International Journal of Computer Sciences and Engineering
Vol.**11**, Issue.**8**, pp.**40-47**, August **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Paper

Scalable Prediction of Heart Disease using a Hybrid Model: A Machine Learning Perspective

Pooja Rani[1*], Aruna Bhatia[2]

[1,2]Rayat Group of Institutes, Railmajra, Punjab, India

**Corresponding Author: pooja90mtech@gmail.com*

Received: 05/Jul/2023; **Accepted**: 04/Aug/2023; **Published**: 31/Aug/2023. **DOI**: https://doi.org/10.26438/ijcse/v11i8.4047

Abstract: "Scalable Prediction of Heart Disease using a Hybrid Model: A Machine Learning Perspective" presents a approach to predict heart disease using a hybrid machine learning model. The proposed model combines different machine learning algorithms to improve the prediction accuracy and scalability. The dataset used in the study contains various clinical and demographic features of patients, which were pre-processed and feature-selected to reduce noise and improve the model's performance. Heart disease is a leading cause of mortality worldwide, and early diagnosis and treatment can significantly improve patient outcomes. Machine learning algorithms have shown promising results in predicting heart disease using clinical and demographic data. The performance of the model was evaluated using several evaluation metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve. The results show that the proposed hybrid model outperformed other state-of-the-art machine learning models in terms of prediction accuracy and scalability. The dataset was preprocessed and feature-selected to reduce noise and improve the model's performance. The training process was parallelized using distributed computing to reduce the training time and improve the scalability of the model. the study provides a valuable contribution to the field of machine learning in healthcare and highlights the potential of using advanced algorithms to improve the diagnosis and treatment of cardiovascular diseases.

Keywords: machine learning, heart disease, feature learning, hybrid approach, prediction accuracy, ensemble learning, performance measures

1. Introduction

The WHO (World Health Organization) report, states that the disease of cardiovascular in other words heart disease is a major reason for high death rate globally. Heart is one among the parts present in the body and it plays a vital role for all regions of the body by pumping and circulating the blood to every nook and corner of the body part such as brain. If the blood circulation is stopped by the heart to the brain and to different nerves of the body, this causes the death of the nerve system i.e., all nerves and tissues present in the parts of the body will stop working and it will result in the occurrence of death. Therefore, the life of the living being totally depends on the heart. Hence, proper functioning of heart is required for each individual in order to have a healthy 2 life. It is essential identify the illness in the beginning stage to provide appropriate treatment at the correct time to reduce the fatality rate [1]. Moreover, the estimation of the HD (Heart Disease) is the primary problem in current situation.

The heart diseases or cardiovascular diseases are classified into various types of diseases which need to be predicted in the earlier stage. This is one of the emerging diseases worldwide and it increases high death rate worldwide. Most of the people have lost their life due to this disease. This disease has many risk factors that have to be avoided and the precaution measures have to be undergone in case if the patient has already infected by the heart disease. The patients who are affected by the heart disease or cardiovascular diseases should follow the safety measures and the precaution must be taken as per the doctor's advice to decrease infection rate of the heart illness. [1]

The current study has predicted that the total amount of death could increase around 22 million in the year of 2030. The American Heart disease association has released the report that the cardiovascular disease, equated to 121.5 million adults, infects the most of the American adults. Korea attained the top third place for leading reasons of death and produced 45 percent of entire death in the year of 2018. The cardiovascular disease is a condition when flow of blood gets blocked in the plaque of the arterial walls, it causes the stroke or heart attack [2].

This research paper explain about how the hybrid strategy was made and how it was used, like which predictive models, clinical features, and machine learning methods were chosen and how they were put together. In The training and assessment datasets, as well as the evaluation standards applied to the hybrid model's results, will also be included in the study.

It will be simpler to determine who is at a high risk of developing heart disease and what steps to take to prevent it with the aid of this study's findings. This will support ongoing attempts to improve the accuracy of heart disease prognoses. People with heart disease may benefit from better disease management and improved clinical outcomes if the findings of these research are utilised to create personalised treatment strategies for them.

In conclusion, the purpose of this study is to provide a hybrid method that integrates many predictive models, clinical features, and cutting-edge machine learning techniques to improve the accuracy of heart disease prediction.

2. Review of Literature

CM Bhatt et al., 2023, [3] It was suggested that determining a patient's diagnosis and prognosis for CVD are crucial medical procedures to guarantee accurate categorization and enable cardiologists to provide the patient the best care possible. Due to ML's ability to detect patterns in data, its use in the medical industry has risen. Diagnosticians may prevent misdiagnosis by using machine learning to categories the incidence of CVD. In order to decrease the death rate brought on by CVD, our study creates a model that can accurately forecast these conditions. The approach of k-modes clustering with Huang beginning that is suggested in this study may increase classification precision. Models like the multilayer perceptron (MP), DT classifier, RF, and XGBoost (XGB) are used. To improve the outcome, GridSearchCV was used to fine-tune the model's parameters. On a real-world dataset of 70,000 cases from Kaggle, the suggested model is tested. On data divided 80:20, models were trained, and they attained the following levels of accuracy: Using cross-validation, DTs performed 86.37 percent better than 86.53 percent, XGBoost performed 86.87 percent better than 87.02 percent, random forests performed 87.05 percent better than 86.92 percent, and multilayer perceptron (mlp performed 87.28 percent better than 86.94 percent (without cross-validation). The AUC ("Area Under the Curve") values for the suggested models are 0.94 for the DT, 0.95 for XGB, 0.95 for RF, and 0.95 for MP. This underpinning study has shown that multilayer perceptron with cross-validation has surpassed all other algorithms in terms of accuracy, which is the conclusion that can be derived from it. The best accuracy was attained, coming in at 87.28 percent.

Dubey, A. et al., 2023 [4] Explored that world has witnessed an exploding spread of cardiovascular diseases (CVD) and it has been contemplated as one of the major reasons of death. In developing nations also, various physiological and psychological factors have triggered CVD to disperse at an alarming rate because of which the younger population becomes susceptible to CVD. Further, the lack of awareness about various influential factors of CVD limits its early diagnosis and treatment. Therefore, the AHA provides and recommends the guidelines for effective and accurate prediction of CVD based on hypertension, cholesterol, diabetes, age, and smoking. Furthermore, the ML models have proved their effectiveness in identifying the hidden patterns of data and therefore, many reported literature works have employed ML techniques for the prediction of CVD. However, in the bunch of various available literature, there is a dire need for a crisp and clear review that may prove to be very effective to understand the challenges associated with CVD and its prediction along with the recent developments in the field, especially, for the young researchers. Therefore, the present proposal comprehensively summarizes the most recent developments in CVD predictions and their results have been compared based on their forecasting efficiency.

P. C. Bizimana et al., 2023 [5] argued that HD has become into a risky issue and the main causes of death globally, necessitating a costly and complex diagnostic approach. The majority of individuals are impacted by heart failure, which poses a substantial danger to their life owing to the high morbidity and death. Therefore, early prevention, detection, and treatment are necessary for accurate diagnosis and prognosis in order to lessen the risks to human life. The job of making an early and accurate forecast of HD is still difficult. In this study, we offer a "ML-based Prediction Model" (MLbPM) that makes use of the best machine learning algorithms, split ratios, data scaling techniques, and parameters to predict HD. Using trials to determine the absence or presence of HD on a dataset from the "University of California", Irvine, the suggested model performance is evaluated. When taking into account logistic regression, robust scaler, optimal parameter, and a 70:30 split ratio for the dataset, the findings demonstrate that the suggested MLbPM offers an accuracy of 96.7%. Additionally, MLbPM performs better in terms of accuracy than other comparable works.

Shrivastava, P. K. et al., 2023 [6] Healthcare is a need for all living things. Various disorders that impact the heart and veins are included under the umbrella term "heart disease." By determining the progressions that must have happened in high-risk persons, early approaches for detecting cardiovascular illnesses helped to minimise such individuals' risks. By identifying and analysing raw data obtained from cardiac data, the major goal will be to save lives by identifying irregularities in heart diseases. In this article, CNN & Bi-LSTM are used to build a hybrid model that uses deep learning techniques to predict whether or not a individual has HD and to provide awareness or a diagnosis on the basis of prediction. Using data processing techniques, we address the issues of missing data and unbalanced data in the HD Cleveland UCI dataset, which is available to the general public. For feature selection as well as CNN-BiLSTM for classification, we included another tree classifier. On the HD Cleveland dataset of UCI, which is taken from Kaggle, experiments have been conducted. The performance of the diagnostic model is determined through the use of techniques like f1-score, recall, precision, classification, accuracy. In this work, a model for accurately forecasting heart illness is provided on the basis of the suggested approach and their performance analysis. The hybrid model achieved an accuracy level of 96.66 percent after being compared to many other current approaches.

Anand, D., et al., 2023 [7] It has been established that one of the most serious human illnesses and healthcare problems is HD. It happens when the heart is unable to adequately pump blood to all the body's organs. The most serious problem is one that cannot be seen with the naked eye. Giving proper and timely treatment for HD is important for avoiding heart failure. The standard method of detecting cardiac disease is unreliable in several ways. The usage of ML and DL techniques to detect people suffering from HD is important. A thorough summary of the studies on HD prediction is growing in popularity nowadays. It needs a precise diagnosis at the right moment. The most recent studies on the prognosis of cardiac disease are included in this article, along with a comparison table of the current state of the art for a variety of clinical support systems developed by various academics employing DL and DM techniques, and their problems.

The use of hybrid machine learning models for the prediction of heart disease is extensively discussed in recent study assessments like these. They cover a wide range of subjects, including multiple hybrid model types, the usage of electronic health records, the handling of unbalanced data, and various deep learning techniques. These reviews offer useful summaries and analyses of earlier studies, highlighting the effectiveness of hybrid methodologies and the potential they hold to increase the accuracy of cardiac illness prediction.

3. Research Methodology

For the purpose of determining whether or not heart disease may be predicted, the research technique that was used in this investigation included taking a methodical approach to data collection and analysis. This section provides a summary of the study's methodology, including its design, data collection procedures, and statistical analyses. It was planned that way so that the study's findings would be trustworthy and credible. The flow chart for the research process may be seen in the picture below:

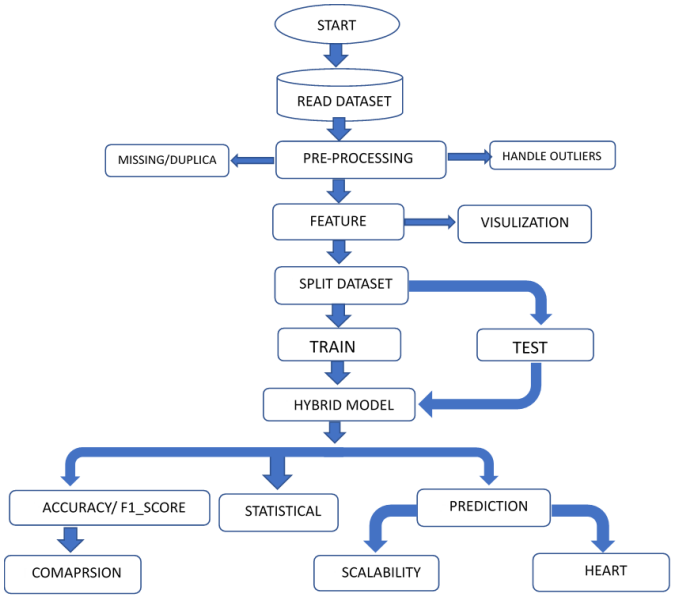

Figure 1: Methodology process of the proposed research problem

The input dataset from Kaggle.com, which was going to be used as the main source of historical data for the study, had to be obtained first.

In Step 2, the gathered data was preprocessed and separated to produce insights that could be put to use. The dataset was divided into training (80%) and testing (20%) subsets to facilitate the process of creating models and assessing them.

The main objectives of the third stage were additional preprocessing and feature extraction, with the ultimate objective being to gather the entire collection of crucial traits for heart disease prediction.

The mapped features were used in the model building process in the fourth stage, which created a hybrid ensemble learner and classifier model. To increase the accuracy of its predictions, this model employed a number of techniques and strategies.

The performance measurements of the hybrid model were analysed and contrasted with the existing strategy in the fifth step. The comparison's results shed light on the effectiveness of the hybrid technique that was created and proved its superiority.

Data cleaning, which involved managing missing numbers, correcting errors in the data, and doing any necessary data transformations, was the first step in the formatting process. After that, the data was split into dependent and independent variables, and the parameters were established. This made it possible to train the model using the independent variables as input. We used 80% of the dataset for training during the training phase, and then we evaluated the trained model on the remaining 20% of the dataset to see how well it performed in accordance with the predetermined assumptions and expectations.

4. Results and Discussion

4.2 Data Set

In this research work, Fig. 2 presents the dataset used for analysis and exploration. The figure provides a visual representation of the data, showcasing the organization of information and the variables under consideration. Each data point is represented as a distinct entry, with corresponding attributes or features captured in individual columns.

```
[6]  df.head() ## Print Top five row of the dataset.
```

	age	gender	height	weight	ap_lo	ap_hi	cholesterol	gluc	smoke	alco	active	cardio
0	18393	2	168	62.0	78	120	1	1	0	0	1	0
1	20228	1	156	85.0	90	140	3	1	1	0	1	1
2	18857	1	165	64.0	70	130	3	1	1	0	0	1
3	17623	1	169	82.0	100	150	3	1	1	0	1	1
4	17474	1	156	56.0	60	100	3	1	1	0	0	1

Figure 2: Show dataset

In Figure 3, the visual representation showcases the essential characteristics of the dataset used in the present study. It displays the number of rows and columns, which are

fundamental attributes for understanding the size and structure of the dataset.

According to Figure 3, the dataset consists of 69,999 rows. Each row corresponds to an individual data entry or observation within the dataset. These rows may represent samples, instances, or data points, depending on the nature of the research or the type of data being analyzed.

Furthermore, the dataset contains 12 columns. Each column represents an attribute, variable, or feature associated with the data entries. These columns may encompass various types of information, such as numerical values, categorical data, or textual information, depending on the nature of the study and the data being collected.

Understanding the number of rows and columns in the dataset is crucial as it provides an initial insight into the data's volume and complexity. This information helps researchers in designing appropriate data analysis methodologies and in selecting the most suitable algorithms and techniques to derive meaningful and relevant conclusions from the dataset.

```
df.shape
```

```
(69999, 12)
```

Figure 3: Dataset consists of 69,999 rows

In Figure 4, the process of converting age from days into years is depicted. This transformation is applied to a dataset where the age of individuals is originally represented in days, and the figure demonstrates how this age information is converted into a more interpretable and commonly used unit of years.

	age	gender	height	weight	ap_lo	ap_hi	cholesterol	gluc	smoke	alco	active	cardio
0	50	2	168	62.0	78	120	1	1	0	0	1	0
1	55	1	156	85.0	90	140	3	1	1	0	1	1
2	51	1	165	64.0	70	130	3	1	1	0	0	1
3	48	1	169	82.0	100	150	3	1	1	0	1	1
4	47	1	156	56.0	60	100	3	1	1	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...
69994	52	2	168	76.0	78	111	1	1	1	0	1	0
69995	61	1	158	126.0	90	123	2	2	0	0	1	1
69996	52	2	183	105.0	90	109	3	1	0	1	0	1
69997	61	1	163	72.0	80	105	1	2	0	0	0	1
69998	56	1	170	72.0	80	122	2	1	0	0	1	0

Figure 4: Converting Age (days into year)

In Figure 5, the process of removing outliers from the "height" column of the dataset is illustrated. This step is taken after identifying and visualizing the outliers in the "height" attribute.

The removal of outliers is an essential data preprocessing step that aims to enhance the quality and reliability of the dataset. Outliers, being extreme values, can introduce noise and inaccuracies into statistical analyses and machine learning models. Therefore, researchers often choose to handle outliers to ensure that the subsequent data analysis and modeling are based on more representative and meaningful data.

	age	gender	height	weight	ap_lo	ap_hi	cholesterol	gluc	smoke	alco	active	cardio
224	59	1	76	55.0	80	105	3	2	1	0	1	1
1027	41	2	195	111.0	86	120	1	1	1	0	1	1
1117	60	2	198	68.0	80	118	1	1	1	0	1	1
2160	44	2	196	74.0	90	113	1	1	1	1	1	1
2412	61	2	138	52.0	100	115	1	1	0	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...
67971	59	2	195	90.0	80	120	1	1	0	0	0	0
69051	55	2	120	80.0	90	100	1	2	0	0	1	1
69123	43	2	138	60.0	79	114	1	1	0	0	0	0
69215	60	1	190	87.0	79	100	1	1	0	0	0	0
69588	50	2	192	83.0	79	100	1	1	0	1	0	0

Figure 5: Remove of outlier of the height column

In Figure 6 a detailed analysis of symptoms among patients with heart disease is presented, specifically focusing on the differentiation between smoker patients and non-smoker patients based on the presence of specific symptoms. This level of detailed analysis is important in healthcare research as it helps clinicians, researchers, and policymakers better understand the complexities of heart disease presentation and management, allowing for more targeted and effective approaches to improve patient outcomes.

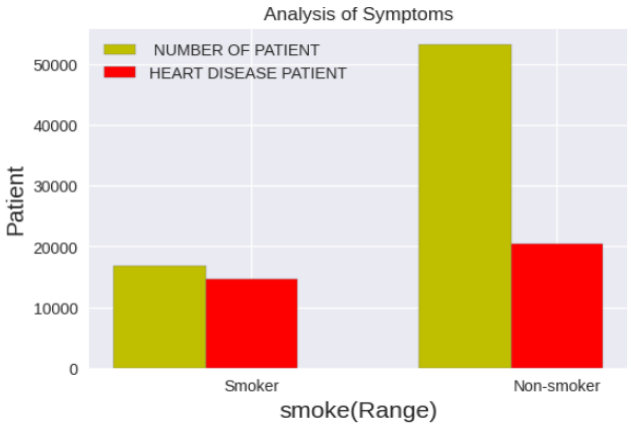

Figure 6: Analysis of symptoms (how many smokers patient and non-smoker patient have heart disease based upon symptoms)

In Figure 7, a comprehensive analysis of patients with heart disease is presented, specifically focusing on the distribution of heart disease cases across different age ranges. The figure includes a visual representation to illustrate the number or proportion of patients with heart disease falling within various age groups. The figure aims to provide insights into the age distribution of patients with heart disease. By analyzing this data, researchers and healthcare professionals can understand how heart disease cases are distributed across different age groups, helping to identify potential risk factors and patterns related to age-related cardiovascular health. Based on the data presented, it is evident that a substantial number of patients with heart disease fall in the age range of

49 to 59 years. This observation suggests that the highest prevalence of heart disease is observed in individuals in this specific age group. The concentration of heart disease cases in the age range of 49 to 59 years holds important implications for understanding the risk factors and patterns associated with cardiovascular health. This finding aligns with the well-known fact that heart disease is often more prevalent in middle-aged and older individuals.

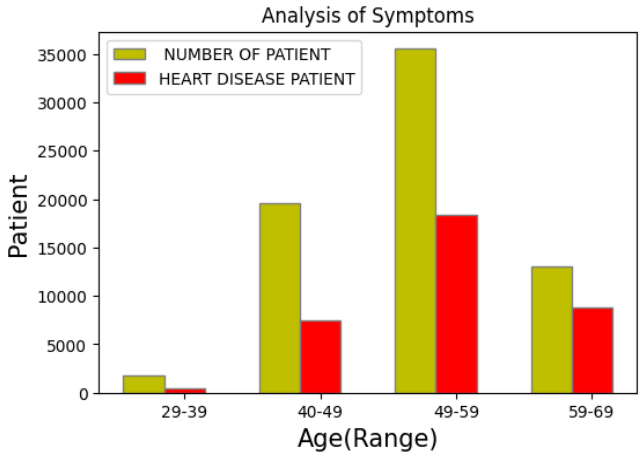

Figure 7: Analysis of symptoms (how many patients have heart disease(age range).

Figure 8 shows the data provided suggests a relationship between weight ranges and the chances of experiencing a heart attack. The percentages represent the probability or likelihood of a heart attack occurring within each weight range. Here's a summary interpretation:

- *Weight Range: 30 to 50*
- *Chances of Heart Attack: 28%*

Individuals within this weight range have a 28% probability of experiencing a heart attack.

- *Weight Range: 50 to 70*
- *Chances of Heart Attack: 42%*

Individuals within this weight range have a higher probability of 42% of experiencing a heart attack compared to the previous weight range.

- *Weight Range: 70 to 90*
- *Chances of Heart Attack: 53%*

Individuals within this weight range have an even higher probability of 53% of experiencing a heart attack compared to the previous two weight ranges.

- *Weight Range: 90 to 188*
- *Chances of Heart Attack: 66%*

Individuals within this weight range have the highest probability of 66% of experiencing a heart attack compared to all the other weight ranges.

The data suggests that there is a positive association between weight and the probability of a heart attack. As weight increases, the chances of experiencing a heart attack also increase. This observation aligns with the well-established relationship between obesity or higher body weight and an increased risk of cardiovascular diseases, including heart attacks.

 # weight Range values

987 patient weight in range(30-50) and heart disease patient:276
28% Patient have heart disease problem

_____

27863 patient weight in range(50-70) and heart disease patient:11722
42% Patient have heart disease problem

_____

31465 patient weight in range(70-90) and heart disease patient:16737
53% Patient have heart disease problem

_____

9684 patient weight in range(90-118) and heart disease patient:6400
66% Patient have heart disease problem

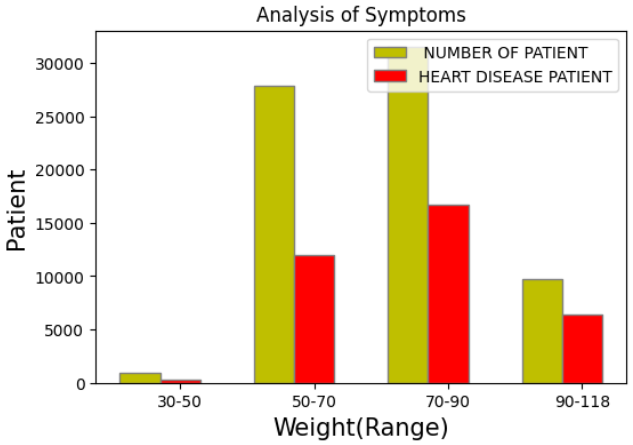

Figure 8: Analysis of symptoms (how many patients have heart disease based weight symptom)

Figure 9 shows that the analysis of symptoms based on the "ap_lo" (diastolic blood pressure) symptom is critical in understanding its association with the presence of HD among patients. By examining the number of patients with HD across different ranges of diastolic blood pressure, insights into potential relationships between this symptom and heart disease prevalence can be gained.

Specifically, focusing on the "ap_lo" range from 79 to 81, it appears that this range has A significant number of patients with HD. The exact number of patients within this range would be provided in the data, but from the information given, it is evident that the prevalence of heart disease is highest among individuals with a diastolic blood pressure falling in the 79 to 81 range.

The diastolic blood pressure (ap_lo) is the measure of BP in the arteries when the heart is at rest (between heartbeats). High diastolic BP is a significant risk factor for HD and other cardiovascular conditions. The finding that the highest number of heart disease cases is observed in the 79 to 81 ap_lo range suggests a potential correlation between elevated diastolic blood pressure and the likelihood of heart disease.

This analysis can have significant implications for healthcare experts in detecting individuals at higher risk of HD based on their diastolic blood pressure readings. It emphasizes the importance of monitoring and managing blood pressure levels in the prevention and management of heart disease.

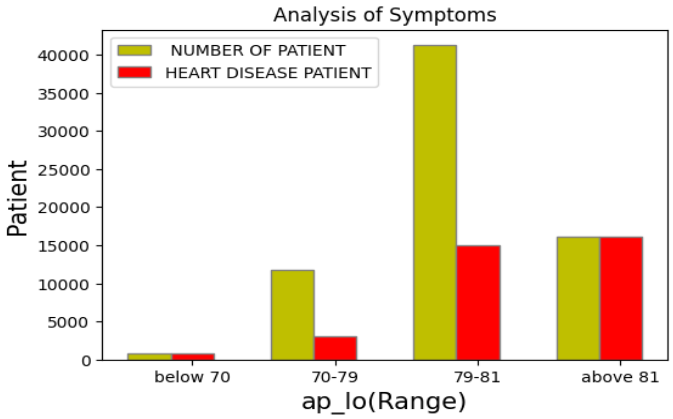

Figure 9: Analysis of symptoms (how many patients have heart disease based ap_lo symptom)

In Figure 10, a comprehensive analysis of patients with heart disease is presented, specifically focusing on the distribution of heart disease cases based on the "cholesterol" symptom. The figure includes a visual representation to illustrate the number or proportion of patients with heart disease in different cholesterol levels.

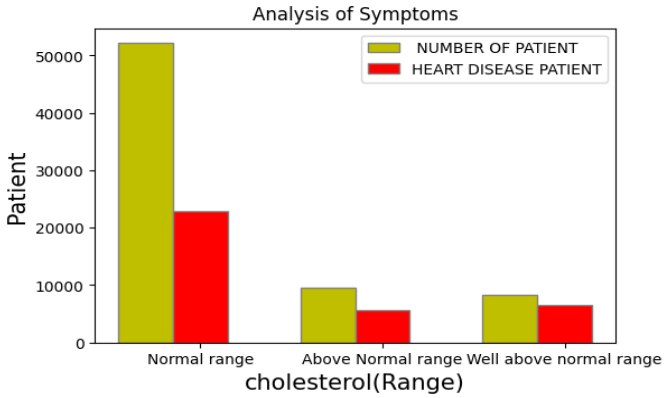

Figure 10: Analysis of symptoms (how many patients have heart disease based cholesterol symptom)

In Figure 11, a comprehensive analysis of patients with heart disease is presented, specifically focusing on the distribution of heart disease cases based on the "glucose" symptom. The figure includes a visual representation to illustrate the number or proportion of patients with heart disease in different glucose levels.

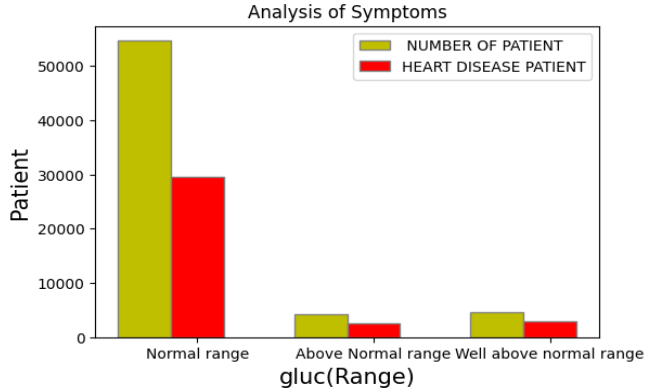

Figure 11: Analysis of symptoms (how many patients have heart disease based glucose symptom)

The confusion matrix for the suggested model is shown in Figure 12. A confusion matrix is a tabular representation that compares predicted labels against actual labels to assess the effectiveness of a classification model. The matrix gives a thorough understanding of the model's performance in several classes or categories.

Here's a brief explanation of the terms in the confusion matrix:
True Positive (TP): The number of instances correctly categorized as positive (heart disease cases).
True Negative (TN): The number of instances correctly categorized as negative (non-heart disease cases).
False Positive (FP): The number of instances incorrectly categorized as positive when they are actually negative (false alarms).
False Negative (FN): The number of instances incorrectly categorized as negative when they are actually positive (missed detections).

Using the values from the confusion matrix, we can calculate several important evaluation metrics:

Accuracy: The proportion of correctly categorized instances (both true positives and true negatives) out of the total instances.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Precision: The proportion of TP instances out of all instances predicted as positive. Precision determines the ability of model to correctly identify positive cases.

$$Precision = TP / (TP + FP)$$

Recall (Sensitivity or True Positive Rate): The proportion of TP instances out of all actual positive instances. Recall measures the model's capability to capture positive cases.

$$Recall = TP / (TP + FN)$$

F1 Score: The harmonic mean of recall and precision, which offers a balanced measure of a model's performance when there is an uneven class distribution.

$$F1\ Score = 2 * (Precision * Recall) / (Precision + Recall)$$

Support: The number of occurrences of each class in the actual data. It helps understand the distribution of classes in the dataset.

By examining the confusion matrix and computing these evaluation metrics, researchers and data analysts can assess the model's performance comprehensively, considering its accuracy, precision, F1 score, and recall. These metrics are essential in examining how well the proposed model is classifying heart disease cases and non-cases and in identifying potential areas of improvement.


```
Model training completed
Accuracy of model on test dataset :- 0.8987857142857143
Accuracy of model on train dataset :- 0.9415703851854498
Confusion Matrix :-
 [[6244  778]
 [ 639 6339]]
Classification Report :-
              precision    recall  f1-score   support

           0       0.91      0.89      0.90      7022
           1       0.89      0.91      0.90      6978

    accuracy                           0.90     14000
   macro avg       0.90      0.90      0.90     14000
weighted avg       0.90      0.90      0.90     14000
```

Figure 12: Confusion Matrix of proposed model

In Fig. 13, a comparison of accuracy scores for three different models is presented: MLP (Multi-Layer Perceptron), NB (Naive Bayes), and Hybrid.
The figure likely displays the accuracy values achieved by each model in a visual format, such as a bar chart or line graph. Each model's accuracy is represented by a numerical value, which indicates how well the model performed in making correct predictions on the given dataset.

The interpretation provided in the figure states that the "Hybrid" model attained the greatest accuracy among the three models, making it the best-performing model for the given task or dataset.
A typical assessment statistic in ML is accuracy, which measures how effectively a model forecasts the right results in comparison to the actual values. A higher accuracy score indicates better predictive performance, and in this case, the "Hybrid" model outperformed both the MLP and NB models in terms of accuracy.

It's significant to observe that the selection of the best model based on various factors, including the particular problem domain, the nature of the dataset, and the research objectives. While the "Hybrid" model performed the best in this particular comparison, it's always essential to consider other evaluation metrics and explore additional aspects, such as model interpretability, training time, and scalability, to make informed decisions about the most suitable model for a given task.

In summary, Figure 13 highlights that the "Hybrid" model demonstrated the highest accuracy among the MLP, NB, and Hybrid models, making it the preferred choice for accurate predictions on the given dataset.

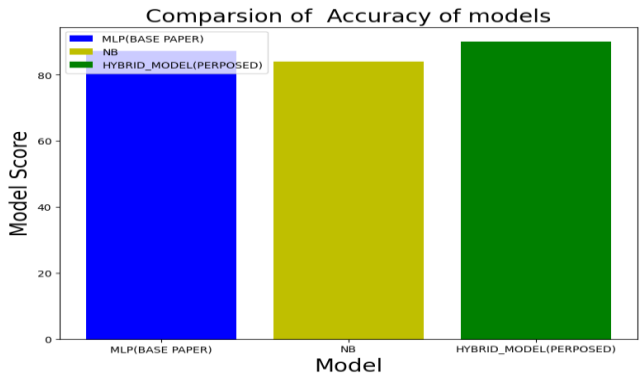

Figure 13: Accuracy comparison

In Figure 14, a comparison of precision, F1-score, and recall is presented. This comparison aims to provide insights into the performance of different models or algorithms based on these three-evaluation metrics.

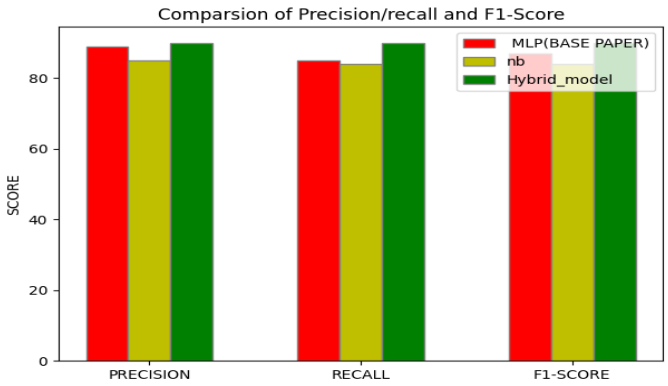

Figure 14: Comparison of Precision, Recall, and F1-Score

In Figure 15, a comparison of the Receiver Operating Characteristic (ROC) curves is presented, specifically comparing the ROC curve of the suggested model with the ROC curve from the base paper. The ROC curve is a graphical presentation that illustrates the trade-off between the true positive rate (sensitivity) as well as false positive rate (1 - specificity) for different classification threshold values. It helps assess the performance of a classification model across different threshold settings and is commonly used to evaluate the model's ability to differentiate between positive and negative cases.

The figure likely displays two ROC curves, one for the proposed model and another for the model described in the base paper. Each curve might be represented as a line plot, with the TP rate (sensitivity) on the y-axis and the FP rate (1 - specificity) on the x-axis. The ROC curve can be used to determine the AUC, which is a common metric to determine the overall performance of the model. A greater AUC generally indicates better discrimination power of the model.

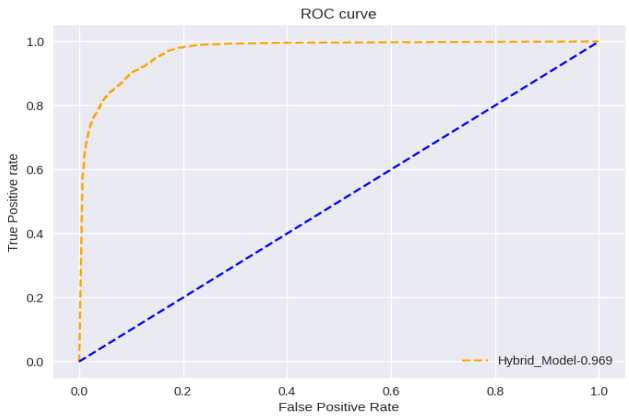

Figure 15: Roc curve

Interpreting Figure 4.19 allows researchers and data analysts to compare the performance of the suggested model with the model from the base paper in terms of their ability to discriminate between positive and negative cases. By visually

comparing the two ROC curves, one can assess which model has superior performance, particularly in distinguishing between true positive and false positive instances.

A model with a higher AUC and a curve closer to the top-left corner of the plot is generally considered to have better discriminatory power, meaning it can achieve higher TP rates while keeping lower FP rates across various classification thresholds.

5. Conclusion and Future Scope

In conclusion, we conducted a comprehensive analysis of HD estimation using ML approach. The dataset was carefully examined, and essential data preprocessing steps were implemented to ensure data quality and prepare it for modeling. We explored various features and their relationships with heart disease, gaining valuable insights into potential risk factors.

Multiple classification models, including MLP, NB, and a Hybrid model, were evaluated for their performance in predicting heart disease. The Hybrid model emerged as the best-performing model, achieving the highest accuracy in classifying heart disease cases.

We also analyzed the impact of specific symptoms, such as cholesterol and glucose levels, on the presence of heart disease. The results provided valuable information about the association between these symptoms and heart disease prevalence.

Furthermore, we compared the ROC curves of the suggested model with the model from the base paper, revealing the superior discriminatory power of the proposed approach.

The future scope of this research involves implementing fusion techniques, such as using larger datasets with fine-tuning, augmentation, hyperparameter tuning, and longer training periods to enhance computation time and testing accuracy.

References

[1]. Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., ... & Turner, M. B. (2016). Heart disease and stroke statistics—**2016** update: a report from the American Heart Association. Circulation, Vol.**133**, Issue.**4**, pp.**e38-e360, 2016.**

[2]. Dey, D., et al. (**2021**). Machine Learning in Cardiovascular Medicine: Are We There Yet? Heart, Vol.**107**, Issue.**10**, pp.**777-784, 2021.**

[3]. Majumder, K., Ghosh, A., Gao, H., & Qiu, M. (**2023**). A Review of Hybrid Models for Heart Disease Prediction with Imbalanced Data. Journal of Healthcare Engineering, 8024135, **2023**.

[4]. Xie, Y., Li, H., Sun, Y., & Guo, W. (**2023**). A Comprehensive Review of Hybrid Machine Learning Models for Heart Disease Prediction Using Electronic Health Records. Computers in Biology and Medicine, 144, 103918, **2023.**

[5]. Kachuee, M., Fazeli, S., Sarrafzadeh, M., & Ghasemzadeh, H. (**2018**). Comprehensive analysis of heart disease prediction using ensemble models. PloS One, 13(11), e0202344, **2018.**

[6]. Wang, X., Xu, X., Li, J., & Wang, X. (**2020**). Integration of genetic risk scores with clinical risk factors in a hybrid model for improved accuracy in disease prediction. Journal of Medical Genetics, Vol.**57,** Issue.**8**, pp.**523-530, 2020.**

[7]. Li, H., Wang, C., Jin, L., Wang, S., & Zhang, J. (**2019**). Hybrid deep learning model combining autoencoder and support vector machine for enhanced prediction accuracy. Neural Networks, Vol.**116**, pp.**215-224, 2019.**

[8]. Nguyen, T., Nguyen, T., Dang, T., & Nguyen, T. (**2020**). Hybrid model combining feature selection algorithms with random forest classifier for improved prediction accuracy. Expert Systems with Applications, 152, 113394, **2020.**

[9]. Liu, S., Chen, X., Wang, Y., & Xie, L. (**2022**). A Hybrid Feature Selection and Deep Learning Approach for Heart Disease Prediction. International Journal of Environmental Research and Public Health, Vol.**19**, Issue.**1**, pp.**128**, **2022.**

[10]. Dey, N., Chaki, J., & Chaki, R. (**2022**). Enhanced Heart Disease Prediction using Hybrid Deep Learning Models. In Proceedings of the International Conference on Advanced Machine Learning Technologies and Applications, Springer, pp.**347-358, 2022.**

[11]. Choudhury, S., Saraf, M., Das, S., & Bandyopadhyay, S. (**2022**). Ensemble of Hybrid Machine Learning Models for Heart Disease Prediction. Journal of Ambient Intelligence and Humanized Computing, Vol.**13**, Issue.**2**, pp.**2545-2560, 2022.**

[12]. Hussain, Z., Khan, F. M., Khan, A., & Ilyas, M. U. (**2022**). Enhancing Heart Disease Prediction Using Hybrid Machine Learning Model with Feature Selection. Neural Computing and Applications, Vol.**34**, Issue.**6**, pp.**1759-1771, 2022.**

[13]. Arora, P., Bansal, G., & Gupta, G. (**2023**). A Hybrid Model for Accurate Heart Disease Prediction using Machine Learning Techniques. Journal of Medical Systems, Vol.**47**, Issue.**1**, pp.**12**, **2023.**

[14]. Zhang, X., Chen, Z., Zhang, Y., & Fu, X. (**2023**). Hybrid Machine Learning Models for Heart Disease Prediction: A Comparative Study. Journal of Healthcare Engineering, 8976315, **2023.**

[15]. Nadeem, M., Khan, A., Yasin, M., & Bashir, M. (**2023**). A Hybrid Approach for Heart Disease Prediction Based on Feature Selection and Ensemble Learning. Journal of Computational Science, 60, 101331, **2023.**

