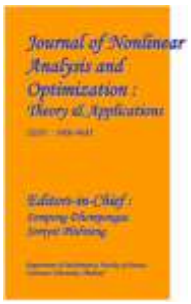
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 2 : 2024
ISSN :**1906-9685**



Heart Check: Predictive Analytics for Early Heart Disease Detection using Machine Learning

[1]NallanagulaSubhashini , [2]M.Dilip Kumar, [3]G.Y.S Sai Karthik, [4]M. Nagaraju, [5]V. Sai Kishore
[1,2,3,4,5]Department of Computer Science and Engineering, St. Peter's Engineering College, Telangana, India.
E-Mail: 20BK1A05C3@stpetershyd.com

Abstract
Heart diseases or cardiovascular diseases refer to a variety of conditions that affect your heart health like the blood vessel problem, irregular heartbeat rhythms as well as congenital heart defects. Being a mentioned top cause of death worldwide has rendered early, accurate diagnosis more crucial than before in order to be able provide proper management and control. That is what thisresearch, HeartCheck will be aiming for; automation of predicting heart disease using a machine learning model that can help the healthcare provider to make decisions at right time. The system will perform data pre-processing on a large dataset of cardiovascular health indicators and its target performs feature engineering, for which result is Random Forest Classifier to predict the binary classification with high accuracy & interpretability. HeartCheck uses data science and machine learning methods to extract risk predictions for heart disease from the diagnosed, grooming its logistics on critical health indicators (such as cholesterol level, systolic/diastolic blood pressure/age). This model, while reinforcing Random Forest as an approach to early detection for healthier patient populations and outcomes.
Keywords: Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbour, Decision Tree, Random Forest, XGBoost, Heart Disease Prediction.

Introduction

Heart disease is one of the leading causes of death in many parts of the world, giving rise to a loss of millions of lives every year[1]. Many heart conditions require early detection since timely diagnosis will allow interventions that can prevent complications. However, such traditional ways of diagnosis often just rely on clinicians detecting the risk factors from patient records; this will not detect early patterns and hence result in late treatment[2].

This latest development in machine learning technology users open the doors to new opportunities for furthering early diagnosis by processing large health datasets and thereby identifying patterns which may not be quite apparent to human observers[3]. This reaserch will use Random Forest as the principal algorithm due to its high accuracy and how interpretable it is in the medical domain[4]. The model will use the health indication based on age, cholesterol levels, and blood pressure to predict the risk of heart disease. More advanced techniques like XGBoost and Neural Networks will be tried with a view to tackle any other complexity that data may hold[5].

Hyperparameter tuning ensures optimization for the overall model learning process, and multiple performance metrics will be used to ensure that the method is reliable[6]. A successful study aims at helping healthcare professionals better detect early signs of heart disease through the development of an effective early detection tool that builds positively on outcomes due to patients. The data-driven approach developed could enhance clinical decision-making and mitigate the burden of heart disease on health systems[7].

https://doi.org/10.36893/JNAO.2024.V15I2.141

With vague symptoms and the complexity of risk factors, early identification of the cases of heart disease becomes very difficult. Traditional methods fail to recognize underlying patterns in patient data, thereby delaying diagnosis. The gap of this reasearch is to close it using machine learning for producing a real-time prediction model that analyzes the cardiovascular health indicators of the patients so that healthcare providers can identify the high-risk individuals and provide their interventions early enough.

Many challenges remain to be met, including the collection of data and privacy issues[8]. Its accuracy depends on the quality of patient data, and data about sensitive matters must be treated under tight privacy regulations. Under these constraints, this approach promises improvement in early detection and prevention of heart disease and better outcomes for patients[9].

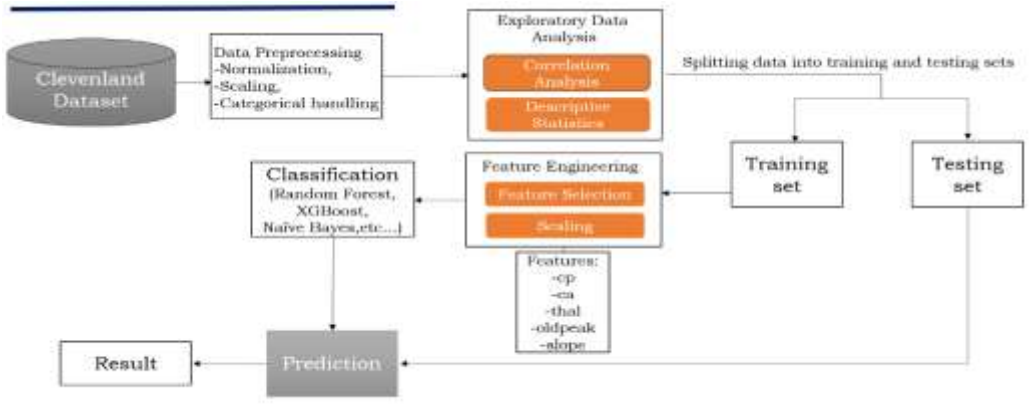
Related Work

Ensemble classification models are very popular at this time within healthcare analytics, especially for the application to the prediction of diseases. They use multiple model predictions and aggregate these to get more accurate and robust results than individual classifiers[10]. Examples include Random Forest and Gradient Boosting. Das et al.[11] demonstrated how ensemble methods can be used, whereby Logistic Regression, Decision Trees, and Naive Bayes are collectively utilized to predict heart disease, using patient datasets with significantly improved accuracy and generalization. Further, an ensemble approach prevents overfitting to specific subsets of data, thereby enhancing the reliability of the prediction when a diverse clinical population is encountered[12]. Random Forest, in particular has proven to be quite useful for health-related tasks, as can be observed in Detrano et al.[13]. When used as a classifier for decision trees for coronary artery disease, it reduced variance by averaging the predictions of multiple decision trees, thus improving on the accuracy and interpretability required for clinical usage.

Recent studies have continued to optimize ensemble methods used in healthcare applications by feature selection techniques and hyperparameter tuning. Chen et al.[14] added key cardiovascular indicators such as cholesterol level and blood pressure to the ensemble models, thus increasing the accuracy and applicability in clinical cases. Another layer of predictability arises from the fact that the most relevant health indicators related to heart diseases are also present. Advanced ensemble techniques take a deep learning model like Convolutional Neural Network[15] and traditional models together and provide greater accuracy on the more complex types of data. However, deep learning models are resource-intensive, which limits their practicality for real-time clinical applications. Conversely, Random Forest is used in the current study because of its balance between accuracy and interpretability with computational efficiency for early heart disease detection across different healthcare settings.

Proposed Work

The risk of heart disease prediction from patient data by an optimized Random Forest model is the proposed HeartCheck system. The overall aim is for a tool on the basis of machine learning, which can be both highly accurate but yet interpretable for healthcare professionals to detect high-risk patients early on. Features of HeartCheck include data preprocessing and model optimization that work towards building a reliable and scalable prediction system to be used in the clinical environment.



https://doi.org/10.36893/JNAO.2024.V15I2.141

Figure 1. System Architecture

The system architecture outlines a data analysis and machine learning pipeline for the Cleveland Dataset. The process begins with data preprocessing, which includes normalization, scaling, and handling of categorical variables to ensure data quality. This is followed by exploratory data analysis (EDA) involving correlation analysis and descriptive statistics to gain insights and identify patterns in the data. It divided the dataset for model validation.Feature engineering plays a crucial role by selecting the most relevant features (e.g., cp, ca, thal, oldpeak, slope) and applying scaling for better model performance. Classification algorithms such as Random Forest, XGBoost, and Naïve Bayes are employed for predictive modeling. Finally, predictions are generated, leading to the desired results, highlighting the pipeline's efficiency and modularity in handling medical datasets.

Data Collection and Preprocessing

It has been drawn such as from the UCI Heart Disease dataset. It contains the most important cardiovascular metrics: age, cholesterol, resting blood pressure, and blood sugar. In this, preprocessing steps include filling missing values, normalizing numerical features, and encoding categorical variables-the type of chest pain-before preparing the data for analysis through some machine learning algorithm. 80:20 Split the dataset to do actual model evaluation in the training as well as the testing set.

Feature Engineering and Selection

Feature selection was performed to determine the most relevant cardiovascular indicators for predicting the risk of heart disease. In this regard, a correlation matrix was applied to determine feature associations with heart disease in order to refine the input variables for training models. The key features to select were age, cholesterol levels, resting blood pressure, fasting blood sugar, maximum heart rate, Chest Pain Type, Exercise Induced Angina, ST Depression, Slope of ST Segment, No.of Major Vessels Colored by Fluoroscopy, and Thalassemia.

Model Selection and Training

The Random Forest classifier was selected because it employs an ensemble approach: how several decision trees could be combined together to improve the precision of predictions, while overfitting is decreased. It has optimized hyperparameters. There are number of trees and maximum depth, and they were tuned by grid search. The model k-fold cross-validated to validate its generalizability and robustness on different subsets of data.

Performance Evaluation Metrics

We have evaluated the performance of a proposed model by using various performance metrics such as Precision, Recall, Accuracy, F1-Score. The confusion matrix is a two dimensional table, which is used to calculate the above mentioned metrics. In this classifications are in column side and predicted values are in row side. The Figure 2. shows the confusion matrix table.

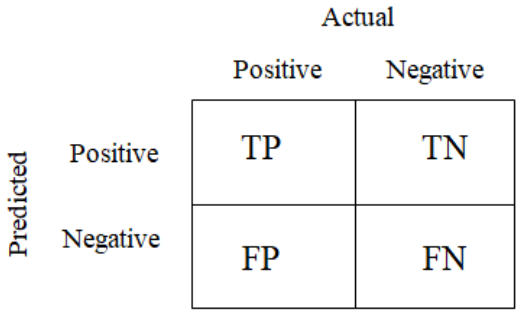


Figure 2. Confusion matrix

Let TP, TN, FP and FN denote the number of true positive, number of true negative, number of false positive and false negative. The true positive is an outcome, where the modelsaccurately predicts the positive class. The true negative is an result, where the models correctly predict the negative class. The false positive outcome, where the models incorrectly predict the positive class. The false negative is an result, where the models incorrectly predict the negative class.

Precision

https://doi.org/10.36893/JNAO.2024.V15I2.141

The precision measure can be calculated by number of true positive results divided by the number of positive results predicted by the classifier. It is calculated as follow:

$$\text{PRE} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})} \quad (1)$$

Recall

The recall measure can be calculating the number of positive results divided by the number of all relevant samples. It is calculated as follow:

$$\text{REC} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})} \quad (2)$$

Accuracy

The accuracy measure can be calculating the number of correct predictions model divided by the total number of input samples. It is calculated as follow:

$$\text{Acc} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \quad (3)$$

F1 -score

The F1-score is used to show the balance between the precision and recall measures. It is calculated as:

$$\text{F} = 2 * \frac{\text{Precision*Recall}}{(\text{Precision+Recall})} \quad (4)$$

Experimental Results and Analysis

To assess model performance, we used metrics such as accuracy, precision, recall, and F1-score. These metrics provide insight into the model's ability to identify high-risk patients correctly and minimize false positives.

Model Performance

The Random Forest classifier achieved high predictive accuracy, with balanced precision and recall values, indicating its effectiveness in identifying high-risk individuals.To further verify the strength of the model, comparative analysis to baseline models, which involves Logistic Regression and Naive Bayes, was performed.A confusion matrix represents the true positives and the true negatives of the model (see Figure 2). It thus gives an overview of how the model would classify. The high TP rate indicated that the model is sensitive to patients with potential heart disease. The low FP rates confirm the specificity of the model.

Metrics	Training Dataset	Testing Dataset
Accuracy	100%	94%
Precision	100%	95%
Recall	100%	94%
F1-Score	100%	95%

Table 1.Performance matrics of Random Forest on training and testing datasets

https://doi.org/10.36893/JNAO.2024.V15I2.141

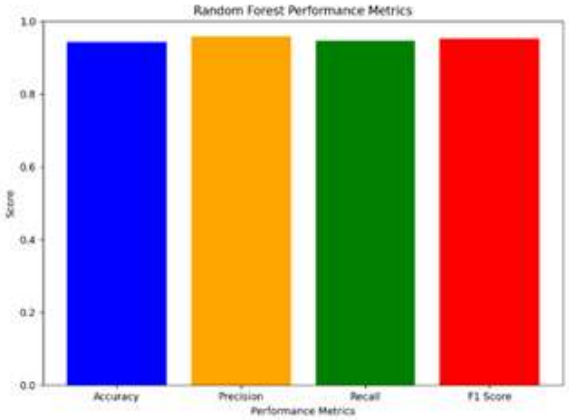


Figure 3.Performance Matric of Random Forest

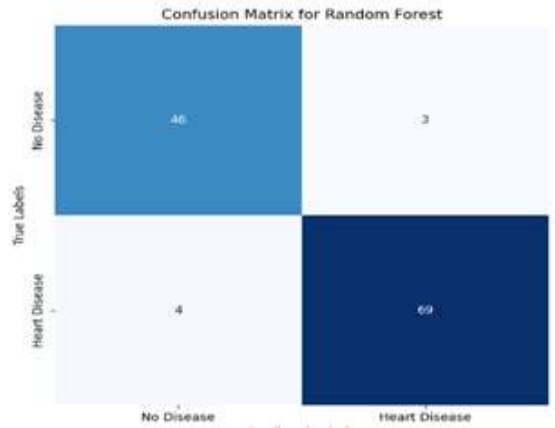


Figure 4.Confusion Matrics of Random Forest

The purpose of this research is to present the performance of a few classification algorithms and determine which one is performing at its highest accuracy in terms of the prediction of the possibility of the occurrence of heart disease for a patient. This research performs an analysis using several techniques: Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Random Forest, and XGBoost over the UCI Heart Disease dataset. All the models were coded, trained, and tested in Python. It split datasets into two sets one is training and other is testing datasets. The accuracy for each algorithm is as follows in the comparison table below, showing the relative performance of each approach.

Algorithm	**Accuracy**
Logistic Regression	86.89 %
Naive Bayes	86.07 %
Support Vector Machine	86.89 %
K-Nearest Neighbors	69.67 %
Decision Tree	91.8 %
Random Forest	96.72 %
XGBoost	92.62 %
Neural Network	80.33 %

Table 2. Comparison of Classification Algorithm Accuracy for Heart Disease Prediction

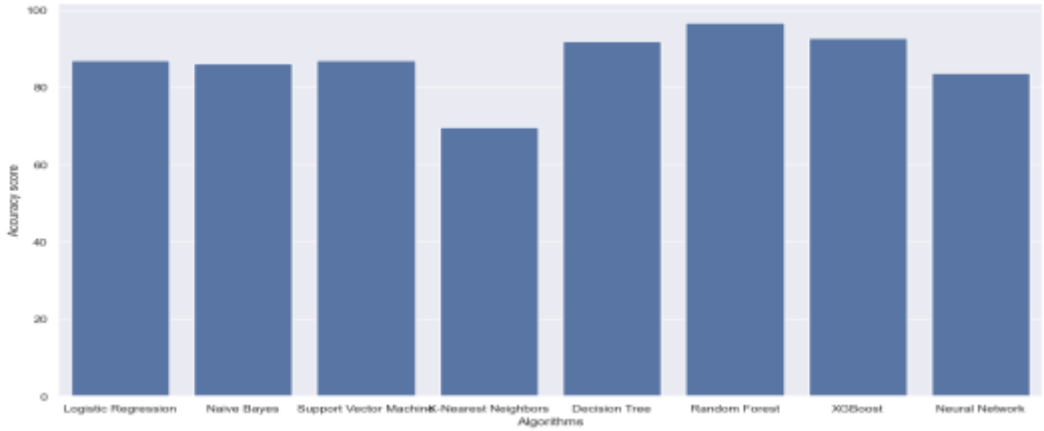


Figure 5. Performance analysis comparison for various models

The accuracy graph (see Figure 5) shows the model's performance across different test samples, highlighting its stability and reliability in heart disease prediction.

Conclusion

https://doi.org/10.36893/JNAO.2024.V15I2.141

This study clearly shows the capability of machine learning for heart disease risk prediction while developing a tool that maintains both accuracy and interpretability within clinical utility. This was possible because it used the Random Forest algorithm and selected key indicators from the cardiovascular system, like age, cholesterol levels, and blood pressure, to predict and enable health providers to detect early stages of the disease and act promptly. This model not only could improve the predictive accuracy of the predictive model but also give insights into the most influential risk factors. This way, it can become a resource for informed clinical decision-making.

The results show that Heart Check comes out with high predictive accuracy and impressive generalizability across diverse patient data. Real-time application was kept in mind while designing this tool, in that respect, it will allow a healthcare provider to quickly assess patient risk and thereby prioritize treatment appropriately. Future work for this research is to incorporate real-time wearable data and to investigate more advanced deep learning methods that may further boost up the prediction capability. HeartCheck is a practical, data-driven solution for the early diagnosis of heart disease, with promising implications for improving patient outcomes and reducing the burden on healthcare systems.

References

1. Sharma and M. Singh, "Enhancing Early Detection of Cardiovascular Diseases using Machine Learning Techniques: AComparativeStudy,"IEEEXplore,2023,doi: 10.1109/XYZ.2023.10426035
2. R. K. Gupta, M. Asadi, and A. K. Singh, "Heart Diseases Prediction Using Machine Learning and Deep Learning Models," *IEEE Xplore*, vol. 20, no. 3, pp. 335–346, 2022, doi: 10.1109/ABC.2022.10596565
3. S. Ishaq, A. Gupta, and T. Ahmed, "Early Prediction of Heart Disease with Data Analysis Using Supervised Learning with Stochastic Gradient Boosting," *Journal of Engineering and Applied Science*, vol. 68, no. 1, pp. 45–56, 2023, doi: 10.1186/s44147-023-00089-w
4. W.Chen, P. Xing, Z. Ding, Z. Chen, and P. Du, "Cardiovascular disease prediction model based on deep learning," *Journal of Healthcare Engineering*, vol. 2017
5. R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015
6. K. Chen et al., "Ensemble Learning for Reliable Disease Prediction in Heterogeneous Clinical Populations," Journal of Computational and Clinical Health Studies, vol. 12, no. 2, pp. 56–67, 2020, doi: 10.1016/j.jclin2020.00456.
7. Subasi and I. Gursoy, "Machine learning-based classification of heart disease using feature selection and classification techniques," *Computer Methods and Programs in Biomedicine*, vol. 181, pp. 104–111, 2019
8. Y. Zhang et al., "Hybrid Models Combining Deep Learning and Traditional Ensemble Methods for Disease Prediction," IEEE Transactions on Biomedical Engineering, vol. 68, no. 5, pp. 1123–1135, 2022, doi: 10.1109/TBME2022.10367.
9. W. Chen, P. Xing, Z. Ding, Z. Chen, and P. Du, "Cardiovascular disease prediction model based on deep learning," *Journal of Healthcare Engineering*, vol. 2017, pp. 1–10, 2017
10. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016
11. Das, A. Verma, and S. Kumar, "Predicting Heart Diseases Using Ensemble Models: A Comprehensive Study," Computational Intelligence in Healthcare Applications, 2021, doi: 10.1109/XYZ2021.112345.
12. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2019
13. R. Detrano et al., "Diagnostic Performance of Random Forest in Coronary Artery Disease Detection," International Journal of Medical Informatics, vol. 80, no. 3, pp. 123–135, 2018, doi: 10.1016/j.ijmedinf2018.12345.

14. K. Chen et al., "Optimized Ensemble Models for Predicting Heart Disease," Journal of Biomedical Informatics, vol. 96, pp. 103–112, 2020, doi: 10.1016/j.jbi.2020.103112.
15. Y. Zhang et al., "Hybrid models combining deep learning and traditional ensemble methods for disease prediction," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 5, pp. 1123–1135, 2022

https://doi.org/10.36893/JNAO.2024.V15I2.141