Journal of Science and Technology 22 (December 2024): 114-121
ISSN 1994-0386 (Print), 2959- 3239 (Online)

BOOSTING HEART ATTACK PREDICTION PERFORMANCE: AN ENSEMBLE LEARNING PERSPECTIVE

Md. Ziaul Hassan*, **Md. Samsul Islam, Arham Aziz Noman and Md. Amir Hamja**
Department of Statistics, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh
*Corresponding author: Email: nirstathstu@gmail.com, Cell Phone: +88-01768932296
DOI: https://www.doi.org/10.59125/JST.22111

ABSTRACT
A cardiovascular event, commonly known as a heart attack, occurs when the blood flow to the heart muscle is abruptly obstructed. Medical research suggests that lifestyle choices primarily contribute to this cardiac issue. Additionally, various crucial factors serve as indicators, signaling whether an individual is at risk or not of experiencing a heart attack. This study aims to predict heart attack based on various Machine Learning (ML) algorithms and proposes an ensemble learning that can best predict heart attack than other models. Additionally provides the top features influencing to have heart attack. The dataset is collected from Kaggle, that contains 14 related features with heart attack. Several ML models including Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), AdaBoost, Gradient Boosting Classifier (GBC) are utilized individually and by combining them four stacking ensemble model are created. Then the trained models are compared based on various metrices such as, accuracy, precision, recall, F1-score, and AUC values. Among the individual models RF, KNN, and LR proved best on the performance measures with accuracy 0.85 for these three. Overall, Stacking Ensemble 1 and Stacking Ensemble 3 outperformed others by achieving accuracy of 0.89 and AUC of 0.88 and leads to conclude that both are equally best. The feature importance reveals exang as the most critical factor contributes to heart attack followed by cp, ca, thal, and sex of the individuals.

Keywords: Boosting, ensemble stacking model, heart attack, machining learning, random forest

INTRODUCTION
Heart disease is a leading cause of death globally, with heart attacks, or myocardial infarctions, occurring when blood flow to the heart muscle is blocked. This blockage is often caused by plaque buildup in the arteries. Without oxygen, the affected heart muscle begins to die. Common symptoms include chest pain, discomfort in other body areas, shortness of breath, fatigue, and dizziness. Risk factors include smoking, high blood pressure, high cholesterol, diabetes, obesity, and family history. Lifestyle changes such as a healthy diet, regular exercise, and stress

management can lower the risk. Early detection is key for prevention and treatment. Machine learning (ML) techniques have become essential in healthcare for disease prediction and diagnosis. In the case of cardiovascular diseases, particularly heart attacks, early and accurate prediction is crucial, as symptoms may not always be apparent, and delays in treatment can result in severe damage to the heart muscle, increasing the risk of mortality. To meet this challenge, Eladham *et al.* (2023) applied five ML algorithms—Multi-Layer Perceptron (MLP), Radial Basis Function (RBF), SVM, KNN, and RF—on the Public Health Heart Attack dataset to enhance prediction accuracy and timely intervention. Among these, KNN emerged as the top performer, showcasing outstanding overall performance. Vynokurova *et al.* (2018) demonstrated significant improvements in predictive model performance using stacking techniques. Their study highlighted the effectiveness of a novel ensemble stacking model, combining RF and eXtreme Gradient Boosting (XGB), with age as a key factor for accurate sleep stage analysis. Satapathy *et al.* (2021) extracted linear and non-linear features from pre-processed signals and used the Relief weight algorithm for feature selection. Testing on the Sleep-EDF (S-EDF) and Sleep Expanded-EDF (SE-EDF) datasets, the model achieved up to 91.10% classification accuracy with the inclusion of age, outperforming existing methods and marking a significant advancement in automated sleep staging techniques. The healthcare industry holds vast data often underutilized for decision-making. This research presents the Intelligent Heart Disease Prediction System (IHDPS), which applies data mining techniques such as DTs, Naïve Bayes, and Neural Networks to uncover hidden patterns in medical data. These techniques help predict heart disease risk based on factors like age, sex, blood pressure, and blood sugar. IHDPS, implemented as a web-based application on the NET platform, provides more precise insights than traditional systems, enhancing healthcare decision-making. Given the critical nature of heart attacks, this study seeks to predict them using ensemble machine learning techniques. The project's goals are as follows:
- Improve the Precision and Accuracy of ML model by using ensemble technique.
- Determine the Most Important factor associated with heart attack.

METHODS AND MATERIALS

ML, a subset of Artificial Intelligence (AI), entails creating algorithms and statistical models that empower computers to enhance their performance on a particular task through experience, without direct programming. This section describes the details about the methods used by the authors in this study.

LR: LR is a statistical method used for binary and multiclass classification tasks. Despite its name, it's primarily for classification rather than regression. It models the probability that an input belongs to a specific class, making it effective for binary outputs (e.g., 0/1, Yes/No). It uses the sigmoid (logistic) function to map real-valued inputs to a range between 0 and 1. The sigmoid function is defined by the formula:

$$\sigma(z)=\frac{1}{1+e^{\wedge}-z}$$

Here, z is a linear combination of the input features and their corresponding weights. Saw *et al.* (2020) used LR to predict the heart attack.

DT: DT is a popular machine learning algorithm used for both classification and regression tasks. According to Eladham *et al.* (2023), it works by recursively partitioning the input space and assigning class labels to the resulting regions. The algorithm's structure resembles a tree, where each internal node represents a feature, each branch corresponds to a decision based on that feature, and each leaf node indicates a class label or final prediction.

SVM: The SVM is a powerful supervised learning algorithm used for both classification and regression tasks. Eladham *et al.* (2023) explained for classification, SVM identifies the optimal hyperplane that separates the data into distinct classes, maximizing the margin between them for better generalization to new data. It is versatile and can handle both linear and non-linear data using different kernel functions, such as linear, polynomial, RBF and sigmoid kernels.

KNN: KNN algorithm is a straightforward and intuitive machine learning approach employed for both classification and regression tasks (Satapathy *et al.* 2021). In the classification version when given a new data point, it finds the 'k' nearest data points in the training set and classifies the new point based on the majority class among its neighbors. The value of 'k' (number of neighbors) is a hyperparameter that can be tuned based on the problem domain.

RF: The RF is a widely adopted ensemble learning technique utilized for both classification and regression tasks. It functions by creating numerous decision trees during the training phase and produces the mode (for classification) or mean prediction (for regression) of the individual trees (Marikani *et al.* 2017). It find extensive applications in various domains, including image recognition, remote sensing, finance, and healthcare, owing to their capability to handle intricate datasets and provide dependable predictions.

GBC: It is an additional ensemble learning technique that sequentially constructs multiple decision trees, with each subsequent tree aimed at correcting the errors of the preceding ones (Eladham *et al.* 2023)**.** Unlike AdaBoost, it builds trees in a sequential manner, where each tree is fitted to the residuals of the combined ensemble. This iterative process results in a strong predictive model.

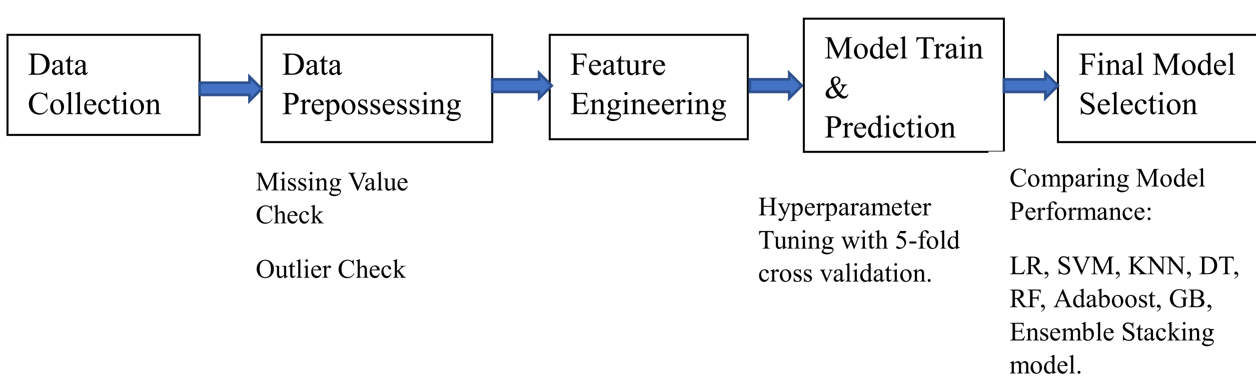
Stacking Model: Stacking is an ensemble learning method that combines multiple base models into a meta-model for improved predictions. Unlike parallel ensemble techniques like RF or GBC, stacking involves sequential training of diverse base models (e.g., DT, SVMs, NNs). Each base model is trained on the original data, and the meta-model learns to make predictions based on the outputs of these models. The diversity of base models is crucial, as they capture different patterns in the data, leading to better overall performance. To prevent data leakage during meta-model training, cross-validation is used, ensuring that base models do not access validation or test data. The choice of the meta-model, whether it be LR, DT, or NN, significantly influences stacking’s performance. Proper experimentation and tuning are essential to achieving the best results, especially for complex machine learning tasks.

Data Source and Dataset Description: This study uses an open-source dataset for analysis namely “Heart Disease Cleveland” is collected from Kaggle. The dataset comprises 14 related factors to heart attack. The "target" field denotes the presence of heart disease in the patient, with integer values: 0 signifies no/less chance of a heart attack, and 1 indicates a higher chance of a heart attack. The detailed description about the dataset is in Table 1.

Table 1. Description of the Variables in the dataset:

Variable	Description
1) age	Ranges from 28 to 82
2) sex	Male / Female
3) chest pain	There are 4 types
4) resting blood pressure	Blood pressure value
5) serum cholesterol in mg/dl	Values in in mg/dl
6)fasting blood sugar > 120 mg/dl	Values > 120 mg/dl
7)Resting electrocardiographic results	values 0,1,2)
8) maximum heart rate achieved	In the diagnostic
9) exercise induced angina	Yes, no
10) oldpeak	ST depression induced by exercise relative to rest
11)the slope of the peak exercise ST segment	Values
12) number of major vessels	(0-3) colored by fluoroscopy
13) thal	: 0 = normal; 1 = fixed defect; 2 = reversable defect
14) target	: 0= less chance of heart attack ,1= more chance of heart attack

We propose a framework for building and testing a predictive model as shown below to predict the heart attack. Firstly, we collect dataset then as preprocessing steps we checked for missingness, and outliers and it reveals there is no missingness and outliers in dataset. Then as feature engineering step we converted the high range columns within smaller range by scaling them and converted categorical columns into numerical by encoding. The Process Diagram is as follows:



RESULTS AND DISCUSSION

Single Model Fitting: To begin with the individual models this study identified hyperparameters for each model aims at getting better performances. For knowing about the exact hyperparameters we provided a range of parameters for each model to Grid Search CV. The tuning parameters are given in Table 2.

Table 2. Hyper parameter with 5-fold cross validation

Model	Best Hyper Parameter
LR	C=0.1, L1 penalty
SVM	C=0.1, kernel='linear'
KNN	n_neighbors=9
DT	max_depth=5
RF	max_depth=5, min_samples_leaf=4, min_samples_split=10, n_estimators=150
AdaBoost	base_estimator: DecisionTreeClassifier, DecisionTreeClassifier(max_depth=5)
GBC	GradientBoostingClassifier(max_depth=5, n_estimators=50

Model Accuracy and different performance Indicator

The performances of the trained and proposed ensemble models are given to the Table 3 in terms of accuracy, precision, recall, F1-socre and AUC values.

Table 3. Model accuracy, f1-Score, precision, recall and ROC AUC value:

Model Name	Accuracy	F1 Score	Precision	Recall	ROC AUC
LR	**0.85**	0.87	0.84	0.91	**0.84**
SVM	0.84	0.86	0.80	0.94	0.82
KNN	**0.85**	0.87	0.84	0.91	**0.84**
DT	0.79	0.81	0.82	0.79	0.79
RF	**0.85**	0.87	0.84	0.91	**0.84**
AdaBoost	0.82	0.83	0.87	0.79	0.82
GBC	0.77	0.79	0.81	0.76	0.77
Stacking Ensemble 1	**0.89**	0.91	0.88	0.93	**0.88**
Stacking Ensemble 2	0.85	0.86	0.81	0.90	0.85
Stacking Ensemble 3	**0.89**	0.91	0.88	0.93	**0.88**
Stacking Ensemble 4	0.85	0.87	0.80	0.91	0.84

Among the individual models LR and RF, exhibit consistent and strong performance with measures of accuracy 0.85, F1 score 0.87, precision 0.84, and recall 0.91, and AUC 0.84 which is closely followed by KNN which gets AUC value 0.82 with similar for others metrices. But GBC achieved the lowest performances in predicting heart attack with accuracy of 0.77 and AUC of 0.77.

In our pursuit of optimizing model performance, we explored several stacking ensembles, each combining distinct base classifiers to harness their individual strengths. Let's delve into our observations on these stacking models, considering both accuracy and AUC values. The Stacking Ensemble 1 (GBM + KNN + SVM) Model emerged as a standout performer, boasting an accuracy of 0.89 and an AUC of 0.88. By combining GBC, KNN, and SVM, this ensemble demonstrated exceptional predictive accuracy and a strong ability to distinguish between classes. Next, the Stacking Ensemble 2 (RF + SVM) model displayed a balanced performance with an accuracy and AUC both measuring 0.85. The Stacking Ensemble 3 (DT + SVM + KNN) model mirrored the performance of the first stacking model, boasting an accuracy of 0.89 and an AUC of 0.88. This ensemble, incorporating DT, SVM, and KNN, demonstrated a better predictive capability. Lastly, the Stacking Ensemble 4 (LR + KNN + RF) model exhibited competitive performance, albeit slightly lower than the other stacking ensembles. With an accuracy of 0.85 and an AUC of 0.84, this combination displayed a balanced trade-off between accuracy and discriminative power.

Feature Importance

The feature importance plot is in Figure 1.

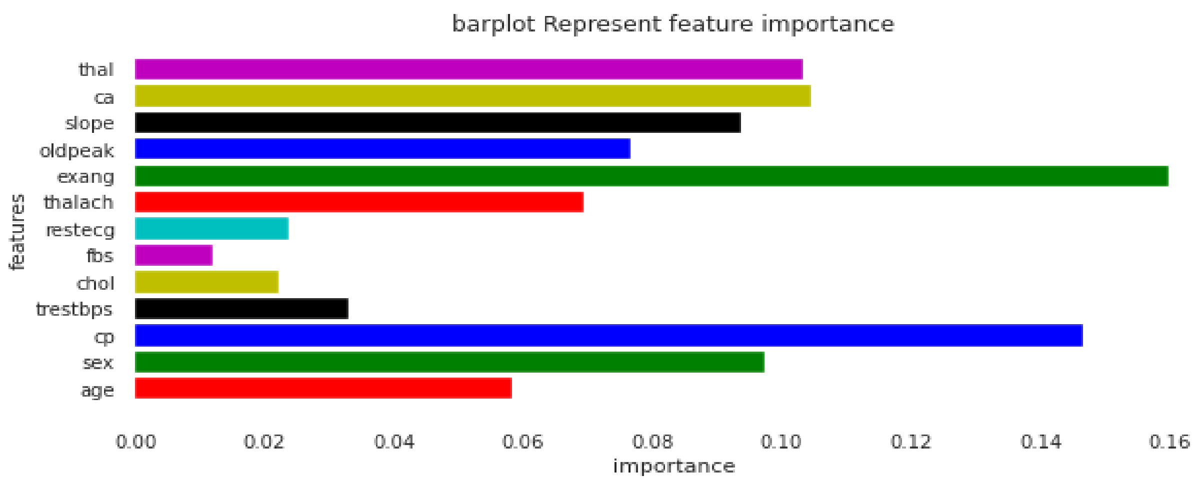


Figure 1. Feature Importance

According to Figure 1, we can conclude that exang is the most critical factor affecting heart attack on individuals. And others like cp, ca, thal and sex are also influencing to having heart attack.

Table 4. Performance of the existing model with heart disease dataset

Reference	Best Model	Performance	Type
Proposed Stacking Ensemble 1	GBM + KNN + SVM	89.00%	Ensemble
Proposed Stacking Ensemble 3	DT + SVM + KNN	89.00%	Ensemble
Reddy *et al.* (2021)	Sequential Minimal Optimization (SMO)	86.48%	Single
Latha *et al.* (2019)	Majority vote with NB, BN, RF and MP	85.48%	Ensemble
Paul *et al.* (2016)	Neural Network with fuzzy	80.00%	Ensemble
Verma *et al.* (2016)	DT	80.68%	Single
Ei-bialy *et al*. (2015)	DT	78.54%	Single
Nahar *et al*. (2013)	NB	69.11%	Single

The proposed stacking ensembles, demonstrate a promising approach with accuracies reaching up to 89.0%. Specifically, the stacking ensemble configurations "GBM + KNN + SVM" and "DT + SVM + KNN" showcase high accuracy levels of 89.0%, indicating a successful integration of diverse classifiers. Reddy *et al.* (2021) achieved 86.48% accuracy using Sequential Minimal Optimization (SMO) algorithm. The Majority Vote ensemble by Latha *et al.* (2019) achieves an accuracy of 85.48%, utilizing NB, BN, RF, and MP classifiers. Individual classifiers also contribute to the landscape, with NN incorporating fuzzy logic from Paul *et al.* (2016) achieving an 80.0% accuracy. Verma *et al*. (2016) employs DT as a standalone classifier, yielding an accuracy of 80.68%. Ei-bialy *et al.* (2015) and Nahar *et al.* (2013) also utilize DT and NB individually, achieving accuracies of 78.54% and 69.11%, respectively.

CONCLUSION

This study aimed to predict the likelihood of heart attacks using a variety of ML models and demonstrated that ensemble learning techniques can provide superior predictive performance compared to individual models. Several ML models, including LR, DT, SVM, KNN, RF, AdaBoost, and GBC, were employed to classify heart attack. Among these individual models, RF, KNN, and LR stood out, each achieving an accuracy of 0.85 on the dataset. However, when four stacking ensemble models were introduced, they proved more effective in improving the overall predictive accuracy. Stacking Ensemble 1 and Stacking Ensemble 3 outperformed the other models, with the highest accuracy reaching 0.89 and an AUC score of 0.88, demonstrating the power of combining diverse models to capture various patterns in the data. This leads to conclude that both Ensemble1 and Ensemble 3 model are equally good. The analysis of feature importance revealed that exercise-induced angina (exang) was the most critical predictor of heart attack, followed by chest pain type (cp), the number of major vessels (ca), thalassemia (thal), and sex. These findings align with existing medical literature, reinforcing the importance of these features in determining cardiovascular risk. Future research could focus on further enhancing model performance by exploring advanced DL techniques and hybrid ensemble methods. Additionally, integrating more diverse datasets from multiple sources could improve the generalization of the models. Moreover, explainability techniques like SHAP or LIME could be employed to provide clearer insights into individual predictions, aiding medical practitioners in decision-making.

REFERENCES

EI-Bialy R, Salamay MA, Karam OH and Khalifa ME. 2015. Feature analysis of coronary artery heart disease data sets. Procedia Computer Science. 65:459-468.

Eladham MW, Nassif AB and AlShabi M. 2023. Heart attack prediction using machine learning. Proceeding of the Smart Biomedical and Physiological Sensor Technology XX, 125480F.

Latha BCC and Jeeva CS. 2019. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked. 16:100203

Marikani T and Shyamala K. 2017. Prediction of Heart Disease using Supervised Learning Algorithms. International Journal of Computers and Applications. 165 (5):41–44.

Nahar J, Imam T, Tickle KS and Chen YPP. 2013. Computational intelligence for heart disease diagnosis: a medical knowledge driven approach. Expert System. 40(1):96–104.

Paul AK, Shill PC, Rabin MRI and Akhand MAH. 2016. Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease, in Proceedings of 5th International Conference on Electronic Commerce. pp. 145–150.

Reddy KVV, Elamvazuthi I, Aziz AA, Paramasivam S, Chua HN and Pranavanand S. 2021. Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators.11(18): 8352.

Satapathy SK, Bhoi AK, Loganathan D, Khandelwal B and Barsocchi P. 2021. Machine learning with ensemble stacking model for automated sleep staging using dual-channel EEG signal. Biomedical Signal Processing and Control. 69: 102898

Saw M, Saxena T, Kaithwas S, Yadav R and Lal N. 2020. Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning. International Conference on Computer Communication and Informatics. pp.1-6

Verma L, Srivastava S, Negi PC. 2016. A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. Journal of Medical System. 40(7):1–7.

Vynokurova O and Peleshko D. 2018. The Autoencoder Based on Generalized Neo-Fuzzy Neuron and its Fast Learning for Deep Neural Networks. IEEE Second International Conference on Data Stream Mining & Processing (DSMP). pp. 113-118.