Multimedia Tools and Applications (2024) 83:68341–68361
https://doi.org/10.1007/s11042-023-18035-5

Applying data science approach to predicting diseases and recommending drugs in healthcare using machine learning models – A cardio disease case study

Muhib Anwar Lambay[1] · **S. Pakkir Mohideen**[2]

Received: 30 November 2022 / Revised: 26 October 2023 / Accepted: 26 December 2023 /
Published online: 25 January 2024

Abstract
Cardiovascular diseases are causing more deaths across the globe. With innovations in Artificial Intelligence (AI) predicting such diseases early is very important research area. With learning based approaches that exploit knowledge from given samples, it is possible to improve disease prediction process. There are many aspects to proper healthcare such as preventing diseases with suitable diet and lifestyle, early detection of diseases if any and efficient treatment. Data is being accumulated in every domain. However, the healthcare industry is on top of the list as it provides large volumes of data pertaining to human health, diet and drug aspects. The existing literature has not shown adequate research in this direction. The Healthcare industry has an unprecedented impact on the well-being of people across the globe. In the recent observations by World Health Organization (WHO), data science approach towards disease prediction greatly complements existing Clinical Decision Support Systems (CDSSs).This research paper presents a comprehensive study on the application of data science techniques for disease prediction and drug recommendation in healthcare, focusing on a case study involving cardiovascular diseases. The primary objective of this study is to develop a robust predictive model that identifies the likelihood of cardiovascular diseases in patients, and subsequently recommends drug interventions for optimal treatment outcomes. Here we propose Disease Prediction and Drug Recommendation Framework (DPDRF). The framework is realized by defining an algorithm known as Cardio Disease Prediction and Drug Recommendation (CDP-DR). The Disease Prediction and Drug Recommendation algorithm in turn uses different supervised machine learning (ML) algorithms such as Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Stochastic Gradient Descent (SGD), Gradient Boosting, and Extreme Gradient Boosting (XGB). Another algorithm known as Entropy and Gain based Hybrid Feature Selection (EG-HFS) is defined to leverage quality of training leading to performance enhancement of prediction models. The experimental results with cardio disease prediction as a case study revealed that the proposed framework is useful in disease prediction and drug recommendations by using different prediction models. Highest accuracy achieved by the proposed system is 96.23%.

Keywords Big data · Big data analytics · Cardio disease prediction · Feature selection · Supervised machine learning · Drug recommendation

Extended author information available on the last page of the article

1 Introduction

Data science, of late, has gotten tremendous attention in academia and research circles. With cloud computing and big data, it became a reality to have data science approach. The application of this approach in the form of big data analytics in the healthcare domain has the potential to make a huge impact on the stakeholders of the industry. Besides, it results in reducing the cost of healthcare service, enhancing Quality of Service (QoS) and reducing error and waste as well [5]. Many contributions were found in this regard. Iqbal et al. [1] presented a data science approach that is used to analyze cyber-physical systems in order to have better security with the usage of computational intelligence and data analytics. They used fuzzy logic in their method to enhance intelligence. Mehta and Pundit [5] explored the concurrence between healthcare requirements and big data analytics. They insist on having a shift in the culture of healthcare units by using technology-driven approaches. Ngiam and Khor [7] explored different algorithms associated with machine learning (ML) for big data in the healthcare industry. They found that algorithms can help in disease diagnosis and different interventions related to healthcare units.

Sahoo et al. [9] studied an "intelligence-based health recommendation system using big data analytics". They found that every recommender system has its phases such as data collection, learning, recommendation and feedback. They reported different kinds of recommender systems that are based on content-based, model-based and hybrid filtering techniques. Dlamini et al. [20] used disease case and oncology case studies to know the benefits of the usage of big data. From the literature, it is understood that there has been the usage of big data analytics in the healthcare domain. However, the literature found prediction systems and recommendations separately. It is more useful if the system is capable of predicting disease and also providing recommendations. Moreover, proposing a novel feature selection also helps in improvement. This is the basis for the work in this paper. Our contributions in this paper are as follows.

We proposed Disease Prediction and Drug Recommendation Framework (DPDRF). This framework is based on supervised machine learning. The framework has two phases namely training and test. In the training phase, which can be done offline, number of ML models are used to learn from labelled data. Such learning process provides required knowledge for automatic disease prediction and drug recommendation. The models used in the training phase are persisted to reuse them later. This process also avoids repetition of training models. The saved models can be used without reinventing the wheel again thus leading to faster convergence in disease detection. In the testing phase, which is considered online, unlabelled data is used to perform desired classification.

We proposed an algorithm known as Cardio Disease Prediction and Drug Recommendation (CDP-DR) to realize the framework. This algorithm uses dataset and also ML pipeline. After pre-processing the given data, the data is splint into training and test sets. Then the training data is subjected to feature selection which plays crucial role in identifying features with higher importance. This process is important as it improves quality of training. The feature selection is done by invoking EG-HFS algorithm proposed in this paper. In other words, CDP-DR makes use of EG-HFS for selecting contributing features. Afterwards, this is an iterative process to train each ML model in the pipeline using training data and perform classification using test data.

An algorithm known as Entropy and Gain-based Hybrid Feature Selection (EG-HFS) is defined to leverage quality of training leading to performance enhancement of prediction models. This algorithm is a filter based approach that considers correlation of features with the target class. It is a hybrid approach where entropy and gain based measures are exploited to reap benefits in identifying features.

The remaining sections of the paper are structured as follows. Section 2 puts a light on big data analytics and its usage in the healthcare domain by reviewing related works. Section 3 presents Disease Prediction and Drug Recommendation Framework (DPDRF). Section 4 presents results of CDP-DR compared with the state of the art. Section 5 concludes the utility of CDP-DR while giving a future scope of the work.

2 Related work

This section reviews the literature on data analytics and predictions in solving real-world applications and particularly in the healthcare domain.

2.1 Data science approaches

Iqbal et al. [1] presented a data science approach that is used to analyze cyber-physical systems in order to have better security with the usage of computational intelligence and data analytics. They used fuzzy logic in their method to enhance intelligence. They intend to apply it to different application areas in future. Iqbal et al. [2] did a similar kind of research as in [1] but extended it to different domains. Galesti et al. [11] explored data science in terms of resources, data types, different techniques for analysis and the potential benefits of using data science in healthcare service units. Pramanik et al. [12] investigated on privacy issues to be considered while using data science analytics. Bag et al. [13] focused on the role of different factors in the manufacturing industry to adopt data science. The factors are associated with resources and economy and institutional pressures. Banerjee et al. [14] explored trends in Internet of Things (IoT) and associated data science for healthcare and biomedical technologies. They found that with data science, it is possible to add more value to healthcare organizations. Mikalef et al. [15] used a hybrid method to ascertain the relationship between data science based analytics and the performance of a given organization. Su et al. [16] used data science approach to leverage the representation of carbon emissions. Patel et al. [17] found that data science has the potential to improve the system of sports. Ma et al. [18] employed data science approach to tourism for obtaining interesting facts. Zhang et al. [19] focused on cognitive data science to ascertain negativity in public emergences. Dlamini et al. [20] used heart disease and oncology case studies to know the benefits of the usage of data science.

2.2 Big data analytics approach

Anisetti et al. [3] proposed big data based methodology as a service for policies associated with public health in urban areas. It could improve the public health policy-making process. Palani Samy and Thirunavukarasu [4] explored in the implications of having frameworks pertaining to the healthcare domain. Different stakeholders benefited from such frameworks include patients, medical practitioners, hospital operators, pharma and clinical researchers and healthcare insurance providers. Mehta and Pundit [5] explored the concurrence between healthcare requirements and big data analytics. They insist on having a shift in the culture of healthcare units by using technology-driven approaches. Wang et al. [6] investigated on the potential benefits of big data analytics and its capabilities pertaining to the healthcare domain. They found that big data analytics is capable of leveraging business intelligence (BI)and use the modern computational infrastructure. Their architecture has different layers like data layer, data aggregation layer, analytics layer and knowledge

exploration layer. These layers are on the top of the data governance layer. Ngiam and Khor [7] explored different algorithms associated with machine learning (ML) for big data in the healthcare industry. They found that algorithms can help in disease diagnosis and different interventions related to healthcare units. Galesti et al. [8] studied big data approach to healthcare and came to know its value to organizations and challenges for society and organizations as well. Sahoo et al. [9] studied on "intelligence-based health recommendation system using big data analytics". They found that every recommender system has its phases such as data collection, learning, recommendation and feedback. They reported different kinds of recommender systems that are based on content-based, model-based and hybrid filtering techniques. Email et al. [10] focused on smart big data analytics that involves clustering in traditional ML and also clustering in distributed architectures.

2.3 Recent methods

Akkem et al. [25] explored AI based methods and their significance in solving problems of the applications in different domains. Rahim et al. [26] proposed an integrated approach for cardiovascular prediction. Their methodology was based on ML techniques. It has provision for learning from historical data and perform classification of diseases in the newly given test data. however, their methodology lacks feature engineering approach. Bertsimas et al. [27] also studied the utility of ML models in heart disease prediction. Their approach includes the usage of best performing ML models that are used for detection of heart diseases. In [28], the research not only includes heart disease prediction but also has provision for severity identification. Ali et al. [29] used many ML models for cardiovascular disease diagnosis while Das et al. [30] used ML models to investigate on the diagnosis potential of the same. From the literature, it is understood that there has been the usage of big data analytics in the healthcare domain. However, the literature found prediction systems and recommendations separately. It is more useful if the system is capable of predicting disease and also providing recommendations. Moreover, proposing a novel feature selection also helps in improvement. This is the basis for the work in this paper.

3 Proposed framework

This section presents the details of the proposed framework, underlying algorithms and evaluation methodology. A data science approach for disease prediction and drug recommendations under healthcare system is followed with a cardio disease case study. The framework presents the methodology involved in the supervised learning based approach in the process of identification of presence of disease. Based on the disease identification, drug recommendations are generated. Since healthcare industry is crucial for the human wellbeing, the healthcare domain is chosen, particularly, cardio disease detection and drug recommendations.

3.1 Disease prediction and drug recommendation framework

Machine learning models have been around for prediction and classification to solve different kinds of problems. Machine learning algorithms have required modus operandi to learn from training samples and render desired business intelligence. Supervised learning has two phases such as training phase and testing phase. In the training phase, the machine learning model

is trained from the training samples. In the testing phase, the unlabelled samples are used as training set and the algorithm predicts class labels. In case of unsupervised learning, there is no training given to algorithm explicitly. On the other hand, semi-supervised methods will have both supervised and unsupervised learning possibilities. The proposed framework in this paper is based on supervised learning. However, the problem with the supervised learning approach is that it depends on quality of training data for better performance. Therefore, if the training data has no good quality, then the performance of supervised learning models gets deteriorated. In order to overcome this problem, in this paper, we proposed a feature selection algorithm that could identify features that can contribute to class label prediction.

The framework is realized by defining an algorithm known as Cardio Disease Prediction and Drug Recommendation. This algorithm in turn uses different supervised machine learning (ML) algorithms. They are known as Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), SGD, Gradient Boosting (GB) and XGB. DT generates understandable rules in the form of tree which is meant for disease classification. RF is an ensemble model which makes use of multiple DTs and voting approach to determine class label. LR model is based on the probability of given instance belonging to a class. SGD model has needed optimization of internal models for better performance. GB has ensemble approach to determine class labels. XGB makes use of gradient boosted decision trees for leveraging classification performance. Another algorithm known as Entropy and Gain-based Hybrid Feature Selection is defined to leverage quality of training leading to performance enhancement of prediction models. The overview of the framework is shown in Fig. 1.

The framework makes use of training set which has class labels mentioned in one of the attributes. Such data, in other words, has ground truth or diagnosis or class label known beforehand. Every instance (record of a patient) has different attributes and also diagnosis column in the training set. Such training set is used to train a classifier. However, before training the classifier, the hybrid feature selection algorithm is used in order to find features that can contribute to prediction of class labels. Such features are only used for training the classifier. This will improve quality of training as the feature selection gets rid of irrelevant and redundant features. Once training is completed, a knowledge model is created for further processing in testing phase. This knowledge model is used by the classifier to predict class labels when test data is given. In the testing phase, the test data which has no class labels mentioned is given as input. It is subjected to hybrid feature selection to identify useful features. Then features are given to the knowledge model or prediction model resulted in the training phase. The model is capable of predicting class labels. In essence, the given testing samples are labelled or classified into cardio disease and no cardio disease.

The input dataset is dataset [21] is divided into the training set and testing set. The training set is subjected to a hybrid feature selection method known as Entropy and Gain based Hybrid Feature Selection (EG-HFS) which takes all features extracted as input and result in features that are highly relevant. Different classifiers are used to have prediction models. In the training phase, the extracted features are learned by the proposed algorithm known as Cardio Disease Prediction and Drug Recommendation (CDP-DR) that is capable of providing predictions and recommendations. In the testing phase, the extracted features are learned by the resultant model and the labelling is made. In EG-HFS different metrics are used as expressed in Eqs. 1 to 4.

$$\mathrm{SU} = \frac{2 * Gain}{H(x) + H(y)} \tag{1}$$

As in Eq. 1, there is composite metric derived by using entropy and gain metrics. The computation of H(X) is expressed in Eq. 2.

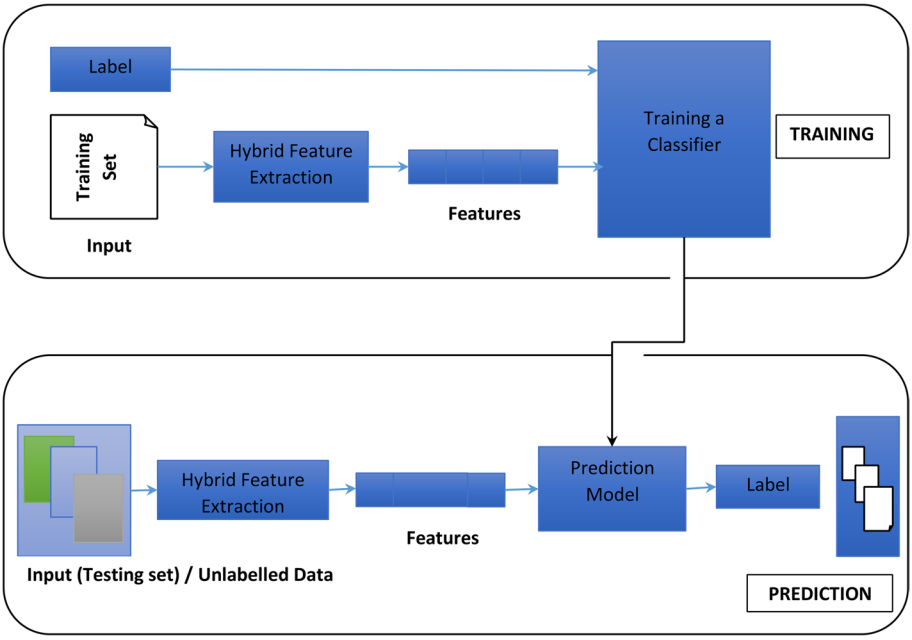

Fig. 1 Overview of the proposed data analytics framework based on data analytics

$$H(X) = -\sum_{x \in X} p(x) log\ p(x) \tag{2}$$

In the same fashion, the computation of H(Y) is done as expressed in Eq. 3.

$$H(Y) = -\sum_{y \in Y} p(y) log\ p(y) \tag{3}$$

The gain metric is computed as in Eq. 4.

$$Gain = H(y) - H(y/x) \tag{4}$$

The entropy and gain measures when combined will have symmetric uncertainty which is the combined metric that is used to determine whether a feature can contribute to the class label prediction or not. As the two measures are combined, it results in a hybrid metric or hybrid approach in class label prediction. Hence, the proposed feature selection method is named as Hybrid Feature Selection (HFS).

The methodology is based on supervised learning phenomenon which has two distinct phases. In the training phase, the proposed system has mechanisms to pre-process the given data and split into training set (T1) and testing set (T2). The system makes use of number of ML models for disease diagnosis. These ML models are trained with the training data. however, prior to training, the training data is subjected the proposed feature selection method. The feature selection method makes use of a composite filter method which computes importance of each feature. Based on the feature importance, only contributing features are chosen. In fact, training is given to ML models based on the selected features only. It has potential to improve quality of training.

The proposed methodology also has provision for feature selection. Section 3.3 provides more details on feature selection. The feature selection method exploits entropy and gain measures that are based on filter method. They correlate features with the target class label in order to compute feature importance. As every feature has different feature importance, only the contributing features are selected. This feature selection method is reused by the disease diagnosis algorithm.

3.2 Cardio disease prediction and drug recommendation

An algorithm known as Cardio Disease Prediction and Drug Recommendation (CDP-DR) is proposed. It is evaluated suing dataset collected from [21]. It makes use of the entropy and gain metrics in order to have better determination of the features that are useful in predicting class labels.

Algorithm 1 Cardio disease prediction and drug recommendation algorithm

Inputs: Healthcare dataset $\boldsymbol{D}$
Pipeline of ML models **M**
Outputs:
Disease prediction results P
Generated recommendations R

```
1. Begin
2. (T1, T2) ← Pre Process(D)
3. F ← EG-HFS(T1)
4. For each model in pipeline of models M
5.    m ←Train Model(F)
6. End For
7. For each tuple t in test data T2
8.    result ←    Predict(m, t)
9.    Add result to R
10.       End For
11.       End For
12.       P ← Evaluate Models(M)
13.       Display R
14.       Display P
15.       End
```

Algorithm 1 takes Healthcare dataset $\boldsymbol{D}$ and Pipeline of ML models $\mathbf{M}$ as inputs and generates disease detection results and recommendations. In the process, it makes use of EG-HFS for selecting best features prior to training different classifiers in the pipeline. There is an iterative process to train all classifiers and another interactive process to perform detection process against all test instances. The rationale behind this algorithm is to improve quality of training in the prediction of Cardio disease. Unless feature selection algorithm is used, even the best classification algorithms yield mediocre results. In order to get rid of this kind of problem, the proposed EG-HFS is used as part of the main algorithm.

3.3 Entropy and gain based hybrid feature selection

This algorithm is defined to improve the performance of prediction models. The prediction models used in the experiments are known as Random Forest (RF) [22], Logistic Regression (LR) [23], Decision Tree (DT) [25], Stochastic Gradient Boosting (SGB), Gradient Boosting and Extreme Gradient Boosting (XGB). The algorithm takes dataset [21] as input and produces selected features. The proposed algorithm acts as pre-processing step to classification or disease prediction. It is based on the gain and entropy measures that are combined into a hybrid metric in order to have better possibilities in determining useful features. Provided a dataset containing details of patients, the algorithm finds useful features and such features are used in the training phase of the proposed framework. The Cardio Disease prediction models aforementioned are expected to have better performance in disease prediction.

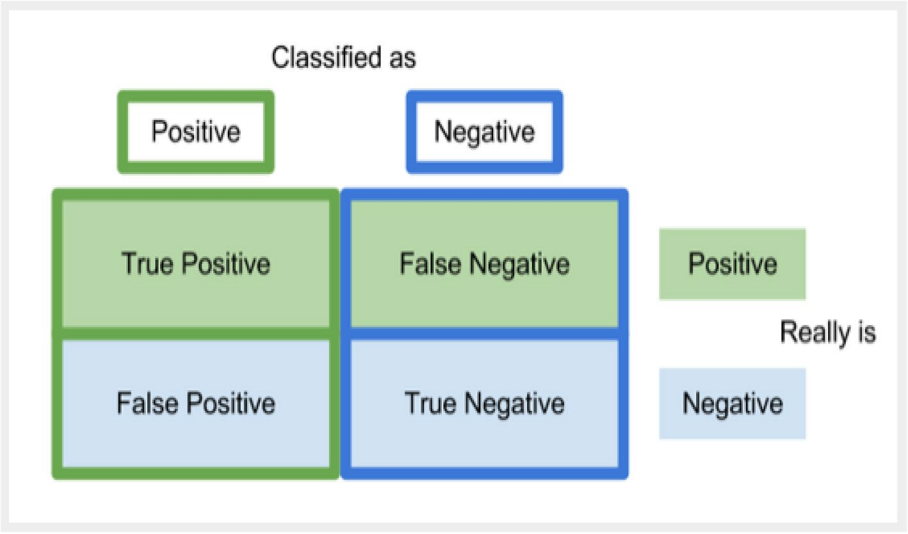

Fig. 2 Confusion matrix

Algorithm 2 Entropy and gain based Hybrid feature selection

Inputs:
Healthcare dataset ***D***
Feature importance threshold **th**
Outputs: Contributing features F

1. Initialize vector S to hold composite metric value
2. F←Get All Features(D)

   Importance Computation
3. For each feature of dataset in F
4. en ← Calculate Entropy Measure (F, f)
5. g ← Calculate Gain Measure (F, f)
6. su ← Compute Composite Measure SU(F, f)
7. Update S with su
8. End For

   Feature Importance Verification
9. F1←F
10. F ←null // empty F
11. For each composite metric value s in S
12. If s value is greater than th Then
13. f ← Obtain Feature(F1)
14. Add f to F
15. End If
16. End For
17. Return final feature set F

Algorithm 2 takes Healthcare dataset ***D*** and Feature importance threshold **th** as inputs and determine contributing features. In the process, there is computation of different measures like entropy and gain besides the composite metric SU. Finally based on the feature important satisfying threshold, only satisfied features are identified and they are used for further processing in disease prediction.

3.4 Dataset

FAERS (FDA Adverse Event Report System) is collected from [21] and used for experimental results. The dataset contains data pertaining to adverse events and medication errors that are notified to FDA. The dataset contains information that can help in safety

surveillance programs, recommendation systems and products that are biological in nature with drug and therapeutic features. The dataset adheres to the guidelines of International Conference on Harmonisation (ICH E2B).

3.5 Evaluation metrics

The confusion matrix as in Fig. 2 is used to derive metrics presented in this section. The confusion matrix provides information that is useful in getting different metrics. This is done by comparing the ground truth values and algorithm predicted or classified results in terms of positive (presence of cardio disease) or negative (absence of cardio disease).

Four cases such as correct predictions (True Positive and True Negative) and incorrect predictions (False Positive and False Negative) are used for deriving metrics. In essence there are four possibilities for an algorithm to predict. They are in terms of positive negative predictions. True positive does mean that there is cardio disease really and the algorithm also predicts it positively. True negative on the other hand does mean that the patient has no cardio disease and the algorithm also finds negative about it. False positive does mean that there is no cardio disease really but algorithm predicts as cardio disease. This kind of case in prediction is known as false positive. Similarly false negative does mean that there is really cardio disease but the algorithm predicts it as no cardio disease case.

$$\text{Precision (p)} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall (r)} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{F1-score} = 2 * \frac{(p * r)}{(p + r)} \tag{7}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Based on these metrics expressed from Eqs. 5 to 8, experimental results of the proposed algorithms are evaluated with certain baseline algorithms. These measures are widely used in evaluation of machine learning algorithms in the field of healthcare and other domains. Each measure has a value ranging from 0 to 1 reflecting 0% and 100% respectively. Each measure reflects higher performance if the value is closer to 1 and low performance if the value is closer to 0.

4 Experimental results

This section presents results of experiments. The results are presented in terms of exploratory results and performance comparison results. The exploratory results mainly reflect the dynamics in the given datasets while performance evaluation results show the performance in prediction of disease exhibited by different models. Dataset is taken from UCI repository. It is heart disease dataset which has 14 attributes including target of class label. It has details of patients consisting of symptoms useful for cardiovascular disease diagnosis. In

the process of using the dataset pre-processing is carried out which includes finding missing values and handling them considering mean values in case of numeric attributes.

As presented in Fig. 3, the source implemented has different components. EDA component is meant for data exploration while Pre-Process is meant for detecting and treating missing values in the dataset. Feature Selection component is meant for realising the proposed feature selection method while Model Creation takes care of creating ML models. Model Training and Testing play important role such as learning from training data and classifying using test data. Performance Comparison component is used to use metrics to know performance of each ML model.

4.1 Exploratory results

Different explorations on the data are made prior to disease prediction and drug recommendations. The exploratory results include histograms on weight, age and gender, results of different clusters, word cloud representing diseases in the dataset and drug versus scope information.

Since weight of a person has its influence in the probabilities of diseases, weight attribute and its underlying data in the dataset is subjected to exploratory data analysis. As presented in Fig. 4, the weight of person has its influence in the frequency of disease. It does mean that out of all people, those who have more weight have more disease occurrence. The weight and disease frequency analysis presented reflects the disease trends with respect to weight dynamics of people.

Age is an important factor in human life and wellbeing. As the age increases, there is probability of losing immunity and acquiring certain diseases. Figure 3 shows the relation between age of people and frequency of diseases. As per the data available, the age is one of the factors that indirectly influences cause of diseases. Thus it is established that people with age more than 30 years are found to have more disease frequency. Age and disease frequency analysis provides some insights pertaining to occurrence of diseases.

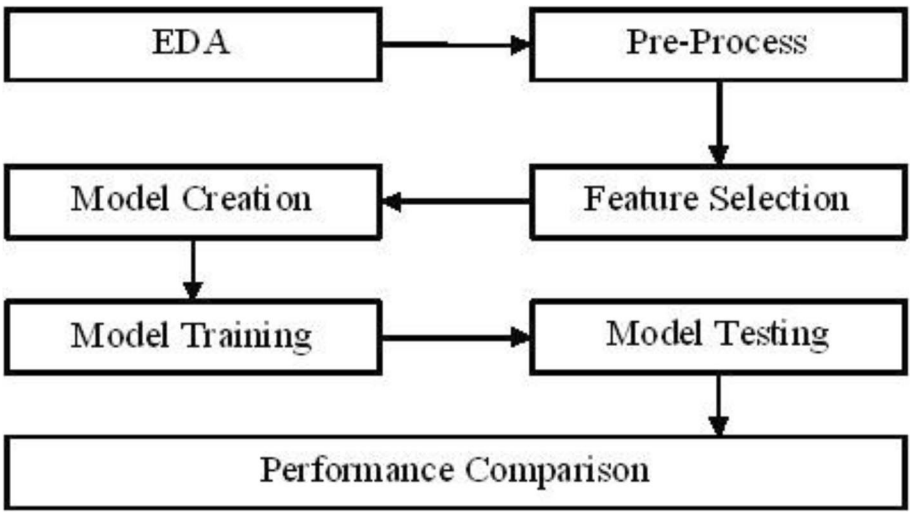

Fig. 3 An outline of components in source code implementation

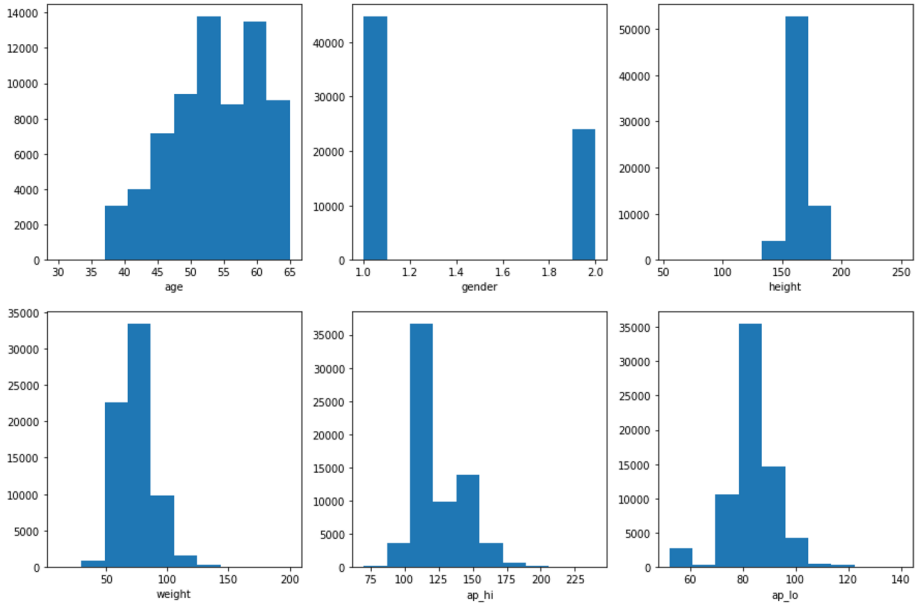

Fig. 4 Histogram of weight and frequency of disease in patients

In the exploratory data analysis with respect to gender, there is an important insight that is female are more vulnerable to cardio disease when compared to male population. Figure 3 shows the frequency of cardio disease for male (1.0) and female (2.0). It also reveals that cardio disease can occur to male population also but it is very less frequent.

As shown in Fig. 5, the different fields in the dataset are shown in horizontal axis and vertical axis to have correlations visualized. The correlation value is presented against each pair of variables.

As shown in Fig. 6, it describes the difference between the number of individuals with heart disease, indicated by 1, and those without heart disease, indicated by 0. The figure shows that the data consists of more non-cardio disease patients as compared with cardio patients.

As shown in Fig. 7 the severity of cardio attack with respect to smoke. The patients who don't smoke will have a less chance of cardio attack other than that it may effects on heart which results in heart disease.

As shown in Fig. 8, it is useful to visualize data with side by side views. Different sub plots are generated to reflect the data dynamics associated with each attribute in the dataset.

Here the Fig. 9 shows the visualizing ROC curve of various proposed models.

4.2 Results of drug recommendations

With regard to cardio disease case study, drug recommendation is made in a subjective fashion. It does mean that, the patient once diagnosed with cardio disease, the drug recommendations given are specific to the patient based on age, weight and gender dynamics.

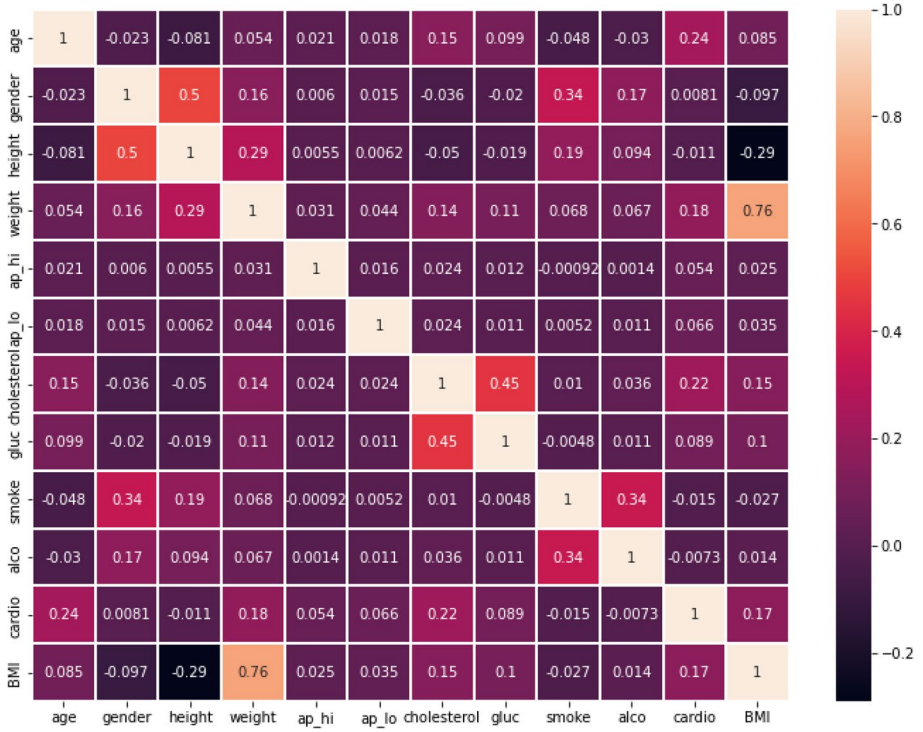

Fig. 5 Correlation matrix

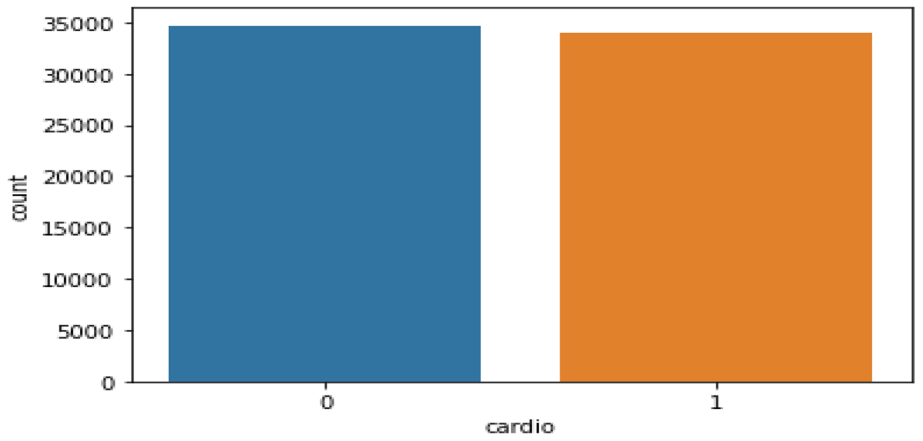

Fig. 6 Visualizing the cardio records in terms of 0 and 1

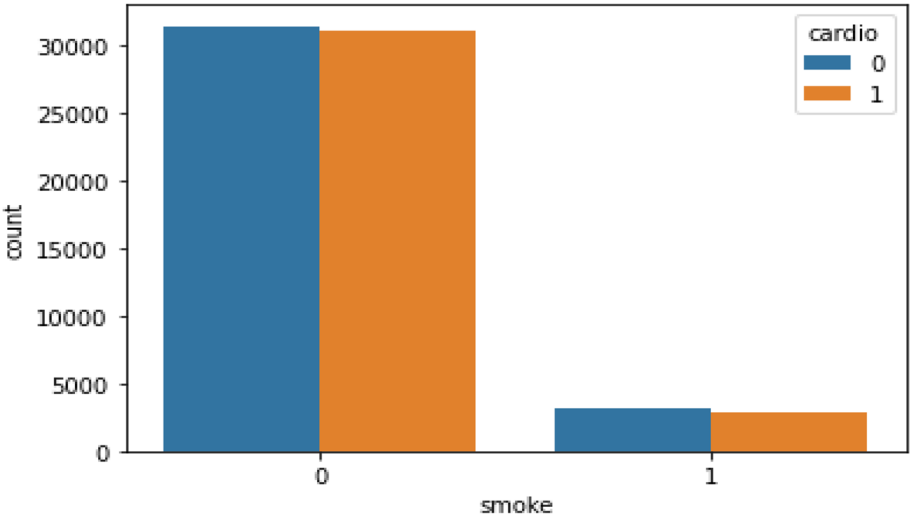

Fig. 7 Visualizing the cardio patients with respect to smoke

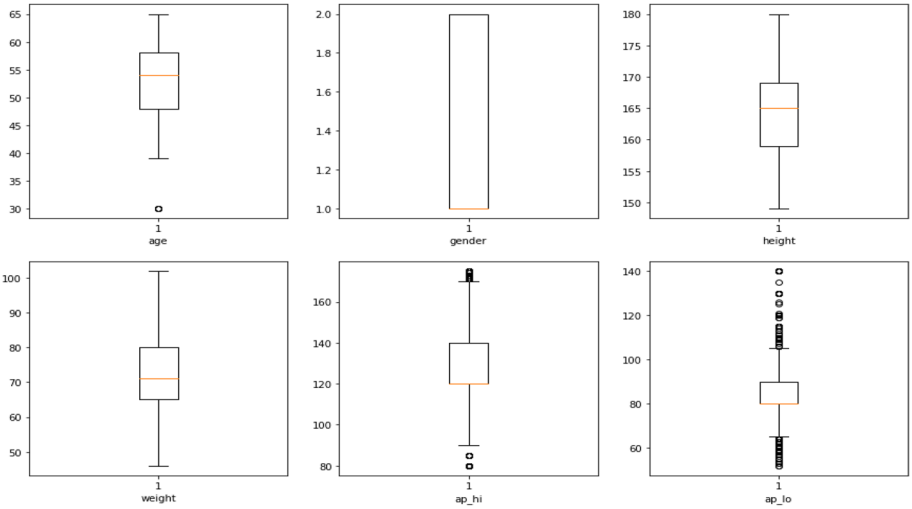

Fig. 8 Visualizing subplots for comparing with each features outer layers

4.2.1 Drug recommendation for cardio disease patient: (Gender: Female; Weight: 78; Age: 58)

Recommended drug for the patient is Acebutolol. Table 1 shows different variants that are the results of recommendations.

As presented in Table 1, the drug recommendation and the score of each drug recommended for given patient with given gender, weight and age. For a patient with Gender: Female; Weight: 78; Age: 58, the table shows the drug recommendations.

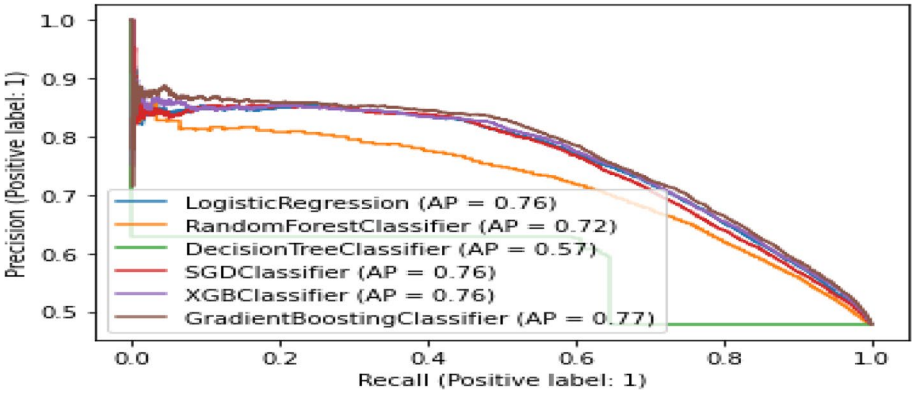

Fig. 9 Visualizing ROC curve of proposed models

Table 1 Drug recommendations for a female patient with weight 78 and age 58

Drug	F1 Score
Acebutolol	0.089831
Metoprolol	0.086331
Propranolol	0.084761
Statins	0.076091
Betaxolol	0.059437

Table 2 Drug recommendations for a female patient with weight 80 and age 75

Drug	F1 Score
Anticoagulants	0.085519
Antiplatelet Agents	0.083589
Calcium channel blockers (CCBs)	0.081289
Statins	0.077519
Nitrates	0.060820

4.2.2 Drug recommendation for cardio disease patient: (Gender: Female; Weight: 80; Age: 75)

Recommended drug for the patient is Statins. Table 2 shows different variants that are the results of recommendations.

4.2.3 Drug recommendation for cardio disease patient: (Gender: Male; Weight: 70; Age: 45)

Recommended drug for the patient is Betaxolol. Table 3 shows different variants that are the results of recommendations.

Table 3 Drug recommendations for a male patient with weight 70 and age 45

Drug	F1 Score
Betaxolol	0.087500
Statins	0.085500
Propranolol	0.084000
Metoprolol	0.083282
Acebutolol	0.075000

As presented in Table 3, the drug recommendation and the score of each drug recommended for given patient with given gender, weight and age. For a male patient with age 45 and weight 70, the table shows the drug recommendations.

4.2.4 Drug recommendation for cardio disease patient: (Gender: Female; Weight: 60; Age: 80)

Recommended drug for the patient is Antiplatelet Agents. Table 4 shows different variants that are the results of recommendations.

As presented in Table 4, the drug recommendation and the score of each drug recommended for given patient with given gender, weight and age. For a female patient with age 80 and weight 60, the table shows the drug recommendations.

4.2.5 Drug recommendation for cardio disease patient: (Gender: Female; Weight: 57; Age: 20)

Recommended drug for the patient is Stanins. Table 5 shows different variants that are the results of recommendations.

Table 4 Drug recommendations for a female patient with weight 60 and age 80

Drug	F1 Score
Antiplatelet Agents	0.103514
Calcium channel blockers (CCBs)	0.093450
Propranolol	0.093317
Metoprolol	0.078505
Acebutolol	0.070150

Table 5 Drug recommendations for a female patient with weight 57 and age 20

Drug	F1 Score
Propranolol	2.06
Metoprolol	2.06
Anticoagulants	2.05
Antiplatelet Agents	2.04
Stanins	1.87

Table 6 Performance measures of various prediction models

Prediction model	Performance (%)					
	Accuracy	Precision	Recall	F1-score	Sensitivity	Specificity
Random forest classification	0.962359	0.97	0.96	0.96	0.969905	0.89243
Decision tree	0.957314	0.96	0.96	0.96	0.963456	0.900398
Logistic regression	0.668606	0.89	0.67	0.74	0.657351	0.772908
SGD	0.642607	0.88	0.64	0.72	0.629836	0.760956
Gradient boosting	0.593713	0.89	0.59	0.67	0.570077	0.812749
XGB	0.552345	0.80	0.52	0.62	0.526758	0.778789

As presented in Table 5, the drug recommendation and the score of each drug recommended for given patient with given gender, weight and age. For a female patient with age 20 and weight 57, the table shows the drug recommendations. The drug recommendation results are provided based on the disease diagnosis, gender, weight and age of the patient.

4.3 Disease prediction and evaluation

Disease prediction performance of different prediction models such as Random Forest (RF) [22], Logistic Regression (LR) [23], Stochastic Gradient Descent (SGD) [24] and Decision Tree (DT) [25], Gradient Boosting and Extreme Gradient Boosting (XGB) is evaluated using different measures.

As presented in Table 6, the performance of prediction models with the proposed hybrid feature selection is evaluated.

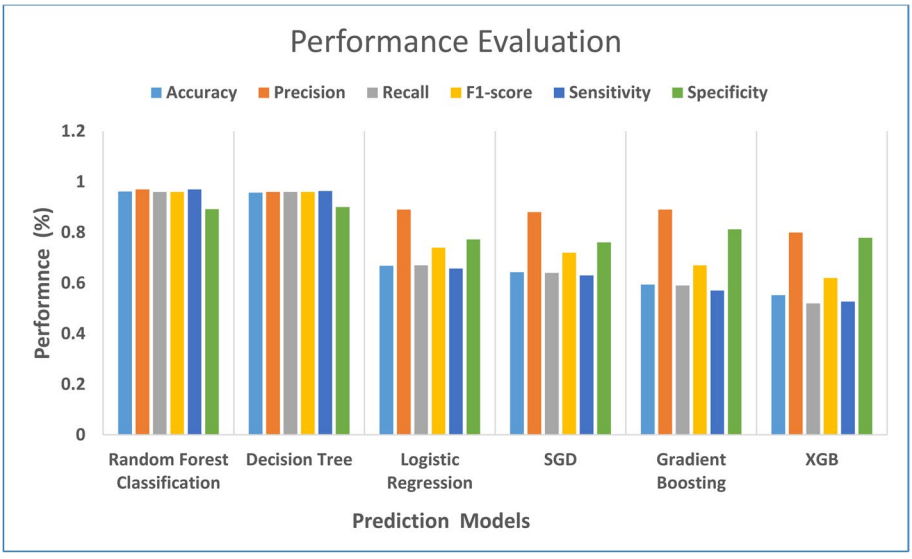

Fig. 10 Performance evaluation of different prediction models with the proposed hybrid feature selection method

As presented in Fig. 10, different prediction models are presented on the horizontal axis and the performance with different metrics is found on the vertical axis. Different prediction models showed varied performance. When accuracy is considered the highest performance is shown by RF with EG-HFS with 0.962359 while the least performance is shown by XGB with 0.5523. DT showed 0.957314, LR 0.668606, XGB with linear kernel 0.642607 while XGB showed 0.5523. From the results, it is understood that RF with the proposed EG-HFS showed significantly improved performance with the proposed CDP-DR method.

The results section throws light on details that are twofold. First it provides exploration of data in terms of discovering data distributions and feature correlations. This part of the results provides useful information about the data used for the empirical study. It also provides data visualization that enables reader to ascertain facts about how data in the dataset is distributed and correlated among different attributes. After exploratory data analysis, the actual results in terms of disease diagnosis and drug recommendation are provided. Disease diagnosis is based on the ML models used in the empirical study. Drug recommendations are based on the gender, age, weight and the kind of the illness of given patient.

5 Conclusion and future work

In this paper we propose a Disease Prediction and Drug Recommendation Framework (DP-DRF). The framework is realized by defining an algorithm known as Cardio Disease Prediction and Drug Recommendation (CDP-DR). Another algorithm known as Entropy and Gain-based Hybrid Feature Selection (EG-HFS) is defined to leverage quality of training leading to performance enhancement of prediction models. The feature selection algorithm improves performance with a hybrid measure known as symmetric uncertainty that is made up of entropy and gain measures. The experimental results with Cardio Disease prediction as a case study revealed that the proposed framework is useful in disease prediction and drug recommendations and shows better performance over the state of the art. When accuracy is considered the highest performance is shown by RF with EG-HFS with 0.962359 while the least performance is shown by XGB with 0.5523. From the results, it is understood that RF with the proposed EG-HFS showed significantly improved performance with the proposed CDP-DR method. In our future work, we intend to improve our framework further to consider more useful data analytics with healthcare data of different kinds.

Funding Authors declare that there is no funding support for this research work.

Declarations

Conflict of interest Authors declare that there are no conflicts of interest among themselves.

References

1. Iqbal R, Doctor F, More B, Mahmud S, Yousuf U (2017) Big data analytics and computational intelligence for cyber physical systems: recent trends and state of the art applications. Future Gener Comput Syst 1–27
2. Iqbal R, Doctor F, More B, Mahmud S, Yousuf U (2018) Big data analytics: computational intelligence techniques and application areas. Technol Forecast Soc Change 1–11

3. Anisetti M, Ardagna C, Bellandi V, Cremonini M, Frati F, Damiani E (2018) Privacy-aware big data analytics as a service for public health policies in smart cities. Sustain Cities Soc 1–36
4. Palanisamy V, Thirunavukarasu R (2017) Implications of big data analytics in developing healthcare frameworks – a review. J King Saud Univ - Comput Inf Sci. https://doi.org/10.1016/j.jksuci.2017.12.007
5. Mehta N, Pandit A (2018) Concurrence of big data analytics and healthcare: a systematic review. Int J Med Informatics 114:57–65
6. Wang Y, Kung L, Byrd TA (2018) Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Soc Chang 126:3–13
7. Ngiam KY, Khor IW (2019) Big data and machine learning algorithms for health-care delivery. Lancet Oncol 20(5):262–273
8. Galitsis P, Katsaliaki K, Kumar S (2019) Values, challenges and future directions of big data analytics in healthcare: A systematic review. Soc Sci Med 1–9
9. Sahoo AK, Mallik S, Pradhan C, Mishra BSP, Barik RK, Das H (2019) Intelligence-based health recommendation system using big data analytics. Big Data Anal Intell Healthcare Manag 227–246
10. Ismail A, Shehab A, El-Henawy IM (2018) Healthcare analysis in smart big data analytics: reviews, challenges and recommendations. Lect Notes Intell Transp Infrastructure 27–45
11. Galetsi P, Katsaliaki K, Kumar S (2020) Big data analytics in health sector: theoretical framework, techniques and prospects. Int J Inf Manag 50:206–216
12. Pramanik MI, Lau RYK, Hossain MS, Rahoman MM, Debnath SK, Rashed MG, Uddin MZ (2020) Privacy preserving big data analytics: A critical analysis of state-of-the-art. WIREs Data Min Knowl Discov 1–26
13. Bag S, Pretorius JHC, Gupta S, Dwivedi YK (2020) Role of institutional pressures and resources in the adoption of big data analytics powered artificial intelligence, sustainable manufacturing practices and circular economy capabilities. Technol Forecast Soc Chang 120420:1–14
14. Banerjee A, Chakraborty C, Kumar A, Biswas D (2020) Emerging trends in IoT and big data analytics for biomedical and health care technologies. Handbook Data Sci Approaches Biom Eng 121–152
15. Mikalef P, Boura M, Lekakos G, Krogstie J (2019) Big data analytics and firm performance: findings from a mixed-method approach. J Bus Res 98:261–276
16. Su Y, Yu Y, Zhang N (2020) Carbon emissions and environmental management based on big data and streaming data: a bibliometric analysis. Sci Total Environ 138984:1–11
17. Patel D, Shah D, Shah M (2020) The intertwine of brain and body: a quantitative analysis on how big data influences the system of sports. Ann Data Sci 1–16
18. Ma S (David), Kirilenko AP, Stepchenkova S (eds) (2020) Special interest tourism is not so special after all: Big data evidence from the 2017 Great American Solar Eclipse. Tour Manag 77:1-13
19. Zhang W, Wang M, Zhu Y (2019) Does government information release really matter in regulating contagion-evolution of negative emotion during public emergencies? From the perspective of cognitive big data analytics. Int J Inform Manage 1–19
20. Atallah R, Al-Mousa A (2019) Heart disease detection using machine learning majority voting ensemble method. 2019 2nd International Conference on New Trends in Computing Sciences (ICTCS), pp 1–6. https://doi.org/10.1109/ICTCS.2019.8923053
21. Nashif S, Raihan R, Islam R, Imam MH (2016) Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. World J Eng Technol 6:854–873
22. Li JP, Haq AU, Din SU, Khan J, Khan A, Saboor A (2020) Heart disease identification method using machine learning classification in E-Healthcare. IEEE Access 8:1–21. Digital Object Identifier. https://doi.org/10.1109/ACCESS.2020.3001149
23. Hazra A, Mandal SK, Gupta A, Mukherjee A, Mukherjee A (2017) Heart disease diagnosis and prediction using machine learning and data mining techniques: a review. Adv Comput Sci Technol 10:1–24. http://www.ripublication.com/acst17/acstv10n7_13.pdf
24. Yadav SS, Jadhav SM, Nagrale S, Patil N (2020) Application of machine learning for the detection of heart disease. 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), pp 165–172. https://doi.org/10.1109/ICIMIA48430.2020.9074954
25. Akkem Y, Biswas SK, Varanasi A (2023) Smart farming using artificial intelligence: a review. Eng Appl Artif Intell 120:105899. https://doi.org/10.1016/j.engappai.2023.105899. (ISSN 0952–1976)
26. Rahim A, Rasheed Y, Azam F, Anwar MW, Rahim MA, Muzaffar AW (2021) An integrated machine learning framework for effective prediction of cardiovascular diseases. IEEE Access. https://doi.org/10.1109/access.2021.3098688
27. Dimitris Bertsimas, Luca Mingardi and Bartolomeo Stellato (2021) Machine learning for real-time heart disease prediction. IEEE J Biomedical Health Inf. https://doi.org/10.1109/jbhi.2021.3066347

28. A Abdellatif, H Abdellatef, J Kanesan, C-O Chow, JH Chuah, HM Gheni (2022) An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods. IEEE. 10, pp 79974–79985. https://doi.org/10.1109/ACCESS.2022.3191669
29. Ali Z, Naseer N, Nazeer H (2022) Cardiovascular disease detection using multiple machine learning algorithms and their performance analysis. IEEE, pp 1–7. https://doi.org/10.1109/ETECTE55893.2022.10007319
30. Chandra Das R, Chandra Das M, Hossain MA, Rahman MA, Hossen MH, Hasan R (2023) Heart disease detection using ML. IEEE, pp 1–5. https://doi.org/10.1109/CCWC57344.2023.10099294

Muhib Anwar Lambay has 15 + years of experience in teaching field. He received his Bachelor of Engineering (B.E.) in Computer Engineering from University of Mumbai, India in year 2008. Master of Technology (M.Tech) in Computer Science & Engineering from Jawaharlal Nehru Technological University, Hyderabad, India in year 2014 and He is currently pursuing his PhD in Computer Science & Engineering from B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai. He presented various academic as well as research-based papers at several national and international conferences. At present he is working as an Assistant Professor in the department of Computer Engineering at Anjuman-I-Islam's Kalsekar Technical Campus, New Panvel, Affiliated to University of Mumbai. His main research work focuses on Big Data Analytics, Machine Learning and NLP.

S. Pakkir Mohideen received his PhD in the field of Personalized Ontology based Adaptive Learning System from Anna University, Chennai in 2017 and M.E. in Computer Science & Engineering from Anna University, Chennai in 2007 He is currently working as a Professor & Head of the Computer Application Department at B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai. He participated in several high profile conferences. He has many publications and is presently working on many more papers. In addition to his academic career, Dr S. Pakkir Mohideen received a "Certificate of Appreciation" award for having produced excellent academic results. His areas of research include Big Data Analytics, Data Mining, and Information Retrieval System.

Authors and Affiliations

Muhib Anwar Lambay[1] · S. Pakkir Mohideen[2]

✉ Muhib Anwar Lambay
lambaymuhib@gmail.com

S. Pakkir Mohideen
pakirmoitheen@crescent.education

[1] CSE Department, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India

[2] CA Department, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India