Dean&Francis
ISSN 2959-6157

Temporal Evolution of Predictive Factors for Heart Disease: A Random Forest Analysis

Gema Zhu

Department of Statistics and Applied Probability, University of California, Santa Barbara, U.S.

gemazhu@ucsb.edu

Abstract:

Heart disease remains the leading cause of death globally, profoundly affecting patients' quality of life and placing a significant burden on healthcare systems and society. Identifying and understanding the key factors associated with heart disease is essential for its prevention, diagnosis, and treatment. This study explores how the significance of these factors has evolved over time by analyzing data from the Behavioral Risk Factor Surveillance System (BRFSS) from 2015 to 2021. This study focused on lifestyle and demographic variables for non-institutionalized adults aged 18 and older, selecting 36 relevant variables from an initial pool of over 300 each year through rigorous data cleaning and normalization. Utilizing a random forest algorithm, this paper evaluated feature importance across the years. The findings consistently highlight BMI, Income, Age, General Health, Education, and Smoking as pivotal predictors of myocardial infarction (MI) and coronary heart disease (CHD). Although High Cholesterol and Arthritis appeared in the top ten features only once during the four years, they maintained a notable presence within the top fifteen, indicating their significant but secondary role compared to the consistently prominent factors. This variability highlights that while some factors retain stable importance, others may vary in relevance due to changing health trends and dataset characteristics.

Keywords: Feature importance, random forest, disease prediction, machine learning.

1. Introduction

Heart disease includes a range of conditions affecting the heart and blood vessels, with the most common types including coronary artery disease, heart failure, arrhythmias, and heart valve disorders. The causes of heart disease are multifaceted, including genetic predisposition, high blood pressure, diabetes, smoking, and unhealthy lifestyle choices such as poor diet and lack of exercise. Symptoms can vary but often

ISSN 2959-6157

include palpitations, shortness of breath, and chest pain or discomfort. Heart disease is the leading cause of death worldwide. According to the World Health Organization, over 17 million people die from heart disease each year, accounting for 31% of all global deaths. In the United States, someone has a heart attack every 40 seconds. Heart disease not only significantly impacts patients' quality of life but also imposes a substantial burden on healthcare systems and society at large. Therefore, identifying and studying early indicators of heart disease is crucial for its prevention, diagnosis, and treatment.

Integrating Artificial Intelligence (AI) into heart disease prediction represents a significant advancement over traditional diagnostic methods, which often rely heavily on manual analysis by physicians. While expert knowledge is invaluable, these conventional methods can be costly, inefficient, and prone to errors. Recent progress in AI has introduced powerful algorithms such as random forests, Support Vector Machines (SVM), and neural networks, which have proven effective across various fields including chemistry, biomedical research, and finance. For example, random forests have been used to predict heart disease risk with notable accuracy [1], and SVM has been successfully applied to classify heart disease data [2]. Additionally, deep learning models like Convolutional Neural Networks (CNNs) have shown potential in automating Electrocardiogram (ECG) analysis for early heart disease detection [3]. By leveraging these advanced AI techniques, researchers aim to enhance diagnostic efficiency, reduce misdiagnosis rates, and optimize resource allocation in healthcare. This integration of AI into heart disease prediction not only promises to improve diagnostic precision but also highlights the transformative role of technology in advancing medical practices.

The application of AI in heart disease prediction not only enhances diagnostic efficiency but also significantly reduces the workload of doctors and the risk of misdiagnosis. In recent years, AI has been increasingly applied in the medical field, with numerous studies demonstrating its potential in disease prediction. For example, Google Health developed a deep learning model that outperformed radiologists in detecting breast cancer in mammograms. This model was able to reduce both false positives and false negatives, making it a powerful tool in breast cancer screening [4]. In Alzheimer's disease prediction, AI has been used to analyze brain imaging data, genetic information, and clinical data to predict the onset of Alzheimer's disease. AI models have shown promise in identifying early biomarkers and patterns associated with the disease, potentially allowing for earlier diagnosis and treatment [5]. However, while past research has largely focused on improving model performance, it is equally important to explore the scientific validity of these models and the significance of feature importance. By analyzing how feature importance changes over time or with patient characteristics, it is feasible to gain a deeper understanding of the mechanisms of heart disease and optimize medical intervention strategies. This type of research not only enriches theoretical discussions but also provides strong support for the practical application of AI in clinical settings.

To achieve the research objectives mentioned above, this study collected data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) spanning different years, from 1988 to 2022. 30 common features across five specific years were identified. Using these datasets from different years, this study trained and predicted heart disease risk using a Random Forest model. This paper then analyzed and compared the prediction accuracies across these different time periods to understand the model's performance and how it evolves with temporal changes in the data, while also evaluating the feature importance from the Random Forest model to identify which factors contributed most significantly to heart disease risk over time.

2. Method

2.1 Dataset Preparation

This study utilized data from the Centers for Disease Control and Prevention's (CDC) Behavioral Risk Factor Surveillance System (BRFSS) [6], a comprehensive telephone survey designed to collect information on a wide range of lifestyle factors and demographic variables from non-institutionalized adults aged 18 and older. The dataset includes a broad spectrum of health-related behaviors and conditions, including chronic health conditions and preventive service usage, with data available from 1984 to 2022. For the analysis, this paper focused on data from four specific years between 2015 and 2021. Each dataset is extensive and complex, with the 2015 dataset alone containing 330 columns and over 400,000 rows of information. Due to the heterogeneous nature of the data, which includes numerous missing values and variations in questions asked across years, meticulous cleaning and preprocessing were essential. For this study, the primary focus was on identifying factors associated with Coronary Heart Disease (CHD) or Myocardial Infarction (MI) as reported by respondents. The target variable, denoted as _ MICHD, indicates whether respondents have ever reported having CHD or MI. Specifically, a value of 1 represents individuals who have reported MI or CHD, while a value of 2 indicates those who have not.

Preprocessing began by removing any blank entries were

removed to ensure the dataset only includes valid responses, with the target variable thus constrained to values 1 and 2. Other variables were assessed as potential factors related to heart disease. For instance, the variable _RFHYPE5 represents high blood pressure as reported by respondents, with values 1 indicating 'yes' and 2 indicating 'no.' Values of 9, which denote unknown or not asked, were excluded from the analysis. Similarly, the diabetes variable, DIABETE3, records whether respondents have ever been told they have diabetes. The values for this variable are 1 for 'yes,' 2 for 'no,' 3 for 'pre-diabetes or borderline diabetes,' and 4 for 'other.' Responses coded as 7, 9, or blank were removed to maintain consistency and relevance. By focusing on variables with relatively complete data and consistent availability across the selected years, this study aimed to mitigate the complexities and variations inherent in the dataset. This rigorous preprocessing was crucial for ensuring a robust and accurate analysis of the factors associated with heart disease.

The preprocessing steps involved removing rows with missing values and eliminating non-independent variables. For instance, instead of keeping multiple variables such as LANDSEX, CELLSEX, and COLGSEX, which all pertain to sex, this paper consolidated these into a single variable, SEX. After selecting 36 relevant variables, this study converted all feature values to numerical formats, treating responses such as 'don't ask' and 'missing'. Some variables were further normalized and standardized to ensure consistency. Given the imbalance in the dataset, where the population with coronary heart disease or myocardial infarction was relatively small, this study created dummy variables for 'Diabetes,' 'Marital Status,' 'Rent Home,' and 'Employment' to address this issue. Finally, the preprocessed data was split into training and testing sets for subsequent analysis.

2.2 Machine Learning-based Prediction

2.2.1 Introduction of Machine Learning Workflow

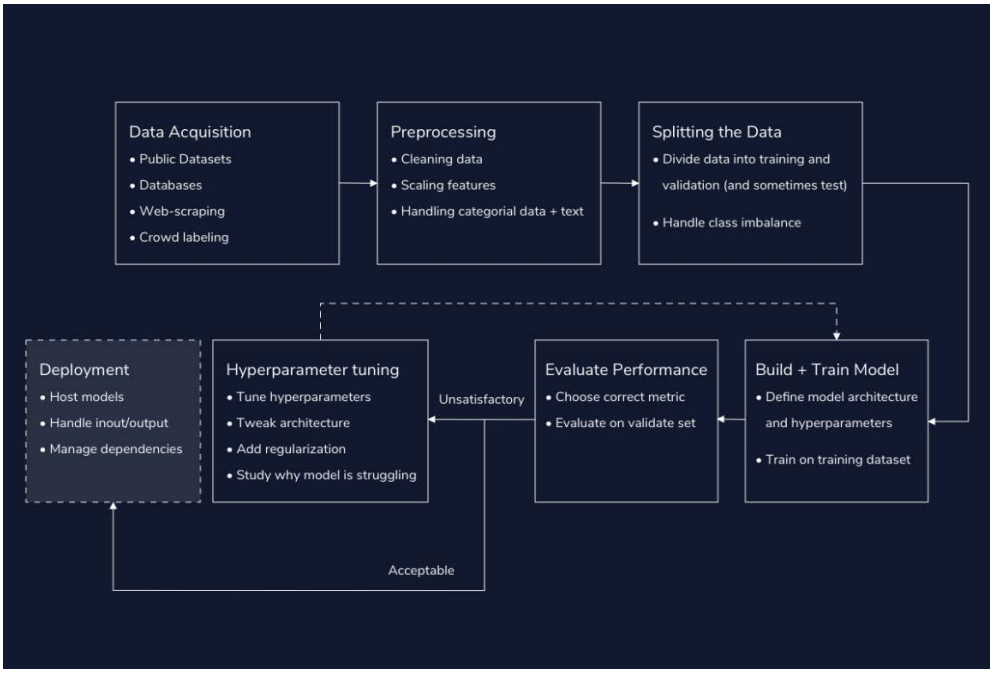

Fig. 1 The machine learning workflow [7].

In discussing the machine learning workflow shown in Fig. 1, it is essential to follow a structured process that spans from problem definition to model deployment [7]. The workflow begins with problem definition, where specific objectives and challenges are clearly identified. This is followed by data acquisition, which involves gathering the relevant data necessary for analysis. Once collected, preprocessing is performed to clean and prepare the data. This includes handling missing values, encoding categorical variables, and normalizing features.

After preprocessing, the data is split into training and testing subsets. This separation allows for a realistic evaluation of the model's performance on unseen data. During the build and train model phase, this study employs a Random Forest model, an ensemble learning method known for its robustness and efficacy in managing complex datasets. The model is trained using the training data.

Subsequent to training, performance evaluation is conducted to determine the model's accuracy and effectiveness. Various metrics such as precision, recall, and F1

ISSN 2959-6157

score are utilized to assess performance. The next step, hyperparameter tuning, involves optimizing parameters like the number of trees in the Random Forest model.

Finally, the model undergoes deployment, where it is integrated into a production environment to make real-time predictions or decisions. This structured approach ensures that each stage of the machine learning workflow is meticulously managed, resulting in the development of effective and reliable models [7].

2.2.2 Random Forest-based Prediction

In the realm of machine learning, the Random Forest algorithm remains a prominent tool due to its effectiveness in both classification and regression tasks. This ensemble method builds multiple decision trees during training and combines their outputs to enhance prediction accuracy and reduce overfitting [8]. Each tree in the forest is constructed from a random subset of data and features, utilizing techniques like bootstrapping and feature randomness to improve model robustness and generalization [9]. Implementing Random Forest using scikit-learn involves fine-tuning hyperparameters such as n_estimators and max_depth to achieve optimal performance. Furthermore, Random Forest provides critical insights into feature importance, which is particularly valuable for tasks such as predicting heart disease, allowing for the identification of key predictors [10]. Model evaluation typically involves metrics such as accuracy and confusion matrix to measure effectiveness and reliability.

This research began by utilizing train_test_split from the sklearn.model_selection module, with parameters set to test_size=0.2 and random_state=42 to ensure reproducibility. This step divided the data into training and testing subsets, preparing it for model evaluation. Subsequently, this study employed the RandomForestClassifier from the sklearn.ensemble module, configuring it with n_estimators=200 and random_state=42 to build a robust ensemble model. After training the Random Forest model, this study examined feature importance using the rf_model.feature_importances_ attribute. To better interpret these results, this paper sorted the features by their importance values, creating a DataFrame importance_df and sorting it in descending order. This sorted importance data allows to assess and visualize the significance of each feature for the year under analysis.

3. Results and Discussion

3.1 The Performance of the Model

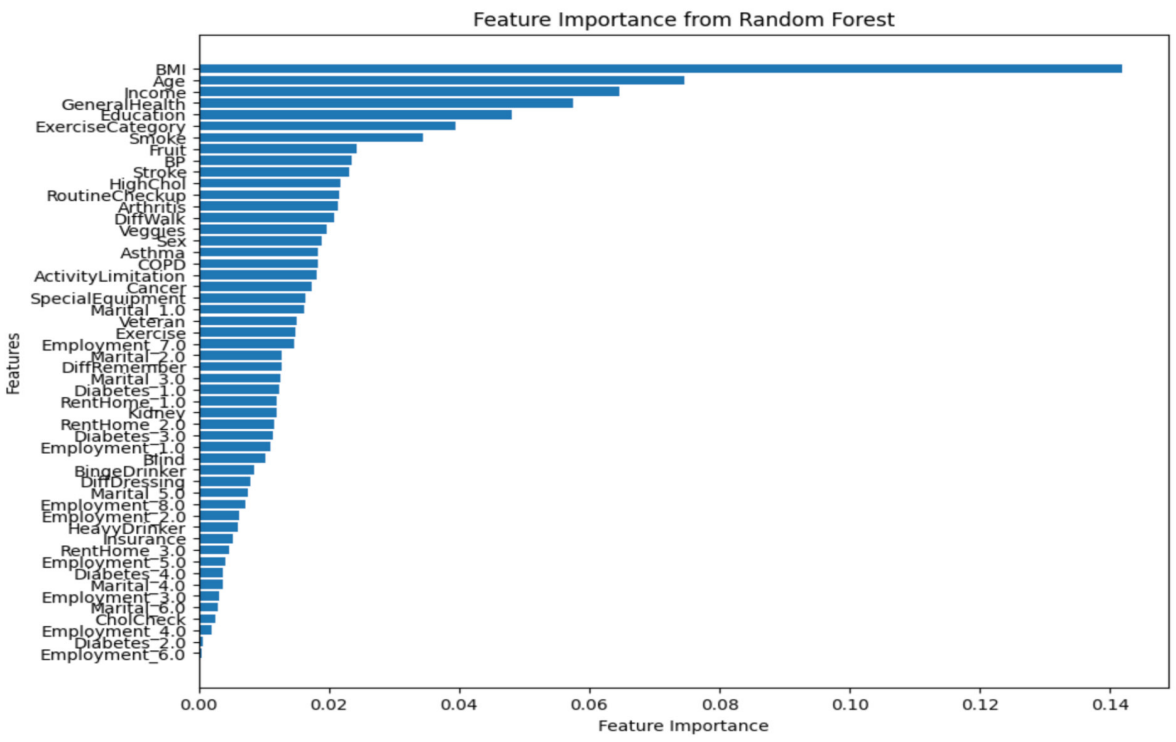

Fig. 2 The feature importance for 2015 (Photo/Picture credit: Original).

Fig. 2 illustrates the feature importance for the year 2015, showcasing the top five features ranked in descending order as Body Mass Index (BMI), Age, Income, General Health, and Education. These primary indicators are fol-

lowed by Exercise Category, Smoke, Fruit, Blood Pressure (BP), and Stroke, which occupy ranks 6 through 10. BMI is a continuous measure of body fat based on height and weight, reflecting overall body composition and health. Age is categorized into fourteen levels from 18 years to 80 years or older, representing different stages of life and related health risks. Income represents annual household income from all sources, providing insight into socioeconomic status and access to healthcare resources. General Health is measured on an ordinal scale from 'excellent' to 'poor,' representing self-reported overall health status and perceived well-being. Education denotes the highest grade or year of schooling completed by the respondent, indicating educational attainment and potential health literacy and understanding. Exercise Category ranges from Highly Active to Inactive, representing varying levels of physical activity and overall lifestyle habits.

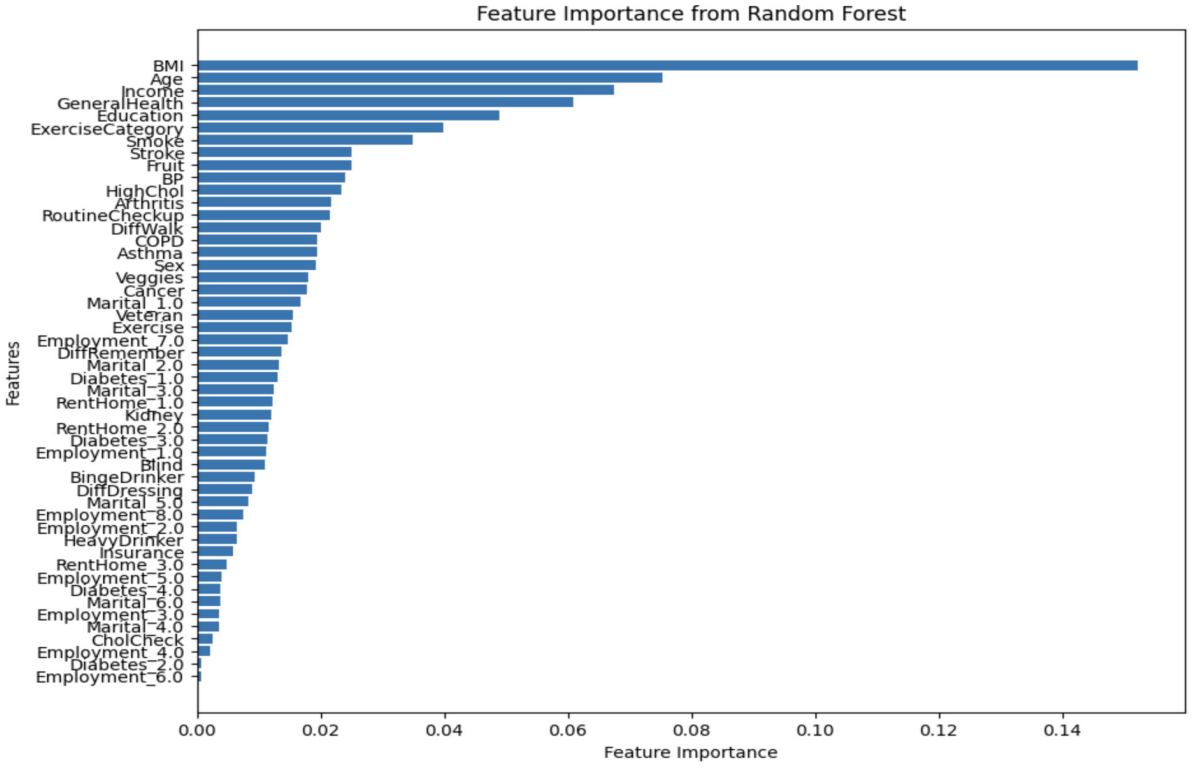

Fig. 3 Feature Importance for 2017 (Photo/Picture credit: Original).

Fig. 3 shows the feature importance for 2017. The top seven features remain consistent with those from 2015. However, the order of features ranked 8th to 10th—Stroke, Fruit, and BP—has shifted, suggesting changes in feature relevance and their impact on heart disease over time.

ISSN 2959-6157

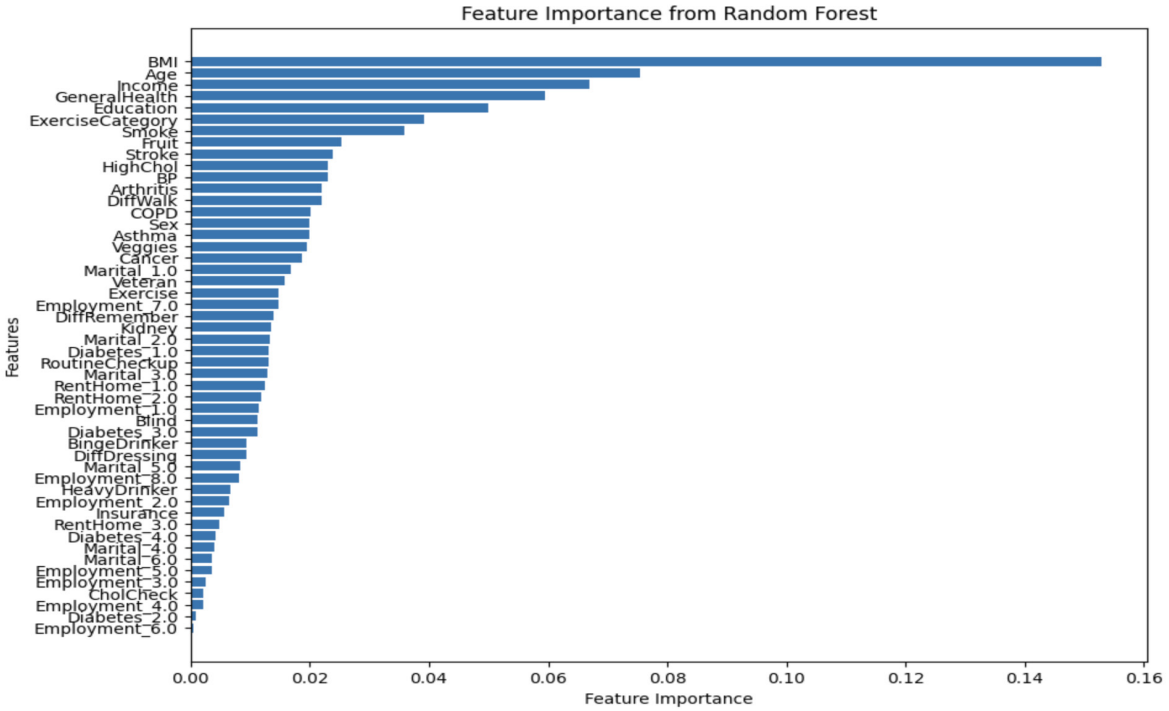

Fig. 4 Feature Importance for 2019 (Photo/Picture credit: Original).

Fig. 4 depicts the feature importance for 2019. The top seven features remain the same as in previous years. Features ranked 8th to 10th are Fruit, Stroke, and High Cholesterol, with High Cholesterol emerging as a significant factor, reflecting its increasing importance in heart disease risk assessment.

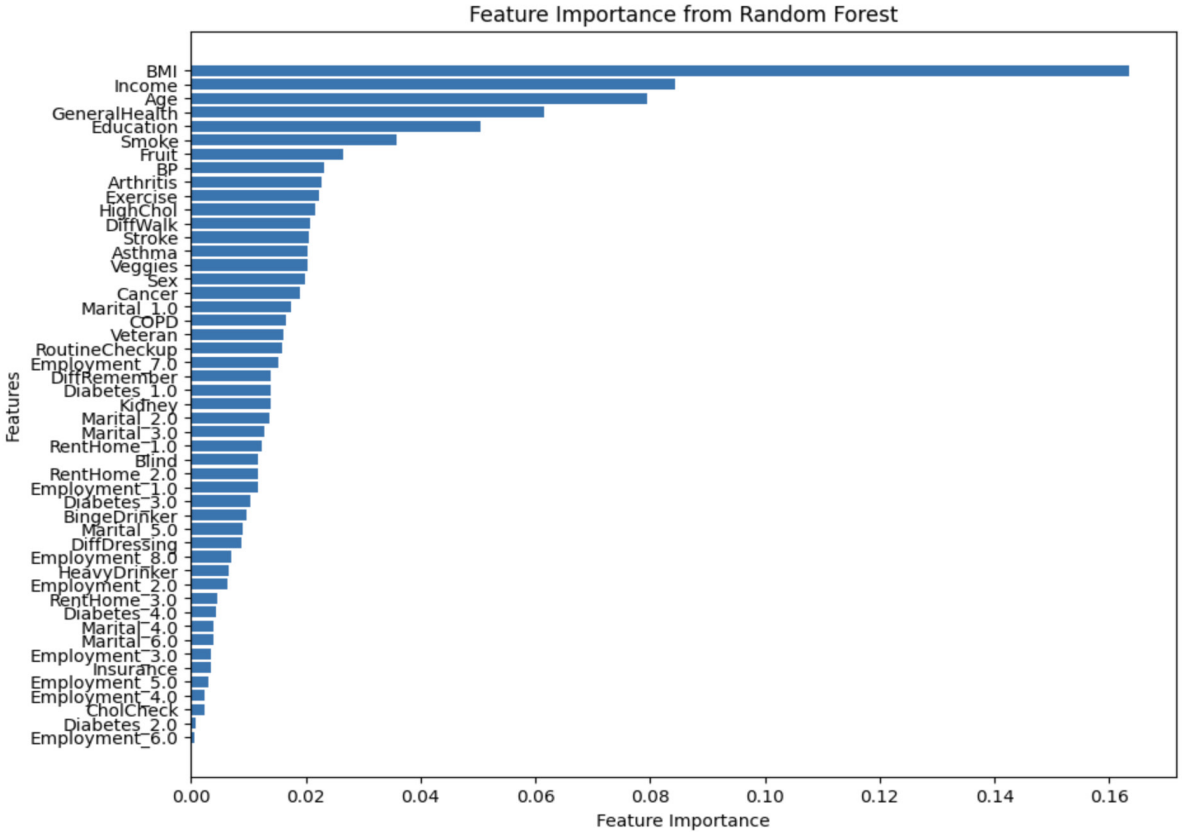

Fig. 5 Feature importance of 2021 (Photo/Picture credit: Original).

Fig. 5 displays the feature importance for 2021. The top five features are unchanged from 2015, 2017, and 2019. Features ranked 6th to 10th include Smoke, Fruit, BP, Arthritis, and Exercise. Notably, Exercise Category is absent from the 2021 dataset, likely removed due to redundancy with the Exercise variable, which captures recent physical activity levels. Despite this, Exercise remains crucial, demonstrating its continued relevance in health assessments. The detailed descriptions for additional features are as follows: 1) Arthritis: Indicates whether respondents have been diagnosed by a doctor with some form of arthritis, reflecting joint health and potential impact on physical activity. 2) Exercise: Represents whether adults have engaged in physical activity or exercise during the past 30 days, excluding occupational activities.

3.2 Discussion

Comparing these four years, BMI, Income, Age, General Health, Education, and Smoke consistently rank as significant factors for individuals with myocardial infarction (MI) or coronary heart disease (CHD). This consistency highlights their ongoing importance in predicting heart disease risk. The absence of Exercise Category in 2021 underscores changes in dataset variables, yet the inclusion of Exercise reinforces the critical role of physical activity in maintaining overall health. In 2019, High Cholesterol appeared among the top ten features, highlighting its relevance in assessing heart disease risk. Similarly, in 2021, Arthritis made it into the top ten. Although both High Cholesterol and Arthritis appeared in the top ten only once across the four years, they consistently ranked within the top fifteen, reflecting their significant but somewhat secondary role compared to the more consistently prominent features. This variability suggests that while certain factors maintain consistent importance, others may fluctuate in relevance based on evolving health trends and dataset characteristics.

4. Conclusion

This article explores how the ranking of important factors related to heart disease has changed over time. Using data from the Behavioral Risk Factor Surveillance System (BRFSS), which includes lifestyle and demographic variables for non-institutionalized adults aged 18 and older from 2015 to 2021. This study selected 36 relevant variables out of over 300 per year, removing rows with missing values and eliminating non-independent variables. The data was further normalized and standardized to ensure consistency. Using a random forest algorithm, this paper identified the feature importance for each year that BMI, Income, Age, General Health, Education, and Smoking consistently rank as significant factors for myocardial infarction (MI) or coronary heart disease (CHD). While the analysis primarily focuses on individual feature importance, it does not delve into potential interactions between features. For example, the combined effects of exercise and fruit consumption on heart disease risk are not fully explored, which could reveal more nuanced insights. Future research could investigate these interactions to provide a more comprehensive understanding of the complex relationships between heart disease risk factors.

References

[1] Pal M, Parija S. Prediction of heart diseases using Random Forest. J Phys: Conf Ser. 2021;1817(1):012009.

[2] Wiharto W, Kusnanto H, Herianto H. Performance analysis of multiclass support vector machine classification for diagnosis of coronary heart diseases. arXiv. 2015.

[3] Hsieh CH, Li YS, Hwang BJ, Hsiao CH. Detection of Atrial Fibrillation Using 1D Convolutional Neural Network. Sensors (Basel). 2020;20(7):2136.

[4] McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94.

[5] Korolev IO, Symonds LL, Bozoki AC. Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification. PLOS ONE. 2016;11(2)

[6] Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System annual survey data. 2015. Available from: https://www.cdc.gov/brfss/annual_data/annual_data.htm

[7] Codecademy Team. Deep learning workflow. Codecademy. 2024, Available from: https://www.codecademy.com/learn/paths/deep-learning

[8] Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media; 2019.

[9] Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2(3):18-22. Available from: https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf

[10] Rashmi S, Kumar P, Reddy K. Feature selection and classification of heart disease using machine learning algorithms. J Healthc Eng. 2021;2021:8834729.