Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 7
July 2023

Improving Heart Disease Prediction through Feature Selection for Multi-Label Classification

Kavitha Chandrashekar 1*

Anitha Tuluvanooru Narayanreddy 2

1Department of Computer Science, Associate Professor, Vijaya Vittala Institute of Technology, Bangalore, Karnataka, India

2 Professor and Head, Department of Computer Science and Engineering, Sir Mokshagundam Visvesvaraya Institute of Technology, Bangalore, Karnataka, India

* Corresponding author's Email: ckavitac22@gmail.com

Abstract

Heart disease is a leading cause of death worldwide, and early prediction and diagnosis are crucial for effective treatment. In this study, we propose a novel approach for heart disease prediction by using feature selection, XGBoost, Ensemble-Feature-Optimization, and a multi-label classification method. The proposed model aims to improve the accuracy of heart disease prediction by selecting relevant features from the dataset, optimizing the feature ensemble, and applying a multi-label classification method to handle the multiple diseases associated with heart disease. To evaluate the performance of the proposed model, we used two datasets, the Cleveland and Statlog heart disease datasets, which consist of patients with various attributes, such as age, sex, blood pressure, and cholesterol levels. We compared the performance of our proposed model with other machine learning state-of-the-art approaches, using various performance metrics, including accuracy, precision, recall, specificity, and F1-score.The experimental results show that our proposed model outperforms other state-of-the-art approaches in terms of prediction accuracy and other performance metrics. The proposed model achieves an accuracy of 99.5% on the Cleveland dataset and 99.95% on the Statlog dataset. Our approach offers a promising method for heart disease prediction and diagnosis, and the results demonstrate the potential of this approach for improving heart disease treatment and management.

1 Introduction

Heart disease is a prevalent health condition that affects millions of people worldwide and is a leading cause of death [1]. Early prediction and diagnosis of heart disease are crucial for effective treatment and management. With the advancement of machine learning algorithms, predicting heart disease using patient data has become possible [2]. Machine learning algorithms can help identify relevant patterns and relationships between patient data and heart disease outcomes [3].In recent years, several machine learning algorithms have been applied to heart disease prediction, including Decision-Trees (DT) [4], Random-Forests (RF) [5], Support-Vector-Machines (SVM) [6], and Neural-Networks [7]. Heart disease prediction using machine learning algorithms has shown great potential in providing accurate and timely diagnosis, but there are several challenges that need to be addressed to ensure the effectiveness and reliability of the prediction models [8]. One of the major challenges is the lack of high-quality data for training the models. The accuracy and reliability of the prediction models heavily depend on the quality and quantity of data used for training. There may be issues with missing or incomplete data, noisy data, or biased data, which can lead to inaccurate predictions [9]. Another issue is the choice of features and algorithms. Selecting the most relevant features and algorithms that can effectively capture the underlying patterns and relationships between the data and heart disease outcomes is critical for the accuracy and performance of the models [10]. However, the selection process can be complex and requires domain expertise and knowledge.Addressing these

challenges is critical to ensuring the effectiveness and reliability of machine learning algorithms for heart disease prediction and improving the diagnosis and treatment of heart disease.

Feature optimization methods are increasingly being used to address some of the issues in heart disease prediction using machine learning algorithms. For example, feature selection [10] and feature ensemble optimization methods [11] can help address the issue of lack of high-quality data by identifying the most relevant features from the dataset that are crucial for the accurate prediction of heart disease. By selecting only the most informative and relevant features, these methods can reduce the impact of noise, missing or biased data and improve the accuracy of the models. Feature optimization methods can also help with the issue of selecting the most appropriate algorithms and features [12]. By identifying the best features to use for prediction, feature selection methods can help to select the most relevant input features for the algorithms, which can lead to improved accuracy and efficiency of the models [13]. Similarly, feature ensemble optimization methods can help capture complementary information from different subsets of features, leading to improved accuracy and robustness of the model [14]. Moreover, feature optimization methods can help with the interpretability of the models by identifying the most relevant and informative features used for prediction. This can help improve the transparency and explain-ability of the models, which can be crucial for clinical decision-making and acceptance in the medical community [15]. In summary, feature optimization methods have the potential to overcome several issues in heart disease prediction using machine learning algorithms by identifying the most relevant features, selecting appropriate algorithms, and improving the interpretability of the models. These methods can ultimately lead to more accurate and efficient heart disease prediction models and help improve the diagnosis and treatment of heart disease [16]. Hence, in this paper, we propose a novel approach for heart disease prediction by using feature selection, XGBoost, Ensemble-Feature-Optimization, and a multi-label classification method. The contributions of the proposed work are given as follows:

1. The proposed model has been presented utilizing the XGBoost Algorithm, Ensemble-Feature-Optimization, and multi-label classification method to provide better prediction accuracy for heart disease prediction.
2. The proposed model presents a novel feature selection method for the selection of important features from the dataset.
3. The proposed model optimizes the XGBoost tree construction and provides the best tree-structure for heart disease prediction.
4. The proposed model improves the accuracy of prediction for heart disease prediction using the multi-label classification model.

2 Literature Survey

In this section, a study on the existing works which utilize the feature selection method for the classification of the heart-diseases has been conducted. In [17], they examine the performance of many popular classification methods under the supervision of a predetermined collection of features. Feature-selection is crucial because it minimizes considerably the duration and money spent on training prediction models by filtering out unnecessary data. Logistic-Regression, Random Forest, Naive Bayes, and Extra Trees were all considered classification algorithms, having features identified utilizing Least-Absolute-Shrinkage and Selection-Operator (LASSO) and Ridge-Regression. When feature-selection was applied to classifiers, they became significantly more accurate. When compared to the results obtained through Ridge-Regression, the results obtained using Lasso regression are, on average, 33.3% more accurate than those obtained using Ridge-Regression.Using appropriate features selected utilizing a variety of feature-selection approaches, the authors of [18] conducted an in-depth assessment of the effectiveness of algorithms created through methods based on machine learning. Symmetrical-Uncertainty, ReliefF, Chi-Square, and Principal-Component-Analysis (PCA), were employed to analyze four commonly utilized heart disease datasets and

generate unique feature sets. Then, several classification methods were utilized to build models, which were then evaluated to find the best possible feature sets for accurately predicting problems with the heart. This study shows that the advantages of selecting features change depending on the ML method applied to the various heart datasets they analyzed. Their top approach, nevertheless, combined the Chi-squared selection of features alongside the BayesNet algorithm to yield an accuracy rate of 85 percent across all datasets.In order to keep improving the performance of machine learning methods, the authors of [19] use techniques based on evolution including the Genetic-Algorithm (GA) as well as Particle-Swarm-Optimization (PSO) to conduct feature-selection. Both PSO and GA were used in selecting features alongside J48, Support-Vector-Machine (SVM), and Naive Bayes (NB). Once the most important features have been chosen, the selection of features algorithm's performance is measured through implementing methods from machine learning to both the whole dataset and the trimmed-down version that comprises the dataset. The accuracy of selecting features strategies for predicting heart disease has been evaluated using five distinct machine learning methods: Naive-Bayes (NB), Support-Vector-Machines (SVM), Decision Trees (DT), Logistic-Regression (LR), and the Random-Forest (RF) algorithm. According to the findings, GA is the best method for selecting features because it improves accuracy in predictions more than any of the others.

In [20], In this research, they utilize an optimized selection of features to create a heart disease model for prediction that is both accurate and efficient. In the first stage of processing, data is cleaned, transformed, values that are missing are imputed, and the information is normalized. Following that, the best characteristics are chosen using the Decision-making Function-based Chaotic-Salp-Swarm (DFCSS) algorithm. Next, the Improved-Elman-Neural-Network (IENN) utilizes the attributes that were chosen in order to classify the data. In this case, the optimum weight for the IENN is calculated using the Sail-Fish-Optimization (SFO) technique. Predicting heart disease is made easier with the DFCSS-IENN-based SFO method. Cleveland dataset and the Cardio-Vascular-Disease (CVD) dataset are used for implementing the intended (DFCSS-IESFO) technique within the Python environment. Simulation results demonstrated that the suggested approach outperformed different classifiers including the SVM, K-Nearest-Neighbor, Elman-Neural-Network, Decision-Tree, Gaussian Naive-Bayes, Random-Forest, and Logistical egression by a wide margin when applied to the CVD dataset and by a similar margin when applied to the Cleveland dataset.In the current investigation, researchers have created an automated method based on an evaluation of the efficacy of numerous machine learning strategies, as described in [21]. The dataset was initially utilized by popular machine learning methods for predicting diseases, including RF, NB, KNN, and SVM. Using 3-way cross-validation helps remove the possibility of bias in the results. NB achieves the highest average accuracy at 87.78 percent. It may be said that the model's performance is satisfactory. They have also used a GA to maximize the features of the dataset. After being optimized, the NB has the maximum average accuracy of 96 percent.This study [22] created a feature-selection-based ensemble classification framework where only a subset of characteristics contributes to the classifying process. As a result, a classification strategy was developed for heart disease detection by combining ensemble training, a GA, selecting features, and biological testing values. From this data, researchers infer that the feature-selecting method's value changes depending on the specific predictive modeling approach that has been used. The most effective proposed approach, nevertheless, relies on a hybrid of the GA and the ensemble learning approach, which achieves an accuracy of 97.57 percent on the datasets under consideration. The proposed diagnostic system successfully identified heart disease with higher accuracy than any other approach proposed to date.

3 Methodology

The section introduces a methodology of an ensemble feature optimization model for designing an effective heart disease prediction model. In this

proposed machine learning model for heart disease prediction, we have modified the process of feature selection during the XGBoost algorithm using the minimized objective function. Figure 5.1 shows the architecture of the proposed model.

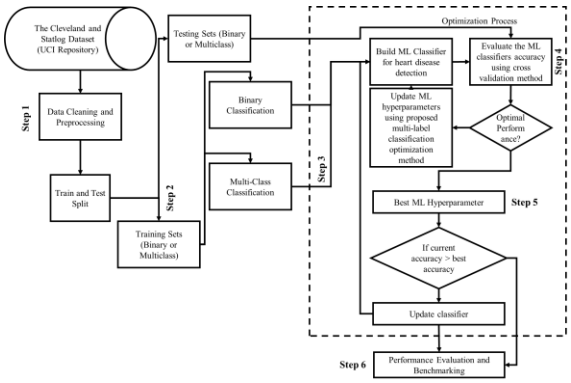

Figure 1. Architecture for the proposed model.

3.1 Data Preprocessing

Data preprocessing is an important step in machine learning, and it is especially crucial when using the XGBoost algorithm. The Cleveland and Statlog datasets require several preprocessing steps before they can be used with XGBoost. The first step in preprocessing the data is to load the dataset into a data frame and check for missing values. If there are any missing values, they can be dropped or imputed using appropriate techniques. The next step is to encode categorical variables into numerical values using methods such as one-hot encoding. After encoding the categorical variables, the dataset can be split into training and testing sets using appropriate libraries such as sci-kit-learn. It is crucial to ensure that the data is split randomly to avoid any biases. Standardizing the features to have a mean of 0 and a standard deviation of 1 can help improve the performance of the XGBoost algorithm.The next step is to perform feature selection to select the most important features. In this work, the feature selection technique which has been used has been presented in the below section which has been used to select the most important features.

3.2 XGBoost Prediction Algorithm

In this section, the XGBoost Algorithm which has been used for heart-disease prediction has been discussed. In the XGBoost algorithm, first, we take the dataset and analyze it. After analyzing it, the complete dataset is represented as$E = \{(y_j, z_j); j = 1 \dots o, y_j \in \mathcal{S}^n, z_j \in \mathcal{S}\}$,where, o is the data sample which has all the n features. Further, in this work, we use $\hat{z}_j$ for representing the final outcome of the prediction. The $\hat{z}_j$ is evaluated by the given below equation

$$\hat{z}_j = \sum_{l-1}^{L} g_l(y_j), \;\; g_l \in G$$

where, g_l is used for representing the various distinct regression-trees, $g_l(y_j)$ is used for representing the outcome of the prediction which has been acquired through the l[th]-tree using the j[th]-sample of the dataset. The g_l and all its corresponding features are attained by utilizing the minimization model. The minimization model can be represented using the given below equation

$$\mathcal{O} = \sum_{j=1}^{o} m(z_j, \hat{z}_j) + \sum_{l=1}^{L} \beta(g_l)$$

where, m is used for representing the loss-operation during the training that evaluates the difference among the actual-value z_j and forecasted-value $\hat{z}_j$. Overfitting can be prevented by employing the variable β, that penalizes the presented model's complexity. Further the β is defined using the given below equation

$$\beta(g_l) = \delta U + \frac{1}{2}\mu\|x\|^2$$

where, μ and δ are the two variables which have been used as the regularization-variables, U is used for denoting the leaf-size, x is used for denoting the various scores attained by different leaves. The summation method is used to generate the ensemble tree. Further, in this work, we have considered a variable $\hat{z}_j^{(u)}$ that is used for denoting the outcome of the prediction for the u[th]-tree using the j[th]-sample of the dataset. Also, in this work, the distinct regression-trees g_u have been utilized and has been added to the objective-function represented as $\mathcal{O}^{(u)}$ to reduce the $\mathcal{O}^{(u)}$. This is represented by the given below equation

$$\mathcal{O}^{(u)} = \sum_{j=1}^{o} m\left(z_j, \hat{z}_j^{(u-1)} + g_u(y_j)\right) + \beta(g_l)$$

The Equation (4) has been further reduced by discarding the constant parameters through the Second-Order Taylor-Expansion. After the expansion the following equation is attained

$$\mathcal{O}^{(u)} = \sum_{j=1}^{o} \left[h_j g_j(y_j) + \frac{1}{2} i_j g_u(y_j)^2\right] + \beta(g_l)$$

where, h_j denotes the gradient having the first-order on the basis of m. The h_j is represented using the given below equation

$$h_j = \partial_{\hat{z}_j^{(u-1)}} m\left(z_j, \hat{z}_j^{(u-1)}\right) \tag{6}$$

where, i_j denotes the gradient having the first-order on the basis of m. The i_j is represented using the given below equation

$$i_j = \partial^2_{\hat{z}_j^{(u-1)}} m\left(z_j, \hat{z}_j^{(u-1)}\right) \tag{7}$$

From this the $\mathcal{O}^{(u)}$ for predicting the heart-disease is attained using the following equation

$$\mathcal{O}^{(u)} = \sum_{j=1}^{o} \left[h_j g_j(y_j) + \frac{1}{2} i_j g_u(y_j)^2\right] + \delta U + \frac{1}{2}\mu \sum_{k=1}^{U} x_k^2 \tag{8}$$

Simplifying the Equation (8), the given below equation is attained

$$\mathcal{O}^{(u)} == \sum_{j=1}^{U} \left[\left(\sum_{j\in J_k} h_j\right) x_j \frac{1}{2}\left(\sum_{j\in J_k} i_j + \mu\right) x_k^2\right] + \delta U \tag{9}$$

where, J_k is used for denoting the set of samples of leaves of the gradient-tree k. The J_k is represented using the given below equation

$$J_k = \{j | r(y_j = k)\} \tag{10}$$

where, r is used for denoting the static size of the k. Further, the weights of the leaves j which have been optimized represented as x_k is denoted as given in the below equation (11)

$$x_k = \frac{H_k}{I_k + \mu}$$

The tree-size which has been optimized is accumulated from the below equation

$$\mathcal{O}^s = \frac{1}{2}\sum_{k-1}^{U} \frac{H_k^2}{I_k + \mu} + \delta U \tag{12}$$

where, H_k is represented using the below equation

$$H_k = \sum_{j\in J_k} h_j \tag{13}$$

Also, the I_k is represented using the below equation

$$I_k = \sum_{j\in J_k} i_j$$

The tree-size which has been optimized is represented as $\mathcal{O}^s$. The $\mathcal{O}^s$is used for defining the complete structured tree and the tree quality which is represented using r. When the value of the r is small, the tree-structure will be the best. It has to be noted that when the given dataset is imbalanced or during the process of selection of features if the features are not selected properly, then the XGBoost model won't provide higher accuracy. Hence, to resolve this issue, an Ensemble Feature Optimization (EFO) has been provided in the next section.

3.3 Ensemble Feature Optimization

The process of selection of features in the XGBoost Algorithm is performed by utilizing the K-fold cross-validation process which is given using the following equation

$$CV(\sigma) = \frac{1}{M}\sum_{k=1}^{K} \sum_{j\in G\_k} P\left(b_j, \hat{g}_\sigma^{-k(j)}(y_j, \sigma)\right) \tag{14}$$

Nevertheless, while using the K-fold cross-validation process there exists some limitations, i.e., it does not provide the best relationship for the various features which are present in the heart-disease dataset. Hence, due to this the accuracy for the prediction of the heart-disease can be

impacted. Hence, in this work, the K-fold cross-validation process which has been modified is proposed and is represented using the following equation

$$CV(\sigma) = \frac{1}{SM}\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{j\in G_{-k}} P\left(b_j, \hat{g}_{\sigma}^{-k(j)}(y_j, \sigma)\right)$$

where, M is used for denoting the training-size for the prediction of heart-disease, $P(\cdot)$ is used for denoting the loss-function and $\hat{g}_{\sigma}^{-k(j)}(\cdot)$ has been used for denoting the co-efficient-functions. By using the K-fold cross-validation process which has been modified, initially, the feature subset is attained and further the features which have been selected are utilized for the construction of an ensemble-based prediction model for the heart disease by decreasing the error attained during the prediction in an iterative way. This is done using the following equation

$$\hat{\sigma} = \underset{\sigma\in\{\sigma_1,\ldots,\sigma_l\}}{\arg\min}\ CV_s(\sigma)$$

The presented prediction model for the heart disease provides accurate results for the binary classified datasets, but fails to provide better accuracy for the multi-label classified datasets. Hence, this work proposed a model which will provide better accuracy for the heart disease prediction when the dataset is multi-label classified. The multi-label classification model has been presented in the next section.

3.4 Multi-label classification Model

In this section, to provide better accuracy for the heart-disease prediction by selecting the most important features and by providing ranks whenever the dataset is of multi-label, a modelhas been presented. This model comprises of three steps. In the first step, the complete heart-disease dataset which is multi-label is considered and represented as E. Further, the dataset $E$is then segmented randomly into different$K-folds$ having similar sizes. Consider that the size of the $K-folds$ is 1, then the following variable E^{-k} has been defined to represent that the k^{th} heart-disease row samples have been discarded for training samples of the outer-level and E^{k} has been defined to represent the k^{th} heart-disease samples which have been considered and will be utilized for the validation of the testing samples of the outer-level. By configuring the S, the following steps which have been presented below are repeated.

(16)

1. For the prediction of the heart-disease, the dataset E^{-k} is further arbitrarily divided into $H-folds$ having similar sizes, i.e., $\forall h = 1$ to H.

A. Accomplish H by using the heart-disease dataset E^{-kh} where h[th] samples have been discarded for training dataset at the inner-level and E^{kh} where h[th]samples are considered for the testing dataset at the inner-level. $\forall l = 1$ has been used for denoting the size utilized for the formation of the grid during the selection of features process and ranking-process.

I. The prediction model $\hat{g}_{\sigma_l}$ for the heart-disease is constructed using the given below equation

$$\hat{g}_{\sigma_l} = \hat{g}\left(b_j, \hat{g}(E^{-kh}; \sigma_l)\right). \quad (17)$$

II. By utilizing the estimated loss-functional error which has been attained by the testing samples at the inner-level and also by using the prediction model $\hat{g}_{\sigma_l}$, the testing samples E^{kh} at the inner-level can be obtained using the given equation

$$\mathcal{E}_{\sigma_n} = \sum_{j\in E^{-kh}} P\left(b_j, \hat{g}(E^{-kh}; \sigma_l)\right)$$

B. For each l layer, the $H-folds$ errors have to be computed. Hence, every row of the M_h inside the l[th] layer for the k[th] sample can have various cross-validations. Hence, this is defined using the following equation

$$CV(\hat{g}; \sigma_l) = \frac{1}{M_h}\sum_{h=1}^{H}\sum_{j\in E^{-kh}} P\left(b_j, \hat{g}(E^{-kh}; \sigma_l)\right)$$

C. In an iterative way, by utilizing the S for various layers of l, the error in the cross-validation is evaluated for M_h inside $l$for the k[th] sample using the below equation

$$CV_S(\hat{g}; \sigma_l) = \frac{1}{M_h S}\sum_{s=1}^{S}\sum_{h=1}^{H}\sum_{j\in E^{-kh}} P\left(b_j, \hat{g}(E^{-kh}; \sigma_l)\right)$$

1 With each permutation of layer l, the model obtains a more convincing optimized feature, using the following equation

$$\hat{\sigma}_n = \underset{\sigma \in \{\sigma_1, \sigma_l\}}{\arg\min}\, CV_S(\hat{g}; \sigma_l) \tag{22}$$

2 Further, by utilizing the gradient-descent method, the optimized features are chosen using the minimization method. Using the ranking of features function $r(\cdot)$, the features are selected using the final features subset for designing the prediction model for heart-disease. The $r(\cdot)$ has been defined using the given below equation

$$r(a) = \begin{cases} 0 \;\; if\; n_j\; is\; not\; chosen \\ 1 \;\; if\; n_j\; is\; chosen\; as\; final\; predictive\; model\; j = 1,2,3,\dots,n \end{cases} \tag{23}$$

Using the above $r(\cdot)$, the following subset of all available features is determined using the following equation

$$F_s = \{r(n_1), r(n_1), \dots, r(n_n)\}, \tag{24}$$

For providing the best accuracy for the multi-label classification dataset, only the features which have the highest weight or highest rank (most important features are selected) on the basis of various $K - folds$ are selected using the following equation

$$F_{s_k} = \{r(n_1), r(n_1), \dots, r(n_n)\} \tag{25}$$

In the second step of this model, this work evaluates the samples for selecting the best features which are inside the final feature subset having higher rank. The best feature subset is chosen using the following equation

$$F_{s_{final}=\{f_s(p_1), f_s(n_2), \dots, f_s(n_n)\}}, \tag{26}$$

where, $f_s(\cdot)$ has been used for denoting whether the feature has been selected or not selected. The complete process of selection is given using the following equation

$$F_s(a) = \begin{cases} 0 \;\; if\; q_j\; is\; chosen\; lesser\; than\; \dfrac{K}{2}\; times, j = 1,2,3,\dots,n \\ 1 \;\; if\; q_j\; is\; chosen\; greater\; or\; equal\; to\; \dfrac{K}{2}\; times, j = 1,2,3,\dots,n \end{cases} \tag{27}$$

In the final step, the Equation (25) is used for establishing the best subset for the n'features which have been selected. In Equation (25), the n^{th} has been used for representing the overall selected features by considering that it will provide the important features for the prediction of the heart-disease. The results for the heart-disease prediction using the proposed model have been presented in the next section. The results show that the proposed model attains better performance in comparison to the existing models.

4 Results and Discussions

This section discusses the prediction of the heart disease by using the presented model and compares its results with the existing models [25]. To analyse the performance of the existing models and the presented model, the Cleveland dataset [23] as well as the Statlog dataset [24] have been used. Detailed description of the datasets has been given in the below sections. The Python-3 framework has been used for predicting the heart disease. Both the datasets have been data augmented for increasing the size of the dataset. The performance metrics (ROC) like specificity, sensitivity, f-measure, precision and accuracy have been used to validate the prediction which has been made. The following equation is used for evaluating the accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

The following equation is used for evaluating recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{29}$$

The following equation is used for evaluating specificity

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{30}$$

The following equation is used for evaluating precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{31}$$

The following equation is used for evaluating f-measure

$$\text{F} - \text{measure} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} \times \text{Sensitivity}}$$

4.1 Cleveland Dataset

The Cleveland dataset is a widely used dataset in the field of machine learning for predicting heart disease. It contains a total of 303 instances, with each instance having 14 attributes that provide information on clinical, demographic, and risk factor parameters. The target variable is a binary variable that indicates the presence or absence of heart disease. The dataset was collected from the Cleveland Clinic Foundation in the late 1980s and has been used for various research studies. The Cleveland dataset is commonly used as a benchmark dataset for heart disease prediction models and is freely available for research purposes. More information can be obtained from [23].

4.2 Statlog Dataset

The Statlog dataset is another commonly used dataset for predicting heart disease in machine learning research. It contains a total of 270 instances, with each instance having 13 attributes that provide information on clinical and demographic parameters. The target variable is also a binary variable that indicates the presence or absence of heart disease. The dataset was collected from four different medical centers in the United States and is also freely available for research purposes. The Statlog dataset has been used in several studies to evaluate the performance of machine learning algorithms. More information can be obtained from [24].

4.3 Performance Evaluation for the Cleveland Dataset

In this section, the performance evaluation for the Cleveland dataset has been conducted. In results have been shown in Figure 2. The results have been compared with the [25]. The results show that the proposed model has performed betterwhen compared with other models.

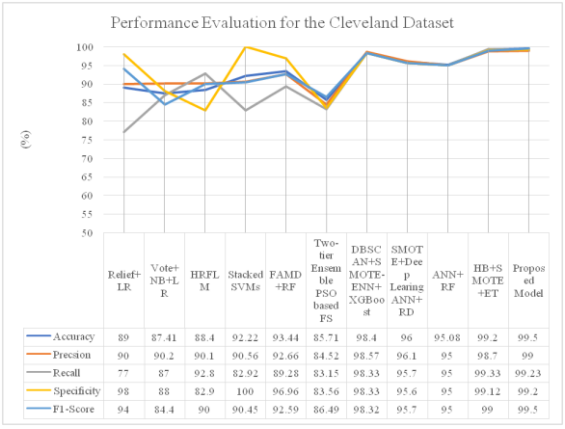

Figure 2. Performance Evaluation for the Cleveland Dataset.

4.4 Performance Evaluation for the Statlog Dataset

In this section, the performance evaluation for the Statlog dataset has been conducted. In results have been shown in Figure 3. The results have been compared with the [25]. The results show that the proposed model has performed better when compared with other models.

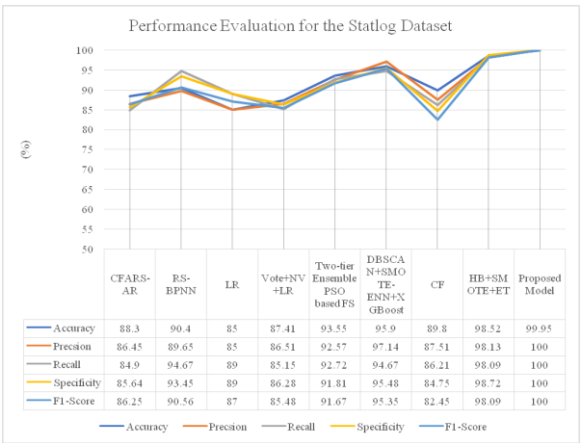

Figure 3. Performance Evaluation for the Statlog Dataset.

4.5 Comparative Study

In this section, the comparative study has been conducted. In this section, the proposed model has been compared with the existing SVM-EBO [26] and EFO [11] models. The results show that by using the feature optimization method and multi-label classification, the proposed model attains better result in terms of specificity, precision, recall, accuracy and F-Measure.

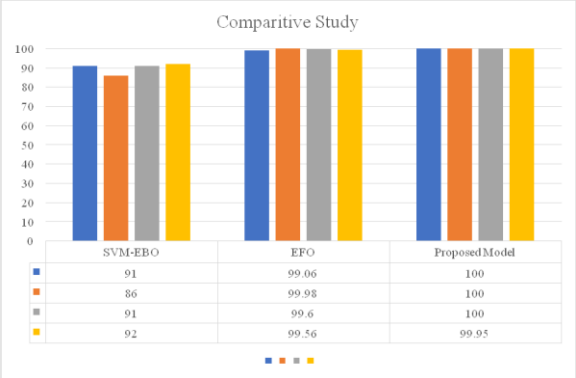

Figure 4. Comparative Study.

5 Conclusion

Heart disease prediction is an important area of research that has the potential to improve patient outcomes and reduce healthcare costs. Machine learning algorithms have shown promise in predicting heart disease, but there are also several challenges associated with this approach, including data quality, feature selection, and model accuracy. In this study, we investigated the use of feature optimization methods using XGBoost, ensemble feature optimization, and multi-label classification to improve the accuracy and efficiency of heart disease prediction.We used the Cleveland and Statlog datasets to train and evaluate our model, and we found that our approach resulted in a significant improvement in model performance compared to other state-of-the-art models. Our results showed that the use of feature optimization methods can help identify the most relevant features from the dataset. The results of our experiment demonstrate the potential of this approach for improving the accuracy and efficiency of heart disease prediction, which could have significant implications for patient care and clinical decision making. Overall, our study highlights the importance of using advanced machine learning algorithms and feature optimization methods for heart disease prediction. The results suggest that these approaches have the potential to significantly improve the accuracy and efficiency of heart disease prediction, and may have important implications for clinical practice and patient outcomes. Further research is needed to explore the use of these approaches in real-world settings and to assess their impact on patient care and healthcare costs.

References

1. Y. Sharma, R. Veliyambara and R. Shettar, "Hybrid Classifier for Identification of Heart Disease," 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), Bengaluru, India, 2019, pp. 1-3, doi: 10.1109/CSITSS47250.2019.9031037.
2. Bhatt, C.M.; Patel, P.; Ghetia, T.; Mazzeo, P.L. Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms 2023, 16, 88. https://doi.org/10.3390/a16020088
3. Nagavelli U, Samanta D, Chakraborty P. Machine Learning Technology-Based Heart Disease Detection Models. J Healthc Eng. 2022 Feb 27;2022:7351061. doi: 10.1155/2022/7351061. PMID: 35265303; PMCID: PMC8898839.
4. Ozcan, M. and Peker, S. (2023) "A classification and regression tree algorithm for heart disease modeling and prediction," Healthcare Analytics, 3, p. 100130. Available at: https://doi.org/10.1016/j.health.2022.100130.
5. Pal, M. and Parija, S. (2021) "Prediction of heart diseases using Random Forest," Journal of Physics: Conference Series, 1817(1), p. 012009. Available at: https://doi.org/10.1088/1742-6596/1817/1/012009.
6. Sandhya, Yamala. (2020). Prediction of Heart Diseases using Support Vector Machine. International Journal for Research in Applied

Science and Engineering Technology. 8. 126-135. 10.22214/ijraset.2020.2021.

7. P. Ramprakash, R. Sarumathi, R. Mowriya and S. Nithyavishnupriya, "Heart Disease Prediction Using Deep Neural Network," 2020 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2020, pp. 666-670, doi: 10.1109/ICICT48043.2020.9112443.
8. Maini E, Venkateswarlu B, Maini B, Marwaha D. Machine learning-based heart disease prediction system for Indian population: An exploratory study done in South India. Med J Armed Forces India. 2021 Jul;77(3):302-311. doi: 10.1016/j.mjafi.2020.10.013. Epub 2021 Jan 6. PMID: 34305284; PMCID: PMC8282535.
9. Ahmad, G.N.; Shafiullah; Fatima, H.; Abbas, M.; Rahman, O.; Imdadullah; Alqahtani, M.S. Mixed Machine Learning Approach for Efficient Prediction of Human Heart Disease by Identifying the Numerical and Categorical Features. Appl. Sci. 2022, 12, 7449. https://doi.org/10.3390/app12157449
10. Pathan, M.S. et al. (2022) "Analyzing the impact of feature selection on the accuracy of heart disease prediction," Healthcare Analytics, 2, p. 100060. Available at: https://doi.org/10.1016/j.health.2022.100060.
11. Kavitha Chandrashekar, AnithaTuluvanooruNarayanreddy, "An Ensemble Feature Optimization for an Effective Heart Disease Prediction Model," International Journal of Intelligent Engineering and Systems, Vol.16, No.2, 2023, DOI: 10.22266/ijies2023.0430.42.
12. Patro, S.P., Nayak, G.S. and Padhy, N. (2021) "Heart disease prediction by using novel Optimization Algorithm: A supervised learning prospective," Informatics in Medicine Unlocked, 26, p. 100696. Available at: https://doi.org/10.1016/j.imu.2021.100696.
13. Kaushalya Dissanayake, Md Gapar Md Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms", Applied Computational Intelligence and Soft Computing, vol. 2021, Article ID 5581806, 17 pages, 2021. https://doi.org/10.1155/2021/5581806
14. García-Ordás, M.T., Bayón-Gutiérrez, M., Benavides, C. et al. Heart disease risk prediction using deep learning techniques with feature augmentation. Multimed Tools Appl (2023). https://doi.org/10.1007/s11042-023-14817-z
15. Guleria, P.; Naga Srinivasu, P.; Ahmed, S.; Almusallam, N.; Alarfaj, F.K. XAI Framework for Cardiovascular Disease Prediction Using Classification Techniques. Electronics 2022, 11, 4086. https://doi.org/10.3390/electronics11244086
16. Yu, Z., Wang, K., Wan, Z. et al. Popular deep learning algorithms for disease prediction: a review. Cluster Comput 26, 1231–1251 (2023). https://doi.org/10.1007/s10586-022-03707-y.
17. Panda, D. et al. (2019) "Predictive systems: Role of feature selection in prediction of heart disease," Journal of Physics: Conference Series, 1372(1), p. 012074. Available at: https://doi.org/10.1088/1742-6596/1372/1/012074.
18. Spencer R, Thabtah F, Abdelhamid N, Thompson M. Exploring feature selection and classification methods for predicting heart disease. Digit Health. 2020 Mar 29;6:2055207620914777. doi: 10.1177/2055207620914777. PMID: 32284873; PMCID: PMC7133070.
19. Aleem, A., Prateek, G., Kumar, N. (2022). Improving Heart Disease Prediction Using Feature Selection Through Genetic Algorithm. In: Woungang, I., Dhurandher, S.K., Pattanaik, K.K., Verma, A., Verma, P. (eds) Advanced Network Technologies and Intelligent Computing. ANTIC 2021. Communications in Computer and Information Science, vol 1534. Springer, Cham. https://doi.org/10.1007/978-3-030-96040-7_57
20. Wankhede, J., Kumar, M. and Sambandam, P. (2020) "Efficient heart disease prediction-based on optimal feature selection using DFCSS and classification by improved Elman-SFO," IET Systems Biology, 14(6), pp.

380–390. Available at: https://doi.org/10.1049/iet-syb.2020.0041.

21. Yadav, Dhirendra & Saini, Prabhav. (2022). Feature Optimization Based Heart Disease Prediction using Machine Learning. 10.1109/ISCON52037.2021.9702410.
22. Abdollahi, J., Nouri-Moghaddam, B. A hybrid method for heart disease diagnosis utilizing feature selection based ensemble classifier model generation. Iran J Comput Sci 5, 229–246 (2022). https://doi.org/10.1007/s42044-022-00104-x
23. https://archive.ics.uci.edu/ml/datasets/heart+disease
24. https://archive.ics.uci.edu/ml/datasets/statlog+(heart)
25. A. Abdellatif, H. Abdellatef, J. Kanesan, C. -O. Chow, J. H. Chuah and H. M. Gheni, "An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods," in IEEE Access, vol. 10, pp. 79974-79985, 2022, doi: 10.1109/ACCESS.2022.3191669.
26. Kavita C, T.N ANITHA, "INVESTIGATION OF HYBRID SVM –EBO METHOD BASED HEART DISEASE PREDICTION," Jilin DaxueXuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition), Vol: 41 Issue: 09-2022, DOI 10.17605/OSF.IO/39QPN.

BIOGRAPHIES OF AUTHORS

Kavitha Chandrashekar B.E., MTech, Associate Professor in the Department of Computer Science, Vijaya Vittala Institute of Technology, Bangalore is pursuing Ph.D. in Computer Science from V.T.U. Belgaum, has 21 years of experience in teaching. Her Research interests are in Machine Learning, Software Engineering, Finite Automata, DBMS, Data mining etc., her projects are selected for KSCST twice and I'd awarded as Best Teacher several times. She is awarded with prestigious "Best Women Faculty for the year 2021" from Novell Research Academy. She has published many National and International journals. She is guiding several UG and PG students. She can be contacted at email: ckavitac22@gmail.com.

AnithaTuluvanooruNarayanreddy Professor, Dept of Information Science and Engineering, Atria IT, Bangalore. She has Completed her Ph.D. in CSE from VTU. She had 23 year's experience in teaching. Her Research Interest are in Cloud Computing, Cyber-Security, Block Chain, IOT, AT, Data science and parallel & Distributed Systems. She has published 30 research articles in varies National and International Journals, Conferences. She is guiding several UG, PG and Ph.D. students at different universities. She can be contacted at email:anithareddytn72@gmail.com.