Risk Prediction of Coronary Heart Disease Using Hybrid Feature Extraction and Artificial Neural Network Model

[1]Arshia Arjumand banu, [2]Shazia Ali

[1] Department of Computer Science,
[2] Department of Information Technology & Security,
College of Computer Science and Information Technology, Jazan University, Jazan, Kingdom of Saudi Arabia
[1]abanu@jazanu.edu.sa, [2]ssali@jazanu.edu.sa

Abstract - *Heart disorders are main causes for global deaths with mortality rates higher than other diseases. DMTs (Data mining Techniques) have been used to anticipate and diagnose various ailments in healthcare domains. Though clinical datasets incorporating inputs from complex lab tests have been proliferated by these techniques to predict heart diseases, they fail to consider risk factors including age, family histories, other ailments like diabetes, hypertensions, high cholesterols and habits like cigarette smoking, alcohol usage. These risk factors are shared by patients with heart diseases and can assist in obtaining accurate diagnoses. Systems based on such risk factors would not only help medical specialists, but also forewarn people about the dangers of heart diseases long before they visit hospitals or undertake costly medical tests. Hence, this research work proposes early prediction of heart diseases. The large dimensions of data in DMTs have been an issue where many studies have attempted to discover optimal numbers for processing reduced numbers. This study proposes a schema based on hybrid feature extractionswith classifications of heart diseases using ANNs (Artificial Neural Networks). Moreover, this work uses pre-processing procedures including replacements of found missing values which are constant or mean values depending on attributes. HPCA (Hybrid Principal Component Analysis) minimizes dataset's dimensions and ANNs classify heart diseases. This work's performances are assessed using the metrics of precisions, recalls, and F-Measures and accuracies as the classifications involve unequal class distributions in input data.*

Keywords *- Heart Disease,hybrid feature extraction, Hybrid Principal Component Analysis (HPCA), artificial neural network, prediction of heart disease*

DOI Number: 10.48047/nq.2022.20.19.NQ99121 **NeuroQuantology2022;20(19): 1337-1347**

1. Introduction

Heart attacks and strokes kill 17.5 million people perennially where three fourths succumb due to CVDs (cardiovascular diseases) in low/middle income countries according to World Health Organization [1]. Heart attacks account for 80% of mortalities related to CVDs. Human hearts are muscular organs that pump blood throughout the body and are essential components of cardiovascular systems including the lungs which are blood vessel networks including veins, arteries, and capillaries[2].Several forms of heart illnesses or CVDs are caused by anomalies in normal cardiac blood flow (cardiovascular diseases). Heart diseases are major causes of global deaths making it imperative to detect them early for saving lives and enabling doctors in developing effective treatment plans and reducing mortalities due to CVDs. Modern healthcare systems capture considerable quantities of patient data [3] (i.e. Big Data in Electronic Health Record System) which can be used to construct prediction models for diagnosis of CVDs.

The use of DMTs in Medical domains can result in identifying hidden patterns (Clinical Diagnosis) from voluminous data. However, raw medical data are fragmented, diverse, and vast [4] and need to be gathered in systematic manners. DMTs offer uncovering novel and hidden patterns in these data. Estimates by World Health say that heart diseases kill12 million people perennially where CVDs account for major portion of mortalities in developed countries including US. They are also the primary causes of death in many underdeveloped nations [5].Heart disease refer to conditions that affect human hearts and top mortality causes including India. US loses one subject every 34 seconds due to heart diseases. Heart disorders can be found in many types including coronary heart diseases,

Following this introductory section, subsequent section reviews latest strategies for predicting heart diseases. Section three details on the proposed methodology while section 4 which displays results of this study's findings along with discussions. This paper concludes with Section 5 with future scope.

Medical diagnoses are challenging procedures as they need to be conducted accurately and expeditiously. Since, not all doctors are competent in all areas with scarcity in certain areas of specializations, automating these methods can be advantageous [7] for clinical tests at lowered costs. For effective and proper deployments of automated systems, comparative examinations of many approaches are essential.

The significant features in anticipating and diagnosing cardiac diseases are doctors' evaluations based on patient medical histories, symptoms, and physical examination reports [8]. Even Doctors sometimes may fail accurately estimate the ate of hearts in patients, but their predictions match up to 67% since diagnoses are based on analogous symptoms observed in previously diagnosed patients. The use of AIs (Artificial Intelligences) and MLTs (Machine Learning Techniques) in these areas can help in effective predictions of diseases specifically from massive amounts of patient's data made available in healthcare [9]. Disease predictions have generated interest of researches mainly due to the evolving computer technologies in healthcare along with the availability of voluminous health datasets. Integrations of DLTs (deep-learning techniques) with AIs have the potential to improve support for healthcare significantly as voluminous medical data allows MLTs to learnor additional gathering of crucial information [10]. Many domains science, technologies, agriculture, businesses, academics and healthcare are accumulating vast amounts of organised/unstructured data (big data) which are unprocessed. Important information need to be retrieved through data analysis where processes store, handle, analyze, manage, and present these massive amounts of data.

Though many algorithms predict heart diseases from clinical datasets, they fail to consider risk factors including age, family histories, otherailments like diabetes, hypertensions, high cholesterols and habits like cigarette smoking, alcohol usage [11]. These risk factors are shared by patients with heart diseases and can assist in obtaining accurate diagnoses. Systems based on such risk factors would not only help medical specialists, but also forewarn people about the dangers of heart diseases long before they visit hospitals or undertake costly medical tests. As a consequence, based on key risk factors, this study provides an approach for predicting heart disease early.The large dimensions of data in DMTs have been an issue where many studies have attempted to discover optimal numbers for processing reduced numbers. This study proposes a schema based on hybrid feature extractions with classifications of heart diseases using ANNs. Moreover, this work uses pre-processing procedures including replacements of found missing values which are constant or mean values depending on attributes.

Following this introductory section, subsequent section reviews latest strategies for predicting heart diseases. Section three details on the proposed methodology while section 4 which displays results of this study's findings along with discussions. This paper concludes with Section 5 with future scope

2. Literature Review

Several existing approaches on the use of classical data mining and optimization algorithms for the prediction of heart disease data have been presented in this area. The function and efficacy of various feature extraction and nature-inspired methodologies used for diagnosis of the supplied heart disease data have been accessible and presented in this part.

Atallah et al. [12] used majority vote based ensembles to forecast occurrences of CVDs. Their forecasts were based on simple, low-cost medical test data and aimed at better trust and accuracy of diagnoses by doctors. The study's model selected patients based on majority votes of multiple MLTs resulting in effective accuracy of 90% and as the models were trained using realtime data of patients.Nashif et al. [13] predicted heart diseases using MLTs and clouds. For proper detections of heart illnesses, their effective MLTs were developed in Java Based Open Access Data Mining Platform, WEKA. This work's e suggested technique was verified on open-access datasets where its performances were assessed using 10-fold cross validations. The study found accuracy levels of SVMs (support vector machines) as 97.53% with 97.50% sensitivity and 94.94% specificity. Moreover, a real-time patient monitoring system capable of detecting multiple parameters was required to continuously monitor heart disease patients where body temperatures, blood pressures, humidity levels and heartbeats were metered and constructed using Arduino. The suggested scheme transmitted recorded data to centralised servers, updated every 10 seconds. The study's use of real-time sensors and live video streaming was of great help to clinicians in times of patient emergencies. The automated systems also notified clinicians when patient's readings crossed defined threshold values using GSM technology. Vaishali et al [14] used big data created a centralized patient monitoring system which took medical records as inputs. The study aimed at extracting key information from medical records of cardiac patients using map reduce approach. Heart diseases are serious public health problems and one of the main causes of global mortalities. Early detections of cardiac diseases have become crucial in medical studies. In diagnosis of CVDs, parameters such as RR, QRS, and QT intervals are assessed. Classification techniques decide if patients are normal or abnormal, and detection phase detects and minimises illnesses using map reduce strategy.This suggested method aided in the classification of vast and complicated medical datasets as well as detections of cardiac diseases.

Polat et al. [15] used kNNs (k-nearest neighbours) for pre-processing data before main classifications. The study proposed artificial immune recognitions using fuzzy resource allocations which were tested on UCI Machine Learning Databases. The system clocked comparative accuracy of 87% which was higher than previous classifications.

Khazaee et al. [16] classified ECG beats into three types namely normal beats and two signs of cardiac arrhythmia. This system had three key modules in feature extractions, classifiers, and optimizations. The efficient pattern characteristics in feature extraction modules were proper collections of form and temporal characteristics. Classifications were executed using multi-class SVMs where PSO (particle swarm optimization) selected optimal parameter values and upstream features for classifications. Their suggested approach has a very high recognition rate in simulations and with only a few characteristics chosen using PSO. Ali et al [17] developed smart healthcare systems that predicted heart ailments using ensemble DLTs and feature fusions. The study initially combined sensor data with electronic medical records (feature fusion) for producing healthcare data. Subsequently, IGs (information gains) were utilized to eliminate unnecessary and redundant features while emphasizing on relevant characteristics resulting in reduced computations and enhanced system performances. In addition, the study computed conditional probabilities of feature weightsand thus considerably improvingtheir system's performances. Finally, DLTs were trained on the processed data to predict heart diseases. The study's suggested system was compared with known classifiers based on feature fusions/selections, and weighing approaches using data from heart disease patients.Their proposed approach achieved 98.5% accuracy, which was higher than most current techniques. Sonawane et al [18] presented NNs (neural networks) based on MLPs (multilayer Perceptrons) for predicting heart diseases. The study used thirteen clinical characteristics as inputs for training NNs while back propagations predicted CVDs with 98% accuracy. Their obtained accuracy was greater and more efficient than prior approaches.

Sabarinathan et al [19] proposed use of MLTs for predicting cardiac illnesses. Their approach based on DTs (decision trees) categorized input characteristics for providing structural information. Variables including ages, genders, chest pains, and heart rates were used for categorisations. The study predicted aetiology of heart disease with 85% accuracy with their use of DTs for selecting features. Anooj et al [20] created weighted fuzzy rule-based CDSS (clinical decision support system) that learnt from clinical data for detecting CVDs where forecasts were separated into two stages:(1) Automated production of weighed fuzzy rules, and (2) creation of fuzzy rule-based decision support systems. In the first step, mining approaches selected and weighed attributes to get weighed fuzzy rules which were the basics for constructions of fuzzy systems. The study's suggested system was evaluated using UCI repository's datasets. The performances were compared with systems based on NNs in terms of accuracies, sensitivities, and specificities. Das et al [21] described a technique for detecting cardiac illness that employs SAS base software 9.1.3.The suggested system is built on an ensemble approach for neural networks. By merging the posterior probability or anticipated values from numerous prior models, this ensemble-based technique generates new models. As a result, more effective models may be developed. Experiments with the suggested tool were carried out here. Their experiments on Cleveland heart disease database showed an accuracy of 89.01% with sensitivity 80.95% and specificity 95.91% in diagnoses of cardiac diseases. According to Reddy et al [22], who used Bayes functions, lazy meta rules, and trees in training Cleveland heart dataset's entire set of attributes, obtained optimal attributes from the afore mentioned attribute evaluators for efficient risk predictions of heart diseases using 10 fold cross validations. The study adjusted instance-based classifier's hyperparameters, nearest neighbours counts, denoted by the letter 'k.' And while utilising complete set of characteristics, their SMO (sequence minimum optimization) achieved an accuracy of 85.148%, and 86.468% when using the optimal attribute sets as found in chi-squared tests. Their use of meta classifier bagging with LR (logistic regression) and ReliefF attribute evaluator yielded ROC value of 0.91 on both entire and optimum attribute sets. Overall, their proposed SMO classifier outperformed other strategies while IBk enhanced accuracies by 8.25% using chi-squared attribute sets when the hyperparameter 'k' was set to 9.

3. Proposed Methodology

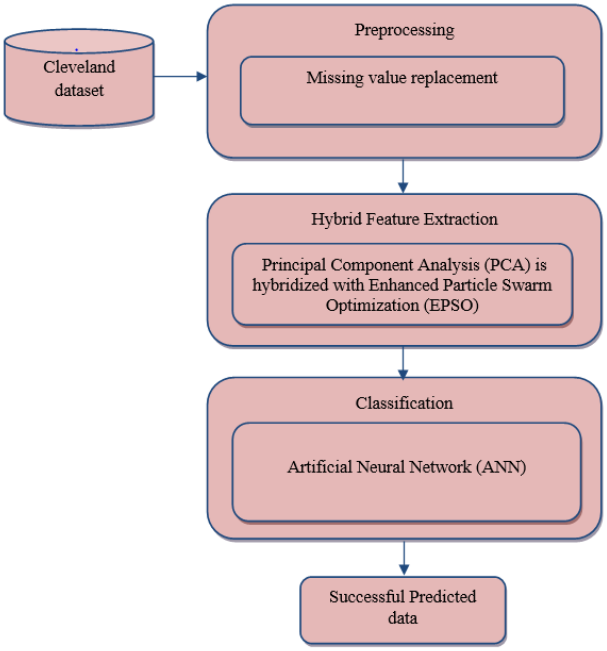

Cleveland dataset
Preprocessing
Missing value replacement
Hybrid Feature Extraction
Principal Component Analysis (PCA) is hybridized with Enhanced Particle Swarm Optimization (EPSO)
Classification
Artificial Neural Network (ANN)
Successful Predicted data

Fig. 1. The suggested heart disease prediction model's procedure

This work proposes a hybrid feature extraction and classification model based on ANNs for predicting heart diseases accurately. Figure 1 displays the proposed technique's procedures. Cleveland heart disease dataset from UCI machine learning library in.csv format was obtained for the study. The data is initially pre-processed where missing values are replaced with a constant value or mean value based on the kind of attributes for better outcomes from MLTs.The proposed HPCA reduces input dataset's dimensionality while PSO is introduced for solving issues of PCA (Principal component analysis) models. ANNs predictheart diseases by classifying these processed instances.

3.2. Dataset Description and Statistics

The Cleveland heart dataset contains 303 examples with 76 features, however only 14 of them are deemed more suited for study and experimentation [22].

Table 1. shows attribute descriptions from the UCI machine learning repository for the Cleveland heart dataset.

Attribute	Description	Type of Attribute	Attribute Value Range
Ages	Age in Years	Numeric	29 to 77
Sexes	Gender	Nominal	0=female, 1=male
Chest Pains	Chest pain type	Nominal	1= typical angina, 2= atypical angina, 3= non-angina pain, 4=asymptomatic
Trest Blood Pressures	Resting blood pressure in mm Hg on admission to the hospital	Numeric	94 to 200
Cholestrol levels	Serum cholesterol in mg/dL	Numeric	126 to 564
Fasting Blood Sugars	Fasting blood sugar levels checked for values > 120 mg/dL	Nominal	0= false, 1=true
Resting ECG	Results of ECGs	Nominal	0= normal, 1=ST-T wave abnormality, 2=definite left ventricular hypertrophy by Estes’ criteria
Thalachs	Achieved Max. heart rates	Numeric	71 to 202
Exangs	Exercises including angina	Nominal	0=no 1=yes
Oldpeaks	ST depressions including exercises relative to	Numeric	0 to 6.2

	rest periods		
Slopes	Peak Slopes attained in ST segments	Nominal	1=upsloping, 2=flat, 3=downsloping
Cas	Counts of colored vessels due to fluoroscopy	Nominal	0-3
Thals	Heart statuses	Nominal	3=normal, 6=fixed defect, 7=reversible defect
Targets	Prediction attributes	Nominal	0=no risk of heart disease, 1 to 4 =risk of heart disease

Nominal or categorical types are attributes with less than 10 classes. The "sexes" feature is divided into two gender categories: 1 for males and 0 for females. The attribute "cps" distinguishes four forms of chest pains: type 1 for classic angina, type 2 for atypical angina, type 3 for asymptomatic, and type 4 for asymptomatic. If the fasting blood sugar levels are more than 120 mg/dL, 'fbs' are divided into two categories: 1 = true and 0 = false. For the attribute "restecg," three categories of resting ECGs were available: 0 for normal, 1 for ST-T wave abnormalities, and 2 for unambiguous left ventricular hypertrophy. The 'exang' characteristic were separated into two groups based on exercise-induced angina: yes (1 = yes) and no (0 = no). The "slopes" provided three types of peaks in ST segments: upslope (1), flat (2), and downslope (3). The property "cas" was divided into four groups based on the number of main vessels (0-3) coloured by fluoroscopy. Three classifications of heart state are included in the property thal: normal (3), fixed (6), and reversible (7). There are five prediction classes for the attribute "target": Heart disease risk ranges from 0 (no risk) to 4 (risk at various stages). Since, the major purpose of this study was to identify CVDs, values from 1 to 4 were changed to 1, resulting in "targets" of two classes namely 0 and 1. Ages, Trestbps, Cho, Thalach, and Oldpeak were all integers or numeric values.

Table 2 (a) shows the statistical Numeric attribute qualities such as lowest, highest, means, standard deviations, missing/distinct/unique values. No values were found missing in the Cleveland dataset's numeric properties.

Table 2. (a) A statistical breakdown of the numerical properties. (b) A statistical breakdown of the nominal properties.

Attribute	Min.	Max.	Mean	StdDev	Missing	Distinct	Unique
age	29	77	54.439	9.039	0	41	4(1%)
trestbps	94	200	131.69	17.6	0	50	17(6%)
chol	126	564	246.693	51.777	0	152	61(20%)
thalach	71	202	149.607	22.875	0	91	28(9%)
oldpeak	0	6.2	1.04	1.161	0	40	10(3%)

Attribute	Lable	Count	Proportion	Missing	Distinct
sex	0 1	97 206	32% 68%	0	2
cp	1 2 3 4	23 50 86 144	7.6% 16.5% 28.4% 47.5%	0	4
fbs	0 1	258 45	85.15% 14.85%	0	2
restecg	0 1 2	151 4 148	49.83% 1.32% 48.84%	0	3
exang	0 1	204 99	67.33% 32.67%	0	2
slope	1 2 3	142 140 21	46.86% 46.20% 6.93%	0	3
ca	0 1 2 3	176 65 38 20	58.08% 21.45% 12.54% 6.6%	4(1.32%)	4
thal	3 6 7	166 18 117	54.79% 5.95% 38.6%	2(0.66%)	3
target	0 1	164 139	54% 46%	0	2

Min.-minimum,Max-maximum,StdDev-standard deviation.

The statistical characteristics of the nominal properties, including label, count, missing values, and distinct values, are shown in Table 2. (b). Six (6) instances out of 303, or 2% of the total dataset, had missing values: four (4) from the 'ca' element and two (2) from the 'thal' field. 164 cases were assigned to label 0 (no risk of heart disease), whereas 139 cases were assigned to label 1 (risk of heart disease), making up 54% and 46% of the dataset, respectively.

3.3. Pre-Processing of Dataset

The dataset is incomplete if there are missing data. When no data value for a variable is maintained in an observation, this is known as missing values or missing data in statistics. Missing values are indicated by blanks or dashes. Respondents' forgetfulness, refusal, or failure to reply to specific questions are main reasons for missing values followed by sensor failures, data losses during uploads, broken internet connections, and incorrect mathematical computations. When missing values are present in a dataset, it is difficult to forecast how they will effect the findings. A dataset may include only a few missing replies for each variable, but the overall number of missing data points may be significant. Although the analysis may be carried on, the results may be statistically insignificant owing to missing data. As a result, instead of eliminating missing values from datasets, they were replaced with user defined constants or mean values. The nominal features 'cas' and 'thal ' having missing values were substituted with user constants based on majority marks in the Cleveland heart dataset, The attribute "ca" had four missing values in 176 of 299 observations, with the mark "0" being the most common. In contrast, the attribute "thal" had two missing values in 166 out of 301 observations and a majority mark of 3. The missing values in "cas" and "thal" were changed to the corresponding majority marks of 0 and 3, respectively, to guarantee that the dataset is complete.

3.4. Feature Extractions using HPCA

Feature extractions are dimensionality reduction methods where most data characteristics are efficiently represented in miniatures that represent the complete data captured effectively. In this work, the PCA is hybridized with the Enhanced Particle Swarm Optimization. The details of this hybrid concept is explained in the below section.

1) PCA

PCA is a technique for reducing data dimensionality. PCA is a method for least-squares projection of high-dimensional data to a lower dimension. It reduces dimensionalities of collected data using variable sets, much smaller than original variables while capturing significant primary variability in data and rejecting little variability [23]. PCA also extracts critical information from confusing data sets while maintaining maximum information in samples. They are simple, non-parametric techniquesfor reducing dimensionalities. As a result, they assist in classifications and data compressions. Many different fields have used PCA including dimensionality reductions, pattern detections, image processes, and compressions.

Finding the direction with the largest variance in the input space and determining the covariance matrix's primary eigenvector are the two basic objectives of PCA. The following algebraic definition applies to PCA:

1) For any data matrix X, calculate covariance and mean of X as:

$$\mu \coloneqq E\{X\} \text{ (1)}$$

Subsequently, compute covariance as:

$$R = Cov\{X\} = E\{(X-\mu)(X-\mu)^T\} \text{ (2)}$$

2) count eigenvalues λ_i and eigenvectors e1,e2, . . . , eN, i= 1, 2, . . . , N of covariance R, sorting eigen values in descending orders for R, solving the equation:

$$|\lambda I - R| = 0 \text{ (3)}$$

Use SVDs to decompose and obtain eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \Lambda, \geq \lambda_p \geq 0 \text{ (4)}$$

and their corresponding eigen vectors $e_i(i = 1,2,\Lambda,p)$

Select λ_i for obtaining component principals by counting proportions of data covered by initial M eigen values.

$$\frac{\sum_{i=1}^{M} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \text{ (5)}$$

M eigen values (85%) are selected from larger cumulative proportions as principal components.

3) project data to lower dimension subspaces

$$P = w^T X \text{ (6)}$$

By using M, reduce variables or dimensions counts from n to M (M« n).

PCA seeks the optimal linear combinations of variables for describing data and this work uses PCA for feature extractions and reducing dimensionalities of data. While using PCA on discrete variables is technically doable. However, if the variables do not belong on a coordinate plane, do not use PCA on them. As a result, our work developed an improved PSO to solve the aforementioned difficulty.

2) EPSO (Enhanced Particle Swarm Optimization)

PSO particles explore the solution space for the best possible solutions, which can be found as single birds in the swarm or as groups of particles in search spaces where the ith particle's position Xi = (x1,x2,...,xd) and velocity Vi = (v1,v2,...,vd) [24] in the D-dimensional search space. Particles in search spaces travel in line with their own experiences as well as the experiences of their neighbours; as a result, the particles are aware of both their best present locations and best positions identified by swarms. The best locations for ith particle, is called PBi, while best positions for the swarms are called gbest. All particles are assigned an initial position and velocity in a randomized manner.Fitness functions are used to assess particles in search spaces which are drawn to the best particles in respective search spaces and with the highest fitness values. The following equation is used to update the velocity and location of particles.

$$v_i^{t+1} = \omega * V_i^t + c1 * rand1 * (PB_i - X_i^t) + c2 * rand2 * (gbest - X_i^t) \quad (7)$$

where, i=1,,,,,,n , N stands for swarm sizes, rand1 and rand2 are random numbers uniformly distributed in the interval [0,1], c1 and c2 are constants. Based on particle, neighbouring particle experiences and inertia of previous velocities, particle's velocitiesare updated. Updates on particle position are provided by:

$$X_i^{t+1} = X_i^t + V_i^{t+1} \quad (8)$$

Every particle's fitness is tested in each iteration. Based on the values of pbest and gbest as well as elements like c1, c2, and w, velocity is updated. Depending on the velocity, the location is updated. Until the maximum number of repetitions or the error requirements are reached, this technique is used..

- **Fitness Function**

Fitness are metrics that determine particle's qualities. Particles are tested for their fitness throughout iterations. The capacity of a particle to promote class separation is used to assess its quality. The binary numbers 0 and 1 indicate the position vector. Let N1,N2,...,NL represent the number of photos in each class, and let c1.c2,...,cL represent the classes. L, where L is the number of courses overall. The overall number of photos is N. Let M0 be the overall mean, or mean of all classes, and M1, M2,..., ML be the means of each class. They are determined as follows.

$$M_i = \frac{1}{N_i} \sum_{j=1}^{i} C_j^i i = 1,2, \dots . L \quad (9)$$

$$M_0 = \frac{1}{N_i} \sum_{j=1}^{i} N_i M_i i = 1,2, \dots . L \quad (10)$$

Fitness function is calculated as:

$$F = \sqrt{\sum_{i=1}^{L} (M_i - M_0)(M_i - M_0)^t} \quad (11)$$

It has been proven that PSOs outperform alternative strategies in terms of efficiency and cost. It may be parallelized as well. Additionally, it does not take advantage of the gradient of the issue. The PSO approach has a low iterative convergence rate and the ability to easily enter local optimum in high-dimensional space. As a consequence, for the optimum outcomes, this study presented a Mutation probability function-based PSO method.

- **Mutation probability function based PSO**

Mutation probability functions in PSOs select optimal features. The purpose of this research was to maximize intra/inter-cluster distancesin developments of mutation probability functions. Intra-cluster distances are distances between two similar sorts of samples Distances between two groups of people are calculated using inter-cluster distances. If X and Y are two vectors (rows) in a feature matrix containing two samples from the same person, and Z is a vector in a separate individual's feature matrix, the mutation probability function for each vector is:

$$M_p = \sum_{c=1}^{w} (X_c - Y_c)^2 - \sum_{c=1}^{w} (X_c - Z_c)^2 \quad (12)$$

w is the vector width and p is the vector number in feature matrix. For all M_p , the mutation probability for the global best position gbest can be adapted to choose the maximum value of M_p as follows:

$$M_{gbest} = \max(M_p) \quad (13)$$

Based on the value of M_{gbest}, the selected vector with best global position. When M_p is maximized, differences between samples from same groups are minimized, while differences between samples from different groups are maximised. This procedure increases recognition precisions.

3.4. Classification using ANNs

Neurons are the processing units that make up ANNs. A synthetic neuron makes an effort to replicate the appearance and functionality of a natural neuron. Dendrites are the neuron's inputs, while synapses along its axon are its outputs [25]. Whether or whether a neuron is activated depends on its function. The inputs to the neuron are x1...xn. A bias is also given to the neuron along with the inputs. The bias value is typically set at 1. Weights range from W0 to Wn. The relationship of the signal is its weight. The weight and input product determines the signal strength. A neuron generates a single output after accepting several inputs from multiple sources. Many different functions are used for activation. The sigmoid function is a popular type of activation function. Weights are inter-unit connection strengths that store processing capabilities. The weight value establishes the input strength. The weight's value might be zero, positive, or negative. A negative weight denotes a weakening or obstruction of the signal. If the weight is 0, there is no connection between the two neurons. The weights are changed to provide the desired result. The ANN weights may be changed using algorithms to produce the desired results.

The process of changing weights is referred to as learning or training. The back-propagation methodology is a methodical approach for training MLPs, as seen in Fig. 2. A number of problems may be addressed using back-propagation techniques. The back-propagation training algorithm consists of three stages:

i. Feed-forward input training pattern

ii. Replication of the relevant mistake

iii. Weight adjustment Only input patterns with arbitrary accuracy may be learned using a multilayer network.

A multilayer feed forward NNs with input, output, and hidden layers are known as an MLPs. Weights on connections from units whose output is always positive are equivalent to biases in the hidden and output layers. The extended delta rule is the foundation of the back-propagation algorithm. It uses a gradient descendent technique to minimise the output of the network's total squared error.

I. Forward the training pattern's input

During feed-forwards, input neurons (Xi) receive input signals and broadcast them to hidden neurons which then compute activations and send signals to output units, which compute activations again to produce net outputs.

II. The linked error's back-propagation

The goal values are compared to net outputs during training, and the appropriate errors are determined. The error factors k obtained from errors are used to disperse errors back to hidden layers.

III. Weight modifications

Weight adjustment, which calls for the weight to be updated whenever an error occurs in order to transfer it to the concealed layer, The error factor j is determined for unit Zj in a similar manner. After obtaining the error factors, the weights are simultaneously updated..

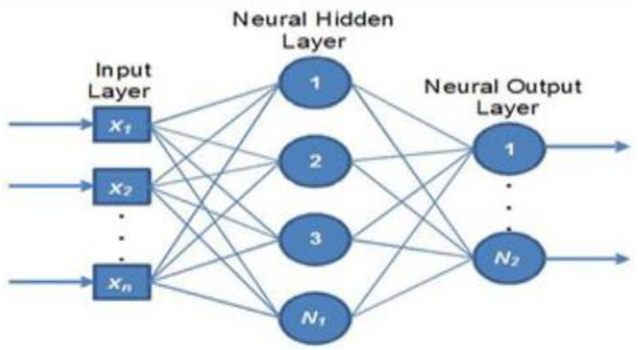


Fig. 2A two layer MLP architecture

The output layer, where the output is computed, gets the outputs of the hidden layers in the BP algorithmThis output is compared to the expected output for the supplied input. Based on this discrepancy, the error is returned from the output layer to the hidden layer and from the hidden layer to the input layer. The weights of the neurons change as the flow is restored. An epoch is a cycle that proceeds from input to output and then returns from output to input. A known set of input data is fed into a neural network, which is then instructed to produce a known output. This is referred to as network training. The network cycles through several of these epochs until the error is discovered. The network is currently thought to have been trained. The weights of all neurons in all layers are determined by this training technique and weights produced by the network's training are used in computing network's reactions to unknown data.

4. Results and Discussion

The proposed HPCA-ANN (Hybrid PCA with ANNs) model's performance is compared to that of the current classifiers, namely SMO and CDSS. Other than classification accuracy, various statistical metrics, as shown in equations (14)-(17), are used to evaluate the classifier, as are the average results for the classifiers..

Precisions are defined as ratios of correctly found positive observations to all positive observations.

Precision = TP/TP+FP (14)

Sensitivities are defined as ratios of correctly identified positive observations to total observations in real classes – yes.

Recall = TP/TP+FN (15)

F1 scores are weighed averages of Precisionsand Recalls and hence consider false positives and false negatives.

F1 Score = 2*(Recall * Precision) / (Recall + Precision) (16)

Accuracies are computed in terms of positives and negatives as follows:

Accuracy = (TP+FP)/(TP+TN+FP+FN) (17)

Where TP, FP, TN, FN are defined as True Positive, False Positive, True Negative, False Negative respectively. These parameter values are calculated for all the standard and proposed classifiers for all five benchmark datasets (UCI, 2010) and a rice disease dataset.

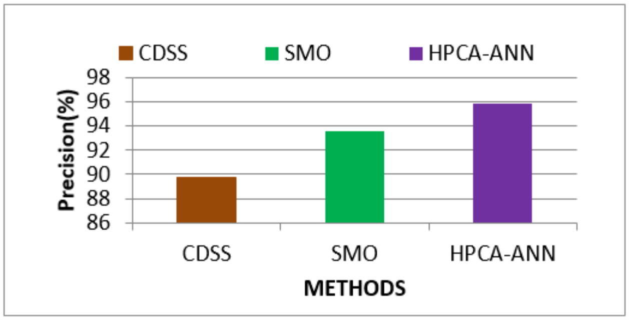


Fig. 3 depicts the precision comparison results of the proposed Hybrid PCA with ANNs (HPCA-ANN)

Figure 3 depicts the precision comparison results of the proposed Hybrid PCA with ANNs (HPCA-ANN). As a result, the findings demonstrate that feature extraction utilising HPCA can be useful in predicting heart disease categorization. As a result, the suggested HPCA contains a variety of valuable characteristics that have no effect on the speed of linear transformation. It is a desirable trait since it eliminates the need to painstakingly modify the regularisation parameter in the classifier. The suggested HPCA employs a highly effective classification algorithm.

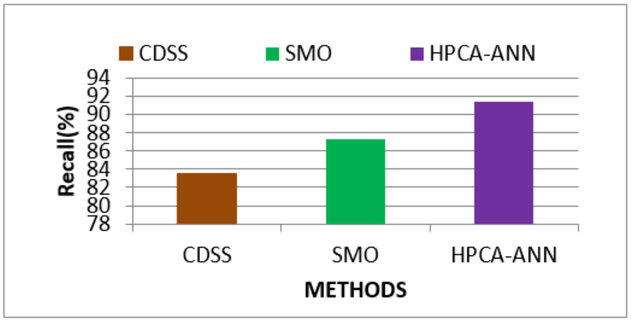
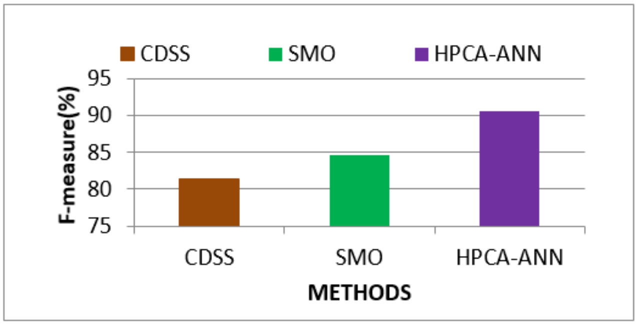


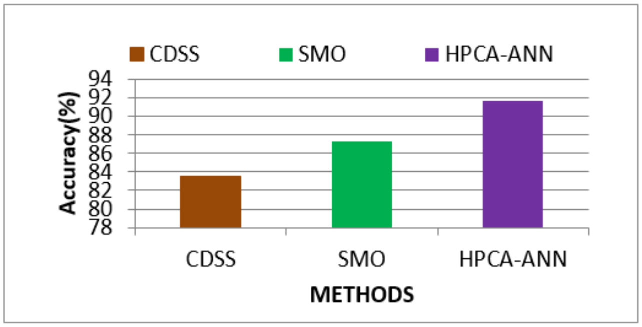
Fig. 4. The suggested Hybrid PCA with ANNs recall comparison findings (HPCA-ANN)

The performance results of the proposed Hybrid PCA with ANNs (HPCA-ANN) based classifier are shown in figure.4. Thus, the findings demonstrate that the suggested approach has a high recall rate of 91.74%, whereas the current techniques have lower recall rates, such as the SMO method metric, which has 89.68%, and the CDSS method metric, which has 87.25%.



F-measure comparison results of the proposed Hybrid PCA with ANNs are shown in Figure.5 (HPCA-ANN)

Figure 5 shows that the proposed Hybrid PCA with ANNs (HPCA-ANN) performs admirably in terms of illness prediction rate, considerably outperforming the SMO and CDSS. The quantitative analysis results in terms of F-measure accord with the qualitative analysis utilising MLTs. In terms of accuracies for heart disease datasets, the proposed Hybrid PCA is compared with many existing classification techniques.



Accuracy comparison results of the proposed Hybrid PCA with ANNs are shown in Figure.6 (HPCA-ANN)

Figure.6 indicates that the proposed Hybrid PCA with ANNs (HPCA-ANN) outperforms the existing classifier in terms of accuracy. The above stated classifiers mentioned were applied on static data and performed badly when compared to HPCA-ANN classifications proving the proposed method's better accuracy and efficacy in classification of CVDs.

5. Conclusion

The healthcare business generates massive volumes of data, which are unfortunately not ";mined"; to disclose hidden information for effective decision-making. It is frequently underutilised for detecting underlying links and patterns. This issue might be remedied by employing advanced DMTs. This study proposes a hybrid feature extraction and ANNs-based classification algorithm to properly forecast heart illness. The Cleveland heart dataset was used in this study to improve MLTs' ability to predict the likelihood of acquiring heart disease. The HPCA-ANN classifier performed admirably when the chi-squared attribute assessment approach was used. Eventually, it was discovered that adequate feature extraction and tweaking the hyper parameters of the classifiers resulted in a considerable improvement in prediction performance. Finally, this study helps in forecasting patients with CVDs by cleaning the dataset and applying CDSS and SMO to reach an accuracy of 91.68% on the recommended HPCA-ANN model, which is better than the earlier models' accuracy of 83.54% and 87.25%. A hybrid model combines numerous well-known classification and selection processes into a single model to get superior results. It has been demonstrated that hybrid models provide very high accuracy when the appropriate mixes of different approaches are applied. In the future, an intelligent system might be developed to assist a patient with heart disease in selecting the optimal MLTs and DLTs.

References

[1] Taneja, A. (2013). Heart disease prediction system using data mining techniques. *Oriental Journal of Computer science and technology*, *6*(4), 457-466.

[2] Jalali, S. M. J., Karimi, M., Khosravi, A., & Nahavandi, S. (2019, October). An efficient neuroevolution approach for heart disease detection. In *2019 IEEE international conference on Systems, Man and Cybernetics (SMC)* (pp. 3771-3776). IEEE.

[3] Kavitha, K. S., Ramakrishnan, K. V., & Singh, M. K. (2010). Modeling and design of evolutionary neural network for heart disease detection. *International Journal of Computer Science Issues (IJCSI)*, *7*(5), 272.

[4] Kirar, A. T. (2022, June). Machine learning based Heart Disease Detection System. In 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) (pp. 1-7). IEEE.

[5] Khan, Y., Qamar, U., Yousaf, N., & Khan, A. (2019, February). Machine learning techniques for heart disease datasets: a survey. In *Proceedings of the 2019 11th International Conference on Machine Learning and Computing* (pp. 27-35).

[6] Learning, M. (2017). Heart disease diagnosis and prediction using machine learning and data mining techniques: a review. *Advances in Computational Sciences and Technology*, *10*(7), 2137-2159.

[7] Dowsley, T., Al-Mallah, M., Ananthasubramaniam, K., Dwivedi, G., McArdle, B., & Chow, B. J. (2013). The role of noninvasive imaging in coronary artery disease detection, prognosis, and clinical decision making. *Canadian Journal of Cardiology*, *29*(3), 285-296.

[8] Devi, A., & Misal, A. (2013). A survey on classifiers used in heart valve disease detection. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2(1).

[9] Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. *Expert systems with applications*, *40*(1), 96-104.

[10] Alizadehsani, R., Zangooei, M. H., Hosseini, M. J., Habibi, J., Khosravi, A., Roshanzamir, M., ... & Nahavandi, S. (2016). Coronary artery disease detection using computational intelligence methods. *Knowledge-Based Systems*, *109*, 187-197.

[11] Verma, L., Srivastava, S., & Negi, P. C. (2018). An intelligent noninvasive model for coronary artery disease detection. *Complex & Intelligent Systems*, *4*(1), 11-18.

[12] Atallah, R., & Al-Mousa, A. (2019, October). Heart disease detection using machine learning majority voting ensemble method. In *2019 2nd international conference on new trends in computing sciences (ictcs)* (pp. 1-6). IEEE.

[13] Nashif, S., Raihan, M. R., Islam, M. R., & Imam, M. H. (2018). Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. *World Journal of Engineering and Technology*, *6*(4),

854-873.

[14] Vaishali, G., & Kalaivani, V. (2016, January). Big data analysis for heart disease detection system using map reduce technique. In *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)* (pp. 1-6). IEEE.

[15] Polat, K., Şahan, S., & Güneş, S. (2007). Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing. *Expert Systems with Applications*, *32*(2), 625-631.

[16] Khazaee, A. (2013). Heart beat classification using particle swarm optimization. *International Journal of Intelligent Systems and Applications*, *5*(6), 25.

[17] Ali, F., El-Sappagh, S., Islam, S. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, *63*, 208-222.

[18] Sonawane, J. S., & Patil, D. R. (2014, February). Prediction of heart disease using multilayer perceptron neural network. In *International conference on information communication and embedded systems (ICICES2014)* (pp. 1-6). IEEE.

[19] Sabarinathan, V., & Sugumaran, V. (2014). Diagnosis of heart disease using decision tree. *International Journal of Research in Computer Applications & Information Technology*, *2*(6), 74-79.

[20] Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University-Computer and Information Sciences*, *24*(1), 27-40.

[21] Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert systems with applications*, *36*(4), 7675-7680.

[22] Reddy, K. V. V., Elamvazuthi, I., Aziz, A. A., Paramasivam, S., Chua, H. N., & Pranavanand, S. (2021). Heart disease risk prediction using machine learning classifiers with attribute evaluators. *Applied Sciences*, *11*(18), 8352.

[23] Richardson, M. (2009). Principal component analysis. URL: http://people. maths. ox. ac. uk/richardsonm/SignalProcPCA. pdf (last access: 3.5. 2013). Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik@ ntf. uni-lj. si, 6, 16.

[24] Poli, R., Kennedy, J., & Blackwell, T. (2007). Particle swarm optimization. *Swarm intelligence*, *1*(1), 33-57.

[25] Zhang, Z. (2018). Artificial neural network. In *Multivariate time series analysis in climate and environmental research* (pp. 1-35). Springer, Cham.