ISSN: 0974-5823 Vol. 7 No. 2 February, 2022

International Journal of Mechanical Engineering

A Novel Pulse Coupled Genetic Particle Swarm Optimization Algorithm with Neural Network classifier for Heart Disease Prediction

[1]A. Sahaya Arthy and [2]G.Murugeswari

[1]Research Scholar, Associate Professor, Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli,

Tamil Nadu, India

Abstract - Heart disease is still a major public health concern around the world. Restricting human knowledge and ability in manual diagnosis results in faulty diagnoses in the health-care system, as well as information about various illnesses collected through various forms of medical technology that is either insufficient or inaccurate. Intelligent resources assist doctors in making better decisions and initiating therapy early since proper identification of a person's condition is critical. Data mining is used as effective support systems in predicting the disease where large volume of data involved. More amounts of attributes helps to predict the disease accurately, but takes more time. It is critical to find the suitable attributes that provide best prediction accuracy. In this paper, we propose a novel hybrid algorithm namely Pulse Coupled Genetic Particle Swarm Optimization Neural Network (PCGPSONN) for feature selection. The proposed algorithm combines Genetic algorithm, Pulse Coupled Neural Network and Particle Swarm optimization algorithm together. A heart disease prediction model is developed using the proposed algorithm and SVM classifier. The prediction model is validated with Real time Data sets, Hungarian data set and UCI machine learning repository using various performance metrics such as accuracy, precision, and recall rate. It is found that our proposed algorithm performs well in feature selection and we achieved a maximum classification accuracy of 96.29% for heart disease prediction.

Index Terms - Heart Disease, Genetic algorithm, Prediction, Pulse Coupled Neural Network, Particle Swarm Optimization, feature selection

INTRODUCTION

Heart disease is one of the leading causes of death worldwide. The patient's signs, symptoms, and physical examination are widely utilised to diagnose cardiac illness. Smoking, high cholesterol, a family history of heart disease, obesity, high blood pressure, and a lack of physical activity are all factors that contribute to heart disease. Heart disease and stroke are the two most common health concerns in the United States today. In order to a recent World Health Organization (WHO) report published in 2018, heart disorders caused 56.9 million deaths worldwide in 2016 [17]. In 2008, 17.3 million individuals died from heart disease [16]. The World Health Organization (WHO) acknowledged data mining's promise for identifying early stages of heart disease and offering precise illness remedies. In order to the World Health Organization (WHO), 14 million of India's 30 million individuals with heart disease live in cities, while 16 million live in rural regions. A multitude of diseases might disrupt the blood flow to the heart, resulting in leakage or incorrect shutdown [27].

Heart disease kills both men and women in the United States every year, despite the fact that it is generally referred to as "men's sickness." In the first year after a heart attack, one out of every four women dies, compared to one out of every five males. Overweight and obesity, poor nutritional condition, physical incapacity, and excessive alcohol use are all factors that raise the danger of heart disease in persons [30].

Various applications make use of data mining methods. In the healthcare industry, data mining has proven to be useful in predicting sickness. The amount of tests will be reduced using mining algorithms. This reduction test is crucial for accuracy and timing. In data mining, there are a variety of learning approaches that may be used to observe large volumes of previously available data. Decision Tree (DT), Multi-layer Perceptron (MLP), Nave Bayes (NB), K-nearest neighbour (K-NN), and Support Vector Machine (SVM) are some of the methodologies [20].

The feature selection method [5] aids in the reduction of irrelevant features. There are a number of ways for selecting features, including WEKA and R. Some of the tools are freely available on the internet. [9] To forecast cardiac disease, the reduced features are sent into classification algorithms. To forecast cardiac disease, the reduced features are sent into classification algorithms. The focus of this research is on feature selection techniques for building predictive models.

The model's performance is then assessed using a heart disease dataset. To find and use the most important and critical features, we apply a feature selection algorithm. These traits are subsequently incorporated into cardiac disease prediction systems. These models' accuracy is then compared.

CONTRIBUTIONS OF THE RESEARCH WORK:

The following are the major contributions of this work:

(i) To reduce the amount of characteristics, a novel feature selection technique called the Pulse Coupled Genetic Particle Swarm Optimization Neural Network (PCGPSONN) algorithm is developed.

(ii) The proposed algorithm's performance in predicting heart disease is tested.

(iii) The suggested algorithm's performance is compared to that of a few current approaches.

RELATED WORK

Using a variety of multiple classifiers and feature selection strategies, several experiments on medical data sets are carried out. Many of them are extremely accurate at classifying [7]. There are very few publications available for classification of cardiac disease datasets.

Tan et al. [24] used a wrapper strategy, which they successfully implemented using a combination of Support Vector Machine (SVM) and Genetic Algorithm (G.A). The results are evaluated and improved using the LIBSVM algorithms and the WEKA data mining tool. The Irvine UC machine learning repository was used to obtain five data sets for this project (Iris, diabetes, breast cancer, heart disease, and hepatitis). Heart disease prediction accuracy was 84.07 percent using a hybrid GA and SVM technique, 78.26 percent for diabetic data, 76.20 percent for breast cancer, and 86.12 percent for hepatitis sickness.

For cardiac disease prediction, Saba Bashir et al. [3] used a variety of data mining approaches, including suppressed Decision Tree, Logistic regression, Logistic regression SVM, Nave Bayes, and Random forest. The author compared the performance of those existing research efforts using fast miner. Decision Trees have an accuracy of 82.22 percent, whereas Logistic Regression has an accuracy of 82.56 percent, and Random Forest, Nave Bayes Logistic Regression SVM have accuracy of 84.17 percent, 84.24 percent, and 84.85 percent, respectively. The authors suggested that the best technique for predicting heart disease is Logistic Regression. PSO and the Genetic Algorithm were integrated by Swati Sharma and Sukhvir Singho [22]. In terms of forecasting cardiac disease, this integrated strategy outperforms the individual techniques. Many academics use the UCI data repository for analysis purposes.

Otoom et al. [19] developed a database for predicting cardiac disease. This dataset comprises 303 cardiac patient instances with 76 attributes/features. Otoom et al. [19] created a feature selection approach that chooses 13 attributes out of 76. Utilizing the WEKA tool, two experiments were conducted out by three algorithms: Nave Bayes, SVM, and Functional Trees (FT). The SVM technique has an accuracy of 88.3 percent, and the cross-validation test showed that SVM plus Navie Bayes net had an accuracy of 83.8 percent. With the help of FT, we were able to reach an accuracy of 81 percent. The Naive Bayes method had an accuracy of 84.5 percent, the SVM algorithm had an accuracy of 85.1 percent, and the FT classifier had an accuracy of 84.5 percent. The top seven traits are chosen using the Best First selection method.

By using a feature selection algorithm on a heart illness data set, Robin Spenceret al. [23] created a model for predicting heart disease. A amount of feature selection strategies are coupled with the Machine Learning algorithm in order to obtain the best potential answer. The author achieved 85.0 percent accuracy, 84.73 percent precision, and 85.56 percent recall using Chi-squared feature selection with Bayes Net classifier.

CengizGazelolu et al. [10] offered 18 machine learning approaches that were grouped into six groups. The authors also offered three distinct feature selection methods. Tools including WEKA, PYTHON, and MATLAB were used to analyse these approaches. PolyKernelSVM was discovered to be the most efficient machine learning method, providing 85.15 percent accuracy, according to the data. After using the Correlation-based Feature Selection (CFS) technique, the most efficient algorithms were discovered to be Naive Bayes and Fuzzy Rough Set. However, employing the Chi-Square function selection yielded the most succinct result. The RBF Network method was discovered to be the most efficient, with an accuracy of 81.19 percent.

According to AshirJaveed et al. [11], he devised a number of sophisticated diagnostic techniques for heart disease diagnosis. The floating window with customizable size is the best feature selection approach for feature deletion (FWAFE). For heart disease prediction, two types of classification frameworks are utilised when feature selection is eliminated: Artificial Neural Network (ANN) and Deep Neural Network (DNN) (DNN). FWAFE-ANN and FWAFE-DNN are two types of hybrid diagnostic systems proposed in this study. The online data collection from UCI Cleveland was used. With an ANN-based system, the best classification accuracy was 91.11 percent, and with a DNN-based diagnostic system, it was 93.33 percent.

YounessKhourdifi and Mohamed Baha [14] The Fast Correlation-Based Feature Selection (FCBF) technique was created to filter duplicate features. Among the categorization algorithms employed are K-Nearest Neighbour, Support Vector Machine, Nave Bayes, Random Forest, and a Multilayer Perception Artificial Neural Network optimised using Particle Swarm Optimization (PSO) with the addition of Ant Colony Optimization (ACO) techniques. This approach has a maximum classification accuracy of 99.65 percent. PSO method was proposed by Azhar Hussein Alkeshuosh et al. [1] to generate the best guidelines for predicting heart disease. It has a higher prediction accuracy and a smaller rule set than C4.5. The accuracy of the PSO approach is 87 percent, while C4.5 is 63 percent.

To forecast cardiac disease, C. S. Dangare and S. S. Apte [4] used categorization algorithms. Decision trees, Naive Bayes, and Neural Networks were used to compare the outcomes of three categorization algorithms. The number of attributes used in Neural Networks to predict heart disease ranges from 15 to 100 percent accurate. When n is large, the accuracy suffers. When the amount of characteristics is increased to 13, however, the accuracy decreases.

The Infinite Latent Feature Selection (ILFS) approach by Le, Hung Minh et al. [15] is used to weight and re-order heart disease variables based on their rank and weights. In their study, they used a public dataset from the UCI Machine Learning Repository for Heart Disease. Experiments reveal that the proposed method is successful in predicting heart disease properly. The performance for distinguishing 'No presence' HD from 'Presence' HD is excellent, with an accuracy of 90.65% and an Area under the Curve of 0.96.

Using Cleveland and statlog project heart datasets, N. Satish Chandra Reddy et al. [21] predicted heart disease. Based on three distinct percentage splits, the random forest method has a 95 percent accuracy in both classification and feature selection. To create a better performance model, the authors identified 8 or 6 features as a minimal feature need.

3. Data Set Description

In this work, a real time data set is used for analysis. The patient's data has been collected from the hospitals and dataset is produced. The detail of the data set is described in Table 1.

Table 1: Data Set Description

S.No	Attributes Name	Description	Range of Values
1	AGE	Patient Age	28-78
2	GENDER	Gender of the Patient	Male/Female
3	GTT	Glucose Tolerance Test	120-180
4	FASTING	Blood test with fasting	80-115
5	POST-PRA	Postprandial blood glucose Test	100-140
6	HBA1C	Glycated Hemoglobin test	5-8
7	HEIGHT	Measurement of human from head to foot(cm)	140-160
8	WEIGHT	Body weight(Kg)	50-100
9	BMI	Body Mass Index	20-30
10	URIC ACID	Waste product found in blood	2-10
11	TG	Triglycerides Level Test	100-500
12	LDL-C	Low-Density Lipoprotein Cholesterol (mg)	80-170
13	VLDL	Very-Low-Density Lipoprotein *Cholesterol*	20-170
14	HDL-C	High-Density Lipoprotein *Cholesterol*	0-80
15	SYSTOLIC	Systolic Blood Pressure	110-150
16	DIASTOLIC	Diastolic Blood Pressure	80-90
17	HSCRP	High-Sensitivity C-Reactive Protein	15-20
18	SMOKERS	Smoking habit	0-No, 1-Yes
19	CHAIN SMOKERS	Habit of Heavy smoking	0-1
20	NON-SMOKERS	Quit smoking habits	0-1
21	ALCOHOLCONSUMPTION	Alcohol drinking habit	0-1
22	DIET	*Eating* habits	0-1
23	PHYSICAL ACTIVITY	Movement of the body that uses Energy	0-1
24	WORK STRESS	Stress at Work	0-1

PROPOSED APPROACH

In this work, we propose a novel feature selection algorithm namely Pulse Coupled Genetic Particle Swarm Optimization Neural Network algorithm which is a hybridization algorithm. This algorithm combines Genetic Algorithm (GA), Pulse Coupled Neural Network (PCNN) Algorithm and Particle Swarm Optimization (PSO) algorithm together. To find the fitness value for all qualities, the proposed technique employs a Genetic algorithm. The PCNN algorithm aids in determining the optimal Pg location. PSO then takes care of the rest. In most cases, both GA operators and the PSO update mechanism work with the same population throughout the PSO startup procedure. All of the values in the population are created at the start and distributed at random.. So this random distribution must lead to slow convergence and leads to more iteration of attribute selection algorithm. But in our proposed approach, GPCNN assigns the initial PSO population. GPCNN and PSO split the total amount of iterations in half.

GPCNN performs the first half of the iterations, and the results are used to populate PSO. Remaining iterations are run by PSO. So it overcomes the problem of slow convergence and decrease the iteration of the attribute selection process. The working stages of the proposed work are listed as below.

Attribute Selection using PCGPSONN

Heart Disease Prediction using SVM

An elaborate account of every stage is provided in the subsequent sections.

I. *Pulse Coupled Genetic Particle Swarm Optimization Neural Network (PCGPSONN) algorithm*

Attribute selection algorithm requires careful selection of attributes for predicting heart disease. An improper choice of the attributes can lead to low accuracy, prediction error or failure. If the starting point of attribute selection was chosen inappropriately, more iterations of the algorithm will only make it move toward a local minimum, never reaching the global one. Our proposed method includes both global search and local search using hybridization. And then best

position Pg is generated for each attribute is given to attribute selection algorithm for refining search process. The outline of the proposed PCGPSONN approach is shown in Figure 1.

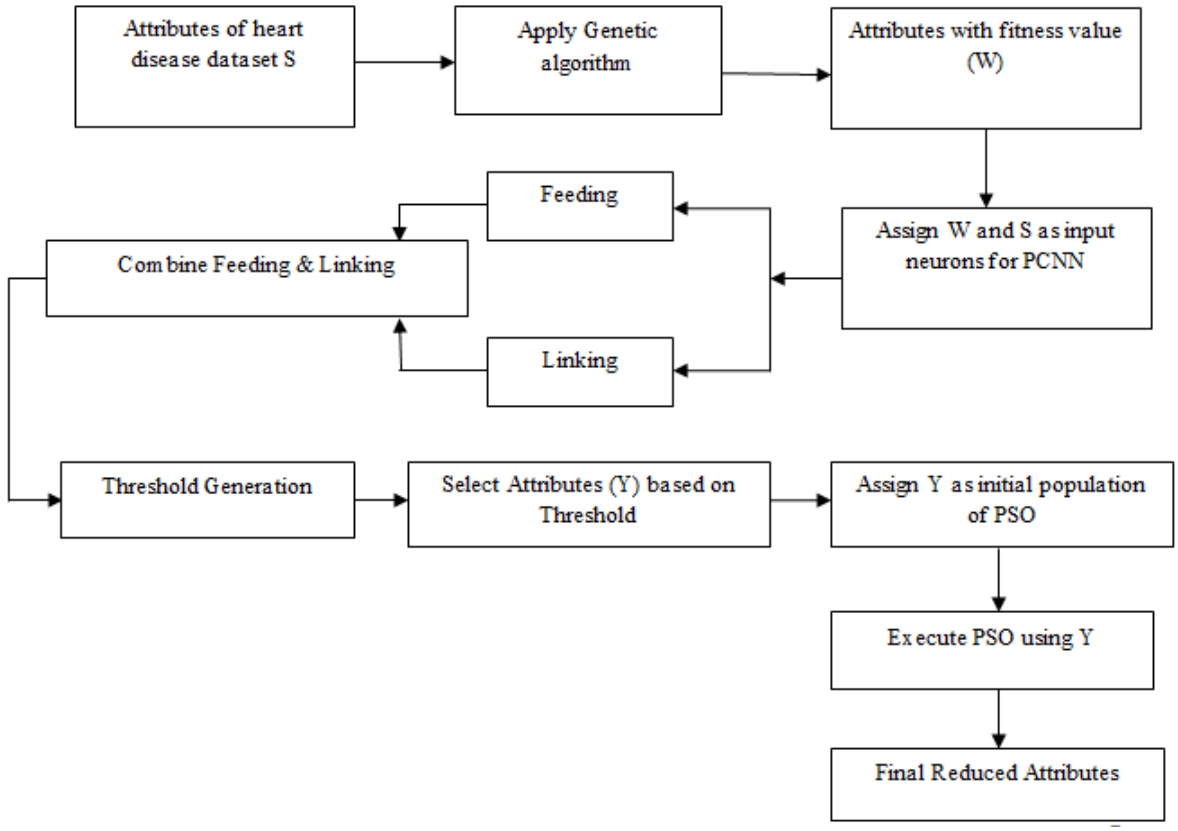

Figure.1 Outline of the Proposed Algorithm

II. *. Genetic Algorithm*

To find the best attributes to predict heart disease, genetic algorithm works on the various populations of attributes. Artificial intelligence in computer science application, Genetic Algorithm (GA) enables to learn and discover the behaviour and process of natural evolution. It also helps to profound solution for problem and optimization. It comes under a wider class of evolutionary algorithms (EA) that provide solution to optimize various problems through various techniques induced by natural evolution like inheritance, selection, mutation, and crossover.

Algorithm

Step 1: Load the dataset of cardiac patients.

Step 2: Verify the dataset so that it may be used as an input for the GA-based computer - aided diagnostic algorithm.

Step 3: Analyse the fitness of all traits in the population to assist us generate better and more accurate findings.

Step 4: Assign the novel pbest value if the existing value of each attribute is better than the pbest value.

Step 5: Keep the preceding pbest value if the existing value of attributes is smaller than the pbest value. We will get the selected properties if these attributes match the constraints.

Step 6: Assign the pbest value as the best attribute to the gbest value. If the qualities do not meet the parameters, they will go through a selection process to find a superior genome.

Step7: The next step is to generate selected attributes through a combination of genetic operators: cross over (also called recombination), and mutation.

Step 8: The fitness value (W) and the best attribute are the final results of the genetic algorithm (S).

III. PCNN Algorithm

Input: Best attribute (S) and the fitness value (W) obtained from Genetic Algorithm

Output: Best attributes (Y)

From beginning to end the synaptic weights M and W, the Feeding and Linking inputs communicate with the neighbouring characteristics. The PCNN Mathematical Model is a mathematical form. The output of the PCNN and the Compartments is defined by the subsequent equations:

$$F_{ij\,[n]} = e^{\alpha_F \beta_n} F_{ij\,[n-1]} + S_{ij} + V_F \Sigma_{kl}\, M_{ijkl} Y_{kl}[n-1] \quad (1)$$

$$L_{ij\,[n]} = e^{\alpha_L \beta_n} L_{ij[n-1]} + V_L \Sigma_{kl}\, W_{ijkl} Y_{kl}[n-1] \quad (2)$$

$F_{ij\,[n]}$: Feedback input of (i,j);

$L_{ij[n]}$: The linking item's output;

S_{ij}: Input value at (i, j)location;

Y_{kl}: result from a preceding step [n 1];

β: Linking coefficient values;

$M_{ijkl and} W_{ijkl}$: $W = M$ Constant Gaussian weight functions with the distance;

The internal state of the neuron is created by combining the states of these two compartments, and its activity U (ij[n]) is computed using Equation (3):

$$U_{ij\,[n]} = F_{ij\,[n]}[1 + \beta L_{ij\,[n]}] \quad (3)$$

The inputs of the attributes are evaluated to a threshold, $U_{ij[n]}$ to generate the pulsating output, Y;

$$y = \begin{Bmatrix} 1, & U_{ij[n]} > T_{ij[n]} \\ 0, & otherwise \end{Bmatrix} \quad (4)$$

When the attributes (Y > T) are the threshold, the value of the attribute greatly increases. The threshold value then decreases until the qualities are restored. As a result, the threshold is fluid.

The procedure is outlined as follows:

$$T_{ij[n]} = e^{\alpha_\theta \beta_n} T_{ij[n-1]} + V_T Y_{ij}[n] \quad (5)$$

$Where T_{ij[n]}$: represents the dynamic threshold and

Where V_T is a large constant which is greater than the average value of U_{ij}

Here W is the genetic algorithms fitness value. And the S is the genetic algorithms best attribute.

IV. *Particles swarm optimization (PSO) algorithm*

Kennedy and Eberhard [6] [29] has developed a method called Particles swarm optimization (PSO) to optimize the emphasize on a social lbehaviour of animals or birds. It interconnects every particle in swarm optimization which informs the nature and to improve performance. PSO is also similar to GA which acts as an optimizer based on population arrived. Each solution in PSO is called a particle, and the entire population of solutions is called a swarm of particles. PSO algorithms are simple to build and have a high probability of achieving global optimal solutions.

A swarm of particles represents the population of candidate solutions in PSO. Every particle in a n–dimensional search space is a point. The current position pi depicts the swarm's I th particle and its current velocity vi. PSO aims to discover the best answer to the difficulty by relocating the particles as well as assessing the fitness of the fresh place.

A particle's location is updated by the Equation (6).

$$X_i(t+1) = P_i(t) + V_i(t) \quad (6)$$

A particle's location is updated by:

$$V_i(t+1) = V_i(t) + (C_1 \times rand\ () \times (P_i^{best} - P_i\ (t))) + (C_2 \times rand\ () \times (P_{gbest} - P_i\ (t))) \quad (7)$$

Where, $V_i(t+1)$, is the novel velocity for the i th particle.

The Weighting Coefficients for the Personal Best and Global Best location are C_1 and C_2 correspondingly.

P_i (t) Stands for the location of the I th particle at time t, while P_i^{best} is well known location of the i th particle

P_{gbest} is the swarm's well known location.

On this update Equation [7], the rand () Function generates uniformly random variables € [0, 1].

The combined procedure for attribute selection is presented in the form of pseudo code as below.

Algorithm: Attribute Selection using PCGPSONN

Input: Attributes (X) and its values

Output: Best attributes (S)

Create a random population of each Attribute and assign P to the Initial Population.

Using Equation (8)., determine the fitness f(X) of every attribute (X) in the population (P)

$$f(x) = f(x) = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (8)$$

Where x_i represents an attribute in X as well as N is the entire amount of attributes

Repeat the steps below to create a new population (NP) until the recent population (NP) is complete.

Choose qualities from a population P based on how fit they are. The qualities are chosen (W) if the fitness value is high, otherwise they are rejected..

Create a new attribute by crossing across the selected attributes with a crossover probability (CP). The new attribute is an identical replica of the selected attributes if no crossover was performed.

Mutate the selected property at each locus using a mutation probability (MP).

Add the attribute to a new population (NP).

For a subsequent run of an algorithm, use fresh generated population (NP).

Stop and return the best solution(S) with the current attributes if the end condition is met.

If not, proceed to step 2.

The fitness value (W) and a collection of best attribute values are the final results of the genetic algorithm (S).

Use the best attribute (S) and fitness value (W) as the PCNN's weight values and run Algorithm 2.

The best attribute (S) is used as the starting population of PSO in the PCNN Algorithm.

By relocating the particles with assess the fitness of the novel place, PSO discovers the best solution to the problem.

V. Heart Disease Prediction using Proposed technique

A prediction model using proposed feature selection technique is also developed in this work. The outline of the prediction model using proposed feature selection method is given in Figure 2. In the prediction model, there are two stages namely the training phase and the prediction phase are involved. The training phase mainly focuses on gathering knowledge on heart disease from the given data set for prediction. Therefore, it is called the training phase. At this stage, the numbers of attributes/features are reduced with the help of proposed PCGPSONN algorithm. And then the features are given to SVM algorithm to train the model. Then the prediction phase is proceeded.

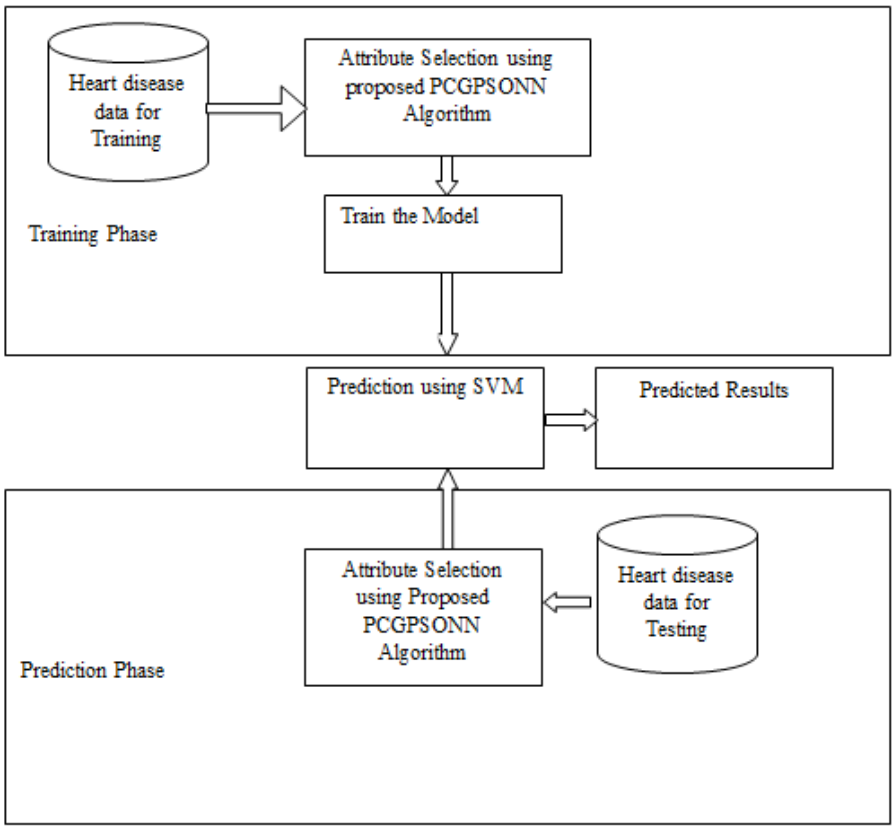

Figure 2.Outline of the prediction model using proposed feature selection algorithm

The prediction phase is same as the training phase. Throughout prediction, all steps of the training phase are repeated for the test data. At last, the predicted results will be obtained from prediction model. The patient dataset contains total of 24 attributes, and the proposed approach selects 14 out of 24 attributes. The total attributes and selected attributes are described in Table 2.

Table 2: Attribute of heart Disease Data Sets and selected attributes using PCGPSONN

S.No	Attribute Name	Selected Attributes
1	AGE	AGE
2	GENDER	
3	GTT	GTT
4	FASTING	FASTING
5	POST PRA	POST PRA
6	HBA1C	HBA1C
7	HEIGHT	
8.	WEIGHT	
9	BMI	BMI
10.	URICACID	URICACID
11.	TG	TG
12.	LDL-C	LDL-C
13.	VLDL	VLDL
14.	HDL-C	HDL-C
15.	SYSTOLIC	SYSTOLIC
16.	DIASTOLIC	DIASTOLIC
17.	HSCRP	HSCRP

18	SMOKERS	
19	CHAIN SMOKERS	
20	NON-SMOKERS	
21	ALCOHOL CONSUMPTION	
22	DIET	
23	PHYSICAL ACTIVITY	
24	WORK STRESS	

The PCGPSONN algorithm chooses an attribute based on its weight values. A population of 250 persons used real-valued representation instead of binary presentation since the attribute weight values were expressed using real-valued numbers rather than merely 0 and 1. Each particle was made up of seven separate attribute weight sets for a total of 24 attributes. Individuals in the initial population were selected using expert-set weights and machine learning techniques. The initial population is supplementary precisely determined. In the PCGPSONN, parent selection was done using a roulette wheel, and offspring creation was completed using a consistent crossover through separate recombination. The crossover was completed with an 80.0 percent chance of success, and the crossover points for each gene were chosen at random and independently. For the gene, mutation was done with a 1.0 percent likelihood and in a consistent way: a random rate was taken as of the assortment and placed as a new value in the present position. During the runs, elitism was also used to keep the best individual inside the population. During the evolution, we didn't desire to misplace the best-performing weight set. A survivor selection was utilised if the amount of people at the end of the generation was greater than 21. Individuals were ranked according to their classification accuracy, and those by the least accuracy were removed from the population. The PCGPSONN method was terminated following 20 generations or if the greatest categorization accuracy remained constant throughout a 10-generation period. Furthermore, if the population's members were all the same, the examination was completed.

SUPPORT VECTOR MACHINE

In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyse data and recognise patterns and are used for classification and regression analysis[26]. The Svm classifier is utilised in this research to predict heart disease based on a set of variables. A non-linear mapping is used to convert the original training data into a higher dimension. Within this new dimension, it looks for a linear best separation hyper plane. A hyper plane with the correct non-linear mapping to a sufficiently high dimension may always segregate data into two classes. To discover this hyper plane, the SVM employs support vectors and margins. SVMs accomplish classification tasks by minimising classification errors while increasing the margin between the two classes.

SVM maximises the geometric margin while minimising the empirical classification error. SVM stands for Maximum Margin Classifiers, and it uses the kernel method to efficiently conduct non-linear classification. An SVM model is an example of the occurrences as points in space that have been projected so that the examples of the multiple types are divided by a considerable margin gap. As data points of the form, given labelled training data

$$M=(x_1y_1),(x_2y_2),(x_3y_3)\ldots\ldots..(x_ny_n) \quad (9)$$

Where y_n=1/-1, is a constant that specifies the class that x n belongs to. n is the number of data samples. x n is a p-dimensional real vector for each x n. The SVM classifier accomplishes classification using an appropriate threshold value after mapping the input vectors into a decision value. The hyper plane will be divided into two portions to visualise the training data:

$$\text{Mapping: } w^T .x + b \quad (10)$$

w is a scalar and b is a p-dimensional weight vector. The separating hyper plane is perpendicular to the vector w points. The margin can be increased using these parameters. If b is not present, the hyper plane is forced to pass through the origin, limiting the solution.

EXPERIMENTAL RESULT ANALYSIS AND PERFORMANCE EVALUATION:

I. Data set

Three types of datasets were used in this research work they are Real-time data sets, UCI Machine Learning Repository Cleveland Clinic Foundation dataset and Hungarian dataset. Real-time dataset are live patient data which has been collected from hospitals, UCI and Hungarian datasets have been collected through resources available online [31]. In Real-time dataset 100 patient records with 24 attributes are available. By applying PCGPSONN algorithm, 14 appropriate attributes have been selected from the total 24 attributes for heart disease prediction. Similarly in UCI dataset 303 patient records were available with 14 attributes and 10 appropriate attributes were selected using the same algorithm. In Hungarian dataset 260 patient records were available with 7 attributes, among this 3 best attributes were selected using PCGPSONN algorithm and applied for heart disease prediction.

II. Performance Metrics

The results of the feature selection with the classification technique are evaluated using the confusion matrix as demonstrated in Table 3.

Table: 3 confusion matrix for heart disease prediction

Actual		Has heart disease	Does not have heart disease
Predicted	With heart disease	TRUE POSITIVE(TP)	FALSE POSITIVE(FP)
	Does not have heart disease	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)

True Positive (TP) is a term that refers to the fact that something is true (Patients with heart illness who have been diagnosed as having heart illness)

False negatives (FN) are a type of error that occurs when data is processed incorrectly (Patients with heart illness who have been diagnosed as not having heart illness)

True Negative (TN) is a term that refers to a negative image that has been (Patients without heart illness who have been diagnosed as not having heart illness)

False Positive (FP) is a term that refers to a situation where something appears to be (Patients without heart illness who have been diagnosed as having heart illness)

Numerous performance indicators are available to assess the performance of heart disease forecasting algorithms employing the SVM algorithm. The performance of the proposed feature selection technique with classifiers is shown in Table 4.

Table 4: Performance Metrics calculation

Performance Measure	Description	Formula
Detection Accuracy	The amount of right forecasts divided by the total amount of guesses in the dataset is accuracy.	$\frac{TP + TN}{TP + FP + TN + FN}$
Precision Rate	The amount of correct positive forecasts divided by the total amount of positive forecasts yields precision.	$\frac{TP}{TP + FP}$
Recall Rate	The amount of correct positive forecasts divided by the total amount of positives yields the recall rate.	$\frac{TP}{TP + FN}$
Specificity	The amount of valid negative predictions divided by the total amount of negatives yields the specificity (SP).	$\frac{TN}{(TN + FP)}$
F1 Score	The weighted average of Precision and Recall is the F1 Score.	$2 * \frac{PrecisionRate * Recallrate}{PrecisionRate + Recallrate}$
Error Rate	The error rate is derived by dividing the total amount of inaccurate forecasts by the total amount of forecasts in the dataset.	$\frac{\text{amount of samples of Fasely Detected samples}}{\text{Total amount of samples}}$

III. . Performance Analysis of Proposed algorithm for heart disease prediction using SVM Classifier

In this experiment, we have analyzed all the attributes of the data sets based on the contribution of each an arrived the appropriate attribute using PCGPSONN algorithm resulting 96.29%. Subsequently the same analysis was applied using GPSO and GPCNN algorithms resulted 93.12% and 94.25% respectively. From this performance analysis it has been identified that PCGOSONN algorithm has to be applied to get the highest accuracy. Table 5 lists the accuracy analysis of PCGPSONN.

Table 5: Performance of the proposed feature selection technique with classifiers

Optimization algorithm	Accuracy (%)	Precision (%)	Recall (%)	Error Rate (%)
GPSO	93.12	76.12	90.11	6.86
GPCNN	94.25	72.34	91.35	5.77
PCGPSONN	**96.29**	**77.27**	**92.30**	**3.71**

IV. Performance Analysis of Proposed algorithm in comparison with existing feature selection techniques

The performance and results of existing feature selection techniques proposed by various authors have been analyzed and it has been found that all the authors have used only UCI datasets and arrived the prediction accuracy. We have applied **PCGPSONN+ SVM** algorithm in real-time data sets resulted 96.29% accuracy. Classifiers and analysis with various feature selection algorithms have been abridged in Table 6.

Table 6: Performance of Existing feature selection techniques and proposed PCGPSONN with SVM classifier

Sl. No.	**Reference**	**Technique(s)**	**Prediction accuracy (%)**
1	Youness et al. [14]	FCBF, PSO and ACO +SVM	84.07
2.	Tan et al. [24]	Hybrid Technique (GA + SVM)	83.55
3.	Ismail Babaoglu [2]	Binary PSO GA–FST SVM	81.46
4.	InduYekkala [28]	PSO + Random Forrest	90.37
5.	Le minh Hung [15]	Infinite Latent Feature Selection+ SVM	90.65
6	BeantKaur et al [12]	Genetic algorithm with SVM	73.46
7.	Zeinab Arabasadi et al. [1]	Genetic algorithm+ Neural Network	93.85
8.	Proposed	**PCGPSONN+SVM**	**96.29**

CONCLUSION

In this research work we have developed a technique namely **PCGPSONN.** A framework for heart disease prediction is also developed for heart disease prediction .We obtained an accuracy of 96.29% with the proposed feature selection technique with usinga real-time datasets. Subsequently this algorithm has been applied in UCI and Hungarian datasets as well resulting highest accuracy. The existing feature selection techniques of various authors have also been analyzed and it has been observed that they have used only UCI data to get accuracy and the work we have done has highest accuracy by using real-time datasets. To enhance the accuracy of the existing prediction techniques, researchers can carry a further study through various machine learning technique and novel feature selection methods be able to be introduced to gain better heart disease prediction results.

REFERENCES

1. Alkeshuosh, Azhar Hussein, et al. "Using PSO algorithm for producing best rules in diagnosis of heart disease." 2017 international conference on computer and applications (ICCA). IEEE, 2017.
2. Babaoglu, İsmail, OğuzFindik, and ErkanÜlker. "A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine." Expert Systems with Applications 37.4 (2010): 3177-3183.
3. Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019, January). Improving heart disease prediction using feature selection approaches. In 2019 16th international bhurban conference on applied sciences and technology (IBCAST) (pp. 619-623). IEEE.
4. Dangare, Chaitrali S., and Sulabha S. Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10 (2012): 44-48.
5. Eberhart, Russell, and James Kennedy. "A new optimizer using particle swarm theory." MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science. Ieee, 1995.
6. El Mountassir, Mahjoub, et al. "Feature selection techniques for identifying the most relevant damage indices in SHM using Guided Waves." Proceedings of the 8th European Workshop On Structural Health Monitoring, EWSHM, Bilbao, Spain. 2016.
7. Fatima, Meherwar, and Maruf Pasha. "Survey of machine learning algorithms for disease diagnostic." Journal of Intelligent Learning Systems and Applications 9.01 (2017): 1.
8. Feshki, Majid Ghonji, and Omid SojoodiShijani. "Improving the heart disease diagnosis by evolutionary algorithm of PSO and Feed Forward Neural Network." 2016 Artificial Intelligence and Robotics (IRANOPEN). IEEE, 2016.
9. Gavhane, A., Kokkula, G., Pandya, I., &Devadkar, K. (2018, March). Prediction of heart disease using machine learning. In 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1275-1278). IEEE.
10. GAZELOĞLU, Cengiz. "Prediction of heart disease by classifying with feature selection and machine learning methods." (2020).
11. Javeed, Ashir, et al. "Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification." Mobile Information Systems 2020 (2020).
12. Kaur, Beant, and Williamjeet Singh. "Analysis of heart attack prediction system using genetic algorithm." Int. J. Adv. Technol. Eng. Sci 3 (2015): 87-94.
13. Kelwade, J. P., and S. S. Salankar. "Radial basis function neural network for prediction of cardiac arrhythmias based on heart rate time series." 2016 IEEE First International Conference on Control, Measurement and Instrumentation (CMI). IEEE, 2016.
14. Khourdifi, Youness, and Mohamed Bahaj. "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization." International Journal of Intelligent Engineering & Systems 12.1 (2019): 242-252.
15. Le, Hung Minh, ToanDinh Tran, and L. A. N. G. Van Tran. "Automatic heart disease prediction using feature selection and data mining technique." Journal of Computer Science and Cybernetics 34.1 (2018): 33-48.
16. Liu, Xiao, et al. "A hybrid classification system for heart disease diagnosis based on the RFRS method." Computational and mathematical methods in medicine 2017 (2017).
17. Malav, Amita, KalyaniKadam, and Pooja Kamat. "Prediction of heart disease using k-means and artificial neural network as hybrid approach to improve accuracy." International Journal of Engineering and Technology 9.4 (2017): 3081-3085.
18. Mohan, Senthilkumar, ChandrasegarThirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE Access 7 (2019): 81542-81554.
19. Otoom, Ahmed Fawzi, et al. "Effective diagnosis and monitoring of heart disease." International Journal of Software Engineering and Its Applications 9.1 (2015): 143-156.
20. Prakash, S., K. Sangeetha, and N. Ramkumar. "An optimal criterion feature selection method for prediction and effective analysis of heart disease." Cluster Computing (2018): 1-7
21. Reddy, N. S. C., Nee, S. S., Min, L. Z., & Ying, C. X. (2019). Classification and feature selection approaches by machine learning techniques: Heart disease prediction. International Journal of Innovative Computing, 9(1).Spencer, Robinson, et al. "Exploring feature selection and classification methods for predicting heart disease." Digital health 6 (2020): 2055207620914777.
22. Sharma, Swati, and Dr Sukhvir Singh. "Heart Disease Diagnosis using Genetic and Particle Swarm Optimization." International Journal of Engineering Research & Technology 3.8 (2014): 1499-1503.

23. Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. Digital health, 6, 2055207620914777.

24. Tan, Kay Chen, et al. "A hybrid evolutionary algorithm for attribute selection in data mining." Expert Systems with Applications 36.4 (2009): 8616-8630.

25. Thomas, J., and R. Theresa Princy. "Human heart disease prediction system using data mining techniques." 2016 international conference on circuit, power and computing technologies (ICCPCT). IEEE, 2016.

26. Wang, Tinghua, et al. "Feature selection for SVM via optimization of kernel polarization with Gaussian ARD kernels." Expert Systems with Applications 37.9 (2010): 6663-6668.

27. World Health Organization. The world health report 2000: health systems: improving performance. World Health Organization, 2000.

28. Yekkala, Indu, Sunanda Dixit, and M. A. Jabbar. "Prediction of heart disease using ensemble learning and Particle Swarm Optimization." 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon). IEEE, 2017.

29. Zhang, Chunkai, and Huihe Shao. "Particle swarm optimisation in feedforward neural network." Artificial Neural Networks in Medicine and Biology. Springer, London, 2000. 327-332.

30. https://www.usfhealthonline.com/resources/key-concepts/data-mining-in-healthcare/

31. http://archive.ics.uci.edu/ml/machine-learning-databases/heart-isease/heartdisease.names.