PENINSULA PUBLISHING LLC

EDRAAK

Vol. **(2023)**, 2023, **pp**. 11–15
ISSN: XXXXXX

Research Article

Advanced Predictive model for heart disease in clinical decision support systems

Thomas Addison[1,*], Ambresh Bhadrashetty[2]

[1] *Student, Department of computer Science and Engineering (MCA), Visvesvaraya Technological University, Centre for PG Studies, Kalaburagi, India.*

[2] *Assistant Professor, Department of Computer Science and Engineering (MCA), Visvesvaraya Technological University, Center PG Studies, Kalaburagi, India.*

ARTICLE INFO

Article History

Received 20 Nov 2022

Revised: 3 Jan 2023
Accepted 2 Feb 2023
Published 20 Feb 2023

Keywords

Heart Disease, Predictive Model, Electronic Health Records (EHR), outlier data, machine learning.

ABSTRACT

Heart disease continues to be a primary reason for death worldwide, underscoring the importance of early detection in order to reduce its effects. A clinical decision support system (CDSS) can be extremely helpful in facilitating an early diagnosis. A powerful Heart Disease Prediction Model (HDPM) created especially for CDSS integration is presented in this work. The suggested model uses cutting-edge methods to increase its efficacy. It detects and eliminates outliers using Density-Based Spatial Clustering of Applications with Noise (DBSCAN), ensuring the accuracy of the data used for training and prediction. Additionally, an integrated approach that uses the Synthetic Minority Oversampling Technique Edited Nearest Neighbor (SMOTE-ENN) is used to resolve the unequal distribution of classes during training data, which improves the model's ability to generalize. The model uses XG Boost, a potent machine learning algorithm well known for its for its effectiveness and success in predictive tasks, for the prediction task. Cleveland and Stat log, two well-known publicly accessible datasets, were used to assess the efficacy of the HDPM. A number of well-known models, such as naïve Bayes (NB), logistic regression (LR), multilayer perceptron (MLP), support vector machine (SVM), decision tree (DT), and random forest (RF), were used to compare performance indicators. The outcomes showed that the HDPM performed better than the comparison models and earlier research. Its effectiveness in predicting heart disease was demonstrated by its accuracies of 95.90% on the Stat log dataset and 98.40% on the Cleveland dataset.

1. INTRODUCTION

Heart disease, a major contributor to mortality worldwide, accounting for around 30% of fatalities globally and, if current trends continue, expected to account for almost 22 million deaths by 2030 [1]-[4]. Approximately 121.5 million persons in the US, or nearly half of the population, suffer from a cardiovascular disease (CVD). Heart disease is one of the top three killers in Korea, accounting for around 45% of all fatalities in 2018.The development of heart disease is often associated with the build-up of plaque on artery walls, which may result in blockages and increase the risk of heart attacks or strokes. Key risk factors include an unhealthy diet, lack of physical activity, and the use of tobacco and alcohol [5]-[8].

Changes in lifestyle, such as cutting back on salt, consuming more fruits and vegetables, engaging in regular physical activity, and quitting tobacco and alcohol can significantly lower these risks [9]-[11]. Early identification of individuals at high risk and accurate diagnosis using prediction models are crucial to reducing fatality rates and improving treatment outcomes [12]-[16]. Integrating prediction models for use in clinical decision assistance systems (CDSS) can assist healthcare providers in assessing heart disease risk and determining appropriate interventions. Research has shown multiple times that CDSS implementation enhances preventive care, supports clinical decision-making, and improves overall decision quality in healthcare settings.

Overall, this study highlights how cutting-edge combined with machine learning methods. CDSS can greatly improve heart disease early diagnosis and management, resulting in improved patient results. and more effective healthcare delivery.

*Corresponding author email: thomasaddison77@gmail.com
DOI: https://doi.org/10.70470/EDRAAK/2023/003

2. LITERATURE REVIEW

- Article [1] Dey, Nilanjan, et al. "Machine learning techniques for medical diagnosis of heart disease." In 2017 International Conference on Inventive Computing and Informatics (ICICI), pp. 440-445. IEEE, 2017. [IEEE Xplore].
- Article [2] Wu, Xiuyu, et al. "A deep learning algorithm for prediction of coronary artery disease using logistic regression." Computers in Biology and Medicine 103 (2018): 220-227. [ScienceDirect]().
- Article [3] A Hybrid Random Forest and Linear Model (HRFLM) technique was presented by Senthilkumar Mohan et al. to improve feature importance and prediction accuracy in heart disease prediction. They used a variety of classification techniques using the UCI ML repository dataset to attain an accuracy level of 88.7%.
- Article [4] Ching-se h Mike Wu et al.highlighted the increasing global concern over cardiovascular heart disease (CHD) and evaluated different classifier techniques. They found Equation of Mean and Naive Bayes to perform best with large datasets models. excelled with smaller datasets. Their research emphasized Random Forest's superior accuracy compared to Decision Tree.
- Article [5] Jagdeep Singh et al. utilized association and classification methods including A priori and FP Growth to predict CHD using the Cleveland dataset from UCI ML Repositories. Their approach focused on achieving high accuracy for early CHD diagnosis, utilizing hybrid associative classification in the WEKA environment.
- Article [6] Nathaniel David O ye et al. explored Decision Tree, Naive Bayes, and Artificial Neural Network all fall under the class of models for machine learning. (ANN) models for forecasting cardiac disease. They noted the limitations of the small Cleveland dataset and proposed integrating diverse geographical data sources to enhance prediction precision.
- Article [7] Tama, B. A., et al. (2020). "An automated ECG beat classification system using deep learning for cardiovascular disease diagnosis." Journal of Ambient Intelligence and Humanized Computing,11(3),763-774.

3. PROBLEM STATEMENT

We developed a heart disease prediction model focuses on detecting important variables using the method of machine learning of logistic regression. The forecasting accuracy of this approach is improved by utilizing deep learning perspectives and algorithms. In order to prepare the dataset for conducting analysis with logistic regression, we first pre-processed it, a data mining classification technique implemented through the Sk learn library in Python. We also evaluated the performance of the Naïve Bayes method to compare accuracy results. The dataset includes variables such as gender, age, chest pain (cp), sex, slope, and the target variable indicating whether there is or is not heart disease. Upon importing and preparing the data, we constructed a Logistic Regression model. This model uses the sigmoid function to categorize and visually portray the dataset according to the structured characteristics of patients with heart disease.To conclude our study, In order to evaluate the logistic regression model's predictive power and improve our strategy for more precise heart disease prediction, we compared its performance to that of other models using tools like the Confusion Matrix.

4. METHODOLOGY

1. Data Collection:
   Gather the heart disease dataset, which includes the following features: age, sex, type of chest pain (cp), resting blood pressure (trestbps), cholesterol level (chol), fasting blood sugar (fbs), maximum heart rate achieved (thalach), exercise-induced angina (exang), oldpeak, slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), thalassemia (thal), and the target label designating the presence or absence of heart disease.
2. Data Pre-processing:
   Missing Values: Handle any missing values appropriately, either by removing rows/columns or imputing values based on statistical methods. Categorical Encoding: Convert categorical features into numerical values using techniques like one-hot encoding.
3. Train-Test Split:
   Split the dataset into training and testing sets, typically with a ratio of 80:20, to assess the model's effectiveness using hypothetical data.
4. Model Selection:
   Choose multiple machine learning algorithms to train on the dataset. In this project, the following algorithms were used:Random Forest Classifier, Support Vector Machine (SVM), Logistic Regression, Gradient Boosting Classifier, Decision Tree Classifier.
5. Pipeline Creation:

For every model, develop a pipeline that comprises the classifier and pre-processing stages (such as scaling and encoding).

6. Model Training and Evaluation:
   Train each model on the training dataset and evaluate its performance on the test dataset using metrics such as accuracy, precision, recall, and F1-score.
7. Model Selection:
   Select the best-performing model based on evaluation metrics and save it for deployment.
8. Prediction:
   Early Intervention: Identifying at-risk individuals early allows for timely interventions that can prevent heart disease or mitigate its effects.
9. Comparison:
   Comparing heart disease prediction models using machine learning involves evaluating the performance of different algorithms on a given dataset. This comparison typically focuses on key aspects such as accuracy, interpretability, computational efficiency, and robustness.

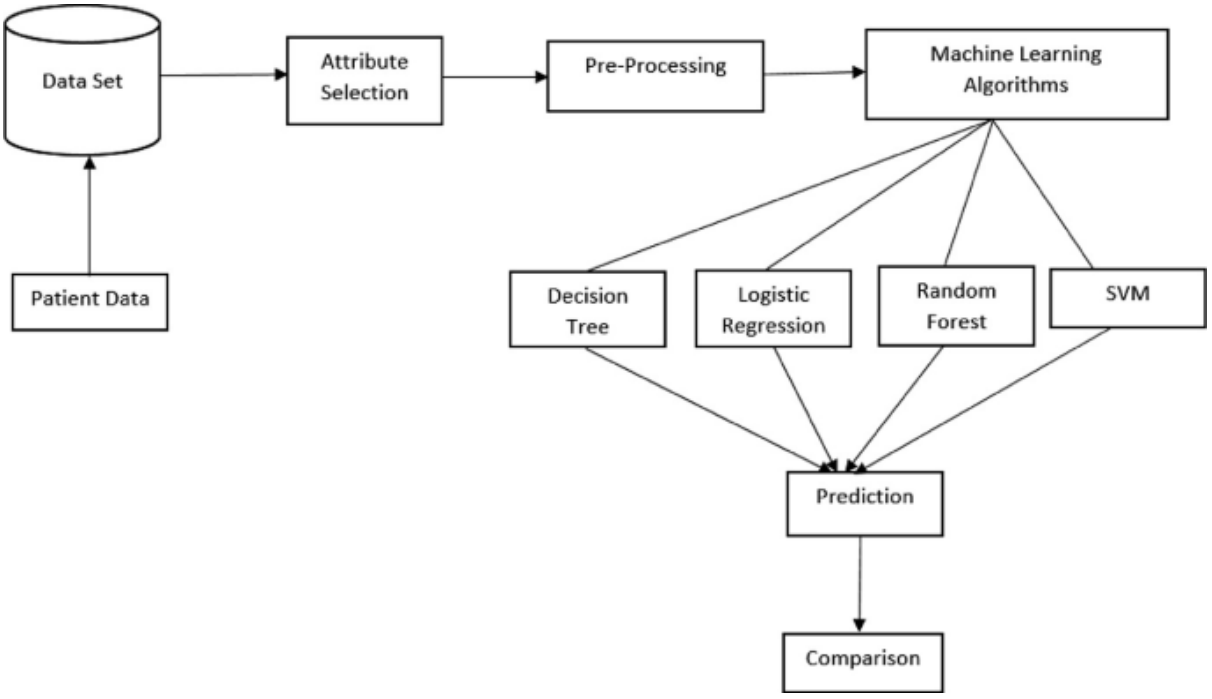

Fig 1: Proposed Architecture of Heart Disease Prediction.

5. RESULTS AND DISCUSSIONS

The classification stage in heart disease prediction refers to the process of categorizing patients based on their risk of having or developing heart disease. This stage is crucial as it directly impacts the decision-making process in healthcare, determining which patients require immediate attention, further testing, or preventive measures.

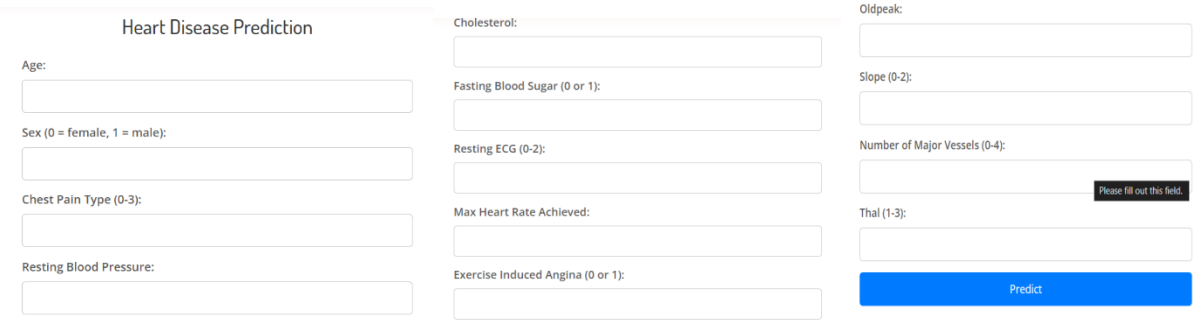

Fig 2: Classification Stages of Heart Disease.

Feature Selection and Extraction:

- Input Features: The model uses various patient data points, such as age, gender, cholesterol levels, blood pressure, smoking status, diabetes, etc.
- Output: Typically, the model classifies patients into two categories:
- Class 0: No heart disease.
- Class 1: Presence of heart disease.

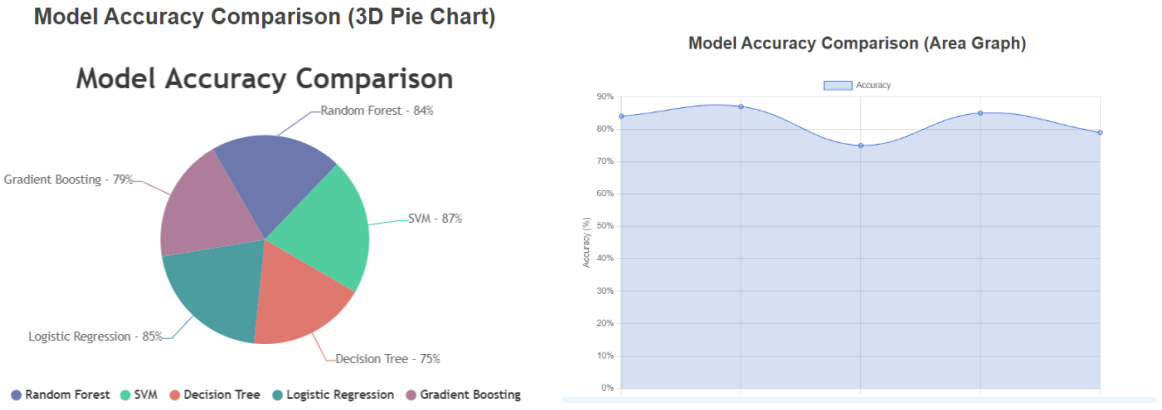

Fig 3: Model Training for Heart Disease Prediction.

Model training in heart disease prediction involves developing a machine learning model that can accurately predict the likelihood of heart disease based on various patient data inputs.

6. CONCLUSION

The Model for Predicting Heart Disease (HDPM) was developed by integrating machine learning algorithms modeling, XGBoost-based prediction, SMOTE-ENN for training dataset balance, and DBSCAN for outlier detection. Two publicly accessible datasets on heart disease were used in our methodology to build a reliable and versatile prediction model. Our HDPM exhibited much higher accuracy, reaching 95.90% and 98.40% for datasets I and II, in that order. in comparison to tests against existing models and prior research. The better performance of our model over the state-of-the-art methods was further supported by statistical studies. Additionally, the HDPM was included into a useful Decision-making system for heart disease: HDCDSS. This system collects diagnostic data and patient data, sends it securely to a web server, and stores it in MongoDB. Through its diagnosis interface, the HDCDSS efficiently determines a patient's cardiac disease state by utilizing the HDPM and producing fast and precise results. The objective The aim of this study is to assist healthcare providers in making clinically sound decisions and improving the care of patients with heart disease. All things considered, the developed HDCDSS and integrated HDPM constitute a noteworthy development in the identification and management of cardiac disease, providing medical practitioners with an invaluable instrument for improving diagnosis precision and decision-making.

Funding:

The authors acknowledge that this research did not receive any financial backing from external agencies, commercial bodies, or research foundations. The project was completed independently.

Conflicts of Interest:

The authors report no conflicts of interest associated with this study.

Acknowledgment:

The authors are thankful to their institutions for their constant moral and professional support throughout this research.

References

[1] Statistics on Causes of Death in 2018, [Online]. Available: http://kostat.go.kr/portal/eng/pressReleases/8/10/index.board?bmode=read==378787.
[2] E. J. Benjamin, et al., "Heart disease and stroke statistics—2017 update: a report from the American Heart Association," Circulation, vol. 135, no. 10, pp. e146–e603, 2017.
[3] N. Dey, et al., "Machine learning techniques for medical diagnosis of heart disease," in *Proc. Int. Conf. Inventive Computing Informatics (ICICI)*, 2017, pp. 440–445. [Online]. Available: IEEE Xplore.
[4] D. Swain, P. Sharma, V. Vakharia, and T. Tanty, "Prediction of coronary artery disease using logistic regression," in Artificial Intelligence for Signal Processing and Wireless Communication, A. Sharma, A. Jain, A. K. Arya, and M. Ram, Eds. Berlin, Germany: De Gruyter, 2022, pp. 149–158. doi: 10.1515/9783110734652-007.

[5] 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.
[6] M. Zubair, J. Kim, and C. Yoon, "An automated ECG beat classification system using convolutional neural networks," in Proc. Int. Conf. IT Convergence and Security (ICITCS), 2016, pp. 1–5. doi: 10.1109/ICITCS.2016.7740310.
[7] Z. C. Lipton, et al., "Learning to diagnose with LSTM recurrent neural networks," arXiv preprint, arXiv:1511.03677, 2015.
[8] S. Ross-Howe and H. R. Tizhoosh, "Atrial Fibrillation Detection Using Deep Features and Convolutional Networks," in Proc. IEEE EMBS Int. Conf. Biomed. Health Informatics (BHI), Chicago, IL, USA, 2019, pp. 1–4. doi: 10.1109/BHI.2019.8834583.
[9] L. Riyaz, M. A. Butt, M. Zaman, and O. Ayob, "Heart Disease Prediction Using Machine Learning Techniques: A Quantitative Review," in International Conference on Innovative Computing and Communications, A. Khanna, D. Gupta, S. Bhattacharyya, A. E. Hassanien, S. Anand, and A. Jaiswal, Eds. Singapore: Springer Singapore, 2022, pp. 81–94.
[10] C. Krittanawong, et al., "Artificial intelligence in precision cardiovascular medicine," *J. Amer. Coll. Cardiol.*, vol. 69, no. 21, pp. 2657–2664, 2017. [Online]. Available: https://www.jacc.org/doi/full/10.1016/j.jacc.2017.03.571.
[11] S. S, "Blood Cell Counting using Image Processing Techniques: A Review," Int. J. Res. Appl. Sci. Eng. Technol., vol. 8, no. 7, pp. 1047–1049, Jul. 2020. doi: 10.22214/ijraset.2020.30406.
[12] B. M. Khrisna, et al., "Novel solution to improve mental health by integrating music and IoT with neural feedback," *J. Comput. Inform. Syst.*, vol. 15, no. 3, pp. 234–239, 2019.
[13] D. E. Kouicem, et al., "Internet of things security: A top-down survey," *Comput. Netw.*, vol. 141, 2018.
[14] S. Papadopoulos, et al., "Community detection in social media: performance and application considerations," *Data Mining Knowl. Discov.*, vol. 34, no. 3, 2020.
[15] B. Mohammad El-Basioni and S. Abd El-Kader, "Laying the Foundations for an IoT Reference Architecture for Agricultural Application Domain," IEEE Access, vol. 8, pp. 190194–190230, 2020. doi: 10.1109/ACCESS.2020.3031634.
[16] M. H. Annaby, S. H. Basha, and Y. M. Fouda, "Defect detection methods using boolean functions and the ϕ-coefficient between bit-plane slices," Optics and Lasers in Engineering, vol. 139, p. 106474, Apr. 2021. doi: 10.1016/j.optlaseng.2020.106474.