Enhanced Stacked Ensemble-Based Heart Disease Prediction with Chi-Square Feature Selection Method

Raniya R. Sarra [1*], Ivan Isho Gorial [2], Reham Raad Manea [3], Ayad E. Korial [4], Mustafa Mohammed [5], Yousif Ahmed [6]

[1, 3, 4, 5, 6] Computer Engineering Department, University of Technology-Iraq, Baghdad 10001, Iraq
[2] Control and Systems Engineering Department, University of Technology-Iraq, Baghdad 10001, Iraq
Email: [1] raniya.r.aziz@uotechnology.edu.iq, [2] ivan.i.gorial@uotechnology.edu.iq, [3] 120127@ uotechnology.edu.iq,
[4] ayad.e.korial@uotechnology.edu.iq, [5] ce.20.280@student.uotechnology.edu.iq, [6] ce.20.279@student.uotechnology.edu.iq
*Corresponding Author

Abstract—Heart disease (HD) is the primary cause of death globally, requiring more accurate, affordable diagnostic technologies. Traditional HD diagnostic methods are adequate but expensive and limited, creating a need for creative alternatives. Machine learning (ML) is one of the many sophisticated technologies healthcare systems use to predict diseases. This work aims to enhance the accuracy and efficiency of HD diagnosis by developing a stacked ensemble classifier that combines predictions from different ML classifiers and uses chi-square feature selection to prioritize significant features. Combining predictions from three basic ML classifiers—decision trees (DT), support vector machines (SVM), and multilayer perceptron (MLP)—the paper creates a stacked ensemble classifier. To raise diagnostic accuracy, this stacked ensemble classifier maximizes the strengths of base classifiers and reduces their errors. Furthermore, applying the chi-square feature selection approach, the study finds five important features for training the classifiers on the Cleveland dataset with thirteen (13) features. Selecting only important features through feature selection minimizes dimensionality, simplifies the classifier, and improves computational performance. This also reduces overfitting, increases generalizability, and speeds up diagnosis, making it more viable for real-time clinical applications. Before and following the feature selection procedure, the ensemble classifier performance is assessed against the base classifiers concerning the accuracy, recall, precision, and f1-score. These metrics are chosen for their ability to validate the effectiveness of the proposed diagnostic tool. With an accuracy of 85.5%, the stacked ensemble classifier exceeded base classifiers before feature selection. After feature selection, the stacked ensemble classifier's accuracy improved to 90.8%. These results underline the proposed method as an inexpensive and more efficient diagnostic tool for HD as compared to current methods, enabling earlier HD detection and lowering healthcare costs. In conclusion, this creative method could alter healthcare systems by providing a highly accurate and affordable diagnostic tool for clinical use.

Keywords—Heart Disease; Machine Learning; Stacking Ensemble; Feature Selection; Chi-Square.

I. INTRODUCTION

Commonly affecting middle-aged and older people, heart disease (HD) is a prevalent condition with often deadly complications [1][2]. One-third of all deaths globally come from HD. It kills over 17 million people worldwide annually [3]. Many bad habits like smoking, drinking alcohol, lack of physical activity, poor eating habits, and stress are making people more prone to HD [4][5]. Apart from the above factors, HD is a complicated condition influenced by several other non-changeable factors, such as age and gender. The complicated interaction of non-changeable and changeable factors produces a varied risk profile that differs substantially between individuals, which makes precise prediction of heart disease difficult. For instance, smoking can raise the risk of HD particularly in older people with a family history of HD. To handle this complexity, improved diagnostic methods are needed to take into consideration these varied and overlapping factors [6][7].

Because of the several risk factors and the fact that symptoms may not show until the disease is somewhat advanced, early identification of HD might be challenging to do manually [8][9]. Also, because conventional diagnostic means, such as electrocardiograms, physical exams, and analysis of vital signs, are so expensive, prone to human error, and time-consuming, there is an urgent need to find more accurate, cost-effective, and efficient diagnostic alternative tools for HD [10][11]. With the help of computer-aided systems, cardiologists can better diagnose heart problems. These systems aid in reducing the risk of HD by predicting it quickly and early on. Also, the use of such systems can help medical experts save time and cost [12][13].

Predictive machine learning (ML) models have been suggested as a solution to the problems with conventional diagnostic approaches in healthcare [14][15]. Unlike conventional approaches, ML-powered diagnostic tools are superior at sifting through complicated and massive datasets with high-dimensional features, a challenge that conventional approaches frequently face. This capacity enables ML to identify minor patterns that might otherwise be ignored. In addition, by automating the diagnosis process and therefore producing faster and more dependable diagnosis results, ML models can improve diagnostic efficiency and help to save costs by decreasing the need for repeated, manual data analysis. On top of automating diagnoses, these models can also detect potential health risks before they manifest [16][17]. ML could use existing data to create innovative diagnostic systems that accurately identify HD [18]. ML techniques such as logistic regression (LR), support vector machine (SVM), k-nearest neighbor (KNN), decision tree (DT), naïve Bayes (NB), and neural network (NN) are among

the most popular for discovering complicated patterns and correlations in data [4][14][19]. While SVM is noted for its ability to handle high-dimensional data, LR is recognized for its ease of use and interpretability, particularly in situations involving medical decision-making. In contrast, KNN is commonly used for the natural way it uses similarity to training data points to classify data. When there isn't a lot of data, NB is good because it works well with small datasets and assumes that features are independent. While NN are more computationally costly, they are well-suited to high-dimensional medical datasets, such as those utilized for predicting heart disease, since they are excellent at capturing complicated non-linear correlations [20].

Researchers are now proposing ML techniques like ensemble classifiers to detect HD better. An ensemble model improves prediction accuracy by combining predictions from multiple base classifiers (C_1, C_2, C_3, ..., C_n) [21][22]. To determine the best tool for predicting HD, some have compared single-base ML classifiers to their ensemble equivalents, where their high classification accuracy has proved the usefulness of ensemble methods [23]. However, ML classifiers learn too much from the training data due to the large number of features, which causes a drop in the classifier's performance [11][24]. Therefore, the prioritization of significant features through feature selection methods is a crucial tactic for improving HD prediction abilities and boosting the accuracy of ML systems [25]. Although feature selection has the potential to improve model performance, it also carries the potential risk of overfitting, a condition in which the model becomes too specific to the training data and loses its capacity to generalize [26].

Implementing the stacked ensemble classifier and feature selection strategy in practical environments could provide numerous difficulties. processing complexity is one major obstacle since stacking several models increases the processing resources needed, hence maybe restricting their viability in contexts with limited resources. Including such complex models into current clinical procedures could also call for major changes including training for doctors and healthcare professionals to guarantee seamless adoption and efficient utilization. Another important consideration is user adoption since medical professionals could be reluctant to embrace new technologies without enough support throughout the changeover period and without strong proof of their superiority over current practices [27].

Hence, this proposed work combines feature selection with ensemble learning. Our work is unique in that it prioritizes both high predictive accuracy and computing efficiency, as opposed to many other research papers that solely concentrate on the model's predictive accuracy. This study builds a simple stacked ensemble classifier for HD prediction using three different base ML classifiers: decision trees (DT), support vector machines (SVM), and multilayer perceptron (MLP). The proposed stacked ensemble classifier was created to determine whether a person has heart disease (binary classification).

This study also tries to improve the accuracy of predicting HD risk using the chi-square feature importance technique that chooses the most correlated features strongly influencing the target variable. This technique drastically reduces the feature space without compromising the prediction accuracy. Given that the chi-square feature selection method is especially meant to assess the independence between categorical features and the target variable, which is very important in HD datasets where many attributes are categorical, it is especially well-suited for this work.

The main objective of our study is to enhance the accuracy of stacked ensemble-based HD prediction with the chi-square feature selection technique. Not only does better identification of HD improve patient outcomes but it also has the potential to greatly lower public healthcare expenses. Dealing with this difficulty might result in improved public health management and more effective allocation of healthcare resources. A summary of this study's main contributions is as follows:

- To improve the accuracy of HD diagnosis using a stacked ensemble classifier that integrates the predictions of many ML base classifiers.

- To extract the most valuable features using a chi-square approach to reduce computation time and improve diagnosis accuracy.

- To assess the proposed method for HD diagnosis and compare it to the existing methods using a variety of performance metrics.

The rest of the paper is organized as follows: Section II summarizes earlier studies on HD detection and prediction using stacked ensemble ML classifiers; Section III describes the methodology, along with thorough explanations of the data collection, the ML base classifiers, and the chi-square method; Section IV shows the experimental findings along with a comparison to current methodologies; Section V summarizes our results and their significance.

II. RELATED WORK

The prediction of heart diseases has been the subject of research using various machine learning (ML) algorithms and ensemble methods [28]. This section reviews and analyzes a few of these studies. The development of expert systems in healthcare makes extensive use of ML techniques and approaches.

A lot of researchers have analyzed various ML methods, including base classifiers, hybrid, and ensemble learning methods for heart disease diagnosis [29]-[31]. However, while examining medical records, many factors could ultimately influence the diagnosis. Thus, a lot of researchers are improving the precision of prediction models using feature selection approaches, which identify which features are the most important [32][33]. According to the review, numerous studies used ensemble classifiers, and few studies have tried improving them with feature selection methodologies.

Researchers in study [34] implemented an ensemble model using the voting classifier. naïve bayes (NB), logistic regression (LR), and support vector machine (SVM) were the components that made up this vote classifier. Using six

selected features by the extra tree classifier for prediction, the accuracy achieved by this ensemble classifier was 87.53%.

Authors in [35] designed a voting ensemble classifier for predicting heart diseases accurately. It included both ML and deep learning classifiers—DT, KNN, RF, KDNN, DNN, and XGB—as base classifiers. On the HD dataset derived from the UCI repository, the ML ensemble classifier exceeded all other classifiers in accuracy—reaching 88.70%. With an accuracy of 86.49%, the DL-stacked classifier also showed dependable performance but was rather less effective.

With an ensemble approach, researchers in [4] improved the accuracy of heart disease prediction. Preprocessing the Cleveland dataset, the study selected the best features using an extra trees classifier. Five basic classifier predictions were combined using ensemble techniques combining bagging and majority voting (NB, K-NN, Artificial Neural Network (ANN), LR, and DT). Proving to be the most efficient, the bagging ensemble method exceeded all other techniques and individual base classifiers with 87.78%.

Likewise, [36] explored ensemble classification—a method combining several classifiers to improve accuracy. Using the Cleveland dataset, researchers assessed several ensemble methods—bagging, boosting, stacking, and majority voting—to predict heart disease. Results reveal that ensemble techniques greatly raise the prediction ability of poor classifiers. With stacking and bagging shortly behind, the majority voting specifically improved their accuracy by up to 7.26%. With a maximum accuracy of 85.48% attained via majority voting, stacking using the RF as the meta-classifier enhanced accuracy by up to 6.93%.

Using an ensemble strategy based on hard voting that combines machine learning classifiers (RF, SVM, and K-NN) and deep learning classifiers (Long Short-Term Memory and Gated Recurrent Unit), authors in [37] evaluated the risk of heart disease. With an accuracy rate of 85.71%, the paper indicates that the proposed voting-based ensemble model is efficient.

Using several ensemble ML approaches—including classification and regression trees (CART), gradient boosting machine (GBM), Adaboost, K-NN, MLP, stochastic gradient descent (SGD), SVM, and NB—the suggested work in [25] CART and GBM were the models with highest performance. The CART model notably attained the best accuracy—87.65%. It also applied light GBM, RF, Chi-square (Chi-2), recursive feature elimination (RFE), logistic regression (L2 penalty), and pearson's correlation coefficient among other feature selection techniques. Using the majority voting approach, seven features were selected following the application of these approaches.

Authors in [38] also used a majority vote approach to predict heart disease. Three separate classifiers—SVM, NB, and ANN—were merged under the ensemble model. With 87.05%, the ensemble model produced better accuracy than individual classifiers.

Stacking three independent classifiers—gradient boosting machine (GBM), extreme gradient boosting machine (XGBoost), and RF—the authors in [39] used an ensemble model to detect cardiac disease. For feature selection they combined particle swarm optimization (PSO) with correlation-based feature selection (CFS). For the Cleveland dataset, seven characteristics were used; the accuracy obtained was 85.71%.

In recent years, multiple researchers have studied and evaluated several ensemble methods. Most of them concentrated on improving the accuracy of heart disease prediction. This work intends to enhance the performance of base classifiers for heart disease prediction using an ensemble approach and feature selection method.

III. METHODOLOGY

The following sections provide a detailed description of this study's methodology, including the description of the dataset used to diagnose heart disease (HD), the preprocessing methods, the machine learning (ML) classifiers, the ensemble classifier proposed to enhance diagnostic accuracy, and the feature selection technique used to improve the ensemble classifier's performance.

A. Heart Disease Dataset and Preparation

In this study, a publicly accessible heart dataset (Cleveland) was retrieved from UCI's ML repository. This dataset contains 303 records and 76 features, yet for HD prediction, most studies only used fourteen (14) features [40]. Thirteen (13) features were input features: age, chest pain type, gender, fasting blood sugar, cholesterol, resting ECG, heart rate, number of fluoroscopically colored vital arteries, ST slope, blood pressure, exercise-induced angina, ST depression, and thalassemia. With binary values of 0 and 1, the final output feature reveals the presence or absence of heart disease [41]. More detailed details regarding the dataset's features are shown in Table I.

TABLE I. CLEVELAND DATASET FEATURES DESCRIPTION [42]

Name	Description
Age	patients in this sample range in age from 29 to 79
Sex	If a patient is male, their gender value will be 1, and if they are female, their gender value will be 0
Cp	Type 1 angina is caused by reduced blood flow to the heart muscles. Type 2 Angina occurs during mental or emotional stress. Non-angina chest pain may be caused by various reasons and may not often be due to actual heart disease. The fourth type, Asymptomatic, may not be a symptom of heart disease.
Trestbps	resting blood pressure
Chol	cholesterol level
Fbs	assigned a value of one if fasting blood sugar is less than 120 mg/dl and a value of zero if it is more significant than
Restecg	electrocardiogram at rest (normal = 0, ST-T wave abnormalities = 1, left ventricular hypertrophy = 2)
Thalach	maximum heart rate
Exang	exercise-induced angina is recorded as one (1) if there is pain and zero (0) if there is no pain.
Oldpeak	Exercise-induced ST segment depression
Slope	An exercise's peak ST segment's slope is represented as three (3) for downsloping, two (2) for flat, or one (1) for upsloping
Ca	number of fluoroscopically colored vital arteries
Thal	The results of the thallium stress test: 3 normal, 6 fixed, and 7 reversible defects
Num	The class attribute is 0 for normal and 1 for patients diagnosed with HD.

As seen in Fig. 1, the Cleveland dataset has 165 individuals diagnosed with HD, whereas 138 individuals do not have the disease. The obtained dataset is balanced since the minority class makes up more than 10% to 20% of the entire data [40].

The Cleveland dataset, with just 303 records, may limit generalizability. Small datasets are prone to overfitting, in which the classifier performs well on the training data but fails to generalize to unseen data. This is so because the classifier might learn noise or patterns from the small dataset that may not exist in a larger population. Also, the minimal diversity in such a dataset limits generalization across populations. To ensure high performance across multiple demographic groups, future study might examine the classifier's generalizability on larger and more diverse datasets.
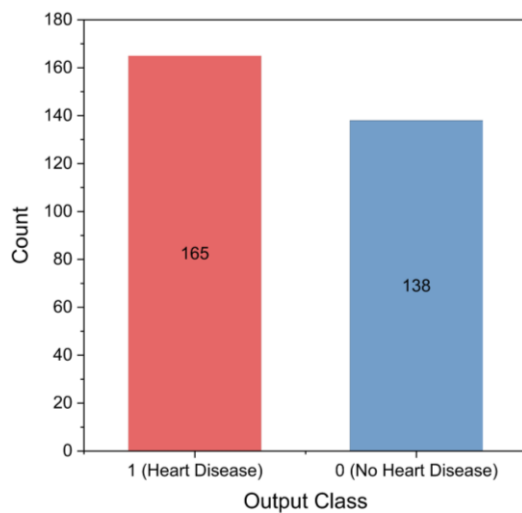

Fig. 1. Target classes of Cleveland dataset

Moreover, certain features had uneven data distribution, which could lead to a less accurate prediction. Thus, a feature scaling method provided in (1) was used to ensure a normal distribution. This method standardizes data points (x_i) by subtracting the mean (μ) from it and dividing by the standard deviation (σ) [43].

$$x_{std} = \frac{x_i - \mu}{\sigma} \tag{1}$$

This research particularly makes use of this feature scaling technique to guarantee that all classifiers are not biassed by the scale of the features. Also, this method is particularly useful when the classifier relies on distance-based algorithms where the scale of the data can considerably affect performance.

B. Ensemble and Machine Learning Classifiers

Algorithm 1 shows the proposed ensemble classifier with a feature selection method developed in Python using the scikit-learn library. The work was conducted in four primary steps: (1) data preparation and scaling (using standard scalar method). (2) selecting important features (using the chi-square technique). (3) training and evaluating the performance of all base ML classifiers in terms of accuracy, recall, precision, and f1-score. (4) Creating the stacking ensemble classifier and evaluating its performance in terms of accuracy, recall, precision, and f1-score.

For this work, the Cleveland heart dataset was used since it comprises important clinical features needed to diagnose HD. In the pre-processing step, the data was scaled using the standard scalar method provided in (1). The chi-square statistical approach was used to identify the most significant features during the feature selection process.

The next step involved training three base ML classifiers using the reduced number of features: DT, MLP, and SVM. Each trained base ML classifier performed its classification using testing data, and a new stacked dataset was created by combining predictions from these classifiers. This stacked dataset was then used to train a meta-classifier. A meta-classifier based on logistic regression (LR) with default parameters is used in this study to produce the final prediction. The training and evaluating the performance of all base classifiers and ensemble classifier was conducted on a machine equipped with Intel i9 CPU, and 16GB RAM.

Algorithm 1 Ensemble Classifier with Feature Selection
1: **Prepare Data:**
2:     Set *X* to all columns except the last from the dataset
3:     Set *y* to the last column of the dataset
4: **Feature Selection:**
5:     Initialize SelectKBest with χ^2 and $k \leftarrow 5$
6:     Fit SelectKBest to *X* and *y*
7:     Transform *X* to get *selected_features*
8: **Feature Scaling:**
9:     Initialize StandardScaler
10:     Fit and transform *selected_features* to *newx*
11: **Split Data:**
12:     Split *newx* and *y* into training and testing sets:
13:         *test_size* ← 0.25
14:         *random_state* ← 44
15: **Define Classifiers:**
16:     Initialize DecisionTreeClassifier parameters:
17:         *criterion* ← *entropy*
18:         *max_depth* ← 3
19:         *random_state* ← 44
20:     Initialize MLPClassifier parameters:
21:         *solver* ← *adam*
22:         *max_iter* ← 1200
23:         *batch_size* ← 32
24:         *hidden_layer_sizes* ← (100, 3)
25:     Initialize SVC with specified parameters
26:         *kernel* ← *linear*
27:         *max iter* ← 10000
28:         *C* ← 1.0
29:         *gama* ← *auto*
30: **Stacking Ensemble:**
31:     Initialize StackingClassifier:
32:         base estimators: DT, MLP, and SVC
33:         final estimator: LR
34:     Fit StackingClassifier to *X_train*, *y_train*
35: **Predict and Evaluate:**
36:     Predict *y_pred* from *X_test* using StackingClassifier
37:     Generate classification report for *y_test* and *y_pred*
38:     Generate confusion matrix for *y_test* and *y_pred*

What follows is a more detailed explanation of how the base classifiers and stacking ensemble classifier work:

1) Decision Tree (DT): is one of the most used versatile tools for classification and regression problems. DT classifier uses branches to establish decision rules and nodes to describe features. Consider it as a flowchart, with every step leading to a decision. To determine what features are most relevant in making these decisions, entropy is used. Entropy indicates the degree to which the data is disordered or unpredictable. Comparing the entropy before and after data is partitioned according to a specific feature allows to observe the improvement in data clarity. This clarity is commonly referred to as information gain. The larger the information gain, the more valuable a feature is for making precise predictions using DT [44][45]. To iteratively construct the decision tree, each node's information gain is calculated. After calculating the information gain of all features from the root node, the node feature with the highest information gain is chosen as the node feature. Then, the final decision tree model is obtained when no feature can be selected and all features have minimal information gains [46].

$$A(B) = -\sum_{i=1}^{|C|} a_i \, log_2(a_i) \qquad (2)$$

$$Gain(B,X) = A(B) - \sum_{b \epsilon Values(X)} \frac{|B_b|}{|B|} A(B_b) \qquad (3)$$

Where the current dataset is denoted as B, with X being a feature within it. $A(B)$ denotes the entropy of the target variable. In the target variable, a_i represents the probability of class i. The subset of B where X has a value of C is denoted as B_b. The number of instances in B_b is denoted by $|B_b|$ and $Values(X)$ represents the set of all potential feature X [44].

In this paper, to achieve a balance between model complexity and overfitting, we adjusted the DT parameters as follows: criterion = entropy, max_depth = 3, and random_state = 44. These adjustments were based on previous experiments and literature. This way, the tree won't be too deep and overfit the data, nor too shallow and underfit.

2) Support Vector Machine (SVM): a supervised binary classifier that helps classify data into different classes by finding the best separation hyperplane in a higher-dimensional space [8][45]. It reduces classification errors by maximising the hyperplane margin between classes. For tasks when the link between features and classes is complicated and can be either linear or non-linear, this is a great approach [38]. Using one of four kernels—linear, polyn, radial, or sigmoid—the SVM model selects the best hyperplane to categorize the data into classes [47]. Since the SVM is driven by the chosen kernel function and allows the transfer of a nonlinear input feature space to a linear feature space with higher dimensions [48], it is important to pick the right one. This work applied a linear kernel, shown in (4) [46].

$$f(x) = w * x + b \qquad (4)$$

where w is the weight vector, x is the feature vector, and b is the bias. Also, the linear kernel had a penalty parameter cost (C), which needed to be tuned. This parameter affects the classifier's accuracy and generalizability. As C decreased, the classifier's accuracy decreased, and its generalizability increased. As C increased, the classifier's accuracy increased, and its generalizability reduced. In our linear kernel, we set the cost (C) equal to 1 to maximize the classifier's efficiency.

In this paper, we set the SVM parameters to make the C, which balances margin and classification error, equal to 1. This value balances underfitting and overfitting to ensure model generalization to new data. A linear kernel can handle complex patterns in our dataset, therefore we set it to linear.

3) Multilayer Perceptron (MLP): a network of interconnected nodes, like how neurons in human brains work together. The interconnected nodes relate to one another and are arranged in several layers. Each layer takes in some information, does some calculations (like multiplying and adding), and then passes the result to the next layer. This process continues until we get the final output [45]. There is more than one hidden layer in the MLP structure to help solve difficult problems that can't be handled by just one hidden layer. The MLP is handy because it can be used for different tasks like classification, predicting values, and finding unusual patterns in data. The following equations show how each layer does its job, combining inputs, weights, biases, and activation functions to produce the final output [44].

$$h_1 = af(W_{m1}v + bv_1) \qquad (5)$$

$$h_2 = af(W_{m2}h_1 + bv_2) \qquad (6)$$

$$h_L = af(W_{mL}h_{L-1} + bv_L) \qquad (7)$$

$$q = ao(W_{m_{L+1}}h_L + bv_{L+1}) \qquad (8)$$

Where the output h_i is generated by the ith hidden layer. In the ith layer, bv_i represents the bias vector, W_{mi} represents the weight matrix and af represents the activation function. The input feature vector is denoted by v, and the number of hidden layers is denoted by L. For the output layer, ao represents the output activation function and q is the final output of the MLP [44].

In this paper, the batch size was set to 32 to balance convergence speed and gradient update stability. Smaller batch sizes update model weights more often, which may speed convergence but increase gradient estimation noise. However, bigger batch sizes provide smoother gradient updates but demand more memory and might hinder training. The solver adam was chosen for MLP training efficiency and resilience. This adaptive learning rate optimization algorithm computes individual learning rates for each feature, making it appropriate for larger datasets and faster convergence. To minimize overfitting and capture data complexity, hidden layer sizes were chosen as (100, 3). Multiple layers with decreasing units typically restrict the network as it goes, allowing the model to learn hierarchical data representations.

4) Stacking Ensemble: The prediction output of multiple separate classifiers is combined via a meta-classifier, a classifier that combines base classifier

predictions, in ensemble learning to provide a robust classifier and better predictive performance [49]-[51]. The stacking ensemble classifier is built with n distinct base prediction classifiers, C_1, C_2, C_3, ... C_n, all of which are trained and evaluated on the same dataset [52]-[54]. The general architecture for stacking is shown in Fig. 2. The test data is used to generate prediction outputs from the base classifiers, and train data is utilized to train the base classifiers. A new stacked dataset is created by aggregating the prediction outputs (P_1, P_2, P_3, ... P_n) of the base classifiers. Once the stacked dataset is created, it is used to train a meta-classifier. The meta-classifier corrects the basis classifiers' predictions and produces the final prediction. This raises the HD diagnostic system's general performance [40][55]. This stacking ensemble classifier is employed to raise the overall prediction accuracy of HD and lower the amount of error that happens with each base classifier [14][56].
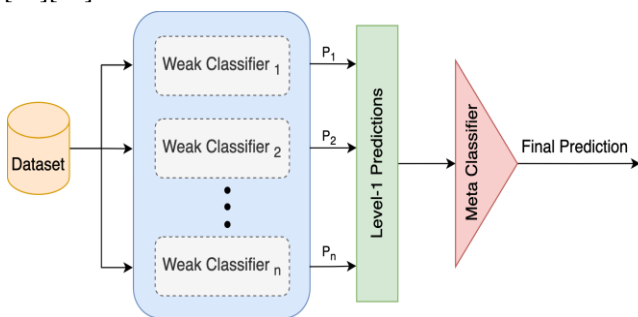

Fig. 2. Stacking ensemble architecture

C. Enhanced Ensemble Classifier With Feature Selection

In any prediction classifier, too many irrelevant features cause the classifier to overfit the training data and not generalize well. Overfitted prediction classifiers learn too much from training data and perform poorly on new data. Therefore, selecting significant features and removing redundant or noisy features improves the performance of the prediction classifier [57].

In this work, we eliminated bias from the training set by checking for noticeable differences in the distribution of features and classes. Then we prioritized features using the chi-square statistical method. This method determines the association between each input feature and predicted output classes. It reveals which features depend on the expected output classes. It ranks the features according to their chi-square test score. With higher chi-square scores, features are more dependent on predicted output classes [40]. Given a positive and negative set of class targets and a total of (t) instances, the chi-square test would compare the predicted count (P) with the observed count (O) as shown in (9). When the observed count significantly differs from the expected, the two features depend on each other. The predicted count (P) can be calculated using (10), assuming P_a, P_b, P_c, and P_d are the predicted values and a, b, c, and d are the observed values. Likewise, P_b, P_c, and P_d are computed.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - P_i)^2}{P_i} \qquad (9)$$

$$P_a = (a + b) \times \frac{(a + b)}{t} \qquad (10)$$

Although the feature selection method lowers noisy features and improves model efficiency, it could also unintentionally ignore interactions between features that would help to provide more accurate predictions. This limitation is recognized; however, future research should investigate ways to preserve important relationships such as applying feature selection strategies considering feature dependencies.

IV. RESULT AND DISCUSSION

In this section, we will review the evaluation results of the stacking ensemble classifier and the three base classifiers (multilayer perceptron (MLP), decision tree (DT), and support vector machine (SVM)) that were utilized to identify HD. We have used the Cleveland dataset to compare all three classifiers and the stacking ensemble classifier before and after applying the chi-square feature selection approach. The dataset was partitioned into train and test sets using a 75:25 split ratio. During all training processes, the test set was kept out and not used. This separate test set was set aside specifically for the performance evaluation of base classifiers and ensemble classifier to reduce overfitting risk resulting from feature selection combined with ensemble classifier.

In the experiments, two approaches were considered. In the first approach, following data standardization, the three base classifiers (MLP, DT, and SVM) were trained and evaluated separately using the thirteen input features. This was followed by constructing a stacking ensemble classifier by combining the prediction outputs of the three base classifiers. Similarly, the stacking ensemble classifier was trained using the training set and evaluated using the test set.

The results are shown in Table II. With an accuracy of 84.2%, MLP and SVM both achieved remarkable results among the base classifiers. On the other hand, the DT classifier achieved an accuracy of 81.8%. The Stacking Ensemble classifier achieved an accuracy of 85.5%, surpassing all base classifiers. This indicates an improvement of 1.54% over the best-performing base classifiers, MLP and SVM.

Using the recall, precision, and F1-score, we further evaluated the stacking ensemble classifier and the base classifiers performance. The classifier's recall evaluates its ability to detect all positive cases. High recall helps identify most HD patients, reducing the probability of misdiagnosis. The classifier's precision reveals its ability to discover true positives from expected positives. In HD prediction, more precision means a lower false positive rate, which is important when false positives cost patients unnecessary concern or medical tests. F1-score measures precision and recall. F1-score is an excellent indication of classifier performance when accuracy and recall are traded off [58].

Table II shows that while using MLP, SVM, and DT, the corresponding precision levels were 85.4%, 83.7%, and 82.9%. The stacking ensemble reached a precision of 85.7%. In terms of recall, MLP was able to reach 85.4%, SVM and DT 87.8%. Also, an 87.8% recall was attained by the stacking

ensemble. F1-scores of 85.4% for MLP, 85.7% for SVM, and 84.7% for DT were attained. However, the stacking ensemble managed to acquire a better F1-score of 86.7%.

TABLE II. BASE AND STACKING ENSEMBLE CLASSIFIERS PERFORMANCE USING TOTAL FEATURE SET

Method	Total features	Accuracy	Precision	Recall	F1-score
MLP	13	84.2%	85.4%	85.4%	85.4%
DT	13	81.8%	82.9%	87.8%	84.7%
SVM	13	84.2%	83.7%	87.8%	85.7%
Proposed Stacking Ensemble	13	85.5%	85.7%	87.8%	86.7%

Following data standardization, the thirteen input features were reduced to the five essential features in the second approach using the chi-square feature selection method. The feature selection process was conducted strictly on the training set, with no information from the test set influencing the selected features. The chi-square method ranked the thirteen features from highest to lowest according to their chi-square score, as shown in Fig. 3. It identified the best-ranking features ($k = 5$) for HD diagnosis. The top features selected by the chi-square method were maximum heart rate, Exercise-induced ST segment depression, number of fluoroscopically colored vital arteries, chest pain type, and exercise-induced angina. Each one of these top features helps the classifier diagnose heart disease more accurately. Eliminating any of these features could impair the model's ability to classify positive and negative cases, reducing its practicality. Since only five features were used for diagnosis instead of thirteen, the componental load was reduced by more than 50%.
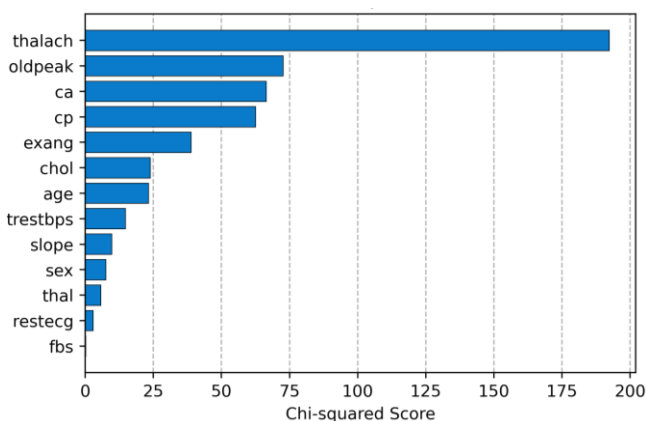

Fig. 3. Features importance using the chi-square method

After that, the reduced-feature dataset was used to train and assess the three base classifiers: MLP, DT, and SVM. The final prediction was generated by feeding the output of each base classifier into the stacking classifier. Table III displays the accuracy of the ensemble stacking classifier and each base classifier following feature reduction.

The results show that feature selection improved the performance of the MLP and SVM classifiers but had little effect on the DT classifier. The MLP classifier achieved an accuracy of 89.5%. The SVM classifier came next, with an accuracy of 88.2%. DT classifier did not improve accuracy; on the contrary, it hit 81.6% for accuracy. With an accuracy

of 90.8%, the stacked ensemble classifier surpassed base classifiers, with a substantial gain of 1.45% over the top-performing base classifier (MLP). Incorporating the suggested stacked ensemble classifier with the chi-square feature selection technique enhances HD diagnosis, as seen here. This improved performance offers clinicians a stronger and more accurate HD diagnostic tool. Since they lower the possibility of misdiagnosis and guarantee that patients get suitable treatments sooner, even such little improvements in prediction accuracy might result in better patient outcomes.

TABLE III. BASE AND STACKING ENSEMBLE CLASSIFIERS PERFORMANCE USING REDUCED FEATURE SET

Method	Total features	Accuracy	Precision	Recall	F1-score
MLP	5	89.5%	88.4%	92.7%	90.5%
DT	5	81.6%	86.5%	78.1%	82.1%
SVM	5	88.2%	86.4%	92.7%	89.4%
Proposed Stacking Ensemble	5	90.8%	90.5%	92.7%	91.6%

Table III demonstrates that the equivalent precision levels while employing MLP, SVM, and DT were 88.4%, 86.4%, and 86.5%, respectively. The precision achieved by the stacking ensemble was 90.5%. When it came to recall, MLP and SVM achieved 92.7%, but DT only managed 78.1%. Additionally, the stacking ensemble had a recall of 92.7%. It was possible to achieve F1-scores of 90.5% for MLP, 89.4% for SVM, and 82.1% for DT. Yet, with a stacking ensemble, a higher F1-score of 91.6% was achieved.

Fig. 4 shows the results of comparing the prediction accuracy of the two approaches. Based on the figure, out of the three base classifiers, the DT classifier performed the worst in both cases, while the MLP and SVM classifiers achieved the best results. Additionally, the stacking classifier performed better than the base classifiers in both cases. This figure shows how the feature selection method improved the accuracy of both the stacking ensemble classifier and the base classifiers.
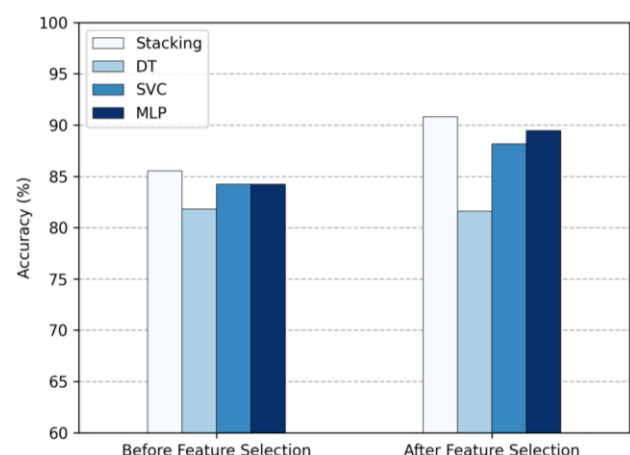

Fig. 4. Prediction accuracy

Using the confusion matrix, we further evaluated the stacking ensemble classifier and the base classifiers for their ability to identify HD. The confusion matrix summarizes the ratio of accurate and inaccurate predictions in a table-like format [59][60]. Four main components make up the

confusion matrix: (1) true positives (TP), which is a metric for the classifier's accuracy in identifying HD cases. A high TP count shows that the classifier is good at recognizing true cases, which is crucial for early intervention and treatment. (2) False positives (FP), which occur when the classifier misdiagnoses a patient with HD. Although a low FP rate is desired, false positives may alarm patients and lead to unnecessary medical procedures. (3) true negatives (TN), indicating the classifier correctly identifies HD-free patients. High TN counts indicate that the classifier can rule out disease. (4) false negatives (FN), which occur when the classifier misses true HD, misclassifying a patient as healthy. The delay of therapy may worsen patient circumstances, making this problematic.

Both Fig. 5 and Fig. 6 show the confusion matrices of the stacking ensemble classifier and the best-performing base classifier (MLP), respectively. Before applying the chi-square feature selection, Fig. 5 compares the confusion matrices of the stacking classifier with the top-performing base classifier (MLP). This shows that the MLP classifier can correctly identify 29 healthy individuals and 35 unhealthy ones. On the other hand, the stacking ensemble classifier correctly identified 29 healthy individuals and 36 unhealthy ones. Thus, it may be concluded that the ensemble stacking classifier is superior.
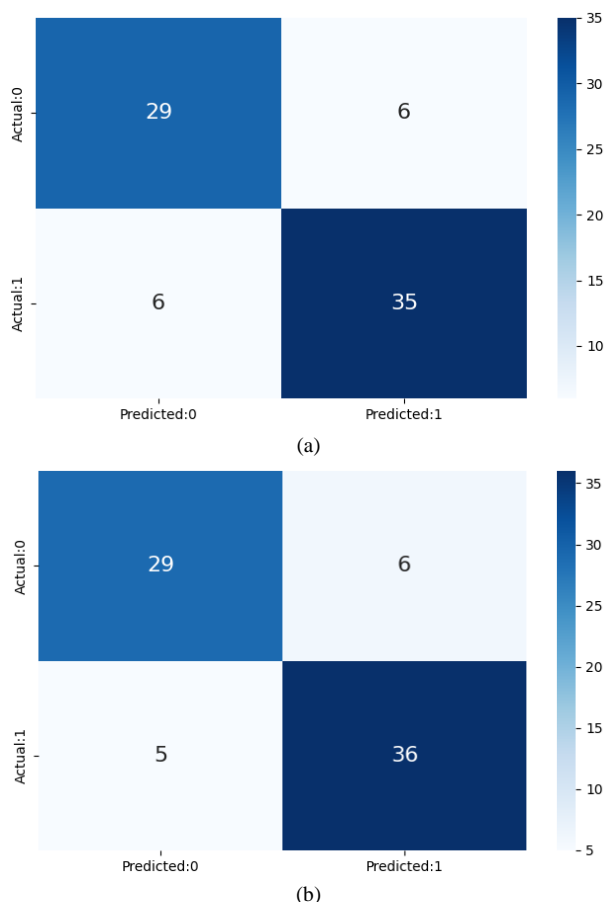

(a)

(b)

Fig. 5. Confusion matrix analysis before chi-square feature selection: a) MLP and b) Stacking ensemble

After applying the chi-square feature selection method, the confusion matrices of the stacking ensemble classifier and

the top-performing base classifier (MLP) are given in Fig. 6. By looking at the figure, we can see that the MLP classifier correctly detects 30 healthy persons and 38 unhealthy ones. In contrast, the stacking ensemble classifier correctly identified 31 healthy persons and 38 unhealthy ones. This proves that the ensemble stacking classifier is more efficient.
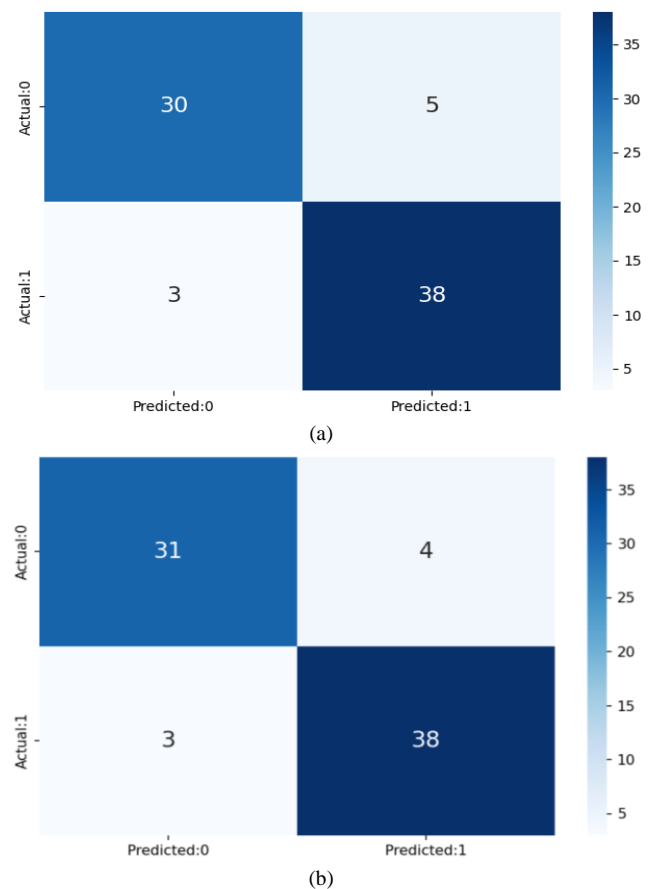

(a)

(b)

Fig. 6. Confusion matrix analysis following chi-square feature selection: a) MLP and b) Stacking ensemble

After using the chi-square feature selection method, the numbers in Fig. 5 and Fig. 6 show that the number of inaccurate predictions made by the MLP dropped from 12 to 8, and by the stacking ensemble, it dropped from 11 to 7.

In Table IV, we compare our improved ensemble stacking classifier for HD diagnosis to several state-of-the-art methods. We investigated the classifier's performance, type of feature selection approach, and number of features selected. With an accuracy of 90.8%, our suggested chi-square-based stacking ensemble classifier outperformed competing approaches.

In conclusion, the results of this work highlight the diagnosis of HD efficacy of the stacking ensemble classifier. Apart from showing better predictive accuracy than individual base classifiers, the suggested classifier underlined the advantages of including feature selection techniques into the ensemble strategy. These findings underline the need of using cutting-edge ML approaches, like stacking ensemble algorithms, to raise clinical diagnosis accuracy. By allowing more accurate and early diagnosis of HD, this strategy could greatly help to improve patient outcomes.

TABLE IV. A COMPARISON OF OUR PROPOSED ENHANCED HEART DISEASE PREDICTION METHOD WITH THE EXISTING STATE-OF-THE-ART

Work	Year	Feature Selection Method	Classifier	No. of Features selected	Accuracy
[4]	2019	extra tree classifier	bagging and majority voting ensemble (NB, KNN, ANN, LR, and DT)	-	87.78%
[36]	2019	brute force feature selection	Boosting, bagging, stacking, and voting ensemble (RF, NB, c4.5, bayes net, MLP, and PART)	9	85.48%
[34]	2019	extra tree classifier	majority vote ensemble (NB, SVM, and LR)	6	87.53%
[37]	2020	-	Hard voting ensemble (KNN, SVM, RF, GRU, and LSTM)	-	85.71%.
[39]	2020	CFS and PSO	stacking ensemble (GBM, RF, XGB)	7	85.71%
[38]	2021	-	majority vote ensemble (SVM, NB, and ANN)	-	87.05%
[25]	2022	Pearson, RFE, logistics, RF, light GBM, Chi-2	ensemble ML (CART, GBM, Adaboost, KNN, MLP, SGD, SVC, and NB)	7	87.65%
[35]	2022	RF	Voting Ensemble (DT, RF, XGB, KNN, DNN, and KDNN)	11	88.70%
Proposed	**2024**	**Chi-2**	**Stacking Ensemble ML (MLP, DT, and SVM)**	**5**	**90.8%**

V. CONCLUSION

This paper introduces an intelligent heart disease (HD) diagnosis method using a stacking ensemble approach and chi-square for feature selection. Multilayer perceptron (MLP), decision tree (DT), and support vector machine (SVM) were among the fundamental machine learning (ML) classifiers used by the stacking ensemble. The Cleveland heart data was the training and testing data for the base and ensemble classifiers. The classifiers' performance was assessed using accuracy, precision, recall, f1-score, and confusion matrix, among other criteria.

The base classifiers MLP, DT, and SVM demonstrated an accuracy of 84.2%, 81.8%, and 84.2%, respectively. By comparison, the stacking ensemble classifier exceeded the base classifiers with a higher accuracy of 85.5%. Five notable features were found by using the chi-square feature selection approach. To a remarkable 90.8%, this enhanced the performance and raised the accuracy of the stacking ensemble classifier even more.

Our enhanced stacking ensemble classifier uniquely combines DT, SVM, and MLP with chi-square feature selection, thereby improving prediction accuracy compared to existing methods. This integration of feature selection with a diverse ensemble of classifiers sets our method apart and exceeds state-of-the-art HD detection techniques.

The proposed method can be applied in healthcare environments to improve the accuracy and efficiency of heart disease diagnosis, therefore facilitating early identification and timely intervention. Incorporating this approach into clinical practice will enable healthcare professionals to use cutting-edge machine learning technologies to offer more accurate and quick diagnosis, therefore enhancing patient outcomes and resource allocation inside healthcare facilities.

Although the proposed work shows encouraging outcomes, its dependence on the Cleveland dataset could restrict generalizability to other populations and the computational complexity of the model could provide difficulties in resource-constrained environments. Addressing these limitations through examining other datasets to confirm the method's generalizability, testing more classifiers to improve prediction accuracy, and applying the method in a clinical setting to evaluate its practicality and integration into healthcare procedures. In conclusion, our HD diagnosis method establishes an entirely novel benchmark for accuracy and efficiency by integrating effective ensemble learning with efficient feature selection.

REFERENCES

[1] R. Rajendran and A. Karthi, "Heart disease prediction using entropy based feature engineering and ensembling of machine learning classifiers," *Expert Systems with Applications*, vol. 207, p. 117882, Nov. 2022, doi: 10.1016/j.eswa.2022.117882.

[2] S. Haseena, S. K. Priya, S. Saroja, R. Madavan, M. Muhibbullah, and U. Subramaniam, "Moth-Flame Optimization for Early Prediction of Heart Diseases," *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1–10, Sep. 2022, doi: 10.1155/2022/9178302.

[3] D. Yewale, S. P. Vijayaragavan, and V. K. Bairagi, "An Effective Heart Disease Prediction Framework based on Ensemble Techniques in Machine Learning," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 2, 2023, doi: 10.14569/ijacsa.2023.0140223.

[4] V. M. Deshmukh, "Heart Disease Prediction using Ensemble Methods," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 3, pp. 521–526, Sep. 2019, doi: 10.35940/ijrte.b2046.098319.

[5] S. Mohammad Ganie, P. Kanti Dutta Pramanik, M. Bashir Malik, A. Nayyar, and K. Sup Kwak, "An Improved Ensemble Learning Approach for Heart Disease Prediction Using Boosting Algorithms," *Computer Systems Science and Engineering*, vol. 46, no. 3, pp. 3993–4006, 2023, doi: 10.32604/csse.2023.035244.

[6] D. Asif, M. Bibi, M. S. Arif, and A. Mukheimer, "Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization," *Algorithms*, vol. 16, no. 6, p. 308, Jun. 2023, doi: 10.3390/a16060308.

[7] N. Chandrasekhar and S. Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," *Processes*, vol. 11, no. 4, p. 1210, Apr. 2023, doi: 10.3390/pr11041210.

[8] R. Rone Sarra, A. Musa Dinar, and M. Abed Mohammed, "Enhanced accuracy for heart disease prediction using artificial neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, p. 375, Jan. 2022, doi: 10.11591/ijeecs.v29.i1.pp375-383.

[9] S. Y. Hera, M. Amjad, and M. K. Saba, "Improving heart disease prediction using multi-tier ensemble model," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 11, no. 1, Oct. 2022, doi: 10.1007/s13721-022-00381-3.

[10] P. Bizopoulos and D. Koutsouris, "Deep Learning in Cardiology," *IEEE Reviews in Biomedical Engineering*, vol. 12, pp. 168–193, 2019, doi: 10.1109/rbme.2018.2885714.

[11] D. Hamid, S. S. Ullah, J. Iqbal, S. Hussain, Ch. A. ul Hassan, and F. Umar, "A Machine Learning in Binary and Multiclassification Results on Imbalanced Heart Disease Data Stream," *Journal of Sensors*, vol. 2022, pp. 1–13, Sep. 2022, doi: 10.1155/2022/8400622.

[12] A. E. Korial, I. I. Gorial, and A. J. Humaidi, "An Improved Ensemble-Based Cardiovascular Disease Detection System with Chi-Square

Feature Selection," *Computers*, vol. 13, no. 6, p. 126, May 2024, doi: 10.3390/computers13060126.

[13] S. Li, F. Li, S. Tang, and W. Xiong, "A Review of Computer-Aided Heart Sound Detection Techniques," *BioMed Research International*, vol. 2020, pp. 1–10, Jan. 2020, doi: 10.1155/2020/5846191.

[14] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Informatics in Medicine Unlocked*, vol. 26, p. 100655, 2021, doi: 10.1016/j.imu.2021.100655.

[15] A. E. Korial, "Brain Tumor Detection from MRI Images Using Artificial Intelligence," *International Journal on Engineering Applications (IREA)*, vol. 10, no. 3, p. 185, May 2022, doi: 10.15866/irea.v10i3.21213.

[16] M. M. Taye, "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions," *Computers*, vol. 12, no. 5, p. 91, Apr. 2023, doi: 10.3390/computers12050091.

[17] B. Abdualgalil, S. Abraham, and W. M. Ismael, "Early Diagnosis for Dengue Disease Prediction Using Efficient Machine Learning Techniques Based on Clinical Data," *Journal of Robotics and Control (JRC)*, vol. 3, no. 3, pp. 257–268, May 2022, doi: 10.18196/jrc.v3i3.14387.

[18] Z. Du *et al.*, "Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation," *JMIR Medical Informatics*, vol. 8, no. 7, p. e17257, Jul. 2020, doi: 10.2196/17257.

[19] S. A. Ebnou Abdem, J. Chenal, E. B. Diop, R. Azmi, M. Adraoui, and C. S. Tekouabou Koumetio, "Using Logistic Regression to Predict Access to Essential Services: Electricity and Internet in Nouakchott, Mauritania," *Sustainability*, vol. 15, no. 23, p. 16197, Nov. 2023, doi: 10.3390/su152316197.

[20] P. Gupta and D. D. Seth, "Improving the Prediction of Heart Disease Using Ensemble Learning and Feature Selection," *International Journal of Advances in Soft Computing and its Applications*, vol. 14, no. 2, pp. 37–40, Jul. 2022, doi: 10.15849/ijasca.220720.03.

[21] L. Miao and W. Wang, "Cardiovascular Disease Prediction Based on Soft Voting Ensemble Model," *Journal of Physics: Conference Series*, vol. 2504, no. 1, p. 012021, May 2023, doi: 10.1088/1742-6596/2504/1/012021.

[22] X.-Y. Gao, A. Amin Ali, H. Shaban Hassan, and E. M. Anwar, "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method," *Complexity*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/6663455.

[23] A. Tiwari, A. Chugh, and A. Sharma, "Ensemble framework for cardiovascular disease prediction," *Computers in Biology and Medicine*, vol. 146, p. 105624, Jul. 2022, doi: 10.1016/j.compbiomed.2022.105624.

[24] F. Aliyar Vellameeran and T. Brindha, "A new variant of deep belief network assisted with optimal feature selection for heart disease diagnosis using IoT wearable medical devices," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 25, no. 4, pp. 387–411, Jul. 2021, doi: 10.1080/10255842.2021.1955360.

[25] S. Diwan, G. S. Thakur, S. K. Sahu, M. Sahu, and N. K. Swamy, "Predicting Heart Diseases through Feature Selection and Ensemble Classifiers," *Journal of Physics: Conference Series*, vol. 2273, no. 1, p. 012027, May 2022, doi: 10.1088/1742-6596/2273/1/012027.

[26] S. J. Pasha and E. S. Mohamed, "Advanced hybrid ensemble gain ratio feature selection model using machine learning for enhanced disease risk prediction," *Informatics in Medicine Unlocked*, vol. 32, p. 101064, 2022, doi: 10.1016/j.imu.2022.101064.

[27] A. Kumar, K. U. Singh, and M. Kumar, "A Clinical Data Analysis Based Diagnostic Systems for Heart Disease Prediction Using Ensemble Method," *Big Data Mining and Analytics*, vol. 6, no. 4, pp. 513–525, Dec. 2023, doi: 10.26599/bdma.2022.9020052.

[28] G. A. Alshehri and H. M. Alharbi, "Prediction of Heart Disease using an Ensemble Learning Approach," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023, doi: 10.14569/ijacsa.2023.01408118.

[29] M. Ahmed and I. Husien, "Heart Disease Prediction Using Hybrid Machine Learning: A Brief Review," *Journal of Robotics and Control (JRC)*, vol. 5, no. 3, pp. 884-892, 2024, doi: 10.18196/jrc.v5i3.21606

[30] S. Abbas, G. A. Sampedro, A. Alsubai, A. Almadhor, and T. Kim, "An Efficient Stacked Ensemble Model for Heart Disease Detection and Classification," *Computers, Materials & Continua*, vol. 77, no. 1, pp. 665–680, 2023, doi: 10.32604/cmc.2023.041031.

[31] S. E. A. Ashri, M. M. El-Gayar, and E. M. El-Daydamony, "HDPF: Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm," *IEEE Access*, vol. 9, pp. 146797–146809, 2021, doi: 10.1109/access.2021.3122789.

[32] B. Srınıvasa Rao, "A New Ensenble Learning based Optimal Prediction Model for Cardiovascular Diseases," *E3S Web of Conferences*, vol. 309, p. 01007, 2021, doi: 10.1051/e3sconf/202130901007.

[33] K. Chandrashekar and A. T. Narayanreddy, "An Ensemble Feature Optimization for an Effective Heart Disease Prediction Model," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 2, pp. 517–525, Feb. 2023, doi: 10.22266/ijies2023.0430.42.

[34] B. Baranidharan, A. Pal, and P. Muruganandam, "Cardiovascular disease prediction based on ensemble technique enhanced using extra tree classifier for feature selection," *International Journal of Recent Technology and Engineering*, vol. 8, no. 3, pp. 3236-42, 2019, doi: 10.35940/ijrte.C5404.098319

[35] A. Alqahtani, S. Alsubai, M. Sha, L. Vilcekova, and T. Javed, "Cardiovascular Disease Detection using Ensemble Learning," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–9, Aug. 2022, doi: 10.1155/2022/5267498.

[36] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, p. 100203, 2019, doi: 10.1016/j.imu.2019.100203.

[37] I. Javid, A. Khalaf, and R. Ghazali, "Enhanced Accuracy of Heart Disease Prediction using Machine Learning and Recurrent Neural Networks Ensemble Majority Voting Method," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 3, 2020, doi: 10.14569/ijacsa.2020.0110369.

[38] N. Harika, S. R. Swamy, and Nilima, "Artificial Intelligence-Based Ensemble Model for Rapid Prediction of Heart Disease," *SN Computer Science*, vol. 2, no. 6, Aug. 2021, doi: 10.1007/s42979-021-00829-9.

[39] B. A. Tama, S. Im, and S. Lee, "Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble," *BioMed Research International*, vol. 2020, pp. 1–10, Apr. 2020, doi: 10.1155/2020/9816142.

[40] S. A. Ali *et al.*, "An Optimally Configured and Improved Deep Belief Network (OCI-DBN) Approach for Heart Disease Prediction Based on Ruzzo–Tompa and Stacked Genetic Algorithm," *IEEE Access*, vol. 8, pp. 65947–65958, 2020, doi: 10.1109/access.2020.2985646.

[41] J. Vijayashree and H. Parveen Sultana, "Heart disease classification using hybridized Ruzzo-Tompa memetic based deep trained Neocognitron neural network," *Health and Technology*, vol. 10, no. 1, pp. 207–216, Jan. 2019, doi: 10.1007/s12553-018-00292-2.

[42] R. R. Sarra, A. M. Dinar, M. A. Mohammed, M. K. A. Ghani, and M. A. Albahar, "A Robust Framework for Data Generative and Heart Disease Prediction Based on Efficient Deep Learning Models," *Diagnostics*, vol. 12, no. 12, p. 2899, Nov. 2022, doi: 10.3390/diagnostics12122899.

[43] T. K. Sajja and H. K. Kalluri, "A Deep Learning Method for Prediction of Cardiovascular Disease Using Convolutional Neural Network," *Revue d'Intelligence Artificielle*, vol. 34, no. 5, pp. 601–606, Nov. 2020, doi: 10.18280/ria.340510.

[44] R. Kapila, T. Ragunathan, S. Saleti, T. J. Lakshmi, and M. W. Ahmad, "Heart Disease Prediction Using Novel Quine McCluskey Binary Classifier (QMBC)," *IEEE Access*, vol. 11, pp. 64324–64347, 2023, doi: 10.1109/access.2023.3289584.

[45] S. Bashir, A. A. Almazroi, S. Ashfaq, A. A. Almazroi, and F. H. Khan, "A Knowledge-Based Clinical Decision Support System Utilizing an Intelligent Ensemble Voting Scheme for Improved Cardiovascular Disease Prediction," *IEEE Access*, vol. 9, pp. 130805–130822, 2021, doi: 10.1109/access.2021.3110604.

[46] X. Wenxin, "Heart Disease Prediction Model Based on Model Ensemble," *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 195-199, May 2020, doi: 10.1109/icaibd49809.2020.9137483.

[47] H. Elwahsh, E. El-shafeiy, S. Alanazi, and M. A. Tawfeek, "A new smart healthcare framework for real-time heart disease detection based on deep and machine learning," *PeerJ Computer Science*, vol. 7, p. e646, Jul. 2021, doi: 10.7717/peerj-cs.646.

[48] S. Richa and S. Shailendra Narayan, "Towards Accurate Heart Disease Prediction System: An Enhanced Machine Learning Approach," *International Journal of Performability Engineering*, vol. 18, no. 2, p. 136, 2022, doi: 10.23940/ijpe.22.02.p8.136148.

[49] S. P. Patro, N. Padhy, and R. D. Sah, "An improved ensemble learning approach for the prediction of cardiovascular disease using majority voting prediction," *International Journal of Modelling, Identification and Control*, vol. 41, no. 1/2, p. 68, 2022, doi: 10.1504/ijmic.2022.127098.

[50] J. M.-T. Wu et al., "Applying an ensemble convolutional neural network with Savitzky–Golay filter to construct a phonocardiogram prediction model," *Applied Soft Computing*, vol. 78, pp. 29–40, May 2019, doi: 10.1016/j.asoc.2019.01.019.

[51] K. Dissanyake and M. G. M. Johar, "Two-level boosting classifiers ensemble based on feature selection for heart disease prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 1, p. 381, Oct. 2023, doi: 10.11591/ijeecs.v32.i1.pp381-391.

[52] N. V. MahaLakshmi and R. K. Rout, "Effective heart disease prediction using improved particle swarm optimization algorithm and ensemble classification technique," *Soft Computing*, vol. 27, no. 15, pp. 11027–11040, May 2023, doi: 10.1007/s00500-023-08388-2.

[53] V. Jain and K. L. Kashyap, "Multilayer hybrid ensemble machine learning model for analysis of Covid-19 vaccine sentiments," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 5, pp. 6307–6319, Sep. 2022, doi: 10.3233/jifs-220279.

[54] D. Mohapatra, S. K. Bhoi, C. Mallick, K. K. Jena, and S. Mishra, "Distribution preserving train-test split directed ensemble classifier for heart disease prediction," *International Journal of Information Technology*, vol. 14, no. 4, pp. 1763–1769, Jan. 2022, doi: 10.1007/s41870-022-00868-2.

[55] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics in Medicine Unlocked*, vol. 20, p. 100402, 2020, doi: 10.1016/j.imu.2020.100402.

[56] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/access.2019.2923707.

[57] L. Ali, C. Zhu, M. Zhou, and Y. Liu, "Early diagnosis of Parkinson's disease from multiple voice recordings by simultaneous sample and feature selection," *Expert Systems with Applications*, vol. 137, pp. 22–28, Dec. 2019, doi: 10.1016/j.eswa.2019.06.052.

[58] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," *IEEE Access*, vol. 9, pp. 106575–106588, 2021, doi: 10.1109/access.2021.3098688.

[59] E. K. Wang, X. Zhang, and L. Pan, "Automatic Classification of CAD ECG Signals With SDAE and Bidirectional Long Short-Term Network," *IEEE Access*, vol. 7, pp. 182873–182880, 2019, doi: 10.1109/access.2019.2936525.

[60] J. Ivan and S. Y. Prasetyo, "Heart Disease Prediction Using Ensemble Model and Hyperparameter Optimization," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 8s, pp. 290–295, Aug. 2023, doi: 10.17762/ijritcc.v11i8s.7208.