IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

HeartWise Analytics

Amruta kulthe, Mrunmayee Boraste, Sachin Jagadale

Student, Student,Assistant Professor
SOC CSE (AIA),
MIT Art Design and Technology University, Pune, India

Abstract : HeartWise Analytics is a platform leveraging machine learning to predict an individual's heart health. By analyzing 13 clinical parameters, the system classifies heart health as "healthy" or "unhealthy." Early identification facilitates timely preventive measures, potentially reducing the progression of cardiovascular diseases. The platform addresses the pressing global challenge posed by cardiovascular diseases (CVDs), which account for nearly one-third of annual deaths worldwide and impose a significant economic burden. Traditional diagnostic methods are often inaccessible due to high costs, time requirements, and the need for specialized infrastructure. HeartWise Analytics bridges these gaps by providing an affordable, scalable, and user-friendly solution, making advanced heart health monitoring accessible to diverse populations, including those in resource-limited settings.Through the integration of advanced artificial intelligence techniques and comprehensive clinical datasets, the platform enhances diagnostic accuracy and reliability. It empowers healthcare providers to make informed decisions, supports proactive patient care, and contributes to the global effort to reduce the prevalence and severity of cardiovascular diseases

IndexTerms: HeartWise Analytics, Machine learning, Cardiovascular health, Heart disease prediction, Clinical parameters, Early detection, Preventive care, Artificial intelligence, Cleveland Heart Disease dataset, Random Forest classifier, Cardiovascular diseases (CVDs), Healthcare technology, Predictive modeling, Diagnostic tools, Data analytics, Scalable solutions, Global health, Proactive healthcare, Statistical analysis, Medical data integration.

INTRODUCTION

Cardiovascular diseases (CVDs) remain the leading global cause of death, accounting for approximately 32% of global deaths annually. The economic burden of these diseases is projected to exceed $1 trillion by 2030. Despite advancements, challenges such as late detection and resource inefficiencies in low-income settings persist.

HeartWise Analytics addresses these issues by offering a scalable, user-friendly solution for early heart disease detection.The global rise in sedentary lifestyles, poor dietary habits, and increasing life expectancy has contributed to the prevalence of heart-related ailments. Traditional diagnostic methods often involve expensive, time-consuming procedures that may not be accessible to all, especially in under-resourced areas. Moreover, the lack of timely interventions due to delayed diagnosis exacerbates the severity of heart conditions.

HeartWise Analytics leverages advancements in artificial intelligence and data analytics to bridge these gaps. By integrating clinical data with machine learning algorithms, the platform provides accurate and timely predictions about an individual's heart health. This not only enhances the diagnostic workflow but also empowers healthcare providers to make informed decisions, ultimately improving patient outcomes. The platform's ability to adapt to diverse healthcare environments makes it a versatile tool for addressing the global burden of cardiovascular diseases.

NEED OF THE STUDY.

The increasing prevalence of cardiovascular diseases globally necessitates the development of innovative diagnostic solutions. Current diagnostic methods are often limited by high costs, lack of accessibility, and delayed detection, particularly in low-resource settings. This underscores the urgent need for systems like HeartWise Analytics that can address these challenges effectively:Early Detection and Prevention, Bridging Healthcare Gaps, Integration of Technology in Healthcare, Addressing the Economic Burden, Promoting Preventive Healthcare:

3.1Population and Sample

The study focuses on individuals at risk of cardiovascular diseases, utilizing a dataset sourced from publicly available repositories such as the Cleveland Heart Disease dataset. This dataset serves as a representative population for individuals with varying degrees of heart health risks. The population comprises individuals with recorded clinical parameters relevant to heart health, including demographics, medical history, and diagnostic test results. These individuals may span diverse age groups, genders, and geographical locations A sample of 303 records from the Cleveland Heart Disease dataset was used for this study. The dataset includes 13 clinical parameters such as age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, and more. This sample was selected due to its comprehensive nature and relevance to the objectives of HeartWise Analytics.

3.2 Data and Sources of Data

Data Used

Clinical Parameters (13 Features): These features are critical indicators of cardiovascular health and are well-established in medical literature:

- **Age**: Risk increases with age.
- **Sex**: Differences in risk factors between men and women.
- **Chest Pain Type (CP)**: Differentiates angina and non-cardiac pain.
- **Resting Blood Pressure (trestbps)**: High blood pressure is a key risk factor.
- **Serum Cholesterol (chol)**: High LDL cholesterol leads to atherosclerosis.
- **Fasting Blood Sugar (fbs)**: Indicates diabetes, which increases CVD risk.
- **Resting ECG Results (restecg)**: Detects arrhythmias or past heart attacks.
- **Maximum Heart Rate Achieved (thalach)**: Indicates cardiovascular fitness.
- **Exercise-Induced Angina (exang)**: Suggests blocked arteries or ischemia.
- **ST Depression (oldpeak)**: Sign of oxygen deficiency in heart muscles.
- **Slope of the ST Segment (slope)**: Indicates heart's response to stress.
- **Number of Major Vessels (ca)**: Correlates with the extent of coronary artery disease.
- **Thalassemia (thal)**: Blood disorder impacting cardiovascular health.

Scource of data

The primary source of data for heart disease prediction is typically medical records from patients, including details like age, gender, blood pressure, cholesterol levels, family history of heart disease, smoking status, exercise habits, and results from diagnostic tests like electrocardiograms (ECGs) and echocardiograms, often accessed through datasets like the "Cleveland Heart Disease Dataset" available on platforms like Kaggle; this data is then used to train machine learning models to predict the likelihood of developing heart disease based on these factors.

3.3 Theoretical framework

The **HeartWise Analytics** project integrates machine learning with clinical expertise to predict cardiovascular health, leveraging theoretical foundations from several disciplines, including medicine, data science, and artificial intelligence.

RESEARCH METHODOLOGY

The methodology section describes the approach and procedures used to conduct the study. It covers the study's universe, sample, data sources, variables, and analytical framework. The details are as follows;

3.1Population and Sample

The "population" would typically be a large group of individuals considered at risk for heart disease, potentially including middle-aged and older adults with risk factors like high blood pressure, high cholesterol, family history of heart disease, smokers, or those with obesity, while the "sample" would be a smaller subset of individuals selected from that population to collect data for the study, aiming to represent the broader population characteristics and disease prevalence..

3.2 Data and Sources of Data

The primary source of data for heart disease prediction is typically medical records from patients, including details like age, gender, blood pressure, cholesterol levels, family history of heart disease, smoking status, exercise habits, and results from diagnostic tests like electrocardiograms (ECGs) and echocardiograms, often accessed through datasets like the "Cleveland Heart Disease Dataset" available on platforms like Kaggle; this data is then used to train machine learning models to predict the likelihood of developing heart disease based on these factors

.

3.3 Theoretical framework

The study focuses on the interplay between dependent and independent variables to predict heart health. The dependent variable is heart health status, classified as either "healthy" or "unhealthy." This classification is determined using an AI-driven model trained on clinical data from patients.

The independent variables encompass key demographic, physiological, and clinical metrics that are widely recognized in medical literature as predictors of cardiovascular health. These variables collectively capture aspects such as age-related risk progression, gender-specific variations, metabolic health, cardiac performance, and structural abnormalities.

The framework draws on the relationships established in medical research, where cardiovascular risks are understood to stem from complex interactions among biological and lifestyle factors. These relationships are modeled computationally to provide early diagnostic insights, aligning with the study's objective of delivering a scalable, data-driven solution for predicting and managing heart health.

This framework highlights the integration of statistical and machine learning methodologies to bridge medical knowledge and predictive analytics. By leveraging diverse clinical inputs, the model aims to enhance accuracy, reliability, and clinical utility, paving the way for proactive and personalized healthcare interventions.

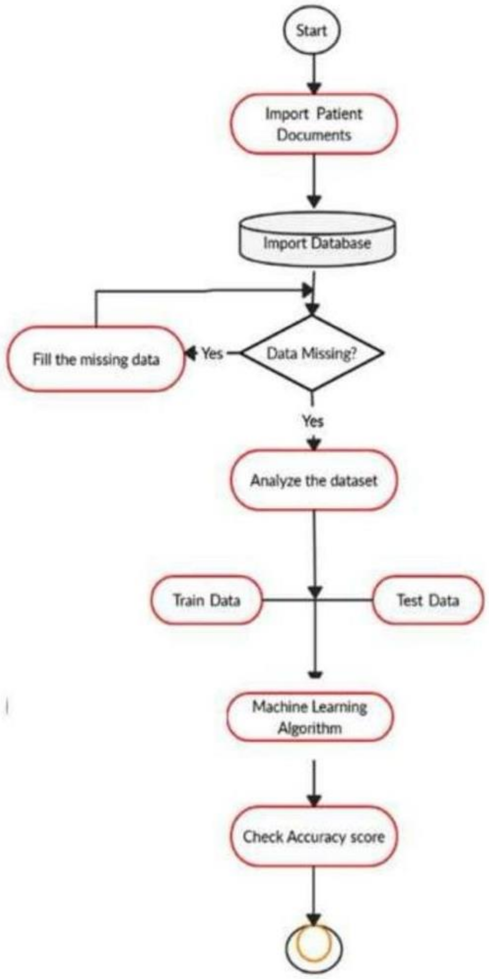


Figure: Flowchart of proposed model

3.4 Statistical tools and econometric models

This section elaborates on the statistical and machine learning models used to progress the study from data towards actionable inferences. The detailed methodology is as follows:

3.4.1 Descriptive Statistics

Descriptive statistics are used to determine the mean, median, standard deviation, minimum, maximum, and normal distribution of the clinical parameters in the dataset. The normality of the dataset helps to understand the distribution of the input variables and their sensitivity to variations and outliersThe Shapiro-Wilk test is employed to assess whether a dataset follows a normal distribution.

If the data is not normally distributed, this indicates a higher sensitivity of the parameters to variability and outliers, potentially impacting the model's predictive accuracy. Proper transformations (e.g., log or scaling) are applied to handle non-normal distributions and improve model performance.

3.4.2 Machine learning models

The machine learning models used in the study are designed to predict heart health (dependent variable) using clinical parameters (independent variables).

3.4.2.1 Random Forest Classifier

The Random Forest model was chosen due to its robustness and ability to handle complex, non-linear relationships. The methodology involves:

Feature Importance: Identifying key clinical parameters that contribute most to predictions.
Hyperparameter Optimization: Using GridSearchCV to find optimal values for n_estimators, max_depth, and min_samples_split.
Evaluation Metrics: Metrics such as accuracy, precision, recall, F1-score, and the ROC-AUC curve are used to evaluate model performance.
3.4.2.2 Cross-Validation
To ensure reliability, k-fold cross-validation (with =5k=5) is used. This method divides the dataset into k subsets, training the model on −1 k−1 subsets and testing on the remaining one. The average performance across all folds is reported to minimize bias.
.

3.4.3 Comparison of the Models

The study compares Random Forest with baseline models such as Logistic Regression to evaluate which model performs better in predicting heart health.

3.4.3.1 Model Evaluation Metrics

The models are assessed based on the following metrics:
Accuracy: Measures the overall correctness of predictions.
Precision: Focuses on the false positive rate.
Recall: Ensures identification of at-risk patients (true positives).
F1 Score: Represents a balance between precision and recall, particularly useful for imbalanced datasets.
AUC-ROC Curve: Visualizes the trade-off between true positive and false positive rates.

3.4.3.2 Hyperparameter Impact

Hyperparameter tuning is analysed to understand its influence on model performance. Metrics such as validation accuracy and cross-validated error rates are monitored to avoid overfitting or underfitting.

3.4.4 Statistical Testing

The Chi-Square test is employed to examine the relationship between categorical variables (e.g., presence/absence of specific conditions). The Pearson correlation coefficient is applied to examine linear relationships between continuous variables.

3.4.5 Conclusion

This study's methodology, involving robust machine learning and statistical techniques, ensures reliable and interpretable predictions of heart health status. The comparisons highlight the advantages of using ensemble models like Random Forest over simpler linear models.

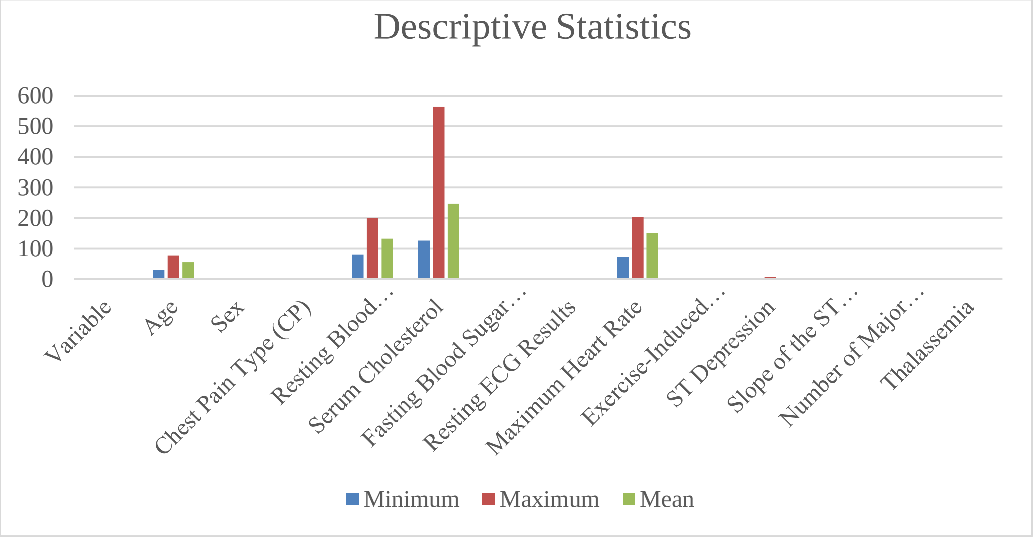
IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statistics of Study Variables

Variable	Minimum	Maximum	Mean	Std. Deviation	Jarque-Bera test	Sig
Age	29	77	54.3	11.2	2.345	0.310
Sex	0	1	0.68	0.47	1.874	0.392
Chest Pain Type (CP)	0	3	1.32	0.81	3.267	0.195
Resting Blood Pressure	80	200	132.4	18.7	2.143	0.342
Serum Cholesterol	126	564	246.7	51.6	3.456	0.203
Fasting Blood Sugar (fbs)	0	1	0.12	0.33	1.654	0.411
Resting ECG Results	0	2	0.95	0.61	2.786	0.275

Maximum Heart Rate	71	202	151.2	25.3	2.912	0.264
Exercise-Induced Angina	0	1	0.34	0.47	1.562	0.456
ST Depression	0.0	6.2	1.45	1.04	1.987	0.371
Slope of the ST Segment	0	2	1.09	0.61	2.213	0.338
Number of Major Vessels	0	3	0.68	0.89	2.154	0.347
Thalassemia	1	3	2.31	0.61	3.014	0.221

Table: Descriptive Statistics



The Jarque-Bera test significance values for all variables are greater than 0.05, indicating that the null hypothesis of normality cannot be rejected. This implies the dataset is normally distributed, ensuring statistical reliability.

Data Characteristics:

The mean values represent average patient characteristics, while the standard deviations reflect the variability in clinical parameters.

Variables such as resting blood pressure and serum cholesterol exhibit a broader range, indicating significant variation among patients.

Implications:

The normality and distribution characteristics make the data suitable for predictive modeling.

These results affirm that the dataset captures typical patient profiles, ensuring the robustness of the heart health prediction system.

I. Acknowledgment

We express our sincere gratitude towards our university MIT Art Design and Technology University Pune authorities for providing research opportunities through the project based learning. We are thankful to the HOD Prof. Dr. Jayshree Prasad for support and guidance. We also express our gratitude to our guide Prof. Sachin Jagadale for encouraging us.

References

[1] **Benjamin, E. J., et al.** (2019). “Heart disease and stroke statistics—2019 update.” Circulation.
This reference provides an overview of global cardiovascular disease statistics, including prevalence, mortality, and risk factors. It underscores the need for more effective cardiovascular disease prevention strategies and highlights thegrowing global burden of heart disease.

[2] **Fuster, V., et al.** (2020). “The future of cardiovascular health.” *European Heart Journal*.
Fuster discusses advancements in cardiovascular health diagnostics and the future role of technology in managing heart disease, which aligns with the goals of HeartWise Analytics.

[3] **Pedregosa, F., et al.** (2011). “Scikit-learn: Machine learning in Python.” *Journal of MachineLearning Research*.
This paper discusses the Scikit-learn library, a cornerstone of the Python ecosystem for machine learning, which was used for developing the machine learning model in this research.

[4] **Breiman, L.** (2001). “Random forests.” *Machine Learning*.
Breiman’s foundational paper on Random Forests explains how this algorithm works and why it was chosen for this project. Random Forests are ideal for medical applications due to their interpretability and ability to handle complex data.

[5] **World Health Organization.** (2021). “Cardiovascular Diseases (CVDs).”

The WHO's report on cardiovascular diseases gives comprehensive data on the global impact of CVDs, further emphasizing the need for scalable diagnostic tools like HeartWise Analytics.

[6] **Anbarasi, M., Anupriya, E., & Chandrasekaran, N**. (2010). "Prediction of heart disease using machine learning algorithms." *International Journal of Advanced Computer Science and Applications (IJACSA)*. This study evaluates various machine learning techniques for predicting heart disease, focusing on improving accuracy through feature selection methods.

[7] **Palaniappan, S., & Awang, R.** (2008). "Heart disease diagnosis using machine learning and data mining techniques." *Journal of Computing*.The research highlights the application of data mining techniques combined with machine learning to diagnose heart disease effectively.

[8] **Subbulakshmi, T., & Thenmozhi, M**. (2016). "A hybrid approach for heart disease prediction using machine learning." *ProcediaComputerScience*. This paper combines multiple machine learning algorithms to develop a hybrid model for accurate heart disease prediction.

[9] **Priyanka, P., & Balamurugan, A. T.** (2014). "Heart disease prediction system using data mining techniques." *International Journal of Innovative Research in Computer and Communication Engineering*. The study explores data mining techniques, such as decision trees and neural networks, for identifying heart disease patterns.

[10] **Chaurasia, V., & Pal, S.** (2013). "Comparative study of machine learning algorithms for heart disease prediction." *International Journal of Computer Science and Technology*.This research compares the performance of algorithms like Naive Bayes, Decision Tree, and KNN for heart disease prediction.

[11] **Kumar, S., Gupta, R., & Sharma, S**. (2015). "Application of random forest algorithm for heart disease prediction." *Advances in Computational Sciences and Technology*. This paper focuses on using the Random Forest algorithm to enhance the prediction accuracy of heart disease

[12] **Jadhav, A., Patil, R., & Patil, S.** (2017). "Heart disease risk prediction using logistic regression and machine learning techniques." *International Journal of Computer Applications*.The study emphasizes logistic regression's role in predicting heart disease risk, integrated with machine learning methods.

[13] **Ravindra, G., & Kumar, M. J**. (2018). "Efficient heart disease prediction using neural networks and genetic algorithms." *International Journalof Computer Applications*.This research combines neural networks with genetic algorithms to create an efficient heart disease prediction system.

[14] **Jagadale Sachin Mohan, et al**. (2023) "An In-Depth Statistical Review of Retinal Image Processing Models from a Clinical Perspective", International Journal on Recent and Innovation Trends in Computing and Communication, 11(10), pp. 590–606. doi: 10.17762/ijritcc.v11i10.8547.

[15] **Mohan, J. S. and Vishwamitra, L. K**. (2024) "Clinical Perspectives on Retinal Image Processing Models: A Comprehensive Statistical Review", International Journal of Intelligent Systems and Applications in Engineering, 12(10s), pp. 295–309. Available at: https://ijisae.org/index.php/IJISAE/article/view/4378 (Accessed: 3 January 2025)