Applying Machine Learning to Predict Heart Risk

Jack Lacy, Felix Zheng, KwanNok Lau, Marcus Tuen Muk, Luis Delotavo

**Abstract—Cardiovascular disease (CVD) remains the leading cause of mortality in the United States, accounting for a significant proportion of healthcare challenges and costs. Its often asymptomatic progression makes early detection and intervention crucial yet difficult to achieve. This study develops a comprehensive machine learning-based framework for predicting heart disease risk using data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS). The proposed methodology encompasses an end-to-end pipeline that includes data preprocessing, exploratory data analysis, feature selection, model development, and evaluation.**

**Seventeen distinct health indicators—spanning physiological measurements (e.g., blood pressure, BMI), behavioral factors (e.g., physical activity, smoking status), and demographic characteristics (e.g., age, gender)—are utilized to capture a wide range of risk determinants. A comparative analysis of predictive models, including logistic regression, random forest, gradient boosting, and ridge regression, is performed to identify the most effective approach. Model evaluation employs multiple metrics, such as the area under the receiver operating characteristic curve (ROC-AUC), precision, recall, F1-score, and calibration plots, to ensure both accuracy and reliability. Robustness and generalizability of the results are validated through cross-validation and testing on unseen data.**

**The results highlight the critical interplay of specific features, such as age, smoking, and physical activity, in determining CVD risk, while demonstrating the superior performance of ensemble methods in predictive accuracy. This study provides a practical and scalable tool for healthcare practitioners to identify high-risk individuals, enabling targeted prevention strategies and timely clinical interventions. By leveraging data-driven insights, this work contributes to improved management of cardiovascular disease, ultimately aiming to reduce morbidity and mortality rates while enhancing overall public health outcomes.**

***Index Terms*— Artificial intelligence, Behavioral Risk Factor, Surveillance System (BRFSS), Cardiovascular disease, Cross-validation, Feature selection, Gradient boosting, heart disease prediction, Machine learning, Multiple linear regression, Random Forest, Ridge regression, Risk factors, Simple linear regression**

# I. INTRODUCTION

H

EART disease remains the leading cause of death in the United States, claiming one life every 33 seconds and accounting for one in five deaths as of 2022 [1]. The silent nature of cardiovascular disease, where symptoms may remain undetected until a critical cardiac event occurs, presents a significant challenge in early diagnosis and prevention. This study develops a machine learning model to predict heart disease risk by analyzing an extensive set of health metrics derived from the CDC's Behavioral Risk Factor Surveillance System (BRFSS).

The proposed model incorporates seventeen distinct variables including both quantifiable measurements and qualitative factors. Quantifiable metrics include BMI and sleep time, while qualitative data encompasses general health status, lifestyle choices (including smoking and alcohol consumption), and demographic information. Multiple machine learning algorithms are implemented in this study: logistic regression serves as a baseline model, while more advanced techniques like random forest, gradient boosting, and ridge regression are employed to capture complex relationships between health indicators.

Through processing these multifaceted health indicators, our approach aims to identify subtle patterns and risk factors that may precede cardiovascular events. Model performance is rigorously evaluated using metrics such as ROC-AUC, precision, recall, and F1-score, with cross-validation ensuring reliable generalization. The significance of this research lies in its potential to provide healthcare providers and individuals with an additional resource for early risk assessment, potentially enabling more timely interventions and lifestyle modifications to reduce heart disease mortality rates.

# II. Dataset

The dataset [2] selected for the study, *Indicators of Heart Disease*, provides insight into the correlation between patients' health conditions and heart disease. The dataset originated from the CDC through their Behavioral Risk Factor Surveillance System (BRFSS). It provides information on patients’ habits, preexisting health conditions and other factors related to their health. For example, the dataset has information on the smoking and alcohol drinking status of the patient. Additionally, it mentions how long the patient sleeps and their physical activity status. These features provide insight into the patients’ day-to-day conditions that potentially relate to their risk of heart attack. The dataset also provides insight on patients’ health through BMI, if they had a stroke or any other pre-existing conditions, like cancer. These features provide insight into possible signs of heart disease and other possible conditions that are affecting the patient.

In all, the dataset provides a large assortment of features. These features are BMI, smoking, alcohol drinking, stroke, physical health, mental health, difficulty walking, sex, age, race, diabetic, physical activity, general health, sleep time, asthma, kidney disease, and skin cancer. The target variable is heart disease, and the options are yes, and no.

The dataset provides both numerical and categorical data which would need to be processed before a model is created. The categorical data, such as smoking and strokes, will need to be encoded to become numerical data.

# III. Methodologies

Our methodology is designed to construct a robust heart disease risk prediction model through detailed data preparation, model selection, optimization, and evaluation. In the preprocessing phase, we focused on ensuring that the dataset was clean, consistent, and in a format suitable for the chosen models. We use techniques such as binary and ordinal encoding, one-hot encoding, scaling, normalization, and feature engineering to effectively handle different types of data and create new features that could enhance our model’s predictive power.

Variables such as heart disease, smoking, alcohol drinking, stroke, difficulty walking, physical activity, asthma, kidney disease, and skin cancer were transformed using binary encoding since they were presented as binary conditions (e.g., “Yes” or “No”).

Some features had a clear, natural ordering that showed the level of progression or hierarchy. For instance, body mass index (BMI) categories (underweight, normal, overweight, obese) were assigned increasing integer values to represent escalating weight categories. Similarly, general health quality (GenHealth) responses were mapped to a scale from 1 (poor) to 5 (excellent), and age categories were encoded as integers from 1 to 13. Additionally, the diabetes feature was represented on an ordinal scale (0 = none, 1 = borderline, 2 = yes), and sleep time categories were assigned values to distinguish undersleeping, normal, and oversleeping patterns. By encoding these variables ordinally, we outlined the relation between the values.

On the other hand, for categorical variables without a natural order, such as race, one-hot encoding was used. Each category within the variable was turned into separate features. This transformation allowed the model to treat each race category as a distinct and independent feature.

To ensure that variables on different scales did not disproportionately influence models, we applied scaling and normalization techniques to continuous features. This approach was especially beneficial for models sensitive to distance measures and gradients, such as neural networks. We used log transformations for features like mental and physical health since the data was skewed left, but used standard scaling for distributions that looked similar to the Gaussian distribution.

Correlation analysis was used to examine the relationship between features and the correlation map revealed generally low correlations between variables. This indicated that multicollinearity was not a major issue and creating interaction features or higher-order features was feasible for our dataset.

To improve our models’ capabilities, we introduced several new variables such as gestational diabetes, BMI squared, summation of chronic conditions, smokes and drinks interaction, and an active good health indicator. The diabetic feature was divided into two separate variables — one that captured the progression and one specific to an individual experiencing diabetes during pregnancy. The BMI squared feature was created to account for potential non-linear effects of body mass on the target variable. The summation of chronic conditions was a new feature that represented the sum of stroke, asthma, kidney disease, skin disease, and diabetes. The smokes and drinks interaction represents individuals who both smoke and drink which could capture an amplified effect on heart disease risk. Lastly, the active good health indicator combined physical activity status with general health quality. By merging these two attributes, we created a measure that may more comprehensively reflect overall health.

Our modeling process includes training two distinct models — a Random Forest Classifier and a Neural Network using the preprocessed dataset.

The Random Forest Classifier is a supervised machine learning algorithm widely used for classification and regression tasks. We selected this model because it is flexible and robust. It is non-parametric and thus can adapt well to various data types and distributions. Additionally, the ensemble of decision trees and how each tree is trained on a randomly selected subset of training data generally reduces the risk of overfitting while maintaining good accuracy and stability.

Neural Networks is a type of machine learning algorithm inspired by the structure and function of the human brain. They consist of nodes/neurons arranged into layers, which collectively learn to recognize patterns and relationships in the input data. We chose the neural network as our second model since it complements the random forest classifier. It leverages parallel processing and can excel in identifying non-linear patterns so our interaction terms and higher-order terms may increase the models performance.

Both the Random Forest and Neural Network models have their own distinct advantages and limitations. The Random Forest Classifier offers robustness and strong baseline performance with relatively less hyper-parameter tuning. The Neural Network provides the capability of modeling highly complex relationships and can potentially be scaled with parallel computation. However, due to the complexity of both models, they are much less interpretable than simpler models like a linear regression model. By using both models, we aimed to achieve a comprehensive understanding of the key factors associated with the likelihood of heart attacks while expanding our own knowledge of complex architectures.

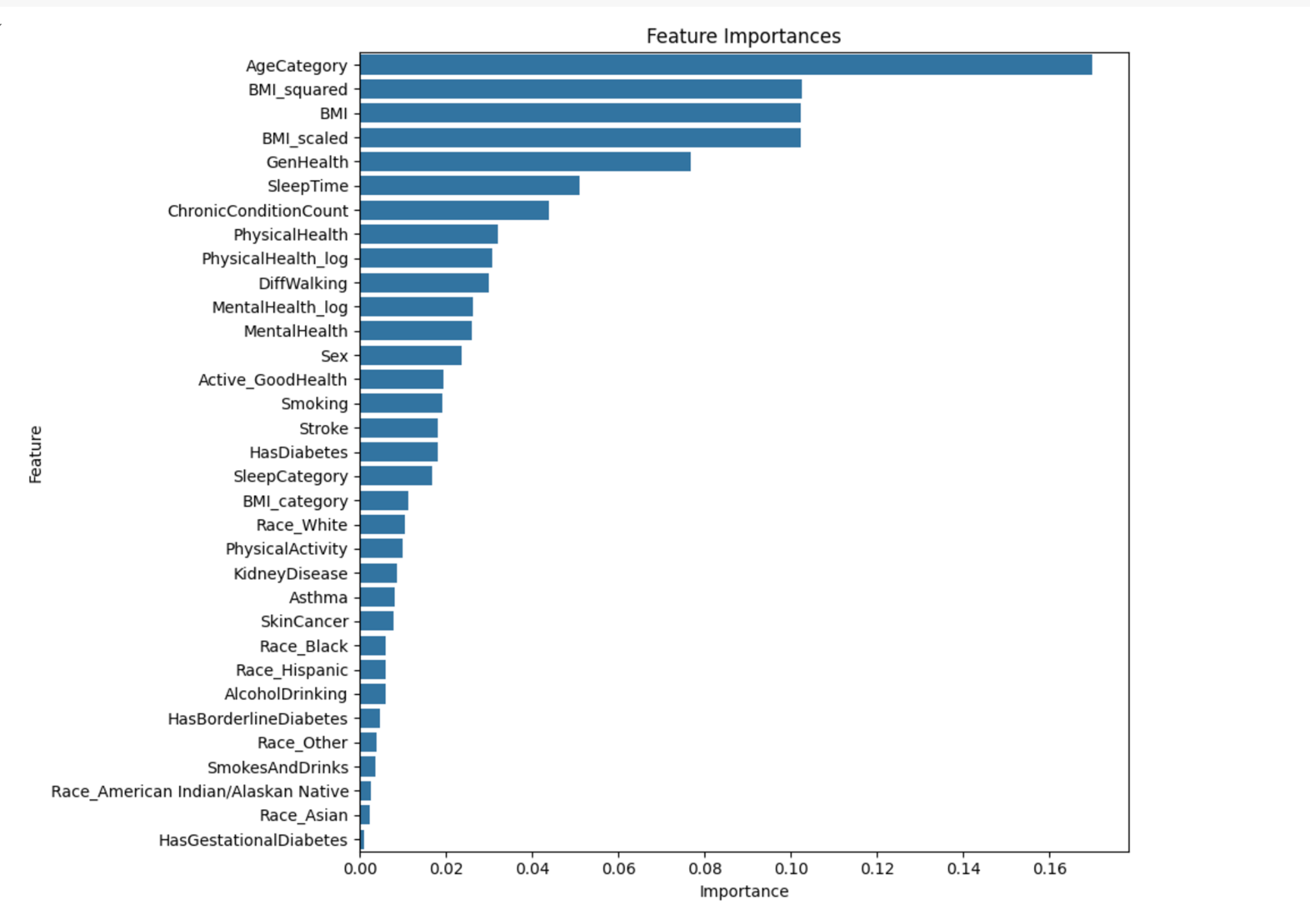
To assess each model's effectiveness, we use metrics such as ROC-AUC, precision, recall, and F1-score, focusing on models that achieve a strong balance between sensitivity and specificity, which is crucial for early disease detection. Cross-validation will ensure that the models generalize well to new data, avoiding overfitting. Additionally, we may employ ensemble techniques, combining the predictions from multiple models to increase stability and reduce the impact of individual model biases.

This comprehensive approach ensures that our final model is not only accurate but also interpretable and reliable, providing actionable insights that can aid healthcare providers in identifying individuals at high risk for heart disease and supporting early intervention measures.

# IV. Results

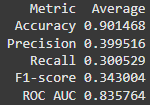
The performance of the neural network and random forest models for heart disease prediction was evaluated using several key metrics including Accuracy, Precision, Recall, F1-score, and ROC AUC scores...

The target variable in our dataset is whether or not a person has heart disease. Our dataset will help us predict this based on lifestyle factors and other illnesses. Prior to jumping straight into the results of the models we expect to see the most impact from certain features as described in Figure 1 with AgeCategory having by far the most impact and diabetes as the least impactful feature.



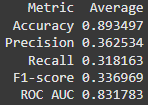
**Fig. 1.** Feature Importance graph

The first result to show is the performance metrics of the Random Forest model (Fig. 2). The random forest model achieved an accuracy of 90.15%. Its precision in identifying positive cases was 39.95%, though the recall of 30.05% was slightly lower. The random forest had a strong ROC AUC score of 0.836, suggesting better ability to distinguish heart disease from non-heart disease patients.



**Fig. 2.** Random Forest Model Results

Secondly, we have the result of the neural network model to show (Fig. 3). The neural network model achieved an overall accuracy of 89.35%. However, the model struggled with precision, correctly identifying positive cases only 36.25% of the time. The recall was also low at 31.82%. The neural network's ROC AUC score of 0.832 indicates reasonably good, but not exceptional, discrimination.



**Fig. 3.** Neural Network Model Results

The low precision and recall for both models suggest the datasets were imbalanced, with more negative (no heart disease) cases than positive (heart disease) cases. This can cause models to struggle identifying the minority class. Techniques like oversampling the minority class or undersampling the majority may improve performance. Evaluating F1-score, which balances precision and recall, could also provide better insight.

The different transformations applied to the BMI feature did not have a significant impact on the performance of the random forest model. The model treated the BMI variable in a similar manner, regardless of whether it was in its raw form or transformed. This is not surprising, as random forests are generally robust to feature scaling and transformations.

However, the impact on the neural network model is less clear. Typically, feature scaling and normalization can have a positive effect on the performance of neural networks, as it helps the model converge more efficiently. Unfortunately, the information provided does not indicate how the neural network model was affected by the various BMI transformations. Further investigation would be needed to understand the neural network's sensitivity to these preprocessing steps. In a similar manner, the exact same could be said for the sleep time vs sleep category features.

# V. Conclusion/Discussion

We have discovered many different things while taking on this project. We have determined that some features in our dataset are much more important than others and the most important by far is the age category. Unfortunately, this clearly is not something a person can change with manual intervention and is just a risk people will have to live with. The good news is that there are other important features that we can have an influence on like BMI, general health, and sleep time. By improving these aspects of one’s life, a person can expect a reduced risk of heart disease.

Through our machine learning efforts, we developed two models. One is a random forest model, and the other is a neural network, with the random forest model being stronger due to its accuracy, precision, and F1 score. Although the precision and recall on both weren’t exceptionally high, our model can still be a good metric for determining potential risk, and its pitfalls just demonstrate how difficult it is to predict heart disease. This is why it kills so many people in North America and around the world.

In conclusion, although heart disease is difficult to classify, the models we developed are quite accurate and can be a step in the right direction for determining heart disease risk. If a person wants to remain healthy, there are clear steps they can take like getting more sleep and trying their best to reduce their BMI and improve their general health.

References

[1] “Heart disease facts,” Centers for Disease Control and Prevention, https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html#:~:text=Heart%20disease%20in%20the%20United%20States&text=Heart%20disease%20is%20the%20leading,people%20died%20from%20heart%20disease. (accessed Dec. 6, 2024).

[2] K. Pytlak, “Indicators of heart disease (2022 update),” Kaggle, https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease (accessed Nov. 17, 2024).