International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Development of a Grey Wolf Optimized-Gradient Boosted Decision Tree Metamodel for Heart Disease Prediction

Narayanan Ganesh[1], M. Balamurugan[2], Jasgurpreet Singh Chohan[3], Kanak Kalita[4,*]

Submitted: 01/10/2023 **Revised**: 27/11/2023 **Accepted**: 09/12/2023

Abstract: In this paper, a comprehensive study of various machine learning (ML) metamodels for heart disease detection is presented. The comparison includes conventional metamodels such as Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Trees, Random Forest as well as more advanced metamodels including Deep Learning, ML, Deep Neural Networks, Gradient Boosted Decision Trees and the proposed Grey Wolf Optimizer-Gradient Boosted Decision Trees (GWO-GBDT). The metamodels are assessed based on their performance in terms of accuracy, recall, precision, F1 measure and specificity. The results reveal that the developed GWO-GBDT metamodel outperforms the other metamodels in most metrics, offering superior prediction capabilities for heart disease diagnosis. This study provides a valuable reference for researchers and practitioners seeking efficient ML metamodels for heart disease prediction.

Keywords: *decision tree; heart disease; optimization; metamodel; machine learning; prediction;*

1. Introduction

The prevalence of heart disease has become a serious concern in global public health, resulting in significant morbidity and mortality rates worldwide. Timely prediction and diagnosis of ailments can significantly affect the treatment efficacy and reduce healthcare costs. Recent advancements in ML and optimization algorithms have opened new avenues for developing predictive models that can assist in the timely diagnosis and identification of diseases.

[1] *School of Computer Science & Engineering, Vellore Institute of Technology, Chennai 600 027, India;*

ganesh.narayanan@vit.ac.in;

[2]*Department of Computer Science, Kristu Jayanti College (Autonomous), Bengaluru, 560077, India; balamurugan@kristujayanti.com*

[3]*Department of Mechanical Engineering and University Centre for Research & Development, Chandigarh University, Mohali, 140413, India; jasgurpreet.me@cumail.in;*

[4]*Department of Mechanical Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi 600062, India; drkanakkalita@veltech.edu.in;*

**Correspondence: drkanakkalita@veltech.edu.in;*

Various ML algorithms and strategies have been explored in depth, ranging from decision tree (DT)-based methods [1] to k-Nearest Neighbor (KNN)-based systems [2], to hybrid approaches that combine multiple algorithms for enhanced performance [3] [4]. Ramesh et al. [5] found that, despite smaller datasets, ML algorithms displayed compelling performance in accurately categorizing cardiac diseases. Dewan and Sharma [1] utilized a DT-based approach using the C4.5 algorithm, emphasizing the reduction of attributes to improve accuracy. Anooj [2] proposed a system based on KNN, focusing on feature selection techniques for performance enhancement. Hybrid metamodels, combining the strengths of multiple ML algorithms, have also shown promise. Sharanyaa et al. [3] demonstrated the superior performance of a hybrid approach that combined Support Vector Machines (SVM) and Naïve Bayes. Similarly, ensemble-based methods have shown their effectiveness in heart disease prediction. Rajendran and Vincent [4] developed a metamodel that integrated several ML algorithms thereby leading to significant improvents in accuracy. This was echoed by Shorewala [6] who showed that KNN, random forest (RF) and SVM stacked models outperformed base classifiers.

Tiwari et al. [7] further bolstered these findings by achieving an accuracy of 92.34% using a stacked ensemble classifier with ExtraTrees Classifier, RF and XGBoost.

In the realm of big data analytics (BDA), Rehman et al. [8] discussed its pivotal role in healthcare and pointed out the challenges that lie ahead, including managing source information, ensuring security and improving data quality and analysis methods. Shaik et al. [9] hybridized a gradient boosted decision tree using a lesser known metaheuristics called squirrel search optimizer. The technological requirements for heart disease prediction systems were addressed by Chang et al. [10] and Nagavelli et al. [11]. Chang et al. [10] proposed using Python for its reliability and diverse data-tracking capabilities in medical research. Meanwhile, Nagavelli et al. [11] suggested the use of a mobile application to reduce complexity and computation time in heart disease detection, given the complexities of interpreting Magnetocardiography (MCG). The need for smart solutions, especially in regions where diagnostic procedures and preventative measures are lacking, was underscored by Ketu and Mishra [12]. They discussed the limitations of wired sensors for cardiac disease detection, underlining the requirement for improved sensor design and calibration.

Reddy et al. [13] conducted an exploration of various classifiers in idetifying heart disease and concluded that sequential minimal optimization classifier is the most accurate. On a similar note, Baccouche et al. [14] and Almulihi et al. [15] focused on the effectiveness of ensemble learning and deep learning (DL) stacking ensemble models respectively, achieving promising results, especially on unbalanced heart disease datasets. Yoon and Kang [16] presented a multimodal approach and Menshawi et al. [17] experimented with several combinationd of ML and DL techniques on different datasets.

In this paper, we present a hybrid approach that incorporates the Grey Wolf Optimizer (GWO) algorithm with the Gradient Boosted Decision Trees (GBDT) metamodel to predict the onset of heart disease. The GWO is a very popular nature-inspired optimization algorithm that draws inpiration from the leadership hierarchy and hunting mechanism of grey wolves in nature. This has been proven to be effective in navigating complex, multi-modal search spaces. On the other hand, GBDTs have gained widespread popularity in both academia and industry for their superior predictive performance in a variety of tasks, including but not limited to, classification and regression problems. The goal of this study is to explore the potential synergies between the GWO's robust optimization capabilities and the GBDT's strong predictive performance. Our approach aims to optimize the hyperparameters of the GBDT metamodel using GWO, thereby enhancing the metamodel's prediction accuracy for heart disease diagnosis. By advancing the fusion of GWO and GBDT for heart disease diagnosis, this research not only contributes to the early detection and prevention of heart disease but also provides a novel perspective on the application of bio-inspired algorithms in optimizing ML metamodels for healthcare predictions.

The rest of the paper is structured as follows: the methodology and experimental design is detialed in Section 2; Section 3 presents the results obtained and a detailed discussion. Section 4 conatins the concluding remarks on the work and proposes future research directions.

2. Methodology

2.1. Dataset

The dataset under consideration is a compilation of interconnected data entries, each represented by a specific report detailing the unique information it holds, along with an attribute for each component integrated within the dataset [18]. Data for this investigation were sourced from various locations including Cleveland, Switzerland, Long Beach and Hungary, supplemented with information retrieved from renowned data repositories, the UCI repository and Kaggle. The dataset encompasses 76 attributes, out of which 14 prove significantly instrumental in the diagnosis of heart disease. Typically, the attribute associated with the predictive class is positioned at the end of the list. A division of the dataset into 200 samples for training and 103 for testing facilitated the metamodel-building and validation process. More details on the dataset can be found in [9].

2.2. Preprocessing using Min-Max normalization

Min-max normalization is a widely adopted technique for data normalization. It facilitates data transformation by mapping each quantitative feature's outcome onto a target value derived from its minimum and maximum values. The utility of min-max normalization lies in its ability to scale data

within a range of 0 to 1, thereby achieving a standardized dataset. This uniform scaling significantly simplifies the process of data analysis. The transformation of data using min-max normalization is achieved through the application of Equation (1).

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x is the values in the original dataset, x_{min} and x_{max} are the minimal and maximum values of x, respectively.

2.3. Grey Wolf Optimizer (GWO)

GWO is a bio-inspired optimization algorithm developed based on the social hierarchy and hunting behaviour of grey wolves in nature. Grey wolves are known for their strategic hunting tactics and collaborative behaviour which involves encircling prey, pursuing it and approaching it, often driven by the social hierarchy among wolves. These behavioural aspects of grey wolves have been modelled into a search algorithm that can solve complex optimization problems. It has been applied in various fields, including machine learning, where it has been utilized for hyperparameter tuning of metamodels, feature selection and various other optimization tasks. The pseudocode of GWO is detailed below:

1. Initialization: Initialize the grey wolf population X_i $(i = 1,2,3,\ldots,n)$ where n is the number of wolves in the population. The fitness of each solution is calculated.
2. Wolves' Ranking and Leadership Determination: Rank the wolves based on their fitness score. The best three wolves are selected and named alpha (α), beta (β) and delta (δ) wolves. These wolves guide the hunt (optimization process).
3. Hunting Process: Update the position of the other wolves (omega wolves) based on the position of the leading wolves (α, β and δ) using the following equations:

$$A_1 = 2a * rand() - a \quad (2)$$

$$C_1 = 2 * rand() \quad (3)$$

$$C_\alpha = abs(C_1 * X_\alpha - X_i) \quad (4)$$

$$X_1 = X_\alpha - A_1 * D_\alpha \quad (5)$$

$$A_2 = 2a * rand() - a \quad (6)$$

$$C_2 = 2 * rand() \quad (7)$$

$$D_\beta = abs(C_2 * X_\beta - X_i) \quad (8)$$

$$X_2 = X_\beta - A_2 * D_\beta \quad (9)$$

$$A_3 = 2a * rand() - a \quad (10)$$

$$C_3 = 2 * rand() \quad (11)$$

$$D_\delta = abs(C_3 * X_\delta - X_i) \quad (12)$$

$$X_3 = X_\delta - A_3 * D_\delta \quad (13)$$

$$X_i = \frac{(X_1 + X_2 + X_3)}{3} \quad (14)$$

Here, $rand()$ is a random number between 0 and 1, a is a coefficient vector that decreases linearly from 2 to 0 throughout iterations, A and C are coefficient vectors, D_α, D_β and D_δ are the absolute distance between the α, β and δ wolves respectively and the current wolf, X_α, X_β, X_δ and X_i are the position vectors of the α, β, δ and the current wolf respectively.

4. Check Termination Criteria: The hunting process is continued until a termination criterion is met, which could be a maximum number of iterations, a minimum error requirement, or any other problem-specific criteria.

2.4. Gradient Boosted Decision Tree (GBDT)

GBDT is a powerful ML technique that involves an ensemble of decision trees. As an iterative algorithm, GBDT's strength lies in its ability to combine multiple weak learners (in this case, decision trees) to create a robust prediction model, minimizing the residuals of the previous tree by training the next tree. In a classification context,

such as heart disease diagnosis, GBDT metamodels can perform exceptionally well as they can handle a mix of categorical and numerical features, as well as missing data. The pseudocode of gradient Boosted Decision Trees is detailed below:

1. Initialization: Train an initial model to predict the labels. This can be a simple model like a single split decision tree. Compute the residuals between the actual labels and predicted labels.
   $F_0(x) = argmin\, \Sigma L(y_i, c)$ (15)
2. Loop over the number of iterations (M): For m in 1 to M:
3. Calculate pseudo-residuals: The negative gradient of the loss function is computed for every instance in the dataset.
   $r_i = -\frac{\partial\, L\,(y_i, F(x_i))}{\partial\, F(x_i)}$ (16)
4. Fit a weak learner: Fit a decision tree to the residuals computed above.
   $h_m(x) = argmin\, \Sigma[r_i - h_m(x_i)]^2$ (17)
5. Compute multiplier (γ): Line search is used to compute the multiplier that minimizes the loss function.
   $\gamma_m = argmin\, \Sigma L[y_i, F_{m-1}(x_i) + \gamma h_m(x_i)]$ (18)
6. Update the model: The final prediction model is updated using the decision tree from the current iteration.
   $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$ (19)
7. Output the final model:
   $F(x) = F_0(x) + \Sigma \gamma_m h_m(x)$ (20)

The algorithm starts with a weak model (typically a decision tree) and iteratively adds more decision trees that learn from the residuals (i.e., errors) of the previous trees. By continuously boosting the model's ability to predict based on the residuals, GBDT reduces bias and variance, producing a metamodel that can generalize well to unseen data.

GBDTs have been popular for their high flexibility, allowing for the optimization of differentiable loss functions, handling of mixed-type data and robustness against outliers in the output space. They also naturally handle missing data and have an inbuilt ability to model complex non-linear relationships.

3. Results and Discussion

The current GWO-GBDT results are compared with several other ML methods from the literature. Bharti et al. [19] had previously used this dataset to develop several solutions using ML methods wherein they use LR, KNN, SVM, RF, DT and DL. Solutions using ML frameworks as presented by Ko et al. [20] for this heart disease diagnosis problem whereas Miao and Miao [21] employed DNN. Solutions obtained with gradient descent optimization (GDO) [22] by Nawaz et al. are also compared with the current GWO-GBDT.

3.1. Accuracy

Accuracy is perhaps the most intuitive measurements of metamodel performance. It is defined as the ratio of the number of correct predictions made by the metamodel to the total number of predictions. From Figure 1, the GWO-GBDT method shows the highest accuracy of 94.37%. This indicates that this metamodel is capable of correctly classifying heart disease cases most accurately among all the metamodels studied. Other high-performing metamodels include the DL metamodel and the GDO, which have accuracies of 94.2% and 97.07% respectively. The ML method shows the lowest accuracy of 70%. This suggests that it may not be as reliable in predicting heart disease cases compared to the other metamodels.

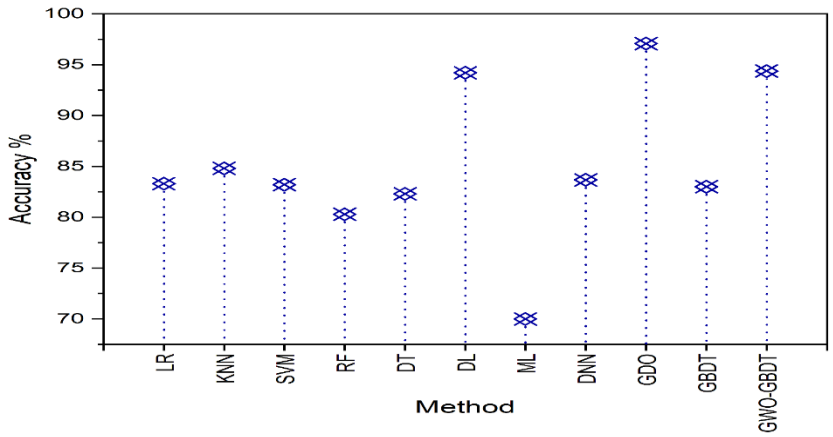

Fig 1. Accuracy of GWO-GBDT compared with other methods [19-22] in literature.

3.2. Recall

Recall which is also referred to as sensitivity or the true positive rate is the measurement of the number of actual positives that was correctly identified by the metamodel. From Figure 2, the GWO-GBDT method also performs well in terms of recall with 92.17%. This means that it can correctly identify a high percentage of heart disease cases out of the total actual heart disease cases. GDO again tops the chart with 97.15%, followed by the DNN metamodel with a recall of 93.51%. It's important to keep in mind that while DNN has lower accuracy than DL and GDO, its recall rate is higher, suggesting it may be a better metamodel for applications where missing a positive case has high costs.

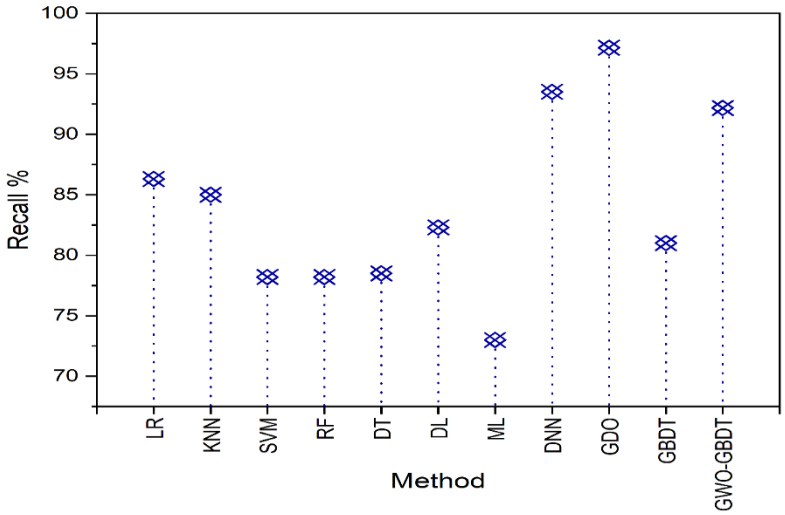

Fig 2. Recall of GWO-GBDT compared with other methods [19-22] in literature.

3.3. Precision

Precision is the measure of the correctness achieved in true prediction. It is calculated as the ratio of true positives achieved by the metamodel to total predicted positives by the metamodel. Among the compared literature, only a few have reported the precision value. From Figure 3, the GWO-GBDT method has the highest precision of 94.11%. This suggests that this metamodel has fewer false positives and is more reliable in its predictions of heart disease cases.

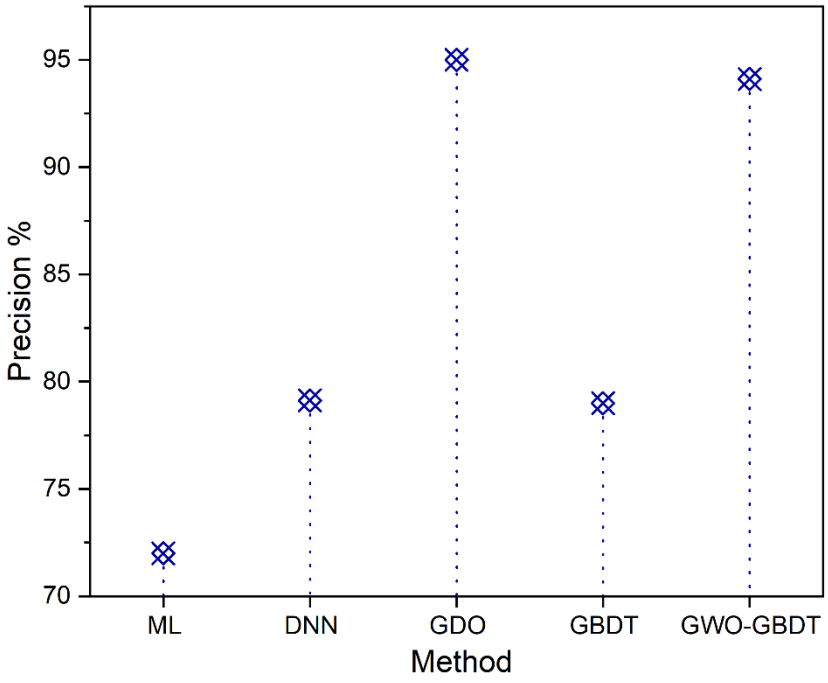

Fig 3. The precision of GWO-GBDT compared with other methods [19-22] in literature.

3.4. F1 score

The F1 metric serves as an indicator of a test's overall reliability, taking into account both its precision and recall for the final calculation. It essentially functions as a balanced mean of these two variables. As seen in Figure 4, the GWO-GBDT metamodel stands out with an F1 metric of 93.13%, signifying an optimal blend of both precision and recall.

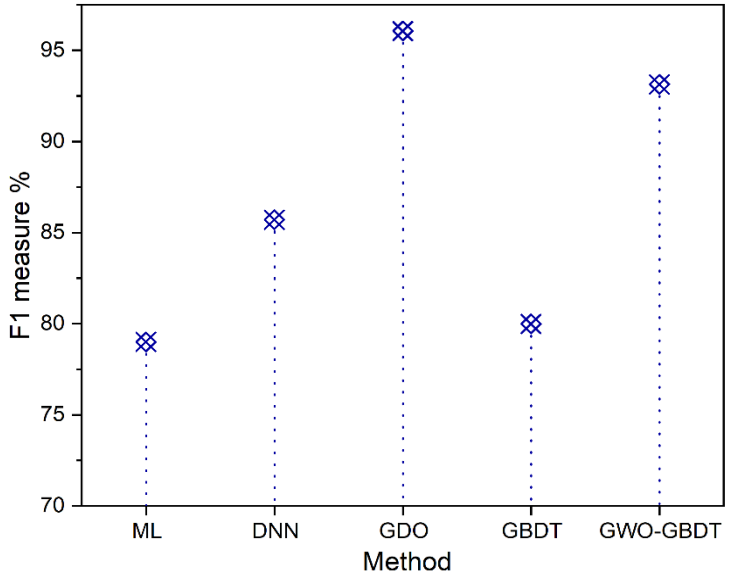

Fig 4. F1 score of GWO-GBDT compared with other methods [19-22] in literature.

3.5. Specificity

Specificity is also referred to as the true negative rate. It is the measurement of the proportion of actual negatives that are correctly identified by the metamodel. From Figure 5, GDO performs best with a specificity of 96.99%. The GWO-GBDT metamodel also shows high specificity at 90.59%. This means these metamodels are less likely to misclassify healthy patients as having heart disease, which is critical in a clinical context to avoid unnecessary treatments.

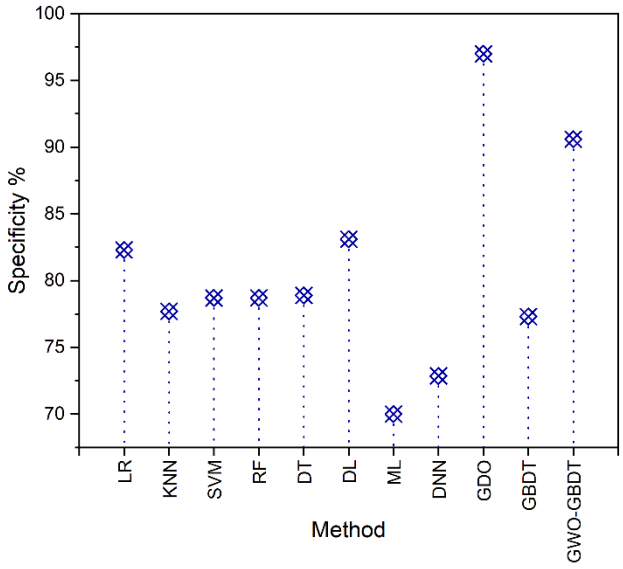

Fig 5. Specificity of GWO-GBDT compared with other methods [19-22] in literature.

Thus the GWO-GBDT metamodel seems to be a promising tool for heart disease prediction. It offers superior performance in terms of accuracy, recall, precision and F1 measure. Its specificity is also reasonably high, though not as high as that of the GDO. The metamodel provides a good balance between identifying positive cases (both in terms of

recall and precision) and correctly identifying negative cases (specificity).

4. Conclusions

This study presented a detailed comparison of different ML metamodels for the prediction of heart disease. It was observed that the GWO-GBDT metamodel showed superior performance, offering a promising tool for heart disease prediction. The GWO-GBDT metamodel demonstrated high accuracy, precision and recall, suggesting a reliable and balanced performance. Moreover, it exhibited a high F1 score, indicating a good balance between precision and recall. The metamodel's specificity was also reasonably high, reducing the likelihood of misclassifying healthy patients as having heart disease. The results highlight the potential of integrating optimization algorithms like GWO with ML metamodels like GBDT in the medical field, providing an effective approach to diagnosing heart disease and potentially saving lives.

References

[1] A. Dewan and M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," in 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015.

[2] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," Journal of King Saud University-Computer and Information Sciences, vol. 24, p. 27–40, 2012.

[3] S. Sharanyaa, S. Lavanya, M. R. Chandhini, R. Bharathi and K. Madhulekha, "Hybrid Machine Learning Techniques for Heart Disease Prediction," International Journal of Advanced Engineering Research and Science, vol. 7, p. 44–8, 2020.

[4] N. A. Rajendran and D. R. Vincent, "Heart disease prediction system using ensemble of machine learning algorithms," Recent Patents on Engineering, vol. 15, p. 130–139, 2021.

[5] R. Tr, U. K. Lilhore, Poongodi, S. Simaiya, A. Kaur and M. Hamdi, "Predictive analysis of heart diseases with Machine Learning approaches," Malays. J. Comput. Sci., p. 132–148, March 2022.

[6] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," Informatics in Medicine Unlocked, vol. 26, p. 100655, 2021.

[7] A. Tiwari, A. Chugh and A. Sharma, "Ensemble framework for cardiovascular disease prediction," Computers in Biology and Medicine, vol. 146, p. 105624, 2022.

[8] A. Rehman, S. Naz and I. Razzak, "Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities," Multimed. Syst., vol. 28, p. 1339–1371, August 2022.

[9] K. Shaik, J. Ramesh, M. Mahdal, M. Rahman, S. Khasim and K. Kalita, "Big Data Analytics Framework Using Squirrel Search Optimized Gradient Boosted Decision Tree for Heart Disease Diagnosis," Applied Sciences , vol. 13, no. 9, p. 5236, 2023.

[10] V. Chang, V. R. Bhavani, A. Q. Xu and M. A. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," Healthcare Analytics, vol. 2, p. 100016, November 2022.

[11] U. Nagavelli, D. Samanta and P. Chakraborty, "Machine learning technology-based heart disease detection models," J. Healthc. Eng., vol. 2022, p. 7351061, February 2022.

[12] S. Ketu and P. K. Mishra, "Empirical analysis of machine learning algorithms on imbalance electrocardiogram based arrhythmia dataset for heart disease detection," Arabian Journal for Science and Engineering, p. 1–23, 2022.

[13] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua and S. Pranavanand, "Heart disease risk prediction using machine learning classifiers with attribute evaluators," Applied Sciences, vol. 11, p. 8352, 2021.

[14] A. Baccouche, B. Garcia-Zapirain, C. Castillo Olea and A. Elmaghraby, "Ensemble deep learning models for heart disease classification: A case study from Mexico," Information, vol. 11, p. 207, 2020.

[15] A. Almulihi, H. Saleh, A. M. Hussien, S. Mostafa, S. El-Sappagh, K. Alnowaiser, A. A. Ali and M. Refaat Hassan, "Ensemble Learning Based on Hybrid Deep Learning Model for Heart Disease Early Prediction," Diagnostics, vol. 12, p. 3215, 2022.

[16] T. Yoon and D. Kang, "Multi-Modal Stacking Ensemble for the Diagnosis of Cardiovascular Diseases," Journal of Personalized Medicine, vol. 13, p. 373, 2023.

[17] A. Menshawi, M. M. Hassan, N. Allheeib and G. Fortino, "A Hybrid Generic Framework for Heart Problem Diagnosis Based on a Machine Learning Paradigm," Sensors, vol. 23, p. 1392, 2023.

[18] D. Cenitta, R. Vijaya Arjunan and K. V. Prema, "Ischemic heart disease prediction using optimized squirrel search feature selection algorithm," IEEE Access, vol. 10, p. 122995–123006, 2022.

[19] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," Comput. Intell. Neurosci., vol. 2021, p. 8387680, July 2021.

[20] Y.-F. Ko, P.-H. Kuo, C.-F. Wang, Y.-J. Chen, P.-C. Chuang, S.-Z. Li, B.-W. Chen, F.-C. Yang, Y.-C. Lo, Y. Yang and others, "Quantification analysis of sleep based on smartwatch sensors for Parkinson's disease," Biosensors, vol. 12, p. 74, 2022.

[21] K. H. Miao and J. H. Miao, "Coronary heart disease diagnosis using deep neural networks," international journal of advanced computer science and applications, vol. 9, 2018.

[22] M. S. Nawaz, B. Shoaib and M. A. Ashraf, "Intelligent Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization," Heliyon, vol. 7, p. e06948, May 2021.