Journal of Research Administration
Society of Research Administrators International

EMPOWERING CORONARY ARTERY DISEASE PREDICTION THROUGH FEATURE OPTIMIZATION WITH ENSEMBLE LEARNING BASED HYBRID BAGGING AND BOOSTING TECHNIQUES

Jasmine Jinitha.A[1], Dr.S.Mangayarkarasi[2]
Research Scholar[1], Associate Professor [2]
Department of Computer Science, Vels Institute of Science Technologies and Advanced Studies (VISTAS), Chennai
jasminejinitha@gmail.com[1], mangai.research@gmail.com [2]

Abstract

Cardiovascular disease, also known as heart disease, has emerged as a prominent global health concern over the past decade. Ensuring the utmost accuracy in predicting cardiovascular events is of paramount importance. Accurate prediction is particularly crucial due to the severe consequences of heart diseases. Detecting heart complications at an early stage significantly enhances the effectiveness of treatment. In the pursuit of achieving precise early detection, various machine learning (ML) methods have been used. However, existing ML methods exhibit limitations in delivering efficient and accurate heart disease detection. These limitations result in challenges such as overcrowding in medical facilities due to unnecessary readmissions and unfortunate fatalities stemming from the discharge of patients requiring urgent medical attention. In this work, we present a method for the early diagnosis of heart disease by employing a combination of feature optimization and ensemble learning using hybrid bagging and boosting techniques. The proposed approach involves several key steps. Following data preprocessing, we harness an enhanced U-Net pre-trained architecture to extract features. This innovative approach utilizes both known and unknown features within the dataset to enhance feature extraction. To address data dimensionality challenges, we introduce a Modified Binary Search (MBS) algorithm for feature optimization. This algorithm effectively identifies the most optimal set of features to address data dimensionality issues. Furthermore, we introduce an ensemble learning-based hybrid bagging and boosting technique for heart disease detection and classification, which leverages the strengths of both approaches. To gauge the performance of our proposed method and compare it with existing techniques, we conduct experiments using the Cleveland and Framingham's heart study datasets. The simulation results of our technique are systematically contrasted with those of state-of-the-art methods. This comparison serves to underscore the effectiveness of our approach in terms of quality measures, thus validating its potential to advance the field of early heart disease diagnosis.

Keywords: early heart disease, cardiovascular disease, machine learning, feature extraction, feature optimization

1. Introduction

Cardiovascular disease (CVD) [1] stands as a significant global health concern, marked by its severe impact on individuals' well-being. Extensive medical evidence has underscored that certain health conditions significantly heighten the risk of developing cardiovascular disease, with heart disease being a prominent example [2]. Key risk factors include a family history of CVDs, hypertension, imbalanced cholesterol levels (low HDL cholesterol and high LDL cholesterol), a high-fat diet, and sedentary lifestyles lacking regular exercise. The repercussions of heart disease extend to increased hospitalizations, readmissions, healthcare expenditures, and reduced patient quality of life [3]. The world health organization (WHO) has designated cardiovascular diseases as the leading cause of death worldwide. The alarming statistics further reveal that annually, over 17.9 million lives are claimed by heart diseases, constituting 31% of global mortality rates [4]. Disturbingly, 33% of these fatalities occur among individuals under the age of 70, with a staggering 80% attributed to heart attacks. The need to predict high-risk patients, potentially vulnerable to readmission, has become paramount to improving healthcare outcomes. Current advances in man-made consciousness have prepared for more exact expectation models and calculations that can anticipate sickness beginning [5]. The escalating threat posed by CVD demands a comprehensive response on a global scale. Consequently, the healthcare sector must devise strategies to mitigate the socioeconomic impact of chronic diseases. The healthcare industry is rich with data pertaining to heart disease, which can be harnessed for informed decision-making [6]. Machine learning and data mining techniques [7][8] have proven indispensable in analyzing the medical data and extracting valuable insights. Researchers are urged to undertake extensive studies to curtail the incidence of cardiovascular diseases and associated mortality rates, fostering a healthier global population.

Machine learning (ML) [9][10] has emerged as a dominant approach in predicting heart disease due to its capacity to efficiently extract valuable insights from extensive datasets, thereby facilitating streamlined and accurate predictions. ML's foundational principles enable the handling of large data volumes with remarkable processing speed, thus enabling early-stage predictions [11]. This serves as the cornerstone for various ML applications that not only prevent medical errors but also enhance healthcare policies, disease prevention strategies, early detection mechanisms, and the prevention of avoidable hospital fatalities. A multitude of studies have explored the realm of heart disease prediction, focusing on identifying ML techniques that excel in diagnosing cardiac conditions [12]. Both regulated and solo learning procedures have been carried out in this scene to further develop identification of coronary illness. Regulated AI models [13]-[15] incorporate calculated relapse, choice trees, irregular woodlands, and brain networks that can be prepared on enormous patient informational collections to recognize risk factors and foster prescient models for readmission. A remarkable benefit of information driven AI models is their capacity to deal with enormous and complex informational indexes, even unstructured information, for example, clinical notes and pictures [16]. This permits medical care suppliers to reveal stowed away components of chance that can evade ordinary factual strategies. Likewise, ML-based models are prepared for persistent learning and variation to evolving information. As new persistent data opens up, these models can be recalibrated to build their precision and general execution. Nonetheless, conventional ML models grapple with issues related to data quality, bias, and interpretability [17]. Furthermore, their limitations in terms of efficacy and precision can lead to overburdened healthcare facilities, housing patients who don't require immediate readmission or, even more troubling, discharging patients in need of urgent care. Given the gravity of these consequences, the pursuit of a highly accurate model becomes a paramount concern [18]. Recent strides have been made in troupe models, showing stamped upgrades over individual classifiers. Be that as it may, group based techniques will quite often win employ rigid voting schemes without the adaptability required to align with the performance of baseline models. This drawback inevitably translates to diminished detection capabilities [19]. In light of these challenges, there exists a pressing need to devise an accurate and adaptive model that addresses the intricacies of heart disease prediction with precision and effectiveness. Nevertheless, there is a compelling need to enhance the efficacy of certain weaker techniques by amalgamating them with multiple classifiers [20]. This approach not only elevates the accuracy rates of these less potent classification methods but also enables the anticipation of heart disease in its early stages. The integration of ensemble techniques, encompassing both bagging and boosting processes, serves to bolster the accuracy levels of predictions derived from these initially weaker classification techniques. Consequently, these ensemble techniques demonstrate superior performance in identifying high-risk individuals prone to heart disease.

Our contributions. This study introduces an approach for the early diagnosis of heart disease, employing feature optimization alongside ensemble learning techniques that encompass hybrid bagging and boosting methodologies. The major contributions of the proposed work are given as follows.

1. Following data preprocessing, our approach involves leveraging an enhanced U-Net pre-trained architecture for the purpose of feature extraction. It effectively captures both known and unknown features from the dataset.
2. Addressing the challenge of data dimensionality, we introduce a novel Modified Binary Search (MBS) algorithm. This algorithm is strategically employed for feature optimization, enabling the identification and selection of the most pertinent and optimal features. By use the MBS algorithm, we aim to enhance the efficiency and effectiveness of our diagnostic model.
3. In an effort to improve heart disease detection and classification, we have devised an innovative ensemble learning approach. This approach encompasses a hybrid strategy that combines both bagging and boosting techniques. By integrating these techniques, we aim to enhance the accuracy and robustness of our model, allowing for more reliable and accurate predictions in heart disease identification.
4. To thoroughly evaluate and validate the performance of our proposed techniques, we conducted experiments using real-world datasets, Cleveland and Framingham’s heart study datasets. These datasets are widely recognized within the cardiovascular research community and provide a suitable foundation for assessing the effectiveness and potential of our approach in comparison to existing methods. Through this validation process, we

aim to demonstrate the superiority and utility of our proposed methodology in the realm of heart disease prediction and diagnosis.

The remainder of this article is coordinated as follows. Part 2 surveys late work on the expectation of coronary illness. In Segment 3, we detail the foundation investigation of the proposed work. The detailed working process of our proposed work is discussed in Section 4 with mathematical models. The results and comparative analysis of proposed work is explained in Section 5. Finally, the paper concludes in Section 6.

2. Literature Review

In this section, we present an overview of the existing body of research that has contributed to the field of early heart disease diagnosis and prediction. Table 1 highlights the various methodologies employed in these studies and provides insights into their strengths and limitations.

Irene et al. [21] have introduced an innovative ensemble approach, combining fuzzy k-modes attribute weighing (FKMAW) and deep belief network kernel extreme learning machine (DBNKELM), aimed at augmenting the medical diagnosis process. This ensemble technique intertwines a DBN driven attribute weighing mechanism with a classification methodology, facilitating effective medical data categorization. The input attributes undergo weighing through the application of the FKMAW technique. Primarily, this weighing process is employed to enhance the efficacy of medical data classification. Furthermore, this approach incorporates the transformation of datasets, originally non-linearly separable, into a form that is linearly separable. The ensemble method exhibits commendable results on various datasets, including heart Hungary, heart Swiss, and heart Cleveland. Notably, the classification accuracy attained through this method stands at 97.56% for the heart Hungary dataset, 97.21% for the heart Swiss dataset, and 97.62% for the heart Cleveland dataset. Aggrawal et al. [22] have introduced a novel sequential feature selection (SFS) algorithm geared towards identifying mortality events in heart disease patients under treatment. Furthermore, the accuracy achieved by the proposed SFS method is juxtaposed against the accuracy yielded by the chosen classifier. For validation, the system applies a K-fold cross-validation methodology. In particular, when employing the random forest classifier alongside the FS algorithm SFS, the achieved accuracy reaches 86.67% with a fivefold cross-validation approach.

Patro et al. [23] have introduced an innovative predictive model designed for ambient assisted living, with a focus on heart attack prediction. When addressing categorical independent variables, it is observed that decision trees exhibit superior performance in comparison to linear regression techniques. In situations involving outliers, the efficacy of a straightforward vector-based ML algorithm surpasses that of linear regression.The exploration of the best-performing algorithm, points to the efficacy of PCA, yields enhanced results within shorter timeframe. Among the various methods evaluated, the SVM emerges as a frontrunner, exhibiting a remarkable maximum accuracy of 92%.

Tougui et al. [24] have introduced a comprehensive approach that harnesses data mining tools and machine learning techniques for the classification of heart disease. The outcomes of this analysis revealed that the artificial neural network emerged as the most effective technique, with Matlab standing out as the preeminent tool. Utilizing the initial dataset, an impressive classification accuracy of 82.3% was achieved when considering all attributes, surpassing accuracy of 65.64% achieved through the selection of specific attributes.

Al-Yarimi et al. [25] have introduced a machine learning-based methodology for predicting heart disease. Your essential objectives are to work out some kind of harmony between the most reduced number of phony problems, the least interaction above, and the most elevated name forecast precision. The overarching aim is to understand and quantify the interplay between the features and labels to enhance the accuracy of heart disease prediction.

Li et al. [26] have introduced a novel approach for the prediction of CVDs harnessing a multi-modal framework that amalgamates both ECG and PCG features. Their methodology involves constructing traditional neural networks to extract deep-coding features from both ECG and PCG sources. To curate the optimal feature subset, they employ a genetic algorithm, which then paves the way for employing a SVM for classification tasks. The outcomes firmly establish that the multi-modal classification approach for predicting CVDs outperforms its single-modal counterparts, substantiating its potential in enhancing the accuracy of CVD prediction methodologies.

Rani et al. [27] have introduced an imaginative half breed choice emotionally supportive network adjusted for the early discovery of coronary illness in light of the patient's clinical boundaries. choice strategy including the hereditary calculation (GA) and recursive element disposal was utilized to choose the main highlights from the given informational index. Thoroughly tried on the Cleveland coronary illness informational collection, hybrid approach yielded remarkable results. Specifically, it achieved an accuracy of 86.6%, signifying its superiority over existing heart disease prediction systems. Zhang et al. [28] have devised an innovative machine learning strategy that revolves around integrating several prominent classification methods through the technique of model stacking. In particular, it exhibits an exactness of 87.7%, a responsiveness of 90.3%, and an explicitness of 84.3%.

Mehmood et al. [29] have introduced cardiovascular disease prediction system "CardioHelp," using CNN. This method centers on harnessing CNN for temporal data modeling specifically targeting the early prediction of heart failure. The approach involved curating a comprehensive heart disease dataset and subsequently comparing its outcomes with those of state-of-the-art methodologies. Remarkably, this approach attains an accuracy of 97%, underscoring its potential for precise cardiovascular disease prediction. Biswas et al. [30] have executed a comprehensive study involving distinct feature selection techniques to identify the most crucial attributes that hold substantial value in the prediction of heart disease. Following the feature selection process, six diverse machine learning algorithms were employed, each utilizing the identified essential features. Notably, every algorithm generated distinct scores based on the selected features. Among the array of algorithms, the SVM and LR exhibited particularly noteworthy performance.

3. Methodologies

This study employs a comprehensive approach encompassing multiple datasets, intricate data preprocessing steps, advanced feature extraction using Improved UNet, feature optimization through MBS algorithm, and diverse ensemble of classification. These steps collaboratively aim to achieve accurate heart disease detection, making significant contributions to the field of medical diagnosis.To bridge the existing research gaps, we present a novel approach for early heart disease diagnosis that encompasses the fusion of feature optimization and ensemble learning through a hybrid combination of bagging and boosting techniques. The central goals of our proposed endeavor are outlined as follows:

1. Developing an optimal model for enhanced feature extraction: Our focus lies in crafting an optimal model that facilitates a more profound extraction of features from the input data provided. This model aims to curtail the false positive rate, thereby enhancing the accuracy of heart disease diagnosis.
2. Addressing data dimensionality with optimization techniques: Tackling the challenge of data dimensionality, we employ advanced optimization techniques to meticulously select the suitable features from an array of possibilities. This approach ensures streamlined and efficient feature set without resorting to manual selection.
3. Leveraging hybrid bagging and boosting techniques: Recognizing the paramount significance of ensemble concepts in heart disease prediction, we introduce a hybrid framework that amalgamates bagging and boosting techniques. The fusion enhances the accuracy of detection, leveraging the strengths of both approaches.
4. Validating performance through benchmark datasets: Our methodology's efficacy is rigorously assessed and validated using benchmark datasets. We systematically evaluate the detection rate and its alignment with diverse quality measures, substantiating the robustness and effectiveness of our approach.

Table 1 Summary of Research gaps

Ref	Methodology	Technique	Dataset	Accuracy	Research gaps
[21]	Heart disease prediction	DBNKELM	Hungary, Swiss, Cleveland	97.62%	The identification of diseases at early stage holds immense significance.
[22]	Prediction of heart disease	LDA, RF, GBC, DT, SVM, KNN	Heart failure clinical record	86.67%	There remains a notable deficit in achieving satisfactory levels of both sensitivity and specificity.

[23]	Ambient assisted living predictive model	KNN, NB, SVM	Cleveland	92%	The efficacy of the performance is closely tied to the choice of the feature selection algorithm.
[24]	Heart disease classification	KNN, NB, SVM	Cleveland	82.83%	The absence of visualized correlations between features is evident.
[25]	Heart disease prediction	DT, KNN, SVM, DSS	Cleveland	92.062%	Their applicability for widespread use in larger populations is constrained.
[26]	Prediction of cardiovascular diseases	Genetic algorithm, SVM	ECG and PCG	93.6%	The consideration of the class imbalance problem is absent
[27]	Heart disease prediction	GA and SMOTE	Cleveland	86.6%	It relies solely on data from a single source.
[28]	Screening of coronary heart disease	Two step stacking	Clinical trial (NCT03905200)	87.7%	It exhibits limitations and proves unsuitable for real-time datasets.
[29]	Prediction of heart disease	DCNN	Cleveland	97%	Attaining an average sensitivity rate of 78% falls significantly short.
[30]	Predict heart disease in early stage	LR, RF, DT, SVM, KNN, NB	Cleveland	94.51%	It is susceptible to challenges posed by data dimensionality.

Fig. 1 shows the overall conceptual structure of proposed work for heart disease detection which leverages two prominent datasets, namely the Cleveland and Framingham, to comprehensively address the research objectives. The initial phase of data preprocessing entails a meticulous examination for missing entries, followed by a thorough cleaning process to ensure data integrity. Subsequently, the collected data from both sources are integrated to form a cohesive dataset, and techniques for data reduction and transformation are employed to streamline its structure. The feature extraction phase is pivotal in this study and is accomplished through an advanced approach called Improved UNet. The Improved UNet's capacity to extract intricate features provides a comprehensive basis for subsequent analysis. To optimize the extracted features, a modified binary search (MBS) algorithm is introduced. This algorithm systematically identifies the best optimal features by iteratively narrowing down the search space. The heart of this study's methodology lies in heart disease detection and classification, utilizing a robust ensemble learning approach. Four distinct ensemble techniques are employed: Adaboost, XGBoost, Bootstrap bagging, and a hybrid approach combining Adaboost, XGBoost, and Bootstrap bagging (AXG-B+BS-B). In terms of classification, two classes are established: "Nil" for instances with no discernible risk of heart disease, and "Risk" for those exhibiting signs of potential heart disease. This binary classification framework is employed across all ensemble techniques, ensuring consistency and comparability in the evaluation process.

3.1 Feature extraction

Feature extraction is a crucial step in ML and data analysis, especially when working with complex and high-dimensional data like medical datasets. In this context, feature extraction involves transforming the raw data into a more compact and informative representation. The feature extraction from medical data could involve identifying specific attributes that are indicative of heart disease risk, such as blood pressure, cholesterol levels, age, and other relevant clinical measurements. Our innovative strategy not only incorporates known features from the dataset but also effectively leverages previously unidentified or "unknown" features. By integrating both types of features, we aim to elevate the efficacy of feature extraction and ensure a comprehensive representation of the underlying data characteristics. The utilization of an enhanced U-Net architecture underscores our commitment to extracting intricate and valuable information from the dataset. This approach aligns with the complexity of medical data, particularly in the context of heart disease detection, where numerous factors contribute to the overall diagnostic picture. By tapping into known and previously unrecognized features, we strive to provide more accurate and informative foundation for subsequent analyses. Through this amalgamation of known and unknown features, our methodology seeks to enhance the precision of feature extraction and contribute to the overall success of our heart disease detection framework.

The U-Net architecture, a convolutional neural network (CNN) model originally devised for image segmentation tasks, follows a distinct sequence of steps to achieve its objectives. The architecture is characterized by its contracting path, which acts as the encoder to capture context, and an expanding path, functioning as the decoder to facilitate precise localization. Each convolutional layer is complemented by a ReLU activation, fostering non-linearity. The integration of max-pooling layers then facilitates spatial down sampling and enhances the receptive fields. The expanding path embarks on the task of restoration and localization. Employing up-convolutional layers, it initiates the process of spatial up sampling. Concatenation with corresponding feature maps from the contracting path establishes critical skip connections, allowing for effective information transfer between the encoding and decoding phases. Convolutional layers within the expanding path contribute to the refinement of features, coupled with ReLU activation. The architecture culminates with an output layer, where a final convolutional layer translates the intricate features into the desired output format, such as a segmentation mask. This sequence of steps facilitates accurate image segmentation by harnessing both contextual information and precise localization cues.

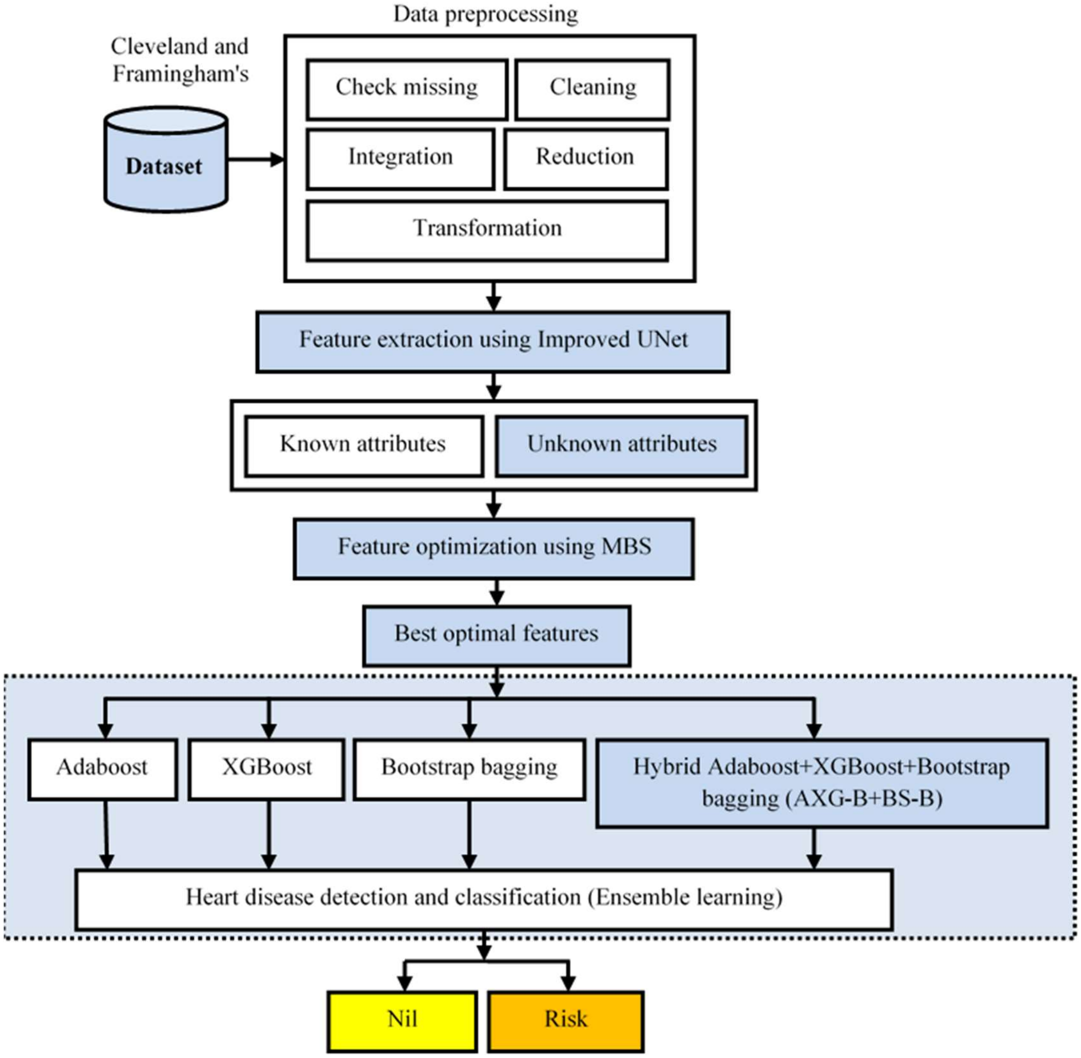

Fig. 1 Overall conceptual structure of proposed work for heart disease detection

The Improved UNet model utilizes the UNet pre trained architecture for feature extraction from the given medical data, and improves upon it by using the optimization algorithm to optimize the weights of the UNet for better performance in extracting hidden features. Let be the initial cost matrix. The standard function for the Hungarian optimization algorithm alters this matrix by successively adding or subtracting entire rows and columns.

$$b[g,i]=Z[g,i]-R[g]-v[i] \tag{1}$$

The exhibits i and v must likewise be every now and again modified for most of straight task issues on the grounds that the calculation transforms them regularly. The partner coefficient, not set in stone by the negligible measurement distance among plots and tracks, is the manner by which every affiliation's worldwide similitude is communicated. Expecting that all state values are assessed through space-time adjustment, the strategy for deciding the state contrast between plot h and combination track u in the time J is as per the following.

$$S_{gi}(h)=\hat{r}_g^1(h\,|\,h)-\hat{R}_i^2(h\,|\,h) \tag{2}$$

Coming up next is a meaning of the measurement distance in time J between plot h and the combination track.

$$\beta_{ji}(h)=S_{ji}(h)^S z_{gi}^{-1}(h)_{gi}(h) \tag{3}$$

$$z_{ig}(h)=q_g^1(h\,|\,h)+q_i^2(h\,|\,h) \tag{4}$$

where is calculated using estimates from plot i's state error covariance and fusion track . To constructs the gathering speed of finding the overall optima in the smoothing out issue.

$$r_i^{'}=r_{a.i} \tag{5}$$

If not, a new value is selected from the set of possible values. The effect on how dynamically the parameter changes in optimal algorithm.

$$Q_a(Gs)=Q_a^{\min}+\left(\frac{Q_a^{\max}-Q_a^{\min}}{\max Ga}\right)\times Ga \tag{6}$$

where $Q^{\max}$ and $Q^{\min}$ are the respective maximum and minimum probabilities.

$$Q_A(Gs)=0.8+\left(\frac{0.15}{\max Gs}\right)\times Gs \tag{7}$$

Since the best arrangement vector will be nearer to worldwide optima in the improvement cycle, it is found that the likelihood is more modest in the underlying runs and expansions in ensuing emphases. The pitch change is then characterized as follows.

$$r_i^{'}=r_i^{'}-Hma(1+2\times u(0,1)\times(r_h^{D_1}-t_l^{D_2})) \tag{8}$$

Where and are two arbitrarily chosen people from the pacemaker populace. The QBR likelihood is portrayed underneath.

$$WVE(ha)=WVE^{\max}-\left(\frac{WVE^{\max}-WVE^{\min}}{\max gs}\right)\times gs \tag{9}$$

where $WVE^{\min}$ and $WVE^{\max}$ are at their lowest and highest probabilities, respectively. This function is one of the most frequently used standard test functions in optimization problems.

In essence, the adoption of the Improved U-Net model as a feature extraction tool introduces a transformative impact on the heart disease detection process. By harnessing its capabilities to capture both known and unknown features, this approach can elevate the accuracy, sensitivity, and overall quality of the diagnostic outcomes, ultimately leading to improved patient care and outcomes. Algorithm 1 describes the working steps involved in the deep feature extraction using Improved UNet model.

Algorithm 1 Deep feature extraction using Improved UNet

```
Input    : Number of known features, maximum iteration and termination condition
Output : Unknown deep features
Set the random population up.
Define total amount removed from column i of the modified matrix.
    $b[g,i]=Z[g,i]-R[g]-v[i]$
Determine the state difference between fusion track i and plot j in the time k is
    $z_{ig}(h)=q_g^1(h|h)+q_i^2(h|h)$
If t=0, while do
    Compute the optimal solution for fitness using
        $r_i^{'}=r_i^{'}-Hma(1+2\times u(0,1)\times(r_h^{D_1}-t_l^{D_2}))$
    Define most frequently used standard test functions for optimization problems
        $f(r)=4r_1^2-2.1r_1^4+\frac{1}{3}r_1^6+r_1r_2-4r_2^2+4r_2^4$
    Update the final neural values
End if
End
```

3.2 Feature optimization

After the process of feature extraction, feature optimization becomes the subsequent crucial step in refining the dataset for further analysis, such as classification or prediction tasks. Feature optimization involves selecting the most relevant and informative subset of features from the extracted set. This selection aims to enhance the performance of ML algorithms by reducing noise, increasing model interpretability, and potentially mitigating the curse of dimensionality. The modified binary search (MBS) algorithm for feature optimization is an innovative approach designed to efficiently and effectively identify the most relevant subset of features from a larger feature space. This algorithm draws inspiration from the traditional binary search technique but is adapted to the context of feature selection in machine learning. The following steps involved in the MBS algorithm for feature optimization.

- Initialization: Start with the entire set of extracted features.
- Ranking features: Rank the features based on their relevance to the target variable. This could involve using metrics like correlation, mutual information, or statistical tests to evaluate the impact of each feature on the model's performance.
- Binary search logic: The algorithm initiates a binary search logic by dividing the ranked feature set into two subsets: a "selected" subset and a "discarded" subset. The selection of this division point is crucial and can be based on a predefined threshold or a statistical criterion.
- Evaluation: Train and evaluate the model using the selected subset of features. This could involve a cross-validation approach to ensure the model's performance is consistent across different folds of data.
- Decision Making: Depending on the performance of the model using the selected subset, the algorithm makes a decision: If the model's performance is satisfactory, the algorithm terminates, and the selected subset of features is considered the optimal feature set. If the model's

performance is suboptimal, the algorithm adjusts the division point and reiterates the evaluation process with a refined subset.

- Iteration: The algorithm iteratively refines the feature subset using the binary search logic, optimizing the feature set based on the model's performance.

In the MBS calculation, every individual position is characterized in a restricted space $R=\{o\in e^{v}\mid km_c\le o_c\le ym_c\}$ $(with\ o=[k_1,k_2,\dots k_c]$ and where Km_v and un_c address the lower and upper limits on aspect V, individually) and show the answer for the given advancement issue.

$$R_g^l=\sum_{\substack{h=1\\h\neq 1}}^{V} r_{gh}^l \tag{11}$$

where r_{gh}^l means the fascination two by two between a person "g" and a few others "h" and is given by:

$$r_{gh}^l=\rho\left(k_g^l,\,k_h^k\right)r\left(e_{gj}^l\right)v_{gh}+rand(1,-1) \tag{12}$$

where ‖ addresses the Euclidean distance between the components "g" and "h". Thusly, the unit vector pointing from to , and rand(1, −1) is an irregular vector, are drawn from the uniform circulation [−1, 1]. The worth is given as follows.

$$r\left(T_{gh}^l\right)=dR^{-e_{gh}^l/k}-R^{-e_{gh}^l} \tag{13}$$

where the boundaries f and l indicate the push size and the length scale, individually. To apply the administrator, it is accepted that every individual is positioned between 0 (best individual) and B-1 (most terrible individual) in light of their separate wellness esteem. Consequently, the predominance worth can be given as:

$$\rho\left(k_g^l,\,k_h^l\right)=\begin{cases}R^{-(rank(k_g^l)/v)} & if\quad rank\left(k_g^l\right)\le\left(k_h^l\right)\\ R^{-(rank(k_g^l)/v)} & if\quad rank\left(k_h^l\right)>\left(k_h^l\right)\end{cases} \tag{14}$$

Because of the impact of complete social power, every person "g" communicates a specific inclination to move towards different individuals from the populace. Under these circumstances, the new place of individual "g" can be communicated as follows.

$$k_g^*=k_h^l+r_g^l \tag{15}$$

The consequence of applying novel move administrators to every individual is another arrangement of up-and-comer choices that address the places of every person because of the impact of any remaining individuals from the gathering. The social stage activity is applied to refine probably the best applicants produced by the solitary stage development administrator. To do this, a subset of arrangements is first characterized, comprising of the endlessly best arrangements in the set. Then, at that point, for every arrangement ∈M, a bunch of u irregular arrangements is produced in the relating subspace ∈T, whose cutoff points are given as follows.

$$F_{g,v}^{lower}=m_{g,v}-T \tag{16}$$

$$F_{g,v}^{upper}=m_{g,v}+E \tag{17}$$

where y signifies the upper and lower limits of every subspace in aspect h, separately, and y means the bth choice variable) of the arrangement n_h ,

$$E=\frac{\sum_{v=1}^{v}\left(m_v^{upper}-m_v^{lower}\right)}{v}\cdot\beta \tag{18}$$

with y signifying the lower and upper limits of the bth aspect, individually, and c meaning the absolute number of choice factors. Besides, ∈ [0, 1] addresses a scalar coefficient that balances the greatness. At long last, the best arrangement of ∈M and its comparing I arbitrary arrangements ( ) are alloted as the area of the person "G" in this emphasis "l+1".

$$k_g^{l+1} = best\left(k_g^*, s_1^g, s_2^g, \ldots s_i^g\right) \tag{19}$$

It is essentially worthless that any arrangement that isn't assembled in the best arrangement set B is wiped out by the administrator of the social stage. In this way, the last position update applied to every person "g" in the whole group can be summed up as:

$$k_g^{l+1} = \begin{cases} best\left(k_g^*, s_1^g, s_2^g, \ldots g_u^g\right) & if \;\; k_g^* \varepsilon \; M \\ l_h^* & if \;\; k_g^* \notin M \end{cases} \tag{20}$$

In these conditions, the period of conduct applied to the populace "L" in every cycle is picked as follows.

$$Z\left(k^l\right) = \begin{cases} solitary & if \quad rand \le z^l \\ social & if \quad rand > z^l \end{cases} \tag{21}$$

where means an irregular number chose from a consistently circulated stretch [0, 1], and the worth , called the conduct likelihood, is given by,

$$z^l = 1 - \frac{l}{itern} \tag{22}$$

where means the most extreme number of cycles considered in the pursuit cycle.

$$Z^l_{kg,m_h} = \frac{M\left(m_h^k\right) R^{-\left\|k_h^l - m_h^l\right\|}}{\sum_{n=1}^{u} n\left(m_m^l\right) r^{-\left\|k_g^l - m_h^l\right\|}} \tag{23}$$

where indicates the Euclidean distance between the person "g" ( ) and the part "g" of the arrangement of best arrangements ( ), and signifies the engaging quality of the arrangement as

$$N\left(m_h^l\right) = \frac{D(m_H^L) - D_{worst}\left(M^L\right)}{D_{best}\left(N^L\right) - D_{worst}\left(M^L\right) + \varepsilon} \tag{24}$$

Considering the abovementioned, while acting in the social stage, every person in the multitude populace refreshes its situation as follows:

$$k_g^{l+1} = k_g^l + 2\left(m_e^l - k_g^l\right) \cdot rand \tag{25}$$

where (with R ∈ [1, ...]. Calculation 2 portrays the work process connected with the advancement of capabilities utilizing MBS.

Algorithm 2 Feature optimization using MBS

```
Input  : No of known features, no of unknown features and maximum iteration
Output : Optimal best features
1.    Initialize the random population
2.    The value $t\left(R_{hg}^K\right)$is given by $r\left(t_{gh}^l\right) = dw^{-t_{gh}^l/k} - w^{-e_{gh}^l}$
3.    If u=0 , h=1
4.    While Do
5.    Define the individual "h" as a result of $T_h^K$ $r\left(T_{gh}^l\right) = dR^{-e_{gh}^l/k} - R^{-e_{gh}^l}$
6     Define the whole swarm may be summarized as follows
      $k_g^{l+1} = \begin{cases} best\left(k_g^*, s_1^g, s_2^g, \ldots g_u^g\right) & if \; k_g^* \varepsilon \; M \\ l_h^* & if \; k_g^* \notin M \end{cases}$
7.    If not discard then
```

```
8.    Compute distance separating both individuals ( $l_h^K$ and $n_g^K$ )
      $Z_{kg,m_h}^l = \frac{M\left(m_h^k\right)R^{-\left\|k_h^l - m_h^l\right\|}}{\sum_{n=1}^{u} n\left(m_m^l\right)r^{-\left\|k_g^l - m_h^l\right\|}}$
9.    Update the swarm population $l^*$ and position $k_g^{l+1} = k_g^l + 2\left(m_e^l - k_g^l\right)\cdot rand$
10.   Update the final values
11.   End
```

3.3 Coronary artery disease detection and classification

Coronary artery issues detection and classification refer to the process of identifying and categorizing different types of heart diseases based on available patient data and diagnostic features. This task involves utilizing various machine learning and data analysis techniques to make accurate predictions about a patient's heart health and the potential presence of cardiovascular conditions. In our research, we present an innovative approach to heart disease detection and classification by introducing an ensemble learning-based hybrid technique that combines the merits of both bagging and boosting methodologies. This hybrid approach aims to capitalize on the strengths and synergies of these two techniques to enhance the accuracy and reliability of our diagnostic model. Ensemble learning is a powerful concept that involves the integration of multiple individual models to create a stronger, more robust predictive model. Bagging and boosting are two prominent ensemble techniques, each offering distinct advantages. Bagging involves training multiple models on different subsets of the data and averaging their predictions, which helps reduce variance and enhance stability. On the other hand, boosting focuses on iteratively improving the model's performance by assigning more weight to misclassified instances, resulting in enhanced predictive capabilities. Our hybrid approach combines these strategies to create a new ensemble technique that inherits the benefits of both bagging and boosting. This hybridization involves training multiple models using both bagging and boosting techniques and subsequently aggregating their predictions. By leveraging the unique strengths of both bagging and boosting, we aim to create a model that excels in capturing complex patterns, reducing errors, and delivering precise predictions in heart disease detection and classification. The integration of this novel hybrid approach into our diagnostic framework has the potential to revolutionize the accuracy and reliability of our model, ultimately leading to more effective early detection and personalized treatment strategies for patients with heart disease.

3.3.1 Adaboost classifier

The Adaboost (Adaptive Boosting) classifier is a prominent machine learning algorithm used for binary classification tasks, including coronary artery disease detection. Adaboost is part of the ensemble learning family and is known for its ability to improve the accuracy of weak learners by combining their predictions in an adaptive and weighted manner. In the context of heart disease detection, the Adaboost classifier works as follows:

1. Initialization: Each training instance is assigned an equal weight initially.

2. A weak learner, often a simple decision tree with limited depth, is trained on the training data. .
3. The Adaboost algorithm calculates the weighted error of the weak learner's predictions. The weight of each training instance is adjusted based on whether it was classified correctly or incorrectly.
4. A weight is assigned to the weak learner's prediction based on its performance. The better the weak learner's accuracy, the higher the weight assigned to its predictions.
5. Instances that were incorrectly classified by the weak learner receive higher weights to emphasize their importance in subsequent iterations.
6. Adaboost combines the predictions of weak learners based on their weights. The final prediction is the result of a weighted majority vote.
7. The process iterates, with subsequent weak learners focusing more on the instances that were misclassified by the previous ones. This adaptive process enhances the overall model's accuracy.
8. After a predefined number of iterations, the predictions of all weak learners are combined to make the final classification.

It provides improved predictive accuracy compared to individual weak learners and has the potential to handle imbalanced data distribution commonly found in medical datasets.

3.3.2 XG boosting classifier

With regards to coronary illness identification, the XGBoost classifier fills in as follows:

1. Initialization: Similar to other gradient boosting algorithms, XGBoost begins with an initial prediction, which is often set to the mean of the target variable.
2. XGBoost aims to minimize a specific loss function, such as logistic loss for binary classification problems like heart disease detection. The loss function quantifies the difference between predicted and actual values.
3. A series of weak learners are sequentially trained on the data.
4. For each instance, the negative gradient of the loss function is calculated to guide the learning process.
5. XGBoost combines the predictions of all weak learners, giving more weight to those that have performed well on instances with higher gradients.
6. To prevent overfitting, XGBoost introduces regularization terms that penalize complex models. This helps maintain model simplicity and generalization.
7. The final prediction is the result of aggregating the predictions of all weak learners. The aggregation process takes into account the weights assigned to each learner.
8. The training process iterates, and each new weak learner is trained to correct the errors made by the ensemble so far.

The XGBoost classifier is a robust tool for heart disease detection, leveraging gradient boosting to create a strong and accurate predictive model. Its ability to handle complex patterns, adapt to challenging instances, and handle imbalanced data distribution makes it a valuable asset in medical diagnostics.

3.3.3 Bootstrap bagging classifier

The bootstrap bagging classifier is powerful ensemble learning algorithm used for classification tasks, including coronary artery disease detection. Bagging is designed to improve the accuracy, stability, and robustness of machine learning models by combining the predictions of multiple base models trained on different subsets of the training data. In the context of heart disease detection, the bootstrap bagging classifier operates as follows:

1. It starts by creating multiple random subsets of the training data through a process called bootstrap sampling. Each subset is generated by selecting instances from the original training data with replacement.
2. For each bootstrap sample, a base model is trained independently on that sample. These base models are known as weak learners because they might not be highly accurate on their own.
3. After training the weak learners, the classifier combines their predictions using a majority voting mechanism for classification. In the case of heart disease detection, the class with the highest number of votes among the weak learners' predictions is chosen as the final prediction.
4. The aggregation of predictions from multiple models helps reduce variance and stabilize the overall model's performance. This is particularly beneficial in cases where a single model might overfit the training data.
5. The final prediction of the ensemble is the result of the majority vote or averaging of predictions from all weak learners. This aggregated prediction is generally more accurate and robust than the prediction of a single model.

This classifier is a valuable tool in heart disease detection, leveraging the power of ensemble learning to improve predictive accuracy and robustness.

3.3.4 Hybrid Adaboost+XGBoost+Bootstrap bagging (AXG-B+BS-B)

The Hybrid Adaboost+XGBoost+Bootstrap Bagging (AXG-B+BS-B) approach is an innovative ensemble learning strategy designed to harness the strengths of three distinct ensemble techniques. By combining these methodologies, we aim to create a robust and accurate predictive model for heart disease detection that leverages the unique advantages of each technique. Compared with other algorithms, AXG-B+BS-B gather a strong classifier from a set of weak classifiers and shows the following advantages: Efficient handling of missing values can reduce the running time of parallel and distributed computation. The reason for AXG-B+BS-B is to utilize the angle drop streamlining strategy and inconsistent differentiable misfortune capabilities to limit the misfortune capability by adding powerless students, that is to say, to characterize and improve the goal capability. AXG-B+BS-B attempts to diminish the conventional objective as follows:

$$obj(\theta) = \sum_{h} k(\hat{z}_h, z_h) + \sum_{j} \Omega(D_J), \quad D_h \in d \tag{26}$$

The larger, Q_a is the more information from the previous moment is sent to the current state. It is E_a "reset gate" that limits the amount of influence of the previous state g_{s-1} on the candidate state $\widetilde{g}_s$. The forward propagation formula for the GRU network is as follows:

$$E_a = \sigma(X_{hE} h_{a-1} + X_{yt} t_a + s_t) \tag{27}$$

where K is a preparation misfortune capability that actions the deviation between the worth anticipated by our model and the genuine worth. Ω is a regularization capability that actions the intricacy of the model and dodges overfitting. In each iteration, the remainder is used to replace the previous weaker students.

$$\hat{z}_{yh} = \sum_{N=1}^{n} D_N(t_h), D_M \in d \tag{28}$$

where signifies the normal worth of the jth test, M signifies the quantity of components in the example, M signifies the normal worth of the jth test in the tree, and f means the element space. The goal capability of AXG-B+BS-B comprises of a misfortune capability and a fitting term.

$$obj = \sum_{h=1}^{\eta} k\left(z_{uh}, z_{th}\right) + \sum_{n=1}^{m} \Omega(D_n) \tag{29}$$

$$\Omega(F_n) = \gamma a + \frac{1}{2} \beta \sum_{u=1}^{a} q_u^2 \tag{30}$$

where η addresses the quantity of tests, l addresses the second request subsidiary of the misfortune capability, which estimates the contrast between the genuine worth and the normal worth. shows the control time frame. S is the quantity of leaf hubs in the tree, the leaf hub score, γ and β are boundaries that control the intricacy of the tree. As a matter of fact, the streamlining of the goal capability comprises of deciding the best division capability and the best division point. The goal capability can be improved to a solitary quadratic condition utilizing a second request Taylor development.

$$obj = \sum_{u=1}^{a} \left[\left(\left(\sum_{h \in u_i} g_h \right) w_i + \frac{1}{2} \left(\sum_{h \in h_i} h_h + \beta \right) q_u^2 \right) + \gamma a \right] \tag{31}$$

where indicates all information tests at leaf hub j and means the first and second subsidiaries of the misfortune capability. Here it tends to be characterized as follows.

$$H_i = \sum_{j \in J_i} h_j, H_i = \sum_{j \in J_i} g_j \tag{32}$$

The ideal score of the leaf hub and the comparing ideal worth of the goal capability are acquired by tackling the above single quadratic capability

$$Q_u^* = -\frac{G_u}{h_u + \beta} \tag{33}$$

$$obj = -\frac{1}{2} \sum_{i=1}^{S} \frac{G_i^2}{F_U + \beta} + \gamma S \tag{34}$$

By synergistically integrating the strengths of Adaboost, XGBoost, and Bootstrap Bagging, this hybrid (AXG-B+BS-B) algorithm aims to deliver a powerful predictive model that excels in capturing complex patterns, reducing errors, and providing reliable results for early diagnosis and treatment of heart disease.

4. Results and Discussion

In this section, we present the results and comparative analysis of proposed and existing heart disease detection techniques. To validate the performance through the two different benchmark datasets such as Cleveland and Framingham's heart study. The simulation results of proposed

Adaboost, XGBoost, bootstrap bagging and AXG-B+BS-B techniques with the Cleveland dataset is compared with the existing state-of-art techniques such as Naïve bayes (NB), decision tree (DT), logistic regression (LR), k-nearest neighbor (K-NN), random forest (RF) and support vector machine (SVM) [31][33][34]. The simulation results of proposed Adaboost, XGBoost, bootstrap bagging and AXG-B+BS-B techniques with the Framingham's dataset is compared with the existing techniques, includes DT, RF, AdaBoost, Gaussian naive bayes (GNB), LR, K-NN, XGBoost, gradient boosting (GBoost) [32]. The performance can be analyzed through different measures such as accuracy, precision, Recall, F-measure, kappa statistics, and area under curve (AUC).

4.1 Dataset description

The effectiveness of both the proposed heart disease detection techniques and existing methods has been rigorously evaluated using two distinct benchmark datasets: the Cleveland dataset and the Framingham's Heart Study dataset.

1. The Cleveland dataset, sourced from the UCI Machine Learning repository, encompasses 300 real-world instances, each characterized by 14 diverse attributes. Among these attributes, 13 serve as predictors, while the remaining one represents the class label. These attributes cover a range of factors such as blood pressure, type of chest pain, and electrocardiogram results, all of which contribute to the dataset's comprehensive nature.
2. In contrast, the Framingham's Heart Study dataset comprises a more extensive set of attributes that delve into the patient's demographic, behavioral, and educational details. Additionally, this dataset includes information about the patient's prior and current medical conditions. Specifically, the dataset encompasses data from 3179 patients without coronary heart disease (CHD) and 572 patients with CHD.

To facilitate a thorough understanding of the attributes, their significance, and data types present in both datasets, Table 2 provides a comprehensive overview. This detailed examination of attributes underscores the breadth of information captured by these datasets, which, in turn, enhances the validity and reliability of the evaluations conducted on the proposed and existing heart disease detection techniques.

Table 2 Description of Cleveland dataset and the Framingham's Heart Study dataset

Attribute	Description	Datasets	
		Cleveland [31]	Framingham's [32]
Age	Age	Age of patients	
Sex	Sex	Female (0), Male (1)	Male
Education	Level of education		4 options
Current Smoker	The patient smokes		No (0), Yes (1)

CigsPerDay	Number of cigarettes		Limits (1-N)
Chest pain	Cp	4 options	
Rest BP	Trestbps	Systolic blood pressure (mm Hg)	Normal (0), Abnormal (1)
DiaBP	Diastolic BP		Normal (0), Abnormal (1)
BPMeds	BP medications		No (0), Yes (1)
BMI	Body mass index		Normal (0), Abnormal (1)
PrevalentStroke	Suffered from a stroke		No (0), Yes (1)
PrevalentHyp	Hyper Tensile or not		No (0), Yes (1)
Serum cholesterol	Chol	Serum cholesterol (in mg/dl)	
Fasting blood sugar	Fbs	False (0), True (1)	
Diabetes	Diabetic or not		No (0), Yes (1)
Rest ECG	Restecg	4 options	
Heart rate	Thalch	Maximum heart rate achieved	
Glucose	Blood glucose levels		Normal (0), Abnormal (1)
Exercise	Exang	No (0), Yes (1)	
ST depression	Oldpeak	Depression induced by exercise	
Slope	Slope	Up sloping (1), Fat (2), Down sloping (3)	
No. of vessels	Ca	No. of major vessels (0–3)	
Thalassemia	Thal	Normal (3), Fixed defect (6), Reverse defect (7)	
Num (class attribute)	Class	Nil (0), Risk (1)	NoCHD (0), CHD (1)

4.2 Results analysis for Cleveland dataset

Table 3 presents a concise comparison of the results achieved by the proposed heart disease detection technique and existing methods using the Cleveland dataset. The table encapsulates the performance evaluation outcomes in terms of various metrics, providing a clear overview of how the proposed technique fares against established approaches. In Fig. 2, an accuracy comparison among various techniques is showcased. Among the existing methods, SVM [33] emerges as a strong contender with a relatively high accuracy of 95.556%. In comparison, other techniques such as NB [33], DT [33], and SVM [34] exhibit accuracies of 74.124%, 82.536%, and 82.00% respectively. Similarly, NB [34], K-NN [34], and LR [34] demonstrate accuracies of 83.00%, 80.000%, and 85.00% respectively. Specifically, NB [31] attains an accuracy of 88.157%, while K-NN [31], DT [31], and RF [31] achieve 90.789%, 80.263%, and 86.840% respectively. The introduction of ensemble techniques marks a substantial improvement in accuracy. Adaboost achieves an accuracy of 95.623%, showing a slight increase compared to certain individual methods. XGBoost further elevates the accuracy to 96.185%, signifying a notable enhancement over previous techniques. Similarly, Bootstrap, another ensemble approach, attains an accuracy of 94.236%. However, the most remarkable advancement is evident with the proposed AXG-B+BS-B hybrid approach, which achieves the highest accuracy of 96.987%. This signifies a substantial increase over individual techniques and even surpasses the ensemble methods. Moreover, the proposed AXG-B+BS-B technique exhibits consistently superior performance, with an average accuracy improvement of 1.127% over DT [32], 1.239% over RF [32], 1.17% over AdaBoost [32], 1.595% over GNB [32], 1.42% over LR [32], 1.183% over K-NN [32], 1.174% over XGBoost [32], 1.152% over GBoost [32], 1.046% over Adaboost, 1.04% over XGBoost, and 1.061% over Bootstrap Bagging techniques.

Table 3 Results comparison for proposed and existing heart disease detection techniques with Cleveland dataset

Techniques	Performance measures (%)					
	Accuracy	Precision	Recall	F-measure	Kappa	AUC
NB [33]	74.124	62.356	58.647	60.445	74.235	59.546
DT [33]	82.536	68.596	54.362	60.655	75.246	57.509
SVM [33]	95.556	71.526	54.235	61.692	77.024	57.963
NB [34]	83.000	75.623	68.985	72.152	65.345	70.568
K-NN [34]	80.000	76.235	79.989	78.067	66.389	79.028
LR [34]	85.000	74.214	84.569	79.054	78.986	81.811
SVM [34]	82.000	81.025	89.678	85.132	75.636	87.405
NB [31]	88.157	83.647	88.745	86.121	76.345	87.433
K-NN [31]	90.789	87.968	86.147	87.048	79.658	86.597
DT [31]	80.263	88.125	84.652	86.354	78.348	85.503
RF [31]	86.840	84.239	88.978	86.544	79.347	87.761

Adaboost	95.623	94.562	95.078	94.819	93.399	94.949
XGBoost	96.185	95.124	95.640	95.381	93.961	95.511
Bootstrap	94.236	93.689	94.018	93.853	94.589	93.935
AXG-B+BS-B	96.987	95.875	95.389	95.631	94.025	95.510

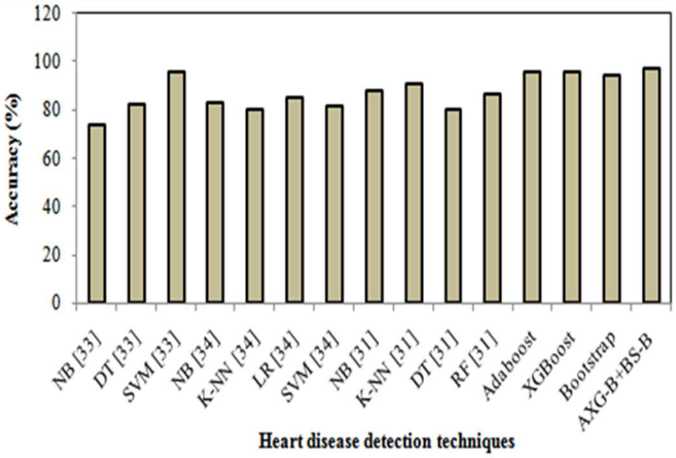

Fig. 2 Accuracy comparison for Cleveland dataset

The recall comparison among heart disease detection techniques using the Cleveland dataset provides in Fig. 4.. Among the existing techniques, SVM [33] demonstrates a recall of 54.235%, whereas NB [33] and DT [33] yield recalls of 58.647% and 54.362%, respectively. Additionally, NB [34], K-NN [34], and LR [34] exhibit recalls of 68.985%, 79.989%, and 84.569%, respectively, reflecting their varying performance levels. The more recent study by authors [31] portrays an upward trend in recall. Particularly, NB [31] attains a recall of 88.745%, while K-NN [31], DT [31], and RF [31] achieve recalls of 86.147%, 84.652%, and 88.978%, respectively. Ensemble techniques significantly enhance recall performance. Adaboost achieves a recall of 95.078%, displaying a substantial increase compared to individual methods. XGBoost further elevates the recall to 95.640%, signifying a remarkable improvement over previous techniques. Similarly, Bootstrap, another ensemble method, attains a recall of 94.018%. Yet, the most remarkable advancement is observed with our AXG-B+BS-B hybrid approach, achieving the highest recall of 95.389%. Furthermore, AXG-B+BS-B technique consistently demonstrates superior performance, with an average recall improvement of 6.922% over DT [32], 6.542% over RF [32], 6.409% over AdaBoost [32], 6.852% over GNB [32], 5.81% over LR [32], 5.242% over K-NN [32], 5.596% over XGBoost [32], 6.075% over GBoost [32], 6.311% over Adaboost, 5.779% over XGBoost, and 6.371% over Bootstrap Bagging techniques.

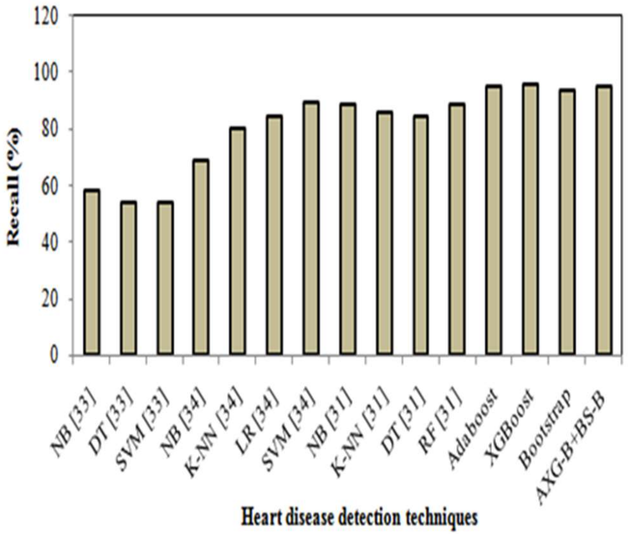

Fig. 3 Recall comparison for Cleveland dataset

The most notable advancement is observed with the AXG-B+BS-B hybrid approach, achieving the highest Kappa of 94.025%. This represents a substantial increase over individual techniques and even surpasses ensemble methods. Furthermore, the proposed AXG-B+BS-B technique consistently demonstrates superior performance, with an average Kappa improvement of 15.274% over DT [32], 14.779% over RF [32], 15.542% over AdaBoost [32], 16.744% over GNB [32], 14.501% over LR [32], 12.997% over K-NN [32], 15.971% over XGBoost [32], 15.214% over GBoost [32], 15.22% over Adaboost, 15.448% over XGBoost, and 15.436% over Bootstrap Bagging techniques.

The AUC comparison among different heart disease detection techniques using the Cleveland dataset demonstrates a substantial increase over individual techniques and even surpasses ensemble methods. Furthermore, AXG-B+BS-B technique consistently demonstrates superior performance, with an average AUC improvement of 16.664% over DT [32], 16.161% over RF [32], 17.441% over AdaBoost [32], 17.292% over GNB [32], 16.946% over LR [32], 14.907% over K-NN [32], 17.69% over XGBoost [32], 17.575% over GBoost [32], 17.227% over Adaboost, 16.866% over XGBoost, and 17.575% over Bootstrap Bagging technique

4.3 Results analysis for Framingham's dataset

Table 4 presents a concise comparison of the results achieved by the proposed heart disease detection technique and existing methods using the Framingham's dataset. The table encapsulates the performance evaluation outcomes in terms of various metrics, providing a clear overview of how the proposed technique fares against established approaches. Fig. 4 shows the comparison of accuracy for heart disease detection techniques using the Framingham's dataset. Among the existing techniques, DT [32] demonstrates an accuracy of 88.700%, while RF [32] and AdaBoost [32] yield accuracies of 80.700% and 85.500%, respectively. GNB [32], LR [32], and K-NN [32] exhibit accuracies of 62.700%, 70.400%, and 84.500%, respectively. XGBoost [32] and GBoost [32] achieve accuracies of 85.200% and 86.800%, respectively, reflecting a diverse range of performance levels. Adaboost achieves an accuracy of 95.623%, indicating a considerable increase compared to individual methods. XGBoost further elevates the accuracy to 96.185%, signifying a substantial improvement over previous techniques. Bootstrap, another ensemble approach, attains an accuracy of 94.236%. However, the most notable advancement is observed with the proposed AXG-B+BS-B hybrid approach, which achieves the highest accuracy of 96.987%. AXG-B+BS-B technique consistently demonstrates superior performance, with an average accuracy improvement of 1.284%, 2.987%, 1.487%, 34.287%, 26.587%, 2.487%, 11.787%, 10.187%, 10.364%, 11.921%, 11.852%, 1.763%, and 2.774% better than DT [32], RF [32], AdaBoost [32], GNB [32], LR [32], K-NN [32], XGBoost [32], GBoost [32], Adaboost, XGBoost, and Bootstrap Bagging techniques, respectively.

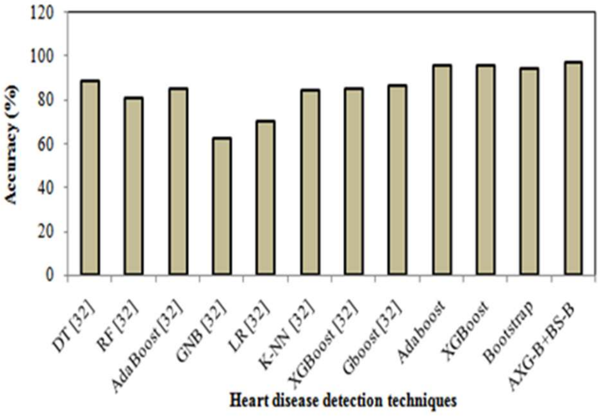

Fig. 4 Accuracy comparison for Framingham's dataset

Table 4 Results comparison for proposed and existing heart disease detection techniques with Framingham's dataset

Techniques	Performance measures (%)					
	Accuracy	Precision	Recall	F-measure	Kappa	AUC
DT [32]	88.700	90.700	89.100	88.600	89.275	88.850
RF [32]	80.700	90.800	90.800	90.700	88.250	90.750
AdaBoost [32]	85.500	86.800	86.000	85.500	85.950	85.750
GNB [32]	62.700	65.800	63.600	63.600	63.925	63.600
LR [32]	70.400	70.400	70.400	70.400	70.400	70.400
K-NN [32]	84.500	85.100	84.500	84.500	84.650	84.500
XGBoost [32]	85.200	86.600	85.700	85.200	85.675	85.450
GBoost [32]	86.800	88.600	87.300	86.700	87.350	87.000
Adaboost	95.623	94.562	95.078	94.819	93.399	94.949
XGBoost	96.185	95.124	95.640	95.381	93.961	95.511
Bootstrap	94.236	93.689	94.018	93.853	94.589	93.935
AXG-B+BS-B	96.987	95.875	95.389	95.631	94.025	95.510

Comparing the precision performance among different heart disease detection techniques using the Framingham's dataset is offers in Fig. 5. Among the existing methods, DT [32] achieves a precision of 90.700%, while RF [32] and AdaBoost [32] yield precision values of 90.800% and 86.800%, respectively. GNB [32], LR [32], and K-NN [32] exhibit precision rates of 65.800%, 70.400%, and 85.100%, respectively. XGBoost [32] and GBoost [32] attain precision values of 86.600% and 88.600%, respectively, showcasing a diverse range of performance levels. Ensemble techniques play a pivotal role in enhancing precision. However, the most remarkable advancement is observed with the proposed AXG-B+BS-B hybrid approach, achieving the highest precision of 95.875%. Furthermore, the proposed AXG-B+BS-B technique consistently demonstrates superior performance, with an average precision improvement of 4.175%, 4.075%, 9.075%, 46.075%, 25.475%, 4.775%, 9.475%, 17.275%, 1.313%, 1.275%, 1.186%, and 1.186% better than DT [32], RF [32], AdaBoost [32], GNB [32], LR [32], K-NN [32], XGBoost [32], GBoost [32], Adaboost, XGBoost, and Bootstrap Bagging techniques, respectively.

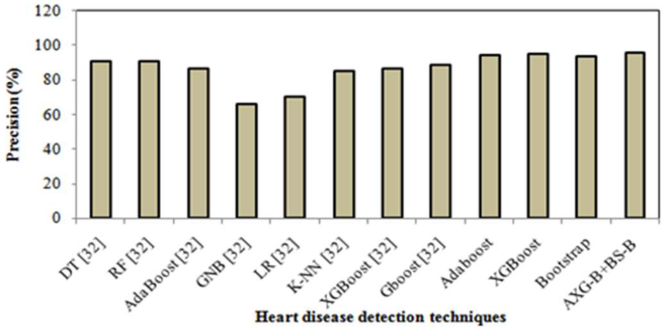

Fig. 5 Precision comparison for Framingham's dataset

The most significant advancement is observed with the proposed AXG-B+BS-B hybrid approach, which achieves the highest recall of 95.389%. Furthermore, the proposed AXG-B+BS-B technique consistently showcases superior recall performance, with an average recall improvement of 6.289%, 5.589%, 9.389%, 49.789%, 25.219%, 5.889%, 10.789%, 18.589%, 0.469%, 0.949%, 1.371%, and 1.130% better than DT [32], RF [32], AdaBoost [32], GNB [32], LR [32], K-NN [32], XGBoost [32], GBoost [32], Adaboost, XGBoost, and Bootstrap Bagging techniques, respectively.

The most significant advancement is observed with the proposed AXG-B+BS-B hybrid approach, which achieves the highest F-measure of 95.631%. Furthermore, the proposed AXG-B+BS-B technique consistently showcases superior F-measure performance, with an average F-measure improvement of 6.789%, 4.931%, 10.131%, 32.031%, 24.331%, 5.789%, 10.431%, 18.931%, 0.609%, 1.251%, 1.423%, and 0.775% better than DT [32], RF [32], AdaBoost [32], GNB [32], LR [32], K-NN [32], XGBoost [32], GBoost [32], Adaboost, XGBoost, and Bootstrap Bagging techniques, respectively.

The most remarkable enhancement is achieved by the proposed AXG-B+BS-B hybrid approach, which achieves the highest Kappa statistic of 94.025%. Moreover, the proposed AXG-B+BS-B technique consistently showcases superior Kappa performance, with an average Kappa improvement of 5.859%, 5.775%, 8.075%, 30.100%, 23.625%, 9.175%, 10.275%, 17.975%, 0.375%, and 0.875% better than DT [32], RF [32], AdaBoost [32], GNB [32], LR [32], K-NN [32], XGBoost [32], GBoost [32], XGBoost, and Bootstrap Bagging techniques, respectively.

Ensemble techniques consistently elevate the AUC performance. Adaboost achieves an AUC of 94.949%, showcasing an increase compared to individual methods. XGBoost further enhances the AUC to 95.511%, emphasizing its significant discriminatory power. Bootstrap, an ensemble approach, achieves an AUC of 93.935%. Yet, the most remarkable advancement is evident in the proposed AXG-B+BS-B hybrid approach, achieving the highest AUC of 95.510%. Furthermore, the proposed AXG-B+BS-B technique consistently demonstrates superior AUC performance, boasting an average AUC improvement of 5.15%, 4.76%, 8.76%, 31.91%, 25.18%, 9.66%, 10.02%, and 17.51% over DT [32], RF [32], AdaBoost [32], GNB [32], LR [32], K-NN [32], XGBoost [32], and GBoost [32] techniques, respectively.

5. Conclusion and future work

We present a novel approach for the early detection of heart disease, encompassing feature optimization and ensemble learning through a hybrid bagging and boosting methodology. Our method integrates an enhanced U-Net pre-trained architecture to facilitate efficient feature extraction. To counteract challenges associated with data dimensionality, we introduce a modified binary search (MBS) algorithm tailored for feature optimization. Additionally, we introduce an ensemble learning paradigm that merges the strengths of both bagging and boosting techniques to enhance heart disease detection and classification. Our approach is rigorously evaluated on the Cleveland and Framingham's heart study datasets to assess its performance against existing methods. Upon analyzing the simulation outcomes, our proposed hybrid bagging and boosting technique (AXG-B+BS-B) attains an impressive accuracy of 96.987% for both the Cleveland and Framingham's datasets. Comparatively, this technique outperforms existing methods by 5.345% and 6.212% for the Cleveland and Framingham's datasets, respectively, in terms of disease detection accuracy. These results not only validate the efficacy of our method but also highlight its potential to significantly enhance heart disease prediction capabilities.

Further research should be done to increase the classification accuracy through the use of advanced algorithms. Also can add more features needed to improve the accuracy implementation of algorithms. Medical practioners can use it as dedicated tool to make better decisions. In future, it can be improved by changing the parameters for the experiment, more work can be done by using more data related to heart disease and by using different data reduction techniques.

References

1. Bizopoulos, P. and Koutsouris, D., 2018. Deep learning in cardiology. IEEE reviews in biomedical engineering, 12, pp.168-193.
2. Mdhaffar, A., Rodriguez, I.B., Charfi, K., Abid, L. and Freisleben, B., 2017. CEP4HFP: Complex event processing for heart failure prediction. IEEE transactions on nanobioscience, 16(8), pp.708-717.

3. Guo, H., Kruger, U., Wang, G., Kalra, M.K. and Yan, P., 2019. Knowledge-based analysis for mortality prediction from CT images. IEEE journal of biomedical and health informatics, 24(2), pp.457-464.
4. Kutyrev, K., Yakovlev, A. and Metsker, O., 2019. Mortality prediction based on echocardiographic data and machine learning: Chf, chd, aneurism, acs cases. Procedia Computer Science, 156, pp.97-104.
5. Balabaeva, K. and Kovalchuk, S., 2019. Comparison of temporal and non-temporal features effect on machine learning models quality and interpretability for chronic heart failure patients. Procedia Computer Science, 156, pp.87-96.
6. Kumar, P.M. and Gandhi, U.D., 2018. A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases. Computers & Electrical Engineering, 65, pp.222-235.
7. 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, 16, p.100203.
8. Mezzatesta, S., Torino, C., De Meo, P., Fiumara, G. and Vilasi, A., 2019. A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. Computer methods and programs in biomedicine, 177, pp.9-15.
9. Abdar, M., Książek, W., Acharya, U.R., Tan, R.S., Makarenkov, V. and Pławiak, P., 2019. A new machine learning technique for an accurate diagnosis of coronary artery disease. Computer methods and programs in biomedicine, 179, p.104992.
10. Al-Makhadmeh, Z. and Tolba, A., 2019. Utilizing IoT wearable medical device for heart disease prediction using higher order Boltzmann model: A classification approach. Measurement, 147, p.106815.
11. Piros, P., Ferenci, T., Fleiner, R., Andréka, P., Fujita, H., Főző, L., Kovács, L. and Jánosi, A., 2019. Comparing machine learning and regression models for mortality prediction based on the Hungarian Myocardial Infarction Registry. Knowledge-Based Systems, 179, pp.1-7.
12. Dwivedi, A.K., 2018. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Computing and Applications, 29, pp.685-693.
13. Venkatesh, R., Balasubramanian, C. and Kaliappan, M., 2019. Development of big data predictive analytics model for disease prediction using machine learning technique. Journal of medical systems, 43, pp.1-8.
14. Narayan, S. and Sathiyamoorthy, E., 2019. A novel recommender system based on FFT with machine learning for predicting and identifying heart diseases. Neural Computing and Applications, 31, pp.93-102.
15. Vijayashree, J. and Sultana, H.P., 2018. A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier. Programming and Computer Software, 44, pp.388-397.

16. Ahmed, H., Younis, E.M., Hendawi, A. and Ali, A.A., 2020. Heart disease identification from patients' social posts, machine learning solution on Spark. Future Generation Computer Systems, 111, pp.714-722.
17. Mienye, I.D., Sun, Y. and Wang, Z., 2020. An improved ensemble learning approach for the prediction of heart disease risk. Informatics in Medicine Unlocked, 20, p.100402.
18. Ali, F., El-Sappagh, S., Islam, S.R., Kwak, D., Ali, A., Imran, M. and Kwak, K.S., 2020. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Information Fusion, 63, pp.208-222.
19. Sharma, P., Choudhary, K., Gupta, K., Chawla, R., Gupta, D. and Sharma, A., 2020. Artificial plant optimization algorithm to detect heart rate & presence of heart disease using machine learning. Artificial intelligence in medicine, 102, p.101752.
20. Beecy, A.N., Gummalla, M., Sholle, E., Xu, Z., Zhang, Y., Michalak, K., Dolan, K., Hussain, Y., Lee, B.C., Zhang, Y. and Goyal, P., 2020. Utilizing electronic health data and machine learning for the prediction of 30-day unplanned readmission or all-cause mortality in heart failure. Cardiovascular Digital Health Journal, 1(2), pp.71-79.
21. Irene, D.S., Sethukarasi, T. and Vadivelan, N., 2020. Heart disease prediction using hybrid fuzzy K-medoids attribute weighting method with DBN-KELM based regression model. Medical Hypotheses, 143, p.110072.
22. Aggrawal, R. and Pal, S., 2020. Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease. SN Computer Science, 1(6), p.344.
23. Patro, S.P., Padhy, N. and Chiranjevi, D., 2021. Ambient assisted living predictive model for cardiovascular disease prediction using supervised learning. Evolutionary intelligence, 14(2), pp.941-969.
24. Tougui, I., Jilbab, A. and El Mhamdi, J., 2020. Heart disease classification using data mining tools and machine learning techniques. Health and Technology, 10, pp.1137-1144.
25. Al-Yarimi, F.A.M., Munassar, N.M.A., Bamashmos, M.H.M. and Ali, M.Y.S., 2021. Feature optimization by discrete weights for heart disease prediction using supervised learning. Soft Computing, 25, pp.1821-1831.
26. Li, P., Hu, Y. and Liu, Z.P., 2021. Prediction of cardiovascular diseases by integrating multi-modal features with machine learning methods. Biomedical Signal Processing and Control, 66, p.102474.
27. Rani, P., Kumar, R., Ahmed, N.M.S. and Jain, A., 2021. A decision support system for heart disease prediction based upon machine learning. Journal of Reliable Intelligent Environments, 7(3), pp.263-275.
28. Zhang, J., Zhu, H., Chen, Y., Yang, C., Cheng, H., Li, Y., Zhong, W. and Wang, F., 2021. Ensemble machine learning approach for screening of coronary heart disease based on echocardiography and risk factors. BMC Medical Informatics and Decision Making, 21(1), pp.1-13.

29. Mehmood, A., Iqbal, M., Mehmood, Z., Irtaza, A., Nawaz, M., Nazir, T. and Masood, M., 2021. Prediction of heart disease using deep convolutional neural networks. Arabian Journal for Science and Engineering, 46(4), pp.3409-3422.
30. Biswas, N., Ali, M.M., Rahaman, M.A., Islam, M., Mia, M.R., Azam, S., Ahmed, K., Bui, F.M., Al-Zahrani, F.A. and Moni, M.A., 2023. Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. BioMed Research International, 2023.
31. Shah, D., Patel, S. and Bharti, S.K., 2020. Heart disease prediction using machine learning techniques. SN Computer Science, 1, pp.1-6.
32. Jinny, S.V. and Mate, Y.V., 2021. Early prediction model for coronary heart disease using genetic algorithms, hyper-parameter optimization and machine learning techniques. Health and Technology, 11, pp.63-73.
33. Deepika, K. and Seema, S., 2016, July. Predictive analytics to prevent and control chronic diseases. In 2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT) (pp. 381-386). IEEE.
34. Dwivedi, A.K., 2018. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Computing and Applications, 29, pp.685-693.