples. Additionally, the focal loss function can improve the overall performance of the model.

**Existing System:**

The existing system likely relies on traditional methods of diagnosing and predicting coronary heart disease (CHD), which may involve manual evaluation of risk factors, symptoms, and medical history by healthcare professionals. These methods might not fully leverage the potential of machine learning and data mining technologies to accurately predict CHD risk and detect it early. Moreover, the existing system might not employ advanced techniques for hyperparameter optimization and loss function optimization, potentially limiting the predictive performance of the models used.

**Disadvantages:**

1. Reliance on traditional diagnostic methods may lead to delays in CHD detection and diagnosis.

2. Lack of utilization of machine learning and data mining technologies may result in suboptimal predictive accuracy.

3. Limited optimization of hyperparameters and loss functions may hinder the performance of predictive models.

4. Unbalanced proportions of positive and negative samples may impact model performance and lead to biased predictions.

**Proposed System:**

The proposed system, HY\_OptGBM, introduces several advancements to address the limitations of the existing system:

1. Leveraging an optimized LightGBM classifier: By utilizing LightGBM, a powerful machine learning algorithm, and optimizing its hyperparameters using the OPTUNA framework, the proposed system aims to enhance the accuracy of CHD prediction.

2. Improvement of the loss function: The proposed system introduces the focal loss function, which addresses the problem of unbalanced proportions of positive and negative samples, potentially improving overall model performance.

3. Evaluation using comprehensive metrics: The proposed system evaluates model performance using various metrics such as precision, recall, F-score, accuracy, MCC, sensitivity, specificity, and AUC, providing a comprehensive assessment of its effectiveness.

4. Enhanced early detection of CHD: By achieving an AUC of 97.9% and outperforming other comparative models, the proposed system demonstrates its capability to improve early identification of CHD within the general population, thereby potentially reducing the economic burden associated with CHD treatment.

**Advantages**:

1. Improved predictive accuracy: By optimizing hyperparameters and utilizing an advanced loss function, the proposed system enhances the accuracy of CHD prediction compared to traditional methods.

2. Comprehensive evaluation: The use of multiple evaluation metrics ensures a thorough assessment of the model's performance, providing confidence in its effectiveness.

3. Early detection capabilities: The proposed system's high AUC indicates its ability to detect CHD early, enabling timely intervention and potentially reducing healthcare costs.

4. Utilization of advanced techniques: By incorporating state-of-the-art machine learning techniques such as LightGBM and OPTUNA, the proposed system represents a cutting-edge approach to CHD prediction, offering superior performance compared to conventional methods.

**literature review**

Overweight and obesity contribute to the development of cardiovascular disease (CVD) in general and coronary heart disease (CHD) in particular in part by their association with traditional and non-traditional CVD risk factors. [1]Obesity is also considered to be an independent risk factor for CVD. The metabolic syndrome, of which central obesity is an important component, is strongly associated with CVD including CHD. There is abundant epidemiologic evidence of an association between both overweight and obesity and CHD.[1] Evidence from post-mortem studies and studies involving coronary artery imaging is less persuasive. Recent studies suggest the presence of an obesity paradox with respect to mortality in persons with established CHD. [1]Physical activity and preserved cardiorespiratory fitness attenuate the adverse effects of obesity on CVD events. Information concerning the effect of intentional weight loss on CVD outcomes in overweight and obese persons is limited.

Overweight and obesity contribute to the development of cardiovascular disease (CVD) in general and coronary heart disease (CHD) in particular in part by their association with traditional and nontraditional CVD risk factors. Obesity is also considered to be an independent risk factor for CVD.[5] The metabolic syndrome, of which central obesity is an important component, is strongly associated with CVD including CHD. There is abundant epidemiologic evidence of an association between both overweight and obesity and CHD. Evidence from postmortem studies and studies involving coronary artery imaging is less persuasive. [5]Recent studies suggest the presence of an obesity paradox with respect to mortality in persons with established CHD. Physical activity and preserved cardiorespiratory fitness attenuate the adverse effects of obesity on CVD events. Information concerning the effect of intentional weight loss on CVD outcomes in overweight and obese persons is limited.

[7]Machine learning (ML) is a burgeoning field of medicine with huge resources being applied to fuse computer science and statistics to medical problems. Proponents of ML extol its ability to deal with large, complex and disparate data, often found within medicine and feel that ML is the future for biomedical research, [7]personalized medicine, computer-aided diagnosis to significantly advance global health care. However, the concepts of ML are unfamiliar to many medical professionals and there is untapped potential in the use of ML as a research tool.[7] In this article, we provide an overview of the theory behind ML, explore the common ML algorithms used in medicine including their pitfalls and discuss the potential future of ML in medicine.

**CONCLUSION AND DISCUSSION**

This paper proposed a CHD prediction method based on the HY\_OptGBM model. Framingham Heart Institute data on CHD was selected as measurements, and the proposed method was trained using the HY\_OptGBM algorithm and the comparison algorithms. Although different algorithms were used for CHD prediction in this study, the best CHD prediction was achieved by the improved LightGBM algorithm. When using data from the Framingham Heart Institute's CHD study, observing all predicted values using the HY\_OptGBM algorithm yielded more successful results, which is the significance of this study. In the experiment, sensitivity, specificity, accuracy, precision, recall, F-score, AUROC, AUPRC and MCC were used as evaluation metrics. The experimental results of the DT, RF, CB, XGB, ADA, BG, GBM and HY\_OptGBM algorithms were compared, and the best results were obtained using the HY\_OptGBM algorithm. The sensitivity was 0.897, the specificity was 0.963, the accuracy was 0.930, the precision was 0.963, the recall was 0.897, the F-score was 0.929, the AUROC was 0.978, the AUPRC was 0.983, and the MCC was 0.861. This study proposed optimizing the hyperparameters of the LightGBM algorithm and improving its loss function (FL). The experimental results will change when changing the alpha and gramma parameters of the FL function. After the experiments were conducted, when the parameter alpha was None and gamma was 1, the accuracy, F-score, AUROC, AUPRC, MCC metrics had the best results. When alpha was 0.1 and gamma was 0, the specificity and precision had the best results. When alpha was 0.9 and gamma was 0, the sensitivity and recall had the best results. When evaluating the performance of a machine learning algorithm, usually multiple evaluation metrics are considered together, so alpha was taken as None and gamma was taken as 1 to obtain the final experimental results. As shown in Tables Ⅴ, Ⅵ, Ⅷ, and Fig. 9 and Fig. 10, the best results can be obtained when making predictions with the proposed method. To compare studies in the literature with the proposed methodology, experimental studies using the Framingham CHD dataset were checked. This dataset has mostly been used to predict the probability of developing CHD within ten years. In 2021, Orit Goldman et al. [12] used ANN models to predict CHD, and the prediction results showed that the lift and gain curves of ANN models were higher than those of FRS models in terms of the highest percentile. For higher risk scores, the ANN model had higher sensitivity and specificity than the FRS model, but the ANN model had lower area under the curve (AUC) values. For the precision-recall measures, ANN models produce significantly better results than FRS models in terms of AUC values. In a 2019 study, Juan-jose Beunza et al. [43] conducted a comparative study of the dataset using machine learning methods. Decision trees, random forests, support vector machines, neural networks and logistic regression were selected for the classification study. The results of the study demonstrated that the support vector machine algorithm had the best AUC value of 0.75. Meeshanthini V Dogan et al. [44], in a 2018 study, used machine learning techniques to construct predictive CHD models. The accuracy, sensitivity and specificity obtained using the random forest classifier were 0.78, 0.75 and 0.80, respectively. In a 2021 study by Meeshanthini V Dogan et al. [45], an ensemble genetic performance genetic model for predicting 3-year coronary events was developed. This model showed a sensitivity of 0.79, a specificity of 0.75, a sensitivity of 0.15 and a specificity of 0.93 on the test set. In 2022, Steven Simon et al. [46] used logistic regression to classify and predict CHD, and AUROC values of 0.71 were obtained. In a study by S. Prabu [47] in 2021, CHD was predicted by using Gaussian process regression (GPR) and kernel ridge regression (KRR) machine learning algorithms and a hyperparametric search of the algorithm, and the final prediction results demonstrated a recall of 0.902, an F1-score of 0.821, and an accuracy of 0.86. Table Ⅸ gives details of CHD prediction studies in the past 5 years using the Framingham Heart Institute's open dataset. Using the synthetic minority oversampling technique (SMOTE) for preprocessing the dataset, optimizing the hyperparameters of the algorithm and improving its loss function in the experimental study were considered, and the success of the proposed method in predicting CHD compared with other methods was demonstrated. In contrast to other studies in the literature, the use of the most advanced algorithm in ensemble learning (LightGBM) in this study, as well as the use of the most advanced hyperparameter optimization framework (OPTUNA) for optimization of the hyperparameters of the algorithm and improvement of its loss function, led to sensitivity, specificity, accuracy, precision, recall, F score, AUROC, AUPRC and MCC enhancements, which are important for diseases such as CHD, which have lethal disease consequences. Due to the lack of similar optimized and improved prediction methods in the literature, the proposed method in this paper provides a new perspective for future CHD prediction studies In future studies, the Framingham Heart Institute dataset should be used to predict CHD, and multiple CHD datasets should be used to build predictive models. When experimenting with the FL function, the alpha and gamma parameters affect the study results. Thus, more accurate results can be obtained by constructing prediction models through multiple trials. The methodology proposed in this study will also be integrated in future studies. As the numbers of trials and datasets increases, it will be necessary to obtain a successful result by adjusting the default parameters presented in this paper. In addition to using a single model to predict CHD, alternatively, one may consider building a prediction model by combining multiple models.

**REFERENCES**

[1] N. Katta, T. Loethen, C. J. Lavie, and M. A. Alpert, "Obesity and coronary heart disease: Epidemiology, pathology, and coronary artery imaging," Current Problems Cardiology, vol. 46, no. 3, p. 100655, Mar. 2021, doi: 10.1016/j.cpcardiol.2020.100655.

[2] G. S. Handelman, H. K. Kok, R. V. Chandra, A. H. Razavi, M. J. Lee, and H. Asadi, "eDoctor: Machine learning and the future of medicine," J. Internal Medicine, vol. 284, no. 6, pp. 603–619, Sep. 2018, doi: 10.1111/joim.12822.

[3] E. L. Romm and I. F. Tsigelny, "Artificial intelligence in drug treatment," Annu. Rev. Pharmacology Toxicology, vol. 60, no. 1, pp. 353–369, Jan. 2020, doi: 10.1146/annurev-pharmtox-010919-023746.

[4] L. Lo Vercio et al., "Supervised machine learning tools: A tutorial for clinicians," J. Neural Eng., vol. 17, no. 6, p. 062001, Dec. 2020, doi: 10.1088/1741-2552/abbff2.

[5] S. Rauschert, K. Raubenheimer, P. E. Melton, and R. C. Huang, "Machine learning and clinical epigenetics: A review of challenges for diagnosis and classification," Clin. Epigenetics, vol. 12, no. 1, p. 51, Apr. 2020, doi: 10.1186/s13148-020-00842-4.

[6] Y. Arfat, G. Mittone, R. Esposito, B. Cantalupo, G. M. De Ferrari, and M. Aldinucci, "Machine learning for cardiology," Minerva Cardiology Angiology, vol. 70, no. 1, pp. 75–91, Mar. 2022, doi: 10.23736/s2724- 5683.21.05709-4.

[7] S. Nematzadeh, F. Kiani, M. Torkamanian-Afshar, and N. Aydin, "Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases," Comput. Biol. Chemistry, vol. 97, p. 107619, Apr. 2022, doi: 10.1016/j.compbiolchem.2021.107619.

[8] M. Liang et al., "Improving genomic prediction with machine learning incorporating TPE for hyperparameters optimization," Biology, vol. 11, no. 11, p. 1647, Nov. 2022, doi: 10.3390/biology11111647.

[9] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "OPTUNA: A next-generation hyperparameter optimization framework," in Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Anchorage, USA: ACM, 2019, pp. 2623–2631.

[10] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," Computerized Med. Imag. Graph., vol. 95, p. 102026, Jan. 2022, doi: 10.1016/j.compmedimag.2021.102026.

[11] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Proc. 31st Int. Conf. Neural Inf. Process. Syst., Long Beach, CA, USA: ACM, 2017, pp. 3149–3157.

[12] O. Goldman, O. Raphaeli, E. Goldman, and M. Leshno, "Improvement in the prediction of coronary heart disease risk by using artificial neural networks," Quality Manage. Health Care, vol. 30, no. 4, pp. 244–250, Jul. 2021, doi: 10.1097/qmh.0000000000000309.

[13] Z. Du et al., "Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: Model development and performance evaluation," JMIR Med. Inform., vol. 8, no. 7, p. e17257, Jul. 2020, doi: 10.2196/17257.

[14] J. K. Kim and S. Kang, "Neural network-based coronary heart disease risk prediction using feature correlation analysis," J. Healthcare Eng., vol. 2017, p. 2780501, Sep. 2017, doi: 10.1155/2017/2780501.

[15] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, and T. Kitai, "Artificial intelligence in precision cardiovascular medicine," J. Amer. College Cardiology, vol. 69, no. 21, pp. 2657–2664, May. 2017, doi: 10.1016/j.jacc.2017.03.571.

[16] A. Akella and S. Akella, "Machine learning algorithms for predicting coronary artery disease: Efforts toward an open source solution," Future Sci. OA, vol. 7, no. 6, p. FSO698, Mar. 2021, doi: 10.2144/fsoa-2020-0206.

[17] L. J. Muhammad, I. Al-Shourbaji, A. A. Haruna, I. A. Mohammed, A. Ahmad, and M. B. Jibrin, "Machine learning predictive models for coronary artery disease," SN Comput. Sci., vol. 2, no. 5, pp. 350, Mar. 2021, doi: 10.1007/s42979-021-00731-4.

[18] C. A. U. Hassan et al., "Effectively predicting the presence of coronary heart disease using machine learning classifiers," Sensors, vol. 22, no. 19, p. 7227, Sep. 2022, doi: 10.3390/s22197227.

[19] captainozlem, "Framingham\_CHD\_preprocessed\_data. Version 1." Accessed: May 5, 2020 [Online.] Available: https://www.kaggle.com/-datasets/captainozlem/framingham-chdpreprocessed-data/download?datasetVersionNumber=1

[20] V. Voillet, P. Besse, L. Liaubet, M. San Cristobal, and I. González, "Handling missing rows in multi-omics data integration: Multiple imputation in multiple factor analysis framework," BMC Bioinf., vol. 17, no. 1, p. 402, Oct. 2016, doi: 10.1186/s12859-016-1273-5.

[21] G. Douzas and F. Bacao, "Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE," Inf. Sci., vol. 501, pp. 118–135, Oct. 2019, doi: 10.1016/j.ins.2019.06.007.

[22] D. Che, Q. Liu, K. Rasheed, and X. Tao, "Decision tree and ensemble learning algorithms with their applications in bioinformatics," in Software Tools and Algorithms for Biological Systems. Advances in Experimental Medicine and Biology, H. Arabnia and Q. N. Tran, Eds. New York, NY: Springer, 2011, pp. 191–199.

[23] L. Yang et al., "Study of cardiovascular disease prediction model based on random forest in eastern China," Sci. Rep., vol. 10, no. 1, p. 5245, Mar. 2020, doi: 10.1038/s41598-020-62133-5.

[24] J. T. Hancock and T. M. Khoshgoftaar, "Catboost for big data: An interdisciplinary review," J. Big Data, vol. 7, no. 1, p. 94, Nov. 2020, doi: 10.1186/s40537-020-00369-8.

[25] W. Wenbo, S. Yang, and C. Guici, "Blood glucose concentration prediction based on VMD-KELM-AdaBoost," Med. Biol. Eng. Comput., vol. 59, no. 11-12, pp. 2219–2235, Sep. 2021, doi: 10.1007/s11517-021-02430-x.

[26] X. Mi, F. Zou, and R. Zhu, "Bagging and deep learning in optimal individualized treatment rules," Biometrics, vol. 75, no. 2, pp. 674– 684, Mar. 2019, doi: 10.1111/biom.12990.

[27] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM)," Diagnostics, vol. 11, no. 9, p. 1714, Sep. 2021, doi: 10.3390/diagnostics11091714.

[28] J. Feng, B. Ni, D. Xu, and S. Yan, "Histogram contextualization," IEEE Trans. Image Process., vol. 21, no. 2, pp. 778–788, Feb. 2012, doi: 10.1109/tip.2011.2163521.

[29] P. Łabędź, K. Skabek, P. Ozimek, and M. Nytko, "Histogram adjustment of images for improving photogrammetric reconstruction," Sensors, vol. 21, no. 14, p. 4654, Jul. 2021, doi: 10.3390/s21144654.

[30] L. Lin, J. Zhang, N. Zhang, J. Shi, and C. Chen, "Optimized LightGBM power fingerprint identification based on entropy features," Entropy, vol. 24, no. 11, p. 1558, Oct. 2022, doi: 10.3390/e24111558.

[31] O. Krivorotko, M. Sosnovskaia, I. Vashchenko, C. Kerr, and D. Lesnic, "Agent-based modeling of COVID-19 outbreaks for New York state and UK: Parameter identification algorithm," Infectious Disease Model., vol. 7, no. 1, pp. 30–44, Mar. 2022, doi: 10.1016/j.idm.2021.11.004.

[32] A. Namoun, B. R. Hussein, A. Tufail, A. Alrehaili, T. A. Syed, and O. BenRhouma, "An ensemble learning based classification approach for the prediction of household solid waste generation," Sensors, vol. 22, no. 9, p. 3506, May. 2022, doi: 10.3390/s22093506.

[33] M. M. Arifin et al., "OLGBM: OPTUNA optimized light gradient boost-ing machine for intrusion detection," in 2021 Int. Conf.Comput. Commun. Chem. Mater. Electron. Eng. (IC4ME2), Rajshahi, Bangladesh: IEEE, 2021, pp. 1–4.

[34] P. Srinivas and R. Katarya, "hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost," Biomed. Signal Process. Control, vol. 73, p. 103456, Mar. 2022, doi: 10.1016/j.bspc.2021.103456.

[35] D. Jensen and J. Neville, "Correlation and sampling in relational data mining," in Proc. 33rd Symp. Interface Comput. Sci. Statist., 2001.

[36] S. Yan, J. M. Peck, M. Ilgu, M. Nilsen-Hamilton, and M. H. Lamm, "Sampling performance of multiple independent molecular dynamics simulations of an RNA aptamer," ACS Omega, vol. 5, no. 32, pp. 20187–20201, Aug. 2020, doi: 10.1021/acsomega.0c01867.

[37] M. Komorowski, D. C. Marshall, J. D. Salciccioli, and Y. Crutain, "Exploratory data analysis," in Secondary Analysis of Electronic Health Records. Cham: Springer International Publishing, 2016, pp. 185–203.

[38] T. R. Vetter, "Descriptive statistics: Reporting the answers to the 5 basic questions of who, what, why, when, where, and a sixth, so what?," Anesthesia Analgesia, vol. 125, no. 5, pp. 1797–1802, Nov. 2017, doi: 10.1213/ane.0000000000002471.

[39] B. Wang, J. J. Klemeš, P. S. Varbanov, and M. Zeng, "An extended grid diagram for heat exchanger network retrofit considering heat exchanger types," Energies, vol. 13, no. 10, p. 2656, May. 2020, doi: 10.3390/en13102656.

[40] M. W. Browne, "Cross-validation methods," J. Math. Psychol., vol. 44, no. 1, pp. 108–132, Mar. 2000, doi: 10.1006/jmps.1999.1279.

[41] S. Parvandeh, H.-W. Yeh, M. P. Paulus, and B. A. McKinney, "Consensus features nested cross-validation," Bioinformatics, vol. 36.