Prediction Modeling for Mortality in ICU Patients with Heart Failure

Mehran Khodadadzadeh
May 2024

Based on: Li, F., Xin, H., Zhang, J., Fu, M., Zhou, J., & Lian, Z. (2021). Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database.

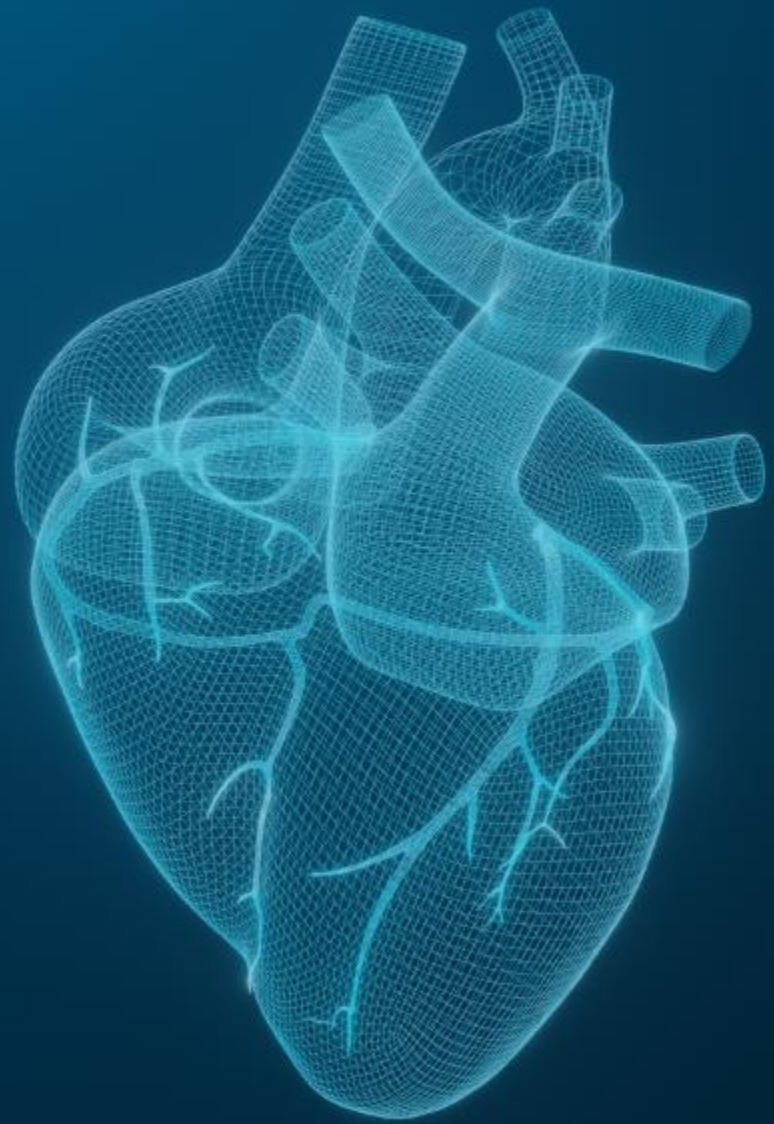
Introduction

Predicting mortality in ICU patients with heart failure helps doctors provide better care, make informed decisions, and efficiently use hospital resources to save lives.

- **Goal**
  - Develop a model to predict in-hospital mortality for ICU patients with heart failure.
  - Identify the most significant features impacting mortality.
  - Analyze the effectiveness of various machine learning models.

Data Source and Features

Data Source: MIMIC-III database.



Features:

- **Demographics**
  - Age, Gender, Ethnicity, etc.
- **Vital Signs:**
  - Heart rate, Blood pressure, Respiratory rate, etc.
- **Comorbidities:**
  - Diabetes, Hypertension, Ischaemic heart disease, etc.
- **Laboratory Results:**
  - Blood glucose levels, Creatinine levels, etc.

Data Preprocessing

Final Dataset:**Shape**: 7280 rows × 37 columns

Steps:

- **Remove Unnecessary Columns**:
- **Remove Duplicates**

- **Extract and Add Units to Feature Names**:
  - Incorporated units into feature column names for clarity.

- **Handle Missing Values**:
  - Removed null entries, keeping the first non-null value.
  - Filtered out rows where missing values exceeded 25%.

- **Outlier Detection and Removal**:
  - Calculated the interquartile range (IQR) and removed outliers beyond 1.5 times the IQR.

- **Feature Reduction**:
  - Dropped features with more than 20% missing values to preserve data quality.

- **Data Imputation**:
  - Replaced missing values with the median of respective columns.

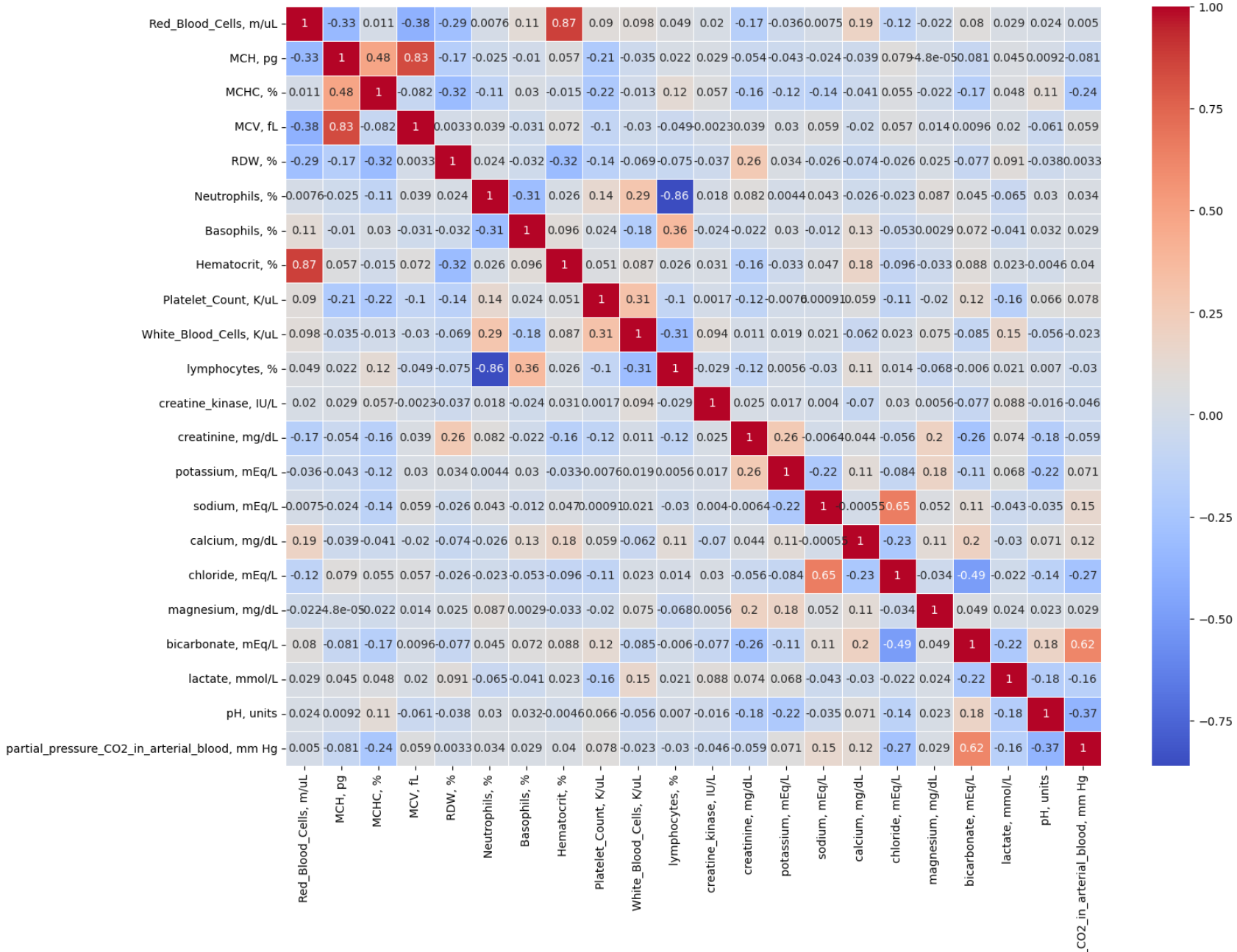
Data Analysis and Visualization

Correlation Matrix, bar plot, box plot

- **Central Tendency Measures**:
  - Calculated mean, median, and mode to understand the typical values in the data distribution.

- **Label Distribution Analysis**:
  - Analyzed the distribution of labels.
  - Noted that labels are not highly imbalanced but considered methods to enhance the training set.

- **Descriptive Statistics and Visualization**:
  - Visualized age, ethnicity, and gender distributions.
  - Found that the dataset consists mostly of white people.
  - Mortality rates are nearly the same between men and women.
  - Mortality rates are similar across different ethnicities.


Red_Blood_Cells, m/uL
MCH, pg
MCHC, %
MCV, fL
RDW, %
Neutrophils, %
Basophils, %
Hematocrit, %
Platelet_Count, K/uL
White_Blood_Cells, K/uL
lymphocytes, %
creatine_kinase, IU/L
creatinine, mg/dL
potassium, mEq/L
sodium, mEq/L
calcium, mg/dL
chloride, mEq/L
magnesium, mg/dL
bicarbonate, mEq/L
lactate, mmol/L
pH, units
partial_pressure_CO2_in_arterial_blood, mm Hg
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Feature Selection

Methods Used:

- **Random Forest Importance:**
  - Identified top features like AGE, RDW, creatinine, etc.
- **XGBoost Importance:**
  - Similar features identified with slight variations.
- **Combined Feature Importance:**
  - Averaged rankings from both methods.

Modeling Techniques

1. **Data Preparation:**
   Loading and Cleaning: removing unnecessary columns ('SUBJECT_ID', 'HADM_ID').
   Separated features (X) and target variable ('EXPIRE_FLAG').
   training (80%) and testing (20%) sets.

2. **Logistic Regression**: A linear model used for binary classification. It predicts the probability of the target variable.
   Implementation: Used a pipeline with `StandardScaler` and `LogisticRegression` to standardize data and train the model.

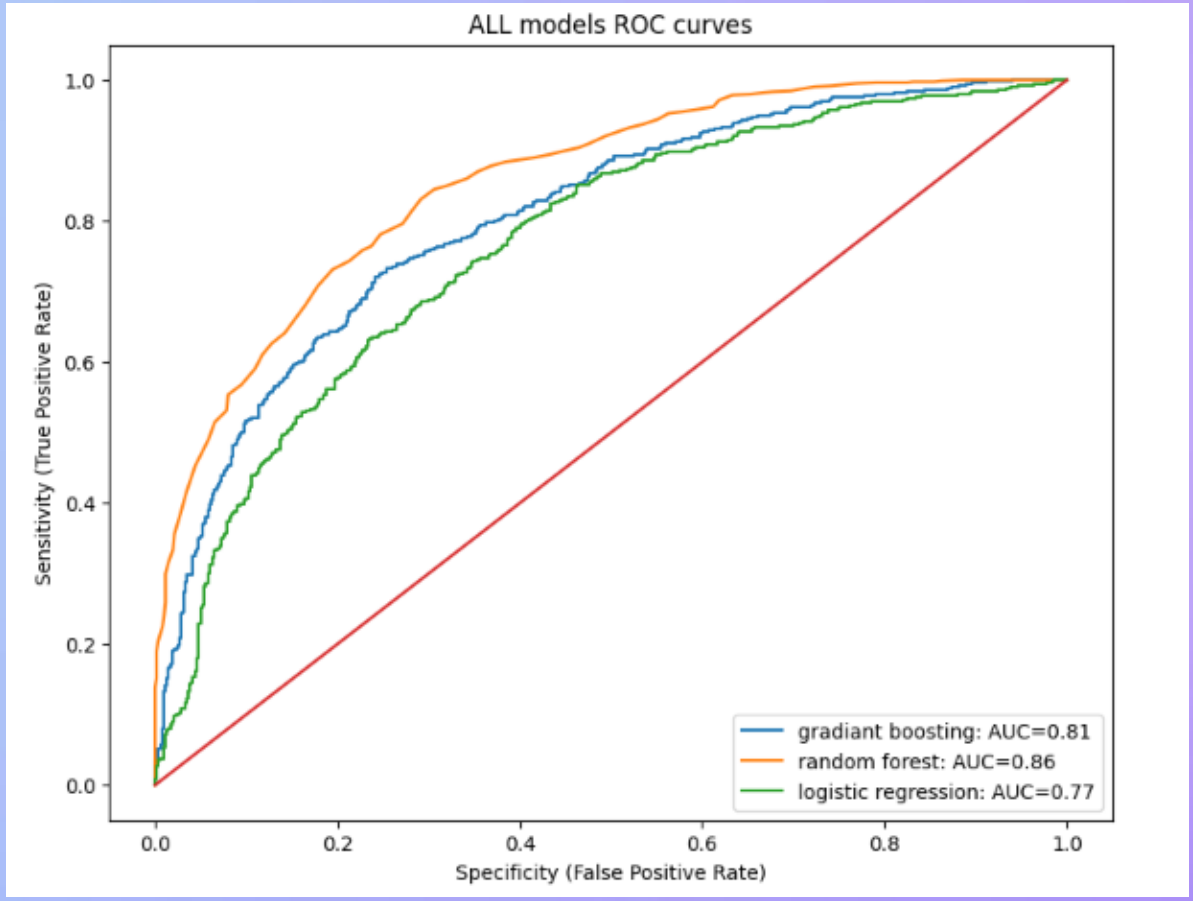
3. **Random Forest**: An ensemble method that uses multiple decision trees to improve classification accuracy.
   Implementation: Trained a `RandomForestClassifier` on the training data with default hyperparameters and evaluated its performance.

4. **Gradient Boosting**: An ensemble technique that builds models sequentially to correct errors of previous models.
   Implementation: Applied `GradientBoostingClassifier` with tuned parameters (`n_estimators=220, learning_rate=0.09`) for training and evaluation..

Results

Compared all models using ROC curves to assess their true positive and false positive rates

- **Logistic Regression:**
  - Accuracy: 71.29%
- **Random Forest:**
  - Accuracy: 77.88%
- **Gradient Boosting:**
  - Accuracy: 73.01%

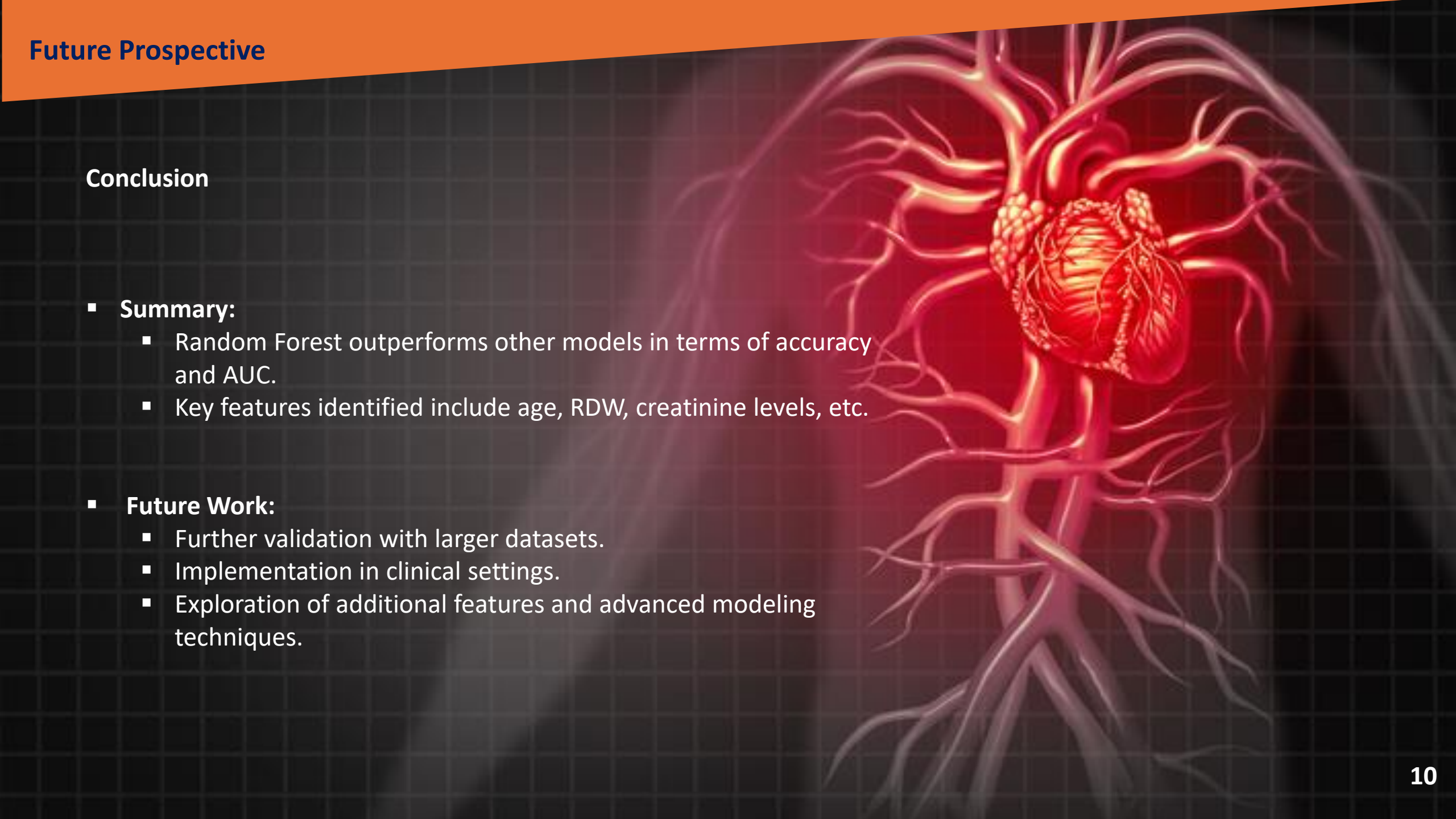


Future Prospective

Conclusion

- **Summary:**
  - Random Forest outperforms other models in terms of accuracy and AUC.
  - Key features identified include age, RDW, creatinine levels, etc.

- **Future Work:**
  - Further validation with larger datasets.
  - Implementation in clinical settings.
  - Exploration of additional features and advanced modeling techniques.



THANK YOU!