An Improved Auto Categorical PSO with ML for Heart Disease Prediction

Animesh Kumar Dubey
Department of Computer Science and Engineering
JK Lakshmipat University
Jaipur, India
animeshdubey@jklu.edu.in

Amit Kumar Sinhal
Department of Computer Science and Engineering
JK Lakshmipat University
Jaipur, India
amit.sinhal@jklu.edu.in

Richa Sharma
Department of Science and Liberal Arts
JK Lakshmipat University
Jaipur, India
richasharma@jklu.edu.in

Received: 17 February 2022 | Revised: 4 March 2022 | Accepted: 27 March 2022

Abstract-Cardiovascular or heart diseases consist a global major health concern. Cardiovascular diseases have the highest mortality rate worldwide, and the death rate increases with age, but an accurate prognosis at an early stage may increase the chances of surviving. In this paper, a combined approach, based on Machine Learning (ML) with an optimization method for the prediction of heart diseases is proposed. For this, the Improved Auto Categorical Particle Swarm Optimization (IACPSO) method was utilized to pick an optimum set of features, while ML methods were used for data categorization. Three heart disease datasets were taken from the UCI ML library for testing: Cleveland, Statlog, and Hungarian. The proposed model was assessed for different performance parameters. The results indicated that, with 98% accuracy, Logistic Regression (LR) and Support Vector Machine by Grid Search (SVMGS) performed better for the Statlog, SVMGS outperformed on the Cleveland, while the LR, Random Forest (RF), Support Vector Machine (SVM), and SVMGS performed better with 97% accuracy on the Hungarian dataset. The outcomes were improved by 3 to 33% in terms of performance parameters when ML was applied with IACPSO.

Keywords-SVMGS; IACPSO; KNN; LR

I. INTRODUCTION

Globally, many people suffer from heart diseases [1]. In 1990, there were 24 million fatalities related to heart disease in the United States, by 2010 that number had risen to 38 million, a 59% increase [2]. According to current forecasts, India will have the world's highest prevalence of cardiovascular diseases and will soon overtake the rest of the world [3]. Heart disorders are responsible for almost 4 million fatalities in Europe and 1.9 million deaths in the EU [4]. In Africa, heart diseases are the leading cause of death among persons over the age of 35 [5]. Massive quantities of data on heart illnesses are collected from hospitals all around the world, which can be used manually to quantify disease rates. However, the data so far have not been efficiently translated to correlate with disease risk and symptoms [6]. Cardiovascular disorder is accompanied by common symptoms of chest tightness, loss of body strength, and swollen legs [7]. Health history examination, clinical test reports, and associated symptoms are usually used by the doctors for diagnosis. But the obtained results by this method are not always accurate, whereas they are costly and difficult to computationally analyze [8]. Researchers have tried to come up with an efficient technique to detect heart diseases since the current diagnostic approaches for heart disease are not very effective in identifying the early stages [9]. It has been reported from different methodological approaches that a combination of ML and optimization methods may be effective in predicting early-stages of heart diseases [10]. Appropriate data are needed for training and testing in ML predictive models. We can increase the performance of ML algorithms by optimal dataset balancing for training and testing [11].

Feature selection is capable of reducing dimensionality, increasing efficiency, and enhancing classification accuracy [12-14]. The data comprising several dimensions may create trouble in feature selection. According to [15-18], classification and clustering methods of ML were proven to be more effective in terms of accuracy rates. Several feature selection evaluation metrics were investigated in [19], to improve the computational efficiency of ML algorithms as well as to discuss the unexpected problems of feature selection. Various data mining and combinations of data mining with optimization algorithms have been proposed as a means of detecting heart diseases. Ant Colony Optimization (ACO) was applied to select an effective subset from a large training set with improved accuracy in [20]. In [21], a combination of optimization and data mining was proposed, based on Glowworm Swarm Optimization (GSO) with k-means to improve the accuracy of image classification. For omics data

Corresponding author: Animesh Kumar Dubey classification, a Particle Swarm Optimization (PSO)-based model was developed in [22] and claimed good accuracy. In [23], a hybrid structure, based on PSO and grid search was proposed to predict heart diseases with 95.95% accuracy. Authors in [24], suggested a combined approach of PSO with Support Vector Machine (SVM) and Convolutional Neural Networks (CNNs). The highest accuracy, 98%, was achieved by PSO with CNN. In [25], the performance of swarm optimization algorithms Artificial Bee Colony (ABC), PSO, and ACO to an Artificial Neural Network (ANN) was studied, and PSO was determined to be the most effective. A Hhybrid Genetic Algorithm (HGA) with k-means was implemented in [26] for classifying and predicting heart diseases with 94.06% accuracy. In [27], ANNs with PCA and PSO were used to predict cardiac diseases, with an accuracy of 98%. In [28], PSO algorithm was proposed for dimension reduction, and several classification methods were used to diagnose heart diseases. Utilizing the dataset from the UCI library, Naïve Bayes (NB), Decision Tree (DT), and k-Nearest Neighbour (KNN) models were applied in [29] to predict heart diseases. In [30], PSO with feedforward backpropagation ANN were implemented to diagnose heart diseases, and achieved 91.94% accuracy. NB with a Genetic Algorithm (GA) were used to eliminate unnecessary information in [31] achieving high accuracy. To deal with the issues related to overfitting and underfitting, as well as selecting attributes, a deep ANN was implemented in [32], and the achieved precision was 93.33%. Similarly, Random Search Algorithm (RSA) for feature selection and Random Forest (RF) for classification were used in [33]. Several DT-based methods along with PSO were used in [34], to identify heart disease occurrence and the highest precision was obtained using a bagged tree with PSO. SVM with a multiclass approach of ML was applied in [35], to detect apple fruit diseases. In [36], crow search with deep learning method were used for the prediction of Parkinson's disease with 96% accuracy. For the detection of various plant diseases, ML algorithms were used in [37]. Various unsupervised learning and optimization methods were used in [38-40] for the prediction of heart diseases.

Most researchers used supervised and unsupervised ML methods and swarm intelligence-based optimization methods such as ACO, GSO, PSO, etc., in conjunction with ML. But their approaches were not stable in handling real-time situations. Hence, there is a need for an automated approach, which can generate the optimal solution based on the current situation. The present research work proposes a combined approach, including ML algorithms with optimization to predict heart diseases. ML methods, such as Logistic Regression (LR), DT, SVM, SVM by Grid Search (SVMGS), RF, KNN, and NB, are used, and the Improved Auto Categorical PSO (IACPSO) method is applied for selecting an optimized set of features. The major objectives of this research are:

- For PSO, to create an automated approach to the selection of the optimal value of control parameters at each iteration.
- To analyze the impact of ML algorithms on different performance parameters in the prediction of cardiovascular diseases.
- To evaluate the combined impact of ML algorithms with optimization for heart disease prediction.

II. Materials and Methods

Three heart disease datasets were taken from the UCI ML library for testing: Cleveland, Statlog, and Hungarian [41]. The datasets consist of a total of 76 attributes, but only 14 relevant attributes including the attributes preferred in most published experiments [42]. Predicted trait values were represented by A and P indicating the absence and presence of heart disease respectively. ML algorithms such as LR [43, 44], DT [45, 46], RF [47], KNN [48], SVM [49, 50], and NB [45, 48], were used for prediction and analysis. IACPSO was used to select an optimal set of features.

A. IACPSO

PSO is a search-based stochastic optimization technique based on population. The particles, which are potential solutions in PSO, follow the current optimum particles through the problem space [51]. The performance of PSO depends on three control parameters which are the inertial weight (w), the acceleration coefficients (C_1 and C_2), and random numbers (R_1 and R_2) [52]. The inertial weight is used for maintaining the effect of convergence and diversity. A large value of w indicates better global exploration, while a small value works on exploitation. Unbalanced values containing the parameters can hurt the results, such that if we take low values for C_1, it tends to acquire a smooth particle trajectory and abrupt movements. Similarly, if C_1 is much greater than C_2, it tends to excessive wander and cause premature convergence [53]. Large inertial weight tends to global search ability and small inertial weight leads to increase in local search power. By dynamic changing w, the acceleration coefficients, efficiently explore the search space [52]. Therefore, in the proposed IACPSO, the control parameters are updated automatically based on the number of particles as well as balancing them at each iteration. The Steps included in IACPSO are given below:

Step I: Initialize the Particle size ($P_0, P_1, P_2, \ldots\ldots\ldots, P_n$) in the D-dimensional space.

Step II: Initialize the velocity $V_x^d(0)$, where $x \in \{P_0, P_1, P_2, \ldots\ldots\ldots, P_n\}$, $d \in \{0, 1, \ldots., D\}$. Calculate the velocity of a particle at the mth iteration by using (1):

$$V_x^d(m) = wV_x^d(m-1) + \emptyset_1\big(L_x(m) - P_x^d(m-1)\big) + \emptyset_2\left(G(m) - P_x^d(m-1)\right) \quad (1)$$

where w is the inertial weight, $\emptyset_1 = R_1C_1$ (local accelerations), $\emptyset_2 = R_2C_2$ (global accelerations), C_1 and C_2 are the acceleration coefficient, and R_1 and R_2 are random numbers. P_x^d is the position of the particles, $L_x(m)$ and $G(m)$ are the local and global best positions. For each iteration, control parameter values have been chosen automatically on the basis of the number of swarm particles (n), which are given in (2-5). For the value of $\emptyset_1$ and $\emptyset_2$, generate the $n/8$ random number between 0 to 2 for the selection of C_1 and C_2.

$$C_1 = \{C_1^1, C_1^2, C_1^3, \ldots\ldots, C_1^{\frac{P_n}{8}} \text{ where } 0 \le C_1 \le 2\} \quad (2)$$

$$C_2 = \{C_2^1, C_2^2, C_2^3, \ldots\ldots, C_2^{\frac{P_n}{8}} \text{ where } 0 \le C_2 \le 2\} \quad (3)$$

$$\emptyset_1 = \sum_{n=1}^{\frac{P_n}{8}} C_1 . R_1 \quad (4)$$

$$\emptyset_2 = \sum_{n=1}^{\frac{P_n}{8}} C_2 . R_2 \quad (5)$$

Then, as per (6), divide the value of $\emptyset_1$ and $\emptyset_2$ into three categories: Low (*L*), Mid (*M*), and High (*H*).

$$\emptyset_1, \emptyset_2 = \begin{cases} \text{L}, & \text{if } (0 \le \text{C}_1, \text{C}_2 \le 0.8) \\ \text{M}, & \text{else if } (0.9 \le \text{C}_1, \text{C}_2 \le 1.2) \\ \text{H}, & \text{else } (1.3 \le \text{C}_1, \text{C}_2 \le 2) \end{cases} \quad (6)$$

Similarly, for the value of *w*, select three values between 0.4 and 0.9 and categorize those into *L*, *M* and *H*, in which *L* is close to 0.4, *M* is nearer to the mean of the other two values, and *H* is close to 0.9. For each iteration select the value of *w* as per the following conditions:

(i) If ($\emptyset_1$ & $\emptyset_2 \in L$) then consider *w* = *H* (value)

(ii) If ($\emptyset_1$ & $\emptyset_2 \in M$) then consider *w* = *M* (value)

(iii) If ($\emptyset_1$ & $\emptyset_2 \in$ H) then consider *w* = *L* (value)

So, the updated velocity at each iteration is:

$$V_x^d(m) = w(L, W, H) V_x^d(m-1) + \sum_{n=1}^{\frac{P_n}{8}} C_1 . R_1 \left(L_x(m) - P_x^d(m-1)\right) + \sum_{n=1}^{\frac{P_n}{8}} C_2 . R_2 \left(G(m) - P_x^d(m-1)\right) \quad (7)$$

Step III: Calculate position of the particles, given in (8):

$$P_x^d(m) = P_x^d(m-1) + V_x^d(m) \quad (8)$$

By using (8), we get *n*/8 number of positions for each particle, then we proceed to step IV.

Step IV: Calculate the current fitness function $\text{Ɣ}(P)$. Based on the *n*/8 position of particles, calculate the fitness function, given in (9) and (10), and choose the best one based on Minimization or Maximization.

$$L_x(m) = P_x(m_L) : \text{Ɣ}\left(P_x(m_L)\right) = \frac{\min}{\max_{0 \le k \le m}} \text{Ɣ}\left(P_x(k)\right) \quad (9)$$

$$G(m) = L_{xb}(m) : \text{Ɣ}\left(L_{xb}(m)\right) = \min/\max_{0 \le x \le n} \text{Ɣ}(L_x(m)) \quad (10)$$

Step V: On the basis of the fitness value, update $L_x(m)$ and $G(m)$.

III. EXPERIMENTAL RESULTS AND ANALYSIS

Figure 1 shows the flow chart of the suggested methodology. Cleveland, Statlog, and Hungarian datasets of heart diseases were considered for the evaluation of the proposed approach. Experiments were run using an x64-based processor with Windows 10 OS and an Intel (R) Core (TM) i5-7200 CPU @ 2.50GHz. The analysis and visual presentation were done using Python and Java7. The experimental results and analysis on heart diseases datasets, based on ML algorithms like LR, SVM, DT, SVMGS, RF, NB, and KNN and the IACPSO optimization method are reported.

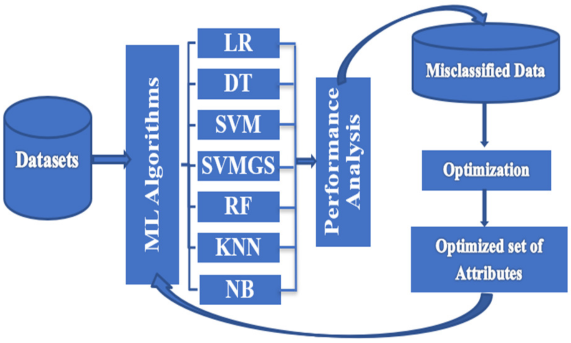

Fig. 1. Flowchart of the proposed methodology.

A. Result Analysis Based on ML Algorithms

We assessed the effectiveness of the ML models by using parameters such as accuracy (AC), precision (PR), Matthews Correlation Coefficient (MCC), sensitivity (SV), and F-score (FS) [11]. Table I presents the AC, PR, SV, FS, and MCC values, where PR, SV, and FS were either A or P. Here, 25% of the data were used for testing and the rest for training. For KNN, the considered values of k were 11, 17, and 15 for Cleveland, Statlog, and Hungarian datasets respectively. According to the comparative analysis shown in Figure 2, SVMGS outperformed the other methods in all aspects and achieved accuracy of 89% for Cleveland and Hungarian datasets, while for Statlog dataset, NB and LR showed better accuracy of 91%.

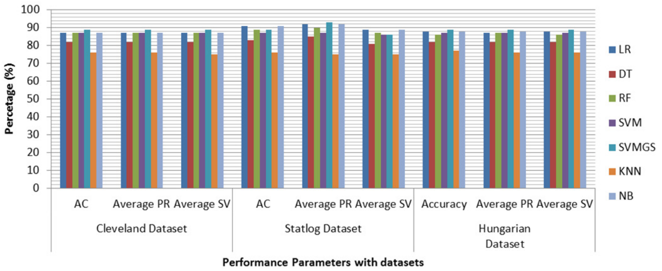

Fig. 2. Comparative analysis of ML algorithms in terms of AC, average PR, and average SV.

B. Result Analysis Based on IACPSO with ML Methods

IACPSO was used for feature selection. The selection of features depend upon their ranks, their values are either 0 or 1, with 0 representing the rejection of a feature and 1 representing its selection. The number of features represents the solution size for each data set. For the optimization process, the selected features along with the performance of the classifier were taken into account. The fitness function was

$$\alpha C(E) + \beta\ (|SF|)/(|TF|) \quad (11)$$

where *C*(*E*) is the misclassification rate, |*SF*| shows the number of the selected features, |*TF*| represents the total features in the data set, and α belongs to [1,0], β = (1-α). The value of *α* and *β* were taken from [54, 55]. For experimentation, the parameters values are: population size=12, number of iterations=100, dimension=7, *w*= 0.4 to 0.9 and *C1* and *C2*= 0 to 2.

TABLE I. EVALUATION RESULTS BASED ON THE PERFORMANCE PARAMETERS OF THE ML METHODS

Classification methods			**LR**	**DT**	**RF**	**SVM**	**SVMGS**	**KNN**	**NB**
Cleveland dataset	AC		0.87	0.82	0.87	0.87	0.89	0.76	0.87
	PR	A	0.86	0.76	0.89	0.86	0.89	0.77	0.84
		P	0.88	0.89	0.85	0.88	0.88	0.74	0.90
	SV	A	0.86	0.90	0.83	0.86	0.86	0.69	0.90
		P	0.88	0.75	0.91	0.88	0.91	0.81	0.84
	FS	A	0.86	0.83	0.86	0.86	0.88	0.73	0.87
		P	0.88	0.81	0.88	0.88	0.89	0.78	0.87
	MCC		0.74	0.65	0.71	0.74	0.77	0.51	0.74
Statlog dataset	AC		0.91	0.83	0.89	0.87	0.89	0.76	0.91
	PR	A	0.89	0.82	0.86	0.88	0.85	0.81	0.89
		P	0.94	0.88	0.94	0.85	1.00	0.68	0.94
	SV	A	0.97	0.94	0.97	0.91	1.00	0.79	0.97
		P	0.81	0.67	0.76	0.81	0.71	0.71	0.81
	FS	A	0.93	0.87	0.91	0.90	0.92	0.80	0.93
		P	0.87	0.76	0.84	0.83	0.83	0.70	0.87
	MCC		0.81	0.65	0.61	0.73	0.78	0.50	0.81
Hungarian dataset	AC		0.88	0.82	0.86	0.87	0.89	0.77	0.88
	PR	A	0.87	0.77	0.88	0.85	0.90	0.77	0.85
		P	0.88	0.88	0.85	0.89	0.87	0.75	0.91
	SV	A	0.87	0.89	0.82	0.85	0.86	0.70	0.91
		P	0.88	0.75	0.90	0.89	0.91	0.82	0.85
	FS	A	0.87	0.83	0.85	0.85	0.88	0.74	0.88
		P	0.88	0.81	0.88	0.89	0.89	0.78	0.87
	MCC		0.77	0.65	0.69	0.74	0.77	0.53	0.76

TABLE II. ACCURACY OBTAINED BASED ON ML METHODS WITH IACPSO

ML methods with IACPSO			**LR+IACPSO**	**DT+IACPSO**	**RF+IACPSO**	**SVM+IACPSO**	**SVMGS+IACPSO**	**KNN+IACPSO**	**NB+IACPSO**
Cleveland	AC		0.96	0.92	0.97	0.97	0.98	0.92	0.92
	PR	A	0.97	0.90	0.98	0.98	0.98	0.92	0.93
		P	0.96	0.95	0.96	0.96	0.95	0.91	0.92
	SV	A	0.97	0.96	0.97	0.97	0.96	0.90	0.94
		P	0.96	0.91	0.98	0.98	0.96	0.91	0.90
	FS	A	0.98	0.95	0.99	0.99	0.97	0.90	0.90
		P	0.95	0.90	0.95	0.95	0.96	0.94	0.91
	MCC		0.89	0.83	0.90	0.90	0.90	0.84	0.83
Statlog	AC		0.98	0.93	0.97	0.97	0.98	0.92	0.94
	PR	A	0.99	0.94	0.97	0.97	0.98	0.91	0.91
		P	0.98	0.97	0.98	0.98	1.00	0.93	0.98
	SV	A	0.99	0.98	0.98	0.98	1.00	0.92	0.98
		P	0.95	0.90	0.97	0.97	0.96	0.93	0.87
	FS	A	0.98	0.98	0.98	0.98	0.97	0.91	0.94
		P	0.99	0.94	0.97	0.97	0.96	0.88	0.90
	MCC		0.95	0.83	0.89	0.89	0.90	0.81	0.85
Hungarian	AC		0.97	0.93	0.97	0.97	0.97	0.91	0.92
	PR	A	0.96	0.91	0.96	0.96	0.98	0.90	0.91
		P	0.99	0.94	0.99	0.99	0.95	0.91	0.93
	SV	A	0.98	0.94	0.96	0.96	1.00	0.91	0.94
		P	0.97	0.92	0.99	0.99	0.94	0.90	0.90
	FS	A	0.99	0.93	0.98	0.98	0.98	0.89	0.91
		P	0.96	0.93	0.97	0.97	0.95	0.89	0.91
	MCC		0.92	0.82	0.88	0.88	0.90	0.80	0.82

The percentage of data based on misclassification rates (shown in Figure 3) were considered for the experiment. Based on the IACPSO, only 7 of the 14 features were chosen for testing, which were thalassemia, chest pain, number of major vessels, ST anxiety exercise-induced relative to rest, exercise-induced angina, maximal heart rate achieved, and exercise-induced angina. Table II presents the AC, PR, SV, FS, and MCC values, based on the combined performance of ML methods and IACPSO. With 98% accuracy, SVMGS outperformed the other methods in the Cleveland dataset. SVMGS and LR with 98% accuracy outscored RF, DT, KNN, NB, and SVM on the Statlog dataset. LR, RF, SVM, and SVMGS with a 97% accuracy outperformed KNN, NB and DT on the Hungarian dataset. For Cleveland, with an optimal set of features, the highest MCC was achieved by SVMGS, SVM, and RF, while LR did better in terms of MCC in Statlog and Hungarian. Figure 4 compares the ML algorithms with IACPSO results based on AC, average PR, and SV. For the

Cleveland dataset, LR, RF, SVM, and SVMGS performed better in terms of average PR, RF and SVM in terms of average SV. LR and SVMGS achieved the highest average PR of 99%, whereas the highest average SV was achieved by RF, SVM, and SVMGS for the Statlog dataset. Regarding the Hungarian dataset, LR, RF, and SVM, with 98% of average PR and SV, performed better than DT, SVMGS, KNN, and NB. With 97%, LR, RF, SVM, and SVMGS outperformed in terms of AC.

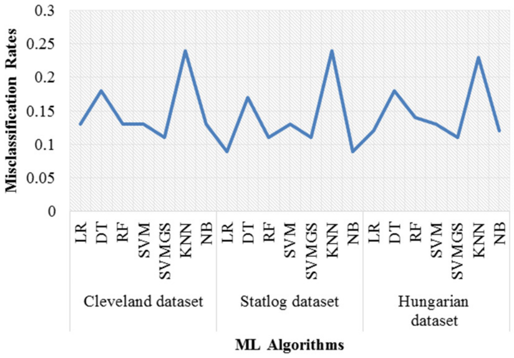

Fig. 3. Comparison of the misclassification rates of ML algorithms.

IV. Discussion

The performance of ML algorithms with and without optimization was examined in this paper. An optimization algorithm (IACPSO) was used for the selection of features to improve accuracy in less time, using the optimum value of the acceleration coefficients in each iteration. In addition, these values were also linked to the graded inertial load, to balance exploration and exploitation. Table III compares the results, based on ML techniques with and without optimization. Figure 5 shows the improvement rates of performance parameters when combined approach (ML algorithms + IACPSO) were used. Outcomes were improved by 3 to 33% in terms of performance parameters when ML was applied with IACPSO. As per Table III and Figure 5, the major findings were as follows:

- When LR was applied separately, the achieved AC was 87% (Cleveland), 91% (Statlog), and 88% (Hungarian) but in the combined methodology of LR with IACPSO, the AC increased by 9% for Cleveland and Hungarian and 7% for Statlog.
- When the DT was applied separately, AC was 82% for Cleveland and Hungarian, and 83% for Statlog, but in the combined methodology of DT with the IACPSO, the AC increased by 10% for Cleveland and Statlog and 11% for Hungarian.
- In the case of RF with IACPSO, AC was 10%, 8%, and 11% higher for Cleveland, Statlog, and Hungarian than for separately applied RF.
- For SVM, the obtained AC with the IACPSO in Cleveland, Statlog and Hungarian was 97%, which was 10% higher than when using only SVM. There was an increase in AC of 1% in Cleveland and Statlog when IACPSO was used with SVMGS.
- For KNN with IACPSO, an increase of 16% in AC occurred for Cleveland and Statlog. A large difference was found in the improvement percentage in the value of MCC, being 33% in Cleveland, 31% in Statlog, and 27% in Hungarian.
- For NB, improvement after IACPSO was 3 to 5% in terms of AC, PR, SV, and FS.

For all the aspects of performance used in the currrent research, ML methods with IACPSO were proven to be superior. ML algorithms were used for the classification and IACPSO method was applied for the selection of effective features. This type of combined approach gave better models for the early prediction of cardiovascular diseases.

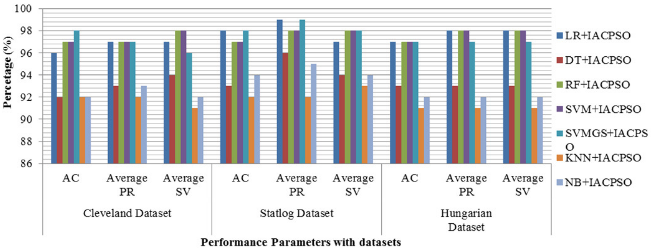

Fig. 4. Comparison of ML methods with IACPSO in terms of AC, average PR, and average SV.

TABLE III. RESULT COMPARISON OF ML METHODS WITH AND WITHOUT IACPSO

Datasets	Cleveland					Statlog					Hungarian				
ML Methods	**AC**	**PR**	**SV**	**FS**	**MCC**	**AC**	**PR**	**SV**	**FS**	**MCC**	**AC**	**PR**	**SV**	**FS**	**MCC**
LR	87	87	87	87	74	91	92	89	90	81	88	88	88	88	77
LR+IACPSO	96	97	97	97	89	98	99	97	99	95	97	98	98	98	92
DT	82	83	83	82	65	83	85	81	82	65	82	83	82	82	65
DT+IACPSO	92	93	94	93	83	93	96	94	96	83	93	93	94	93	92
RF	87	87	87	87	71	89	90	87	88	61	86	87	86	87	69
RF+IACPSO	97	97	98	97	90	97	98	98	98	89	97	98	98	98	88
SVM	87	87	87	87	74	87	87	86	87	73	87	87	87	87	74
SVM+IACPSO	97	97	98	97	90	97	98	98	98	89	97	98	98	98	88
SVMGS	89	89	89	89	77	89	93	86	88	78	89	89	89	89	77
SVMGS+IACPSO	98	97	96	97	90	98	99	98	97	90	97	97	97	97	90
KNN	76	76	75	76	51	76	73	75	75	50	77	76	76	76	53
KNN+IACPSO	92	92	91	92	84	92	92	93	90	81	91	91	91	89	80
NB	87	87	87	87	74	91	92	89	90	81	88	88	88	88	76
NB+IACPSO	92	93	92	91	83	94	95	94	92	85	92	92	92	91	82

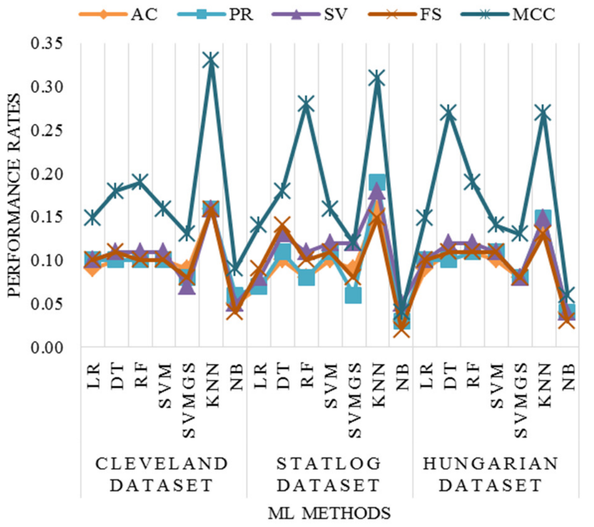

Fig. 5. Improved performance rates of ML algorithms due to IACPSO.

V. Conclusion

In this paper, a combined approach, based on ML algorithms and optimization is proposed to predict heart diseases at an early stage using the history of the patients. ML algorithms, such as DT, LR, SVM, SVMGS, NB, KNN, and RF were used and IACPSO was used for optimization. In IACPSO, the optimum value of control parameters was used, which helped in proper exploration and exploitation. In each iteration, $P_n/8$ number of solutions was generated in terms of local and global best, according to the objective function. The best solution was used for the next iteration. The proposed approach was investigated on Cleveland, Statlog, and Hungarian datasets and the evaluation was performed based on AC, PR, SV, FS, and MCC. The ML algorithms were compared and assessed with and without optimization, and it was found that the former were superior in all parameters. The proposed ML approach with an optimized set of features helped in predicting cardiovascular diseases and yielded better predictive results. In the future, this work can be repeated with more parameters, with real or primary datasets and various other threshold mechanisms towards the use of attributes in detecting different diseases.

References

[1] A. L. Bui, T. B. Horwich, and G. C. Fonarow, "Epidemiology and risk profile of heart failure," *Nature Reviews Cardiology*, vol. 8, no. 1, pp. 30–41, Jan. 2011, https://doi.org/10.1038/nrcardio.2010.165.

[2] D. Prabhakaran, P. Jeemon, and A. Roy, "Cardiovascular Diseases in India," *Circulation*, vol. 133, no. 16, pp. 1605–1620, Apr. 2016, https://doi.org/10.1161/CIRCULATIONAHA.114.008729.

[3] "Alarming Statistics from India - Neo CarDiabCare Heartlicare..." http://neocardiabcare.com/alarming-statistics-india.htm (accessed Mar. 31, 2022).

[4] E. Wilkins *et al.*, *European Cardiovascular Disease Statistics*. Brussels, Belgium: European Heart Network, 2017.

[5] H. Ouyang, "Africa's Top Health Challenge: Cardiovascular Disease," *The Atlantic*, Oct. 30, 2014.

[6] H. Kahramanli and N. Allahverdi, "Mining Classification Rules for Liver Disorders," *International Journal of Mathematics and Computers in Simulation*, vol. 3, no. 1, pp. 9–19, 2009.

[7] M. Durairaj and N. Ramasamy, "A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate," *International Journal of Control Theory and Applications*, vol. 9, no. 27, pp. 255–260, Jan. 2016.

[8] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *Journal of The Royal Society Interface*, vol. 8, no. 59, pp. 842–855, Jun. 2011, https://doi.org/10.1098/rsif.2010.0456.

[9] L. A. Allen *et al.*, "Decision Making in Advanced Heart Failure," *Circulation*, vol. 125, no. 15, pp. 1928–1952, Apr. 2012, https://doi.org/10.1161/CIR.0b013e31824f2173.

[10] A. K. Dubey and K. Choudhary, "A systematic review and analysis of the heart disease prediction methodology," *International Journal of Advanced Computer Research*, vol. 8, no. 38, pp. 240–256, 2018, https://doi.org/10.19101/IJACR.2018.837025.

[11] A. K. Dubey, K. Choudhary, and R. Sharma, "Predicting Heart Disease Based on Influential Features with Machine Learning," *Intelligent Automation & Soft Computing*, vol. 30, no. 3, pp. 229–243, 2021, https://www.techscience.com/iasc/v30n3/44095.

[12] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," *Expert Systems with Applications*, vol. 36, no. 3, Part 1, pp. 5432–5435, Apr. 2009, https://doi.org/10.1016/j.eswa.2008.06.054.

[13] Y. Li, T. Li, and H. Liu, "Recent advances in feature selection and its applications," *Knowledge and Information Systems*, vol. 53, no. 3, pp. 551–577, Dec. 2017, https://doi.org/10.1007/s10115-017-1059-8.

[14] J. Li and H. Liu, "Challenges of Feature Selection for Big Data Analytics," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 9–15, Nov. 2017, https://doi.org/10.1109/MIS.2017.38.

[15] S. Bharti and S. N. Singh, "Analytical study of heart disease prediction comparing with different algorithms," in *International Conference on Computing, Communication & Automation*, Greater Noida, India, Dec. 2015, pp. 78–82, https://doi.org/10.1109/CCAA.2015.7148347.

[16] S. Tahzeeb and S. Hasan, "A Neural Network-Based Multi-Label Classifier for Protein Function Prediction," *Engineering, Technology & Applied Science Research*, vol. 12, no. 1, pp. 7974–7981, Feb. 2022, https://doi.org/10.48084/etasr.4597.

[17] K. Aldriwish, "A Deep Learning Approach for Malware and Software Piracy Threat Detection," *Engineering, Technology & Applied Science Research*, vol. 11, no. 6, pp. 7757–7762, Dec. 2021, https://doi.org/10.48084/etasr.4412.

[18] H. Alalawi, M. Alsuwat, and H. Alhakami, "A Survey of the Application of Artifical Intellegence on COVID-19 Diagnosis and Prediction," *Engineering, Technology & Applied Science Research*, vol. 11, no. 6, pp. 7824–7835, Dec. 2021, https://doi.org/10.48084/etasr.4503.

[19] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, Jul. 2018, https://doi.org/10.1016/j.neucom.2017.11.077.

[20] S. Sharma and K. M. Buddhiraju, "A Novel Ant Colony Optimization Based Training Subset Selection Algorithm for Hyperspectral Image Classification," in *International Geoscience and Remote Sensing Symposium*, Valencia, Spain, Jul. 2018, pp. 5748–5751, https://doi.org/10.1109/IGARSS.2018.8519217.

[21] J. Senthilnath, S. N. Omkar, V. Mani, N. Tejovanth, P. G. Diwakar, and A. Shenoy B., "Multi-spectral satellite image classification using Glowworm Swarm Optimization," in *International Geoscience and Remote Sensing Symposium*, Vancouver, BC, Canada, Jul. 2011, pp. 47–50, https://doi.org/10.1109/IGARSS.2011.6048894.

[22] Z. Xu and J. Yang, "Model selection based on particle swarm optimization for omics data classification," in *5th International Conference on Mechanical, Control and Computer Engineering*, Harbin, China, Dec. 2020, pp. 1338–1341, https://doi.org/10.1109/ICMCCE51767.2020.00293.

[23] L. Demidova and I. Klyueva, "Data classification based on the hybrid versions of the particle swarm optimization algorithm," in *7th Mediterranean Conference on Embedded Computing*, Budva,

Montenegro, Jun. 2018, pp. 1–4, https://doi.org/10.1109/MECO.2018.8406069.

[24] Z. Lu, "Enhanced Accuracy Enabled by Particle Swarm Optimization in Classification Application," in *International Conference on Artificial Intelligence and Computer Engineering*, Beijing, China, Oct. 2020, pp. 146–149, https://doi.org/10.1109/ICAICE51518.2020.00034.

[25] P. Bhavani Shankar and Y. Divya Vani, "Conceptual Glance of Genetic Algorithms in the Detection of Heart Diseases," in *International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies*, Bhilai, India, Feb. 2021, pp. 1–4, https://doi.org/10.1109/ICAECT49130.2021.9392604.

[26] Md. T. Islam, S. R. Rafa, and Md. G. Kibria, "Early Prediction of Heart Disease Using PCA and Hybrid Genetic Algorithm with k-Means," in *23rd International Conference on Computer and Information Technology*, Dhaka, Bangladesh, Dec. 2020, pp. 1–6, https://doi.org/10.1109/ICCIT51783.2020.9392655.

[27] A. T. Saputra, B. Prindo Sugiharto Putro, W. A. Saputro, and M. Muljono, "Optimization Neural Network With PCA And PSO On Heart Disease Classification," in *International Seminar on Application for Technology of Information and Communication*, Semarang, Indonesia, Sep. 2020, pp. 191–195, https://doi.org/10.1109/iSemantic50169.2020.9234276.

[28] S. K. Prabhakar, H. Rajaguru, and S.-W. Lee, "Metaheuristic-Based Dimensionality Reduction and Classification Analysis of PPG Signals for Interpreting Cardiovascular Disease," *IEEE Access*, vol. 7, pp. 165181–165206, 2019, https://doi.org/10.1109/ACCESS.2019.2950220.

[29] S. Hendra Wijaya, G. Timur Pamungkas, M. Burhanis Sulthan, and Muljono, "Improving Classifier Performance Using Particle Swarm Optimization on Heart Disease Detection," in *International Seminar on Application for Technology of Information and Communication*, Semarang, Indonesia, Sep. 2018, pp. 603–608, https://doi.org/10.1109/ISEMANTIC.2018.8549722.

[30] M. G. Feshki and O. S. Shijani, "Improving the heart disease diagnosis by evolutionary algorithm of PSO and Feed Forward Neural Network," in *Artificial Intelligence and Robotics*, Qazvin, Iran, Apr. 2016, pp. 48–53, https://doi.org/10.1109/RIOS.2016.7529489.

[31] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Computational intelligence technique for early diagnosis of heart disease," in *International Conference on Engineering and Technology*, Coimbatore, India, Mar. 2015, pp. 1–6, https://doi.org/10.1109/ICETECH.2015.7275001.

[32] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, "An Automated Diagnostic System for Heart Disease Prediction Based on chi^2 Statistical Model and Optimally Configured Deep Neural Network," *IEEE Access*, vol. 7, pp. 34938–34945, 2019, https://doi.org/10.1109/ACCESS.2019.2904800.

[33] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," *IEEE Access*, vol. 7, pp. 180235–180243, 2019, https://doi.org/10.1109/ACCESS.2019.2952107.

[34] I. Yekkala, S. Dixit, and M. A. Jabbar, "Prediction of heart disease using ensemble learning and Particle Swarm Optimization," in *International Conference On Smart Technologies For Smart Nation*, Bengaluru, India, Aug. 2017, pp. 691–698, https://doi.org/10.1109/SmartTechCon.2017.8358460.

[35] S. Chakraborty, S. Paul, and Md. Rahat-uz-Zaman, "Prediction of Apple Leaf Diseases Using Multiclass Support Vector Machine," in *2nd International Conference on Robotics, Electrical and Signal Processing Techniques*, DHAKA, Bangladesh, Jan. 2021, pp. 147–151, https://doi.org/10.1109/ICREST51555.2021.9331132.

[36] M. Masud *et al.*, "CROWD: Crow Search and Deep Learning based Feature Extractor for Classification of Parkinson's Disease," *ACM Transactions on Internet Technology*, vol. 21, no. 3, Mar. 2021, Art. no. 77, https://doi.org/10.1145/3418500.

[37] M. Kumar, A. Kumar, and V. S. Palaparthy, "Soil Sensors-Based Prediction System for Plant Diseases Using Exploratory Data Analysis and Machine Learning," *IEEE Sensors Journal*, vol. 21, no. 16, pp. 17455–17468, Dec. 2021, https://doi.org/10.1109/JSEN.2020.3046295.

[38] A. Dubey, U. Gupta, and S. Jain, "Medical Data Clustering and Classification Using TLBO and Machine Learning Algorithms," *Computers, Materials and Continua*, vol. 70, no. 3, pp. 4523–4543, Oct. 2021, https://doi.org/10.32604/cmc.2022.021148.

[39] A. K. Dubey, U. Gupta, and S. Jain, "Computational Measure of Cancer Using Data Mining and Optimization," in *International Conference on Sustainable Communication Networks and Application*, Erode, India, Jul. 2019, pp. 626–632, https://doi.org/10.1007/978-3-030-34515-0_65.

[40] J. V. Rosy and S. B. R. Kumar, "Optimized encryption based elliptical curve Diffie-Hellman approach for secure heart disease prediction," *International Journal of Advanced Technology and Engineering Exploration*, vol. 8, no. 83, pp. 1367–1382, 2021, https://doi.org/10.19101/IJATEE.2021.874436.

[41] "UCI Machine Learning Repository." http://archive.ics.uci.edu/ml/index.php (accessed Mar. 31, 2022).

[42] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Association rule mining to detect factors which contribute to heart disease in males and females," *Expert Systems with Applications*, vol. 40, no. 4, pp. 1086–1093, Mar. 2013, https://doi.org/10.1016/j.eswa.2012.08.028.

[43] S. Sperandei, "Understanding logistic regression analysis," *Biochemia Medica*, vol. 24, no. 1, pp. 12–18, Feb. 2014, https://doi.org/10.11613/BM.2014.003.

[44] J. C. Stoltzfus, "Logistic Regression: A Brief Primer," *Academic Emergency Medicine*, vol. 18, no. 10, pp. 1099–1104, 2011, https://doi.org/10.1111/j.1553-2712.2011.01185.x.

[45] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, Jan. 2008, https://doi.org/10.1007/s10115-007-0114-2.

[46] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. 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, Apr. 2013, https://doi.org/10.1016/j.jclinepi.2012.11.008.

[47] L. Yang, *Distance Metric Learning: A Comprehensive Survey*. Michigan, MI, USA: Michigan State University, 2006.

[48] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[49] S.-W. Lin, K.-C. Ying, S.-C. Chen, and Z.-J. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines," *Expert Systems with Applications*, vol. 35, no. 4, pp. 1817–1824, Nov. 2008, https://doi.org/10.1016/j.eswa.2007.08.088.

[50] J. F. Easton, C. R. Stephens, and M. Angelova, "Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: A data mining approach," *Computers in Biology and Medicine*, vol. 54, pp. 199–210, Nov. 2014, https://doi.org/10.1016/j.compbiomed.2014.09.003.

[51] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *International Conference on Neural Networks*, Perth, WA, Australia, Dec. 1995, vol. 4, pp. 1942–1948, https://doi.org/10.1109/ICNN.1995.488968.

[52] W. Y. Dong and R. R. Zhang, "Order-3 stability analysis of particle swarm optimization," *Information Sciences*, vol. 503, pp. 508–520, Nov. 2019, https://doi.org/10.1016/j.ins.2019.07.020.

[53] M. R. Bonyadi and Z. Michalewicz, "Analysis of Stability, Local Convergence, and Transformation Sensitivity of a Variant of the Particle Swarm Optimization Algorithm," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 3, pp. 370–385, Jun. 2016, https://doi.org/10.1109/TEVC.2015.2460753.

[54] S. Ahmed, M. Mafarja, H. Faris, and I. Aljarah, "Feature Selection Using Salp Swarm Algorithm with Chaos," in *2nd International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, Phuket, Thailand, Mar. 2018, pp. 65–69.

[55] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary grey wolf optimization approaches for feature selection," *Neurocomputing*, vol. 172, pp. 371–381, Jan. 2016, https://doi.org/10.1016/j.neucom.2015.06.083.