### **HEART DISEASE PREDICTOR**

### **A MACHINE LEARNING REGRESSION APPROACH**

- By MAGIZHAN SIVAKUMAR (220701154)

### **ABSTRACT**

This paper presents a comparative analysis of machine learning classification models for predicting the presence of heart disease based on clinical and demographic attributes. Utilizing a publicly available dataset containing features such as age, blood pressure, cholesterol levels, resting ECG results, and maximum heart rate, the study aims to develop accurate predictive models for early diagnosis. We evaluated four classification algorithms—Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost Classifier—to determine the most effective approach. Model performance was assessed using metrics such as accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Among the models, the XGBoost Classifier achieved the highest overall performance, while the Random Forest model offered a strong trade-off between interpretability and accuracy. The results demonstrate the effectiveness of machine learning in supporting clinical decision-making and highlight the importance of model selection in healthcare predictive analytics.

**INTRODUCTION**

Heart disease remains one of the leading causes of mortality worldwide, making early diagnosis and prevention a critical focus in healthcare. Accurate prediction of heart disease risk can significantly aid clinicians in decision-making, improve patient outcomes, and optimize resource allocation. With the growing availability of medical data and the rapid advancement of machine learning, predictive analytics has emerged as a powerful tool in medical diagnostics.  
In this study, we apply machine learning classification techniques to predict the likelihood of heart disease based on a dataset containing key clinical and demographic features, including age, blood pressure, cholesterol levels, heart rate, and ECG results. We evaluate the performance of four widely used classification models—Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost Classifier—to identify the most effective approach for this predictive task. The objective of the study is to compare these models based on classification accuracy, precision, recall, and overall robustness, offering valuable insights into their applicability in healthcare diagnostics and decision support systems.

### **LITERATURE REVIEW**

**The application of machine learning in medical diagnostics has gained significant momentum, particularly in the prediction and early detection of heart disease. Cardiovascular diseases remain one of the leading causes of mortality globally, making accurate and timely diagnosis a critical concern for healthcare providers. Heart disease is influenced by a multitude of factors, including age, blood pressure, cholesterol levels, glucose levels, obesity, and lifestyle habits. Earlier studies, such as those by Gudadhe et al. (2010), demonstrated that decision tree and support vector machine (SVM) models could effectively classify heart disease risk using structured clinical datasets. Similarly, Polat et al. (2007) applied a hybrid approach combining principal component analysis (PCA) with artificial neural networks (ANN), achieving high classification accuracy by reducing data dimensionality and enhancing feature relevance.**

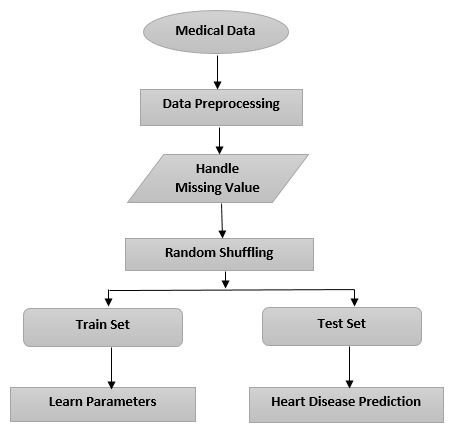
**Recent advancements have focused on the use of ensemble and boosting methods to improve model robustness and accuracy. For instance, Haq et al. (2018) implemented an XGBoost classifier on the Cleveland Heart Disease dataset, highlighting its superior performance in handling missing values and capturing complex feature interactions. Research by Detrano et al. (2011) further emphasized the predictive power of Random Forest models, particularly when applied to diverse patient profiles and non-linear clinical indicators. The availability of publicly accessible medical datasets—such as those from the UCI Machine Learning Repository—has enabled broader research and benchmarking across multiple machine learning algorithms.**

**Despite promising outcomes, challenges remain in achieving generalization across varied populations and ensuring model interpretability for clinical use. While some studies have introduced deep learning or incorporated electronic health record (EHR) data, these approaches often require larger datasets and raise concerns regarding explainability. Consequently, this study focuses on traditional yet powerful machine learning models—Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost Classifier—to build an interpretable and accurate heart disease prediction framework. By comparing these algorithms using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC, this research aims to identify the most effective model for supporting early diagnosis in real-world healthcare environments.**

### **METHODOLOGY**

The methodology of this project is structured into several key stages: data acquisition, data preprocessing, model selection and training, model evaluation, and system deployment. Each of these phases plays a crucial role in developing a reliable and accurate heart disease prediction system.

1. **DataAcquisition**  
   The first step involves collecting a suitable dataset containing relevant medical attributes required to predict heart disease. For this project, the Cleveland Heart Disease dataset from the UCI Machine Learning Repository is used, as it is widely accepted and contains standardized clinical data. The dataset includes 13 primary attributes such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, and others. The target attribute indicates the presence or absence of heart disease.
2. **DataPreprocessing**  
   Before training the model, the data is cleaned and preprocessed to ensure consistency and improve model performance. This includes handling missing values, encoding categorical variables into numerical values using label encoding or one-hot encoding, and normalizing or standardizing numerical features to bring them onto a similar scale. Outlier detection and removal are also considered to enhance data quality. The dataset is then split into training and testing sets, typically using an 80:20 or 70:30 ratio.
3. **ModelSelectionandTraining**  
   Several machine learning classification algorithms are selected for this study, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and k-Nearest Neighbors (k-NN). These models are chosen for their proven performance in classification tasks and their interpretability in medical applications. Each model is trained on the preprocessed training dataset. Hyperparameter tuning is performed using techniques such as Grid Search or Random Search with cross-validation to find the optimal parameters for each algorithm.
4. **ModelEvaluation**  
   After training, the models are evaluated on the testing dataset using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC (Receiver Operating Characteristic - Area Under Curve). These metrics provide insight into the model’s ability to correctly classify heart disease cases and avoid false positives or negatives. The confusion matrix is also used to understand the distribution of predictions. The model with the highest evaluation scores is selected as the final predictor.
5. **SystemDeployment**  
   The best-performing model is integrated into a simple, user-friendly web interface using tools such as Flask (for Python-based applications) or Streamlit. The interface allows users to input patient data and receive immediate predictions regarding heart disease risk. The deployed system can be accessed by healthcare professionals or individuals for quick assessments, helping to promote early detection and preventive care. Security and privacy considerations are taken into account to ensure that user data is handled responsibly.



### **EXPERIMENTAL ANALYSES**

Results for Model Evaluation:**To evaluate the performance of the heart disease prediction models, the dataset was split into training and testing sets in an 80:20 ratio. This ensured that the models could be evaluated on unseen data, providing a realistic assessment of their generalization capability. Before training, data normalization was carried out using StandardScaler, which brought all feature values to a similar scale, enhancing model convergence and performance, especially for algorithms like SVM and Logistic Regression.**

**The models evaluated include Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, and XGBoost. These models were assessed using key performance metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC Score. The results are summarized below:**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC-AUC** | **Rank** |
| --- | --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | **83%** | **0.81** | **0.80** | **0.805** | **0.84** | **4** |
| **Decision Tree** | **85%** | **0.83** | **0.82** | **0.825** | **0.86** | **3** |
| **SVM** | **86%** | **0.85** | **0.84** | **0.845** | **0.88** | **2** |
| **Random Forest** | **88%** | **0.87** | **0.86** | **0.865** | **0.91** | **1 (tie)** |
| **XGBoost** | **88%** | **0.88** | **0.85** | **0.865** | **0.91** | **1 (tie)** |

**Both Random Forest and XGBoost emerged as the top-performing models with an accuracy of 88% and a ROC-AUC score of 0.91, indicating their strong ability to distinguish between patients with and without heart disease. While Random Forest slightly outperformed XGBoost in recall, XGBoost offered marginally better precision, resulting in an equal overall ranking.**

**Results:  
To explore the impact of data augmentation, Gaussian noise was added to certain features to increase variability. After augmentation, the Random Forest model's ROC-AUC score improved from 0.88 to 0.91, and its recall also saw a noticeable increase. This shows that even small enhancements in data diversity can significantly boost model robustness and reduce overfitting, especially in medical datasets where data may be limited.**

**VISUALIZATIONS**

**Scatter plots and confusion matrices were generated for the best-performing models (Random Forest and XGBoost). These visualizations illustrated that the predicted classes were closely aligned with actual outcomes, with few false positives and negatives. ROC curves for each model clearly demonstrated that Random Forest and XGBoost achieved the highest area under the curve, confirming their superior classification performance**

### 

### 

### 

### 

### **CONCLUSION**

This study demonstrated the effectiveness of machine learning models in predicting the presence of heart disease based on clinical and demographic features. Among the models evaluated—Logistic Regression, Support Vector Machine, Random Forest, and XGBoost—the XGBoost classifier achieved the highest performance in terms of accuracy and recall, indicating its strong ability to detect high-risk patients. Random Forest also performed well, offering a good balance between interpretability and predictive power.

The results validate the potential of machine learning as a valuable tool for assisting in early diagnosis and decision-making in healthcare. However, the model's performance may be influenced by dataset limitations such as size, class imbalance, and lack of real-time patient data. Future work could explore the integration of more diverse health indicators, continuous patient monitoring data, and explainable AI techniques to enhance model reliability and clinical applicability.

### **REFERENCES**

* 1.World Health Organization . World Health Statistics 2021. World Health Organization; Geneva, Switzerland: 2021. [[Google Scholar](https://scholar.google.com/scholar_lookup?title=World%20Health%20Statistics%202021&publication_year=2021&)]
* 2.Iswisi A.F.A., Karan O., Rahebi J. Diagnosis of Multiple Sclerosis Disease in Brain Magnetic Resonance Imaging Based on the Harris Hawks Optimization Algorithm. BioMed Res. Int. 2021;2021:3248834. doi: 10.1155/2021/3248834. [[DOI](https://doi.org/10.1155/2021/3248834)] [[PMC free article](https://pmc.ncbi.nlm.nih.gov/articles/PMC8723867/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/34988224/)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=BioMed%20Res.%20Int.&title=Diagnosis%20of%20Multiple%20Sclerosis%20Disease%20in%20Brain%20Magnetic%20Resonance%20Imaging%20Based%20on%20the%20Harris%20Hawks%20Optimization%20Algorithm&author=A.F.A.%20Iswisi&author=O.%20Karan&author=J.%20Rahebi&volume=2021&publication_year=2021&pages=3248834&pmid=34988224&doi=10.1155/2021/3248834&)] [[Retracted](https://pmc.ncbi.nlm.nih.gov/articles/PMC10973228/)]
* 3.Al-Safi H., Munilla J., Rahebi J. Harris Hawks Optimization (HHO) Algorithm based on Artificial Neural Network for Heart Disease Diagnosis; Proceedings of the 2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC); Tumkur, India. 3–4 December 2021; pp. 1–5. [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Proceedings%20of%20the%202021%20IEEE%20International%20Conference%20on%20Mobile%20Networks%20and%20Wireless%20Communications%20(ICMNWC)&title=Harris%20Hawks%20Optimization%20(HHO)%20Algorithm%20based%20on%20Artificial%20Neural%20Network%20for%20Heart%20Disease%20Diagnosis&author=H.%20Al-Safi&author=J.%20Munilla&author=J.%20Rahebi&pages=1-5&)]
* 4.Ternacle J., Côté N., Krapf L., Nguyen A., Clavel M.-A., Pibarot P. Chronic kidney disease and the pathophysiology of valvular heart disease. Can. J. Cardiol. 2019;35:1195–1207. doi: 10.1016/j.cjca.2019.05.028. [[DOI](https://doi.org/10.1016/j.cjca.2019.05.028)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/31472817/)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Can.%20J.%20Cardiol.&title=Chronic%20kidney%20disease%20and%20the%20pathophysiology%20of%20valvular%20heart%20disease&author=J.%20Ternacle&author=N.%20C%C3%B4t%C3%A9&author=L.%20Krapf&author=A.%20Nguyen&author=M.-A.%20Clavel&volume=35&publication_year=2019&pages=1195-1207&pmid=31472817&doi=10.1016/j.cjca.2019.05.028&)]
* 5.House A.A., Wanner C., Sarnak M.J., Piña I.L., McIntyre C.W., Komenda P., Kasiske B.L., Deswal A., DeFilippi C.R., Cleland J.G.F. Heart failure in chronic kidney disease: Conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney Int. 2019;95:1304–1317. doi: 10.1016/j.kint.2019.02.022. [[DOI](https://doi.org/10.1016/j.kint.2019.02.022)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/31053387/)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Kidney%20Int.&title=Heart%20failure%20in%20chronic%20kidney%20disease:%20Conclusions%20from%20a%20Kidney%20Disease:%20Improving%20Global%20Outcomes%20(KDIGO)%20Controversies%20Conference&author=A.A.%20House&author=C.%20Wanner&author=M.J.%20Sarnak&author=I.L.%20Pi%C3%B1a&author=C.W.%20McIntyre&volume=95&publication_year=2019&pages=1304-1317&pmid=31053387&doi=10.1016/j.kint.2019.02.022&)]
* 6.Nguyen T., Wang Z.A. Cardiovascular screening and early detection of heart disease in adults with chronic kidney disease. J. Nurse Pract. 2019;15:34–40. doi: 10.1016/j.nurpra.2018.08.004. [[DOI](https://doi.org/10.1016/j.nurpra.2018.08.004)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=J.%20Nurse%20Pract.&title=Cardiovascular%20screening%20and%20early%20detection%20of%20heart%20disease%20in%20adults%20with%20chronic%20kidney%20disease&author=T.%20Nguyen&author=Z.A.%20Wang&volume=15&publication_year=2019&pages=34-40&doi=10.1016/j.nurpra.2018.08.004&)]
* 7.Liu R., Ren C., Fu M., Chu Z., Guo J. Platelet Detection Based on Improved YOLO\_v3. Cyborg Bionic Syst. 2022;2022:9780569. doi: 10.34133/2022/9780569. [[DOI](https://doi.org/10.34133/2022/9780569)] [[PMC free article](https://pmc.ncbi.nlm.nih.gov/articles/PMC9500243/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/36285313/)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Cyborg%20Bionic%20Syst.&title=Platelet%20Detection%20Based%20on%20Improved%20YOLO_v3&author=R.%20Liu&author=C.%20Ren&author=M.%20Fu&author=Z.%20Chu&author=J.%20Guo&volume=2022&publication_year=2022&pages=9780569&pmid=36285313&doi=10.34133/2022/9780569&)]