Original Article

Parvathaneni Rajendra Kumar*, Suban Ravichandran and Satyala Narayana

Ensemble classification technique for heart disease prediction with meta-heuristic-enabled training system

https://doi.org/10.1515/bams-2020-0033
Received June 19, 2020; accepted October 28, 2020;
published online November 26, 2020

Abstract

Objectives: This research work exclusively aims to develop a novel heart disease prediction framework including three major phases, namely proposed feature extraction, dimensionality reduction, and proposed ensemble-based classification.

Methods: As the novelty, the training of NN is carried out by a new enhanced optimization algorithm referred to as Sea Lion with Canberra Distance (S-CDF) via tuning the optimal weights. The improved S-CDF algorithm is the extended version of the existing "Sea Lion Optimization (SLnO)". Initially, the statistical and higher-order statistical features are extracted including central tendency, degree of dispersion, and qualitative variation, respectively. However, in this scenario, the "curse of dimensionality" seems to be the greatest issue, such that there is a necessity of dimensionality reduction in the extracted features. Hence, the principal component analysis (PCA)-based feature reduction approach is deployed here. Finally, the dimensional concentrated features are fed as the input to the proposed ensemble technique with "Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN)" with optimized Neural Network (NN) as the final classifier.

Results: An elaborative analyses as well as discussion have been provided by concerning the parameters, like evaluation metrics, year of publication, accuracy, implementation tool, and utilized datasets obtained by various techniques.

Conclusions: From the experiment outcomes, it is proved that the accuracy of the proposed work with the proposed feature set is 5, 42.85, and 10% superior to the performance with other feature sets like central tendency + dispersion feature, central tendency qualitative variation, and dispersion qualitative variation, respectively.

Results: Finally, the comparative evaluation shows that the presented work is appropriate for heart disease prediction as it has high accuracy than the traditional works.

Keywords: central tendency; degree of dispersion; ensemble classification; heart disease prediction; optimization.

Introduction

One of the conspicuous illnesses that influence numerous individuals during center or mature age is a coronary illness, and most of the time it in the long run prompts mortal entanglements [1]. Heart illnesses are more common in men than in women [2–4]. As per insights from WHO, it has been assessed that "24% of passing due to non-transmittable sicknesses in India are brought about by heart afflictions." Moreover, one-third of all worldwide demise is because of heart ailments. The Cleveland Heart Disease Database (CHDD) is viewed as the true database for coronary illness examination. Clinical research has called attention to various variables that expanse the danger of Cleveland Heart Disease Database (CHDD) and coronary episode [5–8]. "Sex, age, and family ancestry" are those variables that can't be changed, while factors that are identified with way of lifestyle e.g., "smoking, elevated cholesterol, hypertension and physical idleness" can be changed [9–13]. The latter are hazard factors that can be altered, and in certain cases, it can be disposed of with lifestyle changes and drugs.

The clinical conclusion is considered as a huge yet perplexing assignment that should be done decisively and productively. The underlying issue in the determination creates from the information mining process, where there is a chance for the information to get adulterated. Right now, the

***Corresponding author: Parvathaneni Rajendra Kumar**, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, 608002 Chidambaram, Tamil Nadu, India, E-mail: prajendrakumar@sircrrengg.ac.in
Suban Ravichandran, Department of Information Technology, Faculty of Engineering and Technology, Annamalai University Annamalainagar - 608002, Tamil Nadu, India, E-mail: rsuban82@gmail.com
Satyala Narayana, Gudlavalleru Engineering College Gudlavalleru-521356, India, E-mail: satyala1976@gmail.com

most commonly utilized strategy for the conclusion of CAD by doctors is angiography, which is also viewed as the most exact technique [14–17]. Apart from this, it has a significant consequence and considerable expense is related to it. In addition, dissecting an excessive number of elements, as aforementioned, to analyze a patient, makes the doctor's activity troublesome. Besides, traditionalist procedures for the determination of coronary illness are for the most part dependent on the investigation of the patient's clinical history, the survey of significant manifestations by a clinical professional, and physical assessment report [18–20]. However, these techniques frequently lead to lost determination because of human mistakes [14, 21]. Hence, there is a need for improvement in mechanized demonstrative framework dependent on machine learning for coronary illness finding that can resolve these issues.

To tenacity these complexities in intrusive-based diagnosis of coronary illness, a "noninvasive medical decision support system based on machine learning predictive models such as support vector machine (SVM), K-NN, ANN, DT, LR, AB, NB, FL, and rough set" [2, 3] has been developed by different scientists and generally utilized for coronary illness conclusion. Further, owing to this "machine-learning-based expert clinical decision system," the proportion of coronary illness demise diminished. Further, the choice of the most important features is intricate because of the scourge of dimensionality, and hence hybrid or new optimization algorithms can be detailed [22]. Moreover, the hybrid optimization algorithms have been reported to be advantageous for certain search problems [23–27]. At that point, with the extracted features, the classification using machine learning is a bit complex. In this way, there is a need to have a novel optimization methodology for classifying the information just as to choose the ideal features troublesomely and to enhance the prediction accuracy.

The major contributions of this research work are highlighted as follows:

- Proposed a new combination of features with lower statistical (central tendency) and higher-order statistical features (qualitative variation and degree of dispersion).
- Introduced the optimization assisting ensemble technique for accurate prediction, which includes SVM, random forest (RF), k-nearest neighbor (KNN), and optimized NN.
- A new improved algorithm (S-CDF) is introduced for NN training via tuning the optimal weights.

The rest of the paper is organized as follows: Literature review portrays the recent works undergone in literature corresponding to heart disease prediction. Proposed heart disease prediction framework: an architectural description is provided. Further, phase 1 – feature extraction: central tendency, degree of dispersion and qualitative variation is addressed. Phase 2 – dimensionality reduction via PCA and phase 3 – ensemble-based classification are depicted. Moreover, the proposed S-CDF algorithm for neural network training is discussed. The resultants acquired with the presented approach are discussed in Results and discussions, and a strong conclusion is given to the current research work in Conclusion.

Literature review

Related works

In 2019, Latha et al. [28] had investigated the precision of forecast of coronary illness utilizing an "ensemble of classifiers." The CHDD from the "UCI machine learning repository" was used for preparing and testing purposes. The ensemble calculations "bagging, boosting, stacking and majority voting" were utilized for tests. The outcomes from the proposed work indicated that majority voting created the most elevated improvement in exactness.

In 2018, Mathan et al. [29] had shown a decision tree (DT)-based framework based on neural classifiers for definite coronary illness forecast and early conclusion. The precision of coronary sickness forecast using this system was better than various methodologies. The utilization of the DT in envisioning coronary ailment had helped the authors in directing the prosperity of individuals. The assessment work had provided information on the likelihood of using DT computations in anticipating coronary disease.

In 2018, Vijayashree and Sultana [30] had investigated coronary illness by deploying the HRTM-DNN. Throughout the coronary illness diagnosing process, heart information was gathered from "UCI repository information," and noise present in the information was wiped out by figuring standard component scaling standardization process. From the clamor evacuated information, dimensionality overfitting information was decreased from the informational collection.

In 2019, Ali et al. [31] had built up a computerized symptomatic framework for the finding of coronary illness. The proposed analytic framework utilized two measurable models for feature extraction and χ^2-DNN for classification. The quality of the proposed analytic framework was assessed using six unique assessment measurements.

In 2019, Javeed et al. [32] had featured the issue of overfitting in the cardiovascular breakdown expectation and proposed a novel learning framework to encourage cardiovascular breakdown expectation. The learning framework hybridizes two calculations. The primary calculation was an RSA, which was used to look out for a subset of highlights having correlative data about the cardiovascular breakdown.

The subsequent calculation was RF that was utilized to anticipate cardiovascular breakdown based on the chosen subset of features.

In 2018, Ali et al. [33] had proposed a specialist framework dependent on stacked SVMs for the finding of HF malady. The first SVM model was utilized to eliminate unimportant highlights while the subsequent model was utilized as a prescient model. Both the models were upgraded by a half breed framework search calculation. It was indicated that the proposed technique beat 10 prestigious existing strategies in literature and other condition of the workmanship machine learning models. The proposed strategy was prepared for determining better outcomes with less number of highlights.

In 2019, Maragatham & Devi [34] had proposed a novel long short-term memory models (LSTM) premise prescient model structure for early determination of cardiovascular breakdown by utilizing the strategies of profound learning. It was identified that LSTM models (with and without window length) were best when compared with other basic strategies such as KNN, logistic regression, SVM, and MLP, respectively.

In 2019, Khiarak et al. [35] had considered information characterization of coronary illness and choice of the highlights. The point of this investigation was making a programmed framework to analyze coronary illness, characterize the patients, to have the option to utilize it in facilities. A joined technique with a meta-heuristic approach was used to improve determination, and the KNN algorithm was used to arrange and analyze coronary illness. The required tests actualized to evaluate the viability of the recommended calculation in the information order.

Review

The features and challenges of the existing works are given in Table 1. In the ensemble classifier [28], the prediction accuracy is improved and here the changes in one feature do not affect another feature. This technique needs to reduce the error for better detection accuracy. The Gini index DT data mining in Ref. [29] is less prone to noise and can upgrade the precision of the finding of heart disease. Further, HRTM-DNN in Ref. [30] is accurate and consumes less time. But it needs to solve the overfitting problem and dimensionality issues. The χ^2-DNN in Ref. [31] improves the quality of decision-making and consumes high time. Moreover, random search algorithm in Ref. [32] improves the training accuracy and here the overfitting rate needs to be reduced. Further, the optimized stacked support vector machines in Ref. [33] reduces the time complexity, and this could be much better if the misclassification errors are minimized. Moreover, LSTM model in Ref. [34] exhibits favorable effectiveness and feasibility. But this technique is inconsistent. In the KNN [35], the accuracy of the feature selection technique has been improved. But it requires improvement in feature selection and needs to select the relevant features for better robustness.

Proposed heart disease prediction framework: an architectural description

Figure 1 demonstrates the proposed heart disease prediction framework. The three significant phases of the model are: **proposed feature extraction, dimensionality reduction, and proposed ensemble-based classification**. The overall steps involved are as follows:

Step 1: Initially the input heart data (I_{in}) are normalized. Such that, the higher-order data gets converted into lower-order data (i.e., between 0 and 1 level).

Step 2: In the **proposed feature extraction phase**, the relevant features such as **lower-order statistical and higher-order statistical features** are extracted.

Step 3: The extracted lower-order statistical features or central tendency (f_{CT}) and higher-order statistical features such as dispersion (f_D) and qualitative variation (f_{QV}) are together termed as $F = f_{CT} + f_D + F_{QV}$.

Step 4: Nonetheless, the "curse of dimensionality" is being the greatest issue, **PCA**-based feature reduction approach is used in this work.

Step 5: These dimensional reduced features (F_{opt}) are fed as input to the **proposed ensemble classifier** that encompasses **SVM [36], RF [37], and KNN [38],** respectively. The final classification via optimized NN gives accurate results. As the novelty, the NN training is carried out by a new **S-CDF** model by tuning the optimal weights, which ensures the accurate prediction model.

At first, the input heart data (I_{in}) are normalized, so that the higher-order data reformed to lower-order data. The second phase is feature extraction, where the suitable features such as lower-order statistical of central tendency (Raw data) as well as the higher-order statistical features such as the degree of dispersion of the hard disk data and qualitative variation (raw data and the extracted lower order data) are extracted. Further, the overall extracted features are given by $F = f_{CT} + f_D + F_{QV}$. Here, the PCA-based feature reduction approach is utilized in this work to solve the problem of "curse of dimensionality." Furthermore, the dimensional reduced features (F_{opt}) are given as an input to the proposed ensemble classifier, which includes SVM [36], RF [37], and KNN [38], correspondingly. The final classification through optimized NN provides exact outcomes. Moreover, the NN

Table 1: Features and challenges of existing heart disease prediction framework.

Author [citation]	Methodology	Features	Challenges
Latha et al. [28]	Ensemble classifier	✓ Avoid overfitting ✓ Reduce the similarity differences	× Need to reduce the error × Higher time complexity
Mathan et al. [29]	Gini index decision tree data mining	✓ High accuracy and sensitivity ✓ Noise was reduced	× Requires upgradation in precision
Vijayashree and Sultana [30]	HRTM-DNN	✓ The accurate manner with minimum time	× Requires improvement in prediction rate × Need to solve the overfitting problem and dimensionality issues
Ali et al. [31]	χ^2-DNN	✓ Improve the ✓ Quality of decision-making ✓ Achieved higher detection accuracy	× Time complexity
Javeed et al. [32]	Random search algorithm	✓ Efficient and less complex ✓ Reduces the number of features.	× The overfitting rate needs to be reduced
Ali et al. [33]	Optimized Stacked support vector machines	✓ Reduces the training time ✓ Improve the decision-making process	× Need to minimize the × Misclassification errors
Maragatham & Devi [34]	LSTM model	✓ Diminish cost ✓ High scalability ✓ Less complex ✓ Training time is reduced significantly	× Inconsistent × Needs the lengthiest prediction time for a single patient high costs
Khiarak et al. [35]	KNN	✓ Improves the accuracy of feature selection	× Samples are not balanced × Lack of magnitude for some features

HRTM-DNN, Hybridized Ruzzo–Tompa Mimeti based deep trained neocognitron neural network approach; LSTM, Long short-term memory models; KNN, K-nearest neighbor.

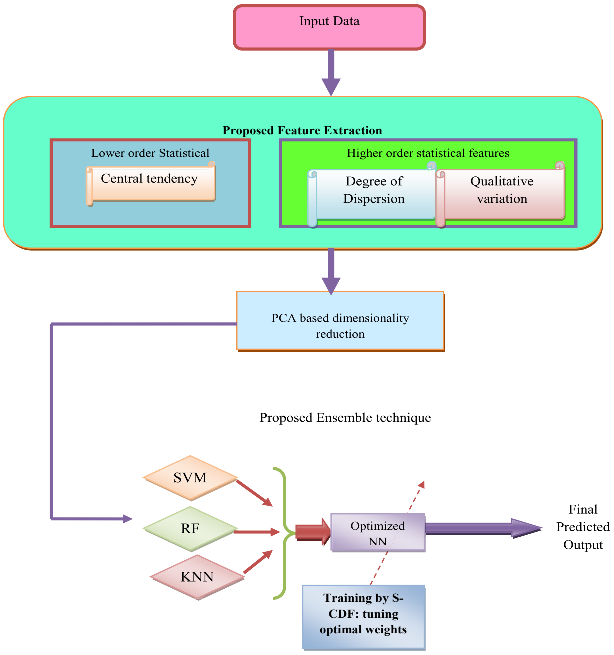

Input Data
Proposed Feature Extraction
Lower order Statistical
Central tendency
Higher order statistical features
Degree of Dispersion
Qualitative variation
PCA based dimensionality reduction
Proposed Ensemble technique
SVM
RF
KNN
Optimized NN
Final Predicted Output
Training by S-CDF: tuning optimal weights

Figure 1: Proposed heart disease prediction framework.

training is executed with the aid of new S-CDF model by tuning the optimal weights, which guarantees the accurate prediction model.

Phase 1 – feature extraction: central tendency, degree of dispersion, and qualitative variation

This is the initial phase, the lower-order and higher-order statistical features are extracted from the normalized input data (I_{norm}).

Central tendency

In statistics, “a central tendency (or measure of central tendency) is a central or typical value for a probability distribution.” The most widely recognized proportions of central tendency are the number-crunching “mean, the median and the mode." A central tendency can be determined for either a limited arrangement of qualities or for a hypothetical conveyance, for example, the theoretical dissemination. The following are the typical central tendency measures [39].

Arithmetic means or average or means: The aggregate of all estimations partitioned by the number of attributes such as age, sex, and BMI in the collected data. Mathematically, the mean can be expressed for the n values of $I_{norm(1),}I_{norm(2),} \dots I_{norm(n),}$ denoted as per Eq. (1). Here, $I_{norm(i)}$ is the value of the data point $i = 1, 2, \dots, n$.

$$\overline{I_{norm(i)}} = \frac{1}{n}\sum_{i=1}^{n} I_{norm(i)} \tag{1}$$

Median: The center worth that isolates the "higher half from the lower half" of the informational collection. In central tendency, the mode and median are referred to as the main proportions. It can be utilized for ordinal information, in which esteems are positioned comparatively with one another yet are not estimated completely. The median can be mathematically expressed as per Eq. (2)

$$Median = \left(\frac{I_{norm(i)} + 1}{2}\right) \tag{2}$$

Mode: The "most frequent value in the dataset." It is the main "central tendency measure" that is utilized with "nominal data," which has absolutely subjective class assignments.

Standard deviation (σ)**:** It quantifies how intently the information bunches about the mean. It is the "square base of the variance."

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(I_{norm(i)} - \overline{I_{norm(i)}}\right)^2} \tag{3}$$

The minimum and the maximum values are the initial and very last-order statistics.

GM: "The nth root of the result of the data values, where there are n of these." This is legitimate for information that is estimated completely on an absolute "positive scale." This can be defined mathematically as per Eq. (4).

$$Geometric\ Mean = \left(\prod_{i=1}^{n} I_{norm(i)}\right)^{\frac{1}{n}} \tag{4}$$

HM: The proportional of the number "arithmetic mean of the reciprocals of the data values." This measure also is legitimate for statistics that are estimated totally on an absolute "positive scale." The mathematical formula for harmonic mean is depicted in Eq. (5).

$$Harmonic\ Mean = \frac{n}{\sum_{i=1}^{n}\left(1/I_{norm(i)}\right)} \tag{5}$$

TM: The Arithmetic Mean (AM) of data values after a specific numeral or degree of the most elevated and least data values have been disposed of.

Interquartile mean: A shortened mean dependent on data inside the assortment of the "interquartile." The mathematical formula for interquartile mean is mathematically expressed in Eq. (6).

$$Interquartile\ mean = \frac{2}{n}\sum_{i=\frac{n}{4}+1}^{3n/4} I_{norm(i)} \tag{6}$$

Midrange: The AM of the most extreme and least estimations of data collection. It is defined mathematically as per Eq. (7).

$$Midrange = \frac{\max\left(I_{norm(i)}\right) + \min\left(I_{norm(i)}\right)}{2} \tag{7}$$

Midhinge: The AM of the "first and third quartiles" (Q).

$$Midrange = \overline{Q_{1,3}\left(I_{norm(i)}\right)} = \frac{Q_1\left(I_{norm(i)}\right) + Q_3\left(I_{norm(i)}\right)}{2} \tag{8}$$

Trimean: The weighted AM of the middle and "two quartiles." It is defined mathematically as per Eq. (9).

$$Trimean = \frac{Q_1 + 2Q_2 + Q_3}{4} \tag{9}$$

Winsorized mean: An arithmetic mean in which outrageous qualities are supplanted by values nearer to the median. The extracted central tendency features are denoted as f_{CT}.

Degree of dispersion

In statistics, "dispersion (additionally called variability, scatter, or spread) is the degree to which a conveyance is extended or squeezed" [40]. Instances of dispersion measures include:

IQR: In eloquent statistics, IQR also called the "misread, middle 50%, or Spread" is a proportion of "statistical dispersion," which is being equivalent to the distinction somewhere in the range of 75th and 25th percentiles or among greater and minor quartiles.

Range: In measurements, the scope of a lot of information is the distinction between the biggest and smallest values. The difference here is explicit; the range of the information set is the result of subtracting the smallest value from the largest value.

Mean absolute difference (otherwise called GMAD): It is a proportion of measurable dispersion equivalent to the AAD of two free qualities drawn from a likelihood distribution. "A related measurement is the relative mean absolute difference, which is the mean outright contrast partitioned by the arithmetic mean, and equivalent to double the Gini coefficient."

MAD: In insights, MAD is a vigorous proportion of the changeability of a univariate test of quantitative information. It can likewise allude to the populace parameter that is evaluated by the MAD determined from a sample.

Average absolute deviation (or simply called average deviation): The “Average absolute deviation” or MAD, of an informational index, is the normal of the total deviations from the middle point.

Distance standard deviation: In insights and likelihood hypothesis, separation distance correlation or distance covariance is a proportion of reliance between two combined irregular vectors of discretionary, not really equivalent, measurement.

Different proportions of dispersion are “dimensionless.” As it were, they have no units regardless of whether the variable itself has units. These include:

Coefficient of variation: In likelihood hypothesis and insights, the CV, otherwise called RSD, is a normalized proportion of scattering of likelihood circulation or recurrence dissemination. It is regularly communicated as a rate and is characterized as the proportion of the “standard deviation to the mean” (or its total worth).

Quartile coefficient of dispersion: In insights, it is an illustrative measurement that estimates dispersion and which is utilized to make examinations inside and between informational collections. The measurement is effectively figured utilizing the first ($Q1$) and third ($Q3$) quartiles for every datum set. The quartile coefficient of scattering is given in Eq. (10).

$$\text{quartile coefficient} = \frac{Q3 - Q1}{Q3 + Q1} \tag{10}$$

The relative mean difference, equivalent to double the Gini coefficient: The MAD (univariate) is a proportion of factual scattering equivalent to the AAD of two free qualities drawn from likelihood dissemination. A related measurement is the AAD, which is the “mean absolute difference divided by the arithmetic mean and equal to twice the Gini coefficient.”

Mean absolute difference: It is characterized as the “normal” or “mean,” officially the expected value, of the total distinction of two arbitrary factors X and Y autonomously and indistinguishably disseminated with the equivalent (obscure) conveyance from now on called Q.

Entropy: The entropy of a discrete variable is “location-invariant and scale-independent” and subsequently not a proportion of dispersion in the above sense, the entropy of a constant variable is area-invariant and added substance in scale: If H_z is the entropy of nonstop factor z and $y = ax + b$, at that point $Hy = Hx + \log(a)$. Here, a, b are the real variables. Here, x, y are two events. The entropy function can be mathematically modeled for the normalized image as: $H(I_{norm(i)}) = -\sum_{i=1}^{n} P(I_{norm(i)}) \log_b P(I_{norm(i)})$. The extracted dispersion features are denoted as f_D.

Table 2: Transformed levels.

Normalized data range	Transformed data level
If 0–0.25	1
If 0.25–0.5	2
If 0.5–0.75	3
If 0.75–1	4

Qualitative variation

An IQV is a “measure of statistical dispersion in nominal distributions” [41]. The normalized data within the bounds 0 and 1 gets transferred into four levels. Table 2 shows the transferred data levels.

Wilcox’s indexes: it encloses “ModVR, RanVR, AvDev, MNDif, MNDif, StDev, HRel, B index, and R packages.” Table 3 shows the features based on Wilcox’s index. The features based on Gibb’s indices and related formulae are illustrated in Table 4. Features based on single-order sample indices are depicted in Table 5.

The extracted qualitative variation features are denoted as f_{QV}. Thereby, the final feature set combining the extracted lower-order statistical features or central tendency (f_{CT}) and higher-order statistical features such as dispersion (f_D) and qualitative variation (f_{QV}) is determined in Eq. (11)

$$F = f_{CT} + f_D + F_{QV} \tag{11}$$

Figure 2A, B portrays the normalized version of the feature map and the features plotted in the logarithmic scale, respectively for better understanding. Here, the red color indicates the normal data and the green color indicates the abnormal data. From the plots, it can be observed that the proposed feature maps distinguish the normal data from the abnormal data, wherever the raw data fails to do. For instance, the attribute ID in Figure 2A, the normal and abnormal raw data overlaps with each other, whereas the qualitative variation features distinguish them. Similarly, attributes 9–10 can be more distinguished by dispersion and central tendency features than the raw data (source: Figure 2B). Even though the proposed features could not distinguish certain attributes, the joint performance of them can facilitate the classifier to recognize the normal and abnormal categories of heart disease more precisely.

Table 3: Features based on Wilcox's index.

Wilcox's indexes	
ModVR	$ModVR = \frac{Bv}{B-1}$ Here, $B \rightarrow$ count of categories and $v = 1 - \frac{fr_m}{N}$, here, v is the Freeman's index and fr_m is the modal frequency $N \rightarrow$ size of the overall count of samples
RanVR	$RanVR = \frac{fr_1}{fr_m}$ Here, $fr_1 \rightarrow$ lowest frequency.
AvDev	The first is based on AvDev. $StDev_1 = 1 - \sqrt{\frac{\sum_{i=1}^{B}\left(fr_i - \frac{N}{B}\right)^2}{\left(fr_i - \frac{N}{B}\right)^2 + (B-1)\left(\frac{N}{B}\right)^2}}$ The second is based on MNDif $StDev_2 = 1 - \sqrt{\frac{\sum_{i=1}^{B}\left(fr_i - \frac{N}{B}\right)^2}{\left(N - \frac{N}{B}\right)^2 + (K-1)\left(\frac{N}{B}\right)^2}}$
MNDif	$MNDif = 1 - \frac{1}{N(B-1)} \sum_{i=1}^{B-1} \sum_{j=i+1}^{B} \left\lvert fr_i - fr_j \right\rvert$ Here, fr_j and fr_i are the jth and ith frequencies.
HRel	$HRel = \frac{-\sum p_i \ \log_2 p_i}{\log_2 B}$ Here, $p_i = \frac{fr_i}{N}$
B index	$B_{index} = 1 - \sqrt{1 - \left[B\sqrt{\prod_{i=1}^{B} \frac{B.fr_i}{N}}\right]^2}$
VarNC	$VarNC = 1 - \frac{1}{N^2} \cdot \frac{B}{B-1} \sum \left(fr_i - \frac{N}{B}\right)^2$

Table 4: Features based on Gibb's indices and related formulae.

M1	$M1 = 1 - \sum_{i=1}^{B} p_i^2$
M2	$M2 = \frac{B}{B-1}(1 - \sum_{i=1}^{B} p_i^2)$ Here, $\frac{B}{B-1} \rightarrow$ standardization factor
M4	$M2 = \frac{B}{B-1}(1 - \sum_{i=1}^{B} p_i^2)$
M6	$M4 = \frac{\sum_{i=1}^{B} \lvert X_i - m \rvert}{2 \sum_{i=1}^{B} X_i}$ $m \rightarrow$ mean

Phase 2 – dimensionality reduction via PCA

PCA

At a certain point, more features or dimensions can decrease a model's accuracy since there is more data that needs to be generalized, and this is known as the curse of dimensionality [42]. The extracted feature set F was offered as input to the subsequent process to minimize the feature's dimensions. Accordingly, in the initial phase, a group of data vectors denoted as $F f_1, \ldots, f_d$ are preferred from the input and f^0 points out the particular cluster examination of i variables. The "empirical mean curve" computation is carried out with all co columns: co=1, …, i. The consequential means value is positioned in $l_{(i)}$ with $i \times 1$ dimensions that are demonstrated in Eq. (12).

$$l_{(i)}[co] = \frac{1}{F} R[v, co] \tag{12}$$

The "mean deviation" is computed as shown in Eq. (13), where J indicates the "$F \times i$ matrix and j is the column vector $F \times 1$ of all 1s: $j[f] = 1$".

$$J = R - jco^T \tag{13}$$

The manipulation of the "covariance matrix" $CV^{(X)}$ is made as expressed in Eq. (14). Together the eigenvalues and eigenvector are computed by valuing the matrix X that furthermore diagonalizes $CV^{(X)}$ as given in Eq. (15), where $f^{(X)}$ signifies the diagonal matrix of eigenvalues.

$$CV^{(X)} = \frac{1}{F-1} J^* J \tag{14}$$

$$X^{-1} CV^{(X)} X = f^{(X)} \tag{15}$$

The "Eigenvalue matrix" minimization of $f^{(X)}$ is obtained by column arrangement of "eigenvector matrix," represented by Y. The manipulation of "cumulative energy content" cu is revealed in Eq. (16) that holds the integration of the "energy content of Eigen values from 1 via co."

$$cu[co] = \sum_{\widehat{D}=1}^{co} f^{(X)}[F, F] \ for \ co = 1, \ldots .i \tag{16}$$

The "subset of eigenvectors" is selected by means of storing the cm column of X as Z matrix. The expense of cm is chosen by means of deploying the vector. "The column of $\widehat{X}$ matrix, $\widehat{X} = C \cdot KLT\{L\}$ is the vector" that verifies the "Kosambi-Karhunen-Loeve transform" in L row matrix. The dimension minimized features are merged to form F_{opt} that is given by Eq. (17).

$$F_{opt} = \frac{J}{j.t'^{\widehat{X}}} \tag{17}$$

where, "$t' = t'(co) = \{\sqrt{F_{opt} V^{(X)} \lvert co, co \rvert}\}$: co=1, …, i" denotes the process as a consequence in the dimensional minimized feature F_{opt}.

Phase 3 – ensemble-based classification

Ensemble techniques: It is a methodology that is typically utilized for enhancing classifier precision. It is a powerful

Table 5: Features based on single-order sample indices.

Shannon–Wiener index	$j = H/\log_e(S)$ J denotes the related index $H = \log_e N - \frac{1}{N}\sum p_i n_i \ \log(p_i)$
Brillouin index of diversity	$E_B = I_B / I_{B(\max)}$ $I_B = \frac{\log(N!) - \sum_{i=1}^{B} \log(n_i!)}{N}$ $N!$ is the factorial of overall count of individuals N in the population n_i is the count of individuals in the ith category E_B is the Brillouin's index of evenness
Hill's diversity numbers	$N_c = \frac{1}{[\sum_{i=1}^{B} p_i^c]^{c-1}}$ when c=0; N_c=species richness when c=1; N_c="Shannon's index" when c=2; N_c="1/Simpson's index (without the small sample correction)" when c=3; N_c=1/"Berger-Parker index"
Margalef's index	$In_{\text{marg}} = \frac{S-1}{\log_e N}$ S represents the count of data types in the sample N is the size of the overall count of samples
Menhinick's index	$In_{Men} = \frac{S}{\sqrt{N}}$
Berger–Parker index	"The Berger-Parker index equals the maximum value $p_i = \frac{fr_i}{N}$ in the dataset, i.e. the proportional abundance of the most abundant type. This corresponds to the weighted generalized mean of the p_i values when q approaches infinity, and hence equals the inverse of true diversity of order infinity (1/∞D)."
Q statistic	$Q = \frac{1/2(n_{R1} + n_{R2}) + \sum_{j=R1+1}^{R2-1} n_j}{\log(R2/R1)}$ where $R1$ and $R2$ are the "25 and 75% quartiles respectively on the cumulative species curve", n_j is the count of species in the jth n_{R1}"category is the number of species in the class where Ri falls (i=1 or 2)".
Nee, Harvey, and Cotgreave's index	"This is the slope of the log(abundance)-rank curve"
Rényi entropy	$H^q = \frac{1}{1-q}\ln(\sum_{i=1}^{B} p_i^q) = \ln(D^q)$ D^q is the "Hill number"
McIntosh's D and E	$D = \frac{N - \sqrt{\sum_{i=1}^{B} n_i}}{N - \sqrt{N}}$ $E = \frac{N - \sqrt{\sum_{i=1}^{B} n_i}}{N - N/\sqrt{N}}$
Fisher's alpha	$B = \sigma \ \ln(1 + N/\sigma) N = \sigma \ \ln(1 - X)$
Strong's index	

Table 5: (continued)

	$D_w = \max\left[\frac{c_i}{B} - \frac{i}{N}\right]$
Simpson's E	$E = \frac{{}^{1}/_{D}}{B}$
Smith & Wilson's indices	$E_1 = \frac{1-D}{1-{}^{1}/_{B}} E_2 = \frac{\log_e D}{\log_e B}$
Heip's index	$Heip' sindex = \frac{e^H}{B}$
Camargo's index	$Camargo's \;\; index = 1 - \sum_{i=1}^{K} \sum_{j=1+1}^{K} \frac{p_i - p_j}{B}$
Smith and Wilson's B	$S \,\&\, W = 1 - \frac{2}{\pi} \arctan(\theta)$
Caswell's V	$V = \frac{H - E(H)}{SD(H)}$ H denotes the "Shannon Entropy" $E\,(H)$ is the "expected Shannon entropy for neural model" $SD\,(H)$ is the "standard deviation for entropy"
Bulla's E	$E_c = \frac{0 - {}^{1}/_{B}}{1 - {}^{1}/_{B}} E_d = \frac{0 - {}^{1}/_{B} - (B - 1/N)}{1 - {}^{1}/_{B} - (B - 1/N)}$
Horn's information theory index	$R_{ik} = \frac{H_{\max} - H_{obs}}{H_{\max} - H_{\min}} \; X = \sum x_{ij} \;\; Y = \sum x_{kj} \;\; H(X) = \sum \frac{x_{ij}}{X} \;\; \log \frac{X}{x_{ij}} \;\; H(Y) = \sum \frac{x_{kj}}{Y} \log \frac{Y}{x_{kj}} H_{\max} = \sum \left(\frac{x_{ij}}{X+Y} \;\; \log \frac{X+Y}{x_{ij}} - \frac{x_{kj}}{X+Y} \;\; \log \frac{X+Y}{x_{kj}} \right)$ $H_{\min} = \sum \left(\frac{x_{ij} + x_{kj}}{X+Y} \log \frac{X+Y}{x_{ij} + x_{kj}} \right) H_{\min} = \frac{X}{X+Y} H(X) + \frac{Y}{X+Y} H(Y) H_{obs} = \sum \left(\frac{x_{ij} + x_{kj}}{X+Y} \log \frac{X+Y}{x_{ij} + x_{kj}} \right)$
Rarefaction index	$fr_n = E[X_n]$
Average taxonomic distinctness index	$TD = 2 \frac{\sum\sum_{i<j} \omega_{ij}}{s(s-1)}$ s is the count of "host species used by a parasite" ω_{ij} is the "Taxonomic distinctness" between host species i and j
Lloyd & Ghelardi's index	$LG = \frac{B}{B}$ B is the count of categories B is the number of categories according to "MacArthur's broken stick model yielding the observed diversity".
Index of qualitative variation	$IQV = \frac{B}{B-1} (1 - \sum_{i=1}^{B} (p_i/100)^2)$
Theil's H	$TD = 2 \sum\sum_{i<j} \psi_{ij} \Big/ s(s-1)$ ψ_{ij} is the taxonomic distinctness between host species s is the "number of host species used by a parasite"

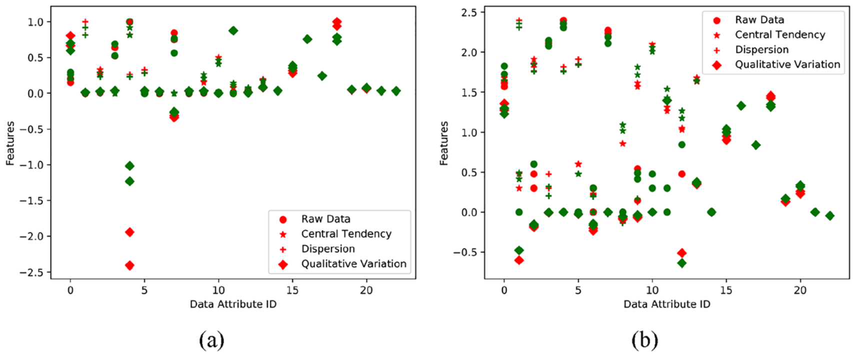


Figure 2: Feature mapping between raw data and central tendency, degree of dispersion, and qualitative variation. (a) Normalized version of the feature map. (b) Features plotted in logarithmic scale.

“Meta characterization strategy” that consolidates feeble beginner (learner) with solid learners to perk up the viability of frail beginner. In this work, the “ensemble strategy” is utilized to enhance the exactness of different algorithms for coronary illness forecast. The point of joining various classifiers is to get enhanced execution as contrasted with an individual classifier. The functioning of the individual classifiers is clarified in the following area:

SVM-based classification

SVM is known for its simple nonlinear regression operation through which classification problems can be well handled [36]. To discover the isolating “hyperplane in the SVM algorithm,” it is important to take care of the double issue of looking through a “saddle” purpose of “Lagrange work,” which can be decreased to the issue of “quadratic programming”:

$$-L(\lambda) = -\sum_{d=1}^{Q}\chi_d + 1/2.\sum_{d=1}^{Q}\sum_{\kappa=1}^{Q}\chi_d\cdot\chi_\kappa\cdot y_d\cdot y_\kappa\cdot\lambda(z_d,z_\kappa)\to\min_{\chi} \qquad (18)$$

$$\sum_{d=1}^{Q}\chi_d\cdot y_d = 0, 0\le\chi\le Cr, d=\overline{1,Q}$$

Here, χ_d denotes the dual variable, z_d denotes the object that is present in the training data set F, y_d represents a number that is +1 or −1, and $\lambda(z_d,z_\kappa)$ denotes the kernel function. Further, Cr is the regularization parameter, Q denotes the count of the object that is said to be available in the training data set.

For the most part, the SVM classifier, even with the default settings, gives a top-notch information grouping. To perk up the exactness of the SVM order, it is important to break down the area of the mistakenly grouped objects. By and large, the mistakenly ordered items are situated close to the isolating hyperplane. Consequently, it is important to utilize the extra instruments to advance the order excellence for the articles inside the isolating strip. The classified result from SVM is denoted as Ca_{SVM}.

RF-based classification

RF algorithm utilizes the group (ensemble) of DT [37]. Here, the excellent grouping is accomplished by means of merging the countless straightforward classifiers (DT). Further, with respect to the reactions total of numerous trees, the acquisition of the last grouping result takes place. The characterization quality is enhanced with a reduction in the overriding issues since the RF correlates with one DT. Moreover, it consolidates the thoughts of packing, the “bootstrap” accumulating, and the “irregular subspace technique.” The characterization of the estimations of the parameters is significant on account of this algorithm. The primary constraints of this algorithm are the accompanying: the “quantity of trees in the forest, the number of attributes such as age, sex, and BMI in the collected data to consider when searching for the best split, the most extreme profundity of the tree, the parting basis.” The Gini debasement can be utilized as the parting model:

$$Gini_t = 1-\sum_{l=1}^{v}U^2(Y_h) \qquad (19)$$

where $U(Y_h)$ Indicates the subset of the element of Y_h class that is said to be available in the tree node t. The splitting criterion is evaluated as per Eq. (20) in case of binary classification

$$Gini_i^{split} = M_1/_M\,Gini_{t_1} = M_2/_M\,Gini_{t_2}\to\min \qquad (20)$$

M indicates the count of objects in the current node $t.M_1$ and M_2 determines the count of objects in the nodes t_1 and t_2 corresponding to the “left and right descendants-nodes in the case of the binary tree.” The classified result from the RF classifier is indicated as Ca_{RF}.

KNN-based classification

The KNN's calculation is a straightforward, simple-to-execute regulated machine learning calculation that can be utilized to tackle both characterization and relapse issues [38]. The KNN calculation is a directed learning technique. This implies all the information is marked and the calculation figures out how to foresee the yield from the information. It performs well regardless of whether the preparation information is enormous and contains boisterous values. The information is separated into preparing and test sets. The train set is utilized for model structure and preparation. A K– value is concluded, which is regularly the square root of the number of attributes such as age, sex, and BMI in the collected data. Presently the test information is anticipated on the model constructed. There are distinctive separation measures. For consistent variables, Euclidean separation, Manhattan distance, and Minkowski distance measures can be used. However, the usually utilized measure is Euclidean distance. It can be mathematically expressed as per Eq. (21). Here, X_l and Y_l are two points on the feature set F.

$$Euclidean\ distance = \sqrt{\sum_{l=1}^{K} X_l - X_l} \tag{21}$$

The classified result from KNN is indicated as Ca_{KNN}.

The classified results (Ca_{SVM}), (Ca_{RF}), and (Ca_{KNN}) are together $(Ca = Ca_{SVM} + Ca_{KNN} + Ca_{RF})$ subjected to optimized NN for final classified results.

Optimized NN-based classification

NN classifier is exploited to classify heart disease data. The network representation is offered in Eqs. (22)–(24) and the predicted label $\widehat{g}_m$ is described in Eq. (24), where g_m is the actual label [22]. In addition, n_h refers to the count of hidden neurons, n_q indicates the count of input neurons, and $w^{(o)}_{(qm)}$ addresses the output weight from qth hidden neuron to mth layer.

$$e^{(H)} = NF\left(w^{(H)}_{(bq)} + \sum_{j=1}^{n_{qi}} w^{(H)}_{(jq)} Ca \right) \tag{22}$$

$$\widehat{g}_m = NF\left(w^{(o)}_{(bm)} + \sum_{q=1}^{n_h} w^{(o)}_{(qm)} e^{(H)}_q \right) \tag{23}$$

$$w^{*} = \underset{\left\{ w^{(H)}_{(bq)}, w^{(H)}_{(jq)}, w^{(o)}_{(bm)}, w^{(o)}_{(qm)} \right\}}{\arg\ \min} \sum_{m=1}^{n_o} g_m - \widehat{g}_m \tag{24}$$

where q points out the hidden neuron, $w^{(H)}_{(bq)}$ specifies the bias weight to qth hidden neuron, $w^{(o)}_{(bm)}$ signifies the output bias weight to mth layer, and as the novelty, the training of NN will be carried out by proposed S-CDF via tuning the weight w by fixing Eq. (24) as objective (fitness).

Proposed S-CDF algorithm for neural network training

Proposed S-CDF model

The Sea lion is considered as one of the savviest creatures. Sea lions live in colossal states, which have a great many individuals [43]. There are a lot of subgroups that encapsulate their own chain of command. The existing SlnO model often suffers from slow convergence, and hence this work improves in the model that aids in tuning the weights precisely. The primary periods of chasing conduct of sea lions are as follows:

- Tracking and pursuing the victim utilizing their bristles.
- Calling different individuals that united their subgroup, pressing together and circling the prey.
- Assault in the direction of the victim.

In this work, this chasing method of sea lions is scientifically demonstrated to structure the S-CDF algorithm and perform streamlining.

Mathematical model of proposed S-CDF approach:

(1) ***Detecting and the following stage*:** Search agent (sea lion) is well thought out as a pioneer for this chasing system and different individuals renew their situations toward objective prey. SLnO calculation expects the objective prey is the present best arrangement or near-ideal arrangement. This is expressed mathematically in Eq. (25). Further, in Eq. (26), the **Canberra distance is computed between the prey and the solution.**

$$\overrightarrow{Dis} = \left| 2\vec{G}\vec{M}(T) - \vec{S}(T) \right| \tag{25}$$

$$\vec{S}(T+1) = \vec{M}(T) - \overrightarrow{Dis}\,.\vec{H} \tag{26}$$

Here, the distance between the objective victim and the search agent is symbolized as $\overrightarrow{Dis}$, and the vector position $\vec{S}(T)$ and target prey $\vec{M}(T)$ assist in position update. The current iteration is represented using the term T and random vector $\vec{G}$ residing in the interval [0, 1] is multiplied by 2 to enhance the search space and to acquire a near-optimal solution.

(2) ***Vocalization stage:*** Sea lions are viewed as creatures of land and water. Also, they have little ears that are

competent to distinguish sounds under or more water. Along these lines, when the search agent distinguishes a victim, he calls different individuals to encompass and assault the prey. This conduct is scientifically defined as in Eqs. (27)–(29). The swiftness of the search agent is depicted as $\overrightarrow{S_{leader}}$ and the "speed of their sound in water and air" is indicated as $\overrightarrow{P_1}$ and $\overrightarrow{P_2}$.

$$\overrightarrow{S_{leader}} = \left|\overrightarrow{P_1}\left(1+\overrightarrow{P_2}\right)\Big/\overrightarrow{P_2}\right| \tag{27}$$

$$\overrightarrow{P_1} = \sin\,\theta \tag{28}$$

$$\overrightarrow{P_2} = \sin\,\phi \tag{29}$$

(3) ***Attacking stage (Exploitation stage):*** Search agents will have the option to perceive the situation of objective victim and enclose them. The pursued strategy is directed by the pioneer (most excellent hunt operator) who identifies the prey and instructs other individuals about them. To numerically display the chasing conduct of search agent, two stages are presented as follows:

(a) ***Dwindling encompassing procedure:*** **This mechanism relies upon the new expression estimation in Eq. (30). Here, the distance function** ***Dis*** **is introduced in addition**. The distance among the best optimal solutions (objective victim and the seekout agent) is determined as $\left|\overrightarrow{M}(T) - \vec{S}(T)\right|$ interval, [−1, 1] the random number is indicated as l. The sea lions move along a circular path while chasing these bait balls and $\cos(2\pi l)$ defines this mechanism.

$$\vec{S}(T+1) = \left|Dis.\,(\overrightarrow{M_i}(T) - \vec{S_i}(T)).\cos(2\pi l)\right| + \overrightarrow{M_i}(T) \tag{30}$$

$$\overrightarrow{M_i}(T) - \overrightarrow{S_i}(T) = 0.5\left(\frac{M_i(T)}{M_{\max}(T)} + \frac{S_i(T)}{S_{\max}(T)}\right)\overrightarrow{M_i}(T) - \overrightarrow{S_i}(T) \tag{31}$$

(b) ***Circle refreshing situation:*** The sea lions pursue trap bundle of fishes and chase them beginning from edges.

(4) ***Searching for prey (Exploration stage):*** Typically, the search agent searches arbitrarily utilizing their whiskers to discover the victim [43]. Subsequently, this circumstance obliges a search agent to scan for other prey. This is mathematically shown in Eq. (32) and (33), respectively.

$$\overrightarrow{Dis} = 2\left|2\vec{B}.\overrightarrow{S_{md}}(T) - \vec{S}(T)\right| \tag{32}$$

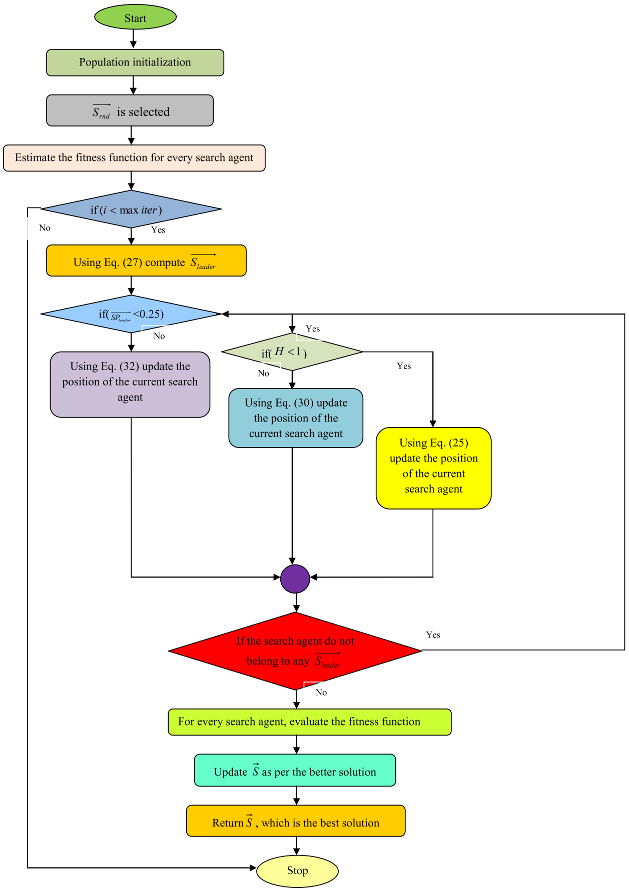

Start
Population initialization
$\overrightarrow{S_{rnd}}$ is selected
Estimate the fitness function for every search agent
if (i < max $iter$)
No
Yes
Using Eq. (27) compute $\overrightarrow{S_{leader}}$
if($\overrightarrow{SP_{leader}}$ <0.25)
Yes
No
if(H <1)
No
Yes
Using Eq. (32) update the position of the current search agent
Using Eq. (30) update the position of the current search agent
Using Eq. (25) update the position of the current search agent
If the search agent do not belong to any $\overrightarrow{S_{leader}}$
Yes
No
For every search agent, evaluate the fitness function
Update $\vec{S}$ as per the better solution
Return $\vec{S}$, which is the best solution
Stop

Figure 3: Flowchart of the proposed S-CDF algorithm.

$$\vec{S}(T+1) = \overrightarrow{S_{md}}(T) - \overrightarrow{Dis}.\vec{H} \tag{33}$$

The flowchart of the presented work is shown in Figure 3. The pseudo-code of the proposed S-CDF is shown in Algorithm 1.

```
Algorithm 1 Pseudo-code of proposed CDF algorithm.
Population initialization
$\overrightarrow{S_{rnd}}$ selection
For each search agent, evaluate the fitness
if1 ($i < \max^{it}$)
  compute the $\overrightarrow{S_{leader}}$ using Eq. (27),
  If2 ($\overrightarrow{SP_{leader}}$<0.25)
    If3 ($H < 1$)
      The position regarding the current search agent is
updated with respect to Eq. (30),
    Else
      The position regarding the current search agent is
updated with respect to Eq. (25)
    End if3
  Else
```

(continued)

```
        The position regarding the current search agent is
updated with respect to Eq. (32)
    End if2
    If the search agent does not belong to any $\overrightarrow{S_{leader}}$
        Go to the first if the condition
    Else
        For each of the search agent, evaluate the fitness
function
        Update $\vec{S}$ as per the better solution
        Return $\vec{S}$, which is the best solution
    Endif
Endif1
Stop
```

Results and discussions

Simulation procedure

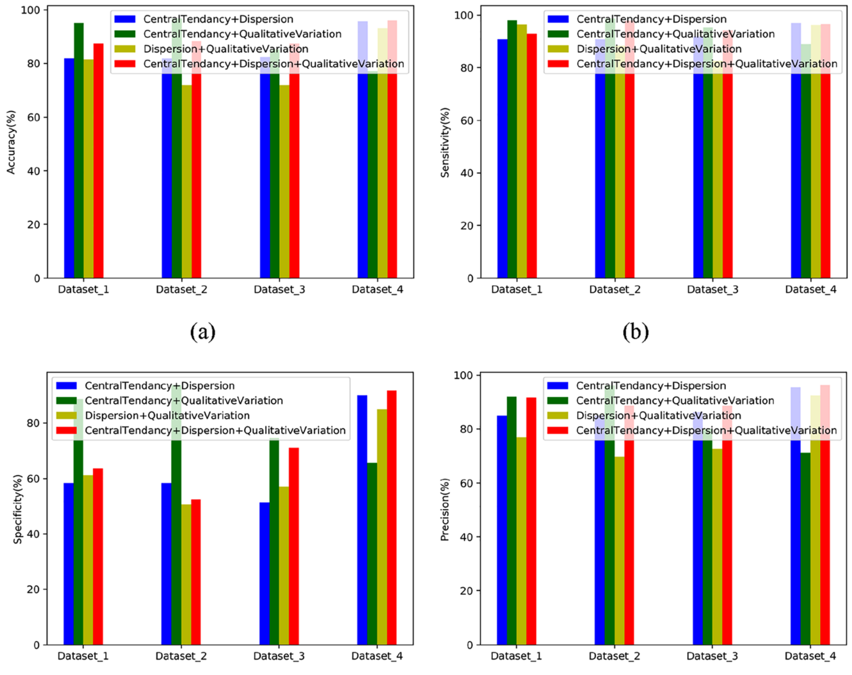
The proposed heart disease prediction approach with ensemble classification was implemented in PYTHON and the resultant acquired was noted. This evaluation was undergone to analyze the classification performance and impact of feature extraction as well. The performance of the proposed work was compared with other traditional models with respect to positive and negative measures such as “accuracy, sensitivity, specificity, precision, FPR, FNR and FDR, respectively.”

Data set description

The evaluation is accomplished for four diverse set of data sets: “dataset-1, dataset-2, dataset-3 and dataset-4,” respectively. The data set for evaluation is downloaded from: “https://archive.ics.uci.edu/ml/datasets/Heart+Disease” [Access date: 2020-05-11]. “This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The goal field refers to the presence of heart disease in the patient. It is integer-valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1, 2, 3, 4) from absence (value 0).”

Impact of feature extraction

The feature extraction plays a major role in this research work. Since, lower- and higher-order statistical features are deployed here, it is essential to analyze its impact on the achievement of the objective. The impact of feature extraction in the proposed disease prediction model is graphically shown in Figure 4. The analysis is done for all



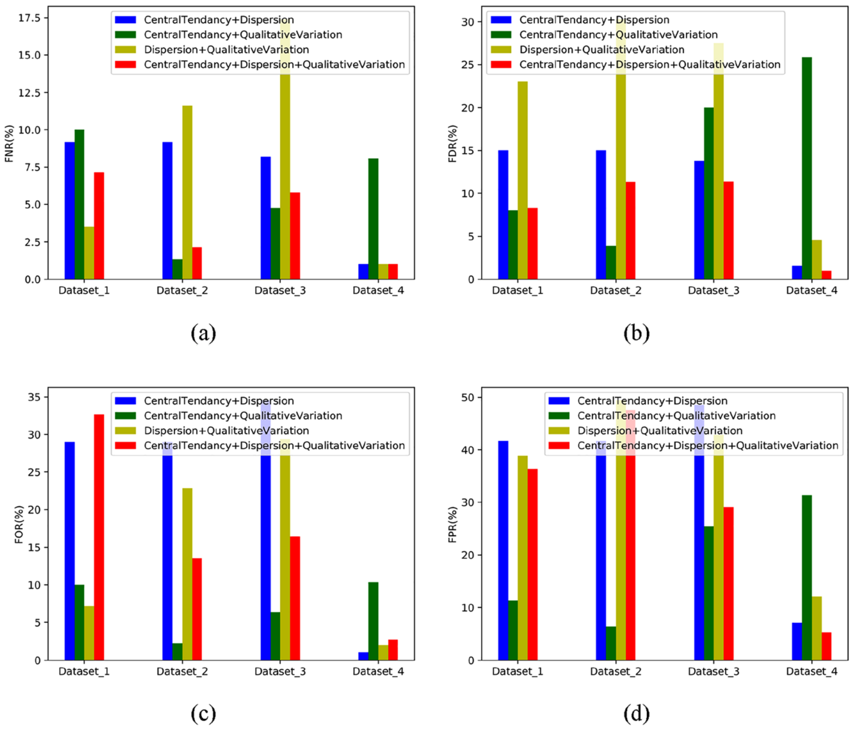
(a)

(b)

(c)

(d)

Figure 4: Impact of proposed feature extraction on presented approach for data set 1, data set 2, data set 3, and data set 4. (a) Accuracy, (b) sensitivity, (c) specificity, and (d) precision.



(a) (b)

(c) (d)

Figure 5: Impact of proposed feature extraction on presented approach for data set 1, data set 2, data set 3, and data set 4. (a) FNR, (b) FDR, (c) FOR, and (d) FPR.

the four data sets individually with different variations in the feature set. Accordingly, better prediction results are obtained by the proposed classifier with the proposed feature combination (central tendency + degree of dispersion + qualitative evaluation) particularly under data set 4. From this, it is observed that the accuracy of the proposed work with the proposed feature set is 5, 42.85, and 10% better than the performance with other feature sets such as central tendency + dispersion feature, central tendency qualitative variation, and dispersion qualitative variation, respectively. Similarly, the analysis is done for all other measures under different data sets.

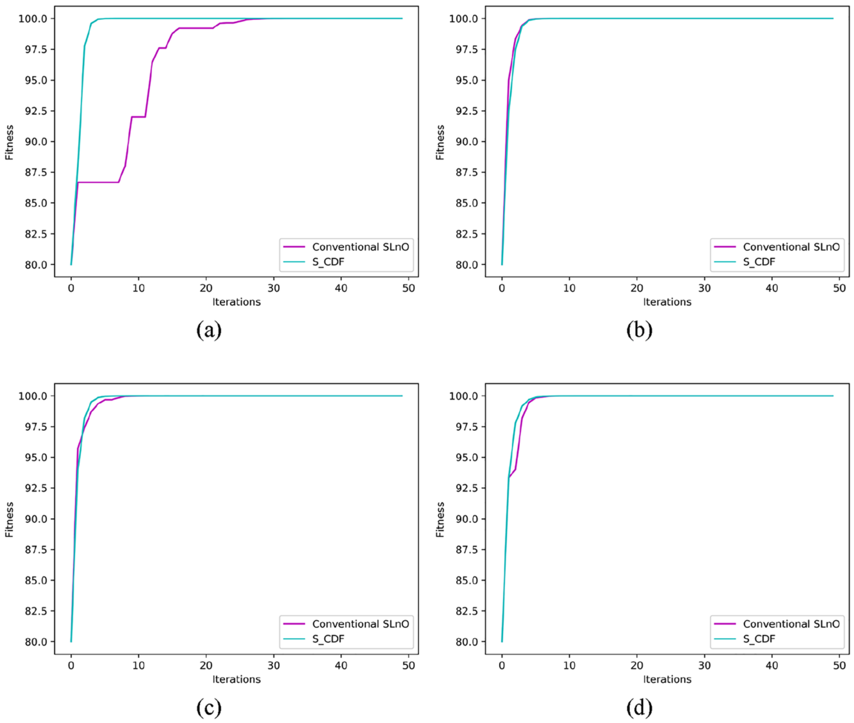
The impact of the proposed feature extraction for presented features as well as existing features in terms of negative performance is graphically illustrated in Figure 5. The presented features are lower for data set 4, and in the case of other data sets, there is small fluctuation. On the other hand, the FDR being the negative measure is lower for data set 4 in the case of the presented features, and it is 33.3, 92, and 60% better than the performance with existing features such as central tendency + dispersion feature, central tendency qualitative variation, and dispersion qualitative variation, respectively. The FOR of the presented features for data set 4 is 50, 93.3, and 66.6% better than the performance with conventional features such as central tendency + dispersion feature, central tendency qualitative variation, and dispersion qualitative variation, respectively. Altogether, it is proved that the proposed feature set is highly influenced in attaining better prediction performance.

Convergence analysis

The convergence tells about the achievement of the objective function by the presented optimization algorithm (S-CDF). The objective of the current research work deals with the maximization of the detection accuracy, which leads to the minimization of errors. Since, accuracy is the inverse of errors. The resultant of convergence for four sets of data sets: data set 1, data set 2, data set 3, and data set 4 is shown graphically in Figure 6. This evaluation is undergone by varying the count of iterations for both the presented work and the existing SLnO. On observing Figure 6A corresponding to data set 1, the presented work shows the maximal convergence, while compared to the existing one. At the 10th iteration, the presented work is 14% better than the existing SLnO. Alike this, in the case of data set 2, data set 3, and data set 4, the presented work shows the higher convergence and hence becomes more robust for classification.

Analysis of classifiers: proposed vs. conventional models

This evaluation is undergone by comparing the proposed work to standard classifiers such as NB, BN, RF, c4.5, PART,



(a) (b) (c) (d)

Figure 6: Convergence analysis: (a) data set 1, (b) data set 2, (c) data set 3 and (d) data set 4.

MLP, ANN, CNN, respectively [28] with the combination of the averaging and majority voting approach with the grouping of the classifier performance is evaluated for methods M1 [28] – NB + BN + RF + c4.5 + averaging, M2 [28] – NB + BN + RF + c4.5 + majority voting, M3 [28] – NB + BN + RF + PART + averaging, M4 [28] – NB + BN + RF + PART + majority voting, M5 [28] – NB + BNN + RF + MLP + averaging, M6 [28] – NB + BNN + RF + MLP + majority voting, M7 [31] – proposed feature + ANN + CNN, M8 – proposed ensemble classifier with S-CDF. This evaluation is done for four sets of data sets in terms of positive and negative measures. The resultant acquired is tabulated in Tables 6–9, respectively. The accuracy of the presented classifier (M8) is higher (=0.957152), and it is 17.11, 8.5, 17.11, 2.6, 17.11, 2.6, and 13.7% better than the M1, M2,M3, M4, M5, M6, and M7, respectively. Similar to this, all other positive measures such as sensitivity, specificity, and precision of the proposed work are higher, while compared to the existing works. The FNR of the presented work (M8=0.01) is lower, while compared to other approaches M1=0.071411, M2=0.015709, M3=0.1164, M4=0.015709, M5=0.133137, M6=0.011781, and M7=0.066325. Thus, as a whole, the presented classifier is much better for data set 1.

Alike this, the accuracy of the presented work is much better than the existing works for data set 2, data set 3, and data set 4, respectively. In the case of data set 4, the accuracy of the presented work is 0.960068 (highest),

Table 6: Analysis on classifiers: proposed vs. conventional models for data set 1.

Methods	Accuracy	Sensitivity	Specificity	Precision	Recall	FMS	TS	NPV	MCC	FNR	FDR	FOR	FPR
M1	0.79	0.89	0.67	0.75	0.90	0.82	0.71	0.86	0.14	0.07	0.21	0.10	0.30
M2	0.88	0.95	0.55	0.88	0.95	0.91	0.87	0.86	0.40	0.01	0.08	0.11	0.42
M3	0.79	0.85	0.71	0.79	0.85	0.82	0.71	0.79	0.06	0.11	0.17	0.17	0.26
M4	0.93	0.95	0.83	0.94	0.95	0.95	0.92	0.89	0.23	0.01	0.03	0.07	0.14
M5	0.79	0.84	0.75	0.77	0.84	0.80	0.68	0.82	0.09	0.13	0.20	0.14	0.22
M6	0.93	0.96	0.82	0.93	0.96	0.95	0.92	0.91	0.24	0.01	0.03	0.05	0.15
M7	0.83	0.90	0.56	0.86	0.90	0.88	0.80	0.69	0.01	0.06	0.11	0.28	0.41
M8 (proposed)	0.96	0.97	0.90	0.95	0.97	0.97	0.95	0.97	0.16	0.01	0.01	0.01	0.07

FMS, Fowlkes–Mallows index; TS, Threat score; NPV, Negative predictive value; MCC, Matthews correlation coefficient; FNR, False negative rate; FDR, False discovery rate; FOR, False omission rate; FPR, False positive rate.

Table 7: Analysis of classifiers: proposed vs. conventional models for data set 2.

Methods	Accuracy	Sensitivity	Specificity	Precision	Recall	FMS	TS	NPV	MCC	FNR	FDR	FOR	FPR
M1	0.68	0.79	0.51	0.71	0.79	0.75	0.61	0.62	0.26	0.18	0.26	0.35	0.45
M2	0.81	0.85	0.74	0.83	0.85	0.85	0.74	0.77	0.61	0.11	0.14	0.20	0.22
M3	0.60	0.67	0.48	0.64	0.67	0.66	0.5	0.52	0.16	0.30	0.32	0.45	0.48
M4	0.72	0.80	0.60	0.75	0.80	0.77	0.64	0.67	0.32	0.16	0.21	0.30	0.38
M5	0.64	0.72	0.51	0.69	0.72	0.71	0.56	0.55	0.24	0.24	0.27	0.42	0.45
M6	0.79	0.85	0.72	0.80	0.84	0.82	0.71	0.78	0.55	0.12	0.17	0.19	0.25
M7	0.76	0.87	0.65	0.71	0.87	0.78	0.66	0.84	0.01	0.09	0.25	0.12	0.32
M8 (proposed)	0.77	0.89	0.66	0.71	0.89	0.79	0.66	0.86	0.21	0.08	0.25	0.10	0.31

FMS, Fowlkes–Mallows index; TS, Threat score; NPV, Negative predictive value; MCC, Matthews correlation coefficient; FNR, False negative rate; FDR, False discovery rate; FOR, False omission rate; FPR, False positive rate.

Table 8: Analysis on classifiers: proposed vs. conventional models for data set 3.

Methods	Accuracy	Sensitivity	Specificity	Precision	Recall	FMS	TS	NPV	MCC	FNR	FDR	FOR	FPR
M1	0.89	0.91	0.80	0.90	0.91	0.91	0.86	0.83	0.14	0.05	0.07	0.14	0.17
M2	0.86	0.92	0.70	0.87	0.92	0.89	0.83	0.82	0.30	0.05	0.10	0.14	0.28
M3	0.85	0.90	0.72	0.86	0.90	0.88	0.81	0.81	0.13	0.06	0.11	0.16	0.25
M4	0.90	0.94	0.80	0.90	0.97	0.92	0.88	0.88	0.23	0.03	0.06	0.10	0.17
M5	0.87	0.92	0.75	0.88	0.91	0.89	0.83	0.83	0.13	0.06	0.09	0.14	0.22
M6	0.88	0.94	0.72	0.88	0.94	0.90	0.85	0.87	0.28	0.04	0.90	0.10	0.24
M7	0.82	0.94	0.65	0.80	0.94	0.86	0.77	0.90	0.01	0.03	0.18	0.06	0.32
M8 (proposed)	0.93	0.97	0.85	0.92	0.96	0.94	0.91	0.95	0.19	0.01	0.05	0.01	0.12

FMS, Fowlkes–Mallows index; TS, Threat score; NPV, Negative predictive value; MCC, Matthews correlation coefficient; FNR, False negative rate; FDR, False discovery rate; FOR, False omission rate; FPR, False positive rate.

Table 9: Analysis of classifiers: proposed vs. conventional models for data set 4.

Methods	Accuracy	Sensitivity	Specificity	Precision	Recall	FMS	TS	NPV	MCC	FNR	FDR	FOR	FPR
M1	0.85	0.94	0.52	0.86	0.94	0.90	0.83	0.78	0.47	0.03	0.11	0.18	0.44
M2	0.85	0.96	0.66	0.81	0.96	0.88	0.80	0.94	0.35	0.01	0.16	0.02	0.31
M3	0.85	0.92	0.75	0.82	0.92	0.87	0.79	0.89	0.26	0.05	0.14	0.08	0.22
M4	0.88	0.97	0.54	0.86	0.96	0.91	0.86	0.92	0.11	0.01	0.10	0.05	0.43
M5	0.85	0.92	0.54	0.88	0.92	0.90	0.83	0.68	0.47	0.05	0.09	0.29	0.43
M6	0.85	0.96	0.55	0.83	0.96	0.89	0.83	0.93	0.17	0.01	0.14	0.04	0.42
M7	0.91	0.96	0.66	0.90	0.96	0.94	0.90	0.89	0.01	0.01	0.06	0.08	0.31
M8 (proposed)	0.96	0.97	0.92	0.96	0.97	0.96	0.96	0.94	0.20	0.01	0.01	0.02	0.05

FMS, Fowlkes–Mallows index; TS, Threat score; NPV, Negative predictive value; MCC, Matthews correlation coefficient; FNR, False negative rate; FDR, False discovery rate; FOR, False omission rate; FPR, False positive rate.

whereas the accuracy of the existing works is M1= 0.847509, M2=0.847509, M3=0.847509, M4=0.870683, M5=0.847509, M6=0.847509, and M7=0.907099, respectively. Thus, from the overall evaluation, it is vivid that the presented classifier is suitable for heart disease prediction as it is more accurate than the existing models.

Conclusion

This paper had developed a novel heart disease prediction framework by following three major phases, namely proposed feature extraction, dimensionality reduction, and proposed ensemble-based classification. Initially in the proposed feature extraction phase, the relevant features such as statistical (central tendency) and higher-order statistical features (degree of dispersion and qualitative evaluation) are extracted. However, in this scenario, the "curse of dimensionality" seems to be the greatest issue, such that there was a necessity to lessen the higher dimensionality features into lower ones. Hence, the PCA-based feature reduction approach was deployed here. These dimensionality-reduced features were fed as input to the proposed ensemble classifier that encompasses SVM, RF, and KNN. Subsequently, the classified outcome from all these three classifiers was fed as input to the optimized NN, where the training is carried out using a new S-CDF optimization algorithm, an improved sea lion algorithm via tuning the optimal weights. The final outcome from optimized NN gives more accurate resultants. The accuracy of the presented classifier (M8) is higher (=0.957152), and it is 17.11, 8.5, 17.11, 2.6, 17.11, 2.6, and 13.7% better than the M1, M2,M3, M4, M5, M6, and M7, respectively. Similar to this, all other positive measures such as sensitivity, specificity, and precision of the proposed work are higher, while compared to the existing works. In future, the dimensionality of data can be extended for inspecting the performance of heart disease prediction system and optimized strategies are applied to enhance the prediction rate.

Research funding: None declared.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication. The authors declare that they have no conflict of interest.

Ethical approval: The conducted research is not related to either human or animal use.

References

1. Bojja GR, Ofori M, Liu J, Ambati LS. Early public outlook on the coronavirus disease (COVID-19): a social media study; 2020.
2. Mienye ID, Sun Y, Wang Z. Improved sparse autoencoder based artificial neural network approach for prediction of heart disease. Inf Med Unlocked 2020;18:100307.
3. Al-Makhadmeh Z, Tolba A. Utilizing IoT wearable medical device for heart disease prediction using higher order Boltzmann model: a classification approach. Measurement 2019;147: 106815.
4. Rodríguez J, Prieto S, Lópe LJR. A novel heart rate attractor for the prediction of cardiovascular disease. Inf Med Unlocked 2019;15: 100174.
5. Baggen VJM, Venema E, Živná R, Bosch AE, Roos-Hesselink JW. Development and validation of a risk prediction model in patients with adult congenital heart disease. Int J Cardiol 2019;276: 87–92.
6. Ong KL, Chung RWS, Hui N, Festin K, Kristenson M. Usefulness of certain protein biomarkers for prediction of coronary heart disease. Am J Cardiol 2020;125:542–8.
7. Patel J, Rifai MA, Scheuner MT, Shea S, Evoy JWM. Basic vs. more complex definitions of family history in the prediction of coronary heart disease: the multi-ethnic study of atherosclerosis. Mayo Clin Proc 2018;93:1213–23.
8. Rajakumar BR, George A. On hybridizing fuzzy min max neural network and firefly algorithm for automated heart disease diagnosis. In: 2013 fourth international conference on computing, communications and networking technologies(ICCCNT); Tiruchengode, India, IEEE 2013:1–5 pp.
9. Praveena MDA, Bharathi B. Cognitive learning based missing value computation in cardiovascular heart disease prediction data. Procedia Comput Sci 2019;165:742–50.
10. Beunza J-J, Puertas E, García-Ovejero E, Villalba G, Landecho MF. Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). J Biomed Inf 2019;97. 103257.
11. Amin MS, Chiam YK, Varathan KD. Identification of significant features and data mining techniques in predicting heart disease. Telematics Inf 2019;36:82–93.
12. Ahmed H, Younis EMG, Hendawi A, Ali AA. Heart disease identification from patients' social posts, machine learning solution on Spark. Future Generat Comput Syst 2020;111:714–22.
13. Bonacaro A, Morgan L. Simulated mindfulness meditation: a major breakthrough in the management of chronic pain; 2016.
14. Harel-Sterling L, Wang F, Cohen S, Liu A, Marelli A. Risk predictions in adult congenital heart disease patients with heart failure: a systematic review. J Am Coll Cardiol 2019;73:656.
15. Hamed MB, Farah A, Abdeljalil O, Garmazi S. Metabolic factors of coronary arteries restenosis formation and unfavourable outcomes prediction of stent angioplasty in patients with chronic coronary heart disease. Arch Cardiovasc Dis Suppl 2019;11:188–9.
16. Kinoshita T, Abe A, Yao S, Yano K, Ikeda T. Risk stratification with non-invasive techniques for prediction of cardiac mortality in patients with ischemic heart disease. J Electrocardiol 2019;53: e17–8.
17. Bossolasco M, Fenoglio LM. Yet another PECS usage: a continuous PECS block for anterior shoulder surgery. J Anaesthesiol Clin Pharmacol 2018;34:569.
18. Yang J, Xiao W, Lu H, Barnawi A. Wireless high-frequency NLOS monitoring system for heart disease combined with hospital and home. Future Generat Comput Syst 2019;110:772–80.
19. Samuel OW, Yang B, Geng Y, Asogbon MG, Li G. A new technique for the prediction of heart failure risk driven by hierarchical neighborhood component-based learning and adaptive multi-layer networks. Future Generat Comput Syst 2019;110:781–94.
20. Kinoshita T, Abe A, Yao S, Yano K, Ikeda T. Risk stratification with non-invasive techniques for prediction of cardiac mortality in patients with ischemic heart disease. J Electrocardiol 2018;51: 1179.
21. Wang Z, Wang B, Zhou Y, Li D, Yin Y. Weight-based multiple empirical kernel learning with neighbor discriminant constraint for heart failure mortality prediction. J Biomed Inf 2020;101:103340.
22. George A, Rajakumar BR. On hybridizing fuzzy min max neural network and firefly algorithm for automated heart disease diagnosis. In: Fourth international conference on computing, communications and networking technologies. Tiruchengode, India: IEEE; 2013:1–5 pp.
23. Singh G, Jain VK, Singh A. Adaptive network architecture and firefly algorithm for biogas heating model aided by photovoltaic thermal greenhouse system. Energy Environ 2018;29:1073–97.
24. Bojja GR, Ambati LS. A novel framework for crop pests and disease identification using social media and AI. In: Proceedings of the fifteenth midwest association for information systems conference. Des Moines, Iowa; 2020:28–9 pp.

25. Manassero A, Bossolasco M, Ugues S, Bailo C. An atypical case of two instances of mepivacaine toxicity. J Anaesthesiol Clin Pharmacol 2014;30:582.
26. Desogus M. The stochastic dynamics of business evaluations using Markov models. Int J Contemp Math Sci 2020;15:53–60.
27. Thangam T, Kazem HA, Muthuvel K. SFOA: Sun Flower Optimization Algorithm to Solve Optimal Power Flow J Comput Mech Power Syst Control 2019;2. Resbee Publishers.
28. Latha CBC, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inf Med Unlocked 2019;16:100203.
29. Mathan K, Kumar PM, Panchatcharam P, Manogaran G, Varadharajan R. A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease. Des Autom Embed Syst 2018;22:225–42.
30. Vijayashree J, Sultana HP. Heart disease classification using hybridized Ruzzo-Tompa memetic based deep trained Neocognitron neural network. Health Technol 2018;10:207–16.
31. Ali L, Rahman A, Khan A, Zhou M, Javeed A, Khan JA. An automated diagnostic system for heart disease prediction based on ${\chi^{2}}$ statistical model and optimally configured deep neural network. IEEE Access 2019;7:34938–45.
32. Javeed A, Zhou S, Yongjian L, Qasim I, Noor A, Nour R. An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. IEEE Access 2019;7:180235–43.
33. Ali L, Niamat A, Khan JA, Golilarz NA, Xingzhong X, Noor A, et al. An optimized stacked support vector machines based expert system for the effective prediction of heart failure. IEEE Access 2019;7:54007–14.
34. Maragatham G, Devi S. LSTM model for prediction of heart failure in big data. J Med Syst 2019;43:111.
35. Nourmohammadi-Khiarak J, Feizi-Derakhshi M-R, Behrouzi K, Mazaheri S, Zamani-Harghalani Y, Tayebi RM. New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection. Health Technol 2019;10:1–12.
36. Avci E. A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier. Expert Syst Appl 2009; 36:10618–26.
37. Masetic Z, Subasi A. Congestive heart failure detection using random forest classifier. Comput Methods Progr Biomed 2016; 130:54–64.
38. Jabbar MA, Deekshatulu BL, Chandra P. Classification of heart disease using K- nearest neighbor and genetic algorithm. Procedia Technol 2013;10:85–94.
39. Central tendency. Available from: https://en.wikipedia.org/wiki/Central_tendency [Accessed 11 May 2020].
40. Statistical dispersion. Available from: https://en.wikipedia.org/wiki/Statistical_dispersion [Accessed 11 May 2020].
41. Qualitative variatoin. Available from: https://en.wikipedia.org/wiki/Qualitative_variation [Accessed 11 May 2020].
42. Gárate-Escamila AK, Hassani AHE, Andrès E. Classification models for heart disease prediction using feature selection and PCA. Inf Med Unlocked 2020;19:100330.
43. Masadeh R, Mahafzah BA, Sharieh A. Sea Lion optimization algorithm. Int J Adv Comput Sci Appl 2019;10:388–95.