Prediction Of Heart Disease Using Hybrid Linear Regression

K. Srinivas[1], B. Kavitha Rani[2], M. Vara Prasad Rao[3], Raj Kumar Patra[4], G.Madhukar[5], A. Mahendar[6]

[1, 2,3,4,6] Professor in CSE, CMR Technical Campus, Hyderabad, Telangana, India
[5]Assistant Professor, CMR Technical Campus, Hyderabad, Telangana, India.

E-mail: [1]phdknr@gmail.com, [2]phdknr1@gmail.com, [3]varam78@gmail.com, [4] patra.rajkumar@gmail.com [5]madhu@gmail.com, [6]mahi.adapa@gmail.com

ABSTRACT- Heart disease (HD) is one of the most common diseases, and early diagnosis of this disease is a vital activity for many health care providers to avoid and save lives for their patients. Heart disease accounts to be the leading cause of death across the globe. Health sector contains hidden information which helps in making early decisions by predicting existing disease such as coronary heart disease using machine learning methods. The proposed Hybrid Linear Regression Model (HLRM) implemented in two phases. Initially, data preprocessing is done; missing values are imputed with KNN and simple mean imputation and next Principal Component Analysis is used to extract the most contributing attributes for the cause of disease. Second, Stochastic Gradient Descent is the linear regression used to record the probability values of dependent variables, in order to determine the relationship between the dependent and independent variables. The overall prediction accuracy of the proposed model is observed as 89.13%. The outcome of this study will help as a reference for medical practitioners and also as a research platform for the academia

Keywords: machine learning; heart disease, association, Linear Regression Model, principal component analysis, Decision tree;

1. INTRODUCTION

Significant numbers of patient records and their medical conditions are found in health databases. A lot of valuable information is hidden within this data such as disease relationships and patterns. Extraction of this valuable information could provide& enhance new medical knowledge to identify and learn more about diseases and their pattern. In the previous research, various methodologies have been developed and applied to discover this unexplored knowledge [1]. The Naïve Bayes algorithm was used by authors in[19] to diagnose HD cases and to propose the Heart Disease Prediction System (HDPS) with the evaluation of certain algorithm parameters. There is an extensive availability of novel computational tools and methods for the analysis of data. Huge medical data is needed to develop predictive models for identification of diseases at an early stage. These models enable practitioners and researchers to select the most appropriate strategy to handle clinical decisions. Data mining can be described as a collection of such methods. Data mining offers technical and methodological solutions to analyze the medical data for developing prediction models [2][3].

The influence of data mining is more on analysis of healthcare industry data. This enables the health systems to use and analyze the data to identify best practices to improve health care and reduce clinical and investigation costs. Data mining applications do not make predicting the outcome of an illness simple. This can be developed only by clearly understanding the patient information. Various studies including data mining techniques helped to establish new research methods for medical research [4]. These approaches contribute towards the detection and creation of trends and relations of more variables. Historical cases stored within the datasets can be used to build the best models with more accuracy to predict the outcome of a disease. Every technique in data mining facilitates a different objective based on the modelling for medical data analysis. Classification and prediction are the most used common modelling techniques. Classification is commonly used to forecast outcomes on categorical labels and the prediction technique is applied for continuously evaluated features.

A technique to predict the presence of heart disease is proposed in this paper. Three steps are taken into the framework proposed. Initially, pre-processing is performed by using KNN and Simple Mean imputation methods. KNN is used for managing the missing values and Simple Mean imputation is used to address gender reorder problem in data. In the second step, the most contributed attributes are selected using principal component analysis. Finally, in the third step, linear regression is used to plot the probability scores of dependent variables to determine the association between a dependent and an independent variable. The results of linear regression analysis are used to construct a decision tree. This decision tree is used to understand real-time cases present in the dataset.

The following paper has been planned. Section 2 includes the relevant works and principles, Section 3 describes the problem, Section 4 provides the suggested model with simulated results and discussions, and Section 6 outlines the outcome and further work.

2. RELATED WORK AND CONCEPTS

Many researchers worked on health data analysis for developing efficient methods for prediction heart disease at early stages. In this section, some recent papers about heart disease prediction are discussed.

Evanthia et al. [5] proposed machine learning methodologies applied for the assessment of failure in the heart. The authors investigated by using three machine-learning algorithmic-strategies namely neural network, SVM, classification and regression trees and claimed that SVM is generating better performance than the other two algorithms.

Akhil Jabbar et al. [6] proposed a method by combining KNN with genetic algorithm for efficient classification of heart disease. The author claimed that Genetic algorithms provide an optimal solution by performing a global search in large, complex and multimodal landscapes data.

Carels et al [7] used multiple linear regression to observe the connection between depression, the severity of the disease, functional status and quality of life in the congestive heart failure patients. Hanley, et al. [8] used correlation and multiple-linear regression analysis to evaluate the cardiovascular risk factors and the associations between proinsulin.

Srinivas Konda [9] proposed a rough-fuzzy classifier to predict heart disease failure. The author combined rough set theory with the fuzzy set to enhance the prediction performance. The development of rough-fuzzy classifier includes rule generation using rough set theory and prediction using the fuzzy classifier. Nguyen, Thanh, et al. [10] proposed an automated medical data classification using interval type-2 fuzzy logic system and wavelet transformation. The author generated fuzzy rules based on the fuzzy system using fuzzy c-mean clustering algorithm and predicted the class of Cleveland heart disease datasets. K Anooj [11] proposed a weighted fuzzy rule for the diagnosis of heart disease. The fuzzy rules are generated from the patient's clinical data using an attribute selection method and these are used to develop a fuzzy rule-based decision support system.

Kirmani, Mudasirm [12] proposed a model to predict accurately the presence of heart disease using machine learning Multilayer Perceptron data mining algorithm. Das et al. [13] proposed Neural Networks ensemble model by combining the posterior probabilities or the predicted values from multiple predecessor models such as Naive Bayes, MLP, C4.5, AIRS, etc. Tan et al. [14] proposed a hybrid model to diagnosis heart failure disease using Genetic Algorithm and Support Vector Machine.

Authors implemented particle swarm optimization in order to generate evolutionary values for HD. In a comparative analysis of various machine learning algorithms for the diagnosis of heart disease as a survey paper, authors presented good rating accuracy for HD data in [21] and demonstrated how suitable machine learning algorithms and methods are to be used for analysing HD. In[22, 23], the study of cardiac variability and various classification algorithms have been proposed as a new system using non-linearity. In [24] the authors proposed a logical model for predicting HD risk level based on classification rules. In the [25] HD data-set prediction classifiers were shown to use a back-propaganda algorithm for network training and 13 clinical features to be used as inputs to predict the absence or presence of heart disease with 95 percent accuracy. In [26] various techniques for previous processes using ANN and other machine learning approaches were also provided with UCI Laboratory data and the implementation and comparison of algorithms for discovery patterns, including Decision Tree, neural networks, Rough sets, SVM, the Naive Bays. In [27], a classification model for coronary cardiac disease was proposed with a Support Vector Machine and the Artificial Neural Network (ANN) and with the aid of the Cleveland Heart Database and Statlog Datenbank based on UCI Machine Learning Dataset, the implementation of a secure, accurate and fast system for medical selection of coronary cardiac disease characterization. The classification of HD datasets for classification and foresight using voting techniques was suggested in [28]. In [29] the authors presented a technique that uses medically-tested findings as inputs, extract a reduced dimensional characteristic subset with a PPCA (Probabilistic Principal Component Analysis) and use the UCI data set to diagnose cardiac disease. The technology proposed averages the exactness of the used data set by 86.43 percent.

2.1 *KNN imputation*

Researchers often notice missing values in the data set while dealing with real-world data. The missing data must be discussed. A good practice is to identify and replace missing column values before modelling your prediction task. before using your input data. This is known as short-term imputation or missing data. The use of a model to predict missing values is a common approach to the missing data imputation. This includes creating a model for the missing values and an input variable. Although many of the different models can be used to detect missing values, the K-nearest neighbour algorithm (KNN) has proved usually to be efficient.

Values may not be accessible for a range of reasons, mostly specific to the issue area, such as corrupt measurements or lack of availability. For several computer algorithms, numeric input values and each data line and column value are needed. As such, missing values will cause problems with the algorithms for machine learning. A significant solution for the data imputation is a model for predicting missing values. For and function, a model is generated which has missing values, using input values for all other input functions.

Generally speaking, certain observations may be omitted if the proportion of missed comments is small in terms of the number of observations. But most of the time this is not the case. Remove missing lines of values can lead to the separation of useful information or patterns. As a consequence, statistically speaking, the amount of independent knowledge decreases and this leads to less equality.

The KNN model is referred to as the "nearest neighbour's imputation" or the "KNN imputation" to estimate or supply lacking values. KNNimpute may seem to have a more robust and responsive approach for estimating missing value [...] and KNNimpute would reach the common row median (as well as filling missing values with nulls). Configuring the KNN imputation typically involves choosing the KNN algorithm k hyperparameter for each prediction (i.e. euclides) and the number of contributing neighbours [30].

This can be crucial for the analysis and for concluding. In K-Nearest Neighbor Imputation (KNNI) method, every missing value is imputed by considering k most similar neighbour values in the dataset [15]. Missing values in numerical attributes are imputed by the mean value of k neighbours and the mode value is used for categorical attributes. Usually, k value is chosen between 5 and 10 and higher k imposes a negative effect on the performance of the imputation.

Missing values categories may typically be categorised as:

➢ Missing Completely at Random (MCAR)

This occurs if the missing values are unrelated to some other variable or observation characteristics. If a physician forgets to record the age of a tenth patient entering the ICU, the missing value would not depend on the patient's characteristics.

➢ Missing at Random (MAR)

The risk of a missing value in these situations depends on the data characteristics. In survey results, respondents with high incomes are less likely to notify the researchers about how many properties they possess. The missing value depends on the income variable for the variable number of assets.

➢ Missing Not at Random (MNAR)

This occurs when the missing values depend on data features and missing values. In this case, it is difficult to decide the mechanism to generate missing value. For example, missing blood pressure values can depend partially on the values of blood pressure, as patients with low blood pressure are less likely to get their blood pressure tested regularly.

2.2 Correlation Coefficient

The correlation coefficient is used to measure the degree of association [16], which is denoted by r. This is measured on a scale varies from + 1 through 0 to - 1. Complete correlation among two variables can be expressed by either + 1 or -1. When any variable has an increment and another variable also increments then the correlation is said to be positive and the value is near to +1. If there is a decline in one variable as the other increases, it is negative. There is a total lack of correlation of 0.

The estimation of the correlation coefficient - with x as the independent variable's values (in this case height) while dependent variable of the given (in this case anatomical dead space) are represented by the y values is by equation (i).

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{[\sum(x-\bar{x})^2(y-\bar{y})^2]}} \qquad \text{(i)}$$

which can be shown to be equal to equation (ii).

$$r = \frac{\sum XY - n\bar{X}\bar{Y}}{(n-1)\sigma(X)\sigma(Y)} \qquad \text{(ii)}$$

The correlation matrix is constructed by considering each attribute (X_i) in the dataset which is correlated with each of the other attributes in the(X_j). This allows you to extract which pairs have the highest correlation. The diagonal of the table is always a set of unit values because the correlation to itself of a variable and itself is always unit (i.e., 1). The correspondence matrix is simply a value table that displays the relationships. The most common coefficient of correlation is the Pearson coefficient of correlation, which compares two variables. However, several others have to correlate according to the type of data.

2.3 Linear Regression and Gradient Descent

The linear regression tries to model the relationship by a simple equation known as the regression equation which is an average value of Y. Linear regression is often used for the resolution of classification problems. No linear relationship is required for logistic regression among dependent and independent variables. It can deal with various kinds of relationships since a non-linear log transformation is applied to the expected likelihood ratio. Both significant variables and large samples should be included to prevent overfitting and underfitting. A simple linear regression has an equation of the form

$$Y = a + bX \qquad (1)$$

where $\boldsymbol{X}$ is the independent variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is

$\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept. The value of $\boldsymbol{b}$, and$\boldsymbol{a}$ is given by

$$b = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} \quad (2)$$

And

$$a = \bar{Y} - b\bar{X} \qquad (3)$$

The best parameter values for a and b in equation (1) is obtained by using Stochastic Gradient descent (SGD) [17]. The SGD is an algorithm used to minimize the objective function to best fit the linear regression equation for a given set of points. It starts with an initial set of parameter weights and iteratively slides toward a set of parameter weights that minimize the objective function defined by the error equation given below.

$$Error_{(a,b)} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - (bx_i + a)\right)^2 \qquad (4)$$

The equation (4) is partially differentiated w.r.t **a** and **b** and they are given below

$$\frac{\partial}{\partial b} = \frac{-2}{N}\sum_{i=1}^{N} x_i\left(y_i - (bx_i + a)\right) \qquad (5)$$

$$\frac{\partial}{\partial a} = \frac{-2}{N}\sum_{i=1}^{N}\left(y_i - (bx_i + a)\right) \qquad (6)$$

The partial differential equations are used to compute the gradient. The gradient will guide to move towards best parameter values. The equations to compute gradient are

(7)

$$aGradient_i = aGradient_{i-1} + \frac{-2}{N}\sum_{i=1}^{N}(y_i - (bx_i + a)) \quad (8)$$

$$bGradient_i = bGradient_{i-1} + \frac{-2}{N}\sum_{i=1}^{N} x_i\left(y_i - (bx_i + a)\right)$$

The gradient values are used to compute the linear regression parameters **a** and **b** using equations

$$a = a - learningRate * aGradient_i (9)$$

$$b = b - learningRate * bGradient_i (10)$$

Initially, $a, b, aGradient$ and $bGradient$ are set to zero and in each iteration new $aGradient$ and new $bGradient$ are computed with equation (7) and (8) respectively.

These new $aGradient$ and new $bGradient$ are used to compute new parameter values **a** and **b** using equations (9) and (10) respectively. The learning rate is a value used as a step size and usually take a value as 0.01. The above process is repeated for many iterations. Each iteration effects in updating and **b** to a line that yields slightly lower error than the previous iteration.

3. PROBLEM DEFINITION

Medical data contains a lot of uncertainty. By examining one or more symptoms it is inadequate to decide the nature of the disease by which a patient is suffering. A combination of symptoms can only indicate the likelihood of a particular disease. Data mining methods are inadequate in dealing with cognitive uncertainties such as ambiguity and vagueness. This research work focuses on addressing the problem of uncertainty in medical datasets. The goal of this work is to build a model that can predict the probability of heart disease occurrence, powered on a combination of features that describes the disease.

A set of disease-predicting parameters $\{A_k\}, k = 1,2..n$ is identified based on observed correlation with disease diagnosis (heart diseases in this case), where n is the number of such parameters. This model also has a set of N patients $\{x_i\}$ that have been diagnosed with heart disease or not and the dataset consists of the $\{A_k(x_i)\}$ parameter readings, where $\{W_k\}, k = 1,2..n$ is a set of weights for each of the n disease-predicting parameters (to be determined by fitting the dataset to the model). The model is aiming to predict the probability the data given patient x will be diagnosed as heart disease, predict the weights of the features on finding the probability, and accuracy.

4. A HYBRID LINEAR REGRESSION MODEL

This proposed system Hybrid Linear Regression Model (HLRM) is developed by using techniques of data mining, namely, PCA, Linear Regression and Decision trees. The used techniques have its unique strength in realizing the objectives of the defined mining goals. Medical profiles of diseased such as sex, age, hypertension, blood sugar and other symptom are used for prediction. The model is designed to predict the possibility of patients getting a heart disease. This model enables us to obtain significant knowledge such as patterns, relationships among medical factors associated with heart disease [6], and its establishment.

HLRM is a user-friendly model, scalable, expandable reliable. It is implemented on the Python platform.

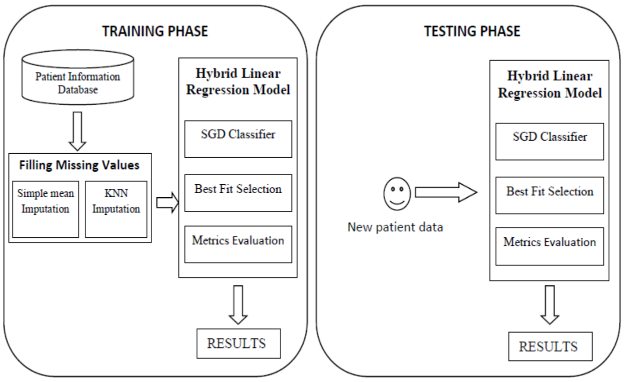


Figure 4.1 Architecture of the HLR Model

The architecture of Hybrid Linear Regression model shown in figure 4.1. The process is divided into two phases. In phase1 the training is conducted to understand the system. First, identify the missing values in the dataset, after that fill the missing values by using simple mean imputation and kNN algorithms. PCA based attributes selection method is used for dimensionality reduction and essential data collection. Secondly, the gradient descent is an iterative algorithm that begins from a random point of a function and travels down in steps to the lowest point of the function by applying the hybrid linear regression algorithm to find the results using the Stochastic Gradient Decent (SGD) process. The correct algorithm is always the difference between success and failure when you look at the best fit model for prediction. Thus the linear regression algorithm for prediction of outcomes has been chosen to experiment with HD datasets. After the experiment, lastly, the model is evaluated with various metrics such as R-Squared, Adjusted R-Squared, Mean Square Errors (MSE) and Root Mean Squared Errors (RMSE) etc. The testing of this model is done in phase2. In this phase, the new patient's data is tested based on the training. The same procedure of phase1 is repeated to get the outcome of the predicted value. And then these outcomes are analyzed and observed the new patterns in identifying of HD with good accuracy, and placed in a Decision Tree.

5. SIMULATION RESULTS AND DISCUSSION

The Cleveland heart disease dataset collected from the UCI repositories [18] and the missing values in it are filled with KNN mean imputation. Latter the dataset is analyzed and we identified gender reorder problem. The gender reorder problem is one in which the records belongs to male behave like female records and vice versa. To overcome this, gender-wise simple mean imputation is used instead of KNN imputation in the records where the gender reorders problem raised.Check all the attributes

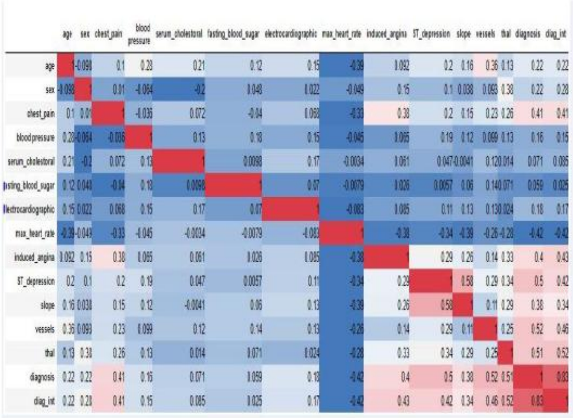


Figure 5.1 Correlation matrixes for Cleveland dataset

Even though the Cleveland dataset contains 14 attributes, but all of their contribution to the cause of heart disease is not equal. Some of the attributes contributions are very high and others contribution is very less. The most important contributed attributes of the dataset are identified by applying Principle Component Analysis (PCA) attribute selection method. It was found that 9 out of the 14 attributes are important and they are age, sex, blood pressure means, ST depression means, vessels mean, thal mean, fasting blood sugar, maximum heart rate and chest pain.

In the next stage, the prediction model is developed by applying a linear regression technique. The linear regression gives better results when the attributes used for developing model is less correlated. The complete correlation of each attribute versus other attributes of Cleveland dataset is calculated and given in matrix form in the above Figure 5.1. It was found that there are no correlations between each pair of attributes (symptoms). So, all the attributes can be used for linear regression model building. But for better efficiency of the model, only 9 attributes identified by PCA are used.

The proposed Hybrid Linear Regression Model was developed with Stochastic Gradient descent (SGD) approach and it is tuned with the four parameters such as loss function, penalty function, learning rate and no. of iterations. The values used in each of the parameters are given in the below table 1.

Table 1: Parameters used in Hybrid Linear Regression Model

Loss Function	Penalty Function	Alpha (Learning Rate)	Number of Iterations
Hinge Loss	L1 Penalty	0.1	500
Log Loss	L2 Penalty	0.05	1000

To measure the difference between expected and current values, the Loss function is used. The objective is to minimise the loss feature and direct the parameter updating process to improve the model efficiency. Two loss functions including hinge loss and log loss are used. Hinge loss uses an approach that is non-differentiated and unlikely while log loss uses a differing and probabilistic approach. Both L1 and L2 penalty features are included. The number of absolute differences (S) between the target values (Yi) and the expected values (f(xi)) is minimised by L1(also known as the least absolute deviations)

$$S = \sum_{i=1}^{n} |y_i - (f(x_i)| \qquad (11)$$

The penalty function L2 is used to minimise squares of the discrepancies (S) between the goal value (Yi) and the expected values (f(xi)). The penalty function is often considered to be least-squares (xi)**:**

$$S = \sum_{i=1}^{n}(y_i - (f(x_i))2 \qquad (12)$$

Table 2: Experimental results of Hybrid Linear Regression Model with different parameters

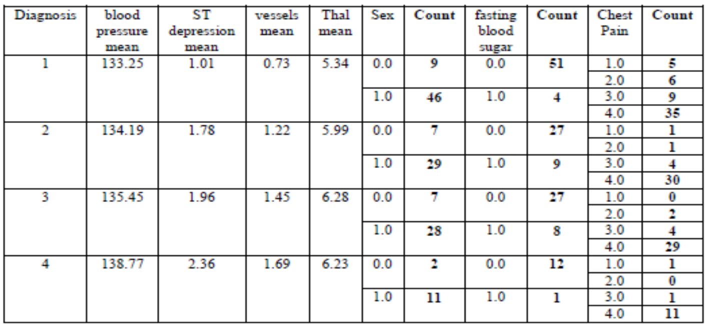
Parameters for Model				**Mean Score**	**Final Score**
Loss Function	**Penalty Function**	**Alpha (Learning Rate)**	**No. of Iterations**		
Hinge	L1	0.1	500	0.77479	0.82729
Hinge	L1	0.1	1000	0.77490	0.83504
Hinge	L1	0.05	500	0.77444	0.82569
Hinge	L1	0.05	1000	0.80447	0.86222
Hinge	L1	0.01	500	0.79780	0.85801
Hinge	***L1***	***0.01***	***1000***	***0.81459***	***0.89131***
Hinge	L2	0.1	500	0.66033	0.70233
Hinge	L2	0.1	1000	0.80135	0.87835
Hinge	L2	0.05	500	0.69618	0.75255
Hinge	L2	0.05	1000	0.74285	0.80585
Hinge	L2	0.01	500	0.69305	0.74832
Hinge	L2	0.01	1000	0.77566	0.84131
Log	L1	0.1	500	0.78501	0.85641
Log	L1	0.1	1000	0.77500	0.84553
Log	L1	0.05	500	0.79814	0.87091
Log	L1	0.05	1000	0.80103	0.88293
Log	L1	0.01	500	0.80769	0.88708
Log	L1	0.01	1000	0.79846	0.87623
Log	L2	0.1	500	0.77835	0.84975
Log	L2	0.1	1000	0.74943	0.82093
Log	L2	0.05	500	0.67439	0.7287

					1
Log	L2	0.05	1000	0.70307	0.76397
Log	L2	0.01	500	0.70832	0.76761
Log	L2	0.01	1000	0.69619	0.75376

The learning rate decides the speed of learning or tuning the function to predict accurately. We used 0.1, 0.01 and 0.05 learning rates. The experiment results for different combinations of parameters used in the development of the proposed model is given below.

Here all the possibilities of the above 4 parameters experimented and found **Loss function= *Hinge*, Penalty function = *L1*, Alpha= *0.01*** and **Iterations=*1000*** gives **89.13** as maximum accuracy. The dataset is analyzed with the proposed model and the following observations are made. The details are given below.

Table 3: Important observations of Hybrid Linear Regression Model on Cleveland Heart dataset



Diagnosis	blood pressure mean	ST depression mean	vessels mean	Thal mean	Sex	Count	fasting blood sugar	Count	Chest Pain	Count
1	133.25	1.01	0.73	5.34	0.0	9	0.0	51	1.0	5
									2.0	6
					1.0	46	1.0	4	3.0	9
									4.0	35
2	134.19	1.78	1.22	5.99	0.0	7	0.0	27	1.0	1
									2.0	1
					1.0	29	1.0	9	3.0	4
									4.0	30
3	135.45	1.96	1.45	6.28	0.0	7	0.0	27	1.0	0
									2.0	2
					1.0	28	1.0	8	3.0	4
									4.0	29
4	138.77	2.36	1.69	6.23	0.0	2	0.0	12	1.0	1
									2.0	0
					1.0	11	1.0	1	3.0	1
									4.0	11

It was observed that

- The greater the mean blood pressure, the greater the type of cardiac disease.
- The more severe the type of heart disease is the greater the ST depression.
- The wider the mean vessels, the higher the heart attack type.
- Because fasting blood sugar is false, the risk of all forms of heart disease is high.
- All forms of heart disease may occur in men more often than in females. Heart disease
- Heart attacks are frequently extricated by individuals with chest pain = 4

The proposed model has achieved satisfactory results with an accuracy of 94% for female and 87% for a male with an overall accuracy of **89.1%**. The proposed model also identified the idle conditions for heart disease. The ***Idle Condition*** for a person with heart disease is

- Age > 38,
- sex=man (gender ambiguous) with “asymptomatic chest pain”,
- Blood pressure > 112,
- Serum_cholestoral > 166 mg/dl,
- Fasting_blood_sugar <120 mg/dl,
- Resting Electrocardiographic showing probable or definite left ventricular hypertrophy by Estes’s criteria,

- Max_heart_rate > 114 (centered),
- ST_depression about 2In mm Hg,
- Slope (The slope of the peak exercise ST segment) is flat or down slopping but not up slopping,
- Vessels about 1.6 and Thallium heart scan more than fixed defect or reversible defect.

Table 4: Comparing the prediction accuracy of the proposed method with other methods

Author	**Proposed Year**	**Method Used**	**Accuracy**
Anooj PK	2012	Weighted Fuzzy rules	57.50%
Robert Detrano	2008	Logistic regression	77.00%
Tu, et al.	2009	J 48 Decision tree	78.90%
Srinivas Konda	2014	Rough-Fuzzy classifier	80.07%
Kirmani, Mudasirm	2017	Multilayer Perceptron	80.85%
Nguyen, Thanh, et al.	2015	wavelet transformation + interval type-2 fuzzy logic system	81.01%
Tan et al.	2009	GA+SVM	84.07%
Das et al.	2009	Neural Networks ensemble	89.01%
Proposed Method	2018	Hybrid Linear Regression	89.13%

Comparing with the eight other methods as seen in Table 4 above the cumulative projection efficiency of the proposed hybrid approach could be shown to be 89.13 percent versus 57.5 percent to 89.01 percent for other methods respectively. This also indicates that the proposed hybrid method would surpass the other approaches to heart disease prediction.

6. CONCLUSION

In this paper, a comparative study was made of various classifications for positive and negative participants for the classification of the Heart Disease data collection. K- Nearest Neighbor (K-NN), Stochastic Gradient Descent, (SGD) algorithms were used and Decision Tree classifiers were used, respectively. It was demonstrated that using various classification algorithms to classify the HD dataset provided very promising results about the accuracy of classification.

Accurate diagnosis of disease and administering effective treatments has become a big challenge for medical practitioners. The recognition of heart disease from diverse signs is a major problem which may encounter several false assumptions frequently accompanied by impulsive effects. Hence for the effective prediction of disease, an efficient approach is needed for extracting significant patterns from the health datasets. Most medical datasets are usually incomplete and possess missing data. Simply removing the missing data from the original datasets can bring more problems than solutions. Imputation kNN is used in the proposed HLR model for missing value to produce quality datasets for better diagnosis which create fitting regression equation for estimating the relationship between cause and effect. A simple mean is used where kNN imputation is anticipated to trigger gender reorder value effects in certain parts of the data collection. The model's output is complemented by an analysis of the area under the ROC curve. The model is complemented by There is a satisfactory overall estimate for the model. Regression has helped to explain synergistically how high the risk factors for serious cardiovascular diseases are. Moreover, more factors and patients with a particular condition must be taken into account in this work.

REFERENCES:

[1] Malterud, Kirsti. "The art and science of clinical knowledge: evidence beyond measures and numbers." The Lancet 358.9279 (2001): 397-400.

[2] Prather, Jonathan C., et al. "Medical data mining: knowledge discovery in a clinical data warehouse." Proceedings of the AMIA annual fall symposium. American Medical Informatics Association, 1997.

[3] Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on. IEEE, 2008.

[4] Srinivas, K., B. Kavitha Rani, and A. Govardhan. "Applications of data mining techniques in healthcare and prediction of heart attacks." International Journal on Computer Science and Engineering (IJCSE) 2.02 (2010): 250-255.

[5] Evanthia E. Tripoliti, Theofilos G. Papadopoulos, Georgia S. Karanasiou, Katerina K. Naka, Dimitrios I. Fotiadis, Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques, Computational and Structural Biotechnology Journal, Volume 15, 2017, Pages 26-47.

[6] M. Akhil Jabbar, B.L. Deekshatulu, Priti Chandra, Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, Procedia Technology, Volume 10, 2013, Pages 85-94.

[7] R. A. Carels. The association between disease severity, functional status, depression and daily quality of life in congestive heart failure patients. Quality of Life Research, 2004, 13 (1): 63–72.

[8] A. J. Hanley, G. Mckeown-Eyssen, et al. Cross-sectional and prospective associations between proinsulin andcardiovascular disease risk factors in a population experiencing rapid cultural transition. Diabetic Care, 1240-1247,24: 1240–1247.

[9] Srinivas, K., G. Raghavendra Rao, and A. Govardhan. "Rough-Fuzzy classifier: A system to predict heart disease by blending two different set theories." Arabian Journal for Science and Engineering 39.4 (2014): 2857-2868.

[10] Nguyen, Thanh, et al. "Medical data classification using interval type-2 fuzzy logic system and wavelets." Applied Soft Computing 30 (2015): 812-822.

[11] Anooj, P. K. "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules." Journal of King Saud University-Computer and Information Sciences 24.1 (2012): 27-40.

[12] Kirmani, Mudasirm. "Heart Disease Prediction using a Multilayer Perceptron Algorithm." International Journal of Advanced Research in Computer Science 8.5 (2017).

[13] Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." Expert systems with applications 36.4 (2009): 7675-7680.

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

[15] Crookston, Nicholas L., and Andrew O. Finley. "yaImpute: an R package for kNN imputation." Journal of Statistical Software. 23 (10). 16 p. (2008).

[16] Mukaka, Mavuto M. "A guide to the appropriate use of correlation coefficient in medical research." Malawi Medical Journal 24.3 (2012): 69-71.

[17] Zhang, Tong. "Solving large scale linear prediction problems using stochastic gradient descent algorithms." Proceedings of the twenty-first international conference on Machine learning. ACM, 2004

[18] Bache, Kevin, and Moshe Lichman. "UCI machine learning repository." (2013).

[19] Vembandasamy K, Sasipriya R, Deepa E. "heart diseases detection using naive Bayes algorithm", IJISET-international journal of innovative science. Eng Technol. 2015; 2:441–4.

[20] Krishnaiah V, Chandra NS. Heart disease prediction system using data mining techniques and intelligent fuzzy approach: a review. Int J Comput Appl. 2016;136(2):43–51.

[21] Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl. 2017;9(01):1.

[22] Lee HG, Noh KY, Ryu KH. “Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV,” Pacific-Asia Conference on Knowledge Discovery and Data Mining, Emerging Technologies in Knowledge Discovery and Data Mining; 2007. p. 218–28.

[23] Tarle B. An artificial neural network-based pattern classification algorithm for diagnosis of heart disease. Int Conf Comput Commun Control Automation (ICCUBEA). 2017:1–4.

[24] Saxenab K, Purushottam RS. Efficient Heart Disease Prediction System. Procedia Comput Sci. 2016;85:962–9.

[25] Karaylan T, Kilic O. Prediction of heart disease using neural network. Int Conf Comput Sci Eng (UBMK) Antalya. 2017; 2017:719–23.

[26] Esfahani HA, Ghazanfari M. “Cardiovascular disease detection using a new ensemble classifier”, IEEE 4th international conference on knowledge-based engineering and innovation (KBEI), Tehran, vol. 2017; 2017. p. 1011–4.

[27] Radhimeenakshi S. Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network. New Delhi: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom); 2016. p. 3107–11.

[28] Uyar K, Ilhan A. Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia Comput Sci. 2017; 120:588–93.

[29] Shah SMS, Batool S, Khan I. Muhammad Usman Ashraf, Syed Hussnain Abbas, Syed Adnan Hussain, “feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis”. Physica A. 2017; 482:796–807.

[30] https://www.analyticsvidhya.com/blog/2020/07/knnimputer-a-robust-way-to-impute-missing-values-using-scikit-learn/