Analysis of Hidden Pattern of Heart Disease Dataset Using Multiple Machine Learning Ensemble Methods

Gyanendra Kumar Pal* and Sanjeev Gangwar

***Abstract*—Heart failure disease, a wide-ranging clinical disorder, is affecting more and more individuals worldwide. The Healthcare Sector (HCS) places a high focus on the early detection of cardiac disease. The creation of a machine learning-based cardiovascular disease prediction system is the main objective of this project. This study's presentation of several machine learning techniques is based on a brief examination of heart disease diagnosis. First, a lasso features selection approach is used to forecast heart disease. The second ensemble strategy is utilized to look into several areas of cardiac disease. We provided two ensemble stages of classifiers such Max Voting and Stacking for XG Boost, Random Forest, and Multilayer Perception in order to enhance results. A variety of factors were used to evaluate the performance of the recommended cardiovascular disease in order to select the best machine learning model. The major objective of this study is to give physicians a tool to help in the early diagnosis of heart problems. As a result, treating patients effectively and preventing negative effects will be easy. This study lasso uses selection-based techniques with machine learning classifiers to investigate several classification algorithms in terms of Mean Average Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Scattered Index in an effort to increase the accuracy of heart disease identification.**

***Index Terms*—XG boost, random forest, and multilayer perception, Mean Average Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), scattered index and heart disease dataset**

I. Introduction

The Cardiovascular (CVD) sickness is a sort of heart disease that is nevertheless a main purpose of mortality across the world. If nothing is done, the world's typical wide variety of fatalities is expected to climb. Plaques on artery partitions can block blood waft and motive a coronary heart assault or stroke. Physical inactivity, a bad diet, and the wonderful use of alcohol and cigarettes are all danger elements for coronary heart ailment [1, 2]. The above-mentioned variables may additionally be diminished by using leading a healthy everyday lifestyle, which consists of limiting salt in the diet, ingesting more veggies and fruits, enticing in ordinary bodily exercise, and abstaining from alcohol and cigarette use, all of which assist to decrease the risk of heart disease [3]. The collecting of patient records from a number health care institutions and hospitals is the solution to these challenges. The choice aid machine is used to reap the findings and to seek a 2nd opinion from an experienced doctor. This approach for analysis eliminates useless test conductions, saving cash and time [4, 5]. Recently, a health center administration system was once used to deal with health care or affected person data, implying that these structures create greater data. In the early degrees of heart failure, a larger variety of neuron hormonal regulating structures are activated (HFD) [6, 7]. In a quick duration, these compensatory mechanisms can reason the HFD consequences, main to accentuated ventricular dysfunction, dyspnea on exertion, peripheral edema, pulmonary, and heart These compensatory methods can lead to improved ventricular dysfunction, dyspnea on exercise, peripheral edoema, pulmonary, and cardiac remodeling, which can produce after load and preload irreversible alterations, in a quick duration of time. The affected person is presented more treatment options with HFD, which include life-style adjustments and implanted or pharmaceutical units such a defibrillator or pacemaker. Given that hospitalization due to acute HFD decomposition is the largest supply of healthcare cost, the key subject is ensuring follow-up in this group. According to data and research, heart sickness is the most serious problem that people confront, mainly these who follow an HFD [8, 9]. Early detection and diagnosis of heart sickness is the first step in care and therapy for a variety of conditions.

Early detection of heart sickness with more suitable prognosis and high-risk sufferers the usage of a prediction model is broadly advised for lowering fatality rates, and decision-making for subsequent remedy and prevention is improved. A prediction model is built and used in Clinical Decision Support Systems (CDSS) to assist doctors in estimating the hazard of coronary heart ailment and imparting appropriate treatments to manage the risk. Furthermore, a range of researchers have found that CDSS deployment can enhance selection quality, scientific selection making, and preventative care [9–11].

In the scientific field, laptop getting to know might also be used to diagnose, detect, and forecast a variety of ailments. The predominant goal of this research is to grant physicians with a device to discover coronary heart problems early on. As a consequence, it will be easier to supply patients with ideal medicine while minimizing predominant side effects. The intention of this study is to study more than one selection timber using desktop getting to know classifiers in the hopes of bettering overall performance in coronary heart sickness detection.

The rest of the paper is organized as follows: Section II presents the related works. Section III describes the proposed method. Next, result and discussion are presented in Section IV. Finally, conclusions are drawn in Section V on the basis of results and evaluation.

Manuscript received December 14, 2022; revised March 3, 2023; accepted July 3, 2023.

The authors are with Department Computer Science & Engineering, Unsietvbs Purvanchal University, Jaunpur, India.

*Correspondence: gyanpal@gmail.com (G.K.P.)

II. Related Work

D'Souza and Wang *et al*. [12], Alharbi and Alosaimi *et al*. [13] used desktop mastering methods to talk about cardiac disease. The heart rate, on the different hand, is altered by the things to do that a man or woman engages in, and as a result, coronary heart charge records are no stationary, unpredictable, and cannot be expected or modeled.

Chen and Lucock [14] developed a method for predicting cardiac disease the use of statistical analysis. Unpredictability variables and different behavioral hazard factors, such as cigarette use, unhealthy meals and obesity, bodily inactivity, and hazardous alcohol use, can all make a contribution to bad health and even double the danger of mortality in CVD sufferers.

Chen and Keravnou-Papailiou *et al*. [15], Su and Chen *et al*. [16], Chen and Shang [17], Knox and Chen [18], Chen and Antoniou [19] used artificial talent to check out cardiovascular illness and found a paradigm. The significance of detecting cardiovascular sickness as soon as possible follows. Artificial intelligence advancements are inflicting a paradigm change in healthcare, from early sickness detection and analysis to individualized therapy and prognosis contrast.

When compared to the ACC/AHA Pooled Cohort Equations (PCE) calculator alone, Machine Learning (ML) approach offers the chance to identify patients at higher risk of Type 2 Diabetes (T2DM) complications, and prediction models built using ML techniques improve cardiovascular disease prediction and reduce the number of screenings required [20].

Kakria and Tripathi *et al*. [21] investigated the use of a Real-Time Health Monitoring System for Remote Cardiac Patients. For example, a patient's cardiovascular problems are now being monitored at home by means of the fitness monitoring gadget in order to make appropriate recommendation to each patients and scientific consultants.

Also, an approach using unsupervised Machine Learning (ML) clustering might identify groups of T2DM patients with various forms of coronary plaque and degrees of coronary stenosis and treat T2DM patients with diverse clinical signs, enabling patient stratification [22].

Ephzibah [23] created a laptop mastering method to become aware of heart sickness characteristics. Heart ailment has been identified as a main reason of mortality in India. The necessity and usefulness of early illness prognosis cannot be overstated. The cautioned mannequin is built on a genetic algorithm, a Neural Networks (NN) classifier, and fuzzy rules. The developed model has an excessive diploma of accuracy in detecting heart sickness.

Haq and Li *et al*. [24] developed an approach for predicting heart ailment the use of a guide vector desktop with unique characteristics that had an accuracy of 84%. C and g were set to 10 and 0.0001 respectively, and the NB classifier had an accuracy of 83%.

Asfaw [25] developed a K-Nearest Neighbor classifier for heart disease prediction the usage of 10 prioritised criteria that good identified the disorder 71.05% of the time.

Khan and Abbas *et al*. [26] investigated cardiac disease diagnosing characteristics the usage of a couple of applied sciences in order to improve sickness prediction. Cloud computing has been built-in with the computer mastering classifier Support Vector Machine (SVM). The accuracy of a cloud-based intelligent gadget driven via the SVM model used to be 93.33%.

Alaa and Bolton *et al*. [27] thought of the use of a desktop studying classifier to estimate cardiovascular risk. In contrast to well-performing systems, Auto-Prognosis significantly extended cardiovascular danger prediction performance. This method used to be created utilising information from over 4 million UK Biobank individuals and 450 special elements for every of them. This method allowed agnostic exploration of novel cardiovascular threat variables. To check therapeutic validity, the Auto-Prognosis model used to be compared to the widespread Framingham model. The Auto-Prognosis algorithm effectively predicted 3357 out of 4801 cardiovascular patients.

A ML analysis of the Action to Control Cardiovascular Risk in Diabetes Study (ACCORD) and Veterans Affairs Diabetes Trial (VADT) trials provided evidence in support of the recommendation in the diabetes treatment guidelines for intensive glucose lowering in diabetics with low cardiovascular risk, and it also suggested benefits of intensive glycaemic control in some people with higher cardiovascular risk [28].

Kasbe and Pippal [29] both respected medical specialists, agree with that a complete tool for diagnosing heart failure based on the information provided is required. Fuzzification, Rule Base, and Defuzzification have been the three important phases of the provided fuzzy specialist system. The machine used to be created with the assist of the MATLAB Fuzzy Logic Toolbox and the Mamdani Fuzzy Interface System (MFIS). The accuracy and sensitivity of the test have been additionally high, at 94.50% and 90.19%, respectively.

The deep neural network learning model, which used to be created by using integrating two subsystems acknowledged as the deep neural community prediction (diagnosis) model and the deep neural community education classification model, was once investigated by Miao [30]. A deep neural community coaching category was used in the first phase, and then last weights had been assigned to deep neural network diagnostic. In contrast to a preferred multilayer perception neural community classification, a deep neural network mannequin contains greater hidden layers. The accuracy, sensitivity, and specificity have been every 83.67%, 93.51%, and 72.86%.

Hashi and Shahid [31] put together a system primarily based on desktop gaining knowledge of strategies that have been extensively studied for heart sickness prediction. On the dataset accumulated from the Union Cycliste Internationale (UCI) machine learning repository, statistics mining methods such as Naive Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Random-Forest have been utilised. The Random Forest Algorithm supplied the excellent accuracy for coronary heart disorder prediction, 90.16%, in accordance to these overall performance approaches.

Cardiovascular sickness analysis has been regarded fundamental for life-saving by way of Siddiqui and Athar *et al.* [32]. The accuracy of cardiovascular disease detection utilising deep extreme computing device gaining knowledge

of Diagnosis Cardiovascular Disease and Deep Extreme Machine Learning (DCD-DEML) with back-propagation was 92.45%, which was higher than that of the Developmental Co-ordination Disorder (DCD) Mamdani Fuzzy Inference System and DCD Artificial Neural Network.

Almustafa [33] investigated the Gradient Descent Algorithm for Cardiovascular Disease Diagnosis. The gradient descent algorithm is an approach for optimising many loss features and is utilised to do so (linear). Based on the root-finding function for cardiovascular illnesses, stochastic gradient descent used to be used. For each generation of the Stochastic Gradient, Descent samples are chosen at random using batch, which is a pattern dimension instead than the whole dataset size. Selected batches resource in the gradient computation for each cycle. Global Distribution System (GDS) has a rather incredible accuracy of 84.39% in diagnosing cardiovascular disease.

III. Proposed Method

A. Methodology

In the area of cardiovascular medicine, computing device studying is becoming increasingly more popular. Despite the truth that there are an extra of machine getting to know algorithms available, selecting the highest quality approach for cardiovascular ailment datasets remains a difficulty. The essential motive of the proposed lookup challenge is to propose a machine learning-based cardiovascular disease prediction device that is extraordinarily accurate. For expanded outcomes, we introduced two ensemble phases of classifiers such as most vote casting and Stacking for XG Boost, Random Forest, and Multilayer Perception. Using a variety of modelling techniques or training data sets, ensemble modelling is the process of building numerous varied models to predict a result. The ensemble model then combines each base model's forecast into a single overall prediction for the unobserved data. A dataset is used to train a variety of models, and the individual predictions made by each model form the basis of an ensemble model. The ensemble model then combines the outcomes of different models' predictions to get the final outcome.

The top of the line machine learning algorithm for dealing with cardiovascular disorder instances is consequently chosen based totally on the performance of the selected classification method.

1) Data pre-processing

The first step of data mining involves a significant number of missing and noisy values in real-world data. To avoid such issues and create reliable forecasts, these data are pre-processed. The raw data is unreliable and inadequate. The missing data can be eliminated or the mean value can be used to fill in the gaps. In this research, we have organized dataset from UCI repository with 1025 instances and 15 attributes as shown in Fig. 1.

In this paper, we have used total instances (1025) and total number of attributes are (14)+1 (Target variables). Total 9 categorical attributes of heart disease dataset are used in this paper as ['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal', 'target'] in Fig. 2. As a result, the data acquired must be somewhat adjusted using the Lasso features selection approach in order to conduct a successful analysis. Two ensemble phases of classifiers such as max voting and averaging for XG Boost, Random Forest, Multilayer Perception, and Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 for XG Boost, random forest, and multilayer perception.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

Fig. 1. Attributes of cardiovascular disease.

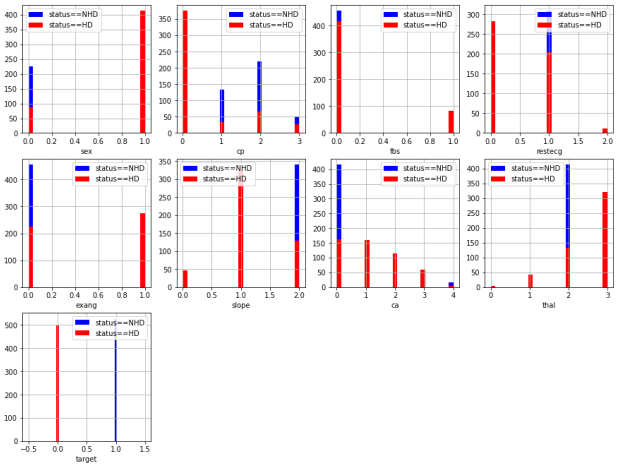

Fig. 2. Representation of heart disease dataset attributes.

2) Machine learning

In data science, machine learning is used to tackle a varietyy of challenges. In machine learning, existing data contributes in the prediction of outcomes. The authors used a strong machine learning approach to study several classes in order to enhance statistical analysis.

3) Features selection

Reduce the quantity of input attributes before beginning data analysis. Not all of the characteristics are equally important in predicting success. The existence of several characteristics adds complexity to the equation while lowering performance. We selected Lasso Regression features as follows [34]:

4) Lasso regression

1. Lasso regression is a regularisation approach for reducing mannequin complexity. Least Absolute and Selection Operator is what it stands for.
2. It's comparable to the Ridge Regression, except instead of a rectangular of weights, the penalty term only incorporates absolute weights.

3. Because it makes use of absolute data, it may minimize the slope to zero, whereas Ridge Regression can solely reduce it to near-zero.
4. L1 regularisation is another identify for it. The Lasso regression equation for the cost function will be utilised.
5. For model evaluation, some of the houses in this method are entirely ignored.
6. As a result, the Lasso regression can aid us in reducing mannequin over fitting as well as function selection.

5) XGBoost

XGBoost was created with the aim of enhancing computing velocity and model efficiency. XGBoost's key features include [35]:

1. Creates decision trees in parallel.
2. Using distributed computing techniques to evaluate big, problematic models.
3. Analyzing large datasets with Out-of-Core Computing.
4. Making the most use of assets by means of enforcing cache optimization.

6) Random forest

The random wooded area is formed in two phases: the first is to mix N choice trees to construct the random forest, and the second is to make predictions for each tree created in the first phase.

The steps of the working method are as follows [36]:

1. Pick K statistics factors at random from the education set.
2. Create selection bushes for the statistics factors you have chosen (Subsets).
3. Decide on the variety N for the choice timber you want to create.
4. Repetition of Steps 1 and 2.
5. Find the forecasts of every choice tree for new facts points, and then allocate the new statistics points to the category with the most votes.

7) Multi-layer perceptron

A perception receives *n* features as input ($x = x_1, x_2, \ldots, x_n$), and every of these features is associated to a weight. Input aspects ought to be numeric. So, nonnumeric enter facets have to be transformed to numeric ones in order to use a perception. The Working system can be defined in the beneath steps [37]:

1. Make a list of characteristics to use as input variables.
2. They accomplish this through merging quite a few neurons grouped in at least three layers:
    a. A single enter layer that simply passes the input aspects to the first hidden layer.
    b. One or more perception layers that are concealed.
    c. The characteristics dispersed by means of the enter layer are sent into the first hidden layer as inputs.
    d. The output of each perception from the previous layer is fed into the different hidden levels.
    e. One perception output layer that receives inputs
    f. Each perception of the closing hidden layer's output.
    g. Perception from the preceding layer's output.
    h. One perception output layer that receives inputs
    i. Each perception's output from the remaining hidden layer.

8) Stacking

The Working method can be defined in the beneath steps [38]:

1. Similar to K-fold cross-validation, we divided the coaching statistics into K-folds. The K-1 components are fitted using a fundamental model, and predictions for the Kth section are created.
2. For every issue of the education data, we do so.
3. The overall performance of the fundamental model on the take a look at set is calculated by fitting it to the whole instruct statistics set.
4. We repeat the previous three tiers for the remaining fundamental models.
5. The 2nd degree model comprises predictions from the teach set as features.
6. On the take a look at set, the 2d level mannequin is employed to produce a prediction.

The stacking ensemble method works on the prediction results obtained by various base classifiers on majority voting method. In this method the votes obtained by base classifier are combined to predict the classification accuracy.

9) Evaluation matric

The accuracy, Mean Absolute Error (MAE), and Root Imply Squared Error (RMSE) have been all investigated. The accuracy of continuous variables is calculated the use of MAE and RMSE [39, 40]. The average magnitude y_i of the mistake in a series of predictions is estimated by MAE in Eq. (1).

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_i - \hat{y}_j\right| \quad (1)$$

RMSE is a metric that measures the average size of a mistake. It is the square root of the average of squared deviations between forecast and actual observation, where *n* is the number of observations and underlying physical quantity, such as the exact distance as given in the Eq. (2).

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_i - \hat{y}_j\right)^2} \quad (2)$$

The Relative Absolute Error (RAE) is a simple predictor that averages the actual value, where error is the entire absolute error represented in percentages.

$$E_i = \frac{\sum_{j=1}^{n}\left|P_{ij} - T_j\right|}{\sum_{j=1}^{n}\left|T_j - \bar{T}\right|} \quad (3)$$

The response variable for the relevant components is calculated using the prediction Eq. (3), where P_{ij} is the predictor for model I with *j* records. T_j is the goal value for each of the *j* records, and *T* is defined in the equation.

B. Method

In this research, we have organized dataset from UCI repository with 1025 instances and 15 attributes. All the training dataset are prepared by lasso features selection method in Fig. 3. The experimental setup is prepared in four stages. First and second stage run by XG Boost, Random Forest, Multilayer Perception classifiers. These classifiers trained with 70% instances of disease attributes and test with 30% instances. The Stage I & II, calculated error values as MAE, MSE, RMSE and Scattered Index values. In Stage III & IV, we prepared two proposed two ensemble techniques as max voting and stacking by different classifiers such as XG Boost, Random Forest, Multilayer Perception to detect better results and test on 50% & 70% instances. Finally compare Stage I, II, III & IV, prediction model and check effect on calculated error values of MAE, MSE, RMSE and Scattered Index values. Thus, based totally on the overall performance of the selected classification algorithm, the fantastic computer learning algorithm is identified for dealing with cardiovascular sickness cases. The proposed coronary heart disease prediction aim is to help specialists in making knowledgeable choices and predictions thru the use of laptop learning techniques.

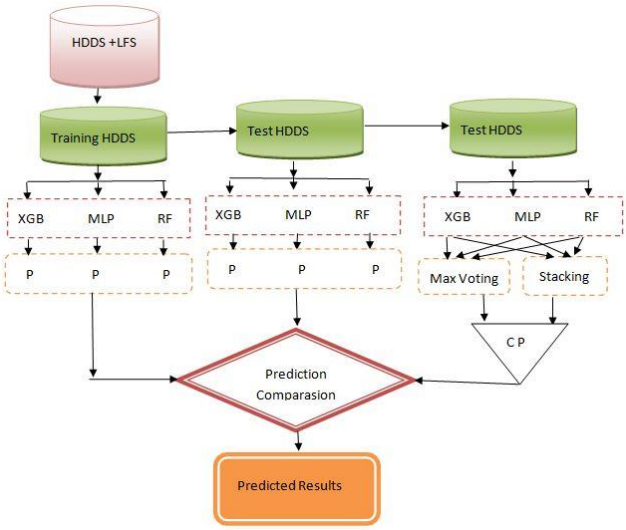

Fig. 3. Representation of proposed model for heart disease dataset attributes prediction.

IV. Result and Discussion

A. Result

Lasso regression technique reducing over fitting of the model so the positive coefficients correspond to the high-skilled features (cp, slop, thalach, fbs and restecg) while negative coefficients are typical for the low-skilled features (age, chol, trestbps, oldpeak, exang, thal, sex and ca). Large absolute value means that feature is more important. Fig. 4 demonstrates that the motion to right and to left is the most characteristic of HDD. In Fig. 4, Lasso regression is a regularization approach represents high and low values for each feature of heart disease dataset and reducing model complexity.

Because it makes use of absolute data, it may minimize the slope to zero, whereas Ridge Regression can solely reduce it to near-zero. As a result, the Lasso regression can aid us in reducing mannequin over fitting as well as function selection. From Fig. 4, it is clear that disease features stand between +10 to −10.

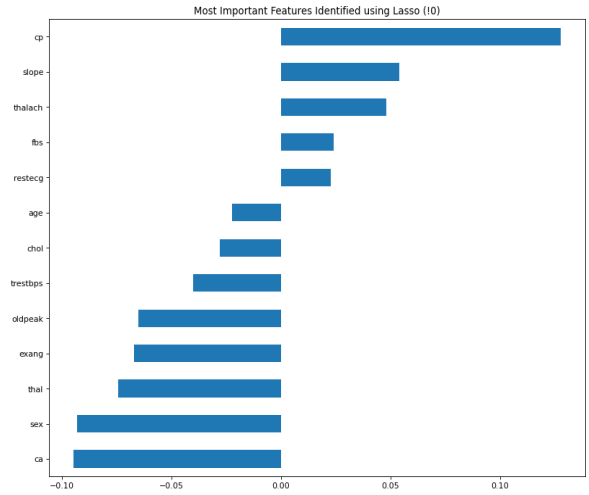

Fig. 4. Representation of Lasso features selection techniques computation for HDD.

We have train used classifiers on 70% heart disease dataset instances and generate prediction model. The analysis of machine learning techniques for the heart disease prediction. The machine learning model performance as in Table I:

- The MAE, MSE, RMSE and Scattered Index values obtained for the XGBoost, are 073.19, 5358.239, 73.18, and 41.36, respectively.
- The MAE, MSE, RMSE and Scattered Index values obtained for the Multilayer Perception, are 016.01, 0256.001, 16.03 and 09.04, respectively.
- The MAE, MSE, RMSE and Scattered Index values obtained for the Random Forest, are 038.49, 1482.09, 38.47 and 21.75, respectively.

TABLE I: Representation of Training and Testing Computation Error Model for Classifiers

	Model	MAE	MSE	RMSE	SI
HDDS Training Model (70%)	XGB	073.01	5350.29	74.17	42.39
	MLP	017.03	0258.09	17.09	10.02
	RF	039.47	1484.93	39.69	22.69
HDDS Test Model (30%)	XGB	173.19	5358.239	73.18	41.36
	MLP	016.01	0256.001	16.03	09.04
	RF	038.49	1482.09	38.47	21.75

We have test used classifiers on 30% heart disease dataset instances and generate prediction model. The machine learning model performance in the heart disease.

- The MAE, MSE, RMSE and Scattered Index values obtained for the XGBoost, are 073.01, 5350.29, 74.17 and 42.39, respectively.
- The MAE, MSE, RMSE and Scattered Index values obtained for the Multilayer Perception, are 017.03, 0258.09, 17.09 and 10.02, respectively.
- The MAE, MSE, RMSE and Scattered Index values obtained for the Random Forest, are 039.47, 1484.93, 39.69 and 22.69, respectively.

In this study, we organized two different ensemble techniques as Max Voting and Stacking for these three classifiers (XGBoost, Multilayer Perception and Random Forest). Finally test Max Voting and Stacking ensemble techniques on 70% & 50% on heart disease dataset instances and calculated results as in Table II:

- With the Max Voting Ensemble strategy, the MAE, MSE, RMSE, and Scattered Index values were 4.06, 12.37, 4.35, and 4.63 accordingly in the first iteration. For the Stacking Ensemble technique, the values were 3.47, 8.79, 3.68, and 3.72.
- In the second experiment, we tested the Max Voting and Stacking ensemble techniques on instances from a dataset of 50% heart disease, and the calculated results were as follows: In the first iteration, we discovered that the MAE, MSE, RMSE, and Scattered Index values obtained for the Max Voting Ensemble technique were 3.17, 8.7, 3.78, and 4.01, respectively, and for the Stacking Ensemble technique, they were 2.76, 6.07, 3.25, and 3.45.

TABLE II: REPRESENTATION OF TRAINING AND TESTING COMPUTATION ERROR MODEL FOR ENSEMBLES ON 70% AND 30% HDD

	Model	MAE	MSE	RMSE	SI
HDDS Test Model (70%)	E1	4.06	12.37	4.35	4.63
	E2	3.47	8.79	3.68	3.72
HDDS Test Model (30%)	E1	3.17	8.7	3.78	4.01
	E2	2.76	6.07	3.25	3.45

B. Discussion

The results of the study showed a very close evaluation score for the 70% and 30% heart disease dataset instances. With the results of Tables I and II, Figs. 5–8, we found Multilayer Perception algorithms perform better or evaluated low error values for MAE. The XGBoost and Random Forest perform poor for MSE compare to Multilayer Perception algorithms. With the results, the calculated values of RMSE and Scattered Index are high of XGBoost and Random Forest algorithm but Multilayer Perception calculated very low error values.

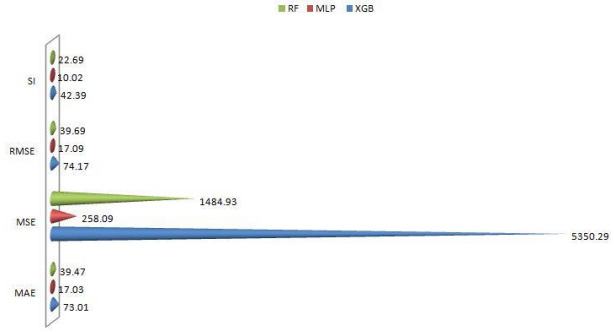

Fig. 5. Representation of training computation error model for classifiers of HDD.

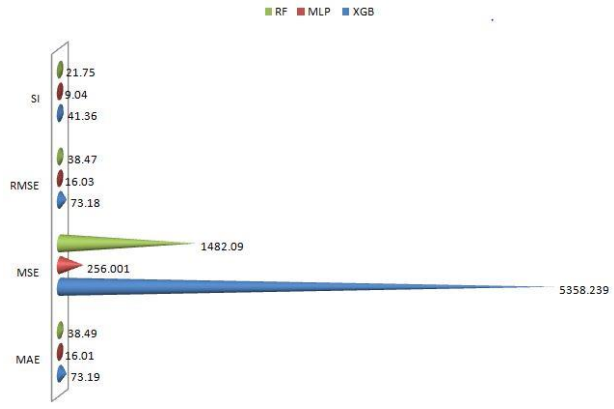

Fig. 6. Representation of testing computation error model for classifiers of HDD.

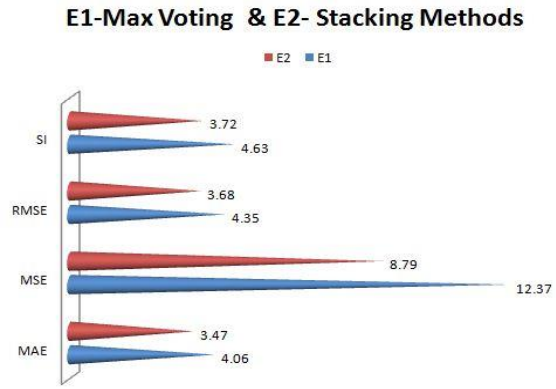

Fig. 7. Representation of training and testing computation error model for classifiers on 70% HDD.

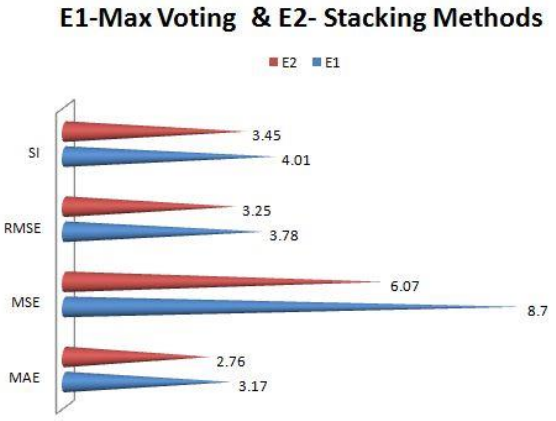

Fig. 8. Representation of training & testing computation error model for ensembles on 70% & 30% HDD.

In the above experiment, it is clear that XGBoost and Random Forest algorithms perform poor for both 70% and 30% heart disease dataset instances. In the second experiment, we compared, organized Max Voting Ensemble technique and Stacking Ensemble technique and test on heart disease dataset instances 70% and 50%. With the results, we found Max Voting technique perform poor and calculated high error values compare to Stacking Ensemble Technique for 70% and 50% heart disease dataset instances. Figs. 7 and 8 represent better values of Stacking Ensemble technique compare to other ensemble Max Voting method.

V. CONCLUSION

In this research, we have proposed two ensemble techniques as max voting and stacking for ensemble different classifiers such as XG Boost, Random Forest, and Multilayer Perception to detect better results. Thus, based on the overall performance of the chosen classification algorithm, the high-quality computing device studying algorithm is recognized for dealing with cardiovascular disease cases. The outcomes of the find out about confirmed a very close assessment rating for the 70% and 30% heart disorder dataset situations in Stage I and II. With the results of Tables I and II, we found Multilayer Perception algorithms perform better or evaluated low error values for MAE, MSE, RMSE and Scattered Index compare to XGBoost and Random Forest algorithms. In the Stage III and IV of experiment, we test on 70% and 30% instances of disease dataset and compared, organized Max Voting Ensemble technique and Stacking Ensemble technique with

the results.

In Figs. 7 and 8, the Stacking Ensemble technique demonstrates superior performance compared to the Max Voting method for ensemble learning. The Stacking Ensemble technique combines the outputs of multiple base models using a meta-model, leveraging their individual strengths to improve overall predictive accuracy. This allows it to capture complex relationships and patterns in the data more effectively.

Future learn about may center of attention on improving the model's effectiveness in the categorization of additional kinds of medical data, ensuing in a greater economical and time-saving answer for patients and clinicians alike. Furthermore, research can be executed to investigate high-dimensional information in coaching for future study.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Gyanendra Kumar Pal has done analysis, coding, and final draft of the manuscript under the supervision of Sanjeev Gangwar. All authors had approved the final version.

REFERENCES

[1] B. A. Snousy, H. M. El-Deeb, K. Badran, and I. A. A. Khlil, "Suite of decision tree-based classification algorithms on cancer gene expression data," *Egyptian Informatics Journal*, vol. 12, no. 2, pp. 73–82, 2011.

[2] S. J. Al'Aref, K. Anchouche, G. Singh *et al*., "Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging," *European Heart Journal*, vol. 40, no. 24, pp. 1975–1986, 2019.

[3] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[4] R. Alizadehsani, M. Roshanzamir, M. Abdar *et al*., "A database for using machine learning and data mining techniques for coronary artery disease diagnosis," *Scientific Data*, vol. 6, no. 1, pp. 227–313, 2019.

[5] A. Quesada, A. Lopez-Pineda, V. F. Gil-Guillen *et al*., "Machine learning to predict cardiovascular risk," *International Journal of Clinical Practice*, vol. 73, no. 10, e13389, 2019.

[6] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 211–215, 2019.

[7] T. Leiner, D. Rueckert, A. Suinesiaputra *et al*., "Machine learning in cardiovascular magnetic resonance: Basic concepts and applications," *Journal of Cardiovascular Magnetic Resonance*, vol. 21, no. 1, pp. 61–14, 2019.

[8] M. Alaa, T. Bolton, E. D. Angelantonio, J. H. F. Rudd, and M. Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants," *PLoS One*, vol. 14, no. 5, e0213653, 2019.

[9] U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, 3860146, 2018.

[10] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *SN Computer Science*, vol. 1, no. 6, pp. 1–6, 2020.

[11] Y. Khourdifi, M. Bahaj, and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 1, pp. 242–252, 2019.

[12] A. D'Souza, Y. Wang, C. Anderson, A. Bucchi, M. Baruscotti, S. Olieslagers, P. Mesirca, A. B. Johnsen, S. Mastitskaya, H. Ni *et al.*, "A circadian clock in the sinus node mediates day-night rhythms in Hcn4 and heart rate," *Hear. Rhythm*, vol. 18, pp. 801–810, 2021.

[13] A. Alharbi, W. Alosaimi, R. Sahal, and H. Saleh, "Real-time system prediction for heart rate using deep learning and stream processing platforms," *Complexity*, 5535734, 2021. https://doi.org/10.1155/2021/5535734

[14] T. Chen and M. Lucock, "The mental health of university students during the COVID-19 pandemic: An online survey in the UK," *PLoS ONE*, vol. 17, e0262562, 2022.

[15] T. Chen, E. Keravnou-Papailiou, and G. Antoniou, "Medical analytics for healthcare intelligence—Recent advances and future directions," *Artif. Intell. Med*., vol. 112, 102009, 2021.

[16] P. Su, T. Chen, J. Xie, Y. Zheng, H. Qi, D. Borroni, Y. Zhao, and J. Liu, "Corneal nerve tortuosity grading via ordered weighted averaging-based feature extraction," *Med. Phys*., vol. 47, pp. 4983–4996, 2020.

[17] T. Chen, C. Shang, P. Su, E. Keravnou-Papailiou, Y. Zhao, G. Antoniou, and Q. Shen, "A decision tree-initialised neuro-fuzzy approach for clinical decision support," *Artif. Intell. Med*., vol. 111, 101986, 2021.

[18] S. A. Knox, T. Chen, P. Su, and G. Antoniou, "A parallel machine learning framework for detecting Alzheimer's disease," in *Proc. International Conference on Brain Informatics*, *Lecture Notes in Computer Science*, Springer, 2021, vol. 12960, pp. 423–432.

[19] T. Chen, G. Antoniou, M. Adamou, I. Tachmazidis, and P. Su, "Automatic diagnosis of attention deficit hyperactivity disorder using machine learning," *Appl. Artif. Intell*., vol. 35, pp. 657–669, 2021.

[20] Q. Li, A. Campan, A. Ren, and W. E. Eid, "Automating and improving cardiovascular disease prediction using machine learning and EMR data features from a regional healthcare system," *Int. J. Med. Inform*., vol. 163, 104786, 2022.

[21] P. Kakria, N. K. Tripathi, and P. Kitipawang, "A real-time health monitoring system for remote cardiac patients using smartphone and wearable sensors," *Int. J. Telemed. Appl*., 373474, 2015. doi: 10.1155/2015/373474

[22] Y. Jiang, Z.-G. Yang, J. Wang, R. Shi, P.-L. Han, W.-L. Qian *et al*., "Unsupervised machine learning based on clinical factors for the detection of coronary artery atherosclerosis in type 2 diabetes mellitus," *Cardiovasc Diabetol*, vol. 21, no. 1, p. 259, 2022.

[23] E. P. Ephzibah, "A neuro fuzzy expert system for heart disease diagnosis," *Comput. Sci. Eng.: Int. J*., vol. 2, no. 1, pp. 7–23, 2012.

[24] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Inf. Syst.*, pp. 1–21, 2018.

[25] T. A. Asfaw, "Performance comparison of k-nearest neighbors and Gaussian naïve bayes algorithms for heart disease prediction," *Int. J. Eng. Sci. Invent. (IJESI)*, vol. 8, no. 8, pp. 45–48, 2019.

[26] M. A. Khan, S. Abbas, A. Atta, A. Ditta, H. Alquhayz *et al*., "Intelligent cloud based heart disease prediction system empowered with supervised machine learning," *CMC-Comp. Mater. Continua*, vol. 65, no. 1, pp. 139–151, 2020.

[27] A. M. Alaa, T. Bolton, E. D. Angelantonio, J. H. F. Rudd, and M. Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants," *PloS One*, vol. 14, no. 5, 2019.

[28] J. A. Edward, K. Josey, G. Bahn, L. Caplan, J. E. B. Reusch, P. Reaven *et al*., "Heterogeneous treatment effects of intensive glycemic control on major adverse cardiovascular events in the ACCORD and VADT trials: A machine-learning analysis," *Cardiovasc Diabetol*., vol. 21, no. 1, p. 58, 2022.

[29] T. Kasbe and R. S. Pippal, "Enhancement in diagnosis of coronary artery disease using fuzzy expert system," *Int. J. Sci. Res. Comput. Sci. Eng. Informat. Technol*., vol. 3, no. 3, pp. 1324–1331, 2018.

[30] K. H. Miao and H. J. Miao, "Coronary heart disease diagnosis using deep neural networks," *Int. J. Adv. Comput. Sci. Appl*., vol. 9, no. 10, 2018.

[31] E. K. Hashi and M. Z. Shahid, "Developing a hyperparameter tuning based machine learning approach of heart disease prediction," *J. Appl. Sci. Proc. Eng.*, vol. 7, no. 2, pp. 631–647, 2020.

[32] S. Y. Siddiqui, A. Athar, M. A. Khan, S. Abbas, Y. Saeed *et al*., "Modelling, simulation and optimization of diagnosis cardiovascular disease using computational intelligence approaches," *J. Med. Imag. Health Informat*., vol. 10, no. 5, pp. 1005–1022, 2020.

[33] K. M. Almustafa, "Prediction of heart disease and classifiers' sensitivity analysis," *BMC Bioinf*., vol. 21, no. 1, pp. 1–18, 2020.

[34] V. Dave, H. Thakker, and V. Vakharia, "Fault identification of ball bearings using fast Walsh Hadamard transform, LASSO feature selection, and random forest classifier," *FME Transactions*, vol. 50, no. 1, p. 203, 2022.

[35] X. Jing *et al.*, "Remote sensing monitoring of winter wheat stripe rust based on mRMR-XGBoost algorithm," *Remote Sensing*, vol. 14, no. 3, p. 756, 2022.
[36] D. C. Yadav and S. Pal, "Analysis of heart disease using parallel and sequential ensemble methods with feature selection techniques: Heart disease prediction," *International Journal of Big Data and Analytics in Healthcare (IJBDAH)*, vol. 6, no. 1, pp. 40–56, 2021.
[37] Y. Zhang *et al.*, "A stacking ensemble algorithm for improving the biases of forest aboveground biomass estimations from multiple remotely sensed datasets," *GIScience& Remote Sensing*, pp. 1–16, 2022.
[38] N. H. Jasni *et al.*, "Prediction of player position for talent identification in association netball: A regression-based approach," *International Journal of Advances in Intelligent Informatic*s, vol. 8, no. 1, 2022.
[39] A. Elbeltagi, C. B. Pande, M. Kumar, A. D. Tolche, S. K. Singh, A. Kumar, and D. K. Vishwakarma, "Prediction of meteorological drought and standardized precipitation index based on the Random Forest (RF), Random Tree (RT), and Gaussian Process Regression (GPR) models," *Environmental Science and Pollution Research*, vol. 17, pp. 1–20, 2023.
[40] A. Elbeltagi, M. Kumar, N. L. Kushwaha, C. B. Pande, P. Ditthakit, D. K. Vishwakarma, and A. Subeesh, "Drought indicator analysis and forecasting using data driven models: Case study in Jaisalmer, India," *Stochastic Environmental Research and Risk Assessment*, vol. 37, no. 1, pp. 113–131, 2023.