ISSN: 2708-7123 | Volume-01, Issue Number-04 | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

A COMPARATIVE ANALYSIS ON CLEVELAND AND STATLOG HEART DISEASE DATASETS USING DATA MINING TECHNIQUES

Nabila Kausar[1], Dr. Hamid Ghous[2]
Institute of Southern Punjab (ISP), Multan, Pakistan
[1]milkywaymn555@gmail.com, [2]hamidghous@isp.edu.pk

ABSTRACT— In today's age deaths due to cardiovascular diseases are turn out to be a major problem. Factors such as high blood pressure, diabetes, high cholesterol level, hypertension, smoking and obesity are high risk to cause cardiovascular disease. Many researchers are using different datasets of heart patients to early diagnose the cardiovascular disease such as Cleveland and Statlog heart disease dataset. This study aims to compare the results of previous studies using Cleveland and Statlog heart disease datasets. We analyzed that different machine learning and deep learning techniques had been applied on these datasets which showed different resultson Cleveland and Statlog datasets.

Keywords— Cleveland dataset, Deep learning techniques, Machine learning techniques, Statlog dataset

I. INTRODUCTION

1.1 Flow of the Study:

In this research work, two heart disease datasets are taken that are most femeliar for the early diagnosis of cardiovascular disease. We compared the results of different data mining techniques on Cleveland and Statlog dataset.

1.2 Data Collection:

Two heart disease datasets are taken for the comparative study of different data mining techniques. Both datasets are taken from UCI machine learning repository contributed by Cleveland[https://archive.ics.uci.edu/ml/datasets/Heart+Disease] and Statlog dataset from the site https://github.com/renatopp/arff-datasets/blob/master/classification/heart.statlog.arff.

These two datasets: Cleveland heart disease dataset and Statlog heart disease dataset are used by many researchers for the prognosis of heart disease.

1.3 Description of Cleveland Dataset

Many researchers have been used Cleveland dataset in their researches for the diagnosis of cardiovascular disease. Cleveland dataset consists of 14 attributes in which one attribute is target value and 13 are input values and 303 instances. In 14 attributes, first two attributes 'age' and 'sex' are non-clinical features, one attribute is class and rest 11 attributes are clinical features.

The description of Cleveland dataset is given in table 1.1

First attribute 'age' describes the age of patients ranges from 29 to 77. Second attribute 'sex' which describes the patient's gender, 1 for male and 0 for female. Attribute 'cp' describes the chest pain type: 1 for typical angina, 2 for atypical angina, 3 for non angina pain and 4 for symptomic pain. 'trestbps' describes resting blood pressure of patients in mmHg. 'chol' describes the serum cholesterol of patients in mg/dl. 'Fbs' describes the fasting blood sugar of patients which should be greater than 120mg/dl ranges from 0 to 1, 0 for false and 1 for true. 'Restecg' describes the electrocardiographic results ranges from 0 to 2.'Thalach' describes the maximum heart rate achieved in patient ranges from 71 to 202. 'Exang' describes that exercise induced engina pain or not 1 for yes and 0 for no. 'oldpeak' tells the ST depression induced by exercise. 'slope' ranges from 1 to 3: 1 for upsloping, 2 for flat and 3 for downsloping. 'ca' describes the number of major vessels colored by fluoroscopy ranges from —100000 to 3. 'Thal' ranges from —100000 to 7; 3 for normal, 6 for fixed and 7 for reversible. 'num' represents the class attribute which describes the presence or absence of cardiovascular disease; 1 shows present and 0 shows absent.

ISSN: 2708-7123 | Volume-**01**, Issue Number-**04** | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

Sr No.	Feature	Description	Type	Range	Mean±StdDev
1	Age	Age of the patient	Real	29-77	54.439±9.039
2	Sex	Gender: 1 for male, 0 for female	Binary	0-1	0.68±0.467
3	Cp	Chest pain: 1=typical angina 2=atypical angina 3=non-angina pain 4=symptomic pain	Nominal	1-4	3.158±0.96
4	Trestbps	Resting blood pressure in mmHg	Numeric	94-200	131.69±17.6
5	Chol	Serum Cholesterol in mg/dl	Numeric	126-564	246.693±51.777
6	Fbs	Fasting blood sugar>120mg/dl True=1 False=0	Binary	0-1	0.149±0.356
7	Restecg	Resting electrocardiographic results	Nominal	0-2	0.99±0.995
8	Thalach	Maximum heart rate achieved	Numeric	71-202	149.607±22.875
9	Exang	Exercise induced angina:	Binary	0-1	0.327±0.47

ISSN: 2708-7123 | Volume-01, Issue Number-04 | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

		1=yes, 0=no			
10	Oldpeak	ST depression induced by exercise	Real	0-6.2	1.04±1.161
11	Slope	1=upsloping 2=flat 3=downsloping	Nominal	1-3	1.601±0.616
12	Ca	No. of major vessels colored by fluoroscopy	Nominal	_100000-3	_1319..469±11432.566
13	Thal	3=normal 6=fixed 7=reversible	Nominal	_100000-7	_655.363±8111.367
14	Num	CVD: 1=present 0=absent	Binary	0-1	0.459±0.499

Table 1.1: Description of Cleveland dataset

1.4 Description of Statlog Dataset

Many researchers have been also used Statlog dataset in their researches for the diagnosis of cardiovascular disease.

Statlog dataset consists of 14 attributes in which one attribute is target value and 13 are input values and 270 instances. In 14 attributes, first two attributes 'age' and 'sex' are non-clinical features, one attribute is class and rest 11 attributes are clinical features.

First attribute 'age' describes the age of patients ranges from 29 to 77. Second attribute 'sex' which describes the patient's gender, 1 for male and 0 for female. Attribute 'chest' describes the chest pain type: 1 for typical angina, 2 for atypical angina, 3 for non angina pain and 4 for symptomic pain. 'resting_blood_pressure' describes resting blood pressure of patients in mmHg. 'serum_cholestoral' describes the serum cholesterol of patients in mg/dl. 'fasting_blood_sugar' describes the fasting blood sugar of patients which should be greater than 120mg/dl ranges from 0 to 1, 0 for false and 1 for true. 'resting_elcetrocardiographic_results' describes the electrocardiographic results ranges from 0 to 2.'maximum_heart_rate_achieved' describes the maximum heart rate achieved in patients ranges from 71 to 202. 'exercise_induced_angina' describes that exercise induced engina pain or not 1 for yes and 0 for no. 'oldpeak' tells the ST depression induced by exercise. 'slope' ranges from 1 to 3: 1 for upsloping, 2 for flat and 3 for downsloping. 'number_of_major_vessels' describes the number of major vessels colored by fluoroscopy ranges from ─100000 to 3. 'Thal' ranges from ─100000 to 7; 3 for normal, 6 for fixed and 7 for reversible. 'class' represents the class attribute which

describes the presence or absence of cardiovascular disease; 1 shows present and 0 shows absent.

The description of Statlog dataset is given in table 1.2.

Sr No.	Feature	Description	Type	Range	Mean±StdDev
1	Age	Age of the patient	Real	29-77	54.433±9.109
2	Sex	Gender: 1 for male, 0 for female	Binary	0-1	0.678±0.468
3	Chest	Chest pain: 1=typical angina 2=atypical angina 3=non-angina pain 4=symptomic pain	Nominal	1-4	3.174±0.95
4	resting_blood_pressure	In mmHg	Numeric	94-200	131.344±17.862
5	serum_cholestoral	In mg/dl	Numeric	126-564	249.659±51.686
6	fasting_blood_sugar	Fasting blood sugar>120mg/dl True=1 False=0	Binary	0-1	0.148±0.356
7	resting_electrocardiographic_results	Resting electrocardiographic results	Numeric	0-2	1.022±0.998
8	maximum_heart_rate_achieved	Maximum heart rate achieved	Numeric	71-202	149.678±23.166
9	exercise_induced_angina	1=yes, 0=no	Binary	0-1	0.33±0.471

ISSN: 2708-7123 | Volume-**01**, Issue Number-**04** | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

10	Oldpeak	ST depression induced by exercise	Real	0-6.2	1.05±1.145
11	Slope	1=upsloping 2=flat 3=downsloping	Numeric	1-3	1.585±0.614
12	number_of_major_vessels	No. of major vessels in colored fluoroscopy	Numeric	0-3	0.67±0.944
13	Thal	3=normal 6=fixed 7=reversible	Numeric	3-7	4.696±1.941
14	Class	CVD: 1=present 0=absent	Nominal		

Table 1.2: Description of Statlog dataset

II. BACKGROUND

2.1 Cleveland Dataset

In this section, the previous work done in the field of heart disease diagnosis using Cleveland dataset is discussed. Many researchers used different data mining techniques on same heart disease dataset i.e, Cleveland but they achieved different accuracy results as disscussed below:

In [11], Luxmi Verma et al. (2016) used multinomial logistic regression (MLR), multilayer perceptron (MLP), fuzzy unordered rule induction algorithm (FURIA) and C4.5 on clinical data of 26 features and 335 instances to predict accuracy and incorrectly classified instances in prediction of coronary artery disease. They found that MLR has highest prediction accuracy which is 83.5%. After that they proposed hybrid method with correlation based feature subset selection (CFS) with particle swam optimization (PSO) search method to reduce the features. After applying CFS and PSO five features are selected. In this way accuracy of MLR is increased 0.67%. After feature selection, K mean clustering is applied so accuracy of MLR is increased to 88.4%. They also applied this proposed method on Cleveland data set with 14 features and 303 instances. After applying hybrid model, features are reduced to seven and accuracy is increased to 92.8%.

In [14], Verma and Srivastava (2016) collected the dataset from UCI machine repository contributed by Cleveland. They used 70% data for training the model and 30% data for testing the model. They used artificial neural network (ANN) based model to predict the coronary artery disease. They used Probabilistic Neural Network (PNN), alternating decision tree (ADTree) and RBF network to predict the CAD with more accuracy. They evaluated the performance of diagnostic model by measuring the difference between actual values and predicting values. They found that the prediction accuracy of PNN is higher than ADTree and RBFN which is 96% and misclassification rate is 4%. They also compared their model with other researchers work and found PNN has highest accuracy.

Amita Malav and Kalyani Kadam (2018) proposed a hybrid system with combination of K-means clustering algorithm and ANN data mining technique for the prediction of cardiovascular disease. They used Cleveland heart disease dataset. First they preprocessed the data and then implemented the hybrid system. They found that this hybrid system showed highest accuracy of 93.52% [17].

Burak et al. (2019) used two datasets; Cleveland and Alizadeh Sani heart disease datasets. They proposed an adaptive ensemble machine learning algorithm. They implemented K-nearest neighbor (KNN), logistic regression (LR), linear discriminant analysis (LDA), naïve bayes (NB), support vector machine (SVM) and ensemble method on two datasets. They found that ensemble method shows highest accuracy of 83.43% on Cleveland dataset and 88.38 % on Alizadeh Sani dataset [22].

Sujata Joshi and Mydhili K. Nair (2015) proposed a model to predict the heart disease by applying classification techniques on Cleveland dataset. They used three classification techniques Naïve Bayes, Decision Tree and K-Nearest Neighbour to predict the heart disease in patients. They used WEKA tool to analyze the dataset. They calculated correctly classified and incorrectly classified instances in dataset. They also calculated the specificity, sensitivity and accuracy on the base of these three techniques. They also compared their accuracy and found that the KNN shows highest accuracy than NB and DT [23].

Prashasti and Disha (2016) proposed a model for the prediction of heart disease in patients. They collected a data from UCI Repository having 15 attributes and 303 instances. They used some preprocessing techniques on dataset like noise removal, discarding records with missing data, filling default values if necessary and classification of attributes for decision making. They also develop a classifier using SVM and Naïve Bayes to predict patient having heart disease or not. They also calculate accuracy, specificity and sensitivity of classifier for training and testing data. They predicted that that SVM shows better accuracy than Naïve Bayes [25].

Sayad and Halkarnikar (2014) developed a system for diagnose of heart disease. They collected a dataset from Cleveland Clininc Foundation. They preprocessed the data by removing filled missing values and duplications. After cleaning the data, they implemented Multi-Layer Perceptron Neural Network (MLPNN) with Back-propagation algorithm for diagnosis of heart disease. After that they measure the performance of MLPNN with Back-propagation by calculated the accuracy, specificity and sensitivity of classifier. They calculated the accuracy 94%, specificity 92.5% and sensitivity 92% [26].

Subhadra and Vikas (2019) proposed a diagnostic system using MLPNN with Back-propagation for predicting heart disease. They also implemented Decision tree, logistic Regression, Naïve Bayes algorithm, Random Forest, Support vector machine, Generalized Linear Model, Gradient boosted trees and Deep learning on Cleveland dataset. They calculated accuracy, specificity, sensitivity and precision of every algorithm and compared these results with proposed algorithm MLPNN. They found that MLPNN shows better accuracy 94% as compared to others [27].

Bhaskura and Devi (2019) proposed a new automated system for accurate diagnose of heart disease, named as Hybrid Differential Evaluation based Fuzzy Neural Network (HDEFNN). The Cleveland heart disease dataset

ISSN: 2708-7123 | Volume-**01**, Issue Number-**04** | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

collected from UCI machine learning repository. First they normalized the dataset then applied the proposed method algorithms on normalized dataset. The simulation results performed in Matlab K-fold cross-validation. After applying HDEFNN, their results compared with J48, NB and RF with reference to accuracy. They found that HDEFNN shows better accuracy than other algorithms [28].

Anitha and Sridevi (2019), they used supervised machine learning algorithms as Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Naïve Bayes (NB) for the prediction of heart disease. They collected the Cleveland heart disease dataset from UCI dataset Repository and performed their experiments on this dataset using R language. They calculated their accuracy on the basis of confusion matrix. They found that KNN shows 76.67%, SVM shows 77.7% and NB shows highest accuracy which is 86.6% [32].

Shylaja and Muralidharan (2019) developed a hybrid classifier by hybridizing of Support Vector Machine (SVM) and Artificial Neural Network (ANN) classifiers for the prediction of heart disease. The Cleveland heart disease dataset was collected from UCI Repository and implemented the data mining techniques on this dataset as ANN, SVM, RIPPER, Decision Support, NB and hybrid SVM-ANN classifier. They performed all these experiments using MATLAB and calculated their accuracy, sensitivity and specificity. They showed that hybrid SVM-ANN is the best classifier with accuracy 88.54%, sensitivity 91.47% and specificity 82.11% [33].

Kathleen and Julia (2018) developed a Deep Neural Network (DNN) classification and prediction model to increase the diagnostic accuracy of heart disease. This model based on deep multilayer perceptron using deep learning technologies. They were collected Cleveland heart disease dataset from UCI machine learning Repository of datasets. They were implemented DNN model on this dataset and calculated diagnostic accuracy as 83.67%, probability of misclassification error 16.33%, sensitivity 93.51%, specificity 72.86%, precision 79.12%, F.score 0.8571, Area under the ROC curve (AUC) 0.8922, K-S test 66.62%, Diagnostic Odd Ratio (DOR) 38.65 and 95% confidence interval of DOR [34].

Latha and Jeeva (2019) proposed an ensemble technique for increasing the accuracy of some weak classifiers for the prediction of heart disease. They collected the Cleveland heart disease dataset from UCI machine learning Repository. They were performed some classification methods like Bayes Net, Naïve Bayes, Random Forest, C4.5, Multilayer Perceptron and PART on this dataset using WEKA tool and calculated their accuracy. After that they were used some ensemble classifiers as bagging, boosting, stacking and majority voting with weak classifiers on this dataset and calculated their accuracies again. They were improved the accuracy with majority voting classifier. After that they were implemented majority voting classifier with feature selection and found that highest accuracy was obtained by using this ensemble classifier [35].

Mustafa Jan et al. (2019) proposed an ensemble approach for the prediction of cardiovascular disease. They proposed ensemble approach by combining five classifiers as Support vector machine (SVM), artificial neural network (ANN), Naïve Bayesian (NB), Regression analysis and Random forest (RF). They used Cleveland and Hungarian dataset collected from UCI Repository for experiments. They were used WEKA tool for implementing the ensemble approach. They calculated highest accuracy 98.13% for Random Forest [42].

Marikani and Shyamala (2017) applied five supervised learning algorithms Decision Tree (DTree), Naïve Bayes (NB), K-Nearest Neighbor (KNN), random forest (RF) and support vector machine (SVM) on Cleveland heart disease dataset. These experiments are implemented using Orange tool. They calculated the accuracies of DT = 95.4%, NB = 81.7%, KNN = 75.7%, RF = 96.3% and SVM shows 100% [52].

Pushkala.V et al. (2019) applied five machine learning algorithms as DT, SVM, RF, KNN and NB on Cleveland

heart disease dataset for the prediction of heart disease. They implemented DT with and without Application Programming Interface (API). They analyzed and evaluated the accuracies of these classifiers. They showed that NB gives highest prediction accuracy which is 91% [55].

Sumit and Mahesh (2020) deployed a model to improve the prediction accuracy of heart disease using Cleveland heart disease dataset. They proposed an optimized Deep Neural Network (DNN) model using Talos. They also applied some other classification models such as Logistic Regression (LR), KNN, SVM NB, RF and then applied proposed model Hyper-parameter Optimization using Talos . The results showed that hyper parameter optimization using Talos performed better with accuracy of 90.78% [57].

Monther and Ossama (2019) proposed a hybrid model for the classification of heart disease. For this purpose, they collected Cleveland heart disease dataset from UCI machine learning repository. They applied classification algorithms such as J4.8, KNN, GA, SVM, RF and NN after preprocessing the dataset and selecting some distinct attributes from the dataset. Highest accuracy achieved by the hybrid model is 89.2% [60].

Radhanath and Bonomali (2019) collected Cleveland heart disease dataset for the classification of heart disease. First the dataset is preprocessed by applying feature selection methods such as entropy and information gain. Then applied decision tree J48, KNN, RBF and NB using WEKA tool. The results showed that J48 is best classifier with accuracy of 87.12%. Then they applied DT, KNN and SVM using Python and found that DT is best classifier with accuracy of 93.4% [62].

Karayilan and Kilic (2017) proposed a medical diagnosis system by using ANN with back propagation algorithm for the prediction of heart disease. They collected Cleveland heart disease dataset from UCI machine learning repository. They implemented their proposed model in MATLAB. They calculated the accuracy of this model which is 86% with original dataset. After that they applied PCA as dimensionality reduction method to reduce neurons of input layer for the performance improvements. After dimensionality reduction, they calculated the accuracy of proposed system which is 95% [63].

Hasan et al. (2018) applied different classification models on Cleveland heart disease dataset by using Anaconda Python (Spyder 3.6) for the diagnosis of heart disease. First they performed data preprocessing by using information gain feature selection method. Then they applied classification techniques such as KNN, DT (ID3), Gaussian NB, LR and RF on original dataset. They evaluated this model and showed that LR gives highest accuracy as 89.5%.After that they applied these techniques on 10 attributes of this dataset. The results showed that LR gives better accuracy of 92.76% [64].

S.Mohan et al. (2019) developed a hybrid model HRFLM (Hybrid Random Forest Linear Model) to enhance the prediction accuracy of model to predict the heart disease. The Cleveland heart disease dataset preprocessed by using R studio Rattle to perform cardiovascular disease classification. After preprocessing, they applied classification methods DT, RF and LM to classify the dataset. The results showed that RF and LM gives better accuracy than DT. So they combined RF and LM to develop hybrid model for the improvement of prediction accuracy. HRFLM shows highest accuracy of 88.4% [66].

R.Suganya et al. (2016) proposed novel feature selection method and applied data mining techniques on Celeveland heart disease dataset for the prediction of heart disease. They applied LR, RF, SMO, DT and NN on selected feature dataset. They found that NN shows better classification accuracy of 93% than SMO which has 89%. But AUC-ROC of SMO is greater than NN which is 0.887 and 0.812 respectively [70].

Rashmi and Kumar (2017) proposed a scalable framework for the prediction of heart disease using Cleveland heart disease dataset. They used Apache Spark and Hadoop plateform for the experiment. After feature selection, they applied random forest (RF) and naïve bayes (NB) on

ISSN: 2708-7123 | Volume-**01**, Issue Number-**04** | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

preprocessed dataset. They found that RF is best classifier with accuracy of 98% [74].

2.2 Comparison of previous researches using Cleveland dataset:

Below is the comparison table between different previous researches using Cleveland dataset. Different authors have used different techniques on Cleveland heart disease dataset in their study. They implemented different methods in different ways on Cleveland dataset to diagnose the heart disease and to classify the patients having heart disease or not. The accuracy comparison and techniques which are used by different researchers are given in table 2.1.

Ref No.	Year	Authors	Techniques	Accuracy
[11]	2016	Luxmi Verma et al.	CFS, PSO, K-mean clustering, MLP, MLR, FURIA, C4.5	92.8%
[14]	2016	Verma and Srivastava	PNN, ADTree, RBFN	96%
[15]	2015	Randa El-Bialy et al.	C4.5, FDT	78.06%
[17]	2018	Amita and Kalyani	K-means clustering, ANN	93.52%
[22]	2019	Burak et al.	KNN, LR, LDA, NB, SVM and ensemble method	83.43%
[23]	2015	Sujata Joshi and Mydhili K. Nair	DT,NB, KNN	100%
[25]	2016	Prashasti Kanikar and Disha Rajeshkumar Shah	SVM, NB	61%
[26]	2014	A.T.Sayad and P.P.Halkarnikar	MLPNN	94%
[27]	2019	K.Subhadra and Vikas	MLPNN	94%
[28]	2019	O.Bhaskura and M.Sree Devi	HDEFNN	69.1%

ISSN: 2708-7123 | Volume-**01**, Issue Number-**04** | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

[32]	2019	Dr.S.Anitha and Dr.N.Sridevi	SVM, KNN,NB	86.6%
[33]	2019	S.Shylaja and R.Muralidharan	Hybrid SVM-ANN	88.54%
[34]	2018	Kathleen H.Miao and Julia H.Miao	DNN	83.67%
[35]	2019	C. Beulah Christalin Latha and S. Carolin Jeeva∗	Ensemble classifier	85.48%
[38]	2020	Safial Islam Ayon, Md. Milon Islam & Md. Rahat Hossain	LR, SVM, DNN, DT, NB, RF and KNN	97.36%
[42]	2019	Mustafa Jan et al.	Ensemble approach	98.13%
[52]	2017	T.Marikani and K.Shyamala	DTree, NB, KNN, RF and SVM	100%
[55]	2019	Pushkala.V et al.	DT, SVM, RF, KNN and NB	91%
[56]	2016	Isra et al.	NB, DT, Discriminant, RF and SVM	99.01%
[57]	2020	Sumit Sharma and Mahesh Parmar	LR, KNN, SVM, NB, RF and Hyper parameter optimization with Talos	90.78%
[60]	2019	Monther Tarawneh and Ossama Embarak	J4.8, KNN, GA, SVM, RF and NN	89.2%

[62]	2019	Radhanath Patra and Bonomali Khuntia	Decision tree J48, KNN, RBF, NB and SVM	93.4%
[63]	2017	Tülay KarayÕlan and Özkan KÕlÕç	PCA + ANN with back propagation	95%
[64]	2018	Hasan et al.	KNN, DT(ID3), GNB, LR and RF	92.76%
[66]	2019	S.Mohan et al.	RF, LM, DT and HRFLM	88.4%
[70]	2016	R.Suganya et al.	LR, RF, SMO, DT and NN	93%
[72]	2012	Chaitrali and Sulabha	NN, DT and NB	100%
[74]	2017	Rashmi and Kumar	RF and NB	98%
[76]	2019	N.Satish et al.	KNN, RF, SVM, NB and NN	90 – 95%

Table 2.1: Comparison of previous researches using Cleveland dataset

At above, comparative analysis of previous researches using Cleveland heart disease dataset is given. This table shows the recent study of different scholars. During last decade, many researchers have been proposed different models that were implemented on Cleveland dataset. Some were used basic classifiers and some implemented these basic classifiers with the combination of different classifiers and different feature selection methods. Some were applied hybrid models on Cleveland dataset to improve the performance of the prediction models.

2.3 Statlog Dataset

In this section, the previous work done in the field of heart disease diagnosis using Statlog dataset is discussed. Many researchers used different data mining techniques on same heart disease dataset i.e, Statlog but they achieved different accuracy results as disscussed below:

In [4], Mukesh et.al (2018) used four classification algorithms, Naïve Bayes, Multilayer Perceptron, Random Forest and Decision Table to classify a patient. Patient is tested positive or negative for heart diseases based on some measurements included into the dataset. They also compare these four algorithms and found that Naïve Bayes has better accuracy for classification of heart disease. They found the maximum

accuracy of Naïve Bayes is 87.20% and minimum accuracy of Random Forest is 83.72% using confusion matrix.

Srabanti and Srishti (2019) used two techniques C4.5 and ANN for prediction of heart disease and develop a hybrid DT by combining ANN with C4.5. Hybrid DT implemented on same dataset and found that its accuracy is better than other two techniques which is 78%. They also found the sensitivity and specificity of C4.5, ANN and hybrid DT [8].

In [15], they collected four datasets having same problems with same disease. They apply C4.5 and fast decision tree (FDT) separately on these datasets with all features and compared the classification accuracy. Then they apply these trees on selected best featured datasets and compared the classification accuracy, execution time and tree size. They found that classification accuracy of selected featured dataset is 78.06% than the all featured dataset which is 75.48%.

Xiao Liu et al. (2017) proposed a system for the improvement of accuracy of diagnosis of heart disease. This system contains the RFRS feature selection system and hybrid classification system. In RFRS feature selection system, extracted the features on the basis of ReliefF algorithm and hybrid classification system with C4.5 classifier which classified the patients of heart disease. These systems implemented on Statlog heart dataset collected from UCI. The highest classification accuracy calculated 92.59% [41].

M.A.Jabbar et al. (2017) analyzed different ensemble classifiers with feature subset selection method as PSO. By using PSO, they reduced least ranked attributes by using Statlog heart disease dataset. After that they applied ensemble methods as Bagged tree, RF and AdaBoost for the improvement of prediction accuracy using R studio. They calculated the accuracy as Bagged tree of 100%, RF of 90.37% and AdaBoost of 88.89% [53].

Saranya and Manavalan (2019) proposed two classification models on Statlog heart disease dataset using MATLAB. Classification models DBN and FDBN, then analyzed by using 10-fold cross validation. They evaluated these classifiers and found that FDBN shows accuracy of 90.74%. They showed that FDBN model is best classification model for the prediction of heart disease [54].

Kernal and Umit (2019) collected two heart disease datasets such as SPECT and Statlog datasets for the prediction of heart disease. They implemented feature selection methods on datasets as Recursive Feature Elimination with Cross Validation (RFECV) and Stability Selection (SS) for the selection of best features. After that they applied Gradient Boosted Machines (GBM), NB and RF algorithms on original datasets and on selected features dataset. The results showed that NB provide highest accuracy for SPECT heart disease dataset provided by RFECV method and for Statlog heart disease dataset provided by SS method which is 77.78% and 86.42% respectively [59].

Moloud Abdar et al. (2015) applied feature analysis on Statlog heart disease dataset to improve the prediction accuracy of a model. After that they applied five algorithms such as C5.0, NN, SVM, KNN and LR on testing and training dataset. They evaluated and compared these algorithms with each other by measuring the accuracy, specificity, sensitivity and precision on both testing and training dataset. The results showed that C5.0 decision tree is the more efficient classifier with prediction accuracy of 93.02% [61].

Jabbar and Samreen (2016) developed a model for the prediction of heart disease. They collected Statlog heart disease dataset from UCI machine learning repository. They preprocessed the dataset by replacing missing values and by applying Inter Quartile Range (IQR). Then they applied Hybrid Naïve Bayes (HNB) on preprocessed dataset using WEKA tool. This preprocessed model shows the accuracy of 100% and overcomes the NB classifier shortfalls [65].

M.A.Jabbar (2017) proposed a model for the prediction of heart disease. He performed preprocessing methods on Statlog heart disease dataset to remove missing and noisy values. After that he used PSO as feature subset selection to reduce noisy features for the increase in performance of model. Then he applied KNN to original dataset and with PSO using WEKA. He evaluated his model and compared the results of KNN without feature subset selection and KNN with PSO. The results showed that accuracy of KNN without FSS is 75% and accuracy of KNN with PSO is 100% [67].

Lakshmi Devasena.C (2016) collected a Statlog heart disease dataset from UCI machine learning repository for the

prediction of heart disease. Then applied IBK classifier, K star classifier and Locally Weighted Learning (LWL) classifier with correlation based feature selection attribute evaluator (CFSAE) for the performance improvement of model. After that she evaluated these models by calculated accuracy. The results showed that IBK classifier, LWL classifier and K star classifier with CFSAE gives the accuracy of 100%, 80.74% and 99.62% respectively on training dataset [68].

Hidayet TAKCI (2018) used machine learning algorithms with feature selection algorithms to investigate best prediction model for the prediction of heart disease. For this purpose collected the Statlog heart disease dataset from UCI machine learning repository. For the selection of distinct features, applied some feature selection models such as regression model as Backward Logit (BL) and Forward Logit (FL), Fisher filtering (FF) and reliefF algorithm. Then he applied classification techniques such as DT, SVM, KNN, MLP, NB and LR models. After that he evaluated these algorithms by measuring accuracy using TANAGRA machine learning tool. The results showed that SVM linear model with reliefF feature selection method is the best prediction model with accuracy of 84.81% [69].

Kumar Pandey et al. (2013) implemented a novel feature selection method on Statlog heart disease dataset to reduce features for better prediction of heart disease. They proposed this method by combining attribute selected classifier (ASC) and maximal frequent pattern (MFP). After that they applied J48 decision tree using WEKA toolkit. They found that this model gives better performance than other traditional models [71].

Indu and Sunanada (2018) used feature selection method using Rough set and then applied random forest (RF), K-nearest neighbor (KNN) and naïve bayes (NB) on Statlog heart disease dataset for the prediction of coronary artery disease. They conducted this experiment using R studio. They applied Rough set feature selection on RF and calculated accuracy of 84%, Rough set on KNN and calculated accuracy of 71%, Rough set on NB which gives accuracy of 75%. So they observed that ensemble classifier RF with feature selection Rough set is best classifier [73].

M.A.Jabbar et al. (2016) proposed a classification model for the classification of heart disease. They implemented random forest (RF) as ensemble classifier with feature selection of chi square and genetic algorithm (GA) to prognosis of heart disease. They collected two heart disease datasets, T.S heart disease dataset from Hyderabad hospitals and heart Statlog dataset for the experiment. They found that RF with chi square shown highest accuracy of 83.7% on Statlog heart disease dataset [75].

2.4 Comparison of previous researches using Statlog dataset:

Below is the comparison table between different previous researches using Statlog heart disease dataset. Different authors have used different techniques on Statlog heart disease dataset in their study. They implemented different methods in different ways on Statlog dataset to diagnose the heart disease and to classify the patients having heart disease or not. The accuracy comparison and techniques which are used by different researchers are given in table 2.2.

Ref No.	Year	Authors	Techniques	Accuracy
[4]	2018	Mukesh et.al	NB, MLP, RF, DT	87.20%
[15]	2015	Randa El-Bialy et al.	C4.5, FDT	78.06%

[8]	2019	Srabanti and Srishti	C4.5, ANN and hybrid DT	78%
[38]	2020	Safial Islam Ayon, Md. Milon Islam & Md. Rahat Hossain	LR, SVM, DNN, DT, NB, RF and KNN	98.15%
[41]	2017	Xiao Liu et al.	RFRS, ensemble C4.5	92.59%
[53]	2017	M.A.Jabbar et al.	Bagged tree, RF and AdaBoost	100%
[54]	2019	S.Saranya and R.Manavalan	DBN and FDBN	90.74%
[56]	2016	Isra et al.	NB, DT, Discriminant, RF and SVM	98.15%
[59]	2019	Kernal Akyol and Umit Atila	Gradient Boosted Machines, NB and RF	86.42%
[61]	2015	Moloud Abdar et al.	Decision tree C5.0, NN, SVM, KNN and LR	93.02%
[65]	2016	M.A.Jabbar and Shirina samreen	IDR and HNB	100%
[67]	2017	Jabbar M.A	PSO + KNN	100%

[68]	2016	Lakshmi Devasena.C	IBK classifier, LWL classifier and K star classifier with CFSAE	100%
[69]	2018	Hidayet TAKCI	BL, FL, FF and reliefF + DT, SVM, KNN, MLP, NB and LR	84.81%
[71]	2013	Kumar Pandey et al.	ASC + MFP + J48	88%
[72]	2012	Chaitrali and Sulabha	NN, DT and NB	100%
[73]	2018	Indu and Sunanda	Rough set + RF, Rough set + KNN and Rough set + NB	84%
[75]	2016	M.A.Jabbar et al.	Chi square + RF	83.7%
[76]	2019	N.Satish et al.	KNN, RF, SVM, NB and NN	90 – 95%

Table 2.2: Comparison of previous researches using Statlog dataset

At above, comparative analysis of previous researches using Statlog heart disease dataset is given. This table shows the recent study of different scholars. During last decade, many researchers have been proposed different models that were implemented on Statlog dataset. Some were used basic classifiers and some implemented these basic classifiers with the combination of different classifiers and different feature selection methods. Some were applied hybrid models on Statlog heart disease dataset to improve the performance of the prediction models.

2.5 Both Cleveland and Statlog Dataset

In this section, the previous work done in the field of heart disease diagnosis using both Cleveland and Statlog dataset is discussed. Many researchers used different data mining techniques on two heart disease dataset i.e, Cleveland and statlog but they achieved different accuracy results as disscussed below:

Sankari et al. (2017) used two data mining techniques as Decision tree and Naïve Bayes to predict the heart diseases using combined form of Cleveland and Statlog dataset. They used J48 algorithm in decision tree. They used confusion matrix for classification. They found that decision tree has better classification accuracy than the Naïve Bayes [6].

In [15], Randa El-Bialy et al.(2015) collected four datasets having same problems with same disease. They applied C4.5 and fast decision tree (FDT) separately on these datasets with all features and compared the classification accuracy. Then

they apply these trees on selected best featured datasets and compared the classification accuracy, execution time and tree size. They found that classification accuracy of selected featured dataset is 78.06% than the all featured dataset which is 75.48%.

Safial Islam et al. (2020) implemented the seven classification techniques Logistic regression (LR), Support vector machine (SVM), Deep neural network (DNN), Decision tree (DT), Naïve bayes (NB), Random forest (RF) and K-nearest neighbor (K-NN) on two heart disease datasets Statlog and Cleveland. These datasets were collected from UCI machine learning repository to early diagnose the Coronary heart disease. They implemented these techniques using Python 3.0. They calculated accuracy, sensitivity, specificity, precision, NPV, F1 score and MCC by using five-fold and ten-fold cross validation. They found that DNN shows better accuracy 98.15% with Statlog dataset and SVM shows better accuracy 97.36% with Cleveland dataset using five-fold cross validation [38].

Isra et al. (2016) collected two heart disease datasets such as Statlog and Cleveland from UCI machine learning repository. First they preprocessed the datasets and then implemented five classifiers such as NB, DT, Discrminant, RF and SVM for prediction of heart disease using MATLAB. They evaluated all classifiers by measuring accuracy, specificity, precision, recall and F-measure on both datasets. They found that DT is best classifier which showed highest accuracy of 99.01% on Cleveland dataset and 98.15% on Statlog dataset [56].

Chaitrali and Sulabha (2012) collected two heart disease datasets as Cleveland ans Statlog for the analysis of data mining classification techniques on heart disease. Both datasets having 13 attributes but they added two more attributes as obesity and smoking in these datasets for the better performance. Then they applied NN, DT and NB on these two datasets. They calculated the accuracy of NN is 100%, DT is 99.62% and NB is 90.74%. They found that NN is best classifier than other classifiers [72].

N.Satish et al. (2019) applied KNN, SVM, RF, NB and NN with three percentage splits for the prediction of classification model of heart disease. They used Cleveland and Statlog heart disease datasets with the combination of three other datasets. They conducted their experiments using R- studio. They calculated the accuracy of RF with feature selection that ranges from 90 to 95% using three different percentage splits. They found that RF is more efficient classifier with better accuracy achieved [76].

III. CONCLUSION

This paper focuses on the comparison of data mining techniques on two heart disease datasets such as Cleveland and Statlog. In this assessment paper, we reviewed the previous studies conducted in the past ten years. In this study, we analyzed data mining techniques applied on Cleveland and Statlog datasets based on their results. We found that same data mining techniques showed different results on these datasets.

In future, there is still a need to explore these datasets with hybrid models using feature selction and machine learning methods.

IV. REFERENCES

[1] Sudhakar, K., & Manimekalai, D. M. (2014). Study of heart disease prediction using data mining. International journal of advanced research in computer science and software engineering, 4(1), 1157-1160.

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

[3] Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases. International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol, 2, 56-66.

[4] Kumar, M., Shambhu, S., & Sharma, A. (2018). Classification of heart diseases patients using data mining techniques.

[5] Kumar, M. N., Koushik, K. V. S., & Deepak, K. (2018). Prediction of heart diseases using data mining and machine learning algorithms and tools. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 3(3), 887-898.

[6] Karthiga, A. S., Mary, M. S., & Yogasini, M. (2017). Early prediction of heart disease using decision tree algorithm.

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST), 3, 1-17.

[7] Aziz, A., & Rehman, A. U. (2017). Detection of Cardiac Disease using Data Mining Classification Techniques. Int. J. Adv. Comput. Sci. Appl, 8(7), 256-259.

[8] Maji, S., & Arora, S. (2019). Decision tree algorithms for prediction of heart disease. In Information and Communication Technology for Competitive Strategies (pp. 447-454). Springer, Singapore.

[9] Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In Proceedings of the world Congress on Engineering and computer Science (Vol. 2, pp. 22-24).

[10] Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019, January). Improving heart disease prediction using feature selection approaches. In 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST) (pp. 619-623). IEEE.

[11] Verma, L., Srivastava, S., & Negi, P. C. (2016). A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. Journal of medical systems, 40(7), 178.

[12] Shaji, S. P. (2019, April). Predictionand Diagnosis of Heart Disease Patients using Data Mining Technique. In 2019 International Conference on Communication and Signal Processing (ICCSP) (pp. 0848-0852). IEEE.

[13] Mirmozaffari, M., Alinezhad, A., & Gilanpour, A. (2017). Heart disease prediction with data mining clustering algorithms. Int'l Journal of Computing, Communications & Instrumentation Engg, 4(1), 16-19.

[14] Verma, L., & Srivastava, S. (2016). A data mining model for coronary artery disease detection using noninvasive clinical parameters. Indian Journal of Science and Technology, 9(48), 1-6.

[15] El-Bialy, R., Salamay, M. A., Karam, O. H., & Khalifa, M. E. (2015). Feature analysis of coronary artery heart disease data sets. Procedia Computer Science, 65, 459-468.

[16] Van der Maaten, L. J. P. (2007). An introduction to dimensionality reduction using matlab. Report, 1201(07-07), 62.

[17] Malav, A., & Kadam, K. (2018). A hybrid approach for heart disease prediction using artificial neural network and K-means. International Journal of Pure and Applied Mathematics, 118(8), 103-10.

[18] Varun Sapra and Madan Lal Saini, "Computational Intelligence for Detection of Coronary Artery Disease with Optimized Features",International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-6C, April 2019

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

[20] Sudha, M. (2017). Evolutionary and neural computing based decision support system for disease diagnosis from clinical data sets in medical practice. Journal of Medical Systems, 41(11), 178.

[21] El Bialy, R., Salama, M. A., & Karam, O. (2016, May). An ensemble model for heart disease data sets: a generalized model. In Proceedings of the 10th International Conference on Informatics and Systems (pp. 191-196).

[22] Kolukisa, B., Yavuz, L., Soran, A., Bakir-Gungor, B., Tuncer, D., Onen, A., & Gungor, V. C. (2020). Coronary Artery Disease Diagnosis Using Optimized Adaptive Ensemble Machine Learning Algorithm.

[23] Joshi, S., & Nair, M. K. (2015). Prediction of heart disease using classification based data mining techniques. In Computational Intelligence in Data Mining-Volume 2 (pp. 503-511). Springer, New Delhi.

[24] Madhura Patil, Rima Jadhav, Vishakha Patil, Aditi Bhawar and Mrs. Geeta Chillarge, "Prediction and Analysis of Heart Disease Using

SVM Algorithm", International Journal for Research in Applied Science & Engineering Technology (IJRASET)

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887

Volume 7 Issue I, Jan 2019- Available at www.ijraset.com

[25]Kanikar, P., & Shah, D. R. (2016). Prediction of cardiovascular diseases using support vector machine and Bayesien classification. International Journal of Computer Applications (0975–8887), 156(2).

[26] A. T. Sayad and P. P. Halkarnikar, "Diagnosis of heart disease usnig neural network approach",International Journal

of Advances in Science Engineering and Technology, ISSN: 2321-9009 Volume- 2, Issue-3, July-2014

[27] Subhadra, K., & Vikas, B. (2019). Neural network based intelligent system for predicting heart disease. Int. J. Innov. Technol. Exploring Eng.(IJITEE), 8(5), 484-487.

[28] O. Bhaskaru and M.Sree Devi, "Accurate and Fast Diagnosis of Heart Disease using Hybrid Differential Neural Network Algorithm", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-3S, February 2019

[29] Chaithra, N., & Madhu, B. (2018). Classification models on cardiovascular disease prediction using data mining techniques. Journal of Cardiovascular Diseases and Diagnosis. doi, 10, 2329-9517.

[30] Shekar, K. C., Chandra, P., & Rao, K. V. (2019). An Ensemble Classifier Characterized by Genetic Algorithm with Decision Tree for the Prophecy of Heart Disease. In Innovations in Computer Science and Engineering (pp. 9-15). Springer, Singapore.

[31] Singh, R., & Rajesh, E. (2019). Prediction of Heart Disease by Clustering and Classification Techniques. International Journal of Computer Sciences and Engineering, 7, 861-866.

[32] Dr. S. Anitha and Dr. N. Sridevi, "Heart Disease Prediction Using data Mining Techniques" Journal of Analysis and Computation (JAC) (An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861 Volume XIII, Issue II, February 2019

[33] S.Shylaja and R. Muralidharan, "Hybrid SVM-ANN Classifier is used for Heart Disease Prediction System", International Journal of Engineering Development and Research (www.ijedr.org), © IJEDR 2019 | Volume 7, Issue 3 | ISSN: 2321-9939

[34] Miao, K. H., & Miao, J. H. (2018). Coronary heart disease diagnosis using deep neural networks. Int. J. Adv. Comput. Sci. Appl., 9(10), 1-8.

[35] Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, 100203.

[36] Ricciardi, C., Valente, A. S., Edmund, K., Cantoni, V., Green, R., Fiorillo, A., ... & Cesarelli, M. (2020). Linear discriminant analysis and principal component analysis to predict coronary artery disease. Health Informatics Journal, 1460458219899210.

[37] Joloudari, J. H., Hassannataj Joloudari, E., Saadatfar, H., GhasemiGol, M., Razavi, S. M., Mosavi, A., ... & Nadai, L. (2020). Coronary artery disease diagnosis; ranking the significant features using a random trees model. International journal of environmental research and public health, 17(3), 731.

[38] Ayon, S. I., Islam, M. M., & Hossain, M. R. (2020). Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE Journal of Research, 1-20.

[39] Ramasamy, S., & Nirmala, K. (2020). Disease prediction in data mining using association rule mining and keyword based clustering algorithms. International Journal of Computers and Applications, 42(1), 1-8.

[40] Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC medical informatics and decision making, 19(1), 211. [41] Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q. (2017). A hybrid classification system for heart disease diagnosis based on the RFRS method. Computational and mathematical methods in medicine, 2017.

[42] Jan, M., Awan, A. A., Khalid, M. S., & Nisar, S. (2018). Ensemble approach for developing a smart heart disease prediction system using classification algorithms. Research Reports in Clinical Cardiology, 9, 33-45.

[43] K.Gomathi and dr. Shanmugapriyaa, "Heart Disease Prediction Using Data Mining Classification", International Journal for Research in Applied Science & Engineering Technology (IJRASET),www.ijraset.com Volume 4 Issue II, February 2016IC Value: 13.98 ISSN: 2321-9653

[44]Kirmani, M. M., & Ansarullah, S. I. (2016). Classification models on cardiovascular disease detection using Neural Networks, Naïve Bayes and J48 Data Mining Techniques. International Journal of Advanced Research in Computer Science, 7(5).

[45] Kautkar Rohit A, "A Comprehensive Survey on Data Mining", IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308 , Volume: 03 Issue: 08 | Aug-2014, Available @ http://www.ijret.org

[46] S. Kiruthika Devi*, S. Krishnapriya and Dristipona Kalita, "Prediction of Heart Disease Using Data Mining Techniques", Indian Journal of Science and Technology Vol, 9(39), DOI: 10.17485/ijst/2016/v9i39/102078, October 2016 , ISSN (Print) : 0974-6846

ISSN (Online) : 0974-5645

[47] H. Benjamin Fredrick David and S. Antony Belcy, "Heart Disease Prediction Using Data Mining Techniques", ICTACT journal on soft computing, October 2018, Volume: 09, ISSUE: 01, ISSN: 2229-6956 (Online) DOI: 10.21917/ijsc.2018.0253

[48] Nikookar, E., & Naderi, E. (2018). Hybrid Ensemble Framework for Heart Disease Detection and Prediction. IJACSA) International Journal of Advanced Computer Science and Applications, 9(5).

[49] KS, D., & Kamath, A. (2017). Survey on Techniques of Data Mining and its Applications.

[50] Sharma, A., Sharma, R., Sharma, V. K., & Shrivatava, V. (2014). Application of data mining–a

survey paper. International Journal of Computer Science and Information Technologies, 5(2),

2023-2025.

[51] Ghorbani, R., & Ghousi, R. (2019). Predictive data mining approaches in medical diagnosis: A review of some

diseases prediction. International Journal of Data and Network Science, 3(2), 47-70.

[52] Marikani, T., & Shyamala, K. (2017). Prediction of heart disease using supervised learning

algorithms. Int J Comput Appl, 165(5), 41-4.

[53] Yekkala, I., Dixit, S., & Jabbar, M. A. (2017, August). Prediction of heart disease using ensemble learning and Particle Swarm Optimization. In 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon) (pp. 691-698). IEEE.

[54] Saranya, S., & Manavalan, R. Computational Framework for Heart Disease Prediction using

Deep Belief Neural Network with Fuzzy Logic. International Journal of Computer Applications, 975, 8887.

[55] PushkalaV, Agalya T and S A Angayarkanni, "Comparative Study of Heart Disease Prediction Using Machine Learning Algorithms", International Journal of Innovations in Engineering and Technology (IJIET) http://dx.doi.org/10.21172/ijiet.124.10 , Volume 12 Issue 4 March 2019, ISSN: 2319-1058

[56] Zriqat, I. A., Altamimi, A. M., & Azzeh, M. (2017). A comparative study for predicting heart

diseases using data mining classification methods. arXiv preprint arXiv:1704.02799.

[57] Sharma, S., & Parmar, M. (2020). Heart Diseases Prediction using Deep Learning Neural

Network Model. International Journal of Innovative Technology and Exploring Engineering

(IJITEE), 9(3).

[58] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, 7(2.8), 684-687.

[59] Kemal Akyol and Ümit Atila, "A Study on Performance Improvement of Heart Disease Prediction by Attribute Selection Methods", Academic Platform Journal of Engineering and Science 7-2, 174-179, 2019

[60] Tarawneh, M., & Embarak, O. (2019, February). Hybrid approach for heart disease prediction using data mining techniques. In International Conference on Emerging Internetworking, Data & Web Technologies (pp. 447-454). Springer, Cham.

[61] Abdar, M., Kalhori, S. R. N., Sutikno, T., Subroto, I. M. I., & Arji, G. (2015). Comparing

Performance of Data Mining Algorithms in Prediction Heart Diseases. International Journal of

Electrical & Computer Engineering (2088-8708), 5(6).

[62] Patra, R., & Khuntia, B. (2019, February). Predictive Analysis of Rapid Spread of Heart Disease with Data Mining. In 2019 IEEE International Conference on

ISSN: 2708-7123 | Volume-**01**, Issue Number-**04** | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

Electrical, Computer and Communication Technologies (ICECCT) (pp. 1- 4). IEEE.

[63] Karayılan, T., & Kılıç, Ö. (2017, October). Prediction of heart disease using neural network.

In 2017 International Conference on Computer Science and Engineering (UBMK) (pp. 719-723).

IEEE.

[64] Hasan, S. M. M., Mamun, M. A., Uddin, M. P., & Hossain, M. A. (2018, February). Comparative Analysis of Classification Approaches for Heart Disease Prediction. In 2018 International

Conference on Computer, Communication, Chemical, Material and Electronic Engineering

(IC4ME2) (pp. 1-4). IEEE.

[65] Jabbar, M. A., & Samreen, S. (2016, October). Heart disease prediction system based on hidden naïve bayes classifier. In 2016 International Conference on Circuits, Controls, Communications and Computing (I4C) (pp. 1-5). IEEE.

[66] S. Mohan et al., "Effective Heart Disease Prediction Using Hybrid ML Techniques", VOLUME 7, 2923707, 2019, IEEE Access

[67] Jabbar, M. A. (2017). Prediction of heart disease using k-nearest neighbor and particle swarm

optimization.

[68] Lakshmi Devasena. C, "Performance Evaluation of Memory Based Classifiers

With Correlation Based Feature Selection SubsetEvaluator For Smart Heart Disease Prediction", IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308

[69] Takci, H. (2018). Improvement of heart attack prediction by the feature selection methods. Turkish

Journal of Electrical Engineering & Computer Sciences, 26(1), 1-10.

[70] R.Suganya et al., "A Novel Feature Selection Method for Predicting Heart Diseases with Data

Mining Techniques", Asian Journal of Information Technology 15 (8): 1314-1321, 2016, ISSN:

1682-3915

[71] Pandey, A. K., Pandey, P., & Jaiswal, K. L. (2013). A novel frequent features prediction model for heart disease diagnosis. International Journal of Engineering Mathematics and Computer Sciences, 1(2).

[72] Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-48.

[73] Yekkala, I., & Dixit, S. (2018). Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection. International Journal of Big Data and Analytics in Healthcare (IJBDAH), 3(1), 1-12.

[74] Saboji, R. G. (2017, August). A scalable solution for heart disease prediction using classification mining technique. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 1780-1785). IEEE.

[75] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2016). Intelligent heart disease prediction system using random forest and evolutionary approach. Journal of Network and Innovative Computing, 4(2016), 175- 184.

[76] Reddy, N. S. C., Nee, S. S., Min, L. Z., & Ying, C. X. (2019). Classification and feature selection approaches by machine learning techniques: Heart disease prediction. International Journal of Innovative Computing, 9(1).