A Comparative Examination of Methodical Approaches for Machine Learning-based Heart Disease Prediction

Asha Dilipkumar Jariwala
Computer Science
Vanita Vishram women's university
Surat,Gujarat

Hemangini Patel, PhD
Computer Science
Vanita Vishram women's university
Surat,Gujarat

ABSTRACT
Heart disease is seen as a contemporary epidemic. People frequently disregard their health as a result of modern lives and work-related stress, which leads to an increase in a number of health problems. Among these, cardiovascular disease has become one of the most common and dangerous illnesses. Numerous risk factors, including diabetes, high blood pressure, high cholesterol, irregular pulse rate, and other associated medical disorders, make heart disease prediction difficult. The primary objective is to identify and process heart-related data in order to identify cardiac problems and save lives. To predict cardiac disease, In this paper, machine learning methods such as KNN, SVM, NB, RF, LR, DT, RF + SVM, RF + DT, RF + KNN, and HRLFM. The UCI repository and Kaggle are the sources of the dataset used to train and evaluate the prediction model. Compare to all other model HRLFM (Hybrid random forest and logistic regression) is outperformed

Keywords— Heart Disease, Linear Model, Random Forest Model, Hybrid Model, Machine learning methods, cardiovascular disease prediction

1.INTRODUCTION
The heart, often considered the second most important organ after the brain, is vital to the body's general functioning. One of the most important organs in the human body is the heart, and any disruption to its operation can have a big impact on general health. Heart disease is sometimes referred to as a "silent killer" since symptoms may not show up right away, making it one of the leading causes of mortality globally. Therefore, early detection and prompt treatment of heart disease are crucial, and encouraging high-risk persons to lead better lives can help minimize difficulties in the future [2]. The primary goal of machine learning in medicine is to provide patients with chronic illnesses with reliable clinical procedures. However, chronic respiratory disorders account for 1.59 million deaths. Different categories of heart disease:

1. Cutting off the blood supply causes stroke damage to the brain.
2. Heart failure that is congestive -The heart's capacity to pump blood effectively is hampered by chronic disease.
3. Coronary artery disease -The main blood vessels in the heart are ill or damaged.
4. Heart arrest: Breathing and consciousness are suddenly cut off.
5. Blood Pressure -It malfunctions because the blood pressure on the artery walls is too high.[1]

According to [3], angiography is one of the most popular medical diagnostic techniques for identifying heart disease since it makes it easier tosee how blood flows through arteries and spot any anomalies or blockages. There are two types of risk factors: Age, family history, and identity are examples of uncontrollable risk factors. Hyperlipidaemia, tobacco use, high blood cholesterol, obesity, higher blood pressure, elevated glucose levels, and being overweight are examples of controllable risk factors. In order to manage and prevent CVD, it is crucial to comprehend these aspects. Learning from data is the main objective of machine learning, which uses a variety of methods to solve data-related problems [5]. Algorithms for machine learning are divided into supervised and unsupervised techniques. The algorithm in supervised learning makes use of labelled data, or trained data. The algorithm employs untrained data, or unlabelled data, in unsupervised learning. In Table 1 Dataset heart dataset from UCI. repository. It focuses on more clinical data

2. DATA AND METHODS
Table 1 Dataset (70,000*12)

Column Name	Feature Name	Category
age	Age	Objective Feature
height	Height	Objective Feature
weight	Weight	Objective Feature
gender	Gender	Objective Feature
ap_hi	Systolic Blood Pressure	Examination Feature
ap_lo	Diastolic Blood Pressure	Examination Feature
cholesterol	Cholesterol	Examination Feature
gluc	Glucose	Examination Feature
smoke	Smoking	Subjective Feature
alco	Alcohol Intake	Subjective Feature
active	Physical Activity	Subjective Feature
cardio	Presence or Absence of Cardiovascular Disease	Target Variable

In Table 1. Cardiovascular dataset, which focus more on blood pressure and lifestyle. In Table 2, dataset is more focus on clinical data of heart disease

Table 2 DATASET

Attribute	Type	Description
Age	Rea1	Age in Years
Sex	Binary	Value 1: Male Value 0: Female
Chest Pain Type	Nominal	Value 1: Typical angina Value 2: Atypical angina Value 3: Non-anginal pain Value 4: Asymptomatic

Resting Blood Pressure	Numeric	(in mm Hg on admission to the hospital)
Serum Cholesterol	Numeric	Serum cholesterol in mg/dl
Fasting Blood Sugar	Binary	If Value > 120 mg/dlValue 1: True Value 0: False
Resting Electrocardiographic Result	Nominal	Value 0: normal Value 1: having ST-T wave abnormality Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Maximum Heart Rate	Numeric	Maximum heart rate achieved
Exercise Induced Angina	Binary	Value 1: YesValue 0: No
Old Peak	Real	ST depression induced induced byexercise relative to rest
Slope of peak exerciseST segment	Nominal	Value 1: Upsloping Value 2: Flat Value 3: Down sloping
Number of major vesselscolored by fluoroscopy	Nominal	Range: 0-3

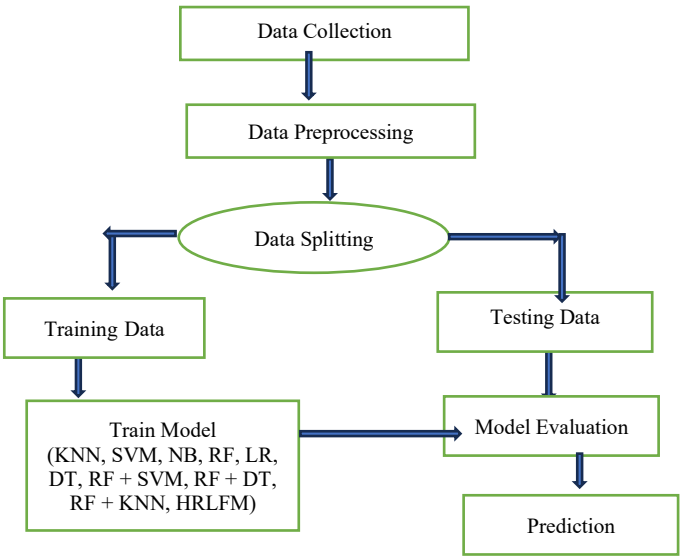

Figure 1. Flowchart of Methodology

3. LITERATURE REVIEW

The heart is vital to human health, and cardiovascular disease remains one of the leading causes of death worldwide. Heart disease can significantly improve patient outcomes and reduce mortality rates with early diagnosis and treatment. Researchers are increasingly using artificial intelligence and machine learning techniques to predict heart illness. Several algorithms have shown promising prognostic results, such as Random Forest, Support Vector Machine, K-Nearest Neighbours, Decision Tree, and Naive Bayes. Combining these algorithms into hybrid models can further increase the precision and reliability of heart disease prediction systems. Cardiovascular disorders account for around 17.5 million deaths yearly, making them one of the major causes of death globally, according to the World Health Organization. The Naïve Bayes classifier performs better on datasets with fewer characteristics, according to comparative study [1]. K-Nearest Neighbors (KNN) achieved the highest overall accuracy [2]. The HRFLM technique combines Random Forest (RF) and Linear Method (LM) for incredibly accurate heart disease prediction. To improve prediction accuracy even more, future research can investigate different machine learning combinations [4]. The best accuracy of 98.53% was attained by Random Forest and the study recommended applying ensemble techniques, feature selection, and hyper-parameter tuning to further improve performance[6].

In [7] Random Forest was shown to be the best successful classi fier for heart disease prediction among KNN, Logistic Regressi on, Naive Bayes, Decision Tree, and Random Forest approache sbecause of its high accuracy and resilience.

The study in [8] found that various datasets yielded varied outcomes from the same data mining approaches, indicating the necessity for hybrid models that combine sophisticated machine learning techniques with feature selection.

The HRFLM approach, which combines Random Forest and Logistic Regression, had notable efficacy in the diagnosis of cardiac attacks in [9]. With an accuracy of 89.50% prior to tuning and 94.96% following tuning utilizing data from Statlog, Cleveland, and Hungary clinics, Random Forest fared better than other classifiers in [10].

The K-Nearest Neighbour (KNN) method produced the best prediction accuracy for the identification of cardiac disease in [12]. Several machine learning methods were used in [13], with Random Forest and MLP achieving the greatest accuracy of 86.96%. The Decision Tree Classifier obtained the highest accuracy of 92% out of five machine learning algorithms in [14]. In [16], Advanced ensemble techniques like Gradient Boosting and XGBoost performed better than conventional models with XGBoost attaining the greatest accuracy of 90%.

4. METHODOLOGY

4.1 Support Vector Machine (SVM)

A supervised machine learning approach called Support Vector Machine (SVM) is primarily employed for classification tasks. It divides several data classes by constructing an ideal hyperplane in an n-dimensional space. In order to increase classification efficiency and accuracy, SVM seeks to maximize the margin between classes [12]

4.2 K-nearest neighbor (K-NN)

One supervised machine learning technique for classification tasks is K-Nearest Neighbor (KNN). By comparing fresh data with trained data using similarity metrics, it forecasts results. By calculating distances between feature points, such as the Euclidean distance, the method classifies unlabelled data [12]

4.3. Random Forest

A supervised machine learning approach called Random Forest can be applied to both regression and classification problems. It is based on ensemble learning, which combines several decision trees to enhance model performance and prediction accuracy. The output of each tree is combined to establish the final prediction, typically by majority vote [23]

4.4. Decision Tree

A supervised machine learning approach called Decision Tree uses decision nodes, chance nodes, and end nodes to represent data in a tree-like form. To produce precise and understandable forecasts, it divides data into various branches according to choices and probabilities [12].

4.5. Logistic Regression

A supervised machine learning technique for binary classification issues is called logistic regression. It divides findings into two groups, usually denoted as 0 and 1, and forecasts the likelihood of an event depending on input data [12]

4.6. Naïve Bayes

Based on Bayes' theorem, Naïve Bayes is a supervised machine learning technique that is mostly applied to classification issues. It is a straightforward and effective probabilistic classifier that works well on high-dimensional datasets and is frequently utilized in text categorization, sentiment analysis, and spam filtering [23].

4.7. Hybrid Ensemble Model (HRFLM).

*Fig 4*shows that combines the strengths of LR and RF use classifier of soft voting. It uses each model's probability. The Random Forest (RF) model's prediction is given twice as much importance as the Logistic Regression (LR) prediction (because of the weight ratio 2:1).

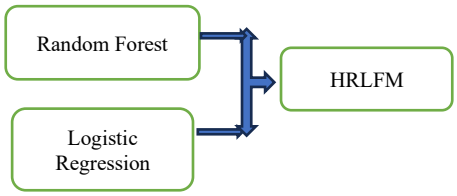

Figure 2. HRLFM (Hybrid RF+LR)

5.EXPERIMENTAL RESULT

In Table 3, The following results are from the 1st data shape. The data shape is (296, 14)

Table 3: Statistics of Algorithms for the 1st Data shape (Dataset2)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
HRFLM (RF + LR)	88.76	89.47	85.0	87.18
Logistic Regression	87.64	89.19	82.5	85.71
Naive Bayes	86.52	85.00	85.0	85.00
Random Forest	82.02	77.27	85.0	80.95
Decision Tree	75.28	69.57	80.0	74.42
SVM	74.16	77.42	60.0	67.61
k-nearest neighbor	61.80	57.50	57.5	57.50
RF + SVM	82.02	78.57	82.5	80.49
RF + DT	79.78	72.92	87.5	79.55
RF + KNN	78.65	74.42	80.0	77.11

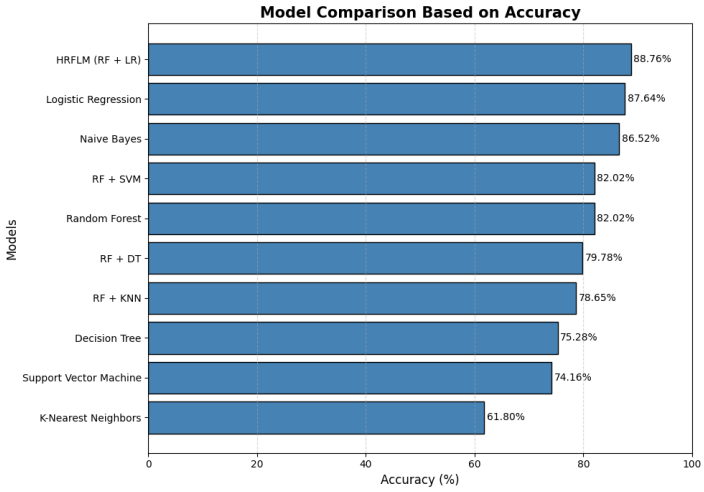

Figure 3. Model comparison with (296,14) data shape

In Figure 3, it shows highest accuracy we got using (296,14) data shape which is 88.76% using HRLFM model.

Table 4 Statistics of Algorithms for the 2nd Data shape (Dataset2)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	98.54	100.00	97.09	98.52
Random Forest	98.54	100.00	97.09	98.52
HRFLM (RF + LR)	98.54	97.17	100.00	98.56
RF + SVM	98.54	100.00	97.09	98.52
RF + DT	98.54	100.00	97.09	98.52
RF + KNN	95.61	94.34	97.09	95.69
Naive Bayes	80.00	75.41	89.32	81.78
Logistic Regression	79.51	75.63	87.38	81.08
K-Nearest Neighbors	73.17	73.08	73.79	73.43
SVM	68.29	66.10	75.73	70.59

In Table 4, The following results are from the second data shape. The data shape is (1025,14). In Figure 4, it shows highest accuracy using (1025,14) data shape which is 98.54% using HRLFM model.

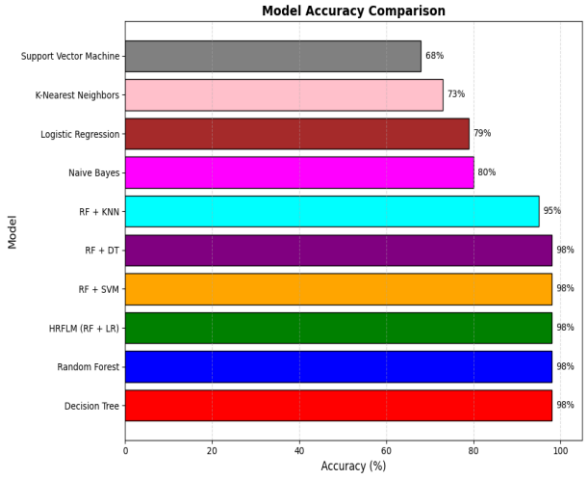

Figure 4. Model comparison with (1025,14) data shape

In Table 5, The following results are from the 1st data shape. The Dataset Shape is (2518, 12). In Table 5,In this dataset 'ca 'and 'thal' are not there. There are Total 12 attributes. In Figure 5, it shows highest accuracy using (1025,14) data shape which is 81.97% using HRLFM model.

Table 5 Statistics of Algorithms for the 3rd Data shape (Dataset2)

Model	Accuracy
HRLFM	81.97%
SVM	81.56%
Random Forest	81.56%
Logistic Regression	79.51%
AdaBoost	78.28%
KNN	77.46%
Decision Tree	75.00%

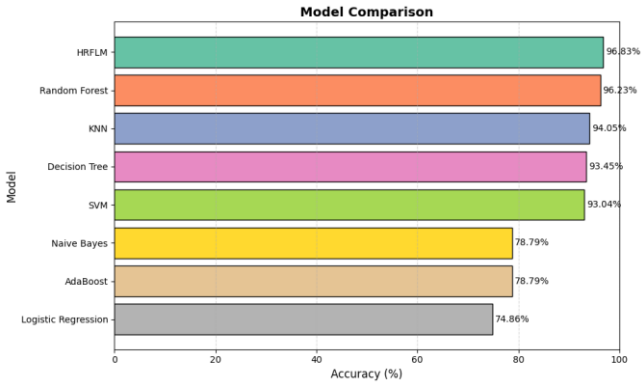

Figure 5. Model comparison with (2518,12) data shape

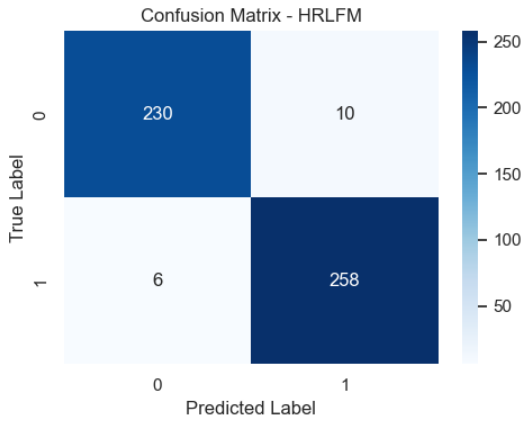

Figure 6. Confusion Matrix (2518,12) data shape

In Figure 6, Where **TP** (True Positive)1, **TN** (True Negative)0, **FP** (False Positive)1, and **FN** (False Negative)0. The number of real instances of the class in the given dataset is known as support.

Classification Report - HRLFM:

	precision	recall	f1-score	support
0	0.97	0.96	0.97	240
1	0.96	0.98	0.97	264
accuracy			0.97	504
macro avg	0.97	0.97	0.97	504
weighted avg	0.97	0.97	0.97	504

Classification Report – HRLFM it shows F1-score (F1), sensitivity (recall) (REC), specificity, accuracy (ACC), and precision (PREC) are the most widely utilized evaluation metrics. These are the results of their calculations: [1].

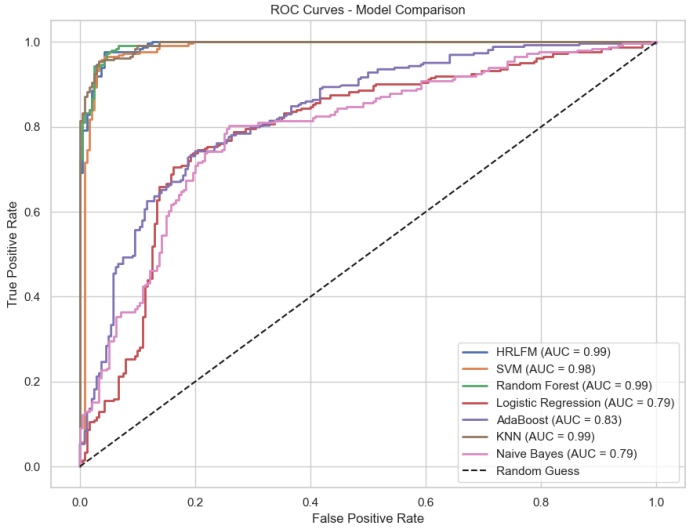

Figure 7. Roc (Receiver operating characteristic) curve for Ml algorithm

Figure 7. AUC-ROC curve is a graph used to check how well a binary classification model works. AUC (Area Under the Curve): measures the area under the ROC curve. A higher AUC value indicates better model performance

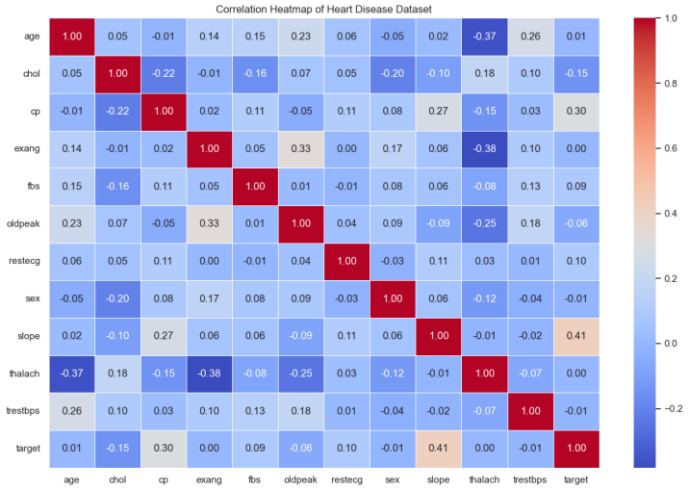

Figure 8 Heatmap of Dataset2

In Figure 8, Heatmap shows one to one correlation with column.

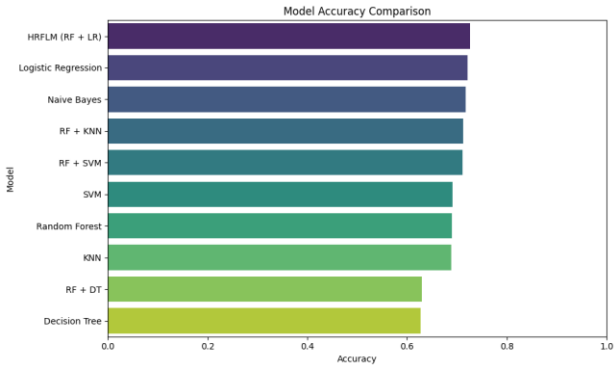

Figure 9. Model comparison with (70000,12) data shape

In Figure 9, it shows highest accuracy using (70000,12) data shape which is 72.68% using HRLFM model. In Table 6,it shows HRLFM has highest accuracy among all ML model.

Table 6 Statistics of Algorithms for the 3rd Data shape (Dataset1)70000 rows x 12 columns (Attribute)

Model	Accuracy	Precision	Recall	F1-Score
HRFLM (RF + LR)	0.7268	0.75	0.67	0.71
Logistic Regression	0.7215	0.76	0.65	0.70
Naive Bayes	0.7178	0.77	0.62	0.68
SVM	0.6908	0.77	0.53	0.63
Random Forest	0.6898	0.69	0.68	0.68
KNN	0.6887	0.69	0.68	0.68
Decision Tree	0.6271	0.63	0.60	0.61
RF + SVM	0.7265	0.76	0.66	0.70
RF + KNN	0.7221	0.74	0.67	0.71
RF + DT	0.6833	0.68	0.68	0.68

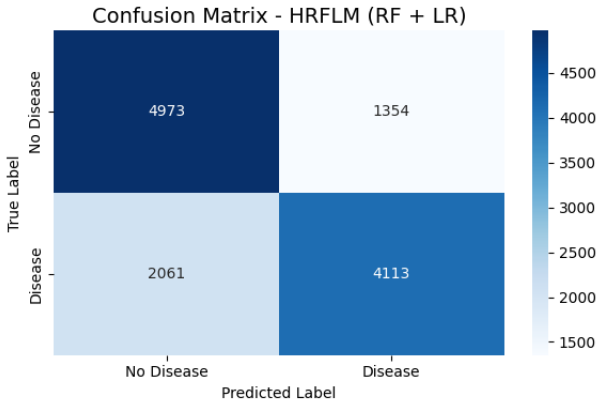

Figure 10. Confusion Matrix (70000,12) data shape

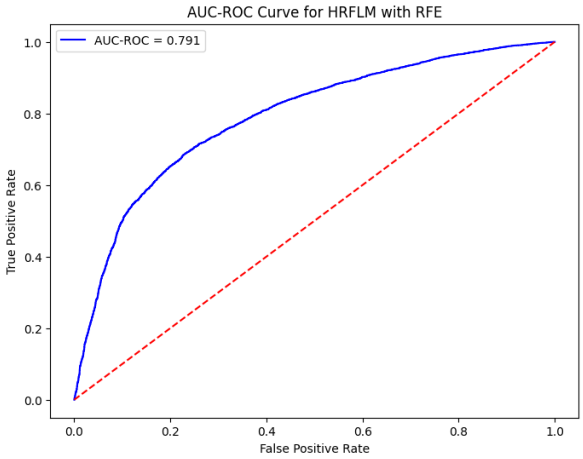

Figure 11. Roc (Receiver operating characteristic) curve for Ml algorithm

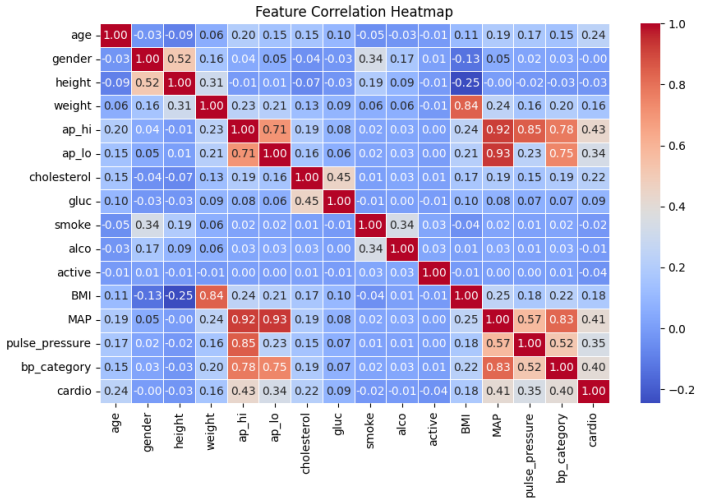

Figure 12. Figure 7 Heatmap of Dataset1

Figure 10, It shows Confusion Matrix (70000,12) data shape. Figure 11, It shows Roc (Receiver operating characteristic) curve for Ml algorithm. Figure 12, It shows Heatmap of Dataset1.

6. CONCLUSION

This study used evaluation criteria such confusion matrix, recall, precision, and F1-score to compare different machine learning algorithms for heart disease prediction. With HRLFM reaching excellent accuracy on the majority of datasets and the results demonstrated that machine learning approaches outperform conventional prediction methods. Compared to previous algorithms, HRLFM is more reliable and accurate handled both continuous and categorical data. In Future, We intend to enhance early heart disease prediction in the future by developing advanced Hybrid and deep learning models. These techniques are expected to improve the system's forecast accuracy, scalability, and robustness beyond what the current HRLFM model can do.

7.REFERENCES

[1] Jayasree, L., & Usha, D. (2022). A Comparison Analysis of Machine Learning Algorithms on Cardiovascular Disease Prediction. International Journal on Future Revolution in Computer Science & Communication Engineering, 8(3), 14-22.

[2] Journal Of Engineering Sciences Vol 14 Issue 04,2023. Heart Disease Prediction Using Machine Learning Mr.Valle Harsha Vardhan 1, Mr.Uppala Rajesh Kumar2 , Ms.Vanumu Vardhini 3 , Ms. Sabbi Leela Varalakshmi 4,Mr.A.Suraj Kumar 5

[3] Adhishayaa, P. V., Gomathi, V., & Mahendran, K. (2023, January). Review On Cardiovascular Disease Prediction Using Machine Learning Algorithm. In 2023 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE.

[4] Kondababu, A., Siddhartha, V., Kumar, B. B., & Penumutchi, B. (2021). WITHDRAWN: A comparative study on machine learning based heart disease prediction.

[5]Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), 381-386.

[6] Patidar, S., Jain, A., & Gupta, A. (2022, May). Comparative analysis of machine learning algorithms for heart disease predictions. In 2022 6th International Conference on

Intelligent Computing and Control Systems (ICICCS) (pp. 1340-1344). IEEE.

[7] Hasan, R. (2021). Comparative analysis of machine learning algorithms for heart disease prediction. In ITM Web of Conferences (Vol. 40, p. 03007). EDP Sciences.

[8] Kausar, N., & Ghous, H. (2020). A Comparative Analysis On Cleveland And Statlog Heart Disease Datasets Using Data Mining Techniques. LC International Journal of STEM (ISSN: 2708-7123), 1(4), 24-43.

[9] Shivadekar, S., Shahapure, K., Vibhute, S., & Dunn, A. (2024). Evaluation of Machine Learning Methods for Predicting Heart Failure Readmissions: A Comparative Analysis. International Journal of Intelligent Systems and Applications in Engineering, 12(6s), 694-699.

[10]Hammoud, A., Karaki, A., Tafreshi, R., Abdulla, S., & Wahid, M. (2024). Coronary Heart Disease Prediction: A Comparative Study of Machine Learning Algorithms. Journal of Advances in Information Technology, 15(1).

[11]Küçükmanisa, A., & Kilimci, Z. H. (2024). Heart Disease Prediction with Machine Learning-Based Approaches. Sakarya University Journal of Science, 28(1), 101-107.

[12] Arghandabi, H., & Shams, P. (2020). A comparative study of machine learning algorithms for the prediction of heart disease. International Journal for Research in Applied Science and Engineering Technology, 8(12), 677-683.

[13] Heart Disease Prediction: A Comparative Analysis of Machine Learning Algorithms.Adarsh Sharma, Himanshu Sharma, Sudeep Varshney, Nutan Gusain.Vol 3 , Issue 2 , July - December 2023 | Pages: 171-192 | Research Paper.https://doi.org/10.17492/computology.v3i2.2309Published Online: February 14, 2024

[14] A Comparative Study of Heart Disease Prediction using Machine Learning Sreekumari S 1 , Rajni Bhalla 1 and Geetha Ganesan 2 1 Lovely Professional University, Phagwara, Punjab, India 2 Jain (Deemed-to-be) University, Bengaluru, India

[15] Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics, 14(2), 144.

[16] Shrestha, D. (2024). Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction Using the Cleveland Heart Disease Dataset.

[17] Islam, R., Sultana, A., & Islam, M. R. (2024). A comprehensive review for chronic disease prediction using machine learning algorithms. Journal of Electrical Systems and Information Technology, 11(1), 27.

[18] Dass, A. K. (2023). Comparison of heart disease prediction using different machine learning algorithms.

[19] https://doi.org/10.1093/ejcts/ezad183

[20] Hammoud, A., Karaki, A., Tafreshi, R., Abdulla, S., & Wahid, M. (2024). Coronary heart disease prediction: a comparative study of machine learning algorithms. Journal of Advances in Information Technology, 15(1), 27-32.

[21] Osei-Nkwantabisa, A. S., & Ntumy, R. (2024). Classification and Prediction of Heart Diseases using Machine Learning Algorithms. arXiv preprint arXiv:2409.03697.

[22] Pushkala, V., Agalya, T., & Angayarkanni, S. A. (2019). Comparative study of heart disease prediction using machine learning algorithms. International Journal of Innovations in Engineering and Technology, 12(4), 64-8.

[23] International Journal of Scientific Research in Computer Science, Engineering and Information Technology ISSN : 2456-3307 (www.ijsrcseit.com) doi : https://doi.org/10.32628/CSEIT228686

.