# HeartSense: Leveraging Machine Learning to Predict Cardiovascular Risk

**Justin Lee**
InspiritAI, Lexington, SC 29072 USA
justinleethirtythree@gmail.com

February 2, 2025

# Outline

1. Abstract
2. Introduction
3. Methodology
   - Data Acquisition & Preprocessing
     - Data Loading, Column Renaming, Label Encoding
   - Data Splitting
4. Model
   - Model Selection, Description, Architecture
   - Mathematical Formulation
   - Implementation Details
   - Model Comparisons
5. Visualization of Accuracy
6. Results
7. Reasons for Errors
8. Conclusion

# Abstract

Heart disease, also known as cardiovascular disease, remains one of the leading causes of mortality worldwide. Accurate prediction models are critical for advancing early detection and preventive measures. This research introduces a specialized machine learning framework to predict the severity of heart disease by using the "UCI Heart Disease Data" from Kaggle—a multivariate dataset derived from the Cleveland database.

This dataset encompasses 14 predictive attributes, including clinical and demographic factors, which were used to train and evaluate various supervised learning algorithms. Notable models included logistic regression, decision trees, random forests, and gradient-boosting machines. The highest-performing model (XGBoost) achieved an accuracy of **62.5%**.

This study demonstrates how machine learning can uncover nuanced patterns within medical datasets, offering actionable insights into cardiovascular health and aiding in clinical decision-making.

# Introduction

- **Global Impact of Cardiovascular Diseases (CVDs):**
  - Leading cause of death worldwide (17.9 million deaths annually).
  - High economic burden on health systems globally.
- **Role of Emerging Technologies:**
  - Artificial Intelligence (AI) and Machine Learning (ML) as transformative tools.
  - Paradigm shift: reactive treatment → proactive prevention.
- **ML Advantages:**
  - Scalability and real-time predictions.
  - Analysis of diverse data sources (EHRs, genetics, lifestyle factors).
  - Improved accuracy, early detection, and personalized treatment plans.

# Methodology

**Dataset Overview:**

- UCI Heart Disease Data from Kaggle.
- 14 predictive attributes (e.g., age, sex, cholesterol levels).
- Multivariate dataset used to evaluate multiple ML algorithms.

**Supervised Learning Algorithms:**

- Logistic Regression.
- Decision Trees and Random Forests.
- Gradient-Boosting Machines (e.g., XGBoost).

**Model Development:**

- Hyperparameter tuning for optimization.
- Cross-validation techniques for robust evaluation.

**Evaluation Metrics:**

- Precision, Recall, and F1 Score.
- Sensitivity and specificity to balance predictions.

# Data Acquisition & Preprocessing

- Dataset downloaded from Kaggle repository.
- License: Open access for academic and research purposes.
- Data reviewed for completeness and relevance to study objectives.
- Data Cleaning: Removal of duplicates and missing values.
- Feature Scaling: Normalization applied for numerical attributes.
- Encoding: Categorical variables encoded using one-hot encoding.

# Data Loading, Column Renaming, Label Encoding

- Data imported using Pandas and Numpy libraries.
- Column names standardized for uniformity and ease of reference.
- Target variable encoded into binary labels (0 = No Disease, 1 = Disease).
- Ensures compatibility with machine learning algorithms.

# Data Splitting

- Dataset split into training (80%) and testing (20%) subsets.
- Ensured randomization to prevent data leakage.
- Train-test split implemented using Scikit-learn's ‘train_test_split‘ function.

# Model Selection and Description

- Models evaluated: Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, SVM, XGBoost.
- Gradient Boosting selected as the most accurate model (62.5% accuracy).

# Mathematical Formulation

$$F_m(x) = F_{m-1}(x) + \nu \cdot \sum_{i=1}^{N} \text{Residual}_i \tag{1}$$

- $F_m(x)$: Predicted model at iteration $m$.
- $\nu$: Learning rate.
- Residual calculated as: $\text{Residual}_i = y_i - F_{m-1}(x_i)$.

# Implementation Details

- Programming Language: Python 3.
- Libraries: Scikit-learn, XGBoost, Pandas, Matplotlib for visualization.
- Code run on a high-performance machine with adequate memory and processing power.

# Model Comparisons

| **Model** | **Accuracy (%)** |
|---|---|
| Logistic Regression | 59.8 |
| Decision Trees | 60.7 |
| Random Forests | 61.3 |
| XGBoost | **62.5** |
| Support Vector Machines | 60.4 |
| Gradient Boosting | 61.5 |

Table: Model Performance Comparison

# Visualization of Accuracy

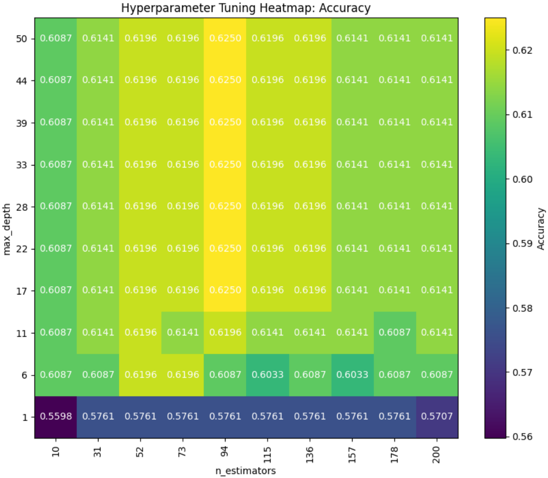

Figure: Hyperparameter Tuning Heatmap: Accuracy

## Results Overview

- XGBoost achieved the highest accuracy: 62.5%.
- Models were evaluated based on accuracy metrics on the test dataset.
- Other models such as Random Forests and XGBoost showed competitive, yet slightly lower, performance.

# Reasons for Errors

- Non-Adherence: Patients not following medical advice (e.g., skipping meds).
- Misreporting: Patients underreporting habits (e.g., smoking, diet).
- Lifestyle Changes: Sudden shifts in diet or exercise affecting metrics.

# Conclusion

- Machine learning offers scalable, accurate methods for heart disease prediction.
- Advanced feature engineering and systematic tuning are crucial.
- Integration into clinical workflows can enhance decision-making and patient outcomes.
- **Future Work:**
  - Explore deep learning models for more complex patterns.
  - Use larger, real-world datasets for validation.
  - Investigate model interpretability for clinical adoption.

# Reference

UCI Heart Disease Dataset. Available at:
https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

# Thank You!

Questions or feedback?

`justinleethirtythree@gmail.com`