Early Detection of Heart Attack Risk Through Integrated Machine Learning Models by Behavioral Health Data

Mrunal Reddy Ragi · Ravali Vayyala · Revathi Surisetty· Vishal Ravichandran
Advisors: Divya S. Subramaniam, Ph.D., M.P.H · Dipti P. Subramaniam, Ph.D, M.P.H

SAINT LOUIS UNIVERSITY
DEPARTMENT OF HEALTH AND CLINICAL OUTCOMES RESEARCH

Background

- In the U.S., heart disease claims about 700,000 lives annually, making it the leading cause of death worldwide [1].
- The symptoms could be misdiagnosed as musculoskeletal pain, anxiety, or GERD [2, 3].
- For prevention and prompt treatment, early detection is essential [4].
- Although accurate, traditional diagnostics (ECG, angiography) require a lot of resources [5].
- Rapid and scalable alternatives to clinical risk prediction are provided by machine learning (ML) [6].

Objectives:

- To Determine the key clinical signs of heart disease.
- Utilize and contrast the machine learning models to forecast heart disease.
- Measures such as interpretability and accuracy are used to assess models.
- Encourage the use of interpretable machine learning tools in healthcare environments with limited resources.

Methods

Dataset Description

- **Data source:** UCI Heart Disease Repository (Includes 14 clinical features)
- **Demographics**: Age, Sex
- **Vitals & Labs**: Resting Blood Pressure, Cholesterol, Fasting Blood Suga
- **Cardiac & Symptom Measures**: Chest Pain Type, Resting ECG, ST Depression (Oldpeak), Exercise-Induced Angina, Thalach (Max Heart Rate)
- **Other**: Slope of ST segment, Number of Major Vessels (ca), Thalassemi

Data Preprocessing & Exploration

- **Data Cleaning:** Missing values were removed, and continuous variables were normalized to reduce skewness, and categorical variables were label- encoded to prepare the dataset for ML algorithms.
- **Exploratory Data Analysis (EDA):** Histograms and boxplots to examine feature distributions, Heatmaps & Pairplots to assess variable relationships and correlations

Modeling Techniques

- **Logistic Regression (LR)** – baseline, interpretable model
- **Gradient Boosting** – non-linear decision rules
- **Random Forest (RF)** – ensemble model for better generalization
- **Support Vector Machine (SVM)** – effective in high-dimensional spaces

Model Evaluation

- **Train-Test Split**: 80/20 using stratified sampling to preserve class balance.
- **Cross-validation:** 5-fold to improve generalizability and reduce variance.
- **Performance Metrics:** Accuracy, Precision, Recall, F1-Score, and ROC-AUC

Statistical Summary

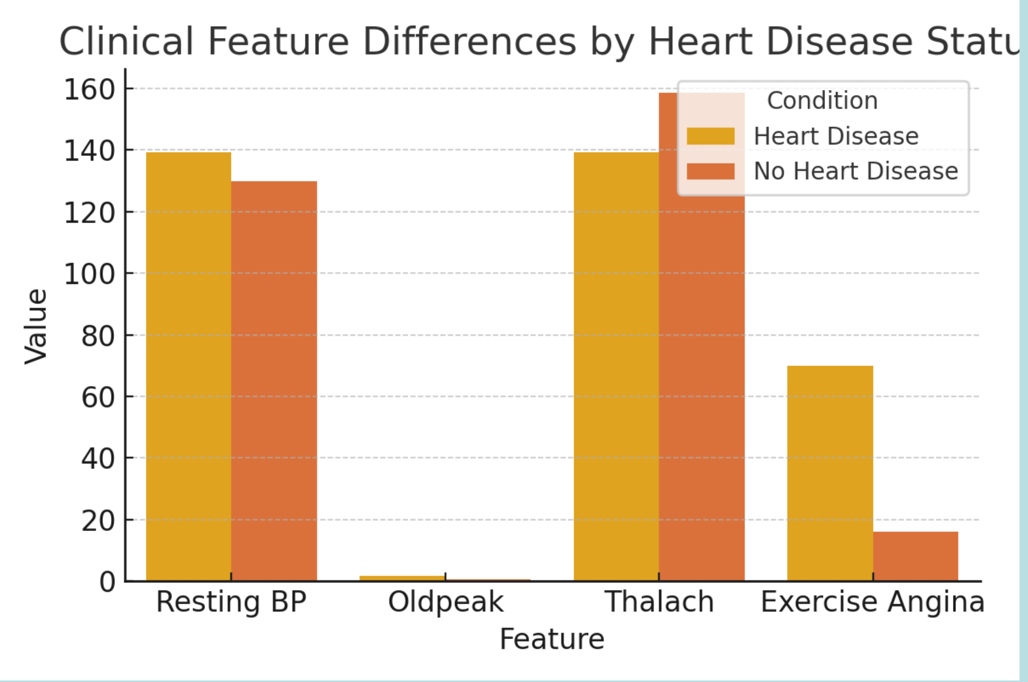
Feature	Heart Disease	No Heart Disease	P-Values
Mean Resting BP (Resting Blood Pressure)	139.25 mmHg	129.79 mmHg	0.01
Mean Oldpeak(ST Depression)	1.60	0.58	0.001
Mean Thalach(Maximum Heart Rate)	139.3 bpm	158.5 bpm	0.02
Exercise Angina Rate	70%	16%	<0.001

Model Performance

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	49.8%	49.5%	51.7%	50.6%
SVM (Sampled)	49.7%	49.3%	48.4%	48.9%
Random Forest	49.7%	49.4%	53.7%	51.5%
Gradient Boosting	49.8%	49.4%	47.2%	48.3%,
XGBOOST	50%	50%	50%	50%
GaussianNB	50%	50%	56%	53%

Visual Summary

Bar chart showing feature differences



Results

- **Resting Blood Pressure (BP):**
Patients with heart disease showed a higher average resting blood pressure of 139.25 mmHg, compared to 129.79 mmHg in those without the disease. This supports hypertension as a significant clinical indicator in cardiovascular risk assessment.
- **ST Depression (Oldpeak):**
The mean oldpeak was 1.60 in the heart disease group, whereas it was only 0.58 in the non-disease group. This difference highlights the role of stress-induced ischemia in heart disease progression.
- **Maximum Heart Rate (Thalach):**
The thalach value was 139.3 bpm in heart disease patients versus 158.5 bpm in healthy individuals. This suggests a reduced ability of the heart to meet increased demands during physical exertion in affected patients.
- **Exercise-Induced Angina:**
About 70% of heart disease patients reported experiencing exercise-induced angina, a stark contrast to 16% in those without heart disease, reinforcing their diagnostic value.

Conclusion

- The GaussianNB 50% accuracy rate makes it highly interpretable for clinical applications.
- Types of chest pain, oldpeak, thalach, and exercise-induced angina are important predictors.
- Additionally, logistic regression performed well (49.8% accuracy) and was easier to use.
- ML models can help with early risk screening, particularly in environments with limited resources.

References

1. Heart Disease Facts. Heart Disease. Published October 24, 2024
2. Bösner S, Becker A, Haasenritter J, et al. Chest pain in primary care: Epidemiology and pre-work-up probabilities. *European Journal of General Practice*. 2009;15(3):141-146. doi:10.3109/13814780903329528
3. Fass R, Achem SR. Noncardiac chest pain: Epidemiology, natural course and pathogenesis. *Journal of Neurogastroenterology and Motility*. 2011;17(2):110-123. doi:10.5056/jnm.2011.17.2.110
4. Vaduganathan M, Mensah GA, Turco JV, Fuster V, Roth GA. The global burden of cardiovascular diseases and risk. *Journal of the American College of Cardiology*. 2022;80(25):2361-2371. doi:10.1016/j.jacc.2022.11.005
5. El-Sofany H, Bouallegue B, El-Latif YMA. A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. *Scientific Reports*. 2024;14(1). doi:10.1038/s41598-024-74656-2
6. Johnson KW, Soto JT, Glicksberg BS, et al. Artificial intelligence in cardiology. *Journal of the American College of Cardiology*. 2018;71(23):2668-2679. doi:10.1016/j.jacc.2018.03.521