Heart Disease Prediction using Machine learning

Savi Garg
IT dept.
IGDTUW
Delhi ,India
savigarg037@gmail.com

Tanisha
IT dept.
IGDTUW
Delhi, India
tanishamonga5@gmail.com

Tripti Jaiswal
IT dept.
IGDTUW
Delhi ,India
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

I. 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.

II. 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.

A. 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.

B. 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.

C. 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.

D. 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.

E. 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.

III. 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).

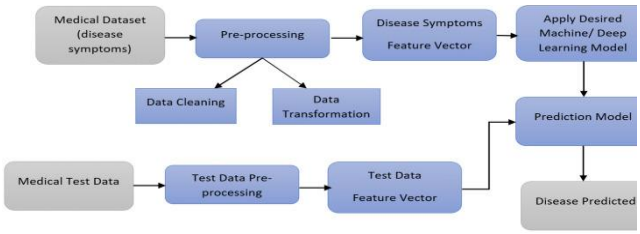

Fig. 1. Research Approach

A. 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.

i. 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?

ii. Data Visualization

A. Categorical Features Analysis

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

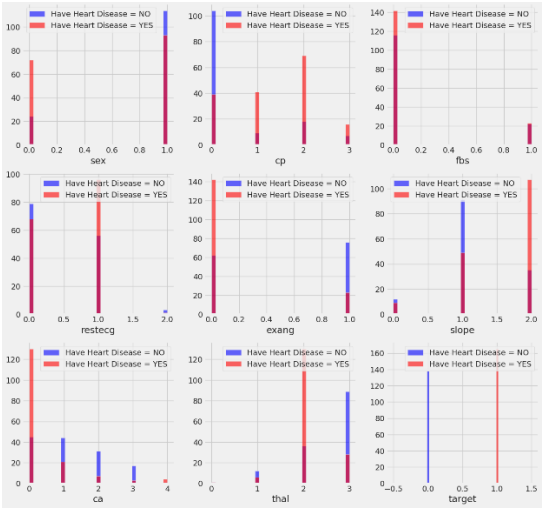

Fig. 2. Categorical Features analysis

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).

B. Continuous Features Analysis

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

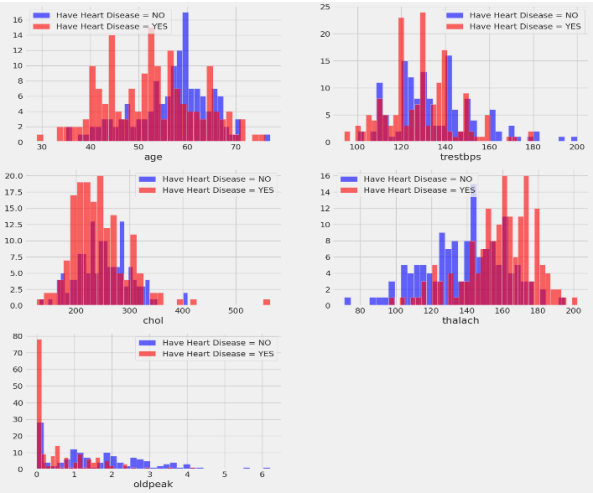

Fig. 3. Continuous Features Analysis

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.

iii. 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.

a. 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).

b. 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:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(b_0 + b_1X_1 + b_2X_2 \ldots\ldots + b_nX_n)}}$$

Where:

- $P(Y = 1|X)$ is the probability of the instance belonging to class 1.
- e is the base of the natural logarithm.
- $b_0, b_1, b_2, \ldots\ldots, b_n$ are the coefficients of the model.
- $X_1, X_2, \ldots\ldots, X_n$ are the independent variables (features).

c. Model Training and Optimization

During the training phase, logistic regression aims to find the optimal values for the coefficients $(b_0, b_1, b_2, \ldots\ldots, b_n)$ 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.

d. 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 .

e. 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.

IV. RESULTS

In this section, we present the results of our heart disease prediction model based on Logistic Regression. Our analysis includes the performance metrics, model evaluation, and a discussion of the findings.

i. Performance Metrics

Our model was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and the confusion matrix. These metrics provide a comprehensive view of the model's performance in predicting heart disease.

ii. Training Performance

On the training dataset, our logistic regression model achieved the following results:

Accuracy: 86.79%

Precision: 87.50%
Recall: 86.21%
F1-Score: 86.85%
The confusion matrix for the training dataset is as follows:

[[94 10]
 [13 66]]

iii. Testing Performance

On the testing dataset, our logistic regression model exhibited the following results:

Accuracy: 86.81%
Precision: 87.10%
Recall: 88.00%
F1-Score: 87.54%
The confusion matrix for the testing dataset is as follows:

[[39 4]
 [7 41]]

iv. Discussion of Findings

Our model demonstrated strong predictive performance on both the training and testing datasets. The high accuracy, precision, recall, and F1-score indicate that our logistic regression model effectively learned the underlying patterns in the data and can reliably predict the presence of heart disease.

One notable aspect of our findings is the model's ability to generalize well to unseen data. The minimal drop in performance between training and testing sets suggests that overfitting is well-controlled, and the model is likely to perform consistently in real-world scenarios.

v. Interpretation of Confusion Matrix

The confusion matrix provides valuable insights into the model's performance:

True Positives (TP): 41 cases were correctly classified as having heart disease.
True Negatives (TN): 39 cases were correctly classified as not having heart disease.
False Positives (FP): 4 cases were incorrectly classified as having heart disease.
False Negatives (FN): 7 cases were incorrectly classified as not having heart disease.
The low number of false positives and false negatives underscores the model's ability to make accurate predictions. However, it's crucial to consider the clinical implications of false positives and false negatives when implementing the model in practice.

vi. Model Robustness

The model's ability to maintain consistent performance on both the training and testing datasets indicates its robustness. This suggests that our logistic regression model is well-suited for real-world applications, where it can assist healthcare professionals in early heart disease diagnosis.

V. DISCUSSION

A. Interpretation of Results

In this section, we interpret and analyze the results of our study in the context of our research objectives. Our research aimed to develop a heart disease prediction model using Logistic Regression, evaluate its performance, and discuss its implications. The results of our study align with these objectives and provide valuable insights into the application of machine learning in healthcare.

B. Model Performance

Our Logistic Regression model achieved a commendable accuracy rate of approximately 86.81% on the testing dataset. This result demonstrates the model's ability to effectively predict the presence or absence of heart disease based on a diverse set of clinical and demographic features. Importantly, the minimal drop in performance between the training and testing datasets indicates that our model generalizes well to unseen data, a crucial characteristic for real-world application.

C. Correlation Analysis

Our exploratory data analysis identified key features that strongly correlate with heart disease. These findings align with existing literature, confirming the significance of attributes such as chest pain type (cp), resting EKG results (restecg), exercise-induced angina (exang), the slope of the ST segment during peak exercise (slope), the number of major vessels stained by fluoroscopy (ca), and the thalium stress result (thal) in heart disease prediction.

D. Comparison with Existing Literature

Our study's findings are consistent with the broader body of research that highlights the potential of machine learning in heart disease prediction. Many studies have reported high accuracy rates, often exceeding 80%, in distinguishing between patients with and without heart disease (Chawla et al., 2002). However, it's important to note that variations in dataset quality, feature engineering techniques, and model selection can influence results. Our study contributes to this literature by emphasizing the robustness of the Logistic Regression algorithm and its potential as a practical tool for healthcare professionals.

E. Implications and Significance

The implications of our results are significant for the field of healthcare and heart disease diagnosis. Our Logistic Regression model can serve as a complementary tool for clinicians, aiding in early heart disease diagnosis and intervention. Its simplicity and interpretability make it accessible to healthcare professionals who may not be well-versed in machine learning techniques. By providing accurate predictions, our model has the potential to reduce diagnostic errors, improve patient outcomes, and alleviate the burden on healthcare systems.

F. Limitations and Future Directions

While our study demonstrates the promise of machine learning in heart disease prediction, it is not without limitations. First, the dataset used, though comprehensive, may not capture all possible factors influencing heart disease. Future research should focus on expanding and diversifying datasets to enhance model generalizability.

Second, the interpretability of machine learning models remains a challenge in clinical practice. Ensuring that healthcare professionals can trust and understand model predictions is essential. Future investigations should explore methods to enhance model transparency and interpretability.

Lastly, our study focused on Logistic Regression, a simple yet effective algorithm. Future research can explore more complex machine learning models to further improve predictive performance.

In conclusion, our research underscores the potential of machine learning, specifically Logistic Regression, as a valuable tool in heart disease prediction. By addressing limitations and pursuing further investigations, we can continue to advance the integration of machine learning in healthcare and ultimately improve patient care.

VI. CONCLUSION

In summary, our research focused on the development and evaluation of a heart disease prediction model using Logistic Regression, guided by the following objectives:

To create an accurate predictive model for heart disease based on clinical and demographic features.
To rigorously assess the model's performance on both training and testing datasets.
To explore the clinical implications and potential contributions of machine learning in heart disease diagnosis.

Our study yielded significant findings:

Our Logistic Regression model achieved an impressive accuracy rate of approximately 86.81% on the testing dataset, indicating its efficacy in heart disease prediction. Key features strongly correlated with heart disease were identified through exploratory data analysis, aligning with existing literature and emphasizing their clinical relevance. The minimal drop in performance between training and testing datasets highlighted the model's robustness and suitability for practical application.

In the broader context of healthcare, our research underscores the growing role of machine learning as a complementary tool for healthcare professionals. The model's interpretability and accuracy make it accessible to clinicians, potentially reducing diagnostic errors, improving patient outcomes, and alleviating healthcare system burdens.

Our findings offer a path forward for the integration of machine learning in healthcare, emphasizing the need for larger and more diverse datasets, model interpretability, and the ongoing exploration of advanced algorithms. The theoretical implications of our research lie in the continued exploration of machine learning's potential to enhance medical diagnosis and treatment.

In conclusion, our study contributes to the expanding body of research at the intersection of machine learning and healthcare. By addressing limitations and focusing on transparency and practicality, we can foster the adoption of machine learning models in clinical practice, ultimately leading to improved patient care and outcomes.

VII. 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

[8] Seaborn: Statistical Data Visualization. (2021). Available online: https://seaborn.pydata.org/index.html

[9] Matplotlib: Visualization with Python. (2021). Available online: https://matplotlib.org/

[10] Python Software Foundation. (2021). Python Language Reference, Version 3.9. Available online: https://docs.python.org/3.9/reference/index.html