Exploring the Factors Linked to Heart Disease in the U.S. during 2020

Dhriti Raval*, Liangjie Lu*, Prakriti Sarkar* and Sara Abril Guevara*
Department of Statistics & Department of Public Health, University of California, Davis, One Shields Avenue, Davis, CA 95616, U.S.
E-mail: {draval, ljlu, psarkar, ssabrilguevara}@ucdavis.edu
*Authors are listed in alphabetical order.

***Abstract*—This paper leverages machine learning techniques to develop predictive models for heart disease using epidemiological data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS). The dataset encompasses 319,795 records across 169 variables related to demographics, lifestyle factors, and medical history. After preprocessing the raw data, two algorithms - Logistic Regression and Random Forest - are employed to construct classification models with the binary target variable of heart disease presence. The study maintains key assumptions underlying these methods regarding relationships, variable independence, model robustness, and feature relevance. Data analysis is conducted, including exploratory analysis and quadratic/interaction term generation, to capture potential non-linearities. The anticipated results exceed 90% accuracy, with Random Forest likely outperforming Logistic Regression. By unraveling the risk profile of heart disease, this research aims to equip healthcare practitioners with personalized assessment tools to curb the incidence and burden of cardiovascular mortality. The methodology and large-scale real-world dataset offer insights for improving preventive strategies.**

***Index Terms*—Heart disease, Machine Learning, Logistic regression, Random forest, Sparsity.**

I. Introduction

Cardiovascular diseases (CVDs), with heart disease as their most prevalent manifestation, remain the foremost cause of mortality globally, imposing significant challenges for both clinical practice and public health policy [1]. Moreover, there were large increases in Year of Life Lost(YLLs) from cardiovascular diseases between 1990 and 2017 [1]. The intricacy of heart disease etiology, combined with its multifactorial risk factors [2], underscores an urgent need for advanced diagnostic and predictive models to facilitate early detection and intervention strategies.

Despite advancements in biomedical research elucidating the pathophysiological underpinnings of heart disease [3], the clinical translation of this knowledge into effective prognostic tools has been hindered by the heterogeneous nature of the disease and the diversity in patient populations. Traditional risk assessment models, while useful, often fail to capture the nuanced interplay between genetic, environmental, and lifestyle factors that contribute to heart disease development [4].

Cardiovascular diseases, including heart disease, continue to be the leading cause of mortality worldwide. To address this challenge, there is a need for advanced diagnostic and predictive models that can facilitate early detection and intervention strategies. Incorporating diverse datasets with representative demographic information is essential to prevent algorithmic biases that could lead to discriminatory models in healthcare applications. Additionally, proactive monitoring and testing during model development are crucial for identifying and resolving any biases before deployment to ensure transparency and trustworthy relationships with end-users in the healthcare sector. Moreover, ensuring transparency alongside fairness is critical for safe deployment of machine learning models in healthcare settings." In this research study, machine learning techniques such as Support Vector Machine (SVM) [5], Logistic regression, Random Forest [6], and XGboost [7] were utilized to develop predictive models for heart disease.

In recent years, the proliferation of large-scale epidemiological datasets, such as the Behavioral

Risk Factor Surveillance System (BRFSS) data provided by the Centers for Disease Control and Prevention (CDC) [8], [9], has offered an unprecedented opportunity to apply machine learning (ML) techniques in unraveling the complexities of heart disease. These ML approaches, renowned for their capability to discern patterns within high-dimensional data, hold the promise of enhancing predictive accuracy and providing personalized risk assessments [10].

This paper seeks to harness the potential of advanced ML methodologies to identify and quantify the impact of key indicators associated with heart disease within the U.S. population. Leveraging the comprehensive BRFSS dataset, which encompasses a wide array of demographic, behavioral, and clinical variables, we aim to construct and validate a series of predictive models that can reliably ascertain an individual's risk of heart disease. In doing so, we not only contribute to the academic discourse on the utility of ML in clinical epidemiology but also endeavor to support healthcare practitioners in devising more effective preventive measures tailored to the risk profile of individual patients [11].

With the ultimate objective of diminishing the prevalence and burden of heart disease, this study adheres to the rigor of computational modeling and statistical validation to offer insights that are both scientifically robust and clinically relevant. Our analysis is poised to enrich the collective understanding of heart disease dynamics and pave the way for the development of intelligent healthcare solutions that are adaptive to the evolving landscape of cardiovascular risk factors.

While this study focuses on predicting cardiac failure, the approach of utilizing machine learning techniques and clinical data could potentially be applied to other medical conditions. The challenges of developing and implementing machine learning models for disease prediction include the complexity of the disease and the need to integrate diverse data sources. However, with careful consideration of these challenges, machine learning models could potentially be developed for other medical conditions. Further research is needed to explore the applicability of this approach to other diseases and to refine the methodology for optimal performance.

II. Problem Definition

The problem being addressed is identifying key indicators or factors associated to heart disease, given its prevalence and significance as a leading cause of death in the U.S. We hope to answer:

1) Which variables or factors have an important association on the likelihood of having heart disease?
2) How can these indicators be used to predict the onset or occurrence of heart disease in individuals?
3) How do these factors stack up in terms of their prediction of heart disease?

Given that about half of all Americans have at least one of the three major risk factors, understanding these key indicators can lead to better preventive measures, healthcare policies, and patient care strategies.

III. Data Description

Our data originates from the CDC's BRFSS, which collects health status information from U.S. residents. The BRFSS was founded in 1984 and initially included 15 states. Currently, it gathers data from all 50 states, the District of Columbia, and three U.S. territories. With over 400,000 adult interviews completed annually, BRFSS is the world's biggest continually operating health survey system. The newest dataset contains information from 2023. Selected from an original pool of 300 variables for their relevance to heart disease [8], our dataset includes 40 key variables. After preprocessing, which involved incorporating all meaningful and second-order interaction terms, we have a data matrix comprising 319,795 rows and 169 columns. Within this matrix, a single column is dedicated to the binary target variable 'HadHeartAttack', denoting the presence or absence of heart disease in respondents. The remaining 168 columns function as predictive variables.

IV. Literature Review

Substantial and esteemed research has been conducted in the field of identifying effective methods to predict potential heart disease in individuals. Below is a selection of these notable contributions:

1) Sumwiza et al. [12] focused on improving the accuracy of cardiovascular disease (CVD)

predictions using machine learning models, particularly the Random Forest (RF) algorithm. Their study used a Kaggle dataset with 14 health-related features, undergoing extensive data preprocessing. Outliers were eliminated using the Interquartile Range (IQR) method, reducing the dataset size from 1025 to 769 records. They further refined the feature set from 14 to 13 by excluding the less significant 'fbs' feature, based on correlation coefficient analysis and feature importance. The dataset was split into training and testing sets to evaluate the RF algorithm against K-Nearest Neighbors (KNN) and Support Vector Machine (SVM). The RF model excelled with a 99% accuracy rate, surpassing Logistic Regression (87%), KNN (95%), and SVM (85%). The study assumes the dataset's representativeness and the validity of outlier removal via the IQR method. It also implicitly relies on the assumption that a dataset of 1025 records is sufficient for developing a reliable, generalizable model and that the RF algorithm is adept at managing high-dimensional data to yield precise, low-variance predictions.
2) Du et al. [13] analyzed electronic health records (EHRs) of 42,676 hypertensive individuals, including 20,156 who subsequently developed Coronary Heart Disease (CHD). They used EHR data from 1-3 years prior to CHD onset for positive cases and from a comparable 3-year disease-free interval for negative cases. The study was predicated on the assumptions that the EHR data were sufficient, high-quality, and generalizable, and that the data contained nonlinear relationships and time-dependent features with clinically relevant predictive factors. They found that the XGBoost ensemble machine learning model demonstrated high accuracy in predicting CHD development within three years, achieving an Area Under the Curve (AUC) of 0.943 on the test dataset.
3) Istiak Mahmud et al. [14] present a metamodel that integrates algorithms like Gaussian Naive Bayes, Random Forest Classifier, Decision Tree, and k-Nearest Neighbor. This metamodel is designed to enhance predictive accuracy and minimize biases by combining the strengths of these individual algorithms. It was evaluated using a comprehensive dataset amalgamated from five well-known cardiac datasets, ensuring a wide representation of patient demographics and clinical features. The study's key contributions include improved predictive models and a diverse dataset, with a comparative analysis highlighting its superiority over existing machine learning models, evidenced by an 87% accuracy rate in predicting heart failure. The study implicitly assumes that an ensemble of varied ML methods will outperform their individual components by leveraging their collective strengths and offsetting weaknesses. Nonetheless, it presumes without empirical evidence that traditional statistical methods and the individual algorithms are less effective than the proposed metamodel. Notably, the study omits specific configurations for the Light Gradient Boosting Machine (LGBM), Ridge Regression, and Bagging methods, raising questions about the validity of the results and necessitating a cautious approach to their interpretation.

V. Methods

A. Algorithms Description

Support Vector Machine (SVM), Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost) are widely cited in scientific literature for heart disease prediction due to their ability to manage high-dimensional data and produce interpretable results [6], [15]–[17]. Here are their descriptions:

1) SVM: The main principle of SVM is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.
   - **Hyperplane:** A decision plane that separates between a set of objects having different class memberships.
   - **Support Vectors:** Data points that are closest to the hyperplane, which influence the position and orientation of the hyperplane.
   - **Margin:** The gap between two lines on the closest class points. A good margin

is one where this separation is as wide as possible.

2) Logistic Regression: This method has already been implemented as per the dataset description. It's suitable because the outcome is binary (had heart disease or did not). Logistic Regression is widely used for binary classification problems [15].

3) Random Forest: The general version of the Random Forest algorithm is delineated in Algorithm 1, following the definition provided by Breiman (2001) [6] in his seminal work on Random Forests.

```
Algorithm 1: Random Forest Algorithm
Input: Training dataset $D = \{(x_i, y_i)\}$,
Number of trees $T$,
Number of features to consider at each
split $f$.
Output: A Random Forest model
        composed of $T$ decision trees.
________________________________________
Procedure:
for $t = 1$ to $T$ do
    1. Create a bootstrap sample $D_t$ from
     $D$.
    2. Grow a decision tree $Tree_t$ from $D_t$:
    repeat
        a. Select $f$ features randomly
         from the feature set.
        b. Determine the best split on
         the selected features to
         partition the data.
        c. Split the node into two child
         nodes.
    until termination condition is met;
    3. Add $Tree_t$ to the forest.
end
Prediction: For a new instance $x$, predict
 by aggregating (majority vote or
 averaging) the predictions of the $T$ trees.
```

Since we train random forest model trained for classification task, let $\hat{C}_t(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{T,\mathrm{rf}}(x) = \text{majority vote}\{\hat{C}_t(x)\}_{t=1}^{T}$ [18].

It's a versatile classifier capable of handling imbalanced datasets and can rank features based on their importance in prediction. It works well with complex datasets that have multiple risk factors [17].

4) XGBoost: The generalized version of the XGBoost algorithm is outlined in Algorithm 2, as defined in the work by Chen et al. (2016) [7] on XGBoost.

```
Algorithm 2: XGBoost Algorithm
Input: Training dataset $D = \{(x_i, y_i)\}$,
Number of iterations (trees) $N$,
Learning rate $\eta$,
Loss function $L(y, \hat{y})$.
Output: An ensemble model composed
        of $N$ boosted trees.
________________________________________
Initialization:
Initialize model with a constant value:
 $f_0(x) = \underset{\theta}{\mathrm{argmin}} \sum_i L(y_i, \theta)$.
for $n = 1$ to $N$ do
    1. Compute the gradients and hessians
     for the loss function:
      $g_i = \partial_{\hat{y}_i} L(y_i, \hat{y}_i)$, $h_i = \partial^2_{\hat{y}_i} L(y_i, \hat{y}_i)$,
     where $\hat{y}_i = f_{n-1}(x_i)$.
    2. Construct and train a tree $T_n$ using
     $\{(x_i, g_i, h_i)\}$.
      - Each leaf of the tree outputs a
     weight value.
    3. Update the model:
      $f_n(x) = f_{n-1}(x) + \eta \cdot T_n(x)$.
end
Final Model:
The resulting model is $f_N(x)$, a sum of $N$
 trees.
```

B. Maintained Assumptions

In our study, we employ four primary algorithms: SVM, Logistic Regression, Random Forest, and XGBoost. These methods are chosen based on their suitability for the dataset and the nature of the problem, which involves predicting the binary outcome of heart disease presence.

SVM assumes (i) a clear margin of separation between classes, where it seeks to maximize the margin to improve classification accuracy; (ii) the data can be linearly separable or transformed to be so using kernel tricks; (iii) the presence of outliers should be minimal as they can significantly impact the margin; and (iv) feature scaling is important to ensure equal consideration of all features.

For Logistic Regression, the assumptions maintained include (i) the binary nature of the outcome variable, fitting the dichotomy of heart disease presence or absence; (ii) independence among observations, ensuring that each patient's data does not influence another's; (iii) a linear relationship between the independent variables and the log odds of the outcome, which is a critical assumption for logistic models; (iv) the absence of multicollinearity among independent variables, as high correlation can skew results; and (v) a sufficiently large sample size, which is essential for the reliability of Logistic Regression models.

Conversely, the Random Forest algorithm operates under a different set of assumptions: (i) individual decision trees within the forest should have minimal correlation to ensure the model's robustness; (ii) the performance improves with an increasing number of trees, though this is contingent on available computational resources; (iii) the model can handle non-linearity and interaction effects between variables, making it suitable for complex datasets; (iv) while capable of managing imbalanced data, extreme imbalances should be avoided as they can still bias the model; and (v) the presence of relevant predictors in the dataset is assumed, highlighting the importance of feature selection for optimal performance.

For XGBoost, key assumptions include (i) the model's capability to handle various types of data, including non-linear relationships; (ii) it performs well with large datasets and numerous features, but requires careful tuning to avoid overfitting; (iii) feature interaction is automatically captured, making complex relationships in the data manageable; (iv) the algorithm is robust to missing values and outliers, though preprocessing can enhance performance; and (v) it benefits from feature selection and regularization to improve model interpretability and prevent overfitting.

These assumptions are critical for the application and interpretation of the Random Forest model in predicting the presence of heart disease.

VI. Data Analysis

A. Original Data

The dataset consists of various columns related to heart health and associated risk factors. Here's a overview of the original columns:

1) HeartDisease: Binary variable indicating if the individual has heart disease ('Yes' or 'No'). Serves as the target variable for prediction.
2) BMI: Continuous variable representing the Body Mass Index of the individual. Reflects the body fat based on height and weight.
3) Smoking: Binary variable indicating whether the individual smokes ('Yes' or 'No').
4) AlcoholDrinking: Binary variable indicating if the individual consumes alcohol ('Yes' or 'No').
5) Stroke: Binary variable indicating if the individual has had a stroke ('Yes' or 'No').
6) PhysicalHealth: Continuous variable representing the number of days in the last month when physical health was not good.
7) MentalHealth: Continuous variable representing the number of days in the last month when mental health was not good.
8) DiffWalking: Binary variable indicating if the individual has difficulty walking ('Yes' or 'No').
9) Sex: Categorical variable indicating the gender of the individual ('Male' or 'Female').
10) AgeCategory: Ordinal variable representing the age category of the individual. Categorized into specific age ranges like '18-24', '25-29', etc.
11) Race: Categorical variable indicating the race of the individual.
12) Diabetic: Ordinal variable indicating if the individual is diabetic. Categories include 'No', 'Yes (during pregnancy)', 'No, borderline diabetes', and 'Yes'.
13) PhysicalActivity: Binary variable indicating if the individual engages in physical activity ('Yes' or 'No').
14) GenHealth: Ordinal variable representing the general health perception of the individual, ranging from 'Poor' to 'Excellent'.
15) SleepTime: Continuous variable indicating the average number of hours the individual sleeps.

16) Asthma: Binary variable indicating if the individual has asthma ('Yes' or 'No').
17) KidneyDisease: Binary variable indicating if the individual has kidney disease ('Yes' or 'No').
18) SkinCancer: Binary variable indicating if the individual has had skin cancer ('Yes' or 'No').

B. Data Preprocessing

The variables are classified into categorical, ordinal, and numerical based on their data types.

The categorical Variables are divided into binary nominal variables and nominal variables with more than two categories. The nominal variables with more than two categories are processed with 'pd.get_dummies' in Python. This function creates a new column for each category, using binary encoding (0 or 1) to indicate the presence of a category. The 'drop_first' parameter is set to True to avoid multicollinearity, which results in creating one fewer dummy variable than the number of possible factor levels. The binary nominal variables are directly mapped to 0 and 1, with a clear identification and assignment process for categories (e.g., 'Yes'/'No', 'Male'/'Female').

The ordinal variables 'AgeCategory', 'GenHealth', and 'Diabetic' are encoded using predefined mappings. These mappings are based on the logical order of the categories:

1) AgeCategory: A mapping is created where each age range is assigned a unique numeric value in ascending order.
2) GenHealth: General health perception is encoded from 'Poor' to 'Excellent', with increasing numeric values.
3) Diabetic: The diabetic status is encoded with distinct numeric values, considering the seriousness and type of diabetes.

The numerical variables are used directly without the need for encoding.

Then, we calculate quadratic (squared) terms for a subset of numerical variables: 'BMI', 'PhysicalHealth', 'MentalHealth', and 'SleepTime'. This is done by squaring the values of these variables and adding them as new columns to the dataset. The move is meant to help capture non-linear relationships within these variables themselves, providing a more nuanced understanding of their effect on the outcome (heart disease).

Before creating the interaction terms, the variables in 'column_to_exclude' (target variable 'HeartDisease' and ordinal variables 'AgeCategory', 'Diabetic', 'GenHealth') are excluded from generating interaction terms. This is to avoid convoluting the model with unnecessary interactions and to keep the focus on relevant variable interactions. We iterate through pairs of variables, generating interaction terms only for those not listed in 'column_to_exclude' and ensuring no interaction is calculated between the same race categories. For each valid pair, an interaction term is created by multiplying the values of the two variables, named in the format 'variable1_x_variable2'. These interaction terms are then concatenated to the original data matrix, expanding it with new columns that represent these interactions. We end up with a data matrix comprising 169 columns in this manner.

C. Exploratory Data Analysis

In the Exploratory Data Analysis (EDA), the distributions of the covariates and their correlations were examined and analyzed in with respect to the response variable. This thorough examination aimed to shed light on their potential influence on the response variable (Heart Disease), providing valuable insights into their impact on the outcomes. Presented here are the most salient findings from this analysis.

- Age Category exhibits a noticeable trend, aligning with a significant positive correlation with the presence of Heart Disease. As illustrated in Figure 1, there is a discernible increase in proportion as the age category advances.
- Diabetes distribution (Figure 2), which is also a categorical variable shows a clear higher proportion of Heart Disease when diabetes diagnosis occurs
- The perception of general health in individuals exhibits a noticeable trend of increasing the proportion of heart disease, particularly when individuals rate their health as poor. This aligns with expectations and suggests that it is likely to be a significant covariate associated with heart disease in the prediction model (3).

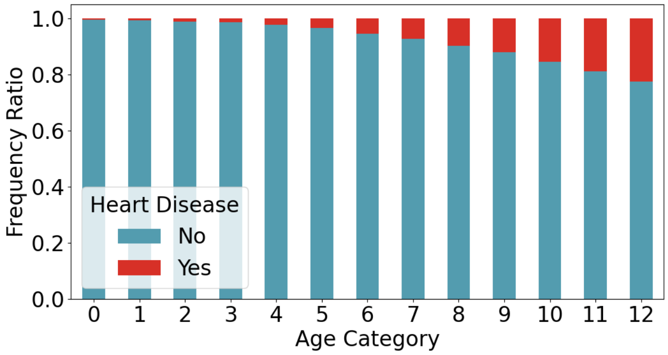

Figure 1. Frequency Ratio of Heart Disease in Different Age Categories

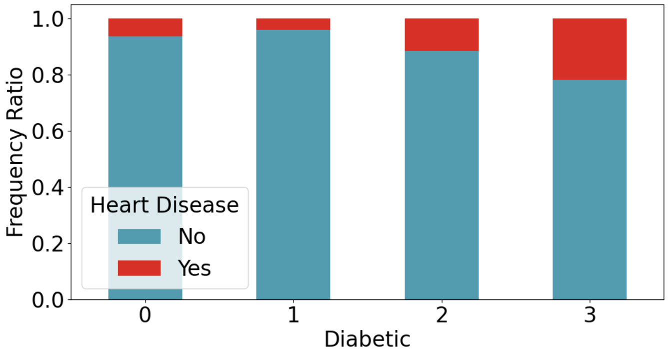

Figure 2. Frequency Ratio of Diabetes in Different Age Categories

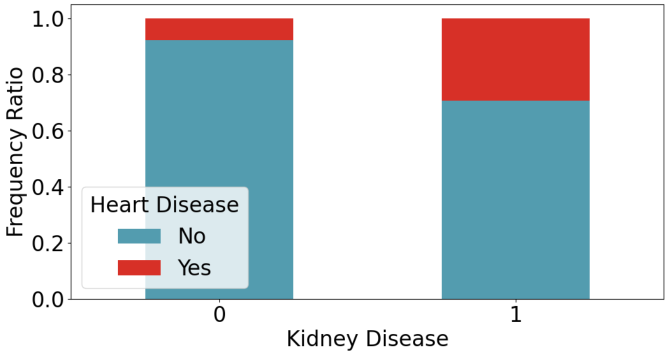

Figure 4. Frequency Ratio of Kidney Disease in Different Age Categories

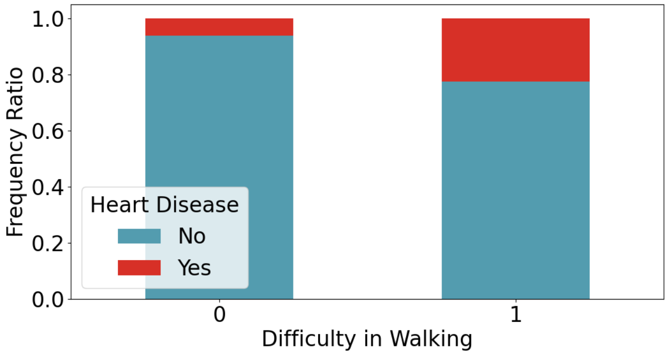

Figure 5. Frequency Ratio of Difficulty Walking in Different Age Categories

- Kidney Disease: Comparable to diabetes, patients diagnosed with kidney disease exhibit an doubled proportion of heart disease cases. This observation indicates that kidney disease may play a pivotal role as a covariate influencing the likelihood of heart disease. The absence of specific categories for kidney disease underscores its potential as a continuous variable impacting heart disease outcomes. The distribution of heart disease cases in relation to kidney disease diagnosis is visualized in Figure 4. This finding emphasizes the importance of considering kidney disease as a significant factor in our predictive modeling, warranting further exploration and analysis.

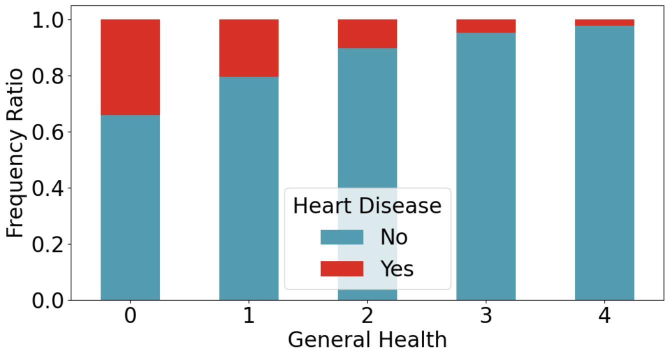

Figure 3. Frequency Ratio of Diabetes in Different Age Categories

- Difficulty Walking: The covariate related to patients' difficulty walking exhibits a higher proportion of heart disease cases when individuals face challenges in walking. This observation suggests that the ability to walk comfortably may be linked to heart health, making it a potential significant predictor in our analysis. The distribution of heart disease cases across different levels of difficulty walking is visually represented in Figure 5**??**. This insight into the relationship between difficulty walking and heart disease will be further explored and considered in our predictive modeling.
- Stroke: The existence of a stroke in a patient

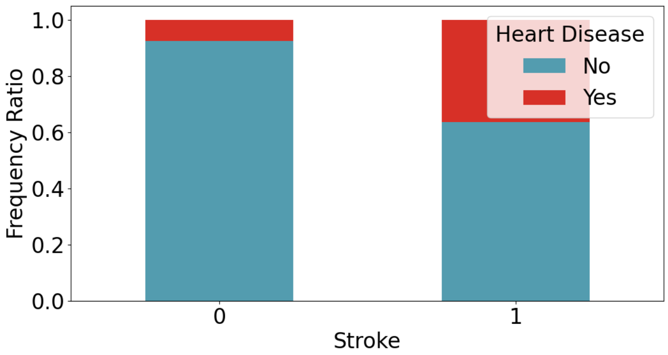

Figure 6. Frequency Ratio of Stroke in Different Age Categories

is a crucial factor to consider when predicting heart disease outcomes. As anticipated, individuals with a history of stroke exhibit a significantly higher proportion of heart disease. Figure 6 visually demonstrates this relationship, revealing an almost fourfold increase in the proportion of heart disease among those with a reported history of stroke. This underscores the importance of including the stroke variable in the prediction model.

- Body Mass Index (BMI): BMI is widely acknowledged for its association with heart health. While the distribution of BMI, as depicted in Figure 7, appears right-skewed, providing limited insights into its direct relationship with heart disease, examining interactions with categorical variables reveals a more nuanced understanding. Figure 8 and 9 showcase the interactions with categorical values such as stroke and difficulty walking. Surprisingly, these interactions exhibit a significantly higher proportion of subjects with heart disease compared to those without, contrasting with their respective single distributions. This suggests that the interplay between these covariates and BMI could be crucial in predicting heart disease outcomes.

- Interaction Term: Male Smoker. Exploring interaction terms, a notable trend emerges when examining the interaction between male gender and smoking status. Although the observed increase in the proportion of heart diseases is not as substantial as in the case of diabetes or kidney disease, it remains

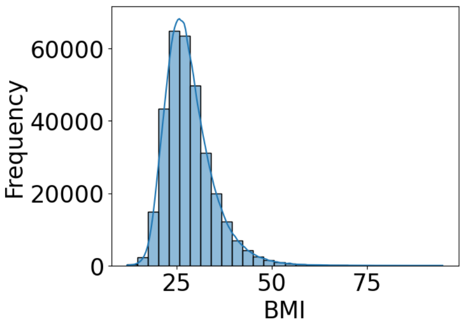

Figure 7. BMI Distribution

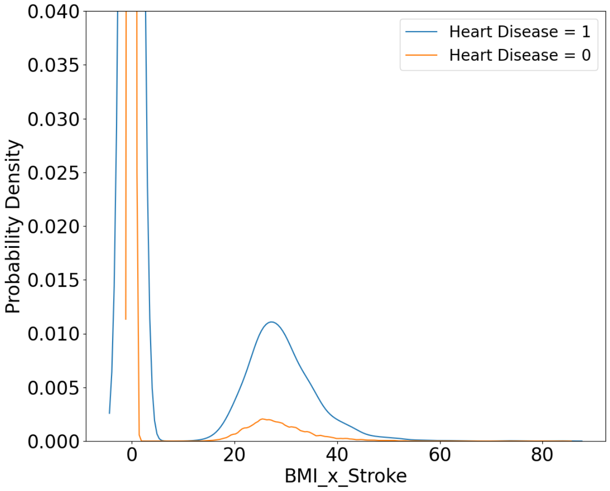

Figure 8. Truncated Probability Density of BMI-Stroke Interaction by Heart Disease Status (Cutoff at 0.04)

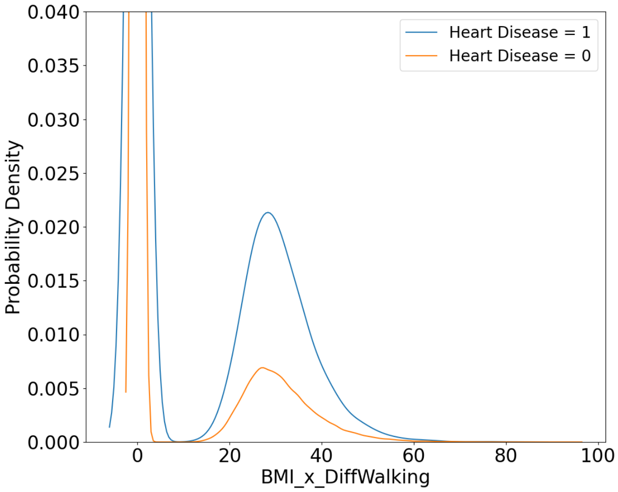

Figure 9. Truncated Probability Density of BMI-Difficulty in Walking Interaction by Heart Disease Status (Cutoff at 0.04)

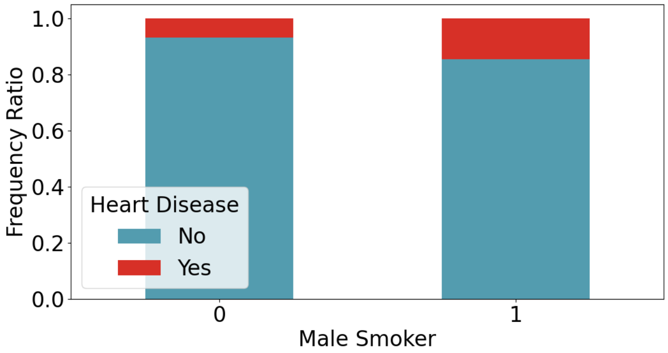

Figure 10. Frequency Ratio of the Gender and Smoking Habit Interaction in Different Age Categories. 1 in the X-axis represents male smokers, 0 Otherwise.

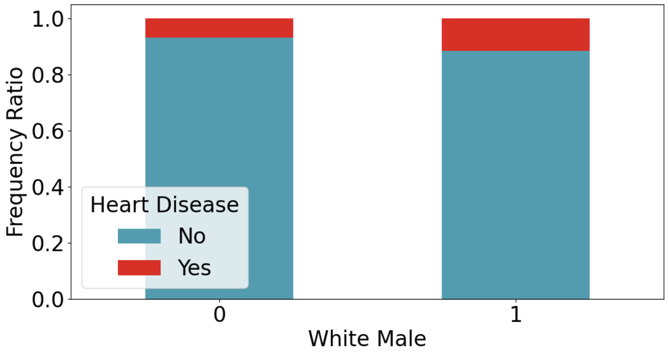

Figure 11. Frequency Ratio of White Individual's Gender in Different Age Categories. 1 in the X-axis Represents a Male.

a significant factor. The interaction effect between being a male smoker is depicted in Figure 10, illustrating its potential impact on heart disease outcomes. While the magnitude of this effect may be comparatively smaller, its significance suggests a nuanced influence that merits consideration in our predictive modeling.

- Interaction Term: Male White. Another noteworthy interaction term is observed when considering the combination of male gender and White ethnicity. Examining this interaction reveals a discernible impact on the proportion of heart diseases. While the effect may not be as pronounced as with certain medical conditions like diabetes or kidney disease, it remains a salient factor. The interaction effect between being a male of White ethnicity is illustrated in Figure 11, shedding light on its potential relevance in predicting heart disease outcomes. However, it's crucial to note that with other ethnicities, such as Hispanic, the proportion of heart diseases appears to be comparatively lower, as evident from the exploration of diverse ethnic interactions.

D. Main Results

Post-training, our models were evaluated on the test set.

1) Critical Analysis of Classifier Efficacy in Heart Disease Diagnostics: Given the critical nature of accurately classifying cases in a medical context, particularly for heart disease detection, we have computed confusion matrices and key medical test criteria for our models. These results are detailed in Table I and Table II, providing insight into the implications of misclassification by each model.

Model	TN	FP	FN	TP
SVM	87649	0	8290	0
Logistic Regression	86936	713	7515	775
Random Forest	87389	260	7927	363
XGBoost	87081	568	7610	680

Table I: Comparative Overview of Confusion Matrices for Various Classifier Models: Detailing True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP) Counts

Metric	SVM	Logistic Reg.	RF	XGBoost
Accuracy	0.9136	0.9142	0.9147	0.9148
Precision	0.0000	0.5208	0.5827	0.5449
Recall	0.0000	0.0935	0.0438	0.0820
Specificity	1.0000	0.9919	0.9970	0.9935
F1 Score	0.0000	0.1585	0.0815	0.1426

Table II: Comparative Analysis of Classifier Performance Metrics: 'Logistic Reg.' denotes Logistic Regression and 'RF' refers to Random Forest.

According to Table II, the Support Vector Machine (SVM) shows high accuracy (0.913591) and perfect specificity (1.0), yet its performance in heart disease detection is significantly hindered due to its very low precision, recall, and F1 score (all 0.0). The model's inability to correctly identify any true positive cases of heart disease, as indicated by the zero recall, renders it unsuitable for this application, despite its otherwise impressive accuracy and specificity metrics.

In contrast, Logistic Regression displays slightly better capabilities in identifying some true positives, with moderate precision (0.520833) and low recall (0.093486). Its accuracy is high (0.914237), and it has a very high specificity (0.991865), which suggests it is effective at identifying negative cases. However, its relatively low recall and F1 score (0.158519) highlight a limitation in effectively detecting positive cases of heart disease, which is a critical aspect of the diagnostic process.

The Random Forest classifier also scores high in accuracy (0.914665) and has the highest precision among the models (0.582665), but it suffers from a very low recall (0.043788) and F1 score (0.081454). Its almost perfect specificity (0.997034) indicates strong negative case identification. However, the extremely low recall means that it misses the majority of positive heart disease cases, which is a crucial drawback for medical diagnostics.

Lastly, the XGBoost model exhibits a high level of accuracy (0.914758) and a moderate level of precision (0.544872). Its recall (0.082027) and F1 score (0.142588), while low, are somewhat comparable to Logistic Regression. The model's very high specificity (0.993520) and high AUC-ROC (0.840125) indicate a reliable performance in classifying negative cases. However, like the other models, its lower recall suggests a limitation in adequately identifying true positive cases of heart disease.

2) Receiver Operating Characteristic Curves: We also compared their performance using Receiver Operating Characteristic (ROC) Curves and calculated the Area Under Curve (AUC) for each model, as depicted in Figure 12. The ROC curves and AUC scores reveal that the Support Vector Machine (SVM) model performed the least effectively, characterized by a ROC curve resembling a seemly diagonal line and an AUC of 0.57. In contrast, the Logistic Regression, Random Forest, and XGBoost models demonstrated similar efficacy, each achieving a high AUC value approximately around 0.84.

3) Feature Importance: Since Logistic Regression, Random Forest, and XGBoost have similar performance, we could find important features by analyzing the feature importance of Random Forest and XGBoost model.

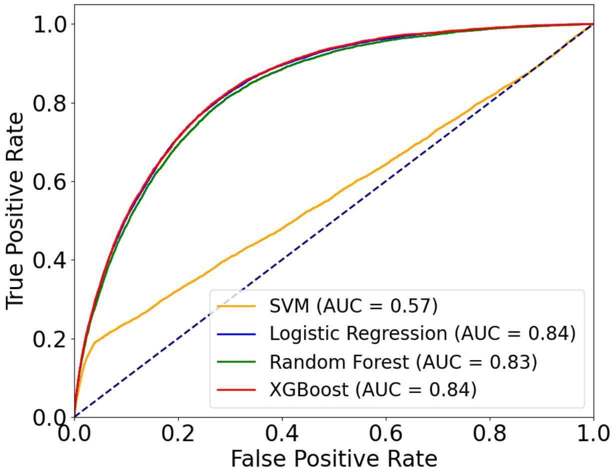

Figure 12. Receiver Operating Characteristic Curves of SVM, Logistic Regression, Random Forest, and XGBoost Algorithms

First, let us define the feature importance of our Random Forest and XGBoost model. Given a Random Forest model with N decision trees, the feature importance for a feature F is calculated as follows:

- For each tree T_j in the Random Forest, and for each feature F:
  - Let $\Delta I_{G,F,T_j,k}$ be the decrease in Gini impurity at node k in tree T_j due to a split made on feature F. Gini impurity is defined as

    $$I_G(p) = 1 - \sum_{i=1}^{J} p_i^2$$

    Here, p_i is the proportion of items labeled with class i in the set, and J is the number of classes.
  - This decrease in impurity is calculated at each split and is a measure of how much the feature F contributes to partitioning the data into purer subsets.
- For each tree T_j, sum the decreases in Gini impurity for all splits made on feature F:

  $$I_{F,T_j} = \sum_{k \in \text{Nodes split on } F} \Delta I_{G,F,T_j,k}$$

- The importance of feature F in the Random Forest, denoted as Importance(F), is then

Feature (RF)	Feature (XGB)
AgeCategory	AgeCategory
GenHealth	GenHealth
Diabetic	Diabetic
BMI_x_DiffWalking	-
BMI_x_Stroke	-
Stroke_x_SleepTime	-
DiffWalking	DiffWalking
BMI_x_SleepTime	-
DiffWalking_x_SleepTime	DiffWalking_x_SleepTime
BMI_x_Sex	-
-	Smoking_x_Sex
-	Stroke_x_DiffWalking
-	Sex_x_Race_White
-	DiffWalking_x_KidneyDisease
-	Stroke

Table III: Comparison of Top-Ranked Features in Random Forest (RF) and XGBoost (XGB) Models. Features are Presented in Descending Order of Their Respective Importance.

the average of I_{F,T_j} over all N trees:

$$\text{Importance}(F) = \frac{1}{N}\sum_{j=1}^{N} I_{F,T_j}$$

The feature importance of our XGBoost model is calculated in terms of gain. However, since we utilized the l_2 regularization terms in training XGBoost, and that Random Forest's trees are built independently, while XGBoost's trees are built sequentially with an error-correcting approach, the feature importance calculation of XGBoost is different from Random Forest.

Table III displays the seven most significant features as determined by their respective importance in the Random Forest and XGBoost models. Notably, the features that consistently rank highly in both models include AgeCategory, GenHealth, Diabetic, DiffWalking, and DiffWalking_x_SleepTime. This consistency underscores the critical influence of an individual's age, general health status, diabetic condition, difficulty in walking, and average sleep duration (especially in the context of walking difficulties) on their heart health.

VII. Conclusions

In the context of heart disease detection, Recall (Sensitivity) emerges as the paramount metric due to its role in accurately identifying true cases of the disease. This prioritization stems from the greater risk associated with false negatives (missed heart disease cases) compared to false positives (incorrectly diagnosed cases).

Upon reviewing the data, Logistic Regression stands out as the most suitable model, primarily because of its relatively higher recall, which is vital for maximizing the detection of true positive cases. However, it is important to acknowledge that all models exhibit room for improvement in recall. This observation underscores the need for potential enhancements through model tuning, data sampling adjustments, or the exploration of alternative modeling approaches to bolster recall performance.

From a statistical perspective, Logistic Regression is also favored due to its impressive balance of high performance and relatively lower complexity, as evidenced by its ROC curves and AUC scores. This dual consideration of effectiveness and simplicity further solidifies its selection as the preferred model for this application.

To identify the key factors influencing heart disease, we focus on the shared top-ranked features in terms of feature importance from both the Random Forest and XGBoost models, given their comparable performance to the Logistic model. The findings highlight the importance of an individual's age, general health status, diabetic condition, difficulty in walking, and average sleep duration, particularly in relation to walking difficulties, as significant determinants of heart health.

VIII. Discussions

We expected the quality of our results to be high, with an anticipated accuracy surpassing 90%, like the results in [12], [13]. The substantial size of our data set, which is closer to the size of Du et al.s' [13], should contribute to stable and solid results, ensuring that the project exhibits strong generalization capabilities.

The evaluation makes clear that maximizing recall is imperative for credibly deploying a heart disease classifier in clinical settings due to the severe risks of false negatives. On this metric, Logistic Regression emerges as the leading model. Still, there is substantial room for improvement considering its recall remains under 10%.

A key priority for further research should be exploring techniques to improve recall without sacrificing precision and accuracy to unacceptable degrees. More aggressive data re-sampling focused specifically on minority positive cases may expose models to a wider breadth of heart disease presentations. Additionally, learning methods that apply higher feature weights on the critical minority class during training could boost sensitivity.

It is also worth investigating how tuning model hyperparameters may influence the recall-precision balance. Allowing Logistic Regression's decision threshold to skew towards more liberal positive labeling could lift recall at the expense of precision. Quantifying this tradeoff curve could reveal an optimal configuration that maximizes life-saving detections while preserving statutory statistical rigor.

Finally, the models' feature importance analyses provide clues into their black-box inner workings. The identified most influential features like chest pain type, cholesterol, and exercise-induced angina align with clinical knowledge on cardiac risk factors. This provides some reassurance of logical modeling. Though notably absent are social determinants of health like income, education and race which are increasingly considered in modern clinical guidelines. Incorporating these non-biological but predictive features could enhance case detection from an ethical, equitable lens.

In summary, while Logistic Regression presents the best available option, truly reliable heart disease prediction remains an open challenge. The models leave much room for improvement before real-world application. Advancing an automated, life-saving cardiovascular diagnostic tool should compel ongoing research, even if human-level performance proves elusive across modeling techniques.

IX. Appendix

All relevant source codes used in this analysis are available on our GitHub repository. For further details and access to the codebase, please visit: https://github.com/jacklj9811/Heart-Disease-Analysis-Kaggle-.

References

[1] Gregory A Roth, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al., "Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017," *The Lancet*, vol. 392, no. 10159, pp. 1736–1788, 2018.

[2] William P Castelli, "Cardiovascular disease and multifactorial risk: challenge of the 1980s," *American heart journal*, vol. 106, no. 5, pp. 1191–1200, 1983.

[3] Russell Ross, "The pathogenesis of atherosclerosis: a perspective for the 1990s," *Nature*, vol. 362, no. 6423, pp. 801–809, 1993.

[4] Scott M Grundy, Richard Pasternak, Philip Greenland, Sidney Smith Jr, and Valentin Fuster, "Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations: a statement for healthcare professionals from the american heart association and the american college of cardiology," *Circulation*, vol. 100, no. 13, pp. 1481–1492, 1999.

[5] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[6] Leo Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[7] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[8] Kamil Pytlak, "Key indicators of heart disease (2022 update)," https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/, 2023, Accessed: 2023-11-02.

[9] Catherine Kim and Gloria L Beckles, "Cardiovascular disease risk reduction in the behavioral risk factor surveillance system," *American journal of preventive medicine*, vol. 27, no. 1, pp. 1–7, 2004.

[10] Khader Shameer, Kipp W Johnson, Benjamin S Glicksberg, Joel T Dudley, and Partho P Sengupta, "Machine learning in cardiovascular medicine: are we there yet?," *Heart*, vol. 104, no. 14, pp. 1156–1164, 2018.

[11] Hayato Tada, Noboru Fujino, Akihiro Nomura, Chiaki Nakanishi, Kenshi Hayashi, Masayuki Takamura, and Masa-aki Kawashiri, "Personalized medicine for cardiovascular diseases," *Journal of Human Genetics*, vol. 66, no. 1, pp. 67–74, 2021.

[12] Kellen Sumwiza, Celestin Twizere, Gerard Rushingabigwi, Pierre Bakunzibake, and Peace Bamurigire, "Enhanced cardiovascular disease prediction model using random forest algorithm," *Informatics in Medicine Unlocked*, vol. 41, pp. 101316, 2023.

[13] Zhenzhen Du, Yujie Yang, Jing Zheng, Qi Li, Denan Lin, Ye Li, Jianping Fan, Wen Cheng, Xie-Hui Chen, Yunpeng Cai, et al., "Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: model development and performance evaluation," *JMIR medical informatics*, vol. 8, no. 7, pp. e17257, 2020.

[14] Istiak Mahmud, Md Mohsin Kabir, M F Mridha, Sultan Alfarhood, Mejdl Safran, and Dunren Che, "Cardiac failure forecasting based on clinical data using a lightweight machine learning metamodel," *Diagnostics (Basel, Switzerland)*, Jul 2023.

[15] Steven S Coughlin, Bruce Trock, Michael H Criqui, Linda W Pickle, Deirdre Browner, and Mariella C Tefft, "The logistic modeling of sensitivity, specificity, and pre-

dictive value of a diagnostic test," *Journal of clinical epidemiology*, vol. 45, no. 1, pp. 1–7, 1992.

[16] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant, *Applied logistic regression*, vol. 398, John Wiley & Sons, 2013.

[17] Steven J Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, 2017.

[18] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2, Springer, 2009.