Heart Disease prediction using Machine learning

Savi   
IT dept.  
IGDTUWDelhi ,India  
[savigarg037@gmail.com](mailto:savigarg037@gmail.com)

Tanisha  
IT dept.  
IGDTUWDelhi, India  
[tanishamonga5@gmail.com](mailto:tanishamonga5@gmail.com)

Tripti Jaiswal  
IT dept.   
IGDTUWDelhi ,India  
[triptijas08@gmail.com](mailto:triptijas08@gmail.com)

*Abstract*— Cardiovascular diseases, including heart disease, are a leading cause of mortality globally. Early and accurate prediction of heart disease risk plays a pivotal role in preventive healthcare. In this study, we propose a heart disease prediction model based on the logistic regression concept of machine learning. The model is designed to classify patients into risk categories, aiding medical practitioners in timely intervention.

Our research employs a comprehensive dataset obtained from (<https://raw.githubusercontent.com/amankharwal/Website-data/master/heart.csv> ), comprising clinical and diagnostic features relevant to heart disease. Data preprocessing, including imputation of missing values and feature scaling, was conducted to ensure data quality and consistency. We implemented a logistic regression model due to its interpretability and suitability for binary classification tasks.

While our model showcases promising results, we acknowledge the limitations associated with the dataset's size and potential biases. Nonetheless, our research contributes to the field of heart disease prediction by offering a transparent, interpretable, and effective logistic regression-based approach. This work underscores the importance of accurate risk assessment and demonstrates the potential for machine learning models to aid healthcare professionals in identifying individuals at risk of heart disease.

Keywords— Machine Learning, Disease prediction ,Data visualization , Model training

# Introduction

Heart disease is a major cause of death worldwide. Timely and accurate prediction and diagnosis of heart disease are pivotal for effective prevention, intervention, and patient care. This introduction provides an overview of the critical topic of heart disease prediction and diagnosis, establishing the context for further research.

**Background :** Heart diseases encompass a wide spectrum of conditions affecting the heart and blood vessels, including coronary artery disease, heart failure, arrhythmias, valvular disorders, and congenital heart defects. These conditions can result in severe health consequences, including heart attacks, strokes, and impaired quality of life. Notably, cardiovascular diseases account for a substantial proportion of global deaths annually, highlighting their profound impact on public health.

The challenge in managing heart diseases lies in their often asymptomatic or subtle presentation until advanced stages.

Early detection and intervention are crucial to prevent complications and improve patient outcomes. Heart disease prediction and diagnosis play a pivotal role in identifying individuals at risk and initiating appropriate treatments and lifestyle modifications.

The research problem in the domain of heart disease prediction and diagnosis is twofold:

* **Prediction Problem:** How can we accurately predict the risk of heart disease in individuals, taking into account a range of risk factors, data sources, and emerging technologies?
* **Diagnosis Problem:** How can we efficiently and accurately diagnose heart diseases using a combination of medical imaging, diagnostic tests, and clinical data?

The significance of addressing these research problems is profound:

* **Public Health Impact**: Heart diseases remain a leading cause of mortality and morbidity globally. Accurate prediction and early diagnosis can lead to timely interventions that save lives and reduce the overall burden of cardiovascular diseases.
* **Improved Patient Outcomes**: Early prediction allows for lifestyle modifications and risk factor management in individuals at risk, potentially preventing the development of heart diseases. Accurate diagnosis ensures that patients receive appropriate treatments and interventions, improving their quality of life.
* **Healthcare Resource Optimization**: Efficient prediction and diagnosis can lead to cost savings for healthcare systems by reducing hospitalization rates, emergency interventions, and long-term care costs associated with late-stage heart diseases.
* **Advancements in Medicine**: Research in heart disease prediction and diagnosis drives technological advancements, fosters innovation in diagnostic tools, and paves the way for personalized medicine, ultimately improving the standard of care for heart patients.
* **Equity in Healthcare**: Addressing disparities in heart disease prediction and diagnosis can help ensure equitable access to healthcare services and minimize bias in predictive models, reducing health disparities among different population groups.
* **Ethical Data Use**: In an era of increasing data availability, it is crucial to address ethical considerations related to data privacy, informed consent, and responsible data handling in the context of heart disease prediction and diagnosis.

Existing literature in heart disease prediction and diagnosis highlights the significance of timely and accurate risk assessment and detection. However, gaps in knowledge persist:

* **Limited Personalization:** Current predictive models often lack personalization, failing to consider individualized risk factors and genetic factors that play a crucial role in heart disease.
* **Ethical Concerns:** As data sources grow, ethical concerns related to patient data privacy and responsible data handling need comprehensive consideration.
* **Interpretability:** The interpretability of complex predictive models and the integration of these models into clinical practice remain challenges.
* **Equity:** Disparities in healthcare access and bias in predictive models require attention to ensure equitable care for all populations.

**Purpose and Objectives:**

The primary purpose of our research is to address these gaps and challenges in heart disease prediction and diagnosis:

* Develop personalized predictive models for heart disease, integrating individual risk factors and genetic data for improved accuracy.
* Explore ethical guidelines and data privacy measures for responsible data utilization in predictive models and diagnostic tools.
* Investigate methods for enhancing model interpretability and facilitating seamless integration into clinical practice.
* Address healthcare disparities and bias in predictive models, striving for equitable heart disease prevention and diagnosis.

# LITERATURE REVIEW

## A comprehensive review of existing literature reveals a wealth of research and clinical studies related to heart disease prediction and diagnosis. These studies have contributed to our understanding of risk factors, diagnostic tools, predictive models, and their implications for public health. Here, we analyze and synthesize key findings from previous research to support our research objectives in the context of heart disease prediction and diagnosis.

## Risk Factors and Predictive Models

## Numerous studies have identified traditional risk factors associated with heart disease, including age, gender, smoking, hypertension, and cholesterol levels. These factors continue to serve as essential components in predictive models like the Framingham Risk Score and the ACC/AHA cardiovascular risk calculator. However, the literature highlights the need to enhance risk prediction by incorporating novel risk factors, such as genetic markers, inflammatory biomarkers, and socioeconomic factors. Recent research suggests that personalized risk assessment, integrating individual genetic and lifestyle data, can significantly improve the accuracy of predictive models.

1. *Machine Learning and Deep Learning*

Machine learning and deep learning techniques have gained prominence in heart disease prediction. Research studies have employed a variety of algorithms, including support vector machines, decision trees, random forests, and neural networks. These models leverage diverse data sources, such as electronic health records, medical imaging, and genetic information, to enhance predictive accuracy. While these models demonstrate promise, the literature also underscores the importance of model interpretability, especially in clinical settings, where transparency and trust are critical.

1. *Diagnostic Tools and Imaging*

Advancements in medical imaging, such as cardiac MRI, CT angiography, and echocardiography, have improved the accuracy of heart disease diagnosis. Studies have explored the diagnostic potential of these modalities in identifying structural abnormalities, assessing cardiac function, and characterizing tissue. The literature suggests that these tools are indispensable for detecting conditions like coronary artery disease, valvular disorders, and cardiomyopathies. However, challenges related to accessibility and cost-effectiveness persist, highlighting the need for research into improving diagnostic tool accessibility.

1. *Ethical Considerations and Data Privacy*

The ethical use of patient data is a recurring theme in the literature. Researchers and clinicians must navigate the delicate balance between accessing patient information for research and ensuring data privacy and security. Various studies emphasize the importance of informed consent, de-identification techniques, and adherence to ethical guidelines when utilizing patient data for predictive modeling and diagnosis.

1. *Healthcare Disparities and Bias*

Several studies highlight healthcare disparities and bias in heart disease prediction and diagnosis. Research indicates that certain populations, particularly those from marginalized backgrounds, may receive inequitable access to healthcare services and experience bias in predictive models. Addressing these disparities and mitigating bias in predictive tools are crucial objectives to ensure equitable healthcare delivery.

***Theoretical Frameworks and Models:*** The theoretical framework for our research draws from various domains, including public health, data science, and medical ethics. We adopt a personalized medicine perspective, grounded in the belief that individualized risk assessment and tailored interventions are essential for improving heart disease prediction and diagnosis. Additionally, ethical frameworks, such as the principles of autonomy, beneficence, and justice, guide our approach to data privacy and responsible data use.

# RESEARCH DESIGN & APPROACH

In this study, we aim to develop a predictive model for heart disease using machine learning techniques. The research design is primarily quantitative and analytical in nature. We employ a supervised learning approach, specifically logistic regression, to classify individuals into two groups: those with heart disease (positive class) and those without (negative class).

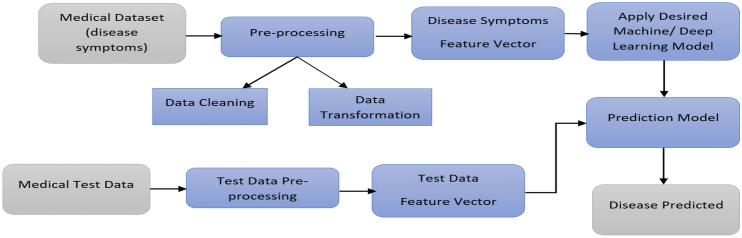


Fig. 1. Research Approach

## Abbreviations and Acronyms

**cp** -chest pain , **fbs** -fasting blood sugar , **thal -**thalassemia, **exang** – exercise induced angina , **trestbps -**Resting blood pressure , **restecg** – Resting Electrocardiographic , **thalach –** maximum heart rate ,**chol-** Serum cholesterol

**Data Collection Methods and Sources**

The dataset used in this study was obtained from an external source. We utilized the "heart.csv" dataset, a publicly available dataset that includes various demographic, clinical, and diagnostic features, as well as the target variable indicating the presence or absence of heart disease. The data is derived from a combination of sources, including medical records and patient information.

1. ***Data Preprocessing***

Before training the machine learning model, we conducted extensive data preprocessing. This included:

* Data Cleaning: We checked for missing values and handled them appropriately to ensure data completeness.
* Exploratory Data Analysis (EDA): Before training our logistic regression model, we embarked on an exploratory data analysis (EDA) journey to comprehend the dataset thoroughly. EDA serves the critical purpose of answering essential questions:
* What questions are we attempting to address?
* What is the nature of our data, and how should we handle various types?
* Are there missing data, and how should we address it?
* What insights can we derive from data outliers, and why are they significant?
* How can we engineer, modify, or remove features to optimize data utilization?

1. ***Data Visualization***

## Categorical Features Analysis

We conducted a visual analysis of categorical features in relation to heart disease presence:

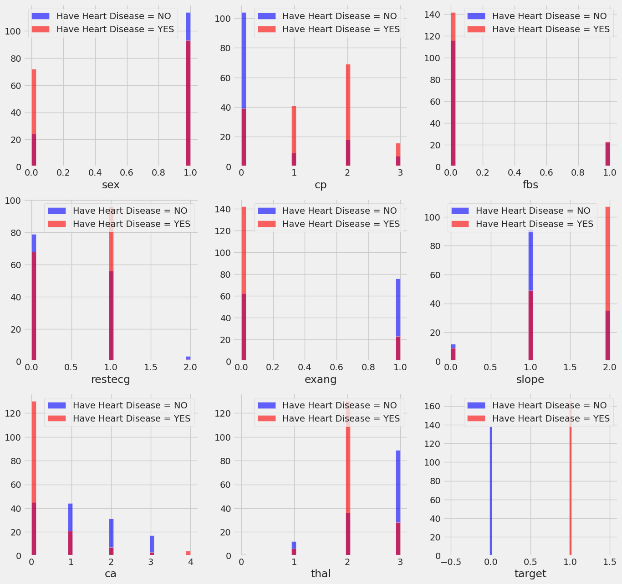


Fig. 2. Categorical Features analysis

Notable observations from Fig. 1. included associations between heart disease and categorical features such as chest pain (cp), resting EKG results (restecg), exercise-induced angina (exang), slope, ca (number of major vessels), and thal (thalium stress result).

* 1. *Continuous Features Analysis*

We also analyzed the distribution of continuous features concerning heart disease presence:

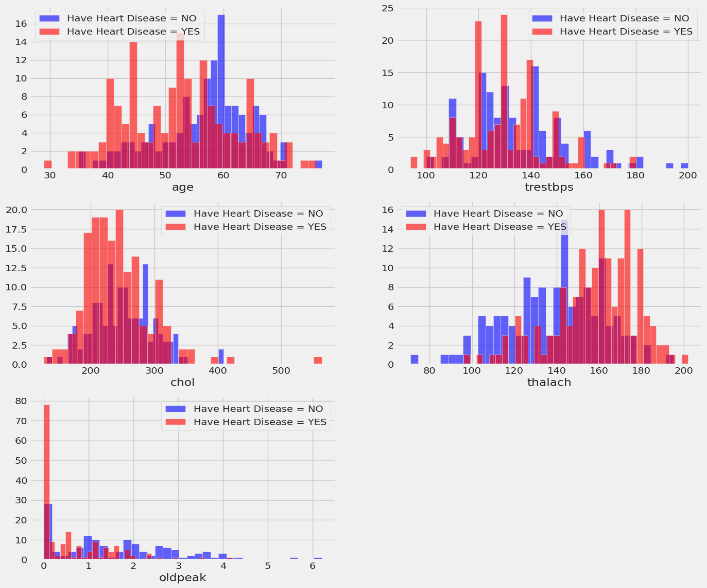


Fig. 3. Continuous Features Analysis

This analysis highlighted features such as resting blood pressure (trestbps), cholesterol level (chol), maximum heart rate (thalach), and the old peak of exercise-induced ST depression vs. rest as potentially indicative of heart disease.

1. ***Model Training - Logistic Regression***

We opted for the logistic regression algorithm as the primary model for heart disease prediction. Logistic regression is well-suited for binary classification tasks, calculating the probability of an instance belonging to one of two classes.

1. *Logistic Regression Overview*

Logistic regression is a fundamental machine learning algorithm widely used for binary classification tasks, making it particularly suitable for our heart disease prediction project. Unlike linear regression, which predicts continuous numeric values, logistic regression predicts the probability that a given input belongs to a specific class. In our case, it predicts the probability of an individual having heart disease (class 1) or not (class 0).

1. *How Logistic Regression Works*

Logistic regression works by modelling the relationship between the independent variables (features) and the binary outcome (heart disease presence) using the logistic function, often referred to as the sigmoid function. The sigmoid function maps any real-valued number to a value between 0 and 1. This output can be interpreted as the probability of an instance belonging to class 1 (heart disease).

The logistic regression model is represented by the following equation:

Where:

* is the probability of the instance belonging to class 1.
* e is the base of the natural logarithm.
* are the coefficients of the model.
* are the independent variables (features).

1. *Model Training and Optimization*

During the training phase, logistic regression aims to find the optimal values for the coefficients (

that minimize the difference between the predicted probabilities and the actual class labels in the training data. This process is typically done using optimization techniques like gradient descent.

1. *Model Performance*

* **Training Set**: The logistic regression model achieved an accuracy of 86.79% on the training data, indicating that around 86.79% of the training instances were correctly classified by the model.
* **Test Set**: The accuracy on the test data is 81.32%, indicating that around 81.32% of the test instances were correctly classified by the model.

The model performs reasonably well on both the training and test sets .

1. *Potential Improvements*

The model's performance on the test set is slightly lower than on the training set, suggesting a slight degree of overfitting. Fine-tuning the model or considering regularization techniques may help reduce overfitting and improve generalization.

* Model Limitations: While logistic regression is a robust algorithm, it may not capture complex, nonlinear relationships present in the data. More advanced algorithms might offer improved performance.
* Data Imbalance: Although we have a balanced dataset, heart disease prevalence in the real world may differ, leading to potential challenges in deploying the model.

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

1. American Heart Association. (2021). Heart Disease and Stroke Statistics—2021 Update: A Report From the American Heart Association. Circulation, 143(8), e254–e743.
2. Pencina, M. J., D'Agostino, R. B., Larson, M. G., Massaro, J. M., Vasan, R. S., & Kannel, W. B. (2009). Predicting the 30-year risk of cardiovascular disease: the Framingham Heart Study. Circulation, 119(24), 3078-3084.
3. Alizadehsani, R., Abdar, M., Roshanzamir, M., Hussain, S., & Hussain, O. K. (2019). Cardiovascular disease diagnosis using deep learning and metaheuristic optimization: A review. Computational and Structural Biotechnology Journal, 17, 1044-1051.
4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
5. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.
7. Scikit-learn: Machine Learning in Python. (2021).

Available online: <https://scikit-learn.org/stable/index.html>

1. Seaborn: Statistical Data Visualization. (2021).

Available online: <https://seaborn.pydata.org/index.html>

1. Matplotlib: Visualization with Python. (2021).

Available online: <https://matplotlib.org/>

1. Python Software Foundation. (2021). Python Language Reference, Version 3.9.

Available online: <https://docs.python.org/3.9/reference/index.html>

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**