applied sciences

MDPI

Article

Big Data Analytics Framework Using Squirrel Search Optimized Gradient Boosted Decision Tree for Heart Disease Diagnosis

Kareemulla Shaik [1], Janjhyam Venkata Naga Ramesh [2], Miroslav Mahdal [3], Mohammad Zia Ur Rahman [4], Syed Khasim [1] and Kanak Kalita [5,*]

[1] School of Computer Science & Engineering, VIT-AP University, Amaravati 522237, India
[2] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram 522302, India
[3] Department of Control Systems and Instrumentation, Faculty of Mechanical Engineering, VSB-Technical University of Ostrava, 17. Listopadu 2172/15, 70800 Ostrava, Czech Republic
[4] Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Guntur 522302, India
[5] Department of Mechanical Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi 600062, India
* Correspondence: drkanakkalita@veltech.edu.in

Abstract: Disease detection is a critical issue in the field of medical diagnostics. Failure to identify heart disease (HD) at an early stage can lead to mortality. The lack of access to expert physicians in remote areas further exacerbates the problem. Big data analytics (BDA) is an emerging area that can help extract valuable information from vast amounts of data and improve medical diagnosis while reducing costs. Machine learning (ML) algorithms have been effectively employed in many fields, including medical diagnostics. ML methods can help doctors detect and forecast illnesses at an early stage by creating classifier systems. In this article, we propose a unique ML- and BDA-based squirrel search-optimized Gradient Boosted Decision Tree (SS-GBDT) for the detection of heart disease. The effectiveness of the proposed method is demonstrated through various performance indicators. The results show that the proposed method is highly efficient in medical diagnosis, with 95% accuracy rate, 95.8% precision, 96.8% recall and 96.3% F1-measure achieved by the SS-GBDT. The use of BDA and ML can greatly enhance medical diagnosis and this proposed method is a significant step in this direction.

Keywords: medical diagnosis; disease detection; big data analytics; machine learning; heart disease; optimization

check for updates

Citation: Shaik, K.; Ramesh, J.V.N.; Mahdal, M.; Rahman, M.Z.U.; Khasim, S.; Kalita, K. Big Data Analytics Framework Using Squirrel Search Optimized Gradient Boosted Decision Tree for Heart Disease Diagnosis. *Appl. Sci.* **2023**, *13*, 5236. https://doi.org/10.3390/app13095236

Academic Editor: Luis Javier Garcia Villalba

Received: 1 March 2023
Revised: 18 April 2023
Accepted: 20 April 2023
Published: 22 April 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Heart disease (HD) is one of the major issues facing present society and is among the leading causes of death globally. Arterial HD, cerebral disease, radial arterial disease, rheumatic HD and fetal HD are some of the most well-known types of HD. The World Health Organization (WHO) estimates that 17.9 million people worldwide lose their lives to heart disease and its consequences each year. Heart attacks and strokes account for more than four out of every five HD fatalities. Heart attacks are caused when blood supply to the heart is blocked, whereas strokes are caused by interruption of blood supply to the brain. Both these conditions are life-threatening and need immediate medical attention. Poor diets, inadequate fitness, alcohol consumption and smoking are a few of the major risk factors that can accelerate heart-related problems. However, unforeseen and premature deaths can be prevented by identifying people at higher risk of HD earlier on and providing the appropriate therapies [1].

Diagnosing diseases is the most crucial aspect of providing medical treatment. It is possible to save human lives by diagnosing a disease earlier than the typical or expected amount of time. Nowadays, heart disease is a prevalent ailment that causes the death of a significant number of people and also shortens a person's lifespan. The functioning of the heart is essential to life, as the heart is a necessary component of the human body; without it, life would not be possible. Heart disease disrupts normal heart function and may either directly result in death or make the patient's last days more uncomfortable. One of the most crucial aspects of HD is determining the likelihood of a particular individual suffering from insufficient blood supply to the heart [2].

The application of big data in the medical sector is expanding globally and there is no denying its promise and advantages. Large amounts of data can be used in machine learning (ML) and big data analytics (BDA). These databases can be utilized to improve detection, guide preventive medical procedures, and lessen the negative effects of medication and other forms of therapy. Big data's effects can be seen in a range of healthcare contexts and sectors, including emergency rooms, critical care, cardiac illnesses, mental well-being and pneumonia. As analytics assist or enable judgments that are essential to health and liberty, they increasingly govern human life [3].

In recent years, ML-based techniques have shown great potential in diagnosis and prediction of various medical conditions among, which heart diseases, tumors and cancers have received the most attention. The electrocardiogram, also known as an ECG, is used to identify cardiovascular disease. However, visually diagnosing long-term ECG irregularities requires a significant amount of time and work. Since the emergence of ML uses in the medical field, many academics and operators have discovered that machine learning-based cardiac disease detection systems are cost-effective and versatile ways. In comparison to the previously conducted investigation, the risk assessment results from ML-based algorithms have been characterized as being superior and more encouraging [4].

The analysis of large amounts of data is recognized as one of the most rapidly developing techniques in the world. It offers a wide variety of medical uses, such as the distribution of medical personnel and resources, remote health surveillance, detection, illness prognosis at earlier phases, critical treatment solutions and care for the elderly, among many others. The use of BDA enables a more comprehensive examination of an individual's health history, including medication reminders and potential consequences. Additionally, it provides information on past treatments and assists in ongoing illness management. The development of systems that can effectively acquire, store and analyze large amounts of data is facilitated by advancements in data administration and analytics [5].

However, there is still significant scope of improvement in terms of accuracy, efficiency and applicability of ML based techniques in real-world scenarios. In this study, a novel big data analytics framework that fuses a squirrel search (SS) optimization algorithm with Gradient Boosted Decision Trees is used for early diagnosis of heart disease. The proposed framework aims to overcome the limitations of existing methods by leveraging the iterative improvement ability of squirrel search optimization and metamodeling ability of Gradient Boosted Decision Trees to achieve better classification performance. The main contributions of this paper are as follows:

- Development of a novel squirrel search-optimized Gradient Boosted Decision Tree (SS-GBDT) framework for the early diagnosis of heart disease;
- Development of comprehensive methodology that includes data preprocessing, feature extraction using word2Vec and classification using SS-GBDT;
- Validation of the proposed SS-GBDT method using various performance indicators and comparing it with other state-of-the-art ML-based techniques, thereby highlighting the superiority and applicability of the proposed approach.

The article is structured as follows: Section 2 provides a literature review where the gaps in the existing literature are explored. Section 3 presents the proposed methodology. The discussions on the dataset, preprocessing, feature extraction and squirrel search-optimized Gradient Boosted Decision Tree classifier are included in this section.

Section 4 reports the performance analysis and compares the proposed SS-GBDT with several state-of-the-art methods. Section 5 presents the conclusion, succinctly summarizing the key findings, limitations and future research directions.

2. Literature Survey

In recent years, ML and BDA for diagnosing heart disease have garnered significant interest. Various studies have proposed numerous methodologies, each with their own merits and limitations. In this section, some of the most relevant pieces of literature are reviewed to identify the research gaps which the proposed SS-GBDT aims to address.

Ramesh et al. [6] worked on four alternative strategies for performing comparison evaluations and attaining favorable performance. In their study, they concluded that ML approaches performed much better than statistical methods. From the investigations of many researchers, they demonstrated that the use of ML models to predict and categorize cardiac disease is the best option, even with a smaller database. Chang et al. [7] developed a Python-based application for medical research since it is more reliable, helps track data and makes it easier to construct different types of health monitoring applications. However, there were not many parameters used in that study. This suggests that more comprehensive methods might yield better results in heart disease diagnosis.

Rehman et al. [8], in their review article, have spoken about BDA approaches, tools, methods and structures in the healthcare industry. Due to the vast volume of data, BDA is essential to healthcare and biomedicine. Although the field has incredible promise, there are some significant problems. These include numerous source information management problems, as well as assuring security, shielding productivity, setting up models and management, improving analyzing approaches and data quality. Nagavelli et al. [9] used four ML modeling strategies for the identification of heart disease. Their article provides an overview of several ML-based approaches for heart disease identification. Since MCG interpretation is time-consuming, heavily reliant on interpreting expertise and has little appeal in clinics despite its great signal quality, the authors carried it out by using a mobile application with less complexity and computation time. Ketu and Mishra [10] argue that solutions are essential for developing smart robotic solutions as well as for reducing the effects of illnesses via wise decision-making. Poor diagnostic procedures, insufficient medical personnel, ineffective medical assistance, inadequate preventative measures and technological improvements have had a significant negative influence on emerging nations. Wired sensors for cardiac disease have more complications. The research presented showed that the smart sensor's design necessity by which it must use pre-programmed integrated functionality causes insufficient detection. Sensor calibration must be managed by an external microcontroller.

Anooj [11] proposed a k-nearest-neighbor-based clinical decision support system and focused on feature selection techniques to enhance performance. Dewan and Sharma [12] employed a decision tree-based approach for heart disease prediction, focusing on reducing the number of attributes to improve accuracy and efficiency. The authors used the C4.5 algorithm for classification and found that it performed well in comparison to other popular ML algorithms. Sharanyaa et al. [13] explored the use of a hybrid ML approach for heart disease prediction, combining the strengths of both Naïve Bayes and Support Vector Machines (SVM). Their findings demonstrated that a hybrid approach could lead to better performance than individual algorithms alone.

Rajendran and Vincent [14] developed an ensemble-based heart disease prediction system that combined multiple ML algorithms and improved prediction accuracy. Shorewala [15] compared base classifiers and ensemble techniques, showing that stacked models involving KNN, random forest and SVM achieved the highest accuracy. Tiwari et al. [16] proposed a stacked ensemble classifier using ML algorithms such as extra trees Classifier, random forest, and XGBoost for heart disease prediction. They achieved an accuracy of 92.34%.

Yoon and Kang [17] presented a multi-modal stacking ensemble approach using ResNet-50 and logistic regression for diagnosing CVDs from 12-lead ECG data. This method combined scalogram and ECG grayscale images and outperforms LSTM, BiLSTM, individual base learners, simple averaging ensemble and single-modal stacking ensemble methods in various metrics. Menshawi et al. [18] proposed a hybrid framework that combined multiple ML and deep learning techniques, which produced unbiased predictions and was adaptable to different datasets.

Reddy et al. [19] evaluated ten ML classifiers for heart disease risk prediction and found that the sequential minimal optimization classifier achieved the highest accuracy. Baccouche et al. [20] proposed an ensemble-learning framework combining deep neural network models and random under-sampling for classifying unbalanced heart disease datasets, achieving about 92% accuracy. Almulihi et al. [21] proposed a deep stacking ensemble model that outperformed five machine learning and hybrid models on two heart disease datasets.

Thus, it is evident that the literature on heart disease diagnosis using ML and BDA is extensive and diverse. It is seen that various studies have explored different algorithms and methodologies for this diagnostic problem. The current proposed squirrel search-optimized Gradient Boosted Decision Tree (SS-GBDT) framework aims to contribute to this field by addressing the gaps and limitations of existing approaches and offering a more accurate, efficient and applicable solution for heart disease diagnosis.

3. Research Methodology

In this section, the proposed methodology is described. The datasets were collected from hospitals and data preprocessing was performed using Min–Max normalization. Big data analysis was conducted on the dataset and HD diseases were detected using the ML-based SS-GBDT. The suggested model flow is depicted in Figure 1.

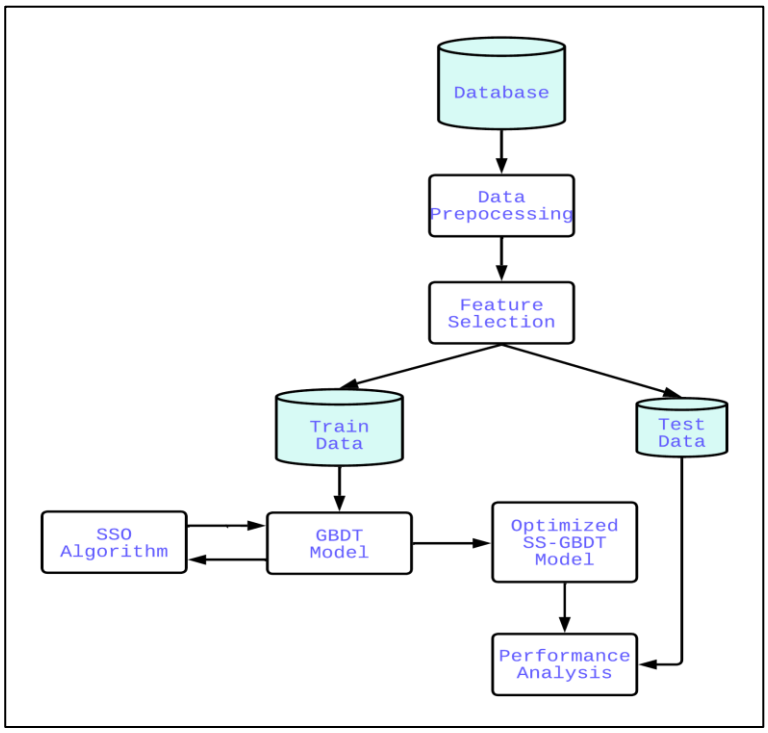

Figure 1. Proposed model flow.

3.1. Dataset

The dataset is a collection of linked data that consists of a report for each instance that is specific to the data that it represents, as well as an attribute for every attribute that is included in the dataset [22]. The data used in this study were gathered from Cleveland,

Switzerland, Long Beach and Hungary, in addition to information obtained from the UCI repository and Kaggle for data analysis. The diagnosis of cardiac disease can be greatly aided by 14 of the 76 features that are included in the dataset. In most cases, the predictive class feature is stated at the very end of the list. In our study, 200 and 103 samples are used as training and testing data, respectively. The parameters of the dataset that describe the features are represented in Table 1.

Table 1. Characteristics of heart disease.

Feature	Symbol	Value
Age	age	29 to 77
Sex	sex	Female (0) Male (1)
Chest Pain	cp	Angina (typical) (1) Angina (atypical) (2) Non-anginal (3) Asymptomatic (4)
Resting blood sugar	trestbps	94 to 200 mm Hg
Serum cholesterol	chol	126 to 564 mg/dL
Fasting blood sugar	fbs	<120 mg/dL (0) >120 mg/dL (1)
Resting ECG result	restecg	Normal (0) ST-T wave abnormality (1) LV-hypertrophy (2)
The highest possible heart rate attained	thalach	71 to 202
Angina caused by exercise	exang	No (0) Yes (1)
ST depression induced by exercise relative to rest	oldpeak	0 to 6.2
Peak workout ST segment slope	slope	Upsloping (1) Flat (2) Down sloping (3)
Major vessels colored by fluoroscopy	ca	0–3
Defect type	thal	Normal (3) Fixed defect (6) Reversible defect (7)
Heart disease	target	0–4

3.2. Preprocessing Using Min–Max Normalization

This is one of the most commonly used methods for data normalization. The method used for data transformation is Min–Max normalization, which transforms the result of each quantitative characteristic into a goal value based on the minimum and maximum values. Min–Max normalization is a useful tool for data normalization as it scales the data between 0 and 1. This uniformity makes it easier for us to analyze the data. Equation (1) is used to perform data transformation:

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

where x is a group of the projected values shown in the data collection, and x_{min} and x_{max} are the minimal and maximum values of x, respectively.

3.3. Feature Extraction Using Word2vec

Word2vec (W2V) has been extensively used in both conventional and deep learning investigations. The procedure involves two different models: continuous bag-of-words (CBOW) and skip-gram, both of which are basic multiple-layer perceptrons. W2V preprocesses training corpora by splitting them into text windows of a predetermined size and initializing word embeddings for the appropriate vocabulary at random. During training, CBOW predicts the central word from context words given a text window, while skip-gram

predicts the central word from the center word. Cross-entropy is used to measure training loss and word embeddings are progressively changed during backpropagation. After considerable training, the embeddings usually converge and are prepared for downstream tasks. In contrast to bag-of-words, which solely makes use of frequency data, W2V can extract rich, real-valued and low-dimensional abstract semantic and grammatical properties.

3.4. Classification Using Squirrel Search-Optimized Gradient Boosted Decision Tree

Squirrel search optimization is a population-based method in which each squirrel explores a multivariate search space while foraging for food. The positions of the squirrels are treated as different design variables and the distance between the food and the squirrel individual is analogous to the fitness value (FV) of the objective function. Individual squirrels in SS move to new, potentially better locations. The optimization process of SS based on the foraging behavior of flying squirrels (FLS) can be hypothetically described in the following stages:

Stage 1 (Initialize the Variables): The total number of iterations ($Iter$), population size (N_P), the total number of decision variables (N_D), the likelihood that a predator would be present (P_{PP})), the scaling factor (F_s), the gliding constant (C_g) and the upper and lower bounds for the decision variables (U_D) and (L_D). At the outset of the squirrel search optimization procedure, certain decisions are made.

Stage 2: Initializing flying squirrels randomly, the starting point for squirrel search optimization, as in other population-based algorithm, is a haphazard position for flying squirrels. There are a certain number of flying squirrels (FLS) in a forest and their locations may be determined. A consistent distribution is used to establish each flying squirrels' starting location inside the forest. The FLS coordinates are initialized at random in the search process as follows in Equation (2):

$$FLS_{id} = L_D + rand() * (U_D - L_D) \qquad i = 1, 2, \ldots, N_P;\ j = 1, 2, \ldots N_D, \tag{2}$$

where $rand()$ provides a random integer in the range [0, 1] with a uniform distribution.

Stage 3 (Fitness Evaluation): By inputting the decision variable's values into a user-defined FF and calculating the associated values, each FLS's fitness is assessed. The sort of food supplies an FLS is seeking—whether an ideal, typical or nonexistent one—and, therefore, its chances of survival, are indicated by the FV of that FLS's location. The location of a flying squirrel's FV $f = f_1, f_2, \ldots\ f_{N_P}$ is evaluated by inserting the number of choice variables into an FF and is determined in Equation (3):

$$f_i = f_i(FLS_{i,1,}FLS_{i,2,} \ldots\ldots FLS_{i,n}) \qquad i = 1, 2, \ldots, N_P. \tag{3}$$

Stage 4: Statement, organizing and Random collection:

The list is sorted ascendingly after saving the FVs for each FLS point. The hickory nut tree is home to an FLS with a subpar FV. The three best FLS after that are believed to be on acorn nut trees and, subsequently, to move to trees that produce hickory nuts. The FLS that are still there need to be on typical trees. Given that they have likely satisfied their daily caloric requirements, some squirrels are believed to be randomly choosing to go to the hickory nut tree. The squirrels who make it to the acorn nut trees will do so. Predators constantly influence the FLL's foraging behavior. We sort the index from highest to lowest. The standard of the feed ingredients is then ranked in order of increasing FV according to the FLS locations in Equation (4):

$$\left[Sorted_f,\ sort_index\right] = sort(f). \tag{4}$$

Step 5: Use aerodynamic gliding to create new positions:

For every case, it is hypothesized that while there is no predator around, the FLS glides and efficiently explores the tree for its favorite food, but when there is a predator there, it is

compelled to do a small random walk search of a neighboring secret place. Gliding to a new site involves Equation (5):

$$FLS_{at}^{new} = \begin{cases} FLS_{at}^{old} + D_g C_g \left(FLS_{ht}^{old} - FLS_{at}^{old} \right), & if\ R_1 \geq P_{PP} \\ random\ location, & otherwise \end{cases}. \quad (5)$$

The terms R_1, C_g and D_g refer to functions that return values from the equal distribution on the range [0, 1], a C_g and a random gliding distance, respectively. Data from the equal distribution on the range [0, 1] are returned by the function R_1. Aerodynamic gliding, or with velocity or random values, is used to estimate the new location. We limit the lower and upper boundaries of the new locations when moving them.

Stage 6: Ending criteria:

A common convergence criterion called the function tolerance criterion allows for a negligible and allowable discrepancy between the final two outputs. There are times when the longest execution period is used as a pause condition. The majority of iterations are employed in this experiment as a halting constraint. Squirrel search optimization is represented by Algorithm 1.

Algorithm 1: Squirrel Search (SS)

Input: Set the majority locations at random starting points about the lower and higher bound limits.
Result: Highly integrated solution.
Stage 1: Produce a random position for n flying squirrels.
Stage 2: Evaluate the FV for the supplied feature value for N samples based on the k-neighbors and error rate.
Stage 3: According to their FV, arrange the flying squirrel sites in increasing order.
Stage 4: Create new regions by gliding arbitrary regions,
Else

$$FLS_{at}^{new} = \begin{cases} FLS_{at}^{old} + D_g C_g \left(FLS_{ht}^{old} - FLS_{at}^{old} \right), & if\ R_1 \geq P_{PP} \\ random\ location, & otherwise \end{cases}$$

Stage 5: For the most iterations possible, repeat stages 1 through 4.

The GBDT model, initially introduced by Friedman, is an ensemble model that combines weak classifiers to form a strong model. Unlike other similar methods, GBDT employs function space for optimization, making it more adaptable and scalable for non-linear situations. The model uses gradient boosting, which involves optimizing the loss function, predicting from weak learners and reducing the loss function by adding weak learners. The decision trees used in the model act as weak learners and are added in a sequential manner, with each new tree minimizing the residual loss from the previous trees. The model's convergence is achieved by following the negative gradient's direction, rather than using weighted data as in traditional boosting methods. The GBDT model can efficiently describe non-linear decision boundaries, as shown in Figure 2, due to its hierarchical nature.

Generally, the model uses the gradient descent approach to minimize loss and prevent over-fitting, with the learning rate determining the step size used to mix the weights of different trees. The minimum loss reduction required for a subsequent partition on a leaf node is represented by γ_m.

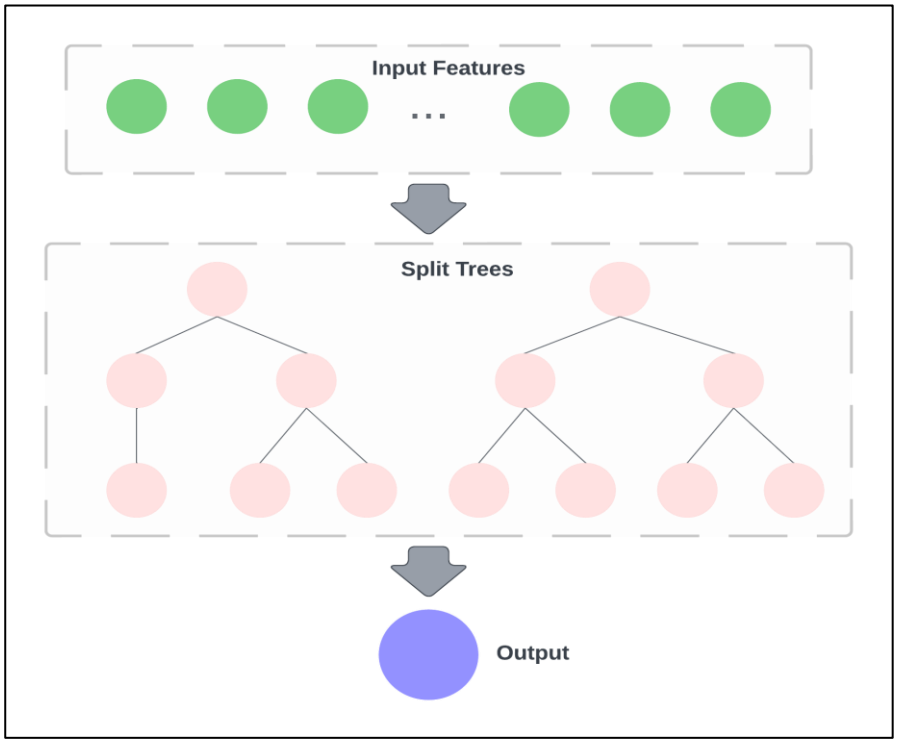

Figure 2. Example of GBDT.

Stage 1: The model's starting constant value is supplied.

$$f_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma) \tag{6}$$

Stage 2 decides how many iterations there will be; m = 1 to M.
Stage 2.1: Based on Equation (7), the step size and minimal loss reduction for averaging the weights of different trees may be determined as follows:

$$(\gamma m, \eta m) = argmin_{\gamma,\eta} \sum_{i=1}^{N} L(y_{i,} f_{m-1}(x_{k,i}) + \eta b(x_{k,i}; \gamma)) + vT + \frac{1}{2}\beta\|\gamma\|^2, \tag{7}$$

where T stands for the tree's leaf count. It must be noticed that the failure function $L()$ uses training data to calculate the model fitness and uses the word $\eta b(x_{k,1}; \gamma)$ to describe the model complexity. Additionally, the term $vT + \frac{1}{2}\beta\|\gamma\|^2$ penalizes the model's difficulty.
Step 2.2 involves updating the method as follows:

$$F_m(x) = F_{m-1}(x) + \eta m^b \, \eta b(x_{k,1}; \gamma). \tag{8}$$

Step 3, after applying M additive functions to produce the result, returns $F_m(x)$.

To determine data X, GBDT employs M additive functions as shown in the following equation to obtain the desired result:

$$\hat{y}_{GBCT} = \sum_{i=1}^{M} \eta m^b(x_{k,1}; \gamma). \tag{9}$$

In a GBDT model, each successive tree predicts the pseudo-residuals of the previous trees, given an arbitrary differentiable loss function. The user defines both the loss function and the function that calculates the associated negative gradient. The loss function is minimized by combining the predictions and training additional trees. The number of trees is a critical parameter in gradient boosting, as too few or too many trees can result in underfitting or overfitting, respectively.

The pseudocode of the proposed squirrel search-Optimized Gradient Boosted Decision Tree (SS-GBDT) is presented in Algorithm 2.

Algorithm 2: Squirrel Search-Optimized Gradient Boosted Decision Tree (SS-GBDT)

1. Initialize variables:

Iterations ($Iter$), Population size (N_P), Decision variables (N_D), Likelihood of predator presence (P_{PP}), Scaling factor (F_s), Gliding constant (C_g), Upper and lower bounds for decision variables (U_D, L_D)

2. Initialize Flying Squirrels (FLS) randomly:

for i in range $(1,\ N_P + 1)$:

for j in range $(1, N_D + 1)$:

$FLS_{id} = L_D + rand() * (U_D - L_D)$

3. Evaluate Fitness (FV) for each Flying Squirrel (FLS):

for i in range $(1,\ N_P + 1)$:

$f_i = f_i\left(FLS_{i,1}, FLS_{i,2}, \ldots\ldots FLS_{i,n}\right)$

4. Sort FLS positions based on their FV:

$\left[Sorted_f,\ sort_index\right] = sort(f)$

5. Create new positions by aerodynamic gliding or random walk:

for each FLS:

if $R_1 \geq P_{PP}$:

$FLS_{at}^{new} = FLS_{at}^{old} + D_g C_g\left(FLS_{ht}^{old} - FLS_{at}^{old}\right)$

else:

FLS_{at}^{new} = random location

Limit the new positions to the lower and upper bounds

6. Repeat steps 2–5 for the maximum number of iterations ($Iter$)
7. Initialize GBDT model with the optimized parameters obtained from squirrel search
8. Train GBDT model with training data:

Stage 1: Set the initial constant value for the model

Stage 2: For m in range $(1,\ M + 1)$:

2.1: Determine step size and minimal loss reduction for averaging the weights of different trees

2.2: Update the model

Stage 3: Return the final model $F_m(x)$

9. Use GBDT model for prediction and classification

4. Results and Discussion

As discussed earlier, this paper proposes a novel approach, named the squirrel search-optimized gradient-boosted decision tree (SS-GBDT), which is based on machine learning (ML) and big data analytics (BDA). As such, it is important to validate the performance of the proposed SS-GBDT against the state of the art. For comparison with existing literature, several performance measures such as accuracy, precision, etc., are reported in this section.

The ratio of all forecasts to corrected predictions Is what Equation (10), which calculates accuracy, is defined as:

$$Accuracy = \frac{(true\ negative + true\ positive)}{(true\ positive + false\ positive + true\ negative + false\ negative)}. \quad (10)$$

The precision may be determined by solving Equation (11), which is shown below.

$$Precision = \frac{true\ positive}{(true\ positive + false\ positive)}. \quad (11)$$

Calculating the recall is possible due to Equation (12), which can be found here:

$$Recall = \frac{true\ positive}{(true\ positive + false\ negative)}. \quad (12)$$

Similarly, the F-measure can be calculated based on Equation (13):

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (13)$$

Based on the above performance metrics, the current SS-GBDT is compared with various machine learning methods that have used by other researchers to solve this problem. For example, Bharti et al. [23] employed several ML methods such as logistic regression (LR), K-nearest neighbor (KNN), support vector machine (SVM), random forest (RF), decision tree (DT) and deep learning (DL). Ko et al. [24] also used ML frameworks for the problem. Deep neural networks were employed by Miao and Miao [25]. On the other hand, dedicated optimization algorithms such as gradient descent optimization (GDO) [26], genetic algorithm (GA) [27] and swarm optimized artificial neural networks [28] were used by other researchers. Kasbe and Pippals [29] developed a fuzzy expert system for the problem. Multiple cardiac sensors for the management of heart failure (MANAGE-HF) [30] and body sensor network (BSN) [31] results are also compared. The modified self-adaptive Bayesian algorithm (MSABA), reported by Subahi et al. [32], is also used for comparison. The comparison of the SS-GBDT with the state of the art is reported in Table 2.

Table 2. Comparison of SS-GBDT with the state-of-the-art.

Method	Source	Accuracy %	Recall %	Precision %	F1-Measure %
LR	Bharti et al. [23]	83.3	86.3	-	-
KNN	Bharti et al. [23]	84.8	85	-	-
SVM	Bharti et al. [23]	83.2	78.2	-	-
RF	Bharti et al. [23]	80.3	78.2	-	-
DT	Bharti et al. [23]	82.3	78.5	-	-
DL	Bharti et al. [23]	94.2	82.3	-	-
ML	Ko et al. [24]	70	73	72	79
DNN	Miao and Miao [25]	83.67	93.51	79.12	85.71
GDO	Nawaz et al. [26]	97.07	97.15	-	-
GA	Eisa and Alnaggar [27]	64	66	65	70
Swarm-ANN	Nandy et al. [28]	95.78	95.21	95.21	95.21
Fuzzy Expert System	Kasbe and Pippals [29]	94.5	-	-	-
MANAGE-HF	Hernandez et al. [30]	56	60	58	63
BSN	Shakya and Joby [31]	82	85	84	88
MSABA	Subahi et al. [32]	90	91	92	95
GBDT	Current paper	83	81	79	79.99
SS-GBDT	Current paper	95	96.8	95.8	96.3

Figure 3 illustrates the accuracy of the proposed SS-GBDT with respect the results of the state of the art. Accuracy is the by far most commonly reported metric in the literature. When compared with the conventional ML methods reported by Bharti et al. [23], the current SS-GBDT is found to be 11.7%, 10.2%, 11.8%, 14.7% and 12.7% better than LR, KNN, SVM, RF and DT, respectively. DL reported by Bharti et al. [23] is found to be on par with the current SS-GBDT, with the current method being about 0.8% better. Similarly, with respect to the ML framework reported by Ko et al. [24], the SS-GBDT is 25% better. With respect to DNN [25] and MANAGE-HF [30], the SS-GBDT is 11.33% and 39% superior. However, the dedicated optimization algorithm-tuned ML results i.e., GDO [26] and swarm-optimized artificial neural networks [28], are found to have 2% and 0.78% better classification accuracy than the current SS-GBDT. Nevertheless, it is quite evident from the comprehensive comparison that the current SS-GBDT is capable of achieving very high rates of classification accuracy, which is a prerequisite in critical applications such as healthcare.

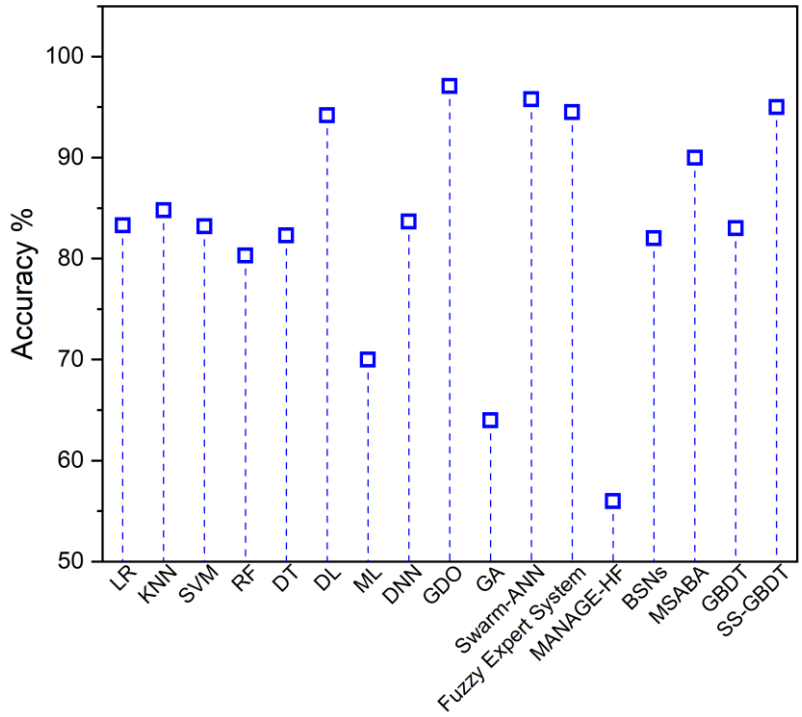

Figure 3. Comparison of accuracy of the proposed SS-GBDT and the state of the art, compared with LR [23], KNN [23], SVM [23], RF [23], DT [23], DL [23], ML [24], DNN [25], GDO [26], GA [27], Swarm-ANN [28], Fuzzy Expert System [29], MANAGE-HF [30], BSN [31] and MSABA [32].

As shown in Figure 4, where the current SS-GBDT has a precision of 95.8%, the swarm optimized artificial neural networks [28] has marginally less precision (95.21%). However, the precision of SS-GBDT is about 38% better than MANAGE-HF [30] and 31% better than GA [27]. With respect to the ML framework [24] and DNN [25], the present precision is 24% and 17% better, respectively.

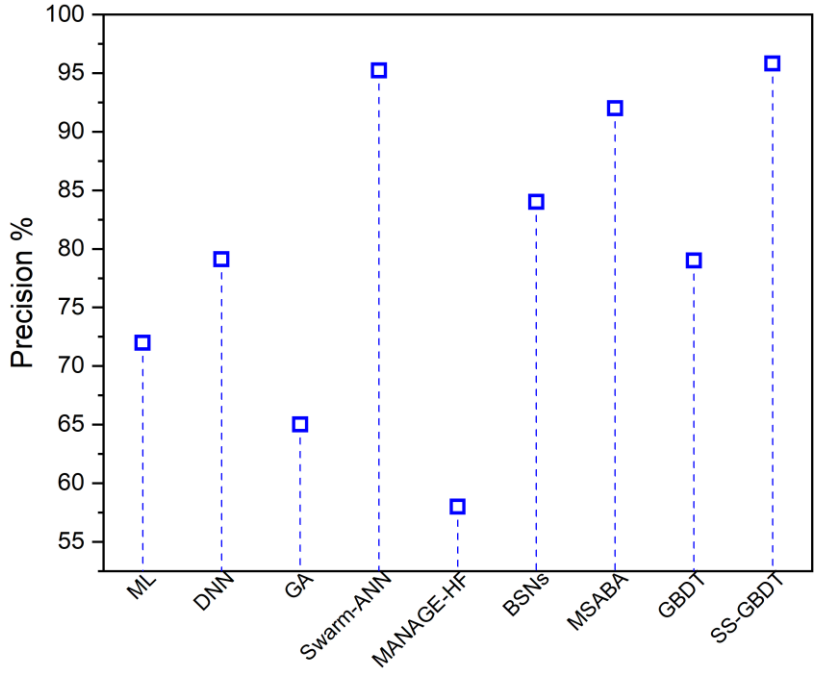

Figure 4. Comparison of precision of the proposed SS-GBDT and the state-of-the-art, compared with ML [24], DNN [25], GA [27], Swarm-ANN [28], MANAGE-HF [30], BSN [31] and MSABA [32].

The new methodology has a higher level of recall when compared to other methods; this can be seen in Figure 5, where the MANAGE-HF [30], GA [27], ML [24], BSNs [31], MSABA [32] and Swarm-ANN [28] have 36.8%, 30.8%, 23.8%, 11.8%, 5.8% and 1.6% lesser recall than the current SS-GBDT. The comparison of the F1-measure for various states of the art and the SS-GBDT is shown in Figure 6. In terms of the F1-measure, the present SS-GBDT is found to be about 33%, 26%, 17%, 8%, 1.3% and 1.1% superior to MANAGE-HF [30], GA [27], ML [24], BSNs [31], MSABA [32] and Swarm-ANN [28], respectively.

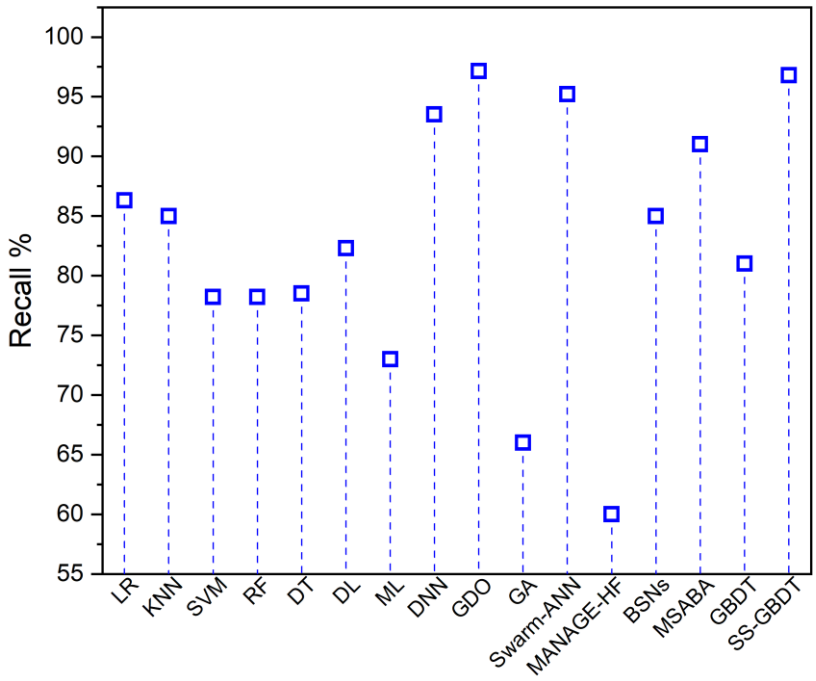

Figure 5. Comparison of recall of the proposed SS-GBDT and the state of the art, compared with LR [23], KNN [23], SVM [23], RF [23], DT [23], DL [23], ML [24], DNN [25], GDO [26], GA [27], Swarm-ANN [28], MANAGE-HF [30], BSN [31] and MSABA [32].

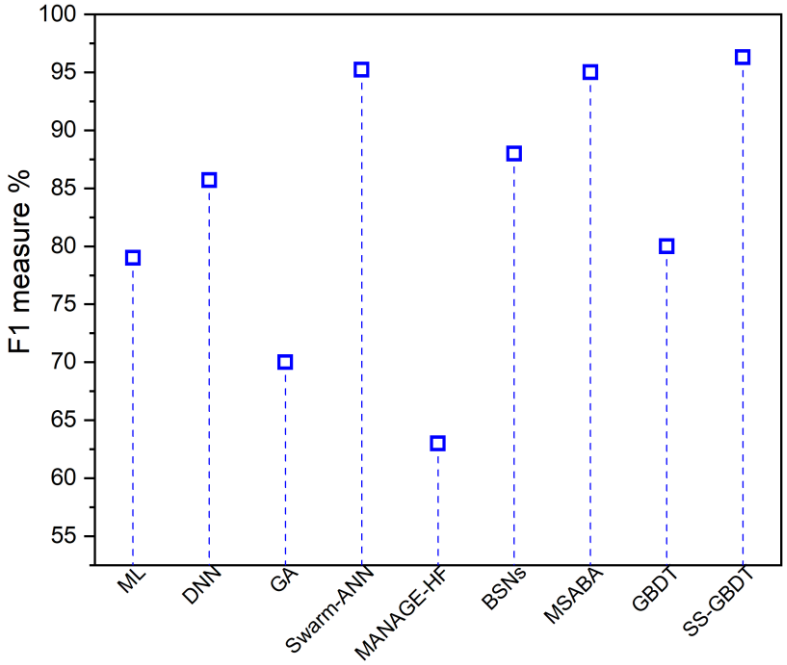

Figure 6. Comparison of the F1-measure of the proposed SS-GBDT and the state of the art, compared with ML [24], DNN [25], GA [27], Swarm-ANN [28], MANAGE-HF [30], BSN [31] and MSABA [32].

5. Conclusions

A novel squirrel search-optimized gradient-boosted decision tree (SS-GBDT), built on ML and BDA, is suggested in this article for the accurate classification of heart disease. The most important characteristics that could be employed as reliable predictors for the categorization of heart disease were identified by the SS-GBDT model. In this study, heart disease prediction was successfully accomplished in an experiment with a 95% accuracy rate, 95.8% precision, 96.8% recall and 96.3% F1-measure.

The current results are compared with a host of conventional ML as well as hybrid ML methods. The absolute superiority of the current SS-GBDT over conventional ML results and its on-par performance with complex hybrid ML methods are established. Therefore, using the provided method to predict heart disease is effective.

While the SS-GBDT model yields promising results, further development of the classification method is needed to improve heart disease prediction. Additionally, the platform's capabilities can be expanded for future deployment with the integration of more cutting-edge technologies. This would enable researchers and practitioners to continue refining and enhancing the performance of heart disease prediction models.

Author Contributions: Conceptualization, J.V.N.R., M.M. and K.K.; Formal analysis, K.S. and S.K.; Investigation, K.S. and S.K.; Methodology, K.S., J.V.N.R., M.M., M.Z.U.R., S.K. and K.K.; Validation, J.V.N.R. and M.Z.U.R.; Writing—original draft, K.S., J.V.N.R., M.Z.U.R. and S.K.; Writing—review and editing, M.M. and K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This article was supported by the project, SP2023/074, Application of Machine and Process Control Advanced Methods, supported by the Ministry of Education, Youth and Sports, Czech Republic.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available through email upon request to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ahsan, M.; Siddique, Z. Machine learning-based heart disease diagnosis: A systematic literature review. *Artif. Intell. Med.* **2022**, *128*, 102289. [CrossRef]
2. Diwakar, M.; Tripathi, A.; Joshi, K.; Memoria, M.; Singh, P.; Kumar, N. Latest trends on heart disease prediction using machine learning and image fusion. *Mater. Today Proc.* **2021**, *37*, 3213–3218. [CrossRef]
3. Hassan, S.; Dhali, M.; Zaman, F.; Tanveer, M. Big data and predictive analytics in healthcare in Bangladesh: Regulatory challenges. *Heliyon* **2021**, *7*, e07179. [CrossRef] [PubMed]
4. Konstantonis, G.; Singh, K.V.; Sfikakis, P.P.; Jamthikar, A.D.; Kitas, G.D.; Gupta, S.K.; Saba, L.; Verrou, K.; Khanna, N.N.; Ruzsa, Z.; et al. Cardiovascular disease detection using machine learning and carotid/femoral arterial imaging frameworks in rheumatoid arthritis patients. *Rheumatol. Int.* **2022**, *42*, 215–239. [CrossRef] [PubMed]
5. Ahmed, I.; Ahmad, M.; Jeon, G.; Piccialli, F. A Framework for Pandemic Prediction Using Big Data Analytics. *Big Data Res.* **2021**, *25*, 100190. [CrossRef]
6. Ramesh, T.R.; Lilhore, U.K.; Poongodi, M.; Simaiya, S.; Kaur, A.; Hamdi, M. Predictive analysis of heart diseases with Machine Learning approaches. *Malays. J. Comput. Sci.* **2022**, 132–148. [CrossRef]
7. Chang, V.; Bhavani, V.R.; Xu, A.Q.; Hossain, A. An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthc. Anal.* **2022**, *2*, 100016. [CrossRef]
8. Rehman, A.; Naz, S.; Razzak, I. Leveraging big data analytics in healthcare enhancement: Trends, challenges and opportunities. *Multimedia Syst.* **2022**, *28*, 1339–1371. [CrossRef]
9. Nagavelli, U.; Samanta, D.; Chakraborty, P. Machine Learning Technology-Based Heart Disease Detection Models. *J. Healthc. Eng.* **2022**, *2022*, 7351061. [CrossRef]
10. Ketu, S.; Mishra, P.K. Empirical Analysis of Machine Learning Algorithms on Imbalance Electrocardiogram Based Arrhythmia Dataset for Heart Disease Detection. *Arab. J. Sci. Eng.* **2022**, *47*, 1447–1469. [CrossRef]
11. Anooj, P.K. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *J. King Saud Univ.-Comput. Inf. Sci.* **2012**, *24*, 27–40. [CrossRef]

12. Dewan, A.; Sharma, M. Prediction of heart disease using a hybrid technique in data mining classification. In Proceedings of the 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 11–13 March 2015.
13. Sharanyaa, S.; Lavanya, S.; Chandhini, M.R.; Bharathi, R.; Madhulekha, K. Hybrid Machine Learning Techniques for Heart Disease Prediction. *Int. J. Adv. Eng. Res. Sci.* **2020**, *7*, 44–48. [CrossRef]
14. Rajendran, N.A.; Vincent, D.R. Heart disease prediction system using ensemble of machine learning algorithms. *Recent Pat. Eng.* **2021**, *15*, 130–139. [CrossRef]
15. Shorewala, V. Early detection of coronary heart disease using ensemble techniques. *Inform. Med. Unlocked* **2021**, *26*, 100655. [CrossRef]
16. Tiwari, A.; Chugh, A.; Sharma, A. Ensemble framework for cardiovascular disease prediction. *Comput. Biol. Med.* **2020**, *146*, 105624. [CrossRef]
17. Yoon, T.; Kang, D. Multi-Modal Stacking Ensemble for the Diagnosis of Cardiovascular Diseases. *J. Pers. Med.* **2023**, *13*, 373. [CrossRef]
18. Menshawi, A.; Hassan, M.M.; Allheeib, N.; Fortino, G. A Hybrid Generic Framework for Heart Problem Diagnosis Based on a Machine Learning Paradigm. *Sensors* **2023**, *23*, 1392. [CrossRef]
19. Reddy, K.V.V.; Elamvazuthi, I.; Aziz, A.A.; Paramasivam, S.; Chua, H.N.; Pranavanand, S. Heart disease risk prediction using machine learning classifiers with attribute evaluators. *Appl. Sci.* **2021**, *11*, 8352. [CrossRef]
20. Baccouche, A.; Garcia-Zapirain, B.; Olea, C.C.; Elmaghraby, A. Ensemble deep learning models for heart disease classification: A case study from Mexico. *Information* **2020**, *11*, 207. [CrossRef]
21. Almulihi, A.; Saleh, H.; Hussien, A.M.; Mostafa, S.; El-Sappagh, S.; Alnowaiser, K.; Ali, A.A. Refaat Hassan. Ensemble Learning Based on Hybrid Deep Learning Model for Heart Disease Early Prediction. *Diagnostics* **2022**, *12*, 3215. [CrossRef]
22. Cenitta, D.; Arjunan, R.V.; Prema, K.V. Ischemic Heart Disease Prediction Using Optimized Squirrel Search Feature Selection Algorithm. *IEEE Access* **2022**, *10*, 122995–123006. [CrossRef]
23. Bharti, R.; Khamparia, A.; Shabaz, M.; Dhiman, G.; Pande, S.; Singh, P. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. *Comput. Intell. Neurosci.* **2021**, *2021*, 8387680. [CrossRef] [PubMed]
24. Ko, Y.-F.; Kuo, P.-H.; Wang, C.-F.; Chen, Y.-J.; Chuang, P.-C.; Li, S.-Z.; Chen, B.-W.; Yang, F.-C.; Lo, Y.-C.; Yang, Y.; et al. Quantification Analysis of Sleep Based on Smartwatch Sensors for Parkinson's Disease. *Biosensors* **2022**, *12*, 74. [CrossRef] [PubMed]
25. Miao, K.H.; Miao, J.H. Coronary heart disease diagnosis using deep neural networks. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 10. [CrossRef]
26. Nawaz, M.S.; Shoaib, B.; Ashraf, M.A. Intelligent Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization. *Heliyon* **2021**, *7*, e06948. [CrossRef]
27. Eisa, M.M.; Alnaggar, M.H. Hybrid Rough-Genetic Classification Model for IoT Heart Disease Monitoring System. In *Digital Transformation Technology: Proceedings of ITAF 2020*; Springer: Singapore, 2022; pp. 437–451. [CrossRef]
28. Nandy, S.; Adhikari, M.; Balasubramanian, V.; Menon, V.G.; Li, X.; Zakarya, M. An intelligent heart disease prediction system based on swarm-artificial neural network. *Neural Comput. Appl.* **2021**, 1–15. [CrossRef]
29. Kasbe, T.; Pippal, R.S. Enhancement in diagnosis of coronary artery disease using fuzzy expert system. *Int. J. Sci. Res. Comput. Sci. Eng. Informat. Technol.* **2018**, *3*, 1324–1331.
30. Hernandez, A.F.; Albert, N.M.; Allen, L.A.; Ahmed, R.; Averina, V.; Boehmer, J.P.; Cowie, M.R.; Chien, C.V.; Galvao, M.; Klein, L.; et al. Multiple cArdiac seNsors for mAnaGEment of Heart Failure (MANAGE-HF)—Phase I Evaluation of the Integration and Safety of the HeartLogic Multisensor Algorithm in Patients With Heart Failure. *J. Card. Fail.* **2022**, *28*, 1245–1254. [CrossRef]
31. Shakya, S.; Joby, P.P. Heart disease prediction using fog computing based wireless body sensor networks (WSNs). *IRO J. Sustain. Wirel. Syst.* **2021**, *3*, 49–58. [CrossRef]
32. Subahi, A.F.; Khalaf, O.I.; Alotaibi, Y.; Natarajan, R.; Mahadev, N.; Ramesh, T. Modified Self-Adaptive Bayesian Algorithm for Smart Heart Disease Prediction in IoT System. *Sustainability* **2022**, *14*, 14208. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.