A hybrid machine learning approach for heart disease prediction using hyper parameter optimization

Dibakar Sinha  
Shri Vaishnav Vidyapeeth Vishwavidyalaya

Ashish Sharma  
GLA University

Sourabh Sharma (✉ ssharmacse@gmail.com)  
Avantika University  
https://orcid.org/0000-0001-6359-6532

Research Article

Keywords: Hyper parameter optimization, machine learning, cardiac arrest diagnosis, k-fold cross validation, WBAM, pre-processing.

Posted Date: December 22nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2303591/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. 
Read Full License
A hybrid machine learning approach for heart disease prediction using hyper parameter optimization

Dibakar Sinha, Ashish Sharma, Sourabh Sharma*

Department of Computer Science & Engineering, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore (MP) India, dibakar18@gmail.com
Department of Computer Engineering & Applications, GLA University, NH#2, Delhi Mathura Highway, Post Ajhai, Mathura (UP) India, ashishs.sharma@gla.ac.in
School of Engineering, Avantika University, Vishwanathpuram Lekoda, Ujjain (MP) India, ssharmacse@gmail.com
* Corresponding Author

Abstract: Heart disease is a serious terminal condition in most parts of the world. The acute lack of medical professionals, expertise, and technology to identify important signs. So a smart and efficient model and technology is required to lead early diagnosis of heart disease. The current study proposes a new experience-based method namely HSPUCD (Heart stage prediction using clinical data) to forecast cardiac disorders utilizing hybrid machine learning. Type 4 protocols Cross Validation is used to assess a model's competence on unknown data. Hyper parameters were modified to regulate the suggested model's behavior and performance. The suggested system is based on extensive research. Data pre-processing was utilized to discover the most relevant network traffic and generate the initial candidate collection of feature characteristics. The suggested model additionally uses automated feature selection to extract features from a candidate set of features and selects the most relevant subset of features from this candidate set. Crow search algorithm is used to extract data characteristics from the pre-processed data. Finally, an online training classifier is utilized to categorize the data and forecast the result. The proposed strategy greatly increases the system's illness prediction accuracy with less noise and more advanced regression approaches using existing medical datasets. Mathematical results depicts the effectiveness of the proposed approach in comparison with others.

Keywords: - Hyper parameter optimization, machine learning, cardiac arrest diagnosis, k-fold cross validation, WBAM, pre-processing.

1. Introduction

The coronary disease is known as one of the major killers in the whole world, WHO declared that around 31% of the total world population died in 2015 which is going to be increased very rapidly in 2020. Approximately 20 million people die each year showing it is the strong reason of death in the world. Cardiovascular disease (CVD) is an illness which is caused mainly by blood vessels taking the blood, oxygen, etc. to the heart leading to the blockage in the vessels for various reasons [1-4]. Coronary Artery Disease (CAD) or coronary heart disease (CHD) is part of CVD which mainly caused by the veins which are takes blood and oxygen to the heart. These types of illness are known as lifestyle disease such as diabetes, high blood pressure which is mainly caused by the lifestyle changes in today's fast-paced modern world. Another terminology is atherosclerosis which depicts the accumulation of fatty production on the border wall of arteries. Due to this, the amount of blood flow throws the arteries reduced a lot blocking smooth action of heart which ultimately causes angina or heart attack. Angina is usually known as angina pectoris which is chest pain in the heart and typically starts with pain in the limbs then jaws and then ultimately in the heart [4-9]. Some of the names heart attacks are myocardial infarction (MI), cardiac arrest and coronary thrombosis or occlusion. It occurs either when the arteries are torn apart or unable to take blood to the heart due to the narrowing of the arteries. A blood clot formed partially or completely during the repair of blood vessels rupture decreases the flow of blood to the muscles of heart causing heart attacks. Some pains which also radiate to the left are also one type
of heart attack. Some of the causes for such kind of illness are the use of nicotine, family history, diabetes, high cholesterol, high blood pressure, older age, less physical activity or stress [10-13]. Congestive heart failure is also known as heart attack where both the side of the heart wall, valve and arteries are damaged and once it is damaged cannot be repaired. Due to the above-mentioned reason, the heart arteries get damaged on both part of the side and it is not able to supply the required amount of blood which is required by the heart which makes it a serious health concern for the person [14]. In another way the chest trouble compared to a heart attack or different heart problem may be defined by or amalgamated with one or more of the following:-

1. Uneasiness in the chest
2. The severe pain that spreads starting from hand to most part of the body.
3. A pain which comes back after it goes down again and again
4. Brevity of inhalation
5. Cool perspiration
6. Dizziness or faintness
7. Queasiness or vomiting

The heart became so weak that the heart was unable to pump the blood to the heart. Another type of heart disease is peripheral artery disease (PAD) which makes the side of arteries narrower of arms, legs and stomach or head, legs and stomach [15-20]. Some other abnormalities are venous thromboembolism which is the clotting of blood in the vein, and aortic aneurysms which allows the artery to widen suddenly which is already weakened enough to do so [21]. So from the above all the statistics we can safely conclude that it is very difficult to diagnose heart disease just from the visual perspective because it has many risk factors and usually the heart diseases or heart attack doesn’t come with prominent noticeable symptoms [22-25]. Some important signs of the patient like triglyceride, cholesterol, diabetes, heartbeat, blood pressure and then measurement from ECG EEG EMG are very important and represent a important role in diagnosing and detecting the stages of heart and whether in future the patient will suffer a cardiac arrest or not. Some devices like WBAN (Wireless body area network) are used for detecting such vital signs of the patient which is due to the advancement in wireless technology [25-28]. These technologies used some very tiny chip or devices which the patient is able to wear most of the time and it will continuously sense the vital signs of the patient regularly and if there is any abnormality it can immediately inform the patient or the doctor or the relatives via digital communication model [29-31]

![Fig. 1 Wireless Body Area Networks](image.png)
There are three types of BMS devices are used, one is body wearable devices that senses heart pulse and blood pressure, than there is in-body devices which measures kidney, liver, heart vital measurement and then the last one is out body devices which measure length, height, weight, etc. These devices which is connected with the person sends signal to the server available on the mobile of the patient [31-35]. These servers which are connected to the internet via GPRS or LAN, sends signal to health data to different nodes like health care worker, physician, caregiver, emergency worker, who analyses the result and take appropriate action according to their work scope[36].

Body Area Network Coordinator plays a central role to monitor all those devices activity which follows a star topology. Often these data collected from these devices are categorized as an emergency and non-emergency data. The emergency data are those which are like glucose level, and nonemergency data are those like blood pressure and heart bit. A brief architecture of the working of these devices is shown in Fig. 1 [37].

There are 4 stages of heart failure (Stage A, B, C, and D). As the condition worsens, the heart muscle pumps fewer blood into tissues, and the patient heads into the next level of heart failure.

Stage A has signs such as
1. Hypertension
2. Diabetes
3. Coronary artery disease
4. Metabolic syndrome

Stage B has symptoms like
1. Alcohol addition
2. Fever occurrence
3. Cardio Vascular disease in family

Stage C has symptoms like
1. Treatments listed in Stage A
2. Not involved in anything like stage A
3. Have heart attack before with 40% EF score
4. Diabetes
5. Possible surgery

Stage D has symptoms like
1. Breathlessness
2. Tiredness
3. Not physically fit
4. Not strong legs
5. Recurrence Urinating
6. legs or ankle is swollen

Proposed HSPUCD (Heart Stage Prediction Using Clinical Data) is judged by using actual life time-varying attributes or data. Moreover the system is very much useful for the medical professional to detect heart diseases. Additionally it helps the medical professional to make a correct prediction in terms of detecting patient heart attack stage [38-40]. If a medical professional is aware of the chances of the patient is going to suffer the heart attack than it is very easy for him to take necessary precautionary steps to avoid or minimize the loss in terms of death or any other physical disability which may occur to the patient in case of heart attack or cardiac arrest

An enhanced machine learning model is inducted in the paper to detect and guess the cardiac diseases at a very premature stage. This is achieved by applying the pre-processing of sampled data set to characterization technique and later fed into the proposed algorithm with the custom cross validation and iteration value. The details working process is discussed on the Methodology section of this paper of this article [41-43].

The experimental outcome shows the accuracy or correctness of the proposed prediction system than the previously used models which are used in the models used by a different researcher with 82.5% accuracy

The remaining section of current article is modelled into various parts: Literature reviews are explained in Sect. 2.
In Sect. 3, previous lessons have been evaluated. The suggested heart disease prediction system is delivered in section 4. Results and discussions have been explained in Sect. 5. At last, conclusions of this paper are discussed in Sect. 6.

2. Related Works

Extraordinary growth has been done in the way of various ML algorithms utilized to therapeutic datasets for exposing various disorders, e.g. disclosure of different kinds of cancer. The chief clinical systems for evaluation of heart illness involve an electrocardiogram (ECG) [26], echocardiogram [27], Holter monitoring [29], cardiac computer-assisted tomography (CT) scan [28], resonance imaging (MRI) [30], cardiac magnetic, blood tests [31], including cardiac categorization [32]. In this segment, heart-related research is appraised, with a centre on CAD, applying machine learning and data mining procedures.

Ali, Farman, et al [1] suggested Naïve Bayes, Decision Trees, Neural Network, also. Not with standing proper results, each procedure has its extremely personal character. Shrouded cases and relationship among them are employed to create this structure. It is simple to learn, expandable and automatic.

Shah, Devansh, Samir Patel, and Santosh Kumar Bharti [2] suggested the estimate of blood pressure, heart ailment and sugar including the design of neural systems. The histories of 13 fields in all were employed in the data set. For fitting and measuring of data, the managed systems.

Dhwakar, Manoj, et al introduced a different method, to produce up the several parametric component including through and nonlinear properties of Heart Rate Variability (HRV) rare classifiers for SVM (Support Vector Machine), CMAR (Classification dependent on Multiple Association Rules), C4.5 (Decision Tree), standard Bayesian Classifiers has been employed by them. To assess the polluting impact of a section or faction of developing tuples [3].

Ahmed, Hager, et al. [4] suggested foretelling heart disease (HD) practicing quantitative computer-assisted classical Chinese medication and representation-based approaches. The optimized and implemented classifiers (11 algorithms) were utilizing the Collaborative Representation-based Probabilistic passageway. Following utilizing to pre-process including normalization and characteristic wrenching, the recommended practice was connected to a heart disorder dataset utilizing four principal classifiers: SVM, RBFNN, PNN and KNN. The achieved outcomes symbolized that the evolutionary-neural practice with logistic regression had the biggest achievement with 90% efficiency for a dataset with 17 classes.

Abdel-Basset, Mohamed, et al. implemented various machine learning algorithms on CAD examination from 2012 to 2017 [38-45]. The most important efficiency achieved between all these investigations was 94.08%, which was achieved by utilizing the Information Gain-SMO algorithm [5].

Yildirim, Ozal, et al worked on making the prediction technique using Hadoop cluster distribution technique in a large dataset. They used feature selection with information gain and applying it in the neural network neuron with weight based methodology for effective prediction [6].

Yu Sato, Akiomi Yoshihisa, et al. developed a methodology using pipeline hybrid system development which works on rule-based approaches and machine learning-based. They used extracting tags from the clinical data based on its time, phase, and logic used [7].

Cohen, Sarah, et al. developed a model where features are optimized and extracted using different feature selection techniques and classified using an ML classification algorithm [8].

Sharma, Prerna, et al. developed neural networks and Gini index prediction models where information in discretized using direct and supervised discretization, which is fed into the decision tree [9].
Sowmiya, C., and P. Sumitra. developed on a model where they tested different algorithms, and on each data point and algorithm were maintained, where cross-validated is performed using three times randomly shuffling before using the model. The working efficiency of the ML algorithm used accuracy; precision, sensitivity, and F1 score [10].

It has been noted that most of the previous models are usually trained on smaller data known as training data. Those models aim to calculate the result of each training example from the training data. That’s why sometimes these models were not able to predict properly when the training data has significant amount of noises mixed with it. Also in the previous model the features and turbulence are determined as a part of training data to the degree that it influences negatively to the achievement of the model on fresh data. The difficulty is that this process is literally unfit for fresh data and influence negatively to model for effective prediction. In addition to that when dataset contains lot of parameters/features but not too much data for training purposes, model will often tailor the parameter values to fit the data perfectly into the model architecture. However, because those perfect fit data don’t appear all the time, so the model predicts poorly in most cases.

3. Proposed Method

3.1 Methodology

Numerous methods on the cardio-vascular disease diagnosis have been described using AI algorithms using different data sets. The current investigation turns at another approach despite suspicion of CAD (Coronary Artery Disease). The discussion allows further facets of the acknowledged technique. The overview of the model proposed as shown in Fig. 2 is given below.

![Proposed Model Architecture](image)

3.2 Dataset

In the proposed system the data set contains 14 different health attribute of 108 and 90 patients for training and testing purpose respectively, and 2 features (patient_id and heart_disease_present) for predicting
disease as target variable. The training and test dataset contains major feature like Age, Sex, Chest pain type, Blood Pressure, Cholesterol, Electrocardiograph Data, Heart Rate, Exercise related pain, Exercise after rest, Slope of exercise, work hours, and Fluoroscope results of major veins, Diagnosis of heart disease (target label) and target vectors have only two values patient_id and heart disease_present_identifier. This dataset is taken from the UCI website. The feature of the dataset is shown in Table 1 as below.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Data Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient_id</td>
<td>Object</td>
<td>Patient identification number</td>
</tr>
<tr>
<td>Peak Exercise Slope Measurement</td>
<td>Integer</td>
<td>1 flat; 2: upsloping; 3: down sloping</td>
</tr>
<tr>
<td>Thalassemia</td>
<td>Integer</td>
<td>6 = fixed defect; 3 = normal; 7 = reversible defect</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>Integer</td>
<td>in Mm HG</td>
</tr>
<tr>
<td>Chest Pain</td>
<td>Integer</td>
<td>chest pain location (1 = substernal; 0 = otherwise)</td>
</tr>
<tr>
<td>Number of Major Veins</td>
<td>Integer</td>
<td>0-3</td>
</tr>
<tr>
<td>Blood Sugar on present on Fasting</td>
<td>Integer</td>
<td>(1 = true; 0 = false)</td>
</tr>
<tr>
<td>Electrocardiographic results</td>
<td>Integer</td>
<td>0: Normal; 2: Showing probable or definite left ventricular hypertrophy; 1: Having ST-T wave abnormality</td>
</tr>
<tr>
<td>Cholesterol Serum</td>
<td>Integer</td>
<td>mg/dl</td>
</tr>
<tr>
<td>Depression due to exercise</td>
<td>Integer</td>
<td>ST depression caused by exercise relative to rest</td>
</tr>
<tr>
<td>sex</td>
<td>Integer</td>
<td>sex (1 = male; 0 = female)</td>
</tr>
<tr>
<td>age</td>
<td>Integer</td>
<td>age in years</td>
</tr>
<tr>
<td>Maximum Heart Rate</td>
<td>Integer</td>
<td>maximum heart rate achieved</td>
</tr>
<tr>
<td>Angina due to exercise</td>
<td>Integer</td>
<td>1 = yes; 0 = no</td>
</tr>
<tr>
<td>Heart Disease Present</td>
<td>Integer</td>
<td>1=Yes;0=No</td>
</tr>
</tbody>
</table>
3.3 Protocol Type 4-fold Cross Validation

4-fold Cross validation as shown in Fig. 3 was mainly used to test the model's capacity on unseen data in the proposed model. That is, to use an insufficient experiment to decide whether the pattern will be presented in the ordinary and used to offer foresight on evidence not applied in the method's base. The procedure which is followed in this proposed model using cross validation is as below:

- Sort all samples in ascending or descending order of property values, if appropriate
- After sorting, the samples are divided into k equal-sized groups called subsets (S1, S2, ..., Sk).
- Assuming that S2 is the validation set and S1 is the training set, train a model on S1 and evaluate its performance on S2.
- This time, set S3 as the validation set, and train a model on S1 and S2 and assess its performance on S3 as the training data, i.e.
- Once all S2 through Sk have been reviewed, repeat step 4. Figure out how each model stacks up against the others

Here cross validation comes into picture when there is a situation of over fitting i.e. the model missed diagnoses anything helpful on yet undiscovered data. It used for checking a portion of data is accessible as test data or not.

\[
x' = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \\
\]

Where the \( x = 10.5 \) \( x_{\text{min}} = 5 \) and \( x_{\text{max}} = 15 \)

So \( x' = \frac{10.5 - 5}{15 - 5} = \frac{5}{10} = 0.5 \)

Where \( x \) is original current value in the series, \( x_{\text{min}} \) is lowest value and \( x_{\text{max}} \) is highest value.

Scaling is appropriate in the case of proposed algorithm because relative higher and lower bounds on the data is known with less or negligible outliers.

Data is spread relatively consistently scattered beyond that series.

The Data set is trained and tested. This trained labelled data is manipulated and normalized using scaling to a range feature which describes transforming floating-point characteristic from their unusual series (for example, 100 to 900) into a normal range. The subsequent uncomplicated method is used to compare to a variety:

\[
x' = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \\
\]

Fig. 3: Grid Search Work flow using cross validation
3.4 Hyper parameter Optimization

A hyper parameter is a parameter value that is set before the training period starts. Hyper parameters in the suggested model are sent in as arguments to the proposed model. Scale analysis techniques are used for hyper parameter optimisation. It operates by exhaustively scanning through a specified sub-set of hyper parameters. The advantage of grid searching, as shown in Fig 3, is that the optimal combination of provided parameters is guaranteed. It randomly scans the defined hyper parameter sub-set. The different hyper parameter tuned as below.

Given a data set $D$, our goal is to find

$$\lambda^* = \arg \min_{\lambda \in \Delta} E(D_{\text{train}}, D_{\text{valid}}) - D V(L, A, D_{\text{train}}, D_{\text{valid}})$$

(2)

Where $V(L, A, D_{\text{train}}, D_{\text{valid}})$ measures the loss of a model generated by algorithm $A$ with hyperparameters $\lambda$ on training data $D_{\text{train}}$ and evaluated on validation data $D_{\text{valid}}$. In practice, we only have access to finite data $D_{\text{train}}$ and thus need to approximate the expectation in Eq. 2. Where $V(L, A, D_{\text{train}}, D_{\text{valid}})$ evaluates the error of a model created with hyperparameters on training and validation data. In practice, model having finite data $D_{\text{train}}$ must approximate Eq 2

i) Regularization (L1 or L2): This is done to address a problem in which the previous model acquired signal as well as turbulence in the training data and was unable to work on data that it had not been trained on. After the penalty applied to the cost function

$$\frac{1}{m} \sum_{i=1}^{m} l(y^{\text{true}}, y^{\text{pred}})$$

(3)

The training equation would look like this

$$J(w, b) = \frac{1}{m} \sum_{i=1}^{m} l(y^{\text{true}}, y^{\text{pred}}) + \frac{\lambda}{2m} \sum_{i=1}^{m} ||\theta_i||^2$$

(4)

ii) C (Inverse of Regularization Strength): In the proposed model, a function is used that penalises high content in the parameters. The function is $\lambda \sum \theta_j * 2$, where $\theta_j = \text{sum of squared values}$. The larger it is, the less likely it is that the parameters will be increased in quantity simply to compensate for minor errors in the training data. To accomplish this, lambda is used as a hyper parameter to determine the regularisation strength. So, this paper proposes inverse regularisation strength, which is a control variable that keeps regularisation strength changes by being opposite to the Lambda regulator and ranging from 0.0001, 0.01, 0.1, 1, and 10,100, 1000, and 100,000.

iii) In the earlier models, a lot of independent variables were present but not too much data, so the prediction was poor. The reason could be that the model was designed so perfectly that it almost fits all training data, but the same does not hold for test data. In the proposed algorithm, LBG (Limited-memory Broyden–Fletcher–Goldfarb–Shanno Algorithm) optimization process is used where evaluation is practised for the inversing component, which is nothing but a square matrix of second-order partial derivatives of order NXN.

iv) An iterative algorithm is assumed to converge as the result is ever closer to the given value as the iterations continue. An algorithm will diverge in certain circumstances; its performance will suffer bigger and bigger oscillations, never reaching a useful result. In the proposed algorithm, the Maxiter value is set to 100 to ensure the proposed model doesn’t spend much time on convergence.

3.5 Data Manipulation & Normalization

Data Cleaning:-

- Some data in the cardiac dataset used in the proposed algorithm have many undesired and missed parts which is a significant obstacle for heart disease prediction. To deal with this, cleaning of data is necessary. It involves handling of noisy data and missing data etc.
Noisy Data:-
- Sometimes in the cardiac dataset some meaningless data can be present due to data entry errors, faulty data collection etc which should be dealt with.

Table 2: Handling of Noisy data in cardiac arrest data set

<table>
<thead>
<tr>
<th>Thalasemia</th>
<th>Blood Viscosity</th>
<th>Heart_Disease_Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>$3.26 \times 10^{-3}$</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>$3.59 \times 10^{-3}$</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>?</td>
<td>1</td>
</tr>
</tbody>
</table>

As can be seen in Table 2 some non-numerical data (?,+,=) present. These are known as noisy data and had to be replaced with some other value before the training starts.

Missing Data:-
Dealing with missing values is a crucial action while creating the proposed model. It influenced the outcome of the proposed model. The missing record in the dataset happened in the file for various caused like difficulties happened during the uprooting or data gathering method.
Here in this proposed model pair wise method have been adopted by neglecting the disappeared value and viewing only the possible advantages. The process details are show in Fig. 4.

<table>
<thead>
<tr>
<th>Blood Pressure</th>
<th>re</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systolic (mm/Hg)</td>
<td>Diastolic(mm/Hg)</td>
</tr>
<tr>
<td>110</td>
<td>75</td>
</tr>
<tr>
<td>115</td>
<td>73</td>
</tr>
<tr>
<td>105</td>
<td>1</td>
</tr>
</tbody>
</table>

Fig. 4 Replacing Missing values with random data

3.6 Pre-processing Data

Pre-processing (data normalization and manipulation):
It is a data mining procedure that comprises reconstructing fresh data into an acceptable format. Real-world information is typically incomplete, irregular, and/or lacking in specific ways or leans, and it is expected to contain numerous errors. Data pre-processing is a proven technique for solving the problems mentioned above.
In the proposed model, the pre-processing algorithm (GridSearchCV) performs as a common estimator API while "fitting" the model on a dataset for all the feasible successions of parameter values and is assessed such that only the best succession is maintained.

Fig. 5 Data Prepossessing Technique
Data is pre-processed through transformation, reduction, cleaning, noise removal, and missing data removal and this data is mined and collected, and then the result is produced for further training. A brief flow of the process is shown in Fig. 5. Further discussion is done below.

**Transformations of space**

If there is requirement to reduce the number of dimensions, feature selection is not the only process. Space transformation approaches don’t choose the most promising features, but rather combine the original ones into a new collection of features. Different criteria can be used to make such a combination. Factor analysis [28] and PCA [29] were the first linear methods used.

Use of instance reduction (IR) approaches has become common in data mining algorithms for reducing the effect of extremely huge data sets. Data sets can be reduced without sacrificing the quality of information that can be gleaned from them. For feature selection, instance reduction serves as an advantage for performance enhancement of the proposed algorithm. This technique removes or creates new instances to decrease the amount of data.

**Need of Pre-processing of the cardiac data for the proposed model**

**Data Transformation:**

In the proposed algorithm the cardiac dataset need to be pre-processed, cleaned and to be dealt with missing values. Mostly it is done by changing string objects to numerical attribute i.e. 0, 1, 2, 3 and so on. For this purpose Grid based pre-processing is applied. The reason behind it as follows:

- Most classifier can only work with numerical data, and will raise an error when categorical values in the form of strings is used as input, in the proposed model cardiac dataset contain some string values so it is required to convert it. For this purpose two things are implemented.
  - One-hot encoding of the column such that its categorical values are converted to numerical values and
  - Expand the column into N different columns containing binary values.

**Data Reduction:**

- As cardiac dataset used in the proposed algorithm is huge in nature so it is difficult to work with such huge data, so various data reduction technique is used to reduce data storage, analysis costs and storage efficiency. As data mining technique which has been applied in the proposed algorithm is dealing with huge amount of data so while working with this large volume of records, analysis became difficult in many cases. To avoid this, data reduction technique is applied which is used for reducing data storage, increasing of efficiency and reduce analysis costs.

- In the proposed algorithm some attribute likes, eye colour, region, community are obsolete and not relevant as per the proposed algorithm flow chart so those attribute are discarded.

**3.7 Feature Selection & Classifier Optimization**

A technique known as crow search was used to extract data characteristics from the pre-processed data in this study. Crow Search Algorithm (CSA): The CSA is a metaheuristic algorithm that uses a levy-based approach. Because of their memory and ability to conceal food, crows are inherently clever birds. When searching for new food, our programme mimics the reasoning abilities of a rookery. To discover where their food is, the crows follow other birds. Crows devour or steal food from other birds as soon as they leave their nests. Crows can avoid becoming victims of food theft if they alter the hiding position of their prey if another crow or another bird snatches their meal. For the crow, predicting the thief’s next move is an essential part of determining the safest spot to store its food.

There have been cases where the crow search algorithm has provided the best option. Nevertheless, due to a lack of thoroughness in exploring the search space, this approach does not ensure convergence. Crow algorithm’s search approach has substantial hurdles when presented with exceptionally multimodal formulation. The suggested solution leverages a Levy-based CSA to solve these issues. Random motions are performed using the Levy flying behaviour in LCSA. There’s a distribution of probability called the Levy distribution that governs how big of a step to take when
doing Levy flights. In the search for space exploration, this distribution is more successful than a random distribution.

N crows are randomly seeded into $P_n$, the population. $M_{it}$ is taken as the maximum iteration count by the search method. The $e$th crow's location in the $d$th-dimensional search space is determined by the following equation:

$$C_{1,e,k}, C_{2,e,k}, \ldots, C_{n,e,k} e = 1, 2, \ldots, P_n, K = 1, 2, \ldots, M_{it}$$

(5)

When it comes time for the following iteration, each crow should be able to recall the most frequently visited place, designated as $B(e,k)$, where the food will be hidden.

$$B_{e,k} = [b_{1,e,k}, b_{2,e,k}, \ldots, b_{n,e,k}]$$

(6)

In order to find where the food is concealed, the two birds are chasing each other while the other bird is ignorant that the other bird is chasing it. The goal of crow $c$ is achieved in this way. Crow $E$ assumes that the presence of crow $C$ is known to it, thus it utilises random paths to protect the food. In the crow search method, the irregular mobility of the crows represents this activity. The suggested approach uses the Levy flying technique for random motion. According to the AP, the sort of behaviour displayed by each crow $e$ is taken into account. The number $R_{e}$ is drawn at random from a uniform distribution between $0$ and $1$. The Levy dispersion is depicted in the following equation.

$$E_{e} = \text{Levy} \sim u = t^{-\lambda}, (1 < \lambda \leq 3)$$

(7)

The movement's functioning may be modelled as an equation by using the following method:

$$C_{e,k+1} = \begin{cases} C_{e,k} + R_{e} \cdot \text{Fl}_{e,k} \cdot (B_{c,k} - C_{e,k}) & \text{if } \text{move to random place}, \\ C_{e,k} & \text{otherwise} \end{cases}$$

(8)

The duration of time spent in the air in the optimum feasible point $B_{c,k}$ for crow $C_{e,k}$, $\text{Fl}_{e,k}$ denotes the amplitude of the movement from $C_{e,k}$. The random variable with a distribution between $[0, 1]$ is denoted by the symbol $R_{e}$. The locations of the crows are reviewed when they change. It is necessary to do the following buffer vector change using Equation (10).

$$B_{e,k+1} = f(x) = \begin{cases} C_{e,k+1} & \text{if } O(C_{e,k+1}) \text{ is better than } f(B_{e,k}) \\ B_{e,k} & \text{otherwise} \end{cases}$$

(9)

The optimization algorithm is indicated by $O(\cdot)$ which is to be minimised.

### 3.8 Online Classifier

Online real-time data collection is used in medical diagnosis applications, where each new piece of information is received one at a time and classified before the next one arrives. An online training method is useful in this situation. Online methods allow simultaneous exercise of model training and data collection. Thus, when dealing with medical data, the HSPUCD is superior, and the layout can be revised to suit the improvement in consumer health data.

Digital binary classification operates on data from health repository. At time $t$, the classification model in a $d$-dimensional vector space is an incoming instance $x_t \in X$, i.e. $X = R^d$. Present classifier predicts the class mark of a new input data,

$$y_t = \text{sgn}(f(x_t, W_t)) = \text{sgn}(W_t \cdot x_t) \in y$$

(10)

Where $y = \{-1, +1\}$ used for binary classification functions. The model measures the loss $L(y_t, y_t')$ by calculating the difference between the expected label and the actual label after making the prediction. The model determines when and how to change his parameters, based on the outcome. Specific online training algorithms enforce numerous failure functions $\mathcal{L}(\cdot)$ and separate upgrade functions $\Delta\mathcal{L}(\cdot)$ so any parameter present in the vector space is classified and chosen suitably for the prediction by minimizing the loss and increasing the accuracy.

**Optimization of the parameters used in classification** - Linked with cross-validation and feature assortment, proposed model has different parameters which are used before training started. One
is the model parameter which is used for learning purpose for training the model, like for example weights in the artificial neural network and another is coefficient regression. The approach used in the optimization of the parameters is parameter tuning and grid search.

Confusion Matrix

In the proposed algorithm confusion matrix as shown in Fig. 6 is used because the details of attribute can be two or more than two classes.

![Confusion Matrix](image)

Fig. 6 Confusion Matrix

It is very much useful for formulating the specificity decision accuracy and record and for doing ROC curve.

“True Positive (TP)” for values of the event which are correctly forecasted.
“False Positive (FP)” for values of the event which are incorrectly forecasted.
“True Negative (TN)” for values of the no-event which are correctly forecasted.
“False Negative (FN)” for values of the no-event which are incorrectly forecasted.

It is considered that while the proposed model was tested according to accuracy and F1-score parameters, according to "accuracy" best results are achieved.

4. Results

Comparison with other algorithm

Initially eight machine learning algorithms were tested against UCI cardiac arrest dataset for finding out their Training and testing score and time. The result has been demonstrated using a Table 3 below.

Table 3. Performance Chart of ML algorithm for the dataset

<table>
<thead>
<tr>
<th>S. No</th>
<th>Classifier</th>
<th>Training Score</th>
<th>Testing Score</th>
<th>Training Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gradient Boosting Classier</td>
<td>1</td>
<td>0.814815</td>
<td>0.576464</td>
</tr>
<tr>
<td>2</td>
<td>Decision Tree</td>
<td>1</td>
<td>0.814815</td>
<td>0.003336</td>
</tr>
</tbody>
</table>
One of the major features of the dataset is that it has less amount of error induced into it and minimum pre-processing is required. As can be seen from the above table the proposed model training and testing score and time outperforms many of the popular machine learning models resulting in performance enhancement.

A graph of the above table can be shown as below in Fig. 7.

![Performance Comparison of Different ML Algorithm](image_url)

**Fig. 7.** Performance Comparison of Different ML Algorithm with the Proposed Algorithm
Using many classification techniques can increase time but accuracy will be much more improved in our case. Some of the popular algorithms which are used are Gradient Boosting, Decision tree, Logical regression, Naive Bays, Random Forest, Nearest Neighbour, Linear SVM, and Neural Network. We are able to identify that train score has been highest for Gradient Boosting, Decision tree, Random Forest and Liner SVM, similarly test score have been highest for gradient boosting, Decision tree, Logistic regression, and Naive Bayes and lowest train time is required for decision tree algorithm.

To assess the achievement of the proposed model, two conventional metrics were adopted: F-measure (F1-score) and Accuracy (ACC).

\[
F - measure = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}
\]  

(11)

Where:
1. N denotes the set no used in the cross-validation (i.e. scaled 100-fold cross-validation),
2. TP is accurately recognized (True Positives) units,
3. TN is rightly declined specimens (True Negatives),
4. FP is mistakenly recognized specimens (False Positives), and
5. FN has wrongly repudiated specimens (False Negatives),
6. C expresses the abundance of sources

Depended on confusion matrix, \( F - measure \) metrics and Correctness is used for calculating the effectiveness of the proposed model.

So in the confusion matrix different parameter value is as below

<table>
<thead>
<tr>
<th>N=165</th>
<th>Predicted False</th>
<th>Predicted Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual No</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>Actual Yes</td>
<td>5</td>
<td>100</td>
</tr>
</tbody>
</table>

So In simplicity we can deduce like below

<table>
<thead>
<tr>
<th>N=165</th>
<th>Predicted False</th>
<th>Predicted Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual False</td>
<td>TN=50</td>
<td>FP=10</td>
</tr>
<tr>
<td>Actual True</td>
<td>FN=5</td>
<td>TP=100</td>
</tr>
<tr>
<td>55</td>
<td>110</td>
<td></td>
</tr>
</tbody>
</table>

**Classification Rate/Accuracy:**

\[
\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100
\]  

(12)

= 90%

**Recall**

Recall gives us an idea about when it’s actually TRUE for Heart Disease, how often does it predict TRUE for heart disease.

\[
\text{Recall} = \frac{TP}{TP + FN} \times 100
\]  

(13)

= 95%

**Precision**

Precision shows the percentage of relevance of the cardiac results as per the proposed model principle
\[ \frac{TP}{TP+FP} \times 100 = 91\% \]  

**F-measure**

Here the F score calculated as the harmonic weighted mean of the proposed model’s recall and precision.

\[ \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \times 100 = 92\% \]

Many of the earlier models utilised by the majority of the authors are computationally costly because of the variety of normalisation and feature selection techniques. Table 4 compares the proposed method’s prediction accuracy to state-of-the-art methodologies used for cardiac illness data sets, as indicated in the figure. In terms of overall accuracy, the suggested work exceeds the competition, as can be shown in Table 4.

**Table 4. Comparison of proposed method with existing models**

<table>
<thead>
<tr>
<th>Authors / Year</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ali, Farman, et al/2020</td>
<td>87</td>
<td>70.5</td>
<td>79.3</td>
<td>77.9</td>
</tr>
<tr>
<td>Shah, Devansh, Samir Patel, and Santosh Kumar Bharti/2020</td>
<td>88</td>
<td>75</td>
<td>78.1</td>
<td>78.7</td>
</tr>
<tr>
<td>Diwakar, Manoj, et al/2021</td>
<td>86</td>
<td>78.45</td>
<td>88.14</td>
<td>77.56</td>
</tr>
<tr>
<td>Ahmed, Hager, et al/2020</td>
<td>84</td>
<td>76.4</td>
<td>84.4</td>
<td>80.95</td>
</tr>
<tr>
<td>Abdel-Basset, Mohamed, et al/2020</td>
<td>86</td>
<td>79.3</td>
<td>87.9</td>
<td>87.59</td>
</tr>
<tr>
<td>Yildirim, Ozal, et al/2019</td>
<td>88</td>
<td>78.7</td>
<td>89.1</td>
<td>86.78</td>
</tr>
<tr>
<td>Yu Sato, Akiomi Yoshihisa, et al/2020</td>
<td>84</td>
<td>84.14</td>
<td>88.32</td>
<td>84.67</td>
</tr>
<tr>
<td>HSPUCD</td>
<td>90.90</td>
<td>90.90</td>
<td>95.23</td>
<td>92.95</td>
</tr>
</tbody>
</table>

Below Fig. 8 shows a graphical representation between different models by different authors in comparison with the proposed model.
Fig 8 Performance Comparison of Different ML Models with the Proposed Algorithm (HSPUCD)

Below Table 5 shows a brief comparison with HSPUCD with various popular machines learning algorithm in terms of various statistical parameters namely accuracy, precision, recall, f-measure. It can be seen that the proposed HSPUCD outperform popular machine learning model in different performance parameter.

Table 5. Existing approach versus proposed work

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient Boosting Classier</td>
<td>67</td>
<td>70.5</td>
<td>69.3</td>
<td>67.9</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>68</td>
<td>69</td>
<td>68.1</td>
<td>68.7</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>76</td>
<td>68.45</td>
<td>68.14</td>
<td>77.56</td>
</tr>
<tr>
<td>Random Forest</td>
<td>80</td>
<td>81.4</td>
<td>81.4</td>
<td>80.95</td>
</tr>
<tr>
<td>Nearest Neighbours</td>
<td>86</td>
<td>87.3</td>
<td>87.9</td>
<td>87.59</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>88</td>
<td>88.7</td>
<td>89.1</td>
<td>86.78</td>
</tr>
<tr>
<td>Neural Network</td>
<td>87</td>
<td>85.14</td>
<td>88.32</td>
<td>84.67</td>
</tr>
<tr>
<td>Proposed work</td>
<td>90.90</td>
<td>90.90</td>
<td>95.23</td>
<td>92.95</td>
</tr>
</tbody>
</table>

Fig 9. Overall comparison of accuracy with proposed work
Fig. 9 indicates the overall performance relative to our suggested hybrid model of current classifiers. The Gradient Descent is around 67 per cent less effective than any other classifier. The Decision tree classifier has 68 per cent accuracy compared to other classifiers in the second lowest results. The existing hybrid model, which is an enhanced regression algorithm with hyper parameter optimization, has an improvement in accuracy compared to the classifier’s separate analysis, and it has 90 per cent precision. Many methods such as Naïve Bayes, Random Forest and Nearest Neighbour perform more than previous current works with an accuracy average of 76, 80 and 86%.

In fact, our proposed hybrid model displays greater precision and efficiency than other current models with 90.90 per cent accuracy. Table 5 indicates the average precision, precision, recall, and F-measure of classifier with the proposed research being contrasted.

5. Conclusion & Future scope

A new algorithm called Heart Stage Prediction Using Clinical Data (HSPUCD) is introduced in this paper for heart disease analysis. The recommended HSPUCD is relatively examined with numerous present machine learning techniques such as Gradient Boosting, Decision tree, Logical regression, Naive Bayes, Random Forest, Nearest Neighbour, Linear SVM, and Neural Network. Accuracy is measured for estimating the proposed HSPUCD method. The outcomes show that the proposed HSPUCD method offered high accuracy (90%) for the dataset.

In HSPUCD, the health dataset of cardiac patients are split into several parts, which are scrutinized and treated as pre-processing. Then those pre-processed report extricated from the dataset has been established within the earlier specified suggested prototype for analytical properties or specialties. This output info is utilized for decision-making arrangement for foretelling heart attack scaffold of a cardiac patient.

The limitation of this work is that the proposed HSPUCD method involves only to heart disease. Additionally, very less number of parameters is utilized to analyse the cardiac disease. Nevertheless, the upcoming work of this article is to produce a composite classification design based on an added number of characteristics to distinguish heart disease.

Conflict of interest

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript. The authors also declare that there are no conflicts of interest.

Declarations

The authors have no connection with any group with a primary or auxiliary commercial concern in the topic matter presented in the script

Ethical Approval

Both the authors declare that they abide by the ethics of this journal. The authors also follow the policies of ethical committees, Internal Review Boards. We also give the consent to publish our article whenever it is accepted.

Competing interests

This research work does not stand any conflicting of any interests whether local or international within any community region religion ethnicity caste creed and religion

Funding

No Funding Received

Availability of data and materials

The dataset used in this research article is available in UCI repository (https://archive.ics.uci.edu/ml/datasets/Heart+Disease) and the code is available in GitHub repository which will be produced as and when demanded.
References


