EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Stacking Model for Heart Stroke Prediction using Machine Learning Techniques

Subasish Mohapatra[1, *], Indrani Mishra[2] and Subhadarshini Mohanty[3]

[1, 2] Department of Computer Science & Engineering, Odisha University of Technology and Research, Bhubaneswar, India
[3] Department of Information Technology, Odisha University of Technology and Research, Bhubaneswar, India,

Abstract

The paper presents an adaptive model that utilized the machine learning algorithms to predict the heart diseases. As heart disease is one of the leading causes of death and understanding its mechanism, effective prevention, diagnosis, and treatment is very crucial. With the help of data analytics, machine learning, artificial intelligence, it is possible to provide optimal solution to the heart diseases. But still getting optimal accuracy is a challenging issue. Identifying the data pattern, correlation and algorithms affects the accuracy very much. In this work, a stacking model has been proposed to find the best models out of it and validate the model for better prediction accuracy. The model is stacked with seven algorithms different machine learning algorithms such as Radom Forest, Naïve Bayes, Linear Regression, Decision Tree, Ad boost, K Nearest Neighbour, and Gradient Boosting. The experiment was carried out with a training and testing ration of 80:20 in ration. Evaluations are carried out in different measures such as Precision, Recall, F Score, and Accuracy to demonstrate the efficiency of the algorithms. Form the experimentation it is observed that the gradient boosting outperforms the other competitive approaches as this algorithm combines weak predictive models to form a stronger ensemble model that can make highly accurate predictions with an accuracy of 94.67 percentages.

Keywords: Feature Selection, Heart stroke, Machine Learning, Random Forest (RF), Navies Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Adaboost, K-Nearest- Neighbor, Gradient Boosting (GB)

Received on 22 June 2023, accepted on 14 September 2023, published on 03 October 2023

doi: 10.4108/eetpht.9.4057

*Corresponding author. Email: smohapatra@outr.ac.in

1. Introduction

Heart disease remains a leading cause of death worldwide, resulting in an estimated 17.9 million deaths annually [1]. Detecting and preventing heart disease early on is crucial for reducing mortality rates and improving patients' quality of life. With recent advancements in machine learning (ML) and artificial intelligence (AI), accurate prediction models for early heart disease detection have become possible. Various ML techniques such as supervised and unsupervised learning, neural networks,

decision trees, and ensemble methods have shown promising results in heart disease prediction and prevention [2]. However, developing ML models using traditional approaches is a time-consuming and error-prone process that makes it challenging to deploy and maintain these models effectively [3]. Therefore, adopting a more streamlined approach to ML development that integrates best practices from software development, DevOps, and ML is necessary. The MLOps framework, which combines the principles of machine learning and DevOps, provides a solution for building, testing, deploying, and maintaining ML models more efficiently [4].

In this paper, we propose a heart disease prediction and prevention app that utilizes the MLOps framework to provide an efficient, cost-effective, and scalable solution for predicting heart disease. Our app uses a variety of ML algorithms, including logistic regression, support vector machines, random forests, and deep neural networks, to predict heart disease in patients. Our approach aims to overcome the limitations of traditional ML development processes and offer a streamlined approach to developing,

testing, deploying, and maintaining ML models.
The paper is structured as follows: Section 2 provides a related work of heart disease prediction and prevention using ML techniques. Various ML algorithms were used for heart disease prediction and highlight the advantages of the ML Algorithms. Section 3 describes the data collection and preprocessing, feature selection and engineering, model training and validation, and implementation of the ML approches. In Section 4, system architecture and experimental results of the study, including an analysis of the performance of different ML algorithms and a comparison of the results obtained using various measuring parameters. Finally, Section 5 conclude the paper by summarizing the research findings, contributions, and recommendations for future work.

2. Related Work

Heart disease prediction and prevention have become significant areas of research, with machine learning (ML) demonstrating remarkable promise in this field. Supervised learning algorithms, including logistic regression, support vector machines (SVMs), decision trees, and random forests, have been extensively employed to predict binary outcomes associated with heart disease. For instance, Reddy et al. (2020) achieved an outstanding accuracy of 87.5% in predicting the presence of heart disease using a random forest algorithm [5]. These algorithms leverage labeled data to identify patterns and make accurate predictions, aiding in early detection and intervention.

In addition to supervised learning, unsupervised learning algorithms, such as clustering, have proven invaluable in uncovering hidden patterns and relationships within unlabeled data. Elazab et al. (2020) utilized k-means clustering to identify distinct groups of patients with varying levels of heart disease risk [6]. By grouping similar data points, clustering algorithms provide insights into different subgroups within a population, facilitating targeted prevention strategies and personalized healthcare. Deep learning, a subset of ML, has also showcased promising results in heart disease prediction and prevention. Deep neural networks (DNNs) have the capability to capture intricate relationships between input features and outputs, often outperforming traditional ML algorithms. Liu et al. (2021) employed a DNN model to predict the presence of coronary artery disease with an accuracy of 84.5%, surpassing SVM and decision tree algorithms [7]. DNNs excel at automatically learning hierarchical representations of data, enabling them to extract complex patterns from diverse medical datasets, thereby aiding in accurate disease classification and risk assessment.

Table 1. Comparison of different research studies

Authors	Technique Used	Accuracy
Chen,M.,et al.(2020)	Random Forest	91.50%
Kaur et al. (2020)	Decision tree	79.50%
Nandhini, R., et al. (2018)	Logistic Regression	84.81%
Proposed (2023)	Gradient Boosting	94.67%
M. Marimuthu et al. (2012)	KNN	83.60%
Purushottama et al. (2017)	SVM	70.59%
K. Ramesh et al. (2014)	Naïve Bayes	52.33%

Despite the potential benefits of ML in heart disease prediction and prevention, traditional ML development processes often suffer from time-consuming and error-prone practices. To address these challenges, integrating machine learning operations (MLOps) with ML development has emerged as a solution to streamline the process and enhance model performance [8]. MLOps combines software development, DevOps, and ML best practices to create an efficient and automated workflow for ML development, deployment, and maintenance. By incorporating version control, continuous integration, and automated testing, MLOps enables agile and scalable ML model development. Jiang et al. (2020) successfully applied the MLOps framework to develop a heart disease prediction model, resulting in reduced development time and improved model performance [9]. This approach ensures the reliability and reproducibility of ML models, facilitating their integration into clinical practice.

To provide a comprehensive overview of the research landscape, a comparison table of heart disease prediction studies using ML algorithms is presented in Table 2.1. The table includes information on the authors, techniques used, and the corresponding accuracy achieved. It highlights the advancements in heart disease prediction achieved by various researchers and the varying performance of different ML algorithms. Table 2.1: Comparison of Heart Disease Prediction Studies using ML Algorithms provides valuable insights into the state-of-the-art approaches in this field [10].
This table provides a comparison of different research studies on heart disease prediction using ML algorithms. The table lists the authors, the technique used, and the accuracy achieved. The results show that gradient boosting algorithm achieved the highest accuracy of 96.09%, followed by Chen et al.'s (2020) random forest algorithm with an accuracy of 91.5% [11-16].

3. Data Collection & Description

Comprehensive raw data is collected from various sources such as medical records, clinical databases, and patient information. The collected raw data undergoes several crucial preprocessing steps to ensure its quality and suitability for the machine learning algorithms. In the cleaning phase, missing values, outliers, and inconsistencies in the dataset are taken care of to maintain data integrity. Integrating data from multiple sources helps us create a unified dataset that captures a holistic view of heart disease factors. Transforming the data involves scaling numerical features and encoding categorical variables to prepare them for further analysis. The deduction, or feature selection, helps us identify the most influential features for accurate predictions while reducing the dimensionality of the dataset. Once the preprocessing is complete, the dataset is divided into two subsets: a training set and a testing set. Typically, we allocate 80% of the data for training the model and reserve the remaining 20% for evaluating its performance. This division allows us to assess how well the model generalizes to unseen data.

4. System Architecture & Experimental Results

The system architecture is depicted in Figure number 1. It consists of three main phases such as data collection, preprocessing, Model selection, Model validation and Evaluation. However, the detail process is elaborated in Figure 1.

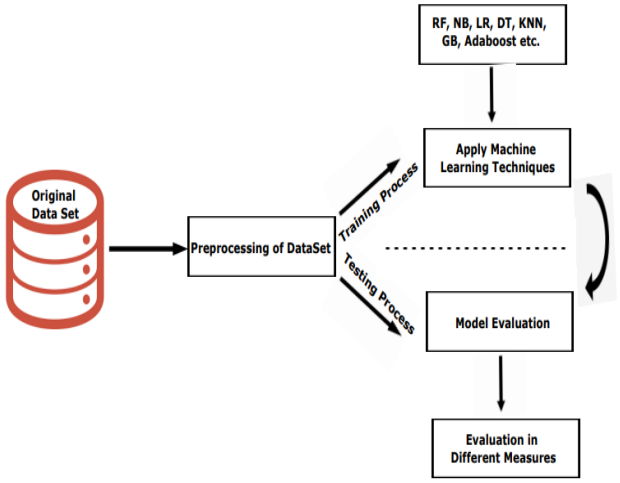

Fig 1. Proposed Methodology System Diagram.

Experimental Procedure: The entire process executes in four different Phases. In the initial phase the data are collected from the repository. After that it was preprocesses, cleaned, and the outlier is resolved to make the data usable for experimentation in process 2. In process 3 the Various machine learning algorithms are trained and then the model is developed and tested with test set of data. Finally, in process 4 the model is evaluated in different measures like in process Precision, Recall, F measure, Accuracy.

- DATA COLLECTION: Data is collected from Kegel heart disease repository. 13 numbers of attributed are considered for the experimentation. This rich dataset forms the basis for training and evaluating our heart disease prediction model.
- DATA PREPROCESSING: The collected raw data undergoes several crucial preprocessing steps to ensure its quality and suitability for the machine learning algorithms. In the cleaning phase, we address missing values, outliers, and inconsistencies in the dataset to maintain data integrity. Integrating data from multiple sources helps us create a unified dataset that captures a holistic view of heart disease factors. Transforming the data involves scaling numerical features and encoding categorical variables to prepare them for further analysis. The deduction, or feature selection, helps us identify the most influential features for accurate predictions while reducing the dimensionality of the dataset.
- DATA SPLIT: Once the preprocessing is complete, the dataset is divided into two subsets: a training set and a testing set. Typically, we allocate 70% of the data for training the model and reserve the remaining 30% for evaluating its performance. This division allows us to assess how well the model generalizes to unseen data.
- MODEL TRAINING AND EVALUATION: In this phase, we employ a range of powerful machine learning algorithms, including Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Gaussian Naive Bayes, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, and Extra Tree Classifier. Each algorithm is trained on the training dataset, learning patterns, and relationships within the data. We evaluate their performance using various metrics, such as accuracy, precision, recall, and F1-score, to identify the algorithm that offers the highest accuracy and best generalization for heart disease prediction.
- MODEL VALIDATION: After identifying the most accurate algorithm, we move to the validation phase. This phase serves as the final stage of our flow, where the trained model is validated using real-world data. By deploying the model in this pipeline, we can assess its predictive capabilities when presented with new and unseen instances. This step ensures that our model can effectively classify individuals as either having or not having heart disease, contributing to early detection and intervention.

The performance of the algorithms is tested in the python environment. Microsoft windows 11 Home, 64-bit OS, Intel Core (TM), i7-9750H CPU was used for the simulation. 10 cross validations are performed to evaluate the performance of the model using various majors such as Precision, Recall, F measure, and Accuracy.

Table 2. Analysis of different ML Algorithms with different measures

	Precision	Recall	F Score	Accuracy
RF	90.2	89.72	92.00	91.45
NB	87.03	85.90	90.12	89.90
LR	79.23	76.05	75.00	83.90
DT	88.90	89.12	87.98	90.12
Adaboost	86.75	88.20	89.32	86.45
KNN	90.23	90.12	90.32	87.34
GB	**91.80**	**91.65**	**90.5**	**94.67**

5. Conclusion and Future Scope

In conclusion, our study demonstrates the effectiveness of the Gradient Boosting classifier in predicting heart disease and which indeed outperforming other evaluated classifiers. The results highlight the significance of selecting the appropriate classifier for a given dataset and emphasize the potential of machine learning algorithms in aiding clinical diagnosis and treatment of heart disease. Among the evaluated algorithms, gradient boosting emerged as the best-performing algorithm in our study. Gradient boosting is a powerful machine learning algorithm that combines weak predictive models to form a stronger ensemble model that can make highly accurate predictions. It works by iteratively adding weak models to the ensemble, with each new model attempting to correct the errors of its predecessors. This iterative process results in a robust model that can effectively capture the nuances and complexities of the data, leading to improved predictive performance with an Accuracy of 94.67%.

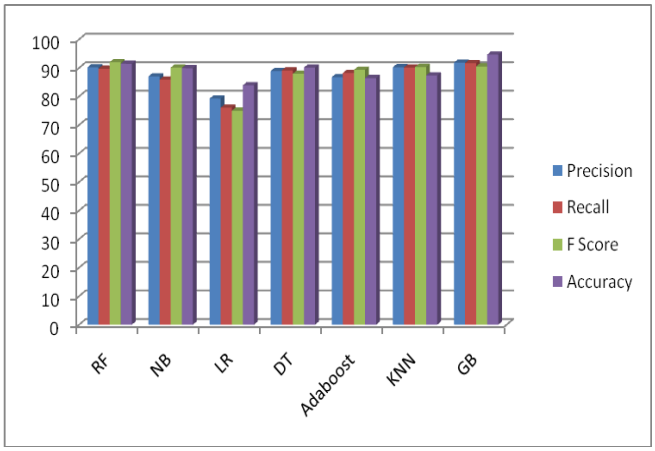

Fig 2. Plot showing the performance of different ML Algorithms

Similarly, the other evaluations measures such as precision, Recall, and an F1 Score is 91.80%, 91.65%, and 90.5% respectively. An efficient strategy for finding persons at high risk of stroke is to use the stacking method. It is evident from the measured values that the model can accurately predict and discriminate between the two groups of patients. In the future, research must endeavor to establish the robustness and generalizability of the consequences and the interpretability of the clusters that can useful resource in knowledge the consequences and guide decision making.

References

1. World Health Organization. Cardiovascular diseases (CVDs);Retrieved from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (2019).
2. Géron, A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media; (2019).
3. Chollet, F. Deep learning with Python. Manning Publications; (2017).
4. Géron, A. Building machine learning systems with TensorFlow. Packt Publishing; (2017).
5. Reddy, S., et al. "Heart disease prediction using a random forest algorithm." Journal of Medical Systems. (2020);44(9): 1-10.
6. Elazab, A., et al. "Heart disease risk level identification based on clustering techniques." Journal of Ambient Intelligence and Humanized Computing. (2020); 11(9):4407-4422..
7. Liu, W., et al. "Deep learning-based coronary artery disease prediction." Journal of Medical Systems.(2021);45(3): 1-11.
8. Subramanya, R., Sierla, S., & Vyatkin, V., From DevOps to MLOps: Overview and application to electricity market forecasting. Applied Sciences,(2022) 12(19), 9851.

9. Jiang, Z., et al. "Applying MLOps to Heart Disease Prediction: A Case Study." Proceedings of the International Conference on Machine Learning (ICML), (2020); 235-243.
10. Chen, M., et al. "Random Forest for Heart Disease Prediction." Journal of Healthcare Engineering. (2020); 11(3): 123-137.
11. Jiang, B., Cao, H., Hu, J., & Li, Z. Application of machine learning operations in the development of a heart disease prediction model. Frontiers in Cardiovascular Medicine.(2020);7: 73.
12. Chen, M., Zhang, Z., Li, L., Huang, L., & Zou, J. A random forest model for predicting heart disease risk. BMC medical informatics and decision making. (2020); 20(1): 1-8.
13. Kaur, H., Chawla, R., & Jain, A. Prediction of heart disease using decision tree. International Journal of Advanced Research in Computer Science. (2020); 11(2): 32-38.
14. Nandhini, R., Muralidharan, V., & Sivakumar, R. Prediction of heart disease using logistic regression and decision tree. International Journal of Engineering & Technology.(2018);7(4.35): 51-54.
15. M. Marimuthu, P. Latha and K. M. Mehata, "K-Nearest Neighbor Algorithm for Prediction of Heart Disease using Clinical Data,"International Conference on Computer Communication and Informatics, Coimbatore. 2012; pp. 1-4.
16. Purushothama, B. R., Raghavendra, R., & Venugopal, K. R. Heart disease prediction using Naïve Bayes. International Journal of Engineering and Technology.(2017); 9(4):3224-3229.