MULTIDISCIPLINARY SCIENCE JOURNAL

RESEARCH ARTICLE
Published Online: August 31, 2023
https://doi.org/10.31893/multiscience.2024028

Heart Failure prediction on diversified datasets to improve generalizability using 2-Level Stacking

Madhuri Dubey[a] | **Jitendra Tembhurne**[a] | **Richa Makhijani**[a]

[a]Indian Institute of Information Technology, Nagpur 441108, Maharashtra, India.

Abstract Heart disease is a leading cause of death worldwide, and early detection is crucial for improving patient outcomes. Machine learning classifiers have shown promise in predicting heart disease using patient-specific factors such as demographic information, medical history, and lifestyle habits. In this paper, we aimed to evaluate the performance of machine learning classifiers in predicting heart disease using a combination of 5 different datasets such as Cleveland, Hungarian, Switzerland, Long Beach VA, and StatLog (Heart) Datasets available on IEEE data port. The two significant challenges are addressed in this work: 1) predicting heart failure using machine-learning models without eliminating any clinical features, which increases the risk of overfitting and can result in poor performance metrics, and 2) we propose a model that will provide remarkable accuracy regardless of the type of data, to offer model generalizability. Machine-learning algorithms such as logistic regression, decision trees, random forests, support vector machine, extreme gradient boosting, extra-tree, and K-nearest neighbor are applied for heart disease prediction. In addition, ensemble approaches such as majority voting, boosting, bagging, and stacking are employed. The performance of the classifiers is evaluated using performance metrics such as accuracy, precision, recall, and F1-score. The results showed that the ensemble approach of stacking outperformed individual models, with an accuracy of 93.67%.

Keywords: ensemble, heart disease prediction, machine learning, stacking

1. Introduction

Millions of individuals around the world are affected by heart disease, which imposes a heavy strain on healthcare systems. According to the World Health Organization (WHO) heart disease causes 16% of all deaths worldwide (WHO 2021). The heart disease diagnosis, treatment, and long-term medication will add more financial burden to the patient and family, which leads to psychological imbalance. In this technological era, many health professionals, researchers, and data scientists worked together to identify more accurate, reliable algorithms for estimating the risk of heart disease. Centers for disease control and prevention (CDC) suggested that heart disease can be prevented by embracing lifestyle habits like regular exercise, a balanced diet, stress management, and abstaining from alcohol and tobacco (Prevent Heart Disease cdc.gov 2023). A sufficient volume of data is now available from several medical institutes, trials, and researchers for studying and developing accurate methodologies for early disease detection and its risk factors. Machine learning (ML), deep learning (DL), data mining, and the internet of medical things (IoMT) have contributed significantly to predictive analytics. Clinical data integrates various information such as demographics, medical background, lifestyle details, lab test (ECG/EKG) results, and biomarkers, which are used to develop accurate mathematical, statistical, and analytical models by investigating the patterns, and relationships of samples with one another (Ahmed et al 2020). Machine learning is a promising technology that has made progress in the predictive analytics of clinical data in the healthcare industry (Javaid et al 2022); nevertheless, it still faces significant challenges related to data security, privacy, and generalization, which can be resolved by federated learning (Dang et al 2022), effective feature learning, wise model selection, and thorough model training.

Latha and Jeeva (2019) grouped the clinical features into some feature sets and assessed how well various classifiers performed on feature sets. Further, majority voting with Naive Bayes (NB), Bayes Net, and Multilayer Perceptron (MLP) delivered 85.48% accuracy on selected features, outperforming any other ensemble strategy. A scaled conjugate gradient back propagation algorithm of artificial neural network (ANN) was implemented by Paul et al (2022) using MATLAB environment; it provided varying accuracy of 88.47% with the number of hidden neurons on the "IEEE comprehensive heart disease" database (IEEE CHD) but required more computation time and numerous iterations. Data mining is also underlined by its success in prediction analysis. In Reddy et al (2019), the two strategies as feature selection and ratio identification of train-test splits are highlighted. In addition, while assessing the effect of these strategies on model accuracy, authors demonstrated that the random forest with an 80:20 split provided a higher accuracy score than other splits with significant features of the IEEE CHD database. Khennou et al (2019) integrated three datasets from the UCI repository (Cleveland, Hungarian, and Switzerland) and used the SVM classifier to obtain an accuracy of 87%. Authors also employed the KNN

Received: April 21, 2023 | Accepted: July 31, 2023

classification algorithm for missing value imputation. In Shah et al (2020), the KNN classification method on a small Cleveland dataset after data preprocessing procedures such as data cleaning, transformation, integration, and reduction recorded 90.789% accuracy using data mining technology. Doppala et al (2022) suggested a weighted majority voting ensemble approach that involved classifiers like NB, RF, SVM, and gradient boosting which reported the good accuracy of 93.39% with a reduced feature count of 9 on IEEE CHD. Because of the limited feature count, Miriyala et al (2021) achieved the maximum validation accuracy of 93.26% using the Extra tree classifier. A feature score greater than 0.5 was chosen as an important feature for prediction under the absolute correlation approach. Some researcher compared the performance of ML and DL models on the heart disease dataset for clinical data. Wu et al (2021) used multiple datasets with variable sample sizes, feature counts, and domains. Additionally, they used XGBoost as a machine learning classifier and Multilayer Perceptron (MLP) as a neural network model; using the IEEE CHD database, XGBoost performed better than the DL model with a 7% increase in accuracy. Using a back propagation algorithm with data mining association rules, Mohan et al (2019) suggested a "Hybrid Random Forest with Linear Model" (HRFLM) that classified features with 88.4% accuracy on the Cleveland dataset. Vellameeran and Brindha (2022) proposed a novel feature selection technique that combined GWO with PSO (Particle Swarm Optimization). Additionally, they used a Deep Belief Network (DBN) and tuned the number of hidden neurons and activation function to create a good optimized model, recording 83.8% accuracy on the IEEE CHD database. Table 1 highlights the limitations of previous research work, leading to the need for a novel approach.

Table 1 Review of literature of previous research work.

Ref. Studies	Approach used	Features Count	Dataset Used	Accuracy	Challenges
(Paul et al 2022)	Scaled Conjugate Gradient Back Propagation of artificial neural networks (ANN)	11	IEEE-CHD	88.47%	Requires more computation time and number of iterations
(Khennou et al 2019)	Support vector machine and Naïve Bayes ML classifiers	13	UCI-ML Repository (Cleveland, Hungarian and Switzerland)	87% and 86%	Data preprocessing and feature engineering were not done
(Doppala et al 2022)	Weighted Majority voting ensemble model	9	IEEE-CHD	93.39%	High accuracy is achieved but with reduced feature count
(Vellameeran and Brindha 2022)	PS-GWO-DBN	--	IEEE CHD	83.83%	Features were optimally selected from the dataset could lead to loss of clinical information also the sensitivity is less
(Miriyala et al 2021)	Light Gradient Boosting machine	5	IEEE-CHD	93.26%	Number of Clinical features considered for training the model were very less
(Deb et al.2022)	Random forest	7	IEEE CHD	93.10%	1.47% decreased in accuracy when full features used for training
(Mohan et al 2019)	HRFLM uses Random Forest (RF) and Linear Method (LM)	13	UCI-ML Repository (Cleveland)	88.4%	Features required for training were selected based on the error rate computed by classifiers

1.1. *Motivation*

The motivations to develop the generalized heart disease prediction model on the diversified dataset with all clinical features are listed below:

- Many studies solely used traditional classification techniques for model training, limiting the ability to recognize the risk factor and enable early intervention.
- To obtain high accuracy and minimal complexity, the researcher would lower the number of features, but they would also skip some essential components, so the model with limited features can lead to inaccurate predictions and potentially missed diagnoses.
- Several models performed better on one dataset yet failed to maintain the performance on other dimensions (called poor generalizability), highlighting the need to develop an accurate heart disease generalized prediction model with all clinical variables for improving patient outcomes and deepening the understanding of the disease.

1.2. *Challenges*

- The dataset contains a wide range of variables, such as nominal (categorical) and numerical. The categorical attributes include gender, type of chest pain, ECG slope, and many others. Furthermore, the numerical attribute like age, cholesterol level, blood pressure, blood sugar, heart rate, and others. To develop a model that includes all categories of features with improved performance is a big challenge.
- Every patient is unique in their medical background, way of life, stress level, and demographics. As a result, there will be a different association between clinical parameters and the outcome. Henceforth, we desired a model that considers all clinical variables and yields precise results without leading to overfitting.

- Build a model that performs well on every heart disease dataset (balanced/unbalanced, categorical/numerical, or mixed) without compromising any clinical attributes to address the issue of poor generalizability.

1.3. Contributions

- To study and implement effective feature engineering, data preparation, analysis, model training and hyperparameter optimization for improving the performance metrics on diverse clinical comprehensive data of five datasets, which is available on IEEE data port (Siddhartha 2020).
- To propose the generalized 2-level stacking model with effective base model and suitable meta-learner design, that includes seven different machine learning fine-tuned classifiers.
- To compare the performance of proposed model on different datasets with diverse attributes (i.e. numeric, categorical, and mixed) for validating the generalizability of proposed work.

Several researchers performed the heart disease prediction with various classifiers and received variable accuracy with limited feature count. The potential requirement for heart disease prediction is to treat with different clinical features, data source, variable dimensions or data types. A stacked generalization approach can potentially improve the accuracy of heart disease prediction using diverse clinical data. Stacked generalization is an ensemble learning technique that combines multiple models to improve predictive performance. It is particularly effective when using diverse sources of data or when there are various types of features that may require different modeling techniques for capturing different aspects of complex relationships within the diverse clinical data. The key aspect of proposed model is choice of base model, and the design of the meta-learner. Proper feature engineering, data preprocessing, and hyperparameter tuning is employed to build an effective stacked model.

This paper proposes, the generalized 2-level stacking model, which includes seven different machine learning classifiers with properly tuned hyperparameters using grid search, hyperopt sklearn, randomized, and manual validation approaches. The study included five datasets, with the majority of the features being categorical, and utilized one-hot encoding to capture the complex relationship between the variables. The proposed model is tested on a different dimension of the datasets to validate the performance.

2. Materials and Methods

Predicting heart failure using the machine learning models without removing any clinical attributes can be challenging task as it increases the risk of overfitting and can lead to poor performance metrics. However, some strategies can be used to improve the model performance such as data pre-processing, feature engineering, model selection, hyperparameter tuning, model training and validation/testing.

2.1. Dataset Description

The dataset used in this study is obtained from the IEEE data port namely, "IEEE Comprehensive Heart Disease Dataset" (Siddhartha 2020). It is a collection of 5 datasets: Cleveland, Hungarian, Switzerland, Long Beach VA, and StatLog (Heart). It includes demographic information, medical history, lifestyle information, and lab test results that can be used to predict the presence of heart disease. As shown in Table 2**,** the dataset contains 1190 instances of the patient with 11 clinical features and 1 target attribute.

Table 2 The Dataset's description.

SN	Attribute	Variable Name	Measured unit	Data type
1	age	Age	years	Numeric
2	sex	Sex	1, 0	Categorical
3	chest pain type	chest pain type	1,2,3,4	Categorical
4	resting blood pressure	resting bp s	mm Hg	Numeric
5	serum cholesterol	cholesterol	mg/dl	Numeric
6	fasting blood sugar	fasting blood sugar	1,0 > 120 mg/dl	Categorical
7	resting electrocardiogram results	resting ecg	0,1,2	Categorical
8	maximum heart rate achieved	max heart rate	71–202	Numeric
9	exercise induced angina	exercise angina	0,1	Categorical
10	old peak =ST	oldpeak	depression	Numeric
11	slope of the peak exercise ST segment	ST slope	0,1,2	Categorical
12	Class	target	0,1	Categorical

Table 3 describes various categorical clinical variables with critical levels in medical terminology. The chest pain is classified as typical angina, caused by chest discomfort, triggered by exertion or stress, and eased by rest or medication

(UpToDate 2022). Type 2 atypical angina most likely did not involve the type 1 trait but shows other types of symptoms, type 3 indicates non-cardiac chest pain, and type 4 suggests silent (asymptomatic) myocardial ischemia (SMI), which includes the lack of chest discomfort but leading to SMI that may be evaluated in ST-segment ECG testing (AlBadri et al 2017). The fasting blood sugar, sex, exercise-induced angina, and class were binary attributes.

Table 3 Description of categorical attributes.

Attribute	Description
Sex	1 = male, 0= female;
Chest Pain Type	1: typical angina, 2: atypical angina, 3: non-angina pain, and 4: asymptomatic
Fasting Bloodsugar	(Fasting blood sugar > 120 mg/dl) 1 = true; 0 = false
Resting electrocardiogramresults	0: normal, 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), and 2: showing probable or definite left ventricularhypertrophy by Estes' criteria
Exercise induced angina	1 = yes; 0 = no
slope of peak exercise ST segment	1: up sloping, 2: flat, and 3: down sloping
class	Target: Heart disease; 1 = Present, 0 = Absent

2.2. Data Preprocessing

This dataset is one of the biggest heart disease datasets available in the public domain for research purposes. Enormous volumes of information on heart disease from various sources needs to be pre-processed, and certain factors required to be taken into account to make sure the data is clean, consistent, and prepared for model training.

1. Handling missing values: impute or remove records with missing values. Impute the missing values in the numeric column using the mean, mode, median, forward fill, or interpolation techniques. Additionally, substituting an arbitrary value for the missing variable. For categorical columns, impute with the most frequent value.
2. Handling data imbalance: The number of positive and negative cases is significantly different. Undersampling and oversampling are two strategies that can be used to process data balancing. Oversampling involves cloning minority records to match majority class frequency, whereas undersampling involves removing the records from the majority class to balance it with the minority class. Oversampling is preferred for clinical data since every piece of clinical information is crucial, and it would be inappropriate to remove any patient information. The oversampling also involves the creation of some synthetic samples for unbiased training purposes. Additionally, various oversampling methods are used, including SMOTE (synthetic minority oversampling technique), ADASYN (adaptive synthetic sampling), random oversampling, SMOTENC, and many others.
3. Handling outliers: Remove or modify extreme values of features when handling outliers. There are several ways to eliminate outliers from the dataset to improve the effectiveness of the predictive analysis. Outliers can be visualized using a box or scatter plot. Interquartile range detection (IQR) and Z-score approaches can be used to deal with the outliers.
4. Encoding categorical variables: Convert categorical data into numerical data. The Machine learning model only processed the numerical values hence label, ordinal, target and one-hot encoding techniques can be applied to processed categorical data.
5. Feature scaling: normalize the data to a specific range of 0 to 1 using a normalization or standardization approach

The categorical variables included in the dataset utilized in this study are listed in Table 3 and there are no missing values or duplicate records. The dataset containing total 1190 records of patients and out of these, 561 patients are without heart disease and 629 patients with heart illness. The class ratio is 0.53 and the Figure 1 depicts that the data is balanced.

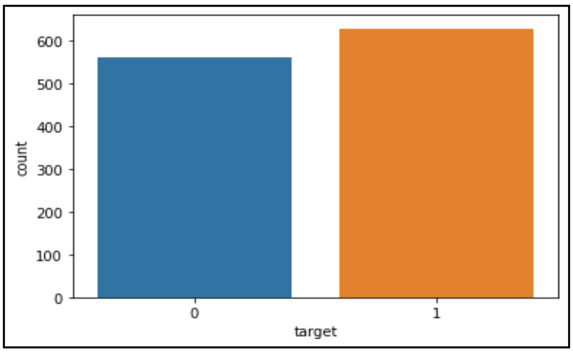


Figure 1 Target class.

The outliers are detected at columns like “resting bp s”, “oldpeak”, and “cholesterol” which can be removed using IQR (Inter quartile range) (Beunza et al 2019). The IQR is evaluates as shown below;

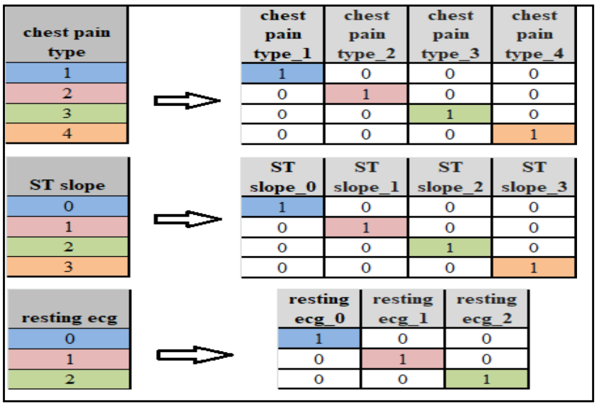
IQR is the range between the 1st and the 3rd quartiles namely Q1 and Q3; IQR = *Q3 – Q1*.

The data point lies in following case are considered as outliers;

Lower bound = *Q1 – 1.5 IQR* or upper bound = *Q3 + 1.5 IQR*

The categorical variables need to be encoded into numerical data before model training. The one-hot encoding, ordinal encoding, numeric encoding, and label encoding are several methods for transforming categorical data into numeric values.

One-hot encoding (Budholiya et al 2022) transforms categorical variables to numerical values as shown Figure 2. It provided more information about each category of attributes and their associated relations. It can cause the problem of overfitting as it increases the dimension of the data frame by increasing the number of columns for each category. Proper hyperparameter tuning and feature engineering could resolve the overfitting issue.



chest pain type
1
2
3
4

chest pain type_1	chest pain type_2	chest pain type_3	chest pain type_4
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

ST slope
0
1
2
3

ST slope_0	ST slope_1	ST slope_2	ST slope_3
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

resting ecg
0
1
2

resting ecg_0	resting ecg_1	resting ecg_2
1	0	0
0	1	0
0	0	1

Figure 2 One-hot encoding of categorical attributes.

2.3. Feature Engineering

The dataset features provide fine-grained information and play an essential role in analysing the target variable. Feature engineering entails developing new features from existing attributes, such as aggregating and summarizing various variables, determining the connection of variables with the predictor, and producing feature importance to prioritize the relevant clinical parameters. After encoding categorical features, every category of clinical attribute provided the actual correlation score with the predictor class “target” as mentioned in table. From Table 4, it is observed that ST slope_1 and 2, chest pain type_4, Exercise induced angina, old peak ST value, and max heart rate shows the highest correlation with the target class.

Table 4 Correlation score of all attributes with respect to “target” attribute.

Attributes	Score
target	**1**
ST slope_1	0.57827
ST slope_2	0.530989
chest pain type_4	0.518223
exercise angina	0.518085
oldpeak	0.488353
max heart rate	0.402171
chest pain type_2	0.350369
sex	0.319549
age	0.285016
chest pain type_3	0.235154
resting ecg_0	0.14939
resting bp s	0.147689
cholesterol	0.116761
fasting blood sugar	0.116026
ST slope_3	0.10395
resting ecg_2	0.097537

2.4. Model Selection

Heart disease prediction can be addressed using a variety of machine learning algorithms. However, choosing the model that will best learn the range of features (nominal and numeric) is a critical task. To achieve the best results, we chose seven models and used a variety of hyperparameter tuning strategies, including randomized search, grid search, hyperopt, and manual as described in the Table 5.

Table 5 Model and their tuned hyperparameter.

SN	Model Name	Description	Hyper parameters
1	Decision tree (DT)	A classification algorithm uses a tree structure to classify categorical and numerical data. Entropy assesses the quality of each split in the data, and tree structures offer the maximized information gain at each split. Grid Search compares the entropy criteria performance, shown in the Figure 3.	{'criterion': 'entropy', 'max_depth': 15, 'min_samples_split': 2}
2	Random Forest (RF)	It involved multiple decision trees, which were ensemble to predict the outcome after rigorous training. The best parameters are retrieved using the RandomizedSearchCV technique. For categorical data, the entropy or impurity of data is calculated based on the number of instances of each category in the set.	{'n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': None}
3	Logistic Regression (LR)	A statistical technique describes data and the association of one dependent variable with one or more independent variables. The independent variables can be nominal, ordinal, or interval in nature.	solver='liblinear'
4	K-Nearest Neighbour (KNN)	KNN algorithm stores the dataset during the training phase and subsequently classifies new data into a category that resembles newly acquired data. For categorical data, the optimum parameters were Manhattan as a metric and distance as a weight. Finding the best value for *k* as shown in Figure 4 concerning the metric is a real challenge.	Best leaf_size: 1 Best p: 1 Best n_neighbors: 29 Best weights: distance Best metric: manhattan
5	Extra Tree	It is the collection of multiple decision trees generated using random subsets of features and thresholds. Compared to Random forest, it performed computations more quickly because of its split selection method. The parameters were tuned using RandomizedSearchCV.	{'n_estimators': 100, min_samples_split': 2, 'min_samples_leaf': 1, max_depth': 52, 'bootstrap': False}

6	Support Vector Machine (SVM)	SVM is a supervised machine learning technique used for classification or regression tasks. In SVM, data was transformed using kernel, and based on these modifications, it determines the optimal boundary. Here, Radial basis function (rbf) selected for data transformation task. The hyperparameter were tuned using GridSearchCV	{'C': 100, 'gamma': 0.1, 'kernel': 'rbf'}
			SVC(C=100, gamma=0.1, random_state=42)

7	Extreme Gradient Boosting (XGB)	It is optimized tree ensemble gradient boosting algorithm. It involves parallelized tree building, regularization for overfitting, a depth-first approach for tree pruning, and optimized use of hardware resources. The hyperparameter are tuned using hyperopt technique.	{ 'colsample_bytree': 0.539697348025, 'gamma': 2.76635165, 'max_depth': 14.0, min_child_weight': 2.0, 'reg_alpha': 41.0

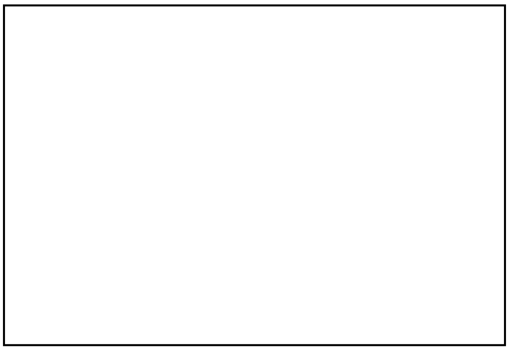


Figure 3 Decision tree entropy on maximum depth.

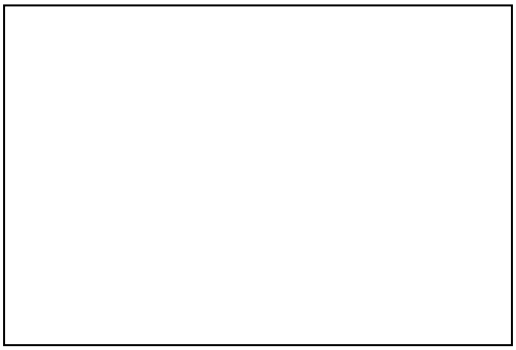


Figure 4 The value of "K" for KNN.

Each model in machine learning has its significance in terms of strengths and weaknesses; when these models with excellent learning strategies combine to give the best-trained model, that model is the ensemble model. There are various ensemble models as listed in Table 6.

Table 6 Ensemble model and their significance.

SN	Ensemble	Description
1	Majority Voting	Multiple models are trained on various subsets of the training data before being combined to make predictions on the entire dataset in majority voting. The main benefits of majority voting are lowering the risk of overfitting by integrating the predictions of various models to get a more reliable result and minimizing bias. The majority voting includes hard, weighted, and soft voting methods.
2	Adaptive Boosting	AdaBoost (a.k.a Adaptive Boosting) optimizes the performance metrics of a machine-learning model by iteratively training weak classifiers on the data and reweighting training samples based on their classification errors. The parameters were tuned using GridSearchCV.{'learning_rate': 0.1, 'n_estimators': 50}
3	Bagging	Bootstrap aggregating (a.k.a bagging) is used to create several bootstrap samples from training data and train a classifier using each sample. Also, it is useful when the individual classifiers are relatively simple, but the ensemble of classifiers can capture complex relationships between inputs and output.

2.5. *Proposed Method*

Stacking or stacked generalization is the proper integration of diversity derived from several classifiers and optimization of the best model. It solves the generalization problem by combining multiple models (linear and non-linear) and training the meta-model based on the prediction score of base models. Instead of dividing the train data into blocks for each model, we gave them all the clinical features, as shown in the Figure 5 to enable better learning. The raw data aggregates the five diverse datasets, involving 7 categorical and 4 numerical variables.

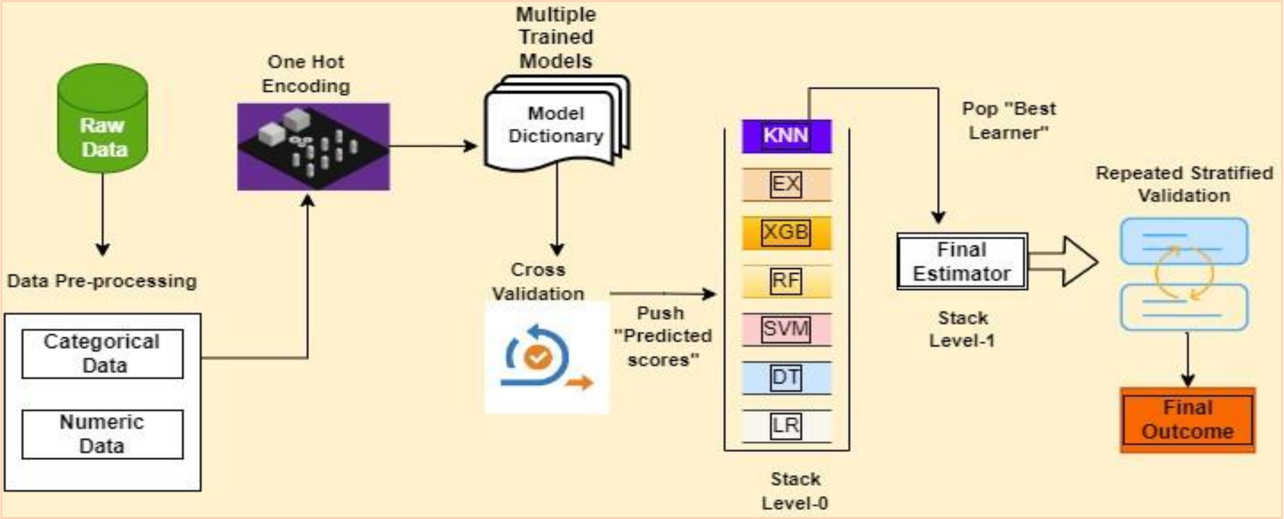


Figure 5 The proposed approach: 2 - level Stacking.

The proposed approach involved data pre-processing before model training; one-hot encoding (OHE) to transform categorical data into numerical. Further, outlier elimination, locating missing or null values, and determining data imbalance, as described in section 0. Seven classifiers are employed separately on the datasets after tuning the hyperparameter (mentioned in section 0), and all the trained models are pushed into the stack with the cross-validated predicted score at level-0. A model with the best-predicted score is popped out as a Meta learner and placed at level-1 as the final estimator. After repeated stratified validation, the predicted output is achieved.

3. Results and Discussion

This section highlights the statistical analysis of various classifiers, ensembles, and the proposed approach of stacking on IEEE comprehensive heart disease dataset. Also, we present the performance of stacking approach on variety of databases such as small set of database “Cleveland” (UCI machine learning Repository: Heart disease dataset 1988), imbalanced database “Framingham heart study” (Framingham Heart Study 2021; Framingham Heart Study-Cohort (FHS-Cohort) 2023), and comprehensive database of “IEEE heart disease” (Siddhartha 2020).

3.1. *Performance metrics*

- *Confusion matrix*: The matrix representation in which a row indicates the ground truth and columns show the model prediction. It contains four terms as shown:

i. True Positive (TP) - Predicted “Yes” and the actual outcome is also “Yes”
ii. True Negative (TN) - Predicted “No” and the actual outcome is also “No”
iii. False Positive (FP) - Predicted “Yes” but the actual outcome is “No”
iv. False Negative (FN) - Predicted “No” but the actual outcome is “Yes”

- *Precision*: The number of true positive to the total positives predicted by the model. It is also called as positive predicted value (PPV) represented in the equation as;

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \qquad (1)$$

- *Recall / Sensitivity*: The ratio of the true positives to the total number of actual positive values.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \qquad (2)$$

- *F1 – score:* The weighted average of precision and recall which balances both the metrics as shown in equation;

$$\boldsymbol{F1\ score} = \frac{\boldsymbol{2*Precision*Recall}}{\boldsymbol{Precision+Recall}} \qquad (3)$$

- *Accuracy:* The ratio of the total number of true value predicted to the sum of all the predicted values.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \qquad (4)$$

3.2. Performance analysis

Table 7 lists various models’ average accuracy, precision, recall, and F1-score performance on IEEE heart disease data. With 93.37% accuracy, the K-nearest neighbor outperformed the other classifiers. In comparison to other methods, the performance of DT, LR, and SVM are a little lower. For XGB, Extra tree, RF, and KNN; the Precision, recall, and F1-score remain consistent at approximately >=0.90. It demonstrates that the features are learned effectively because the model predicted instances more accurately.

Table 7 Performance of Classifiers after validation on IEEE heart disease data.

Models	Average Accuracy	Precision/ PPV	Recall / Sensitivity	F1-score
XGB	91.01%	0.90	0.91	0.90
Extra Tree	92.73%	0.92	0.93	0.92
DT	88.29%	0.88	0.87	0.88
RF	92.95%	0.92	0.94	0.92
LR	85.67%	0.85	0.85	0.85
KNN	93.37%	0.92	0.94	0.93
SVM	88.98%	0.87	0.90	0.88

Table 8 shows the effectiveness of various ensemble models on IEEE heart disease dataset, and the proposed strategy of "stacking" exhibits the highest average accuracy (93.57%) after validation and the highest True positive rate or recall (0.95).

Table 8 Performance of ensemble after validation on IEEE heart disease data.

Models	Average Accuracy	Precision/ PPV	Recall / Sensitivity	F1-score
Voting	92.21%	0.91	0.92	0.91
Adaboost	85.68%	0.85	0.84	0.84
Bagging	90.15%	0.89	0.90	0.89
Stacking	**93.67%**	**0.92**	**0.95**	**0.93**

Following repeated stratified validation, Figure 6 uses a box plot from the "matplotlib.pyplot" module to show the average accuracy and standard deviation of models. Compared to other models, the Adaboost (ada) has the highest standard deviation (0.040) and the lowest mean accuracy (85.68%). After validation, stacking had a maximum accuracy of 97.4% with a variation of 0.03. However, stacking had an excellent average accuracy of 93.67%.

To address the issue of poor generalizability, we investigated several datasets with varying sample sizes, data types (categorical, numeric, and mixed), and data sources (diversity in location). After pre-processing, the multiple models are applied to different datasets, and it is clear from Figure 7 that regardless of type, source, or size of the dataset, the proposed approach of stacking with proper hyperparameter tuning achieved excellent accuracy without removing any clinical features.

Several datasets as listed in Table 9, were used to investigate the other performance indicators such as recall, precision, and F1-score. The stacking approach provides greater precision (>= 0.75), as shown in Figure 8**.** The recall (in Figure 9) and F1-score (in Figure 10) reflect good results, with the exception of the unbalanced FHS dataset.

Figure 6 Box plot representations of various models on IEEE heart dataset.

Figure 7 Comparison of accuracies of various classifiers on different datasets.

Table 9 Datasets used for performance validation.

Datasets	Feature Count	Description
IEEE Heart disease	11	Combination of 5 datasets with 7 Categorical and Numerical clinical features
FHS	16	Highly imbalanced datasets
Cleveland	14	Small data set with categorical features
SMOTE FHS	16	Balanced FHS dataset with synthetic oversamples

Figure 8 Precision comparisons of various classifiers on different datasets.

Figure 9 Recall comparisons of various classifiers on different datasets.

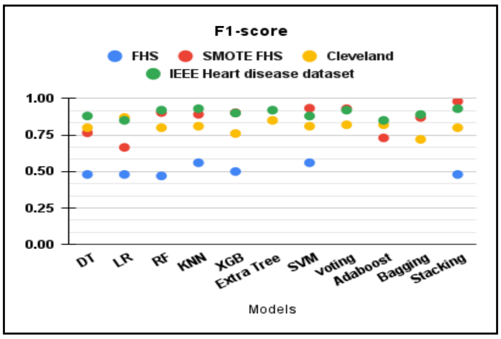


Figure 10 F1 score comparison of various classifiers on different datasets.

Table 10 lists several researches that were used to predict cardiac disease on the IEEE comprehensive dataset. By implementing artificial neural networks (ANN) (Mohan et al 2019), support vector machines (SVM) (Hossain et al 2021; Khennou et al 2019), and K-nearest neighbor (KNN) algorithms (Shah et al 2020) were managed to obtain less than 90% accuracy with 11 clinical features. Other researchers decreased the number of features to achieve more accuracy. Here, we have utilized a 2-level stacked strategy and used seven different classifiers with tuned hyperparameters without discarding any clinical variables, resulting in the highest accuracy of 93.67%.

Table 10 Comparison of Stacked generalization approach on IEEE comprehensive dataset with other studies.

Ref.	Year	Feature count	Classifiers	Average Accuracy (%)
Proposed approach	**2022**	**11**	**Stacked generalization (stacking)**	**93.67%**
(Mohan et al 2019)	2019	11	ANN	88.47%
(Deb et al 2022)	2022	7	Bagging (Decision Tree)	91.68%
(Khennou et al 2019)	2019	11	SVM	87%
(Hossain et al 2021)	2021	11	SVM	85.49%
(Vellameeran and Brindha 2022)	2022	–	PS-GWO algorithm + deep belief network	83.83%
(Shah et al 2020)	2020	11	KNN	77%
(Reddy et al 2019)	2019	6	RF	92.44%

All clinical features of the IEEE-CHD dataset were used by (Mohan et al 2019), (Shah et al 2020), (Khennou et al 2019), and (Hossain et al 2021), who also employed different classifiers such as ANN, KNN, and SVM. Due to the absence of appropriately tuned hyperparameters and the ensemble approach, the average accuracy is < 90%. The proposed method with complete clinical features performed extremely well and received > 90% of accuracy with good precision of 0.92, recall of 0.95, and f1 score of 0.93 due to proper choice of the base model, effective data handling, hyper-parameter tuning with several techniques like randomized search, grid search, and hyperopt sklearn. Proper feature engineering, data pre-processing, and hyperparameter tuning are crucial steps in building an effective stacked model.

4. Conclusions

This study develops a generic model for heart disease prediction that can handle diverse clinical data irrespective of sources, sample sizes, data types, and origins. We have utilized stacking which incorporates numerous models and takes advantage of their strengths for feature learning. The model is tested on multiple databases with all clinical features to determine efficacy. The detailed analysis of the IEEE-CHD database shows that stacking recorded a remarkable average accuracy of 93.67% (maximum 97.64%). In future, as technology advances, more chances for collecting diverse clinical data will emerge, such as wearable devices, continuous monitoring, genetic information, and electronic health records. Integrating such data efficiently into the stacking approach can result in more accurate forecasts and a better knowledge of risk factors for heart disease prediction. The proposed approach can be deployed in real-time clinical settings as a generalized heart disease prediction model that can detect diseases early to assist healthcare professionals in making timely and accurate decisions. While prediction analytics holds immense potential to revolutionize various sectors, its real-world implementation requires addressing challenges and limitations related to data, models, interpretability, and ethics such as data quality and availability, ethical concerns of data privacy and sharing, model interpretability and understanding.

Ethical considerations

Not applicable.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This research did not receive any financial support.

References

Ahmed Z, Mohamed K, Zeeshan S, Dong XL (2020) Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. DOI: 10.1093/database/baaa010

AlBadri A, Leong D, Merz CNB, Wei J, Handberg EM, Shufelt C, Mehta PK, Nelson ML, Thomson L, Berman DS, Shaw LJ, Cook-Wiens G, and Pepine CJ (2017) Typical angina is associated with greater coronary endothelial dysfunction but not abnormal vasodilatory reserve. DOI: 10.1002/clc.22740

BeunzaJ J, Puertas E, García-Ovejero E, Villalba GP, Condes E, Koleva G, Hurtado C, Landecho M F (2019) Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). DOI: 10.1016/j.jbi.2019.103257

BioLINCC: Framingham Heart Study-Cohort (FHS-Cohort) Available in: https://biolincc.nhlbi.nih.gov/studies/framcohort/ Accessed on: January 22, 2022.

Budholiya K, Shrivastava SK, Sharma V (2020) An optimized XGBoost based diagnostic system for effective prediction of heart disease. DOI: 10.1016/j.jksuci.2020.10.013

Dang T, Lan X, Weng J, Feng M (2022) Federated Learning for Electronic Health Records. DOI: 10.1145/3514500

Deb A, Koli MNY, Akter B, Chowdhury AA (2022) An Outcome Based Analysis on Heart Disease Prediction using Machine Learning Algorithms and Data Mining Approaches. DOI: 10.1109/aiiot54504.2022.9817194

Doppala BP, Bhattacharyya D, Chakkravarthy M, Baik N (2022) A Reliable Machine Intelligence Model for Accurate Identification of Cardiovascular Diseases Using Ensemble Techniques. DOI: 10.1155/2022/2585235

Framingham heart study dataset (2022, April 19) Kaggle Available in: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset Accessed on: April 24, 2022.

Hossain AI, Sikder S, Das A, Dey AK (2021) Applying Machine Learning Classifiers on ECG Dataset for Predicting Heart Disease. DOI: 10.1109/acmi53878.2021.9528169

Javaid M, Haleem A, Singh RP, Suman R, Rab S (2022) Significance of machine learning in healthcare: Features, pillars and applications. DOI: 10.1016/j.ijin.2022.05.002

Khennou F, Fahim C, Chaoui H, Chaoui NEH (2019) A Machine Learning Approach: Using Predictive Analytics to Identify and Analyze High Risks Patients with Heart Disease. DOI: 10.18178/ijmlc.2019.9.6.870

Latha CBC, Jeeva S (2019) Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. DOI: 10.1016/j.imu.2019.100203

Miriyala GP, Sinha AK, Jayakody DNK, Sharma A (2021) A Review on Recent Machine Learning Algorithms Used in CAD diagnosis. DOI: 10.1109/iciafs52090.2021.9605854

Mohan S, Thirumalai C, Srivastava G (2019) Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. DOI: 10.1109/access.2019.2923707

Paul B, Karn B (2022) Heart disease prediction using scaled conjugate gradient backpropagation of artificial neural network. Soft Computing, 1-16.

Prevent Heart Disease | cdc.gov (2023) Centers for Disease Control and Prevention Available in: https://www.cdc.gov/heartdisease/prevention.htm. Accessed on: February 12, 2023

Reddy NSC, Nee SS, Min LC, Ying CS (2019) Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction. DOI: 10.11113/ijic.v9n1.210

Shah D, Patel SN, Bharti SK (2020) Heart Disease Prediction using Machine Learning Techniques. DOI: 10.1007/s42979-020-00365-y

Siddhartha M (2020, November 6) Heart Disease Dataset (Comprehensive) IEEE DataPort Available in: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive#:~:text=This%20heart%20disease%20dataset%20is,so%20far%20for%20research%20purposes. Accessed on: October, 2022.

Silent myocardial ischemia: Epidemiology, diagnosis, treatment, and prognosis (2022) UpToDate. Available in: https://www.uptodate.com/contents/silent-myocardial-ischemia-epidemiology-diagnosis-treatment-and-prognosi. Accessed on:February 12, 2023.

UCI Machine Learning Repository: Heart Disease Data Set Available in: https://archive.ics.uci.edu/ml/datasets/heart+disease Accessed on: August 24, 2022.

Vellameeran FA, Brindha T (2022) A new variant of deep belief network assisted with optimal feature selection for heart disease diagnosis using IoT wearable medical devices. DOI: 10.1080/10255842.2021.1955360

World Health Organization: WHO (2021) cardiovascular diseases (CVDs) Available in: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed on: January 21, 2022.

Wu J, Li Y, MacDonald AW (2021) Comparison of XGBoost and the Neural Network model on the class-balanced datasets. DOI: 10.1109/icftic54370.2021.9647373