MESHCHERYAKOV
LABORATORY OF
INFORMATION
TECHNOLOGIES

JOINT INSTITUTE FOR
NUCLEAR RESEARCH

Machine Learning Fundamentals: From Theory to Practice

Heart Disease Classification Using Machine Learning Algorithms

By
Rabab Asr, Enjy Fargally, Zeyad Radwan,
Aliaksei Hurnovich and Esraa Hussien.

Supervised By:
Dr. Mohammed Ibrahim
Prof. Andrey Nechaevsky
Meshcheryakov laboratory of information technologies, JINR

Outlines

1. Introduction

Background:

- Medical diagnoses often rely on expert opinions, but in heart disease cases, consensus is difficult due to varying patient symptoms. To improve early detection and treatment outcomes, researchers are developing new methods to identify heart disease in its early stages.
- Machine Learning (ML) is a field of AI that enables computers to learn from data and make predictions without being explicitly programmed.

Project Objective:

- Develop a predictive model using clinical and demographic data to identify patients at risk for heart disease.
- This project uses supervised learning, specifically classification, to predict heart disease based on patient medical data.

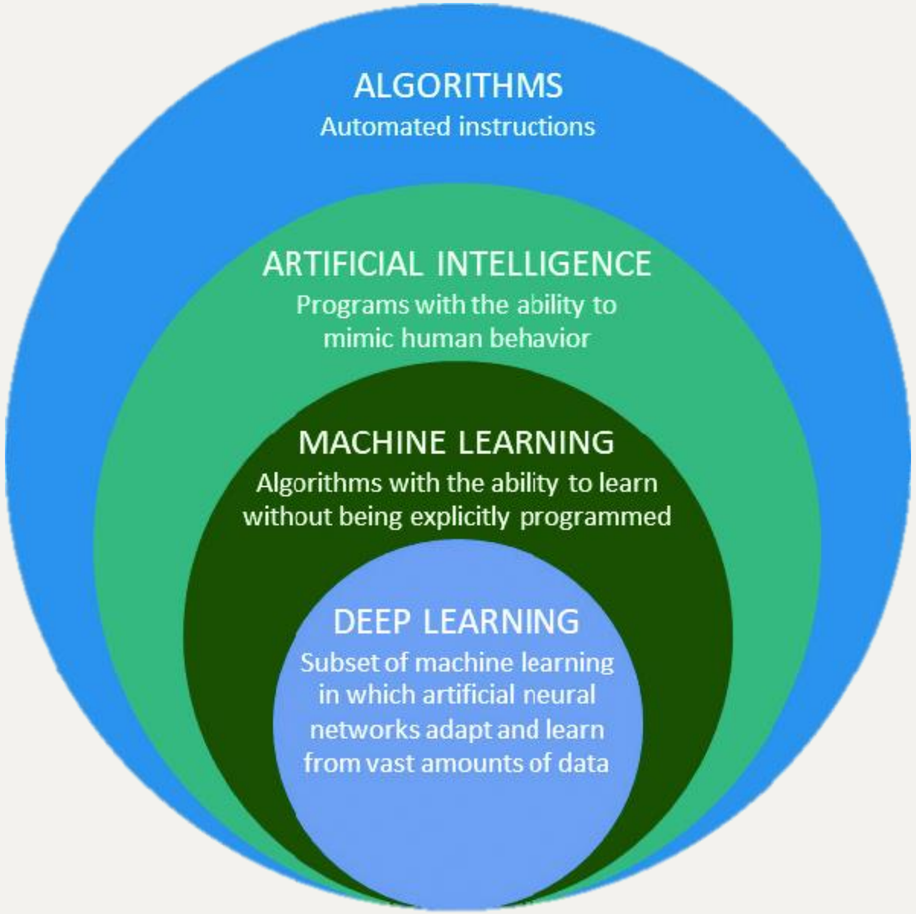
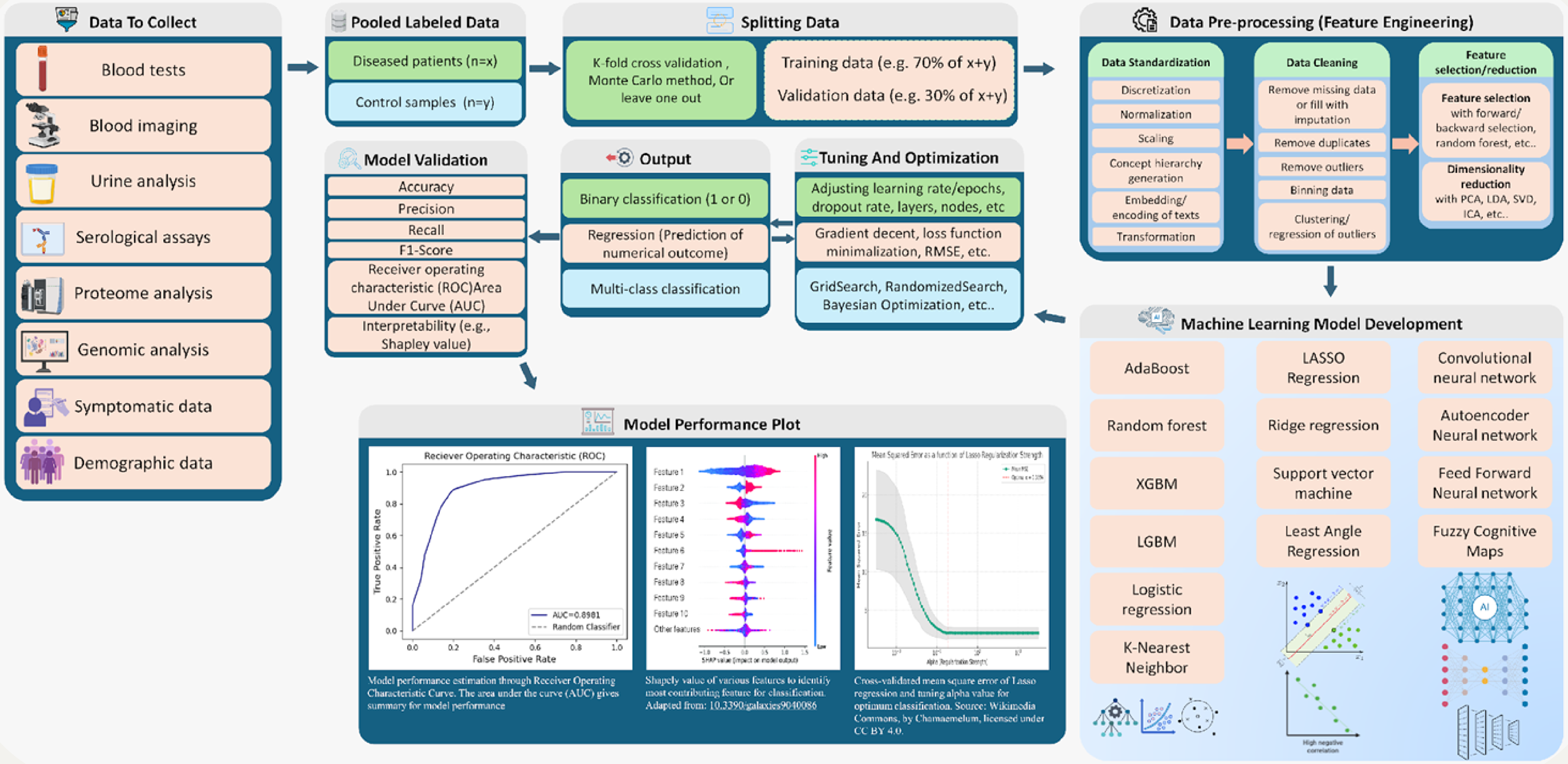


FIG.1: AI and its subcategories

1.1. General Workflow For Classification AI dev.

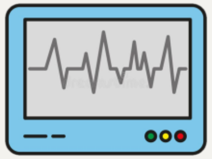


2. Dataset Overview

- **Source:**
  - Dataset of 303 patients with 13 clinical features and a binary target variable (target: 1 = heart disease, 0 = no heart disease).
  - Obtained from Machine Learning Repository at https://www.openml.org/search?type=data&status=active&id=43672.
- **Key Features:**

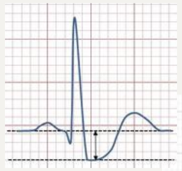
Feature	Description & Value Meaning
sex	1 = male, 0 = female
chest_pain_type	1 = Typical angina (classic heart-related pain), 2 = Atypical angina, 3 = Non-anginal pain, 4 = Asymptomatic (no pain)
fasting_blood_sugar	1 = true (>120 mg/dl), 0 = false (≤120 mg/dl)
target	1 = heart disease present, 0 = no heart disease

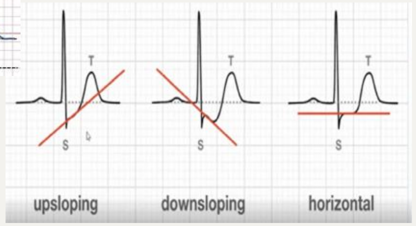
Feature	Description
Age	Age of participants
resting blood pressure	blood pressure upon admission
cholesterol	serum cholestoral conc (mg/dl)
fasting blood sugar	>120 mg/dl = 1 | <120 mg/dl =0
resting_ecg (electrocardiograph)	**0:** normal **1:** having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) **2:** showing probable or definite left ventricular hypertrophy by Estes' criteria
max_heart_rate	maximum heart rate achieved
exercise_angina	exercise induced angina (1/0)
oldpeak	ST depression induced by exercise relative to rest
ST_slope	slope of the peak exercise ST segment **1:** upsloping **2:** flat **3:** downsloping

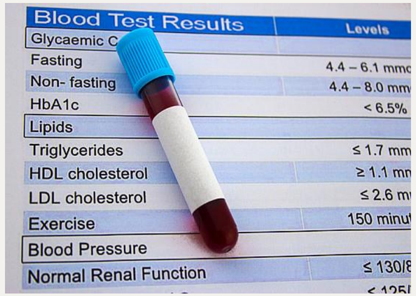


ECG measurements

Blood measurements

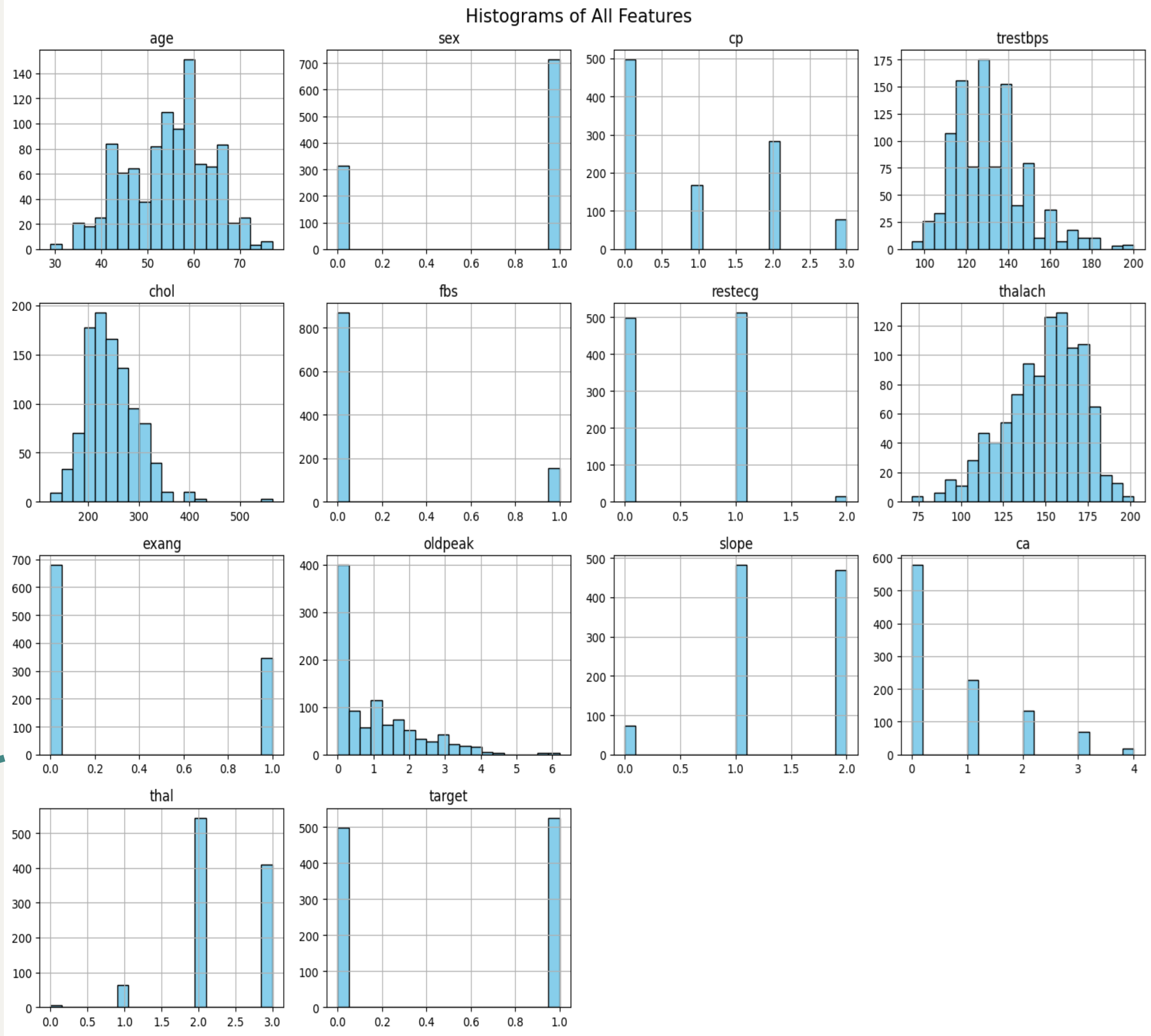






2.1. Exploratory Data Analysis

age	sex	chest_pai	resting_bp	cholesterc	fasting_bl	resting_ec	max_hear	exercise_a	oldpeak	ST_slope	target
40	1	2	140	289	0	0	172	0	0	1	0
49	0	3	160	180	0	0	156	0	1	2	1
37	1	2	130	283	0	1	98	0	0	1	0
48	0	4	138	214	0	0	108	1	1.5	2	1
54	1	3	150	195	0	0	122	0	0	1	0
39	1	3	120	339	0	0	170	0	0	1	0
45	0	2	130	237	0	0	170	0	0	1	0
54	1	2	110	208	0	0	142	0	0	1	0
37	1	4	140	207	0	0	130	1	1.5	2	1
48	0	2	120	284	0	0	120	0	0	1	0
37	0	3	130	211	0	0	142	0	0	1	0
58	1	2	136	164	0	1	99	1	2	2	1
39	1	2	120	204	0	0	145	0	0	1	0
49	1	4	140	234	0	0	140	1	1	2	1
42	0	3	115	211	0	1	137	0	0	1	0
54	0	2	120	273	0	0	150	0	1.5	2	0
38	1	4	110	196	0	0	166	0	0	2	1
43	0	2	120	201	0	0	165	0	0	1	0
60	1	4	100	248	0	0	125	0	1	2	1
36	1	2	120	267	0	0	160	0	3	2	1
43	0	1	100	223	0	0	142	0	0	1	0
44	1	2	120	184	0	0	142	0	1	2	0
49	0	2	124	201	0	0	164	0	0	1	0
44	1	2	150	288	0	0	150	1	3	2	1
40	1	3	130	215	0	0	138	0	0	1	0
36	1	3	130	209	0	0	178	0	0	1	0



- Checking if the data was unbiased, and non-randomly generated.
- Checking distribution, range, outliers, duplicates, etc..

3. Methodology

Preprocessing:

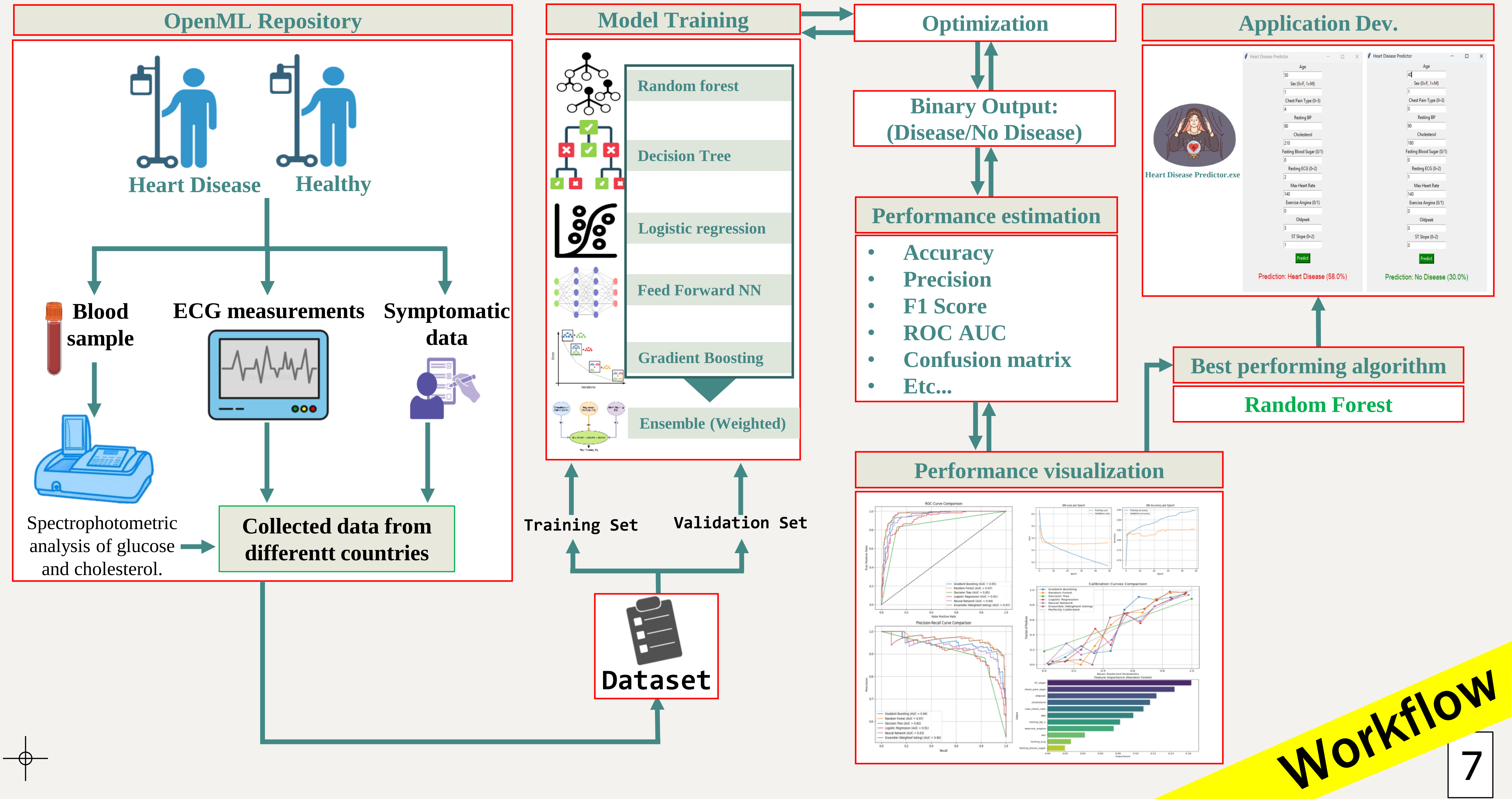
- Checked and cleaned data (no missing values in provided sample)
- Standardized numerical features

Model Selection:

- Random Forest Classifier for its accuracy and interpretability
- Gradient Boosting
- Neural network
- Decision Tree
- Logistic regression
- Ensemble model

Training & Evaluation:

- Data split into training and testing sets (e.g., 80/20)
- Performance measured by accuracy, precision, recall, F1-score, and ROC-AUC


OpenML Repository
Heart Disease
Healthy
Blood sample
ECG measurements
Symptomatic data
Spectrophotometric analysis of glucose and cholesterol.
Collected data from differentt countries
Model Training
Random forest
Decision Tree
Logistic regression
Feed Forward NN
Gradient Boosting
Ensemble (Weighted)
Training Set
Validation Set
Dataset
Optimization
Binary Output: (Disease/No Disease)
Performance estimation
Accuracy
Precision
F1 Score
ROC AUC
Confusion matrix
Etc...
Performance visualization
ROC Curve Comparison
Precision-Recall Curve Comparison
Calibration Curves Comparison
Feature Importance (Random Forest)
Application Dev.
Heart Disease Predictor.exe
Heart Disease Predictor
Prediction: Heart Disease (58.0%)
Prediction: No Disease (30.0%)
Best performing algorithm
Random Forest
Workflow

Results

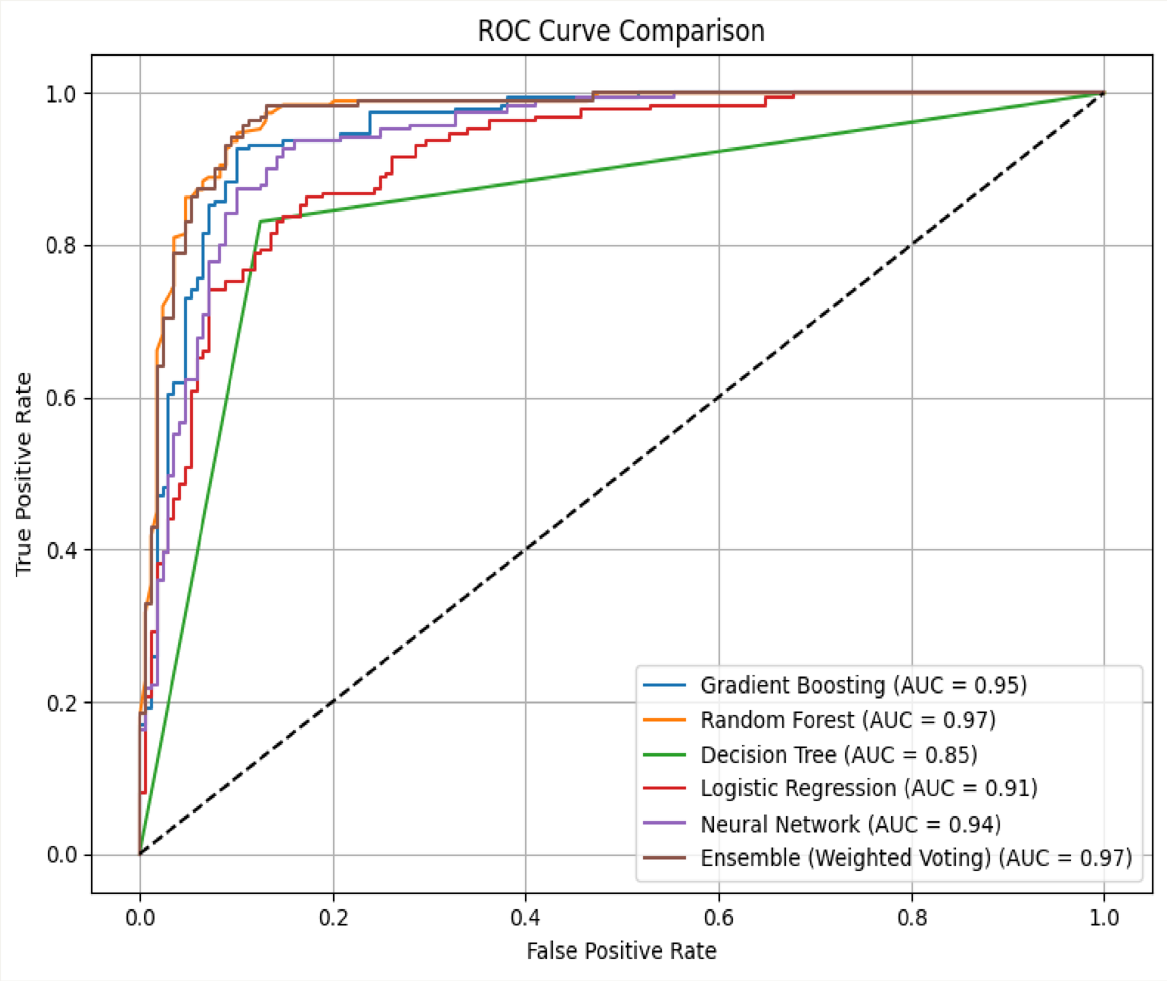


FIG. 2 Compares the performance of different machine learning models using **Receiver Operating Characteristic (ROC)** curves.

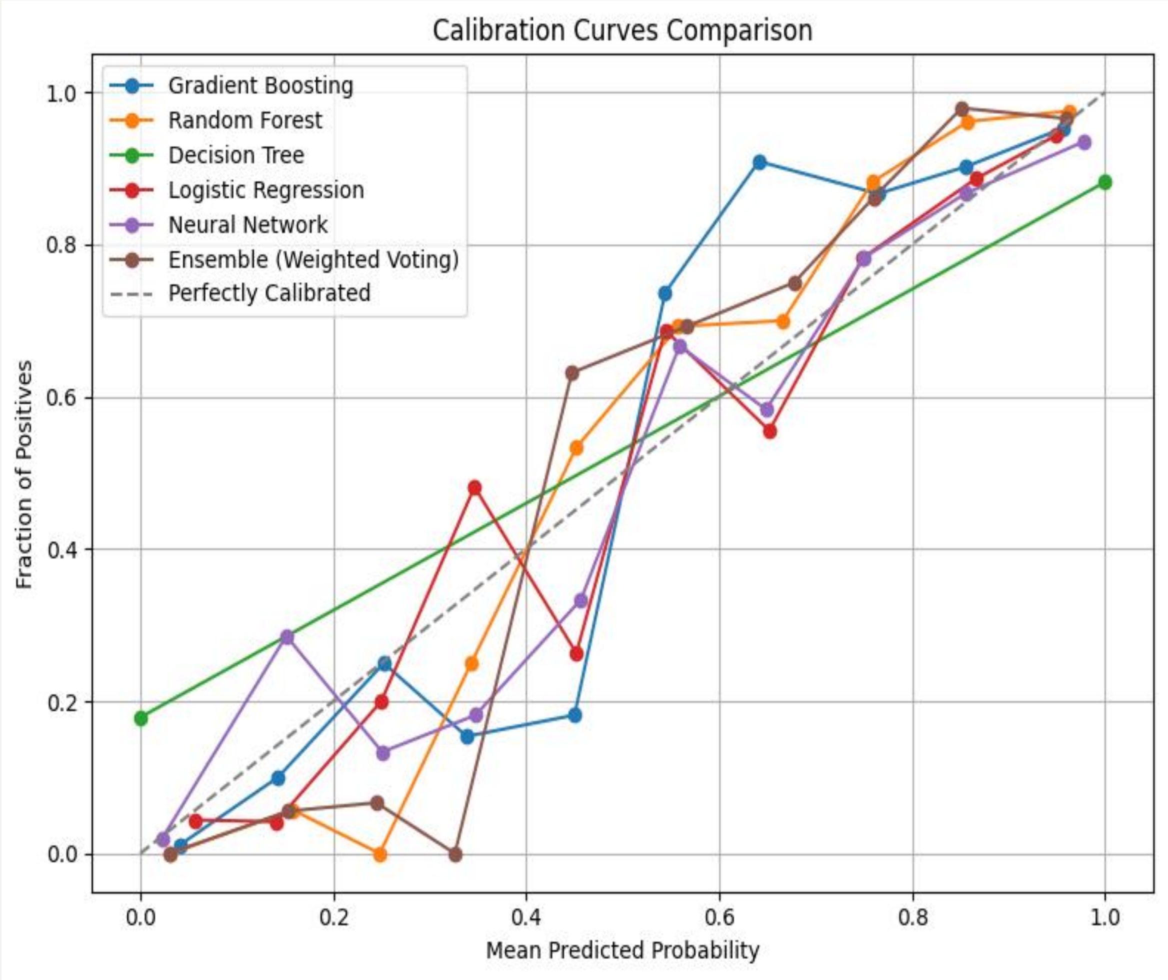


FIG. 3 Assesses how well each model's predicted probabilities match real-world outcomes, which is crucial for clinical decision-making.

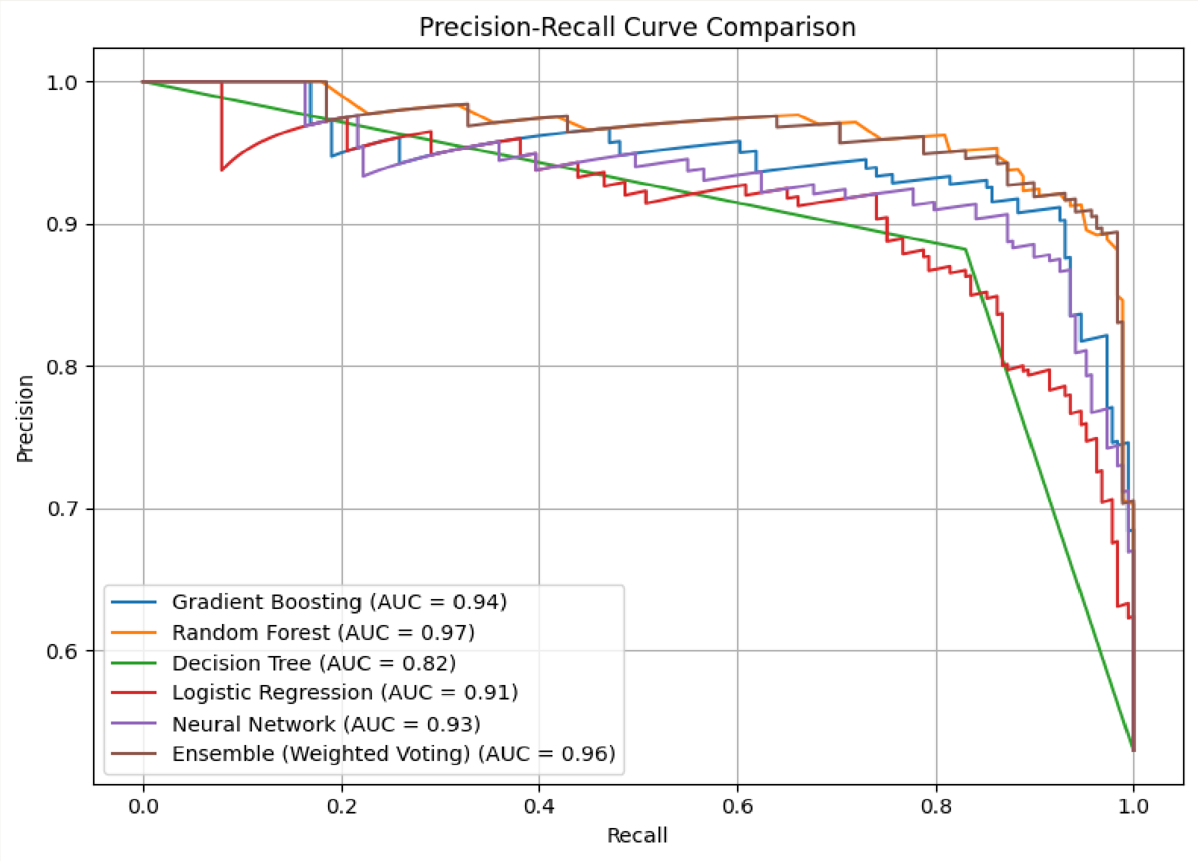


FIG. 4 Evaluates model performance using **precision-recall curves**, which is particularly informative for imbalanced medical datasets.

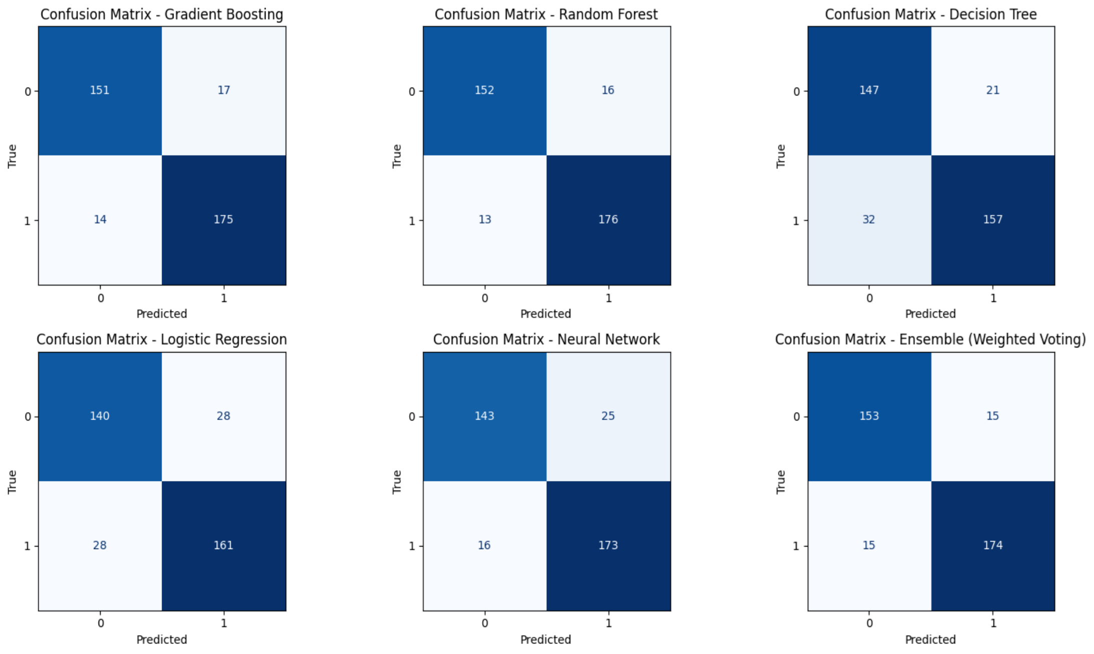


FIG. 5 Compares the classification performance of five different machine learning models through their confusion matrices, showing how each model distinguishes between patients with and without heart disease.

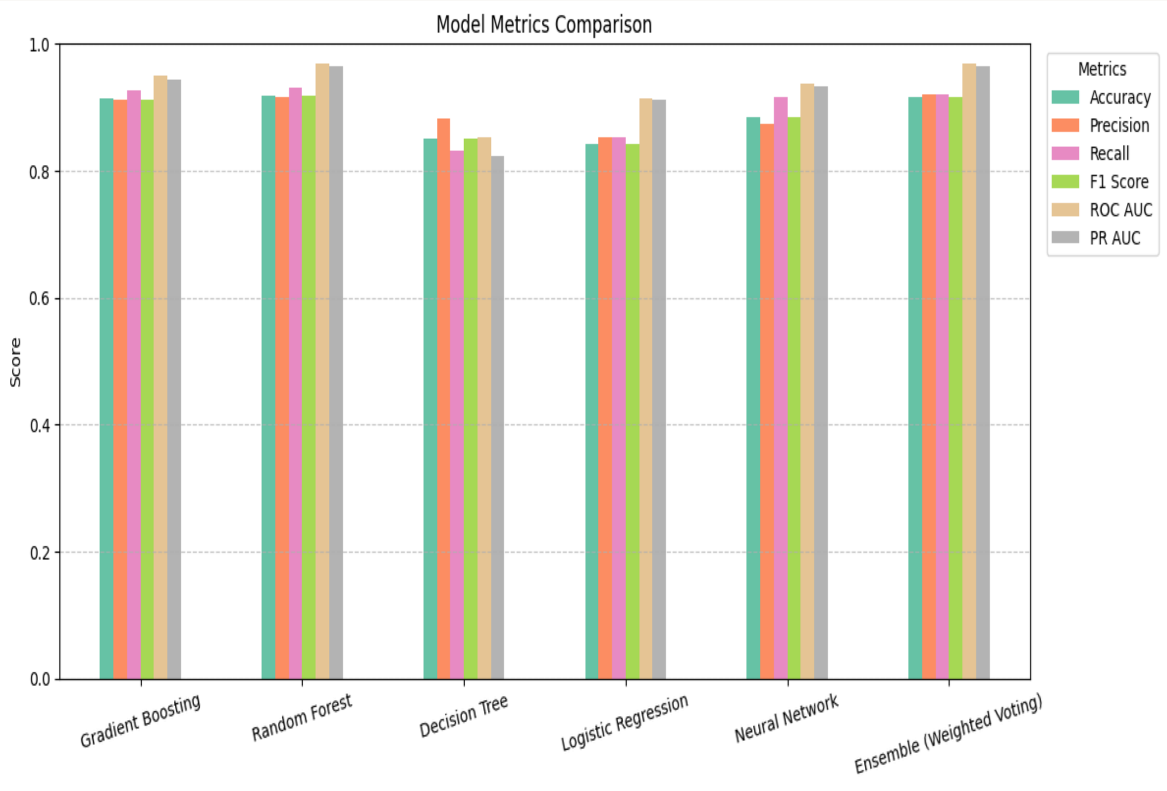


FIG. 6 Presents a side-by-side evaluation of six machine learning models across six key performance metrics for heart disease detection

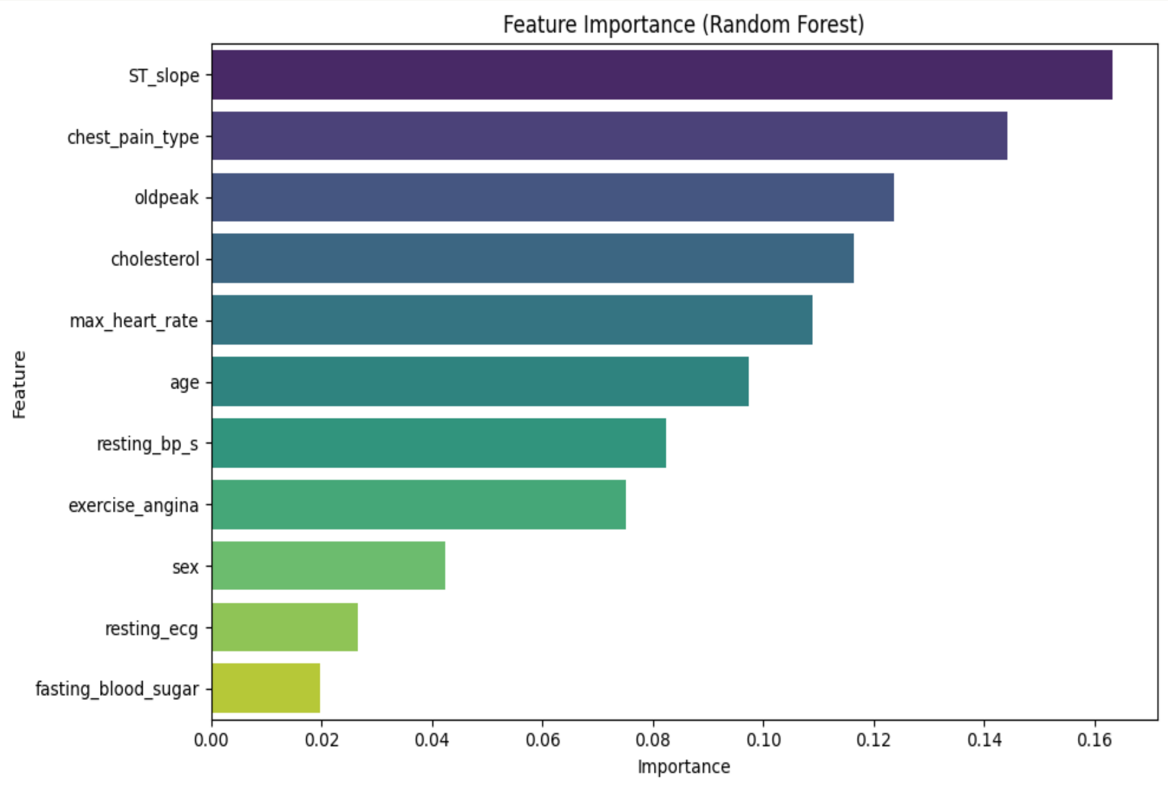


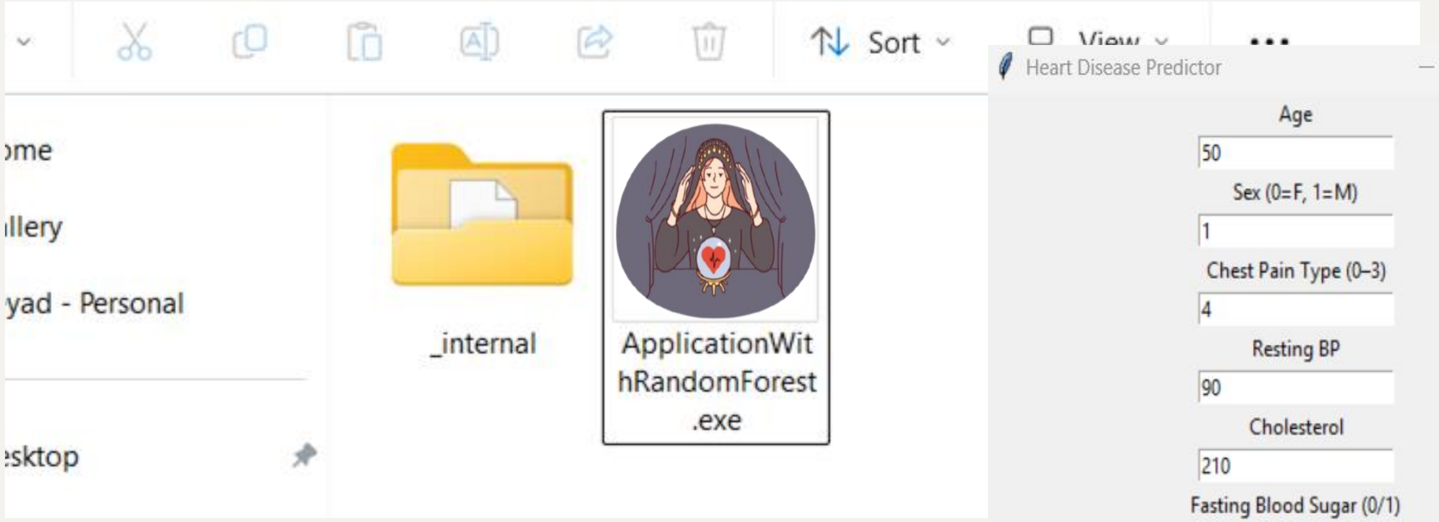
FIG. 7 Reveals the most influential medical factors for predicting heart disease, as determined by a Random Forest algorithm.

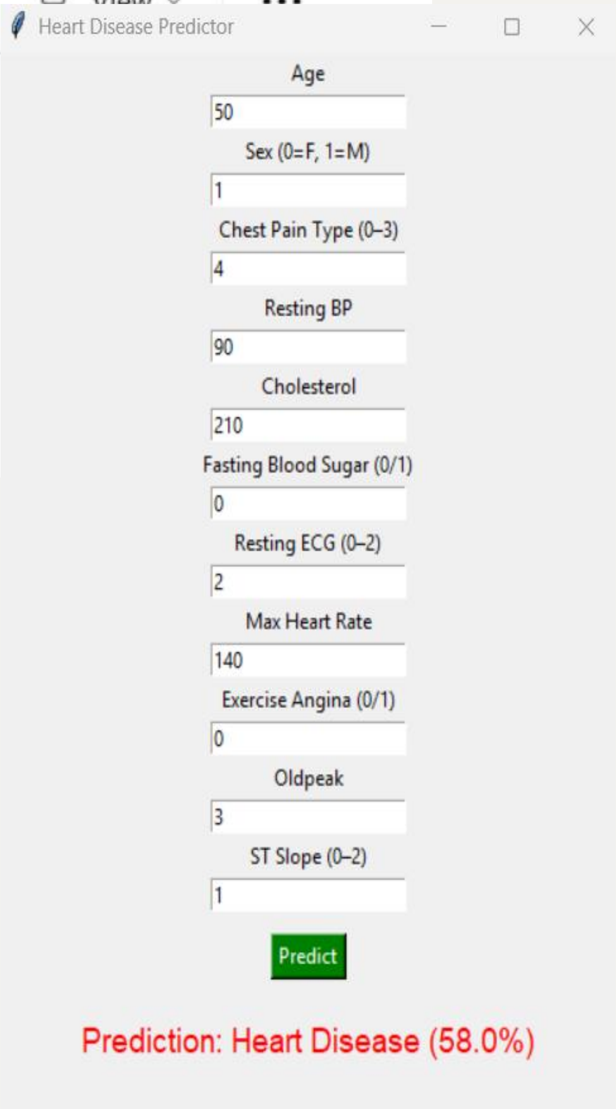
* Summary Results For Model Performance

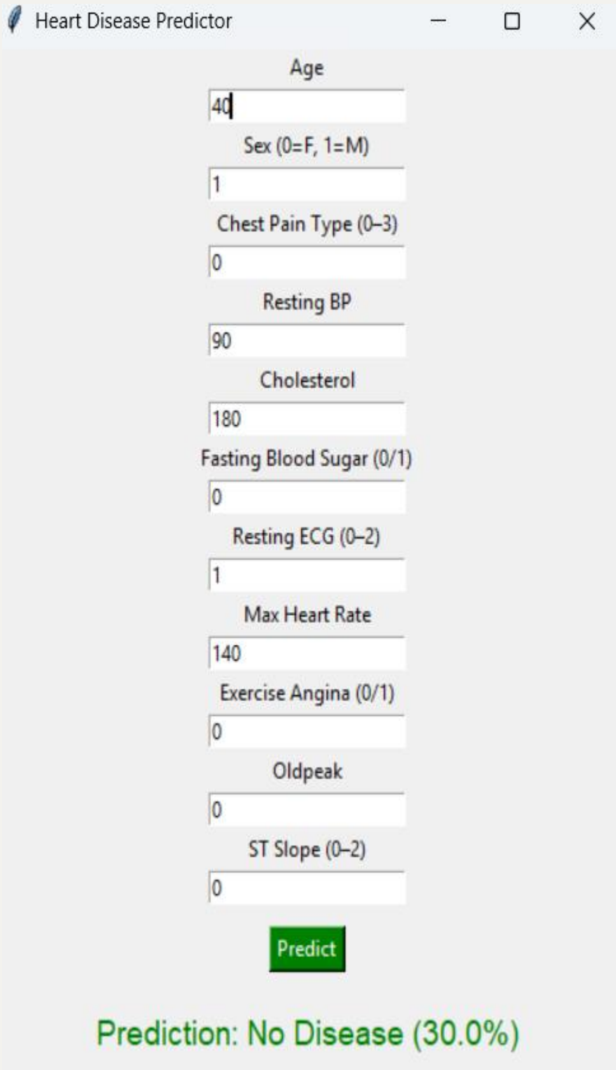
	Accuracy	Precision	Recall	F1 Score	ROC AUC	PR AUC
Gradient Boosting	0.913165	0.911458	0.925926	0.912771	0.950145	0.944635
Random Forest	0.918768	0.916667	0.931217	0.918399	0.969309	0.965629
Decision Tree	0.851541	0.882022	0.830688	0.851424	0.852844	0.822321
Logistic Regression	0.843137	0.851852	0.851852	0.842593	0.913045	0.911826
Neural Network	0.885154	0.873737	0.915344	0.884337	0.938146	0.932695
Ensemble (Weighted Voting)	0.915966	0.920635	0.920635	0.915675	0.968191	0.964948

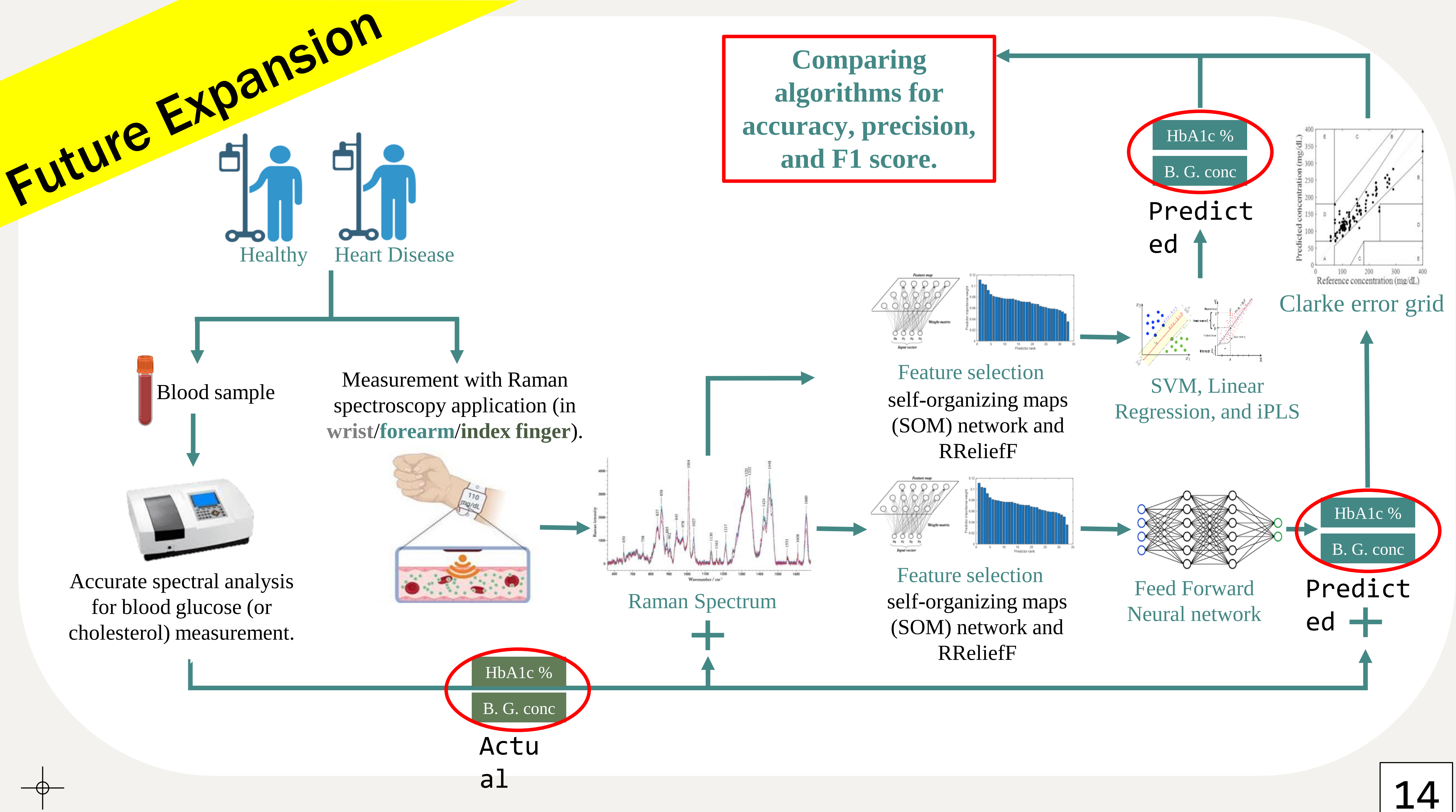
Comprehensive quantitative comparison of six machine learning models across seven key evaluation metrics.

Working Desktop Application Draft



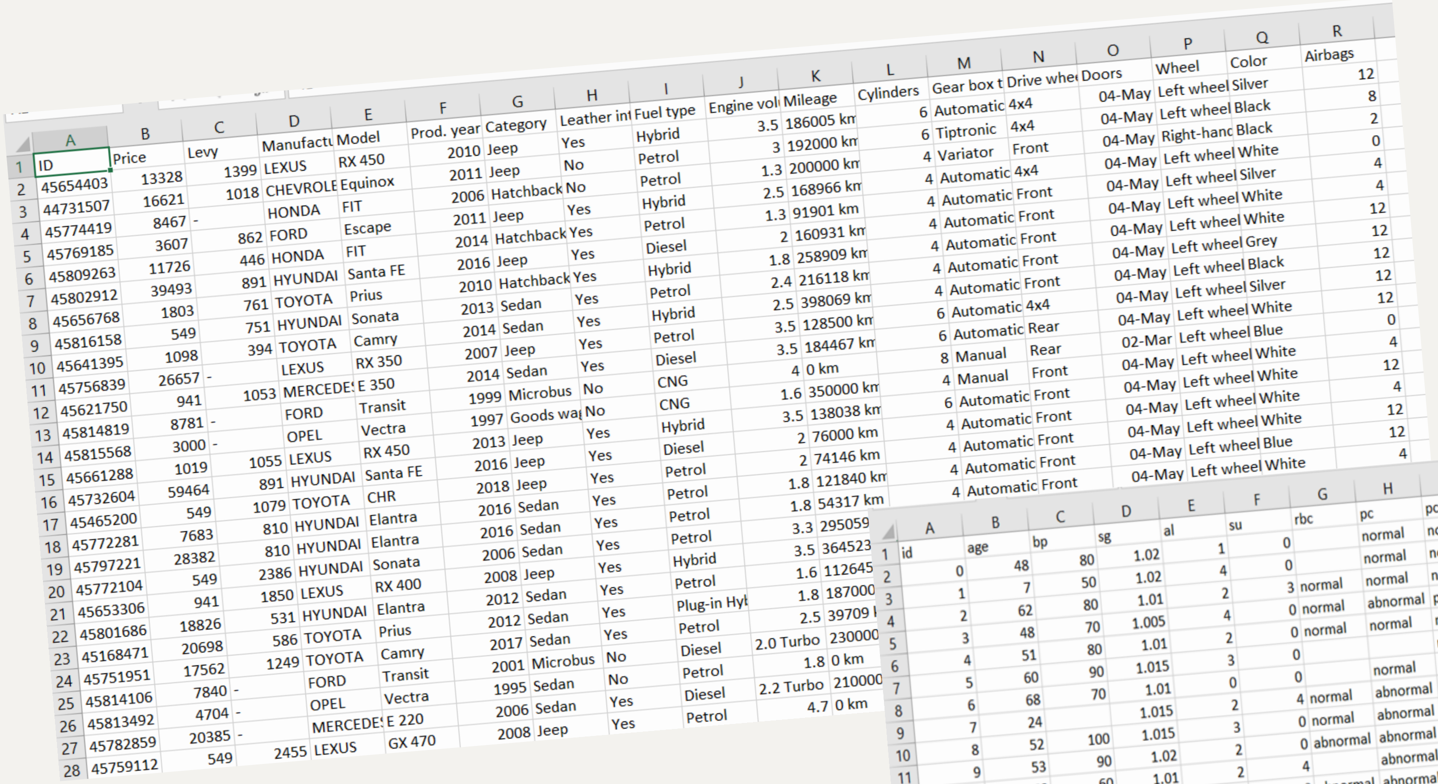




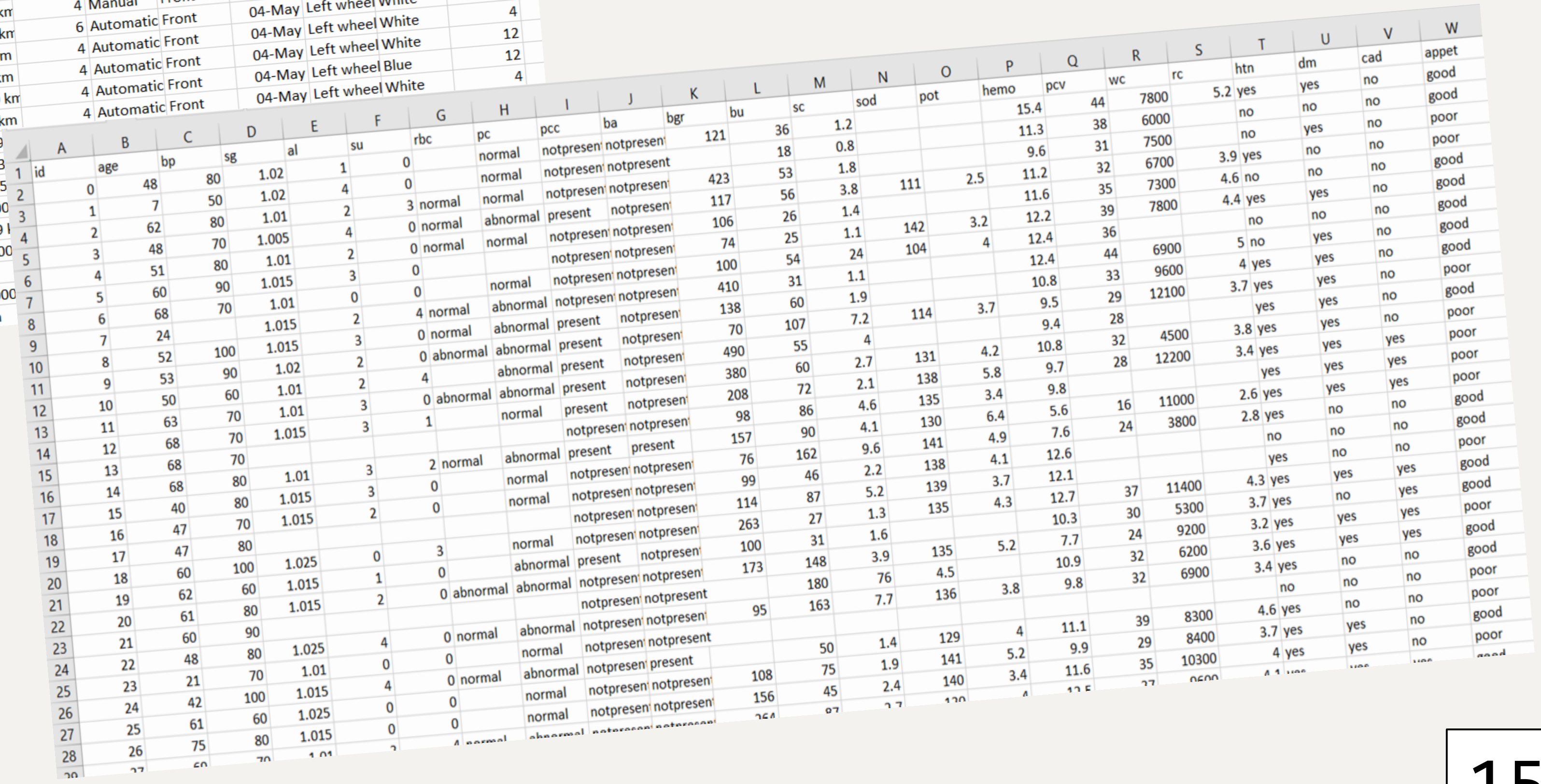


Future Expansion
Comparing algorithms for accuracy, precision, and F1 score.
Healthy
Heart Disease
Blood sample
Measurement with Raman spectroscopy application (in wrist/forearm/index finger).
Accurate spectral analysis for blood glucose (or cholesterol) measurement.
110 mg/dL
Raman Spectrum
Feature selection
self-organizing maps (SOM) network and RReliefF
Feature selection
self-organizing maps (SOM) network and RReliefF
SVM, Linear Regression, and iPLS
Feed Forward Neural network
HbA1c %
B. G. conc
Predicted
HbA1c %
B. G. conc
Predicted
HbA1c %
B. G. conc
Actual
Clarke error grid
Reference concentration (mg/dL)
Predicted concentration (mg/dL)

5.Other Real-World Projects



Used-cars prices prediction



Kidney Disease prediction

Thank you.

For more information, please contact:
zeyad.mansour@aucegypt.edu
alexbe323@gmail.com
Rabab.asar17@fsc.bu.edu.eg
Enjy_sayed@aun.edu.eg
Esraa_Hussien@aun.edu.eg