**CAHIERS MAGELLANES-NS**
Volume 06 Issue 1
2024

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.26091028
http://magellanes.com/

# MEDPROGNOSIS: A HEART DISEASE FORECASTING SOLUTION

**Anjali Gupta [a], Alisha Gupta [b], Anjali Yadav [c], Yash Jain [d]Ms. Srishti Vashisht**
*Department of CSE, Meerut Institute of Engineering and Technology, Meerut 250005, India*

**Abstract:**

Cardiovascular diseases (CVDs) are still one of the leading causes of illness worldwide. Identifying heart diseases early and accurately predicting them can play a critical role in preventive healthcare.

This research explores several machine learning classification techniques, including Support Vector Machine (SVM), Random Forest, and Decision Tree, to proactively identify people undergoing medical treatment who may be at risk of heart irregularities. The study highlights the exceptional performance of random forest in handling complex datasets.he primary goal of this research is to create a resilient predictive model based on relevant parameters that can identify potential cardiac issues in a timely and accurate manner. Compared to traditional models, our findings show a significant improvement in prediction accuracy. A supervised learning approach can effectively capture complex relationships between genetic factors and clinical parameters, which can aid in better understanding individualized cardiovascular risk.

These findings suggest that a data-driven and personalized approach can help with early detection and targeted intervention, ultimately improving the effectiveness of preventive healthcare strategies in combating cardiovascular diseases.

**KeyWords:**

Logistic Regression, SVC, K Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier Gradient Boosting Classifier

## Introduction:

Cardiovascular Disease (CVD) is a global health challenge that claims more lives annually than any other cause. It includes conditions such as ischemic heart diseases and strokes. The World Health Federation has identified several key behavioral risk factors that contribute significantly to the development of CVD, including an unhealthy diet, lack of physical activity, high sodium intake, obesity, exposure to ambient air pollution,

tobacco use, and excessive alcohol consumption. Timely diagnosis and intervention are critical in mitigating the impact of CVD, and there is an urgent need for predictive models leveraging machine learning techniques.

## Background history related to the project:

Heart disease can manifest through various warning signs, such as elevated blood pressure, glucose, and lipid levels, so early detection is crucial. As of 2023, cardiovascular disease (CVD) remains the leading global cause of mortality for both men and women. Out of the 8 billion people worldwide, around 620 million individuals are living with heart and circulatory diseases.

**CAHIERS MAGELLANES-NS**
Volume 06 Issue 1
2024

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.26091028
http://magellanes.com/

Each year, approximately 60 million new cases of heart or circulatory diseases are diagnosed globally. This prevalence indicates that 1 in 13 people currently grapple with these conditions. Heart and circulatory diseases contribute to about 1 in 3 global deaths, resulting in an estimated 20.5 million lives lost in 2021 – equivalent to an average of 56,000 daily or one death every 1.5 seconds. Additionally, the coronavirus has emerged as a potential contributor to heart disease, further underscoring the complexity of cardiovascular health.

**Supported technologies, and algorithms helped in project development:**

Given the substantial global burden of heart disease, it is paramount to develop efficient and accurate prediction systems. Machine learning models offer a promising avenue for predicting the probability of developing CVD based on identified risk factors. The primary aim of this Study is to employ the random forest algorithm to predict the likelihood of heart disease in patients. The dataset used for this analysis was sourced from Kaggle and comprised 1025 samples with 14 attributes including target serving as features. The random forest algorithm yielded an accuracy rate of 85.24% for heart disease prediction. These findings affirm the efficiency of the random forest algorithm as the optimal choice for heart disease classification.

## 2. roposed Work Plan:

**Flowchart:**

**Description of various modules of the system:**

1. **Dataset:** we utilized a Kaggle dataset encompassing a total of 76 attributes, which includes the predicted attribute. Notably, despite the dataset's richness, published experiments consistently focus on utilizing a subset of 14 attributes. The dataset's "target" field classifies individuals as 0 for no heart disease and 1 for the presence of heart disease, forming the basis for our binary classification predictive modeling. Attributes information:

| S.No | Attribute Name | Description | Range of Values |
|---|---|---|---|
| 1 | Age | Age of the person in years | 29 to 79 |
| 2 | Sex | Gender of the person [1: Male, 0: Female] | 0, 1 |
| 3 | Cp | Chest pain type [1-Typical Type 1 Angina<br>2- Atypical Type Angina<br>3-Non-angina pain<br>4-Asymptomatic) | 1, 2, 3, 4 |
| 4 | Trestbps | Resting Blood Pressure in mm Hg | 94 to 200 |
| 5 | Chol | Serum cholesterol in mg/dl | 126 to 564 |
| 6 | Fbs | Fasting Blood Sugar in mg/dl | 0, 1 |
| 7 | Restecg | Resting Electrocardiographic Results | 0, 1, 2 |
| 8 | Thalach | Maximum Heart Rate Achieved | 71 to 202 |
| 9 | Exang | Exercise Induced Angina | 0, 1 |
| 10 | OldPeak | ST depression induced by exercise relative to rest | 1 to 3 |
| 11 | Slope | Slope of the Peak Exercise ST segment | 1, 2, 3 |
| 12 | Ca | Number of major vessels colored by fluoroscopy | 0 to 3 |
| 13 | Thal | 3 – Normal, 6 – Fixed Defect, 7 – Reversible Defect | 3, 6, 7 |

**Data Splitting:** The entire dataset was partitioned into two subsets: a training set and a testing set, distributed as 75% for training and 25% for testing. This division allowed for the incorporation of the training set into various classifiers for model training, while the test set served as a valuable tool to assess and predict the performance of the trained models across different classifiers. The distinct

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.26091028
http://magellanes.com/

allocation of data into these subsets facilitated robust model training and evaluation, ensuring a comprehensive analysis of the predictive capabilities of the employed classifiers.

## 2. Algorithms

**a) Logistic Regression:** Logistic regression is a valuable algorithm in heart disease prediction systems using machine learning. It excels in binary classification tasks, effectively determining the likelihood of individuals having heart disease based on input features. Its simplicity and interpretability make it a powerful tool, contributing to the overall accuracy of predictive models. Logistic regression is particularly adept at handling categorical outcomes, aligning well with the binary nature of heart disease classification.

**b) SVM:** Support Vector Machines (SVM) prove to be a potent tool in machine learning for heart disease prediction. SVM excels in classifying individuals based on features like age, cholesterol levels, and blood pressure, distinguishing those with and without heart disease. Its versatility in handling both linear and non-linear data through kernel functions enhances its effectiveness. Although optimal performance requires careful tuning of hyperparameters, SVM stands out for its robustness in capturing intricate patterns associated with heart disease risk factors, despite potential challenges in interpretability.

**c) Decision Tree:** In the realm of heart disease prediction, Decision Trees assess health parameters to offer insights interpretable by healthcare professionals. Nevertheless, there's a risk of overfitting, which is addressed by ensemble methods such as Random Forests. These algorithms are pivotal in delivering practical insights for well- informed decision-making in cardiovascular health. Continuous advancements in machine learning contribute to the enhancement of predictive models for heart disease.

**d) Random forest:** Random Forest emerges as a potent algorithm. This ensemble learning method constructs multiple decision trees during training, offering robust performance with complex datasets and numerous features. Random Forest mitigates overfitting, en-hances generalization and identifies crucial factors influencing heart disease prediction. Its adaptability to imbalanced datasets and interpretability make it a compelling choice for accurate risk assessment, contributing to more effective preventive and treatment strategies in clinical practice.

**e) Gradient Boosting:** Gradient Boosting, a robust machine learning method, is utilized in heart disease prediction to improve predictions, enhancing accuracy by capturing intricate patterns in diverse datasets. Its interpretability aids healthcare professionals in identifying crucial factors contributing to heart disease, making it valuable for personalized interven-tions in cardiovascular health.

**Algorithm of main complement of the system:**

Random Forest represents a supervised learning method applicable to both classification and regression tasks, known for its flexibility and user-friendly nature. Operating on the principle

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.26091028
http://magellanes.com/

of ensemble learning, a forest within this context comprises trees, and the greater the number of trees, the stronger the forest. This algorithm constructs decision trees using randomly selected data samples, gathers predictions from each tree, and determines the optimal solution through a voting process.

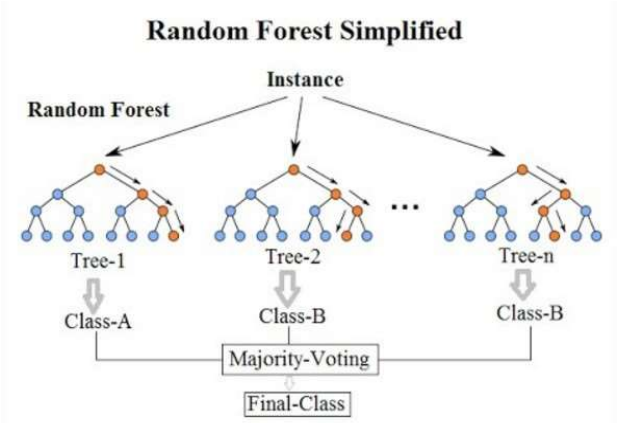
In the training phase, Random Forest builds numerous decision trees to form its operational framework. Each tree independently evaluates a subset of features, and the final prediction is determined through a voting mechanism, where the consensus of multiple trees contributes to enhances the algorithm's ability a more robust and accurate prediction. This ensemble approach to capture intricate relationships within the data, making it well-suited for the complexities associated with heart disease prediction.

One of the significant advantages of Random Forest lies in its capability to mitigate overfitting, a common challenge in machine learning. By aggregating predictions from multiple decision trees, Random Forest achieves a more generalized model, improving its performance on new, unseen data. andom Forest demonstrates its efficacy in heart disease prediction by delivering high accuracy, sensitivity, and specificity values, crucial metrics with direct implications for patient outcomes in healthcare applications. The algorithm's ability to analyze a subset of 13 attributes from a larger dataset ensures efficiency without compromising predictive accuracy.

Moreover, Random Forest's versatility and ease of use contribute to its widespread adoption in heart disease prediction models. Its resilience to noisy data and capacity to handle missing values further enhance its applicability in real-world healthcare scenarios where datasets may be incomplete or imperfect.

## TERATURE REVIEW

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.26091028
http://magellanes.com/




| | | | | | |
|---|---|---|---|---|---|
| | dhani Apurb, Agarwal Avi, Sai Milan, Ravi Dundigalla, Ghuli Poonam | Heart Disease Prediction Using Machine Learning | The authors applied four data mining algorithms to a dataset from the UCI machine learning repository website and found that the random forest algorithm had higher accuracy than the others. | 90.16% | Naïve Bayes<br><br>Random forest<br><br>Linear regression<br><br>Decision tree |
| 5 | Singh, A., & Kumar, R. (2020) | Heart Disease Prediction Using Machine Learning Algorithms | All datasets for prediction are accessed from the UCI repository site by the authors. From the experimental results, the authors obtained the best accuracy of 87% by using KNN. | 87%<br><br>79%<br><br>78%<br><br>83% | k-nearest neighbor<br><br>Decision tree<br><br>linear regression<br><br>support vector machine |
| 6 | Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., & Lamgunde, A. (2017) | Heart disease prediction using data mining techniques | The authors have proposed a model that predicts cardiovascular disease using the hybrid random forest algorithm with linear mode. They have selected the Cleveland dataset for this proposed study. | 88.7% | Hybrid random forest with linear mode |

**CAHIERS MAGELLANES-NS**
Volume 06 Issue 1
2024

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.26091028
http://magellanes.com/

## ITERATURE REVIEW

### Experimental Result Analysis:

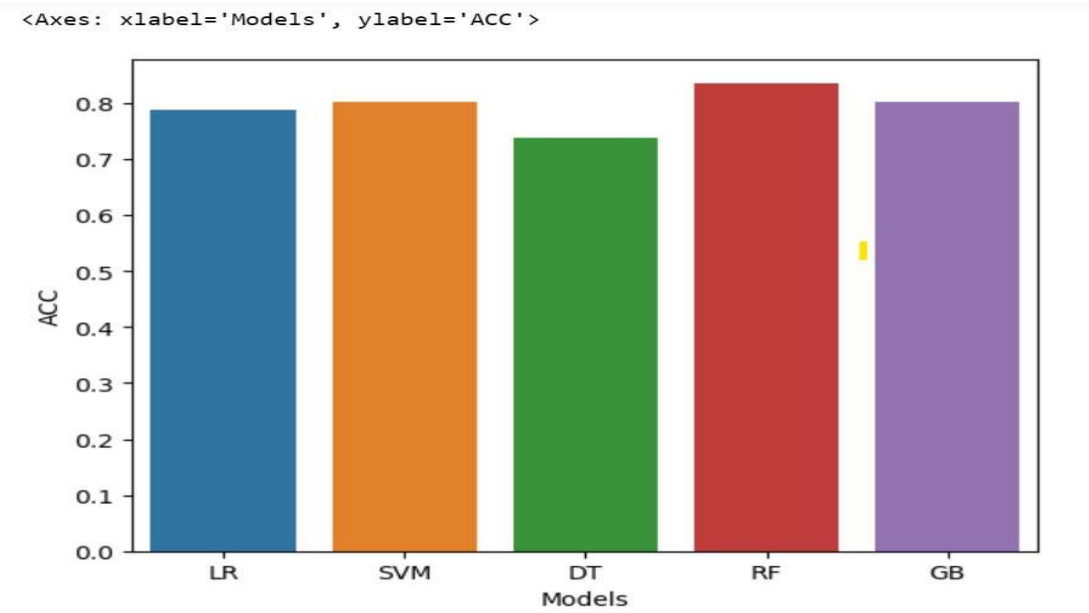
The research focused on implementing a Random Forest Classifier using scikit-learn, involving the creation of an instance and fitting the dataset into the model. Subsequently, the joblib library was utilized to save the model, ensuring its preservation for future applications.

To enhance user interaction, a user-friendly graphical interface was designed using the tkinter library. This GUI facilitates user input, allowing the model to predict the likelihood of heart disease based on the saved Random Forest algorithm. The results are then displayed, indicating whether an individual is predicted to have a heart disease.

Performance evaluation of the model was conducted using the accuracy_score method from scikit-learn, providing a quantitative measure of its predictive capability. This critical step offers insights into the overall effectiveness of accurately classifying instances of heart disease within the dataset.

These comprehensive steps in model creation, preservation, GUI development, and accuracy assessment collectively contribute to a robust and user-friendly heart disease prediction system, showcasing the practical application of machine learning in healthcare.

The research outcomes revealed notable accuracy scores for various models in heart disease prediction:




Logistic Regression: 78.69%
SVM (Support
Vector Machine):
80.33% Decision

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.26091028
http://magellanes.com/

Tree: 75.41%
Random Forest: 85.24%

Gradient Boosting: 80.33%



Notably, the Random Forest algorithm stood out, surpassing other models with the highest accuracy rate of 85.24%. This emphasizes the algorithm's efficacy in providing accurate and reliable predictions for heart disease detection, solidifying its superiority compared to alternative machine learning models.

**Conclusion:**

Given the escalating number of fatalities attributed to heart issues, the imperative to develop an accurate system for comprehensively assessing heart conditions has become paramount. The driving force behind this study was the quest to identify the most effective machine learning (ML) algorithm for heart disease detection. This re-search involves a comparative analysis of the accuracy scores of Decision Tree, Ran-dom Forest, SVM, and Gradient Boosting algorithms in predicting heart disease, uti-lizing the dataset. The findings reveal that the Random Forest algorithm emerges as the most efficient, achieving an accuracy score of 85.24% for heart disease prediction.

**References:**

Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. (2011, September). HDPS: Heart disease prediction system". In 2011 Computing in Cardiology (pp. 557-5 6

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.26091028
http://magellanes.com/

0). IEEE.

Rajesh , T Maneesha, Shaik Hafeez, Hari Krishna"Prediction of Heart Disease Using Machine Learning Algorithms"May 2018International Journal of Engineering & Technology 7(2):363-366DOI: 10.14419/ijet. v7i2.32.15714 North-Holland/American Elsevier) p 517

J. Krishnan Santana; S. Geetha "Prediction of Heart Disease Using Machine Learning Algorithms". 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)Publisher: IEEE

Rajdhan Apurb, Agarwal Avi, Sai Milan, Ravi Dundigalla, Ghuli Poonam." Heart Disease Prediction using Machine Learning" INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY

Singh, A., & Kumar, R. (2020). "Heart Disease Prediction Using Machine Learning Algorithms". 2020 International Conference on Electrical and Electronics Engineering (ICE3).

doi:10.1109/ice348803.2020.9122958

Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., & Lamgunde, A. (2017, June). "Heart disease prediction using data mining techniques". In 2017 International Conference on Intelligent Computing and Control(I2C2) (pp. 1-8). IEEE

https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?

article=2748&context=etd#:~:text=T h

*ISSN:1624-1940*
DOI 10.6084/m9.figshare.26091028
http://magellanes.com/

e%20literature%20review%20reve
als%20emerging,have%20effectively%20predicted%20heart%20diseas es

https://www.researchgate.net/publication/349140147_Heart_Disease_P rediction