Original Article

Check for updates

Available online at
www.heca-analitika.com/ijds

Infolitika Journal of Data Science

Vol. 1, No. 2, 2023

Cardiovascular Disease Prediction Using Gradient Boosting Classifier

Rivansyah Suhendra [1,*], Noviana Husdayanti [2], Suryadi Suryadi [1], Ilham Juliwardi [1], Sanusi Sanusi [1], Abdurrahman Ridho [1], Muhammad Ardiansyah [1], Murhaban Murhaban [1] and Ikhsan Ikhsan [3]

[1] Department of Information Technology, Faculty of Engineering, Universitas Teuku Umar, Aceh Barat 23681, Indonesia: rivansyahsuhendra@utu.ac.id (R.S.); suryadi@utu.ac.id (S.S.); ilhamjuliwardi@utu.ac.id (I.J.); sanusi@utu.ac.id (S.A.) ; abdurrahmanridho@utu.ac.id (A.R.) ; muhammadardiansyah@utu.ac.id (M.A.) ; murhaban@utu.ac.id (M.M.)
[2] Teuku Umar Hospital, Aceh Jaya 23654, Indonesia: novianahdy@acehjayakab.go.id (N.H.)
[3] Department of Sport Education, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; Ikhsanhamdani93@gmail.com (I.I.)

***** Correspondence: rivansyahsuhendra@utu.ac.id

Article History

Received 30 October 2023
Revised 9 December 2023
Accepted 18 December 2023
Available Online 24 December 2023

Keywords:
Cardiovascular disease
Gradient boosting classifier
Classification
Healthcare analytics

Abstract

Cardiovascular Disease (CVD), a prevalent global health concern involving heart and blood vessel disorders, prompts this research's focus on accurate prediction. This study explores the predictive capabilities of the Gradient Boosting Classifier (GBC) in cardiovascular disease across two datasets. Through meticulous data collection, preprocessing, and GBC classification, the study achieves a noteworthy accuracy of 97.63%, underscoring the GBC's effectiveness in accurate CVD detection. The robust performance of the GBC, evidenced by high accuracy, highlights its adaptability to diverse datasets and signifies its potential as a valuable tool for early identification of cardiovascular diseases. These findings provide valuable insights into the application of machine learning methodologies, particularly the GBC, in advancing the accuracy of CVD prediction, with implications for proactive healthcare interventions and improved patient outcomes.

1. Introduction

Cardiovascular disease, commonly known as heart disease, stands as the predominant global cause of mortality [1, 2]. Recent findings from the World Heart Federation indicate that CVD is responsible for one in four deaths [3]. Projections from the World Health Organization (WHO) estimate that heart failure and stroke will contribute significantly to the anticipated 25 million CVD-related deaths by 2030 [4]. Mitigating CVD's impact is challenging due to its irreversible nature, emphasizing the crucial role of early prevention [5]. The complexity of CVD diagnosis arises from various risk factors, including high cholesterol, high blood pressure, smoking, diabetes, and obesity [6]. Researchers have explored diverse methods for detecting CVD, reflecting the urgency in developing effective predictive models.

Within healthcare monitoring, cardiac health monitoring holds paramount importance [7]. A predictive system for heart disease aids medical practitioners in making informed decisions about their patients' cardiac health [8]. Anomalies such as improper signal conduction or irregular heart rhythms, known as arrhythmias, can have severe consequences, including fatality. The intricate nature of assessing heart alignment may lead to oversights by medical experts. In this context, a machine learning-based approach for heart alignment prediction becomes instrumental, offering a potential solution to enhance the precision of cardiac assessments and improve patient outcomes [9].

In recent studies on cardiovascular disease prediction, notable contributions have been made by various researchers employing diverse methodologies and GBC has been proved to be useful in classifying medical data [10–14]. Geweid et al. advanced the field by constructing identification procedures for cardiovascular diseases using an improved Support Vector Machine (SVM)-based duality optimization approach. Despite the progress, existing prediction algorithms for cardiovascular diseases exhibit limitations in forecast accuracy and computational efficiency [5]. Javid addressed this challenge by combining different classifiers with a voting-based model, resulting in enhanced accuracy for heart disease identification, as demonstrated on the UCI Heart Disease dataset [15].

Furthermore, researchers such as Chen et al. proposed a prediction method based on physical investigation markers, utilizing clinical valuation signals to classify patients with hypertension. Employing Extreme Gradient Boosting (XGBoost), they aimed to accurately predict patient outcomes by isolating crucial components through recursive feature removal and cross-validation [16]. The ensemble technique developed by Latha and Jeeva stands out for its commitment to improving prediction accuracy through bagging and boosting strategies. They aggregated classifiers from various algorithms, including Naive Bayes, Bayes Net, C 4.5, Multilayer Perceptron, PART, and Random Forest, creating a hybrid model that achieved an accuracy of 85.48 percent for heart disease risk identification [17].

Amin et al. undertook an approach that involved combining various variables in their model to forecast cardiovascular disease. Using datasets from the Cleveland database available in the UCI machine learning repository, they implemented a range of classification models, including Decision Tree, Logistic Regression, Support Vector Machine, Neural Network, Vote, Naive Bayes, and k-NN. Their findings indicated a prediction accuracy of 87.4% for heart illness [18]. Meanwhile, U. Haq et al. introduced a novel method for recognizing heart disease by integrating feature selection and classification algorithms. Leveraging a sequential reverse feature selection algorithm and the K-Nearest Neighbors (KNN) classification model, their approach demonstrated remarkable accuracy, providing an innovative avenue for improved disease recognition [19].

This study aims to improve the accuracy of cardiovascular disease prediction using the Gradient Boosting Classifier machine learning method. By applying GBC, we seek to enhance the precision of identifying individuals at risk of CVD. Additionally, we aim to compare the performance of GBC across datasets to assess its generalizability. This research contributes insights into the potential of machine learning, particularly GBC, in early CVD detection, supporting proactive healthcare interventions and reducing CVD-related morbidity and mortality.

2. Materials and Methods

This study employs the Gradient Boosting Classifier for cardiovascular disease prediction. The method involves data collection, cleaning, GBC classification, and model evaluation using metrics like accuracy, recall, precision, and F1-score. Two datasets undergo preprocessing, and GBC, known for handling non-linear relationships, is chosen for its effectiveness with complex medical data. The goal is to provide a concise and accurate prediction of CVD, contributing insights to cardiovascular health using machine learning.

2.1. Data Collection

This study employs two distinct datasets for binary classification tasks in cardiovascular disease prediction. The Heart Disease dataset, obtained from the University of California Irvine (UCI) machine learning repository, comprises 1025 instances. Additionally, the Cardiovascular Disease dataset, obtained from the reference [20], consists of 70000 instances. Table 1 provides detailed information on the datasets, including their sources, instance counts, and feature dimensions.

Established in 1988, Dataset 1 incorporates four databases: Cleveland, Hungary, Switzerland, and Long Beach V. Initially encompassing 76 attributes, including the predicted attribute. However, all published experiments, including the current study, focus on utilizing a subset of 14 attributes. The datasets used in this study focus on predicting cardiovascular disease with a binary classification setup. Regarding the target variable indicating the presence of cardiovascular disease, Dataset 1 has 526 samples for "no_cardio" and 499 samples for "cardio." Meanwhile, Dataset 2 consists of 35021 samples for "no_cardio" and 34979 samples for "cardio.". Dataset 1 consist of 13 features per instance whereas dataset 2 characterized by 12 features. The specifics of these features, including their data types, can be found in Table 2.

2.2. Gradient Boosting Classifier

Gradient Boosting Classifier is a powerful method employed in the development of classification and regression models, specifically optimized for learning processes in models that exhibit non-linear characteristics [21]. Frequently associated with decision or regression trees, this technique builds a series of weak

Table 1. Dataset information.

No	Datasets	Number of Rows	Number of Features	Target	
				No_cardio	Cardio
1	Heart Disease	1025	13	526	499
2	Cardiovascular Disease	70.000	12	35021	34979

Table 2. Dataset's features information.

Dataset	Feature name	Data Type
1	Age	numerical
	Sex	categorical
	Chest_pain_type	categorical
	Resting_blood_pressure	numerical
	Serum_cholestoral	numerical
	Fasting_blood_sugar	categorical
	Resting_electrocardiographic_results	categorical
	Maximum_heart_rate	numerical
	Exercise_induced_angina	categorical
	Oldpeak	numerical
	Slope_peak_exercise	numerical
	Major_vessel	numerical
	Thal	numerical
2	Age	numerical
	Height	numerical
	Weight	numerical
	Gender	categorical
	Systolic_blood_pressure	numerical
	Diastolic_blood_pressure	numerical
	Cholesterol	numerical
	Glucose	numerical
	Smoking	categorical
	Alcohol_intake	categorical
	Physical_activity	categorical

prediction models, such as regression decision trees, in a gradual sequential manner [22].

Each new learner is added incrementally to the model, refining the ensemble's predictive capabilities. The nodes and leaves within the constructed trees contribute to the decision-making process, yielding predictions based on these decision nodes. Despite the individual weakness of each regression tree, their collective strength as an ensemble significantly enhances predictive accuracy. This incremental and sequential construction of ensembles allows for the rectification of errors in previous ensembles, contributing to the overall robustness of the model [23, 24].

The effectiveness of GBC method in healthcare is notably pronounced due to its ability to handle complex and intricate patterns within medical datasets. In the context of predicting cardiovascular disease, GBC excels in capturing non-linear relationships among various health indicators, contributing to more accurate and reliable predictions. The sequential construction of ensembles, rectifying errors made by prior models, enhances the model's adaptability to diverse medical scenarios. Moreover, GBC's success in surpassing other machine learning algorithms in healthcare data challenges underscores its versatility and robustness [25]. Its proficiency has been demonstrated in various medical applications, showcasing its potential for tasks such as disease prediction, diagnosis, and patient risk assessment [26].

2.3. Evaluation of the Model

In evaluating the Gradient Boosting Classifier's performance in the classification of cardiovascular disease, a comprehensive set of performance metrics was employed. These metrics, including accuracy, recall, precision, and the F1 score, collectively provided a nuanced understanding of the classifier's effectiveness [22, 27]. Accuracy, measuring overall correctness in classifications, revealed the model's general predictive capacity. Concurrently, recall demonstrated GBC's proficiency in correctly identifying true positive cases, a crucial aspect in healthcare applications where accurately identifying instances of cardiovascular disease is paramount. Precision, assessing the classifier's ability to minimize false positives, ensured that the predictions were not dominated by false alarms. Furthermore, the F1 score, balancing precision and recall, furnished a comprehensive evaluation of the GBC's performance, particularly crucial in a medical context where both sensitivity and specificity are crucial for reliable predictions [28, 29].

3. Results and Discussion

The experimental setup for this study involved a comparative analysis between the Gradient Boosting Classifier and several classic machine learning algorithms, including Linear Discriminant Analysis (LDA), K Nearest Neighbors (KNN), Support Vector Machines (SVM), and Naive Bayes (NB). This comparison aimed to discern the relative performance of the Gradient Boosting Classifier in the context of cardiovascular disease prediction. To ensure robust and unbiased evaluations, the k-fold cross-validation resampling approach was adopted, with 'k' set to 10. This choice of the k-fold cross-validation method, a widely accepted practice in machine learning research, aimed at mitigating bias in the prediction model by iteratively

Table 3. Classification results of dataset 1.

Model	Accuracy	AUC	Recall	Precision	F1-Score
SVM - Linear Kernel	0.6224	0.0000	0.5965	0.7533	0.5402
Naive Bayes	0.8187	0.8982	0.8640	0.8005	0.8304
Linear Discriminant Analysis	0.8215	0.9128	0.9048	0.7844	0.8392
K Neighbors Classifier	0.6959	0.8188	0.6899	0.7120	0.6991
Gradient Boosting Classifier	**0.9763**	**0.9861**	**0.9810**	**0.9735**	**0.9771**

Table 4. Classification results of dataset 2.

Model	Accuracy	AUC	Recall	Precision	F1-Score
SVM - Linear Kernel	0.5492	0.0000	0.6521	0.6143	0.5207
Naive Bayes	0.5537	0.7011	0.1530	**0.7746**	0.2527
Linear Discriminant Analysis	0.6478	0.7050	0.6133	0.6585	0.6351
K Neighbors Classifier	0.5567	0.5750	0.5511	0.5571	0.5540
Gradient Boosting Classifier	**0.7363**	**0.8024**	**0.6954**	0.7573	**0.7249**

Table 5. Confusion matrix on validation data.

Dataset	Actual	Predicted	
		No_cardio	Cardio
1	No_cardio	93	2
	Cardio	3	107
2	No_cardio	5513	1597
	Cardio	2042	4848

dividing the dataset into training and validation subsets. The 'k' value of 10 ensured a thorough evaluation by repeating this process ten times, providing a comprehensive assessment of each algorithm's performance.

The GBC model for cardiovascular disease was effectively trained using an 80-20 split for training and validation. Further enhancement involved a 10-fold cross-validation with random search hyperparameter tuning, revealing optimal configurations: criterion='friedman_mse', learning_rate=0.1, loss='log_loss', max_depth=3, subsample=1.0.

The GBC model excels in predicting cardiovascular disease when compared to traditional machine learning algorithms, showcasing its superiority in both datasets. In Dataset 1, GBC achieves an unparalleled accuracy of 0.9763, outperforming SVM (0.6224), Naive Bayes (0.8187), Linear Discriminant Analysis (0.8215), and K Neighbors Classifier (0.6959). GBC's remarkable recall of 0.981 further emphasizes its exceptional ability to correctly identify positive cases, surpassing the recall values of other algorithms. Precision and F1-Score values of 0.9735 and 0.9771, respectively, affirm GBC's precision-recall balance, demonstrating its proficiency in minimizing false positives while maintaining high sensitivity. The detailed results for Dataset 1 are available in Table 3.

In Dataset 2, the GBC demonstrates superior accuracy (0.7363) compared to other algorithms, surpassing SVM (0.5492), Naive Bayes (0.5537), Linear Discriminant Analysis (0.6478), and K Neighbors Classifier (0.5567). While GBC's accuracy slightly decreases compared to Dataset 1, it maintains a considerable lead over alternative methods. GBC's recall of 0.6954, precision of 0.7573, and F1-Score of 0.7249 underline its balanced performance in identifying true positive cases and minimizing false positives, outpacing other algorithms in the context of cardiovascular disease prediction. The detailed results for Dataset 2 are presented in Table 4.

In dataset 1, the GBC exhibits an exceptional Area Under the Curve (AUC) value of 0.9861, reflecting its superior discriminative ability. This signifies the model's capacity to effectively distinguish between individuals with and without cardiovascular disease, validating its robust performance. In dataset 2, the GBC maintains a commendable AUC of 0.8024, indicating strong discriminatory power in a different dataset context. While slightly lower than dataset 1 as presented in table 2 and 3, this AUC value reinforces the GBC's effectiveness in accurately predicting cardiovascular disease across diverse datasets.

Comparatively, GBC's consistently superior performance across both datasets underscores its efficacy in addressing the complexities of cardiovascular health prediction. Its ensemble-based learning, which incrementally corrects errors in previous models, proves advantageous in capturing intricate relationships within the data. While other algorithms demonstrate respectable performances, GBC's comprehensive accuracy, recall, and precision metrics establish it as a reliable and promising tool for healthcare analytics. Further refinements and investigations into GBC's parameters could potentially enhance its utility, making it an even more potent asset for accurate cardiovascular disease prediction.

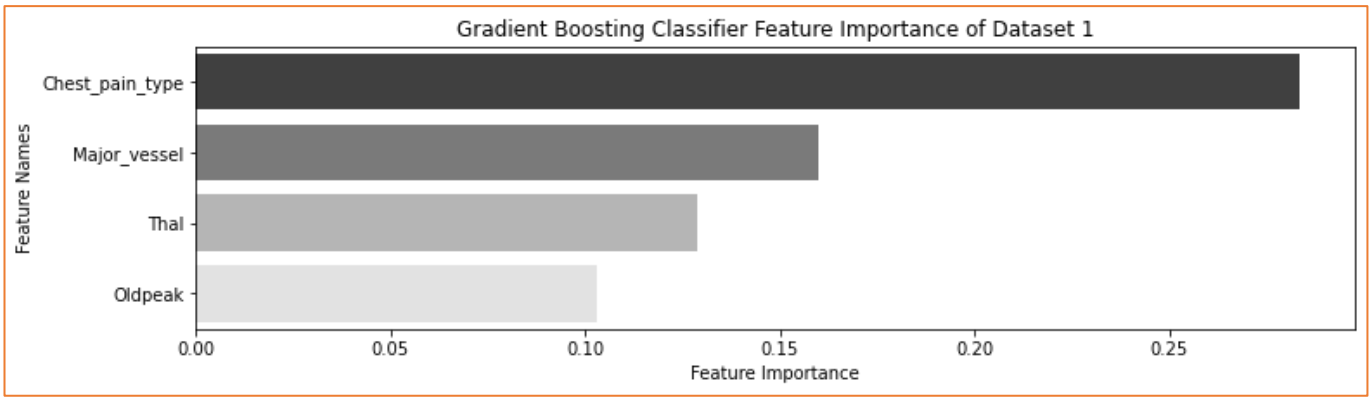

Figure 1. GBC's feature importance of dataset 1.

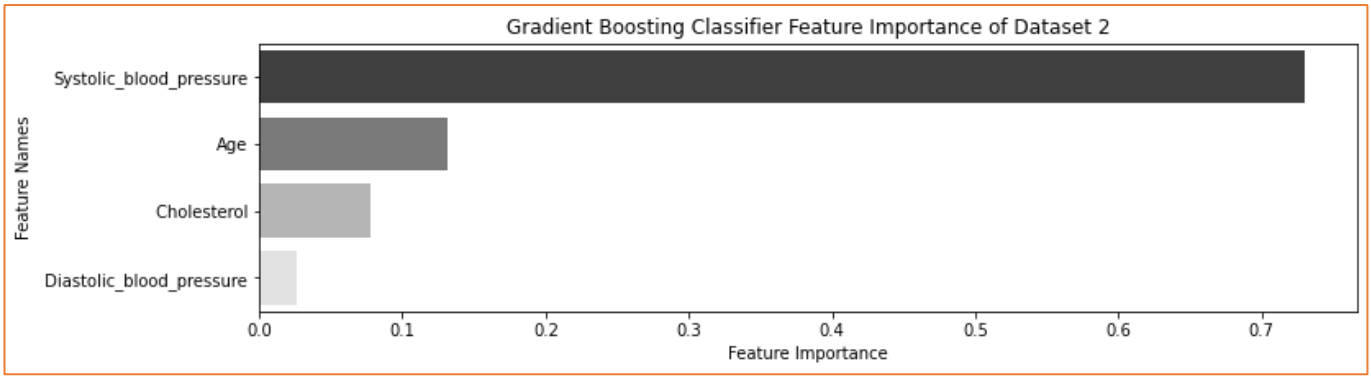

Figure 2. GBC's feature importance of dataset 2.

The confusion matrix encapsulates the classification performance of the predictive model, offering a nuanced breakdown of its outcomes as presented in Table 5. For Dataset 1, the matrix reveals 93 instances correctly identified as positive (True Positives), emphasizing the model's ability to accurately recognize cases of cardiovascular disease. Simultaneously, it correctly identifies 107 instances as negative (True Negatives), demonstrating proficiency in recognizing non-cardiovascular cases. However, the model exhibits a minor tendency towards false positives, marking 2 cases as positive when they are not (False Positives), and 3 false negatives, failing to identify positive instances. This comprehensive breakdown underscores the model's strengths and areas for improvement, crucial for refining its predictive capabilities.

In the case of Dataset 2, the confusion matrix portrays a more extensive evaluation of the model's performance. The model excels in identifying positive instances, with 5513 True Positives, showcasing its robustness in correctly classifying cases of cardiovascular disease. It also accurately identifies 4848 cases as negative (True Negatives), demonstrating its proficiency in recognizing instances without cardiovascular disease. However, the model encounters challenges, as indicated by 1597 False Positives, signifying instances incorrectly classified as positive, and 2042 False Negatives, depicting instances where the model failed to identify positive cases. This intricate analysis enables a thorough understanding of the model's strengths and limitations, providing essential insights for potential enhancements in predictive accuracy.

The feature importance results for Dataset 1 and Dataset 2, as determined by the GBC method, are visually presented in Figure 1 and Figure 2, respectively. For Dataset 1, the feature importance analysis reveals that "Chest_pain_type" holds the highest importance with a weight of 0.2835. This indicates that this feature significantly influences the GBC model's decision-making process in predicting cardiovascular disease. Following closely is "Major_vessel" with a feature importance of 0.1599, underscoring its substantial contribution to the model's predictive power. Additionally, "Thal" and "Oldpeak" exhibit importance values of 0.1287 and 0.1030, respectively, emphasizing their roles in the model's decision hierarchy. This insight into feature importance not only aids in understanding the key contributors to the model's predictions but also provides valuable information for potential feature selection strategies or domain-specific investigations.

Moving to Dataset 2, the GBC method assigns the highest importance to "Systolic_blood_pressure" with a significant weight of 0.7297. This underscores the critical

role of systolic blood pressure in predicting cardiovascular disease within this dataset. Following in importance are "Age" with 0.1318, "Cholesterol" with 0.0776, and "Diastolic_blood_pressure" with 0.0258. These feature importance values shed light on the relative influence of each variable in the model's decision process. Understanding these importance weights is instrumental in comprehending the factors driving the predictive accuracy of the GBC model in Dataset 2, providing valuable insights for healthcare professionals and researchers alike.

The identified importance of specific features in the context of cardiovascular disease prediction aligns with well-established medical knowledge. In dataset 1, the feature "Chest_pain_type" emerges as crucial, resonating with the recognized diagnostic significance of chest pain in cardiovascular assessments. Additionally, the prominence of "Major_vessel" and "Thal" features corresponds to their established roles in cardiovascular health indicators. This alignment reinforces the clinical relevance of the Gradient Boosting Classifier's findings, showcasing its ability to discern features deeply rooted in medical understanding. In dataset 2, the elevated importance of "Systolic_blood_pressure" echoes its well-documented significance as a key contributor to hypertension, a prominent cardiovascular risk factor. This direct linkage between identified important features and established medical insights enhances the interpretability and practical applicability of the model, reinforcing its potential as a valuable tool in cardiovascular health assessments.

While the GBC method has demonstrated remarkable efficacy in predicting cardiovascular disease, certain limitations and avenues for future exploration exist. One limitation lies in the interpretability of the model's decisions, as the ensemble-based nature of GBC makes it inherently complex. Understanding the reasoning behind each prediction can be challenging, hindering the model's interpretability in clinical settings. Additionally, the method's performance may vary across diverse demographic groups or datasets, emphasizing the importance of robust validation across different populations. Future research directions could involve exploring explainability techniques to enhance the interpretability of GBC, investigating the model's generalizability across diverse populations, and further refining hyperparameter tuning strategies to optimize its performance in real-world healthcare scenarios. Addressing these limitations and delving into these research directions will contribute to the continued evolution and applicability of the GBC method in cardiovascular disease prediction.

4. Conclusions

This study highlights the effectiveness of the Gradient Boosting Classifier in accurately predicting cardiovascular disease. The achieved high accuracy, precision, recall, and F1-score underscore the potential of GBC as a valuable tool in the early detection and classification of CVD. The comparison across two distinct datasets further emphasizes the robustness and generalizability of the GBC method, providing valuable insights for future applications in diverse medical contexts. The study's findings contribute to the growing body of knowledge on machine learning applications in healthcare, particularly in enhancing the predictive capabilities for cardiovascular health. However, limitations include the reliance on specific datasets, potentially restricting generalizability, and challenges in the interpretability of the GBC model. Future research should explore larger, diverse datasets, implement explainability techniques, and conduct real-world validations to enhance the practical utility of GBC in healthcare settings. Addressing these limitations will strengthen the applicability of machine learning in cardiovascular disease prediction.

Author Contributions: Conceptualization, R.S. and N.H.; methodology, R.S. and S.N.; software, A.R., I.J. and M.A.; validation, N.H., S.S. and M.M.; formal analysis, R.S.; investigation, I.I; resources, N.H. and S.S.; data curation, M.A.; writing—original draft preparation, R.S. and N.H.; writing—review and editing, R.S., S.S., N.H., A.R. and M.M.; visualization, A.R. and I.J.; supervision, S.S. and M.M.; project administration, I.I; All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset 1 (Hearth disease) from the University of California Irvine (UCI) machine learning repository is accessible at the following link: https://archive.ics.uci.edu/dataset/45/heart+disease and dataset 2 (Cardiovascular disease) is obtained from: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.

Acknowledgments: We express our sincere gratitude to all individuals and institutions whose contributions have enriched this study. Additionally, we appreciate the participants who provided crucial data, enabling the exploration of cardiovascular disease prediction.

Conflicts of Interest: All the authors declare that there are no conflicts of interest.

References

1. Maruyama, K., and Iso, H. (2014). Overview of the Role of Antioxidant Vitamins as Protection Against Cardiovascular

Disease, *Aging*, Elsevier, 213–224. doi:10.1016/B978-0-12-405933-7.00021-4.

2. Teunis, C. J., Stroes, E. S. G., Boekholdt, S. M., Wareham, N. J., Murphy, A. J., Nieuwdorp, M., Hazen, S. L., and Hanssen, N. M. J. (2023). Tryptophan metabolites and incident cardiovascular disease: The EPIC-Norfolk prospective population study, *Atherosclerosis*, Vol. 387, 117344. doi:10.1016/j.atherosclerosis.2023.117344.
3. Lopez, E. O., Ballard, B. D., and Jan, A. (2022). Cardiovascular disease, *StatPearls [Internet]*, StatPearls Publishing.
4. Karageorgou, D., Micha, R., and Zampelas, A. (2015). Mediterranean Diet and Cardiovascular Disease: An Overview of Recent Evidence, *The Mediterranean Diet*, 91–104.
5. Geweid, G. G. N., and Abdallah, M. A. (2019). A New Automatic Identification Method of Heart Failure Using Improved Support Vector Machine Based on Duality Optimization Technique, *IEEE Access*, Vol. 7, 149595–149611. doi:10.1109/ACCESS.2019.2945527.
6. The Lancet Regional Health – Europe. (2023). Navigating disparities in cardiovascular disease outcomes across Europe: a call to action, *The Lancet Regional Health - Europe*, Vol. 33, 100746. doi:10.1016/j.lanepe.2023.100746.
7. Ahmad, G. N., Shafiullah, Fatima, H., Abbas, M., Rahman, O., Imdadullah, and Alqahtani, M. S. (2022). Mixed Machine Learning Approach for Efficient Prediction of Human Heart Disease by Identifying the Numerical and Categorical Features, *Applied Sciences*, Vol. 12, No. 15, 7449. doi:10.3390/app12157449.
8. Ali, L., Niamat, A., Khan, J. A., Golilarz, N. A., Xingzhong, X., Noor, A., Nour, R., and Bukhari, S. A. C. (2019). An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure, *IEEE Access*, Vol. 7, 54007–54014. doi:10.1109/ACCESS.2019.2909969.
9. Gao, X.-Y., Amin Ali, A., Shaban Hassan, H., and Anwar, E. M. (2021). Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method, *Complexity*, Vol. 2021, 1–10. doi:10.1155/2021/6663455.
10. Solomon, D. D., Khan, S., Garg, S., Gupta, G., Almjally, A., Alabduallah, B. I., Alsagri, H. S., Ibrahim, M. M., and Abdallah, A. M. A. (2023). Hybrid Majority Voting: Prediction and Classification Model for Obesity, *Diagnostics*, Vol. 13, No. 15, 2610. doi:10.3390/diagnostics13152610.
11. Suhendra, R., Suryadi, S., Husdayanti, N., Maulana, A., Noviandy, T. R., Sasmita, N. R., Subianto, M., Earlia, N., Niode, N. J., and Idroes, R. (2023). Evaluation of Gradient Boosted Classifier in Atopic Dermatitis Severity Score Classification, *Heca Journal of Applied Sciences*, Vol. 1, No. 2, 54–61. doi:10.60084/hjas.v1i2.85.
12. Rahman, S., Irfan, M., Raza, M., Moyeezullah Ghori, K., Yaqoob, S., and Awais, M. (2020). Performance Analysis of Boosting Classifiers in Recognizing Activities of Daily Living, *International Journal of Environmental Research and Public Health*, Vol. 17, No. 3, 1082. doi:10.3390/ijerph17031082.
13. Bakasa, W., and Viriri, S. (2023). VGG16 Feature Extractor with Extreme Gradient Boost Classifier for Pancreas Cancer Prediction, *Journal of Imaging*, Vol. 9, No. 7, 138. doi:10.3390/jimaging9070138.
14. Nipa, N., Riyad, M. H., Satu, S., Walliullah, Howlader, K. C., and Moni, M. A. (2023). Clinically adaptable machine learning model to identify early appreciable features of diabetes in Bangladesh, *Intelligent Medicine*. doi:10.1016/j.imed.2023.01.003.
15. Javid, I., Alsaedi, A. K. Z., and Ghazali, R. (2020). Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method, *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 3.
16. Chen, Q., Meng, Z., and Su, R. (2020). WERFE: A gene selection algorithm based on recursive feature elimination and ensemble strategy, *Frontiers in Bioengineering and Biotechnology*, Vol. 8, 496.
17. Latha, C. B. C., and Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, *Informatics in Medicine Unlocked*, Vol. 16, 100203. doi:10.1016/j.imu.2019.100203.
18. Amin, M. S., Chiam, Y. K., and Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease, *Telematics and Informatics*, Vol. 36, 82–93. doi:10.1016/j.tele.2018.11.007.
19. Haq, A. U., Li, J., Memon, M. H., Hunain Memon, M., Khan, J., and Marium, S. M. (2019). Heart Disease Prediction System Using Model Of Machine Learning and Sequential Backward Selection Algorithm for Features Selection, *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, IEEE, 1–4. doi:10.1109/I2CT45611.2019.9033683.
20. Janosi, A., Steinbrunn, W., Pfisterer, M., Detrano, R., and Aha, W. D. (1988). UCI machine learning repository, *Heart Disease Data Set*.
21. Suhendra, R., Suryadi, S., Husdayanti, N., Maulana, A., and Rizky, T. (2023). Evaluation of Gradient Boosted Classifier in Atopic Dermatitis Severity Score Classification, *Heca Journal of Applied Sciences*, Vol. 1, No. 2, 54–61. doi:10.60084/hjas.v1i2.85.
22. Noviandy, T. R., Maulana, A., Idroes, G. M., Maulydia, N. B., Patwekar, M., Suhendra, R., and Idroes, R. (2023). Integrating Genetic Algorithm and LightGBM for QSAR Modeling of Acetylcholinesterase Inhibitors in Alzheimer's Disease Drug Discovery, *Malacca Pharmaceutics*, Vol. 1, No. 2, 48–54. doi:10.60084/mp.v1i2.60.
23. Aler, R., Galván, I. M., Ruiz-Arias, J. A., and Gueymard, C. A. (2017). Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting, *Solar Energy*, Vol. 150, 558–569. doi:10.1016/j.solener.2017.05.018.
24. Abdullah-All-Tanvir, Ali Khandokar, I., Muzahidul Islam, A. K. M., Islam, S., and Shatabda, S. (2023). A gradient boosting classifier for purchase intention prediction of online shoppers, *Heliyon*, Vol. 9, No. 4, e15163. doi:10.1016/j.heliyon.2023.e15163.
25. Chen, T., and Guestrin, C. (2016). XGBoost, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, USA, 785–794. doi:10.1145/2939672.2939785.
26. Hong, W. S., Haimovich, A. D., and Taylor, R. A. (2018). Predicting hospital admission at emergency department triage using machine learning, *PLOS ONE*, Vol. 13, No. 7, e0201016. doi:10.1371/journal.pone.0201016.
27. Idroes, R., Noviandy, T., Maulana, A., Suhendra, R., Sasmita, N., Muslem, M., Idroes, G. M., Kemala, P., and Irvanizam, I. (2021). Application of Genetic Algorithm-Multiple Linear Regression and Artificial Neural Network Determinations for Prediction of Kovats Retention Index, *International Review on Modelling and Simulations (IREMOS)*, Vol. 14, No. 2, 137.
28. Han, J., Pei, J., and Kamber, M. (2011). *Data Mining: Concepts and Techniques*, Elsevier.
29. Idroes, G. M., Maulana, A., Suhendra, R., Lala, A., Karma, T., Kusumo, F., Hewindati, Y. T., and Noviandy, T. R. (2023). TeutongNet: A Fine-Tuned Deep Learning Model for Improved Forest Fire Detection, *Leuser Journal of Environmental Studies*, Vol. 1, No. 1, 1–8.