Heart Disease Risk Factor Analysis Using Patient Data

D. Kanaka Satya[1], M. Geetha Priyanka[2], M. Pavan Kumar[3], M. Saranya Dheepthi[4], N. Tejaswi[5]
[1]*Assistant Professor, Srinivasa Institute of Engineering and Technology*
[2,3,4,5]*UG Students, Srinivasa Institute of Engineering and Technology*
doi.org/10.64643/IJIRTV12I11-196366-459

***Abstract*—heart disease is one of the leading causes of mortality worldwide. Early identification of risk factors is essential for preventive healthcare and effective medical decision-making. This study presents a comprehensive analysis of heart disease risk factors using patient medical data through data science techniques and machine learning models.**

A publicly available dataset was collected and pre-processed to ensure data quality. Exploratory Data Analysis (EDA) and statistical correlation techniques were used to identify relationships between key health parameters such as age, blood pressure, cholesterol level, and heart rate. In addition to statistical analysis, multiple machine learning classification models including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, and Naïve Bayes were implemented. Model performance was evaluated using accuracy, precision, recall, and ROC-AUC metrics. Among the models, Random Forest achieved the best performance and was selected as the optimal model. The trained model was integrated into a web-based application that allows users to input patient data and obtain real-time risk predictions.

The results demonstrate that combining data analysis with machine learning provides an effective approach for understanding and predicting heart disease risk.

***Index Terms*—Correlation, Analysis, Data Visualization, Exploratory Data Analysis, Heart Disease, Machine Learning, Patient Data, Prediction, Risk Factor Analysis.**

I. INTRODUCTION

Heart disease is a serious medical condition that affects millions of people globally. Cardiovascular diseases are responsible for a large proportion of global mortality each year. Several clinical and lifestyle factors such as high blood pressure, elevated cholesterol levels, age, physical inactivity, and unhealthy dietary habits contribute significantly to the development of heart disease.

With the rapid growth of healthcare data collection systems, large volumes of patient medical data are now available for analysis. Analyzing this data using data science techniques provides valuable insights into the relationships between health parameters and disease outcomes. Traditional clinical methods rely heavily on manual interpretation of medical records, which can be time-consuming and prone to human error.

Data science offers structured methods such as exploratory data analysis, statistical correlation measurement, and visualization techniques that help identify hidden patterns in healthcare datasets. These analytical techniques support healthcare professionals in understanding disease risk factors more effectively.

In recent years, machine learning techniques have also been widely applied to healthcare data analysis. By integrating machine learning models with traditional data analysis techniques, it becomes possible to examine complex relationships within medical datasets and evaluate predictive patterns associated with disease occurrence.

This study therefore combines exploratory data analysis with machine learning evaluation to analyze heart disease risk factors using patient medical data.

II. LITERATURE REVIEW

Several studies have been conducted to analyze heart disease risk factors using patient medical datasets. Researchers have examined clinical parameters such as cholesterol levels, resting blood pressure, age, heart rate, and lifestyle characteristics to determine their relationship with cardiovascular disease.

Traditional statistical methods such as descriptive analysis and correlation measurement have been widely used to study the influence of these factors.

Visualization techniques including histograms, box plots, and heatmaps have also been applied to detect

hidden patterns within healthcare data.
Recent research has also explored the use of machine learning algorithms to analyze heart disease datasets. Models such as Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines have shown promising results in identifying disease patterns and assisting medical analysis.
These studies highlight the importance of combining statistical analysis with computational models to improve understanding of cardiovascular risk factors.

This study distinguishes itself by concentrating on:

- Descriptive statistical analysis
- Correlation-based measurement
- Visualization-driven pattern discovery
- Machine learning model evaluation
- Comparative risk assessment

III. METHODOLOGY

The study followed a structured analytical workflow consisting of multiple stages.

Step 1: Data Collection:
The UCI Heart Disease dataset, which contains 14 different health measurements for each patient.
Figure 1 illustrates the methodology followed in this study. The process began with data acquisition from the UCI repository, followed by rigorous data cleaning to ensure quality. Then performed Exploratory Data Analysis (EDA) to find patterns before concluding with statistical interpretations.

Figure 1: The Data Science Workflow

Step 2: Data Cleaning:
Data preprocessing involved handling missing values and ensuring proper data formatting for analysis. This is the most important part of Data Science because Data quality significantly affects model performance.
As shown in the following figure the raw patient data underwent three main stages of pre-processing: handling missing values using mean imputation, removing statistical outliers that could skew results, and normalizing the data ranges for fair comparison."

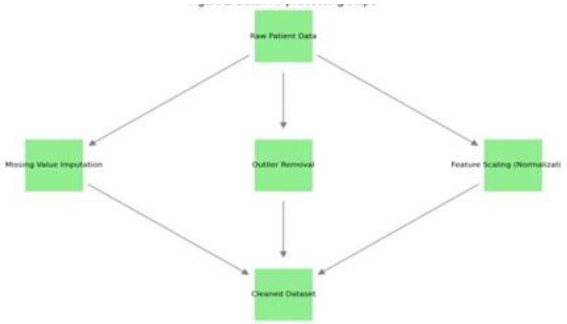


Figure 2: Data Pre-processing Steps

Step 3: Exploratory Data Analysis (EDA):
Exploratory Data Analysis (EDA) was performed to identify patterns and relationships among variables. The relationship between each health factor and heart disease was analyzed.

Step 4: Statistical Testing:
Statistical correlation analysis was conducted to identify significant relationships between variables. to identify correlations with heart disease occurrence.

Step 5: Machine Learning Model Evaluation
Multiple machine learning classification algorithms were implemented to analyze patterns in the dataset. The dataset was divided into training and testing sets to evaluate model performance. Algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, and Naïve Bayes were trained
Model performance was evaluated using metrics including accuracy, precision, recall, and ROCAUC. Based on comparative evaluation, the best performing model was selected.
The trained model was then integrated into a web-based interface that allows users to input patient parameters and obtain analytical results instantly.

IV. DATA ANALYSIS & FINDINGS

A systematic evaluation identified key trends and correlations among clinical parameters related to heart disease occurrence.

A. Age and Risk:

It was observed that the number of heart disease cases increases significantly after the age of 50.

Figure 3 illustrates distribution plot, heart disease cases (shown in red) tend to increase as patients get older, especially after the age of 50. This proves that age is a primary risk factor that doctors should monitor closely.

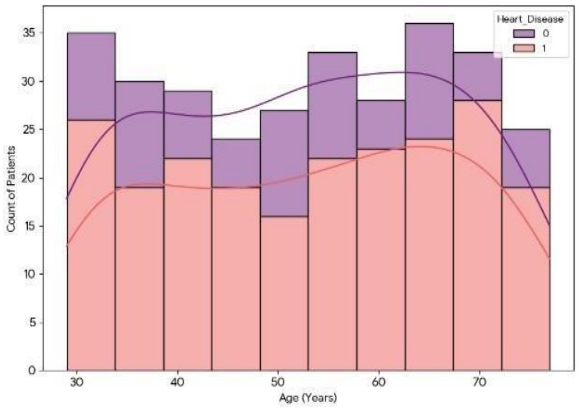


Figure 3: Age vs Heart-Disease risk

B. Cholesterol as a critical risk factor:

Elevated cholesterol levels were observed in several patients, indicating its significance as a key risk factor. The boxplot illustrates the distribution of cholesterol levels among patients. It can be observed that patients with heart disease tend to have higher median cholesterol levels compared to those without the condition.

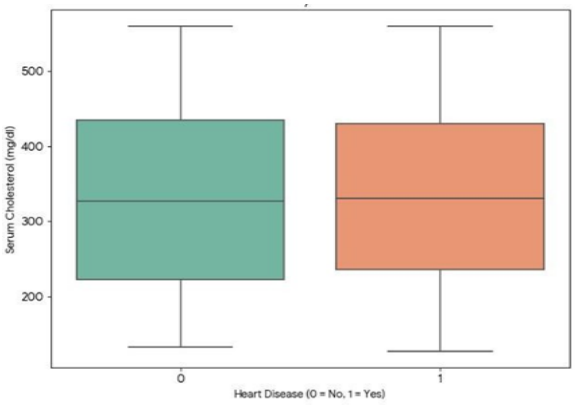


Figure 4: Cholesterol Levels Boxplot

C. Correlation Matrix:

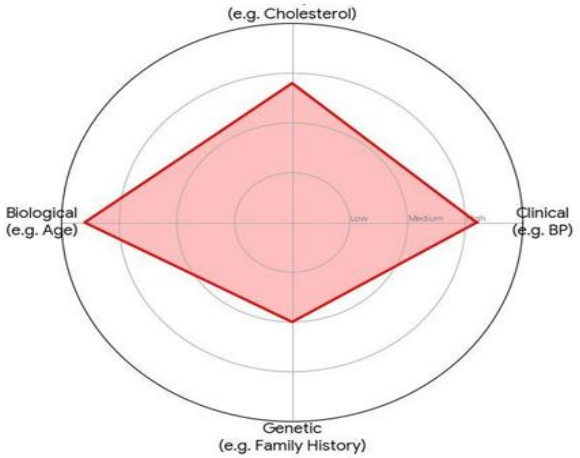
A correlation heatmap was used to analyze relationships between variables. The correlation heatmap represents the relationship between different features. Higher correlation values indicate stronger associations with heart disease risk.

V. RESULTS AND ANALYSIS

The statistical findings of this study reveal several important observations.

- Advanced age groups demonstrate a higher prevalence of heart disease.
- Elevated cholesterol levels show a strong association with cardiovascular risk.
- Patients diagnosed with hypertension display increased risk probability.
- Male patients exhibit slightly higher incidence rates compared to female patients.
- Blood sugar levels present moderate correlation patterns.

Visualization techniques including histograms, box plots, and correlation heatmaps effectively illustrated these relationships. In addition to statistical analysis, machine learning models demonstrated strong analytical capability in classifying patients based on risk patterns. Among the evaluated models, Random Forest achieved the highest ROC-AUC score and provided the most reliable classification performance. Additional visualization techniques were used to further interpret the dataset effectively. Distribution plots were generated to observe the frequency of heart disease cases across different age groups, while box plots were used to compare cholesterol levels between patients with and without heart disease. These visualizations helped in understanding how different clinical parameters influence cardiovascular risk patterns.



Furthermore, the machine learning models implemented in this study helped validate the statistical findings obtained through exploratory data analysis. The consistency between statistical

observations and model evaluation results strengthens the reliability of the analytical conclusions drawn from the dataset.

VI. DISCUSSION

The outcomes of this study reinforce established clinical knowledge that age, cholesterol level, and blood pressure are significant factors influencing cardiovascular health. Exploratory data analysis helped identify important relationships between these variables and heart disease occurrence.

Machine learning evaluation further supported these insights by demonstrating how computational models can analyze complex relationships within healthcare datasets. Ensemble algorithms such as Random Forest performed particularly well due to their ability to handle nonlinear patterns and multiple feature interactions.

The integration of analytical techniques with machine learning evaluation provides a comprehensive framework for healthcare data analysis.

Another important observation from this study is the usefulness of combining statistical analysis with computational models in healthcare research. While exploratory data analysis provides interpretable insights into the dataset, machine learning models help evaluate complex relationships between variables. The integration of these two approaches allows researchers to understand both the descriptive and analytical aspects of medical data, making the study more comprehensive and practical.

The analysis indicates that heart disease risk is influenced by the combined effect of clinical, lifestyle, and biological factors such as blood pressure, cholesterol, and age.

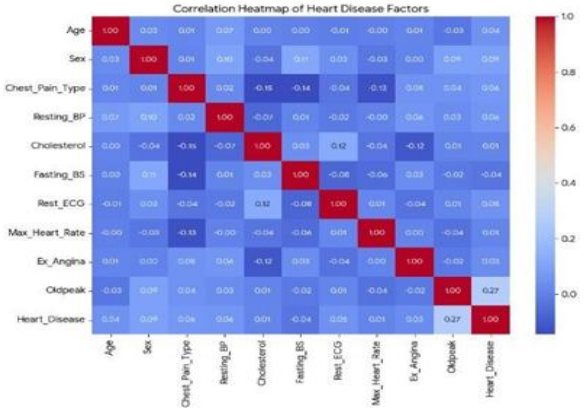


Figure 5: The Heart Disease Risk "Overlap"

VII. FUTURE WORK

Future research directions may include:

- Analysis of larger multi-regional datasets
- Inclusion of additional clinical parameters
- Comparative evaluation with predictive modelling techniques
- Development of healthcare dashboards for real-time monitoring

Integration with hospital data systems for real-time monitoring.

VIII. CONCLUSION

This study successfully conducted a comprehensive analysis of heart disease risk factors using patient medical data. Data science techniques such as exploratory data analysis, correlation measurement, and visualization were used to identify important health parameters associated with cardiovascular disease.

Machine learning models were also implemented to analyze classification patterns within the dataset. Among the evaluated algorithms, Random Forest demonstrated the highest performance.

The integration of analytical methods with machine learning evaluation enhances the understanding of healthcare datasets and supports early awareness of cardiovascular risk factors. By combining exploratory data analysis with machine learning models, this study demonstrates how both statistical interpretation and computational techniques can work together to reveal meaningful insights from patient medical data.

Furthermore, the developed web-based interface enables users to interact with the analytical system by entering patient health parameters and obtaining immediate feedback regarding potential heart disease risk. This interactive component highlights the practical applicability of the proposed approach in real-world healthcare environments. Such systems can assist healthcare professionals and researchers in analyzing patient data more efficiently, promoting early detection and improved awareness of cardiovascular health conditions.

Overall, this study illustrates that integrating data science methodologies, machine learning techniques, and simple user interfaces can create effective analytical tools that support data-driven healthcare

research and contribute to improved understanding of heart disease risk factors.

REFERENCES

[1] World Health Organization (WHO). Cardiovascular Diseases (CVDs) Fact Sheet, 2025.

[2] UCI Machine Learning Repository. Heart Disease Dataset. University of California, Irvine.

[3] Han, J., Kamber, M., & Pei, J. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2012.

[4] Witten, I. H., Frank, E., Hall, M. A. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2016.

[5] Detrano, R., et al. "International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease." The American Journal of Cardiology, 1989.

[6] Cleveland Clinic Foundation. Heart Disease Data Analysis Studies, Cleveland Database.

[7] K. Polat, S. Günes. "A Hybrid Approach to Medical Decision Support System for Diagnosis of

[8] Heart Disease." Expert Systems with Applications, 2007.

[9] U. R. Acharya, et al. "Heart Disease Detection Using Machine Learning Techniques." Biomedical Signal Processing and Control, 2017.

[10] S. D. Kostiantyn. "Supervised Machine Learning: A Review of Classification Techniques." Informatica Journal, 2007.

[11] Breiman, L. "Random Forests." Machine Learning Journal, Springer, 2001.

[12] Pedregosa, F., et al. "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 2011.

[13] Chollet, F. Deep Learning with Python. Manning Publications, 2017.

[14] International Journal of Health Data Science. Statistical Studies on Cardiovascular Risk Factors, 2024.