International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7, Issue-6, March 2019

Optimized Feature Extraction and Hybrid Classification Model for Heart Disease and Breast Cancer Prediction

Sireesha Moturi, S. N. Tirumala Rao, Srikanth Vemuru

Abstract: ***Nowadays, diseases like heart disease and breast cancer are rising day by day due to the life style, hereditary and so on. Particularly, heart disease has become more common these days, i.e. life of people is at risk. Each and every individual has various values for cholesterol, Blood pressure, pulse rate and so on. However, the prediction of heart disease with data mining classification is not up to the mark. Hence, this paper intends to propose a new disease prediction model with advanced and modified classification technique. The proposed prediction model includes three phases: Coalesce rule generation, Optimized feature extraction and hybrid classification. Initially, the given big data is preprocessed by transforming the data to some other form, from which the rules are generated. The optimal features are selected by a new introduced algorithm namely, New levy Update based Dragonfly Algorithm (NL-DA). Finally, the selected optimal features are subjected to the new hybrid classifier, hybridization of Support vector Machine (SVM) and Deep belief Network (DBN), so that the accurate disease prediction is worked out. The proposed NL-DA model is compared to other conventional methods in terms of Accuracy, Specificity, Sensitivity, Precision, F1Score, Negative Predictive Value (NPV) and Matthews Correlation Coefficient (MCC), False negative rate (FNR), False positive rate (FPR) and False Discovery Rate (FDR), and proven the betterments of proposed work.***

Index Terms: ***Disease Prediction; Data Mining; Feature Extraction; Optimization; Classification.***

I. INTRODUCTION

Generally, data mining is considered as an extensive area that combines approaches from various fields such as, statistics, database systems, pattern recognition, machine learning, and artificial intelligence for investigating huge quantities of data [1] [9]. There have been several data mining approaches entrenched in these fields to execute dissimilar data investigation tasks. Data mining is regarded as the procedure of extorting concealed information from huge quantities of original data [10]. Data mining has been described as the non-trifling removal of formerly unidentified, inherent and probably functional information from data. It is one of the tasks which deal with the progression of detecting information from the database. Data Mining is exploited to find information out of data and then, it is depicted in an effortlessly understandable form. It is the process of investigating huge quantities of data which is collected in a routine manner [11] [12] [13]. Data mining is most constructive in an investigative scrutiny because of non-trifling information in huge quantities of data. It is an accommodating endeavour of humans and computers. Preeminent consequences are attained by matching the information of human authorities in defining difficulties and purposes with the exploration abilities of computers. There are two major objectives of data mining approaches such as prediction and description. In the prediction model, it includes several variables in the data set in order to detect the unidentified or potential values of other variables of interest. Moreover, description models are used for evaluating the patterns with respect to the information deduced by humans. The detection of disease plays a significant role in data mining. These types of patterns are exploited for the clinical diagnosis for widely disseminated raw medical data. These data should be gathered in a controlled type. This collected information can be combined to form a hospital information system [16].

Data mining technology offers a user oriented approach for exploiting hidden patterns in the data. The prediction of heart disease is modeled to prop up clinicians in their diagnosis to predict the disease [17] [18] [14] [15], which offers [19] [20] an effective way to retrieve the information obscured in the data. Similarly, Breast cancer is also a common disease in women, nowadays. The mammography is considered as the traditional approach for predicting the breast cancer. However, the radiologists demonstrate substantial unpredictability depending on the interpretation of a mammogram [21] [22]. The main objective is to handle cases for which cancer has not persisted along with the case for which cancer has persisted at a definite time [23] [24].

This paper proposes a new disease prediction model that comprises of three phases: Coalese rule generation, Optimized feature extraction and hybrid classification. At first, the given big data is preprocessed via data transformation, from which the rules are generated. Then, the optimal features are selected by a new introduced NL-DA algorithm. At last, the selected optimal features are subjected to the hybrid classifier, hybridization SVM and DBN; thereby the prediction model grants the classified outcome even more accurately.

Revised Manuscript Received on 30 March 2019.
***** Correspondence Author
Sireesha Moturi*, Department of Computer Science and Engineering, KLEF, Vijayawada, India.
Sireesha Moturi, Department of Computer Science and Engineering, Narasaraopeta Engineering College, Narasaraopet, India.
Dr.S.N.Tirumala Rao, of Computer Science and Engineering, Narasaraopeta Engineering College, Narasaraopet, India.
Dr. Srikanth Vemuru, Department of Computer Science and Engineering, KLEF, Vijayawada, India.

The comparison is made to some other conventional methods with respect to positive and negative measures.

The rest of the paper is arranged as follows: Section II reviews the literature work. Section III explains the working strategy of proposed disease prediction model. Section IV explains the optimized feature extraction phase. Section V describes the classification process using the proposed hybrid classifier. Section VI discusses the obtained results and Section VII concludes the paper.

II. LITEARTATURE REVIEW

A. Related Works

In 2017, Malarvizhi and Gandhi [1] have implemented a machine learning algorithm for predicting heart disease. Health monitoring system was considered to be as the imperative application facilitated by the IoT. Wearable sensor devices deployed in IoT health monitoring system was normally used for producing a massive amount of data on a permanent basis. Here, Tier-1 was used to gather information from IoT wearable sensor devices, Tier-2 was exploited to store huge quantity of information from wearable IoT sensor and Tier-3 was used for developing the logistic approach for predicting the heart diseases. From the experimental results it was clear that, this approach attains better detection rate when compared with the other approaches.

In 2017, Wang *et al.* [2] have developed an ensemble based approach for detecting breast cancer. Here, it uses an SVM based ensemble learning algorithm based on WAUCE model in order to minimize the variance and accuracy in terms of detection process. Three datasets were used to prove the betterment of proposed model.

In 2017, Nilashi *et al.* [3] have implemented a machine learning approach for predicting diseases. Here, awareness based system with Classification and Regression Trees (CART) for detecting several disorders and prediction techniques were employed. This approach was examined based on the outcomes on different datasets. From the simulation results it was apparent that, this approach attains better detection rate in case of diseases when compared with the other classical approaches.

In 2016, Lafta *et al.* [4] have implemented a machine learning ensemble method for prediction of heart disease. Here, an efficient medical recommendation system with Fast Fourier Transformation (FFT) was deployed for predicting the short term diseases. It also provides several approximate suggestions with respect to the medical examination of medical information. The input data of sliding windows with respect to the time series data of patients were disintegrated for the extraction of frequency related data. Moreover, a Bagging based ensemble approach was exploited to detect the circumstance of the patients for generating the final

suggestion. From the experimental results it was clear that, this approach attains accurate and reliable recommendations when compared with the other traditional approaches.

In 2016, Tripoliti *et al.* [5] have developed a machine learning approach for diagnosing heart disease. Here, it describes about the state-of-the-art of the machine 29 learning methodologies in order to detect the disorders in heart. Moreover, several models which predict the evaluation of subtypes were deployed to assess the rigorousness of heart failure and also for predicting purpose. From the experimental results it was evident that, this approach achieves better detection rate when compared with other conventional approaches.

In 2014, Theodora *et al.* [6] have developed a supervised technique for predicting heart disease. In this model, five machine learning algorithms such as SVM, AdaBoost using trees as the weak learner, and so on for predicting the disease. Here, each approach was trained depending on the training set and then it was tested on the test set. Thus, these approaches reveals about the maximum attainable prediction accuracy. From the experimental results it was clear that, this approach achieves considerable amount of potential savings when compared with other existing approaches.

In 2018, Alwidian *et al.* [7] have implemented an association rule for diagnosing heart disease. Here, an improved pruning and prediction approach depending on the arithmetical measures was deployed to produce an efficient association rule for improving the effectiveness of the AC classifiers. Moreover, WCBA was exploited for the classification breast cancer illustrations based on the premise material professionals from King Hussein Cancer Center (KHCC). Additionally, WCBA was deployed for producing effective rules with effective attributes for detecting breast cancer. From the experimental outcome it was clear that, this approach attains better accuracy when compared with the other standard approaches.

In 2018, Vazifehdan et al. [8] have established an improved approach for predicting breast cancer. Here, a combined imputation method was proposed based on the correlation for enhancing the detection rate for predicting breast cancer. First, the misplaced values in the discrete fields were imputed based on the Bayesian network with respect to the segmented discrete and numerical data series. Then, both the uninterrupted misplaced values and the accuracy of imputation was improved based on the Tensor factorization which was composed of several subsets with previous stage and misplaced numerical subsets. From the experimental results it was clear that, this approach attains better prediction rate when compared with the other approaches

TABLE I. FEATURES AND CHALLENGES OF DETECTING DISEASES IN HEART OR BRAIN USING VARIOUS MACHINE LEARNING APPROACHES

Author [Citation]	Adopted Methodology	Features	Challenges
Malarvizhi and Gandhi [1]	Three-tier architecture	• Could process huge data • High scalability and availability.	• Low quantitative accuracy. • Produces irrelevancy in the features.
Wang *et al.* [2]	WAUCE	• Achieve high generalization ability. • Diagnose breast cancer in an accurate manner.	• Issues in setting the model parameter. • Potential risk of over-fitting.

Nilashi *et al.* [3]	Fuzzy logic	• It can be used to filter out the potential noise. • High accuracy	• Analysis of disease is proven to be incomplete. • Loss in relevant information.
Lafta *et al.* [4]	FFT based machine learning ensemble model	• Provide accurate results. • Improves the overall accuracy of the prediction model.	• Complexity in selecting the features. • Susceptible to inter-expert variability.
Tripoliti *et al.* [5]	Machine learning algorithm	• Reduces the associated medical cost. • Assess the severity of heart failure.	• Less proximity limit. • Decrease in classification performance.
Dai *et al.* [6]	SVM and AdaBoost	• This approach can easily scale to a large number of monitored patients. • Prediction variables were calculated.	• Less accuracy. • Issues occur during classification.
Alwidian *et al.* [7]	Association rules	• Allows enhancement of accuracy level. • Reduces the fear of possibility.	• Lots of effort should be made to predict the disorders. • High operation expense.
Vazifehdan *et al.* [8]	Tensor factorization	• Increase the accuracy of prediction. • Inputes the missing values based on the single sample.	• Lack of attention to dependency between attributes. • Computationally expensive task.

B. Review

The literature has come out with several machine learning techniques for the detection of diseases in heart and brain, which is summarized in Table I. However, they require more improvements because of lack of several features in the prediction process. Three-tier architecture [1] was used to process huge data volume of sensor data. But, it produces irrelevancy in the features. WAUCE [2] was used to attain high generalization ability for the prediction of breast cancer in an accurate manner. However, there arise several potential risks related to over-fitting. Fuzzy logic [3] was exploited to filter out the potential noise with high accuracy. But, the analysis of disease was proven to be incomplete. FFT based machine learning ensemble model [4] was used for providing accurate and reliable recommendations to the patients. But, it was s Susceptible to inter-expert variability. Machine learning algorithm [5] was deployed to minimize the associated medical cost for assessing the severity of heart failure. But, there arise decrease in the classification performance. SVM and AdaBoost [6] were easily used to scale large number of monitored patients for the estimation of predicted variables. But, there also arise various issues during classification. Association rules [7] were exploited to permit the enhancement of accuracy level by minimizing the fear of possibility of recurrence of the disease. However, lots of effort should be made to predict the disorders. Tensor factorization [8] was used to impute the misplaced values based on the single sample and thus, increases the accuracy of prediction. But, it was considered to be computationally expensive task. These limitations have highly motivated to develop more improved prediction models.

III. Proposed Model for Predicting Heart Disease and Breast Cancer

The proposed prediction model includes three phases: (i) Rule generation (ii) Feature extraction (iii) Classification. Let the original data be D that includes records and labels. Table I shows the example original data. Here, a_1, a_2 and a_3 are the records about the patients (for instance, age, sex, etc) and the labels be either 1 or 2 (1-affected, 2-no disease). Fig 1 shows the art of proposed disease prediction model.

TABLE I. Original data

a_1	a_2	a_3	label
2	0.5	2	1
1	0.7	10	1
4	0.8	12	2
3	0.9	14	2

Before starting the process, the original data is transformed to some other form, which is simply termed as pre-processing. Since the data originally exist will be in non-uniform, the preprocessing step uniforms the corresponding record in a specific format. For this, initially the number of range, N_{range} is assigned (say N_{range}=3), i.e, the range of data must be from 1to 3. Then, the formulation of normalize parameter is done for all the records as per Eq. (1). Here, $\max(a_1)$ indicates the maximum value of a_1 and $\min(a_1)$ indicates the minimum value of a_1. Subsequently, find the number of levels, N_{levels} as per Eq. (2). Here, as the N_{range} is 3, the N_{levels} will be 2 as per Eq. (2). The evaluation of levels is given in Eq. (3) and (4).

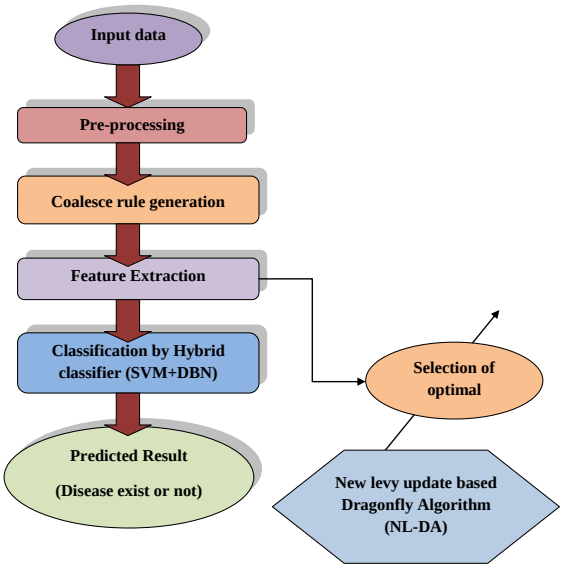


Fig. 1. Art of proposed disease prediction model

$$\Delta_1 = \frac{\max(a_1) - \min(a_1)}{N_{range}} \quad (1)$$

$$N_{levels} = N_{range} - 1 \quad (2)$$

$$Level_1 = \min(a_1) + \Delta_1 \quad (3)$$

$$Level_2 = \min(a_1) + [Level\ no \times \Delta_1] \quad (4)$$

Then, the normalization process of a_1' is evaluated on the basis of these levels ($Level_1$ and $Level_2$), which is given in Eq. (5). Eq. (5) is the normalized data of a_1, and similarly evaluate for the remaining records and attains a_2' and a_3'. The normalized data would be as per Table II.

TABLE II. NORMALIZATION PROCESS OF DATA

a_1'	a_2'	a_3'	label
2	1	1	1
1	2	3	1
3	3	3	2
3	3	3	2

$$a_1' = \begin{cases} 1 & if\ \ a_1 < Level_1 \\ 2\ , if\ \ Level_1 \le a_1 \le Level_2 \\ : \\ : \\ N, if\ \ a_1 \ge Level_{N_{level}} \end{cases} \quad (5)$$

Subsequently, evaluate the pre-processed data a_1'' , a_2'' and a_3'' is evaluated as per the formulation that given in Eq. (6). Table III shows the preprocessed data. This preprocessed or normalized data is given for rule mining.

$$a_2'' = a_2' + \max(a_1'') \quad (6)$$

TABLE III. PREPROCESSED DATA

a_1''	a_2''	a_3''	label
2	1	1	1
1	2	3	1
3	3	3	2
3	3	3	2

A. Coalesce rule generation

In this Algorithm [36], the corresponding transactional database is denoted to the Binary Table. The frequent patterns and the support count of every frequent pattern are directly attained from the Binary Table. At first, this database is denoted as the form of 0's and 1's where 1 indicates the presence of an item and 0 specifies the absence of an item.

The Table includes the gathering of rows and columns. In the transactional database, each row indicates one transaction and each column indicates one item in the given transactional database. Table IV shows the Binary Table that generated.

TABLE IV. GENERATED BINARY TABLE

TId	ITEMS	A	B	C	D	E	F	G
T1	ABCEF	1	1	1	0	1	1	0
T2	ACG	1	0	1	0	0	0	1
T3	E	0	0	0	0	1	0	0
T4	ACDEG	1	0	1	1	1	0	1
T5	ACEG	1	0	1	0	1	0	1
T6	E	0	0	0	0	1	0	0
T7	ABCEF	1	1	1	0	1	1	0
T8	ACD	1	0	1	1	0	0	0
T9	ACEG	1	0	1	0	1	0	1
T10	ACEG	1	0	1	0	1	0	1

Consider TD as the given Transactional Database, $TR = TR_1, TR_2, ...TR_n$ is the set of transactions, $IT = IT_1, IT_2, ...IT_s$ is the set of items. In Eq. (7), $z'' = 1,2,....n$ and $z' = 1,2,....s$.

$$TD = TR_{z'z''} = \begin{cases} TR_{z'z''} = 1, if\ IT_{z''} \in IT_{z'} \\ TR_{z'z''} = 0, if\ IT_{z''} \notin IT_{z'} \end{cases} \quad (7)$$

Once the binary table is represented, the evaluation of support count or frequency of each item is done. The support count of one item set is attained by adding the non zero element in each column and it is shown in Table IV. The support count evaluation is given in Eq. (8).

$$SC = \sum_{z''=1}^{n} TR_{z'z''}, where\ z' = 1,2,....s\ and\ TR_{z''z'} > 0 \quad (8)$$

TABLE V. ATTAINMENT OF SUPPORT COUNT

TId	ITEMS	A	B	C	D	E	F	G
T1	ABCEF	1	1	1	0	1	1	0
T2	ACG	1	0	1	0	0	0	1
T3	E	0	0	0	0	1	0	0
T4	ACDEG	1	0	1	1	1	0	1
T5	ACEG	1	0	1	0	1	0	1
T6	E	0	0	0	0	1	0	0
T7	ABCEF	1	1	1	0	1	1	0
T8	ACD	1	0	1	1	0	0	0
T9	ACEG	1	0	1	0	1	0	1
T10	ACEG	1	0	1	0	1	0	1
SC		**8**	**2**	**8**	**2**	**8**	**2**	**5**

Then, the sorting processes is carried out and to identify the frequency of every transaction, add a new column termed as transaction Frequency column (TF) column to the binary table. Finally, the given TB with TF column is indicated by binary table and it is shown in Table VI.

TABLE VI. TRANSACTIONAL DATABASE WITH TRANSACTION FREQUENCY

B	D	F	G	A	C	E	TF
1	0	1	0	1	1	1	2
0	0	0	1	1	1	0	1
0	0	0	0	0	0	1	2
0	1	0	1	1	1	1	1
0	0	0	1	1	1	1	3
0	1	0	0	1	1	0	1

By following this algorithm, the proposed model evaluates the final rules from the preprocessed data along with their lables, and it is $\begin{bmatrix} A_2 & A_4 & A_7 \\ A_1 & A_5 & A_9 \\ A_3 & A_6 & A_9 \\ A_3 & A_6 & A_9 \end{bmatrix}$. The output frequency rules in respective labels is finalized, which is given below in Table VII.

TABLE VII. OUTPUT FREQUENCY RULES

Output rules	Label 1	Label 2
(A_2, A_4)	1	0
(A_1, A_5)	1	0
(A_2, A_7)	1	0
(A_3, A_6, A_9)	0	1
(A_3, A_6)	0	2

Finally the rules are separated based on the labels, which is nothing but the extracted features. The exemplary table with the extracted features is given in Table VIII.

TABLE VIII. EXTRACTED FEATURES

Sl.no	Label 1	Label 2
1	(A_2, A_4)	(A_3, A_6)
2	(A_1, A_5)	(A_3, A_6, A_9)
3	(A_2, A_7)	

II. OPTIMIZED FEATURE EXTRACTION VIA PROPOSED NL-DA

The extracted features include number of features for both label 1 and label 2. However, it is not so efficient to give all the features for classification. In order to get the accurate classification result, it is aimed to select the optimal features. To select the optimal features from the extracted features, this paper introduces a new NL-DA, which is explained in the subsequent section. The solution encoding of the proposed work is given in Fig 3. Here, $F_{L1}, F_{L2}, ...F_{LN}$ indicates the rules of Label 1 and $F_{L'1}, F_{L'2}, ...F_{L'N}$ specifies the rules of Label 2.

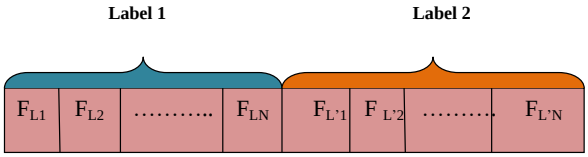


Fig. 2. Solution Encoding

The main condition behind this optimal feature selection is given in Eq. (9) and the objective function of the proposed work is defined in Eq. (10). In Eq. (10), F_1 indicates the fitness of rule 1 of label 1, F_2 indicates the fitness of rule 2 of label 1, F_3 indicates the fitness of rule 1 of label 2. All the three fitness are given in Eq. (11), (12) and (13).

$$No.of\ optimal\ rules\ for\ each\ label < Total\ no.of\ rules\ of\ the\ label \quad (9)$$

$$OB = \frac{F_1 + F_2 + F_3}{3} \quad (10)$$

$$F_1 = \frac{Frequency\ of\ rule\ 1\ of\ label\ 1}{Sum\ of\ rules\ in\ label\ 1} \times \Pr ecision \quad (11)$$

$$F_2 = \frac{Frequency\ of\ rule\ 2\ of\ label\ 1}{Sum\ of\ rules\ in\ label\ 1} \times \Pr ecision \quad (12)$$

$$F_3 = \frac{Frequency\ of\ rule\ 1\ of\ label\ 2}{Sum\ of\ rules\ in\ label\ 2} \times \Pr ecision \quad (13)$$

A. Conventional Dragonfly Algorithm

DA[33] algorithm is the new renowned optimization algorithm that is based on the behaviour of dragonflies. In fact, the major objective of any group is survival. Hence, the individual dragonfly in the group must follow two behaviors: (i) Attraction and (ii) Distraction. This means, the dragonfly must be attractive towards the food source and they should be distracted from external enemies. Five factors are there in position updating, which may work based on the mentioned behaviours. The factors are as follows: Control cohesion, Alignment, Separation, Attraction and Distraction.

Step 1: Separation: This indicates the avoidance of static collision of DA from other neighborhood DAs. In dr^{th} dragonfly separation, NH_{dr} makes its neighbors, and the formulation of separation factor is given in Eq. (14), in which the position of current dragonfly is specified as X, X_t specifies the position of t^{th} closer dragonfly and nm denotes the count of neighboring dragonflies.

$$NH_d = \sum_{t=1}^{nm} (X - X_t) \quad (14)$$

Step 2: Alignment: This specifies the velocity matching of every dragonfly to remaining neighbour dragonflies. Eq. (15) shows the evaluation of alignment, in which VE_t

indicates the velocity of t^{th} neighboring dragonfly.

$$AM_r = \frac{\sum_{t=1}^{nm} VE_t}{nm} \quad (15)$$

Step 3: Cohesion: Cohesion indicates the preference of dragonfly towards the nearby mass center. Eq. (16) shows the formulation of cohesion.

$$CH_{dr} = \frac{\sum_{t=1}^{nm} X_t}{nm} - X \quad (16)$$

Step 4: Attraction: Eq. (17) shows the evaluation of attraction towards the food source, in which *Food* indicates the food position.

$$AT_{dr} = Food - X \quad (17)$$

Step 5: Distraction: The behavior of distraction from the external enemy is modeled as given in Eq. (18). Here, *ene* specifies the position of enemy.

$$EN_{dr} = ene + X \quad (18)$$

All the behaviors of dragonfly are purely based on the afore mentioned patterns. Two vectors termed, ΔX step vector and position X is concerned here to update the position of dragonfly. ΔX specifies the direction movement of dragonfly, and it is determined in Eq. (19), in which nl indicates the weight of separation factor, NH_{dr} specifies the separation of dr^{th} dragonfly, al refers to the alignment weight, AM_{dr} indicates the alignment of dr^{th} dragonfly, the cohesion weight is specified by cn, CH_{dr} indicates the cohesion of dr^{th} dragonfly, the fiood factor is denoted as $food$, AT_{dr} specifies the food source of dr^{th} dragonfly, en specifies the enemy factor, EN_{dr} indicates the enemy's position of dr^{th} dragonfly, δ refers to the inertia weight and it referes to the iteration counter.

$$\Delta X_{it+1} = \left(nlNH_{dr} + alAM_{dr} + cnCH_{dr} + foodAT_d + enEN_d\right) + \delta.\Delta X_{it} \quad (19)$$

With the utilization of afore-mentioned factor, both exploitative and explorative counts of dragon behavior can be defined. Eq. (20) shows the evaluation of position vector, in which it indicates the current iteration.

$$X_{it+1} = X_{it} + \Delta X_{it+1} \quad (20)$$

While exploration, the DAs are assigned with high alignment and minimal cohesion and while exploitation, the dragonflies are allocated with less alignment and high cohesion. If there has no neighbor solutions, the dragonfly do a random walk (levy flight). At this point, the dragonflies position gets updated using Eq. (21), where it indicates the current iteration, dp refers to the position vector dimensions, rn_1 and rn_2 specifies the two random numbers in [0,1], $\wp$ indicates the constant value. Positions are updated as per Eq. (19), Eq. (20) and Eq. (21) respectively. To update` X and ΔX , the dragonfly neighbor is determined by formulating the Euclidean distance. Algorithm 1 shows the pseudo code of conventional DA.

$$X_{it+1} = X_{it} + levy(ds) \times X_{it} \quad (21)$$

$$levy(dp) = 0.01 \times \frac{rn_1 \times \phi}{|rn_2|^{\frac{1}{ds}}} \quad (22)$$

$$\phi = \left(\frac{\psi(1+\wp) \times \sin\left(\frac{\pi\wp}{2}\right)}{\psi\left(\frac{1+\wp}{2}\right) \times \wp \times 2^{\left(\frac{\wp-1}{2}\right)}} \right) \quad (23)$$

$$\psi(z) = (z-1)! \quad (24)$$

Algorithm 1: DA Algorithm

Population initialization, $X_{\bar{s}} : \bar{s} = 1,.....r_S$

Step vector initialization ΔX

while the end criteria is not satisfied
- Access the objective value of all dragonflies
- Perform food source updating and enemy updating
- Update $nl, al, cn, food, en$
- Access the primitive behavior of all NH, AM, CH, AT, EN using Eq. (15)to Eq.(18)
- Neighboring radius update
- **If** the dragonfly has only one neighboring dragonfly
  - Update velocity vector value by Eq. (19)
  - Update position vector value by Eq. (20)
- **else**
  - Update position vector value by Eq. (21)
- **End if**
- Return the new position

End while

Improved Dragonfly-NL-DA

Even though the conventional DA solve the Multiobjective problem, the algorithm has the ability to solve only the continuous problems. Hence, this paper introduces an improved version of DA algorithm to rectify all the problems of conventional one. The pseudo code of proposed algorithm is given in Algorithm 2. The improvement is made in the position update (levy flight) of conventional DA. Here, the new introduced evaluation of levy is given in Eq. (25), where it indicates the current iteration, and Max^{it} indicates the maximum iteration. The flowchart of proposed NL-DA is given in Fig 4.

$$levy(dp) = \frac{it}{2Max^{it}} \times \frac{rn_1 \times \phi}{|rn_2|^{\frac{1}{ds}}} \quad (25)$$

Algorithm 2: Proposed Algorithm for optimal feature selection

Population initialization, $X_{\bar{s}} : \bar{s} = 1,.....r_S$

Step vector initialization ΔX

while the end criteria is not satisfied
- Access the objective value of all dragonflies
- Perform food source updating and enemy updating
- Update $nl, al, cn, food, en$

Access the primitive behavior of all NH, AM, CH, AT, EN using Eq. (15)to Eq.(18)
Neighboring radius update
If the dragonfly has only one neighboring dragonfly
Velocity vector update by Eq. (19)
Position vector update by Eq. (20)
else
Position vector update by Eq. (21) with new Levy evaluation that given in Eq. (25).
End if
Return the new position
End while

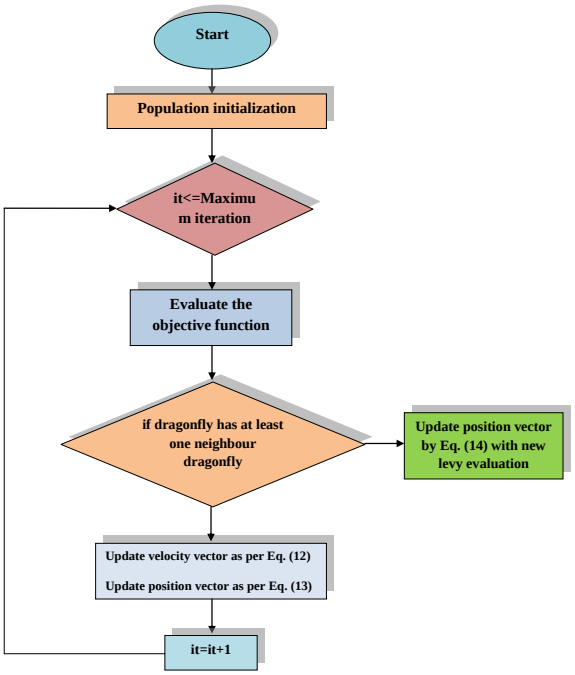


Fig. 3. Flowchart of proposed algorithm

III. NOVEL HYBRID CLASSIFICATION FOR DISEASE PREDICTION

The selected optimal features are subjected for classification process, where it classifies the data (affected or not). For this purpose, his paper introduces a new hybrid classifiers (hybridization of SVM and DBN). This is because, the class results from SVM is actually not so accurate in terms of classification. So the resultant class from SVM is again given as the input to DBN and it gives the classified output.

A. Support Vector Machine

In general, SVM [35] is termed as the two-class classifier, which generates a hyperplane for classifying two data segments. As per the statistical theory, the significant objective of SVM is the identification of optimal (maximize) margin. The respective optimal margin is determined by the minimum distance among hyperplane and any of the sample points. The subset of data point that defines the position or location of hyperplane is named as support vectors. The hyperplane of two-class linearly separable issue in an n-dimensional feature space is as per Eq. (26).

$$K(a) = VT^T Z + g = 0 \qquad (26)$$

where VT specifies the normal vector and g indicates the distance from hyperplane to origin. $K(a)$ is learned by training data set, $\{a_i, c_i\}; i = 1,....h$, where $a_i \in \Re^n$ and $c_i \in \{+1, -1\}$. The training samples are precisely classified by $K(a)$ with the given constraints: if $c_i = +1, K(a) \geq 1$ and if $c_i = -1, K(a) \geq -1$. The point which makes $K(x) = +1$ or-1 is termed as support vector. Eq. (27) defines the distance of perpendicular from a specific point a to hyperplane.

$$r = \frac{VT^T Z_n + g}{\| VT \|} = \frac{y_n (VT^T Z_n + g)}{\| VT \|} \qquad (27)$$

The major objective of SVM is the finding of a hyperplane to maximize the distance among hyperplane and the points of training data that are closest to the hyperplane. The corresponding issue is then altered into the given equivalent convex quadratic issue, and it is given in Eq. (28).

$$\min_{EV,g} \frac{1}{2} \| VT \|^2 \qquad (28)$$

So that $c_i (VT^T a_i + g)^3 1, i = 1,2,3.....N$. With the aid of lagrange multipliers, Eq. (28) is defined as in Eq. (29).

$$\max_h \sum_{i=1}^{N} h_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} c_i \times c_j \times h_i \times h_j \times \langle a_i, a_j \rangle c \qquad (29)$$

$$\text{so that } \sum_{i=1}^{N} h_i c_i = 0 \qquad (30)$$

where $h_i, i = 1,2....N$. The major issue is indicated by $VT = \sum_{i=1}^{N} \alpha_i c_i a_i$ and $0 = \sum_{i=1}^{N} \alpha_i c_i$. Hence, having the observed lagrange multipliers α, the definition of w and g is happened. In general, data could be overlapped, and hence obtaining of accurate training data division is a challenging aspect, and that could lead to least generalization. The resultant class from SVM is given to DBN.

B. Deep Belief Network

The class labels from SVM are considered as the features, and it is the input to DBN [34] classifier. DBN approach is a renowned intelligent model, which is developed in the year 1986. In fact, the model comprises of many layers and every layer includes visible neurons. Furthermore, there exists a deep relation with hidden as well as input neurons. This corresponding neuron model determines the accurate output for the input.

An Eq. (31) show the output and the possibility in sigmoid-shaped function are given in Eq. (32) where t^P indicates the pseudo-temperature. The deterministic model of stochastic approach is given in Eq. (33).

$$\overline{O}_q(\zeta) = \frac{1}{1 + e^{\frac{-\zeta}{t^P}}} \qquad (31)$$

$$\overline{PS} = \begin{Bmatrix} 1 & with\ 1-\overline{O}_q(\zeta) \\ 0 & with\ \overline{O}_q(\zeta) \end{Bmatrix} \quad (32)$$

$$\lim_{t^P \to 0^+} \overline{O}_q(\zeta) = \lim_{t^P \to 0^+} \frac{1}{1+e^{\frac{-\zeta}{t^P}}} = \begin{cases} 0 & for \quad \zeta < 0 \\ \frac{1}{2} & for \quad \zeta = 0 \\ 1 & for \quad \zeta > 0 \end{cases} \quad (33)$$

The diagrammatic illustration of DBN model is given in Fig. 4, where the process of feature extraction takes place through a set of (Restricted Boltzman Machine (RBM) layers and the classification process is carried out via Multi-Layer Perceptron (MLP). The model of energy of Boltzmann machine for the creation of neuron or binary state bi is given in Eq. (33), where $W_{a,l}$ indicates the weights among neurons and θ_a indicates the biases.

$$\Delta ER(bi_a) = \sum_l bi_a W_{a,l} + \theta_a \quad (34)$$

The progression of energy in terms of joint composition of visible as well as hidden neurons (x, y) is defined in Eq. (35), Eq. (36) and Eq. (37). In this, x_a indicates either the binary or neuron state of a visible unit, B_l indicates the binary state of l hidden unit, and k_a.

$$ER(x, y) = \sum_{(a,l)} W_{a,l} x_a y_l - \sum_a k_a x_a - \sum_l B_l y_a \quad (35)$$

$$\Delta ER(x_a, \bar{y}) = \sum_l W_{al} y_l + k_a \quad (36)$$

$$\Delta ER(\vec{x}, y_a) = \sum_l W_{al} x_a + B_l \quad (37)$$

The input data's possibility dissemination is encoded into weight (parameters), which is spread as RBM's learning pattern. RBM training can attain the distributed possibilities, and the consequent weight assignment is defined by Eq. (38).

$$\hat{W}_{(\hat{M})} = \max_{\hat{W}} \prod_{\vec{x} \in N} c(\vec{x}) \quad (38)$$

For the visible and hidden vectors pair $(\vec{x}, \vec{hi})$, the possibility assigned RBM approach is given in Eq. (39), where PT^{FN} specifies the partition function as in Eq. (40).

$$c(\vec{x}, \vec{hi}) = \frac{1}{PT^{FN}} e^{-ER(\vec{x}, \vec{y})} \quad (39)$$

$$PT^{FN} = \sum_{\vec{x}, \vec{y}} e^{-ER(\vec{x}, \vec{y})} \quad (40)$$

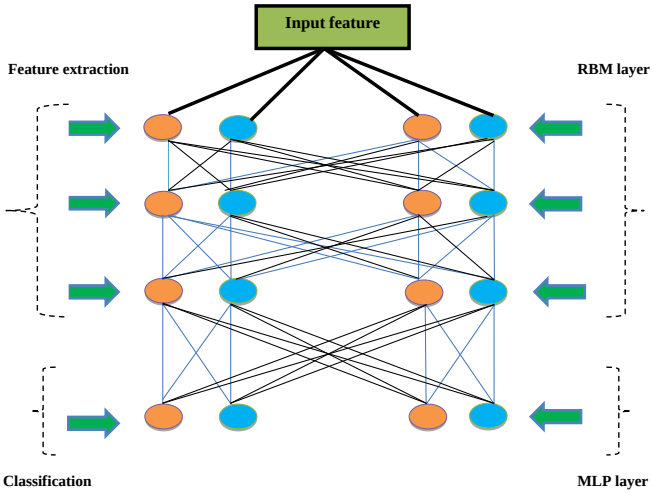


Fig. 4. Architecture of DBN model

Contrastive Divergence (CD) is used by DBN approach for learning purpose, and the achievement of sampling expectations is a complex process under distribution. The CD algorithm steps is as follows:

Step 1: Select the x training samples and brace it into visible neurons.

Step 2: Asses the possibility of hidden neurons c_y by identifying the product of $\hat{W}$ weight matrix and visible vector x as $c_y = \sigma(x.\hat{W})$ based on Eq. (41).

$$c(\vec{y}_l \to 1 | \vec{x}) = \sigma\left(B_l + \sum_a x_a W_{a,l} \right) \quad (41)$$

Step 3: Validate the y hidden states from c_y probabilities.

Step 4: Formulate the x exterior product of vectors and c_y that is measured as positive gradient $\phi^+ = x.c_y^{t^P}$.

Step 5: Validate the reconstruction of x' visible states from y hidden states as Eq. (42). Furthermore, it is needed to formuale y' hidden states from the reconstruction of x'.

$$c(\vec{x}_l \to 1 | \vec{y}) = \sigma\left(k_a + \sum_a x_l W_{a,l} \right) \quad (42)$$

Step 6: Asses the x' and y' 's exterior product, by it as negative gradient $\phi^- = x'.y'^{t^P}$ as given in Eq. (43), where η specifies the learning rate.

$$\Delta \hat{W} = \eta(\phi^+ - \phi^-) \quad (43)$$

Step 8: Eq. (44) defines the update of weight with new values.

$$W'_{a,l} = \Delta W_{a,l} + W_{a,l} \quad (44)$$

Before doing the learning process of MLP algorithm, consider $\left(T^{\hat{M}}, L^{\hat{M}}\right)$ training patterns, where $\hat{M}$ specifies the count of training patterns, $1 \le \hat{M} \le \overline{O}$, $T^{\hat{M}}$ and $L^{\hat{M}}$ refers to the input vector with desired output vectors, respectively. Eq. (45) determines the neuron error in l of output layer.

$$e_l^{\hat{M}} = T^{\hat{M}} - L^{\hat{M}} \tag{45}$$

In this, Eq. (20) shows the squared error of $\hat{M}$ pattern followed by MSE (Mean Squared Error), and it is given in Eq. (46).

$$SE_{\hat{M}}^{mean} = \frac{1}{\tilde{o}_y} \sum_{l=1}^{\tilde{o}_y} \left(e_l^{\hat{M}}\right)^2 = \frac{1}{\tilde{o}_y} \sum_{l=1}^{\tilde{o}_y} \left(T^{\hat{M}} - L^{\hat{M}}\right)^2 \tag{46}$$

$$SE_{avg} = \frac{1}{\tilde{P}} SE_{\hat{M}}^{mean} \tag{47}$$

DBN process is as follows:

Step 1: Initializes the DBN model with biases, weights, and other related parameters, which are randomly chosen.

Step 2: At first, the RBM model initializes its progress through the input data and provides the unsupervised learning.

Step 3: In this step, the input to the subsequent layer process with potential sampling. Furthermore, it follows the unsupervised learning.

Step 4: The afore-mentioned steps are continued for the respective number of layers. Thus, the pre-training stage by RBM is progressed till it reaches the MLP layer.

Step 5: MLP phase indicates the attained learning by supervised format, which continuous till it attains the target.

Finally, the classifier results the predicted results in an accurate manner.

IV. Results and Discussion

A. Simulation Setup

The proposed disease prediction is implemented in Java. Four datasets were used to implement this work and they were Cleveland, heartdata, hungerian and Wisconsin. In this, Wisconsin was collected from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic), datasets like cleveland and hungerian were downloaded from https://archive.ics.uci.edu/ml/datasets/heart+Disease, and heartdata was downloaded from http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart/. The proposed model was compared over other conventional methods like Classification based on Association +Classification based on Predictive Association rule (CBA+CPAR) [28], Multiple kernel Learning + Adaptive Neuro-Fuzzy Inference System (MKL+ANFIS) [29], Random Forest + Evolutionary Algorithm (RF+EA) [30], Weighted Classification based on Association rule algorithm (WCBA) [31] and Interquartile Range +K-Nearest neighbour+ Particle Swarm Optimization (IQR+KNN+PSO) [32] with respect to positive nad negative measures.

B. Overall Performance of Proposed Model under Dataset 1(Cleveland)

The performance analysis of proposed model over other conventional works for dataset 1 is given in Table IX. Here, the analysis is carried out under both the positive and negative measures, where the proposed method proves its superiority over others. More particularly, it is observed that the specificity of proposed model is 99.63%, 51.48%, 20.89%, 19.23% and 3.35% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. The sensitivity of proposed method is 79.39%, 74.05%, 41.62%, 30.81%, and 6.59% better from MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. Similarly, the accuracy of proposed method is 61.91%, 34.81%, 30.28%, 29.55%, and 23.94% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. The precison of proposed method is also very high, which is 30.46% better than CBA+CPAR, 24% better than MKL+ANFIS, 13.01% better than RF+EA, 10.49% better than WCBA and 8.38% better than IQR+KNN+PSO models. While analysing the negative measures, it is observed that the proposed method attains less FNR, FPR, and FDR over other models. In this, the FNR of proposed method is 91.16%, 86.29%, 80.28%, 69.48% and 65.49% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively.

TABLE IX. Overall Performance of Proposed and Conventional Methods for DATASET 1

Measures	CBA+CPAR [32]	MKL+ANFIS [31]	RF+EA [30]	WCBA [29]	IQR+KNN+PSO [28]	NL-DA
Specificity	0.00294118	0.38611111	0.629577	0.64285714	0.76923077	0.79591837
Sensitivity	0.17647059	0.22222222	0.5	0.59259259	0.91304348	0.85652174
Accuracy	0.31868132	0.54545455	0.583333	0.58947368	0.63636364	0.83673469
Precesion	0.58823529	0.64285714	0.735849	0.75714286	0.775	0.84594595
FPR	0.82352941	0.77777778	0.5	0.40740741	0.08695652	0.04347826
FNR	0.79705882	0.51388889	0.357143	0.23076923	0.20408163	0.07042254
NPV	0.17647059	0.22222222	0.492593	0.5	0.81304348	0.85652174
FDR	0.41176471	0.35714286	0.264151	0.14285714	0.125	0.05405405
F1_Score	0.18421053	0.64220183	0.642857	0.66666667	0.76470588	0.89189189
MCC	0.01002761	0.03035958	0.09126	0.14285714	0.35063594	0.56854915

C. Overall Performance of Proposed Model under Dataset 2(Heartdata)

The analysis of proposed method for dataset 2 is given in Table X. In this, it is evident that the proposed method is very much better with high accuracy rate, high specificity, sensitivity and so on over other conventional models. In this, the specificity of proposed method is high, which is 97.91%, 51.90%, 35.97%, 28.06% and 27.36% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. Further, the sensitivity of proposed method is 43.52% better from RF+EA, 10.17% better than WCBA and 10.17% better than IQR+KNN+PSO models. The precision of proposed model is maximum when compared to other conventional methods, and it is 40.08%, 38.84%, 27.49%, 14.60% and 2.47% superior to CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. The accuracy of proposed model is high in terms of disease prediction, which is 41.80%, 39.79%, 37.16%, 32.10% and 15.84% superior to CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. Similarly, the FNR of proposed work is less when compared to other models, and it is 91.90%, 86.65%, 85.39%, 82.27% and 44.93% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively.

TABLE X. OVERALL PERFORMANCE OF PROPOSED AND CONVENTIONAL METHODS FOR DATASET 2

Measures	CBA+CPAR [32]	MKL+ANFIS [31]	RF+EA [30]	WCBA [29]	IQR+KNN+PSO [28]	NL-DA
Specificity	0.01842105	0.42564103	0.56666667	0.63670886	0.642857	0.88505747
Sensitivity	0.14473684	0.46428571	0.5	0.88529412	0.88529412	0.97540984
Accuracy	0.5	0.51724138	0.5398773	0.58333333	0.723022	0.85918367
Precesion	0.53125	0.54225352	0.64285714	0.75714286	0.864706	0.88666667
FPR	0.75526316	0.53571429	0.5	0.02459016	0.02459	0.01470588
FNR	0.78157895	0.47435897	0.43333333	0.35714286	0.114943	0.06329114
NPV	0.14473684	0.46428571	0.5	0.87540984	0.87541	0.88529412
FDR	0.46875	0.45774648	0.35714286	0.14285714	0.035294	0.01333333
F1_Score	0.20809249	0.5483871	0.64285714	0.67248908	0.680498	0.96103896
MCC	0.03110083	0.04436556	0.14285714	0.17529827	0.539679	0.91960949

D. Overall Performance of Proposed Model under Dataset 3(Hungarien)

The performance rate of proposed work is compared over other models for dataset 3 and it is summarized in Table XI. Here, it is evident that the performance of proposed method is really effective than the conventional models in term so both positive and negative measures. In this, the proposed model's specificity is 90.44%, 96.64%, 24.58%, 23.41%, and 1.60% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. The sensitivity of proposed method is 97.33%, 70.32%, 44.11%, 9.63% and 2.94% superior to CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. The accuracy of proposed method is high, which is 61.26%, 45.54%, 32.21%, 26.12% and 25.51% superior to CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. Similarly, the precision of developed method is high when compared to other models, and it is 48.99%, 40.08%, 26.66%, 11.85% and 0.87% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. The negative measures are also proving the betterment of proposed model by attaining less value.

TABLE XI. OVERALL PERFORMANCE OF PROPOSED AND CONVENTIONAL METHODS FOR DATASET 3

Measures	CBA+CPAR [32]	MKL+ANFIS [31]	RF+EA [30]	WCBA [29]	IQR+KNN+PSO [28]	NL-DA
Specificity	0.08148148	0.02857143	0.642857	0.65283	0.83870968	0.852380952
Sensitivity	0.02380952	0.26548673	0.5	0.808511	0.86842105	0.894736842
Accuracy	0.33333333	0.46857143	0.583333	0.635762	0.6409396	0.860544218
Precesion	0.32786885	0.38518519	0.471429	0.566667	0.63722628	0.642857143
FPR	0.77619048	0.73451327	0.5	0.191489	0.03157895	0.005263158
FNR	0.78148148	0.72857143	0.357143	0.16129	0.04761905	0.047169811
NPV	0.02380952	0.26548673	0.5	0.708511	0.86842105	0.894736842
FDR	0.67213115	0.61481481	0.428571	0.357143	0.33333333	0.262773723
F1_Score	0.03603604	0.12698413	0.487805	0.527919	0.64285714	0.83127572
MCC	0.06382761	0.06401844	0.091548	0.118684	0.14285714	0.732825996

E. Overall Performance of Proposed Model under Dataset 4(Wisconsin)

The performance of proposed method over other conventional methods under dataset 4 is given in Table XII. From the summarized value, it is evident that the proposed model attains effective disease prediction, and the specificity of proposed method is 85.47%, 27.48%, 24.11%, 19.06% and 17.81% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. The sensitivity of proposed method is 92.96%, 54.99%, 43.74%, 11.01% and 2.19% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively.

The accuracy of proposed work is high, and it is 52.48%, 30.76%, 28.80%, 19.03% and 10.02% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. Similarly, the precision of proposed work is greater than the conventional models, which is 78.55% better than CBA+CPAR, 52.51% better than MKL+ANFIS, 51.71% better than RF+EA, 19.92% better than WCBA and 1.31%better than IQR+KNN+PSO. The negative measures also proving the betterments of proposed model. Altogether, the proposed work is very effective in predicting the disease.

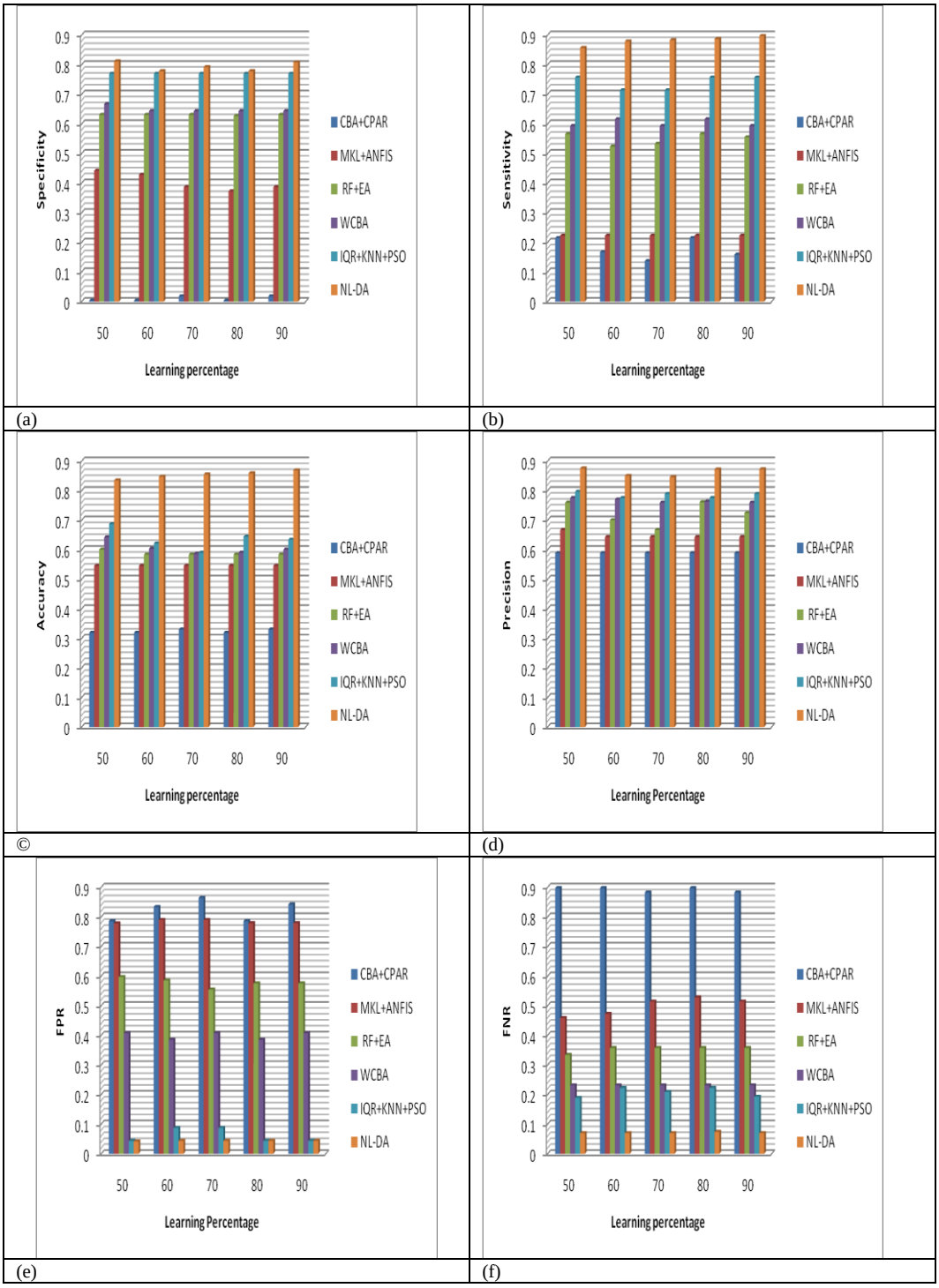
TABLE XII. OVERALL PERFORMANCE OF PROPOSED AND CONVENTIONAL METHODS FOR DATASET 4

Measures	CBA+CPAR [32]	MKL+ANFIS [31]	RF+EA [30]	WCBA [29]	IQR+KNN+PSO [28]	NL-DA
Specificity	0.11962617	0.597248	0.625	0.66666667	0.67695853	0.82366412
Sensitivity	0.0625	0.4	0.5	0.79090909	0.86935933	0.88885794
Accuracy	0.41176471	0.6	0.616984	0.70157068	0.77966102	0.86655113
Precesion	0.17857143	0.395327	0.401993	0.66666667	0.82156863	0.83251534
FPR	0.7375	0.6	0.5	0.20909091	0.03064067	0.01114206
FNR	0.78037383	0.375	0.333333	0.30275229	0.07633588	0.02304147
NPV	0.0625	0.3	0.5	0.79090909	0.86935933	0.88885794
FDR	0.72142857	0.598007	0.504673	0.33333333	0.07843137	0.06748466
F1_Score	0.27777778	0.354717	0.560185	0.65736434	0.66666667	0.79790026
MCC	0.02696487	0.166667	0.245776	0.35416588	0.41721476	0.71787924

F. Performance Analysis by varying Learning Percentage

This section explains the performance of proposed work over other conventional methods in terms of varied learning percentage to 50%, 60%, 70%, 80% and 90%, respectively. Further, the analysis is made for all the used datasets (1,2, 3 and 4) with respect to both positive and negative measures. Here, the graphical representations also show the performance of proposed NL-DA (before feature extraction).The analytical results for dataset 1 under various learning percentage is given in Fig (5). Here, Fig 5 (a) shows the specificity of proposed model over other conventional methods. the graph shows that for 50% of learning, the performance of proposed method in terms of specificity is 5.47%, 21.69%, 28.66%, 45.56%, and 99.63% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. For 60% of learning, the proposed method is 1.11%, 20.98%, 23.34%, 81.81% and 99.62% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively with high specificity. For 70% of learning, the developed method has high specificity, which is 2.91%, 23.14%, 25.55% and 51.22% superior to CBA+CPAR, MKL+ANFIS , RF+EA and WCBA, respectively. The sensitivity of proposed model under dataset 1 is given in Fig 5(b). Here, for 50% of learning, the proposed method is 13.20%, 44.52%, 51.46%, 74.05% and 74.97% better from CBA+CPAR, MKL+ANFIS, RF+EA, WCBA, and IQR+KNN+PSO, respectively with high sensitivity. For 60% of learning, the proposed method attains high sensitivity, which is 23.77%, 48.93%, 65.74%, 74.82% and 84.54% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. The accuracy measure is also analysed to prove the superior performance of proposed methods, and also succeeded in that (Fig 5 (c)). Here, it is evident that the accuracy of proposed method in disease prediction is very high over other models, and for 50% of learning, the proposed method is 21.55%, 29.97%, 39.09%, 53% and 61.81% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. For 60% of learning, the developed method has attained greatest accuracy rate, which is 36.37%, 40.41%, 45.18%, 55.27% and 62.37% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. The proposed method for 70% percentage is 44.98%, 45.91%, 46.51%, 56.68% and 61.42% better from CBA+CPAR, MKL+ANFIS, RF+EA, WCBA, and IQR+KNN+PSO, respectively.

Similarly, Fig 5(e) shows the FPR of proposed model and conventional models. Here, for 50% of learning, the proposed method has attained less FPR, which is 4.42%, 89.80%, 93.03%, 94.65% and 94.71% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. The proposed method for 60% of learning is 49.92%, 88.67%, 92.55%, 94.48% and 94.77% superior to CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. For 70% of learning, the proposed method is 50.03%, 89.33%, 92.16%, 94.49% and 94.96% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. Thus, the performance of proposed model is proven over other models in terms of disease prediction.


Specificity
Learning percentage
Sensitivity
Learning percentage
Accuracy
Learning percentage
Precision
Learning Percentage
FPR
Learning Percentage
FNR
Learning percentage
CBA+CPAR
MKL+ANFIS
RF+EA
WCBA
IQR+KNN+PSO
NL-DA
50
60
70
80
90
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

(a) (b)

© (d)

(e) (f)

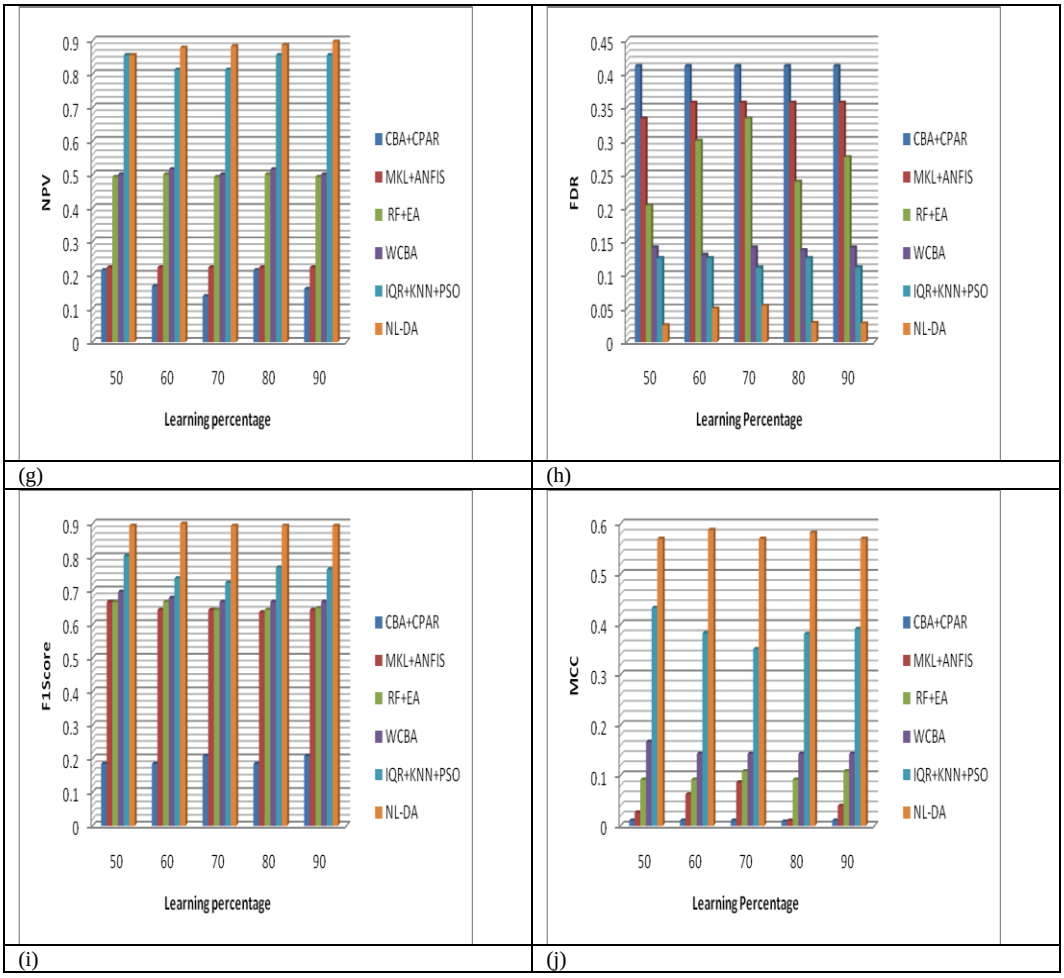
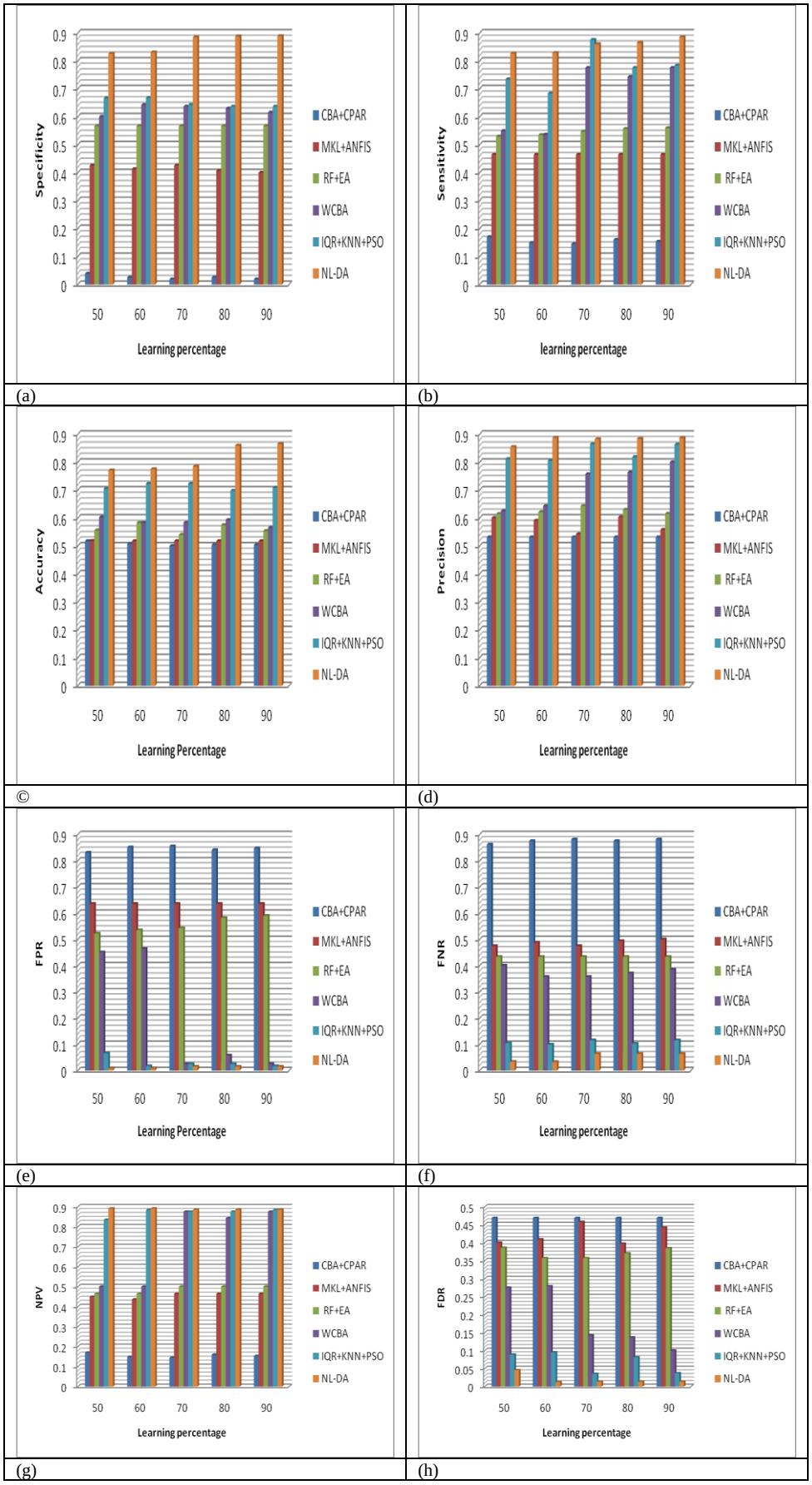


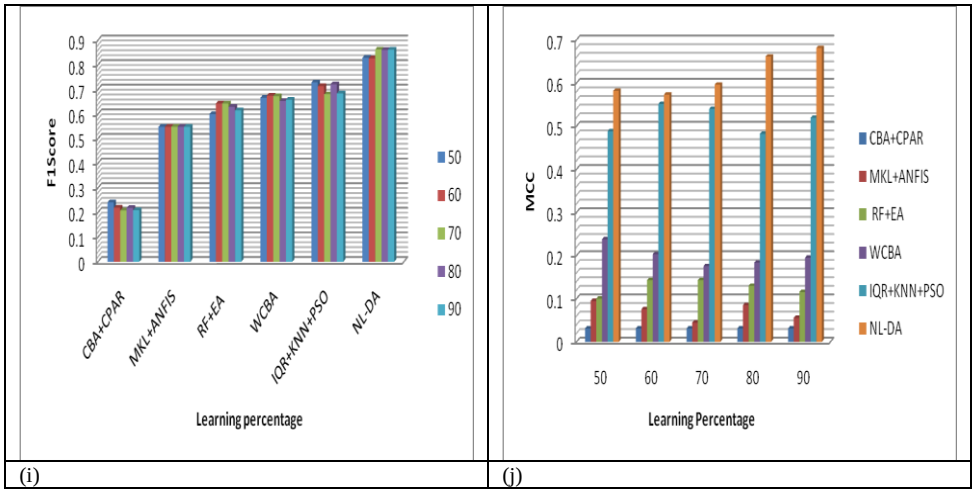
Fig. 5. Performance of proposed model over other conventional models by varying the learning percentage (Dataset 1) (a) Specificity (b) Sensitivity (c) Accuracy (d) Precision (e) FPR (f) FNR (g) NPV (h) FDR (i) F_1Score (j) MCC

Fig 6 shows the analytical results under dataset 2 by varying the learning percentage. Here, specificity of proposed method is given in Fig 6 (a). For 50% of learning, the proposed method is 23.87%, 37.63%, 45.73%, 94.02% and 95.37% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively with high specificity. For 60% of learning, the proposed method is 24.44% better from CBA+CPAR, 29.26% better from MKL+ANFIS, 46.64% better than RF+EA, 56.32% better from WCBA and 96.99% better from IQR+KNN+PSO. For 70% of learning, the proposed method is 37.67%, 39%, 56.18%, 51.90% and 97.91% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. The sensitivity analysis is shown in Fig 6 (b). In this, for 50% of learning, the proposed method is 12.55%, 50.52%, 56.10%, 43.83% and 79.52% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. For 60% of learning, the proposed method attains high sensitivity, and it is 21.23%, 54.62%, 55.01%, 43.98% and 82.06% superior to CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. For 70% of learning, the proposed method is 1.77%, 10.89%, 57.33%, 46% and 83.16% superior to CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively with high sensitivity.

The accuracy analysis is given in Fig 6 (c), from which the efficiency of proposed method is proven in terms of disease prediction. In this, for 50% of learning, the proposed method is 9.27%, 27.87%, 38.67%, 48.65% and 48.94% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. For 60% of learning, the proposed method is 7.18%, 32.85%, 32.96%, 49.83%, and52.76% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively with high accuracy rate. Similarly, the accuracy rate of proposed method for 70% learning is 8.51%, 34.49%, 45.32%, 51.68% and 56.91% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. Similarly, all the remaining learning percentage has also shown the betterment of proposed model with high accuracy rate. The negative measures in Fig 6 (e) shows the betterments of proposed work. More particularly while analysing FPR, it is proved that the proposed method for 50% of learning is 87.50%, 98.18%, 98.43% and 98.71% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. For 60% of learning, the FPR of proposed method is low, which is 50%, 98.23%, 98.46% and 98.71% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively.

For 70% of learning, the proposed method is 40.19% better than CBA+CPAR and MKL+ANFIS, 97.28% better than RF+EA and 97.68% better than WCBA, and 98.28% better than IQR+KNN+PSO with less FPR.



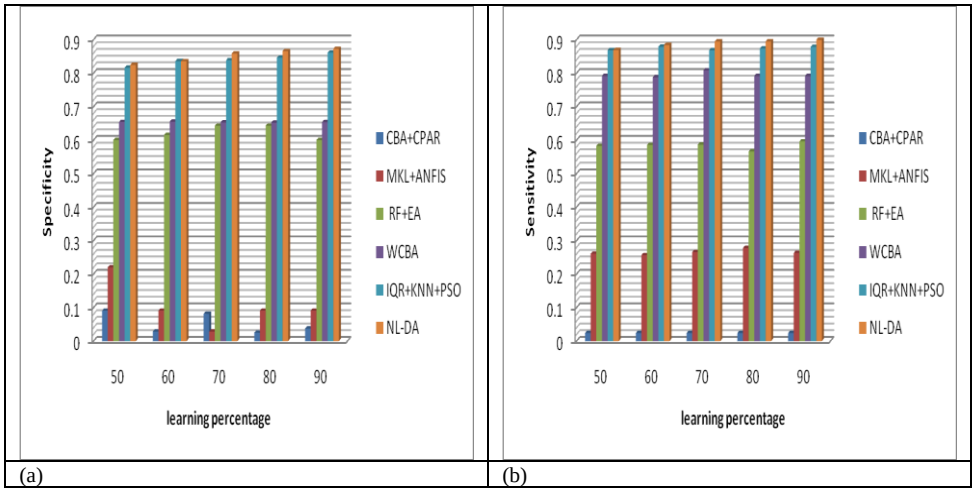


(i) (j)

Fig. 6. Performance of proposed model over other conventional models by varying the learning percentage (Dataset 2) (a) Specificity (b) Sensitivity (c) Accuracy (d) Precision (e) FPR (f) FNR (g) NPV (h) FDR (i) F_1Score (j) MCC

The performance analysis of proposed and conventional methods for dataset 3 is given in Fig 7. From the graphs, it is observed that the specificity of proposed model (in Fig 7 (a)) is high than the conventional models. For 50% of learning, the proposed method is 1.04%, 26.06%, 37.44%, 73.36% and 88.99% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively with high specificity. For 60% of learning, the proposed method attains high specificity, and it is 27.50%, 35.79%, 89.14% and 96.58% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. For 70% of learning, the proposed method is 2.38%, 31.54%, 33.58%, 96.67% and 90.51% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively.

Fig 7(b) shows the sensitivity of proposed model over other models for all the learning percentages. Here, almost in all learning percentage, the proposed model attains better sensitivity. For 50% of learning, the proposed method is 0.06%, 9.76%, 49.15%, 69.97%, and 97.25% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. For 60% of learning, the proposed method is 0.59%, 12.27%, 51.02%, 71% and 97.30% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. For 70% of learning, the proposed method attains high sensitivity, and it is 3.03%, 10.66%, 52.46%, 70.32% and 97.33% superior to CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. Similarly, the negative measure like FPR, FDr and FNR of proposed method is low than other conventional methods. Here, FPR of proposed model (Fig 7(e)) for 50% of learning is 66.66%, 94.94%, 97.89%, 98.57%, and 98.56% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. For 60% of learning, the FPR of proposed method is 25%, 92.56%, 96.84%, 97.87% and 97.85% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively with less value. For 70% of learning, FPR of proposed model is low, which is 19.99%, 86.80%, 94.94%, 96.56% and 96.61% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively.



(a) (b)

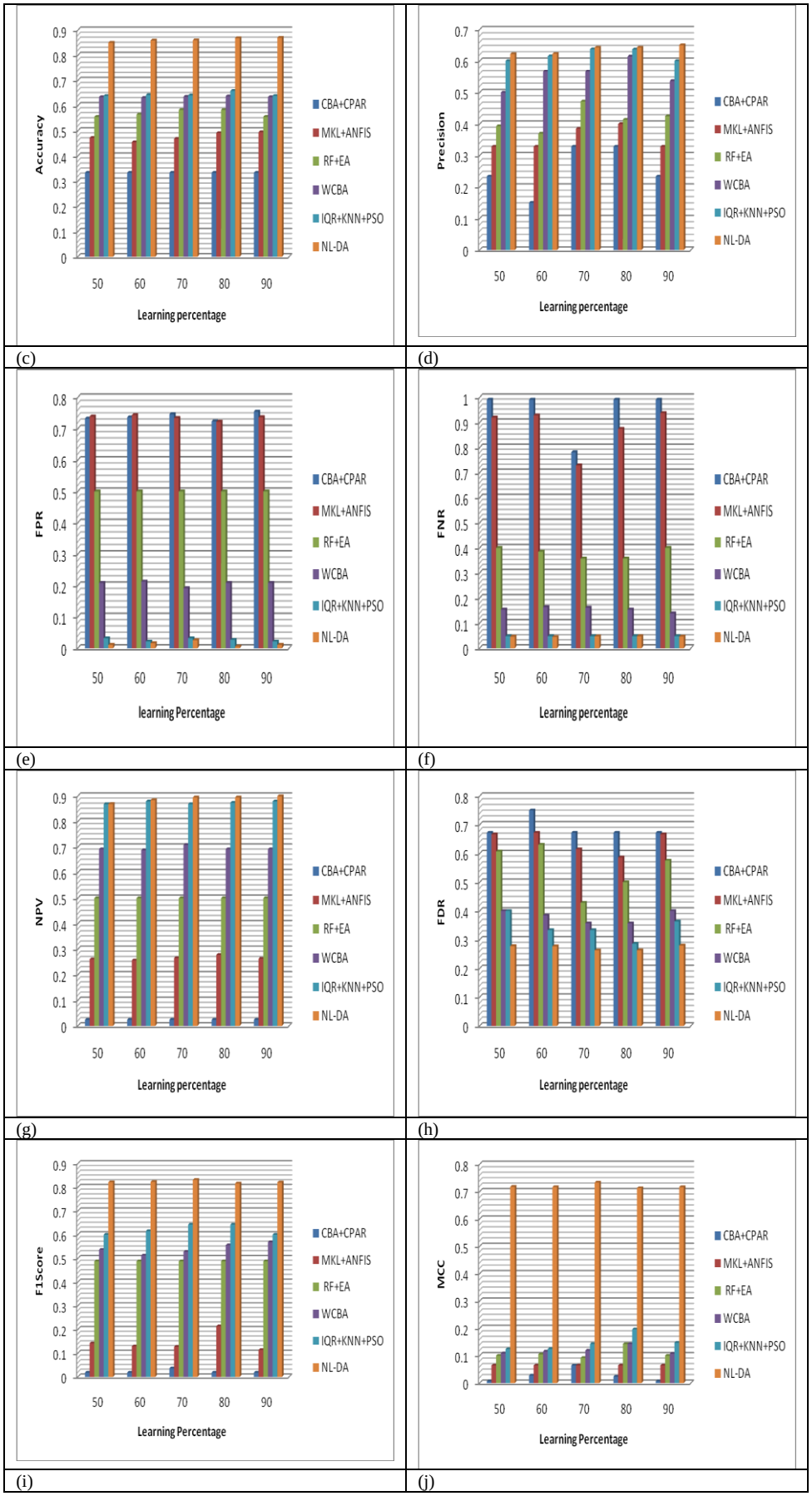
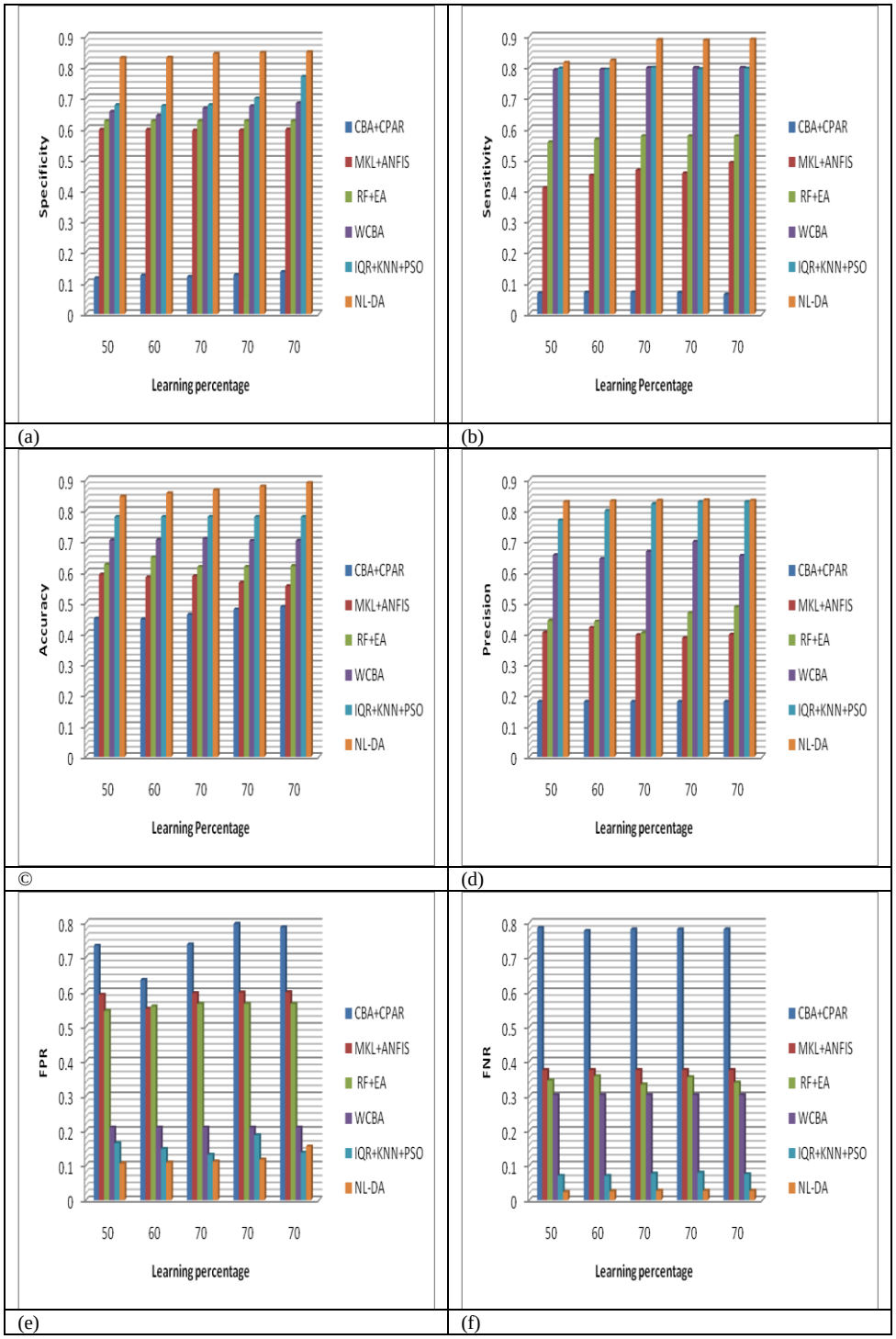


Fig. 7. Performance of proposed model over other conventional models by varying the learning percentage (Dataset 3) (a) Specificity (b) Sensitivity (c) Accuracy (d) Precision (e) FPR (f) FNR (g) NPV (h) FDR (i) F_1Score (j) MCC

The performance analysis of proposed and conventional models for dataset 4 is given in Fig 8. Here, the specificity analysis is given in Fig 8 (a), and for 50% of learning, the proposed model is 22.61%, 26.69%, 32.81%, 38.98% and 86.15% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. For 60% of learning, the proposed model is 23.13%, 29.19%, 32.88%, 39.06% and 85.03% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. For 70% of learning, the proposed method is 24.62%, 26.54%, 34.98%, 41.90% and 85.82% better from CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively with high specificity. Similarly, all other measures have proven the performance of proposed disease prediction model over other conventional methods.



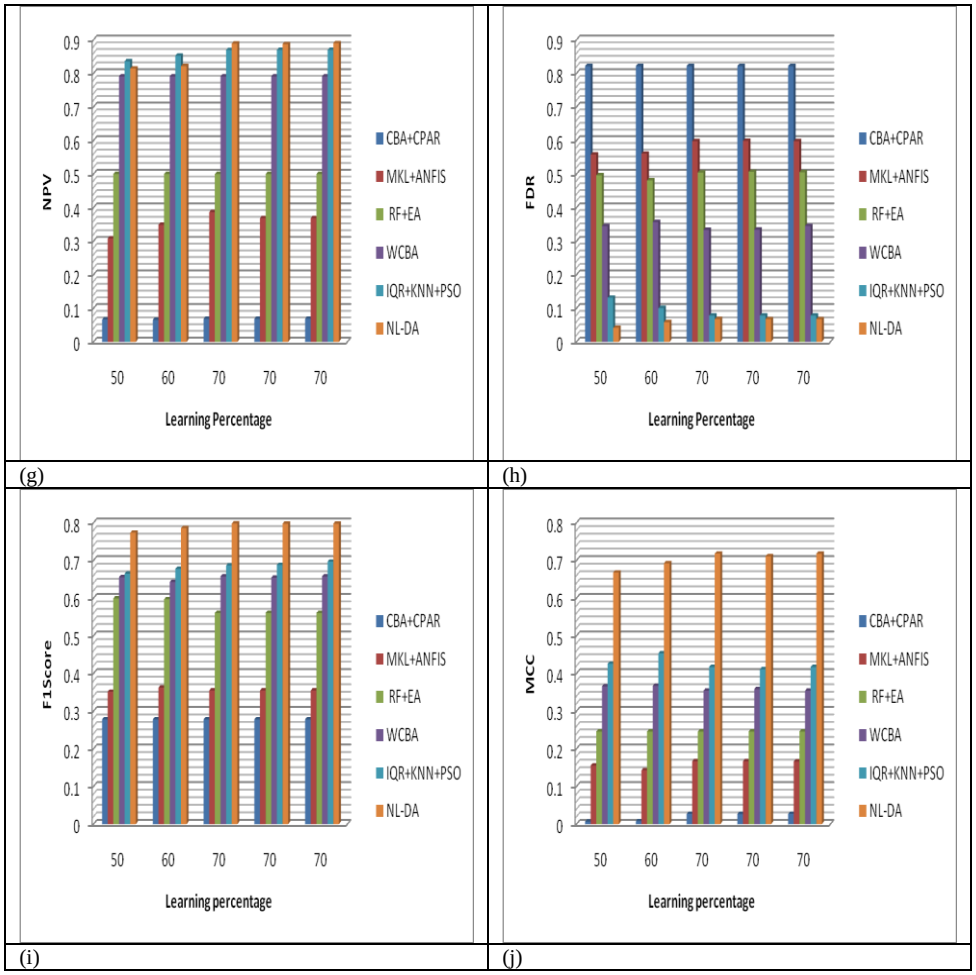


Fig. 8. Performance of proposed model over other conventional models by varying the learning percentage (Dataset 4) (a) Specificity (b) Sensitivity (c) Accuracy (d) Precision (e) FPR (f) FNR (g) NPV (h) FDR (i) F_1Score (j) MCC

G. Time Complexity

This section explains the time complexity of both the proposed and conventional models. Table XIII proves that the proposed model takes less time to complete the process, whereas the remaining conventional models require more time for the same almost for all the datasets. For Cleveland, the proposed model requires only 422 ms, whereas the conventional models require more time. Similarly, for all the datasets, the proposed model proves its betterments over other models with respect to less time.

TABLE XIII. TIME COMPLEXITY OF PROPOSED AND CONVENTIONAL MODELS

Dataset	CBA+CPAR [32]	MKL+ANFIS [31]	RF+EA [30]	WCBA [29]	IQR+KNN+PSO [28]	NL-DA
Cleveland (ms)	317407	2183	69046	24746	734	422
Heartdata (ms)	414215	626	37383	29980	648	441
Hungerian (ms)	338345	749	35971	23847	426	314
Wisconsin (ms)	3325791	1249	32826	16861	774	695

V. CONCLUSION

This paper has proposed a new prediction model that has included three phases: Coalesce rule generation, Optimized feature extraction and hybrid classification. At first, the input data was preprocessed via data transformation, from which the rules were generated. Further, the optimal features were chosen by a new algorithm NL-DA. Finally, the selected optimal features were given as the input to hybrid classifier (SVM+DBN), by which the prediction model was accurate. Theproposed NL-DA model was compared over conventional methods with respect to certain measures. From the results, it was observed that the specificity of proposed method was high, which is 37.67%, 39%, 56.18%, 51.90% and 97.91% better than CBA+CPAR, MKL+ANFIS , RF+EA , WCBA and IQR+KNN+PSO, respectively. Further, the sensitivity of proposed method was 77.05% better from RF+EA, 90.67% better than WCBA and 83.65% better than IQR+KNN+PSO models.

REFERENCES

1. Priyan Malarvizhi Kumar and Usha Devi Gandhi, "A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases," Computers and Electrical Engineering, pp. 1-14, 2017.

2. Haifeng Wang, Bichen Zheng, Sang Won Yoon and Hoo Sang Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," European Journal of Operational Research, vol. 26, no. 2, pp. 687-699, 2018.
3. Mehrbakhsh Nilashi, Othman bin Ibrahim, Hossein Ahmadi and Leila Shahmoradi, "An analytical method for diseases prediction using machine learning techniques," Computers & Chemical Engineering, vol. 106, pp. 212-223, 2017.
4. J. Zhang et al., "Coupling a Fast Fourier Transformation With a Machine Learning Ensemble Model to Support Recommendations for Heart Disease Patients in a Telehealth Environment," IEEE Access, vol. 5, pp. 10674-10685, 2017.
5. Evanthia E.Tripoliti, Theofilos G.Papadopoulos, Georgia S.Karanasiou, Katerina K.Naka and Dimitrios I.Fotiadis, "Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques," Computational and Structural Biotechnology Journal, vol. 15, pp. 26-47, 2017.
6. Wuyang Dai, Theodora S.Brisimi, William G.Adams, Theofanie Mela, Venkatesh Saligrama and Ioannis Ch.Paschalidis, "Prediction of hospitalization due to heart diseases by supervised learning methods," International Journal of Medical Informatics, vol. 84, no.3, pp. 189-197, 2015.
7. JaberAlwidian, Bassam H.Hammo and Nadim Obeid, "WCBA: Weighted classification based on association rules algorithm for breast cancer disease," Applied Soft Computing, vol. 62, pp. 536-549, 2018.
8. Mahin Vazifehdan, Mohammad Hossein Moattar and Mehrdad Jalali, "A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction," Journal of King Saud University - Computer and Information Sciences, 2018.
9. Mehrbakhsh Nilashi, Othman Ibrahim, Hossein Ahmadi, Leila Shahmoradi and Mohammadreza Farahmand, "A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques," Biocybernetics and Biomedical Engineering, vol. 38, no. 1, pp. 1-15, 2018.
10. Mehrbakhsh Nilashi, Othman Ibrahim, Hossein Ahmadi and Leila Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method," Telematics and Informatics, vol. 34, no. 4, pp. 133-144, 2017.
11. Sai Prasad Potharaju, M.Sreedevi, Vinay KumarAnde and Ravi Kumar Tirandasu, "Data mining approach for accelerating the classification accuracy of cardiotocography," Clinical Epidemiology and Global Health, 2018.
12. Eun Young Kim, Min Young Lee, Se Hyun Kim, Kyooseob Ha, Kwang Pyo Kim and Yong Min Ahn, "Diagnosis of major depressive disorder by combining multimodal information from heart rate dynamics and serum proteomics using machine-learning algorithm," Progress in Neuro-Psychopharmacology and Biological Psychiatry, vol. 76, pp. 65-71, 2017.
13. M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," IEEE Access, vol. 5, pp. 8869-8879, 2017.
14. Ilayaraja M and Meyyappan T, "Efficient Data Mining Method to Predict the Risk of Heart Diseases through Frequent Itemsets," 4th International Conference on Eco-friendly Computing and Communication Systems, ICECCS, vol. 70, pp. 586-592, 2015.
15. Qiming Chen, Lina Han and Shuli Guo, "GW28-e0435 Prediction and intervention of coronary heart disease based on data mining," Journal of the American College of Cardiology, vol. 70, no. 16, 2017.
16. Maryam Tayefia, Mohammad Tajfard, Sara Saffar, Parichehr Hanachi, Ali Reza Amirabadizadeh, Habibollah Esmaeily, Ali Taghipour, Gordon A. Ferns, Mohsen Moohebati and Majid Ghayour-Mobarhan, "hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm," Computer Methods and Programs in Biomedicine, vol. 141, pp. 105-109, 2017.
17. Yan S, Wang Y, Aghaei F, Qiu Y and Zheng B, "Improving Performance of Breast Cancer Risk Prediction by Incorporating Optical Density Image Feature Analysis: An Assessment," Academic radiology, 2017.
18. Qinghan Xue and Mooi Choo Chuah,"Incentive design for high quality disease prediction model using crowdsourced clinical data," Smart Health, In press, corrected proof, Available online 21 December 2017.
19. Bikesh Kumar Singh, Kesari Verma, Lipismita Panigrahi and A. S. Thoke, "Integrating radiologist feedback with computer aided diagnostic systems for breast cancer risk prediction in ultrasonic images: An experimental investigation in machine learning paradigm," Expert Systems with Applications, vol. 90, pp. 209-223, 30 December 2017.
20. L. N. Pu, Z. Zhao and Y. T. Zhang, "Investigation on Cardiovascular Risk Prediction Using Genetic Information," IEEE Transactions on Information Technology in Biomedicine, vol. 16, no. 5, pp. 795-808, Sept. 2012.
21. T. Vivekanandan and N. Ch Sriman Narayana Iyengar, "Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease," Computers in Biology and Medicine, vol. 90, pp. 125-136, 2017.
22. Baskaran, A. Guergachi, R. K. Bali and R. N. G. Naguib, "Predicting Breast Screening Attendance Using Machine Learning Techniques," IEEE Transactions on Information Technology in Biomedicine, vol. 15, no. 2, pp. 251-259, March 2011.
23. Hiba Asria,Hajar Mousannif,Hassan Al Moatassime and Thomas Noeld, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," The 6th International Symposium on Frontiers in Ambient and Mobile Systems, vol. 83, pp. 1064-1069, 2016.
24. Peter C.Austin, Jack V.Tu, Jennifer E.Ho, Daniel Levy and Douglas S.Lee, "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes," Journal of Clinical Epidemiology, vol. 66, no. 4, pp. 398-407, 2013.
25. Seyedali Mirjalili, "Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems," Neural Computing and Applications, vol. 27, no. 4, pp. 1053-1073, 2016.
26. Amin Khatami, Abbas Khosravi, Thanh Nguyen, Chee Peng Lim, and Saeid Nahavandi, "Medical image analysis using wavelet transform and deep belief networks", Expert Systems with Applications, vol.86, pp.190-198, November 2017.
27. Bissan Ghaddar, and Joe Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines", European Journal of Operational Research, vol.265, no.3, pp.993-1004, March 2018.
28. M. A. jabbar, P. Chandra and B. L. Deekshatulu, "Prediction of risk score for heart disease using associative classification and hybrid feature subset selection," 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), Kochi, 2012, pp. 628-634.
29. Gunasekaran Manogaran, R. Varatharajan and M. K. Priyan, " Hybrid Recommendation System for Heart Disease Diagnosis based on Multiple Kernel Learning with Adaptive Neuro-Fuzzy Inference System", Multimedia Tools and Applications, vol. 77, no. 4, pp. 4379–4399, 2018.
30. Jabbar Akhil, Bulusu Deekshatulu and Priti Chandra, " Intelligent heart disease prediction system using random forest and evolutionary approach", Journal of Network and Innovative Computing, vol. 4, pp. 175-184, 2016.
31. Jaber Alwidian, Bassam H. Hammo and Nadim Obeid, " WCBA: Weighted classification based on association rules algorithm for breast cancer disease, " Applied Soft Computing, vol. 62, pp. 536-549, 2018.
32. Jabbar MA, "Prediction of heart disease using k-nearest neighbor and particle swarm optimization", Biomedical Research, vol. 28, no. 9, pp. 4154-4158, 2017.
33. Mohammad Jafari and Mohammad Hossein Bayati Chaleshtari, " Using dragonfly algorithm for optimization of orthotropic infiniteplates with a quasi-triangular cut-out", European Journal of Mechanics A/Solids, vol. 66, pp.1-14, 2017.
34. BinbinTang, XiaoLiu, JieLei, MingliSong, DapengTao, ShuifaSun and FangminDong, " DeepChart: Combining deep convolutional networks and deep belief networks in chart classification", Signal Processing, vol.124, pp. 156-161, July 2016.
35. YouliYuan, MinZhang, PengfeiLuo, ZabihGhassemlooy, LeiLang, DanshiWang, BoZhang and DahaiHan, " SVM-based detection in visible light communications", Optik, vol.151, pp. 55-64, December 2017.
36. M. Sireesha, Srikanth Vemuru and S. N. TirumalaRao, " Coalesce based binary table: an enhanced algorithm for mining frequent patterns", International Journal of Engineering & Technology, vol. 7, no. 1.5, pp. 51-55, 2018.