Physica Scripta

PAPER

Integrating machine learning techniques for enhanced prognostic modeling of heart failure risk in the american population

Vishnumon S[1], **Sayooj Aby Jose**[1,2] **and Anuwat Jirawattanapanit**[2]

[1] School of Data Analytics, Mahatma Gandhi University, Kottayam, India
[2] Department of Mathematics, Faculty of Education, Phuket Rajabhat University, Phuket, Thailand

E-mail: sayooaby999@gmail.com

RECEIVED
29 November 2024

REVISED
26 May 2025

ACCEPTED FOR PUBLICATION
2 June 2025

PUBLISHED
23 June 2025

Keywords: F1 score, heart failure, logistic regression, machine learning, naive bayes, random forest, XGBoost

Abstract

Heart failure remains a significant public health challenge, with high rates of morbidity and mortality. Early detection is critical for effective management, yet traditional diagnostic approaches often lack precision and are prone to variability. This study aims to develop a machine learning model specifically designed to predict heart failure onset with high accuracy, enabling timely intervention and personalized treatment strategies. Using a comprehensive dataset of U.S. residents, we analyzed various clinical and laboratory parameters to train and evaluate machine learning algorithms, including Logistic Regression, Random Forest, Naïve Bayes and XGBoost. The top-performing model was selected based on key evaluation metrics— F1-score, Sensitivity, Specificity and the confusion matrix— ensuring a balanced assessment of its predictive capability. By identifying high-risk individuals before symptoms escalate, this model could allow healthcare providers to proactively manage heart failure, potentially reducing hospitalizations and enhancing quality of life. The findings underscore the potential of machine learning to transform heart failure care, supporting a shift toward more data-driven and preventative healthcare in the United States.

1. Introduction

The rising cost of healthcare is influenced by genetic and lifestyle factors, with large volumes of health data being generated. This data poses challenges for extracting valuable insights, but the increasing use of data analytics helps hospitals and Non-Governmental Organizations (NGOs), which are non-profit organizations working for public welfare, better utilize this data to improve healthcare outcomes. Heart disease, particularly heart failure (HF), is a major health issue, with approximately 6.7 million Americans over 20 affected by HF, according to the National Health and Nutrition Examination Survey (NHANES) [1]. HF is a chronic condition in which the heart struggles to pump blood effectively, affecting 26 million people globally, contributing to high morbidity and mortality, and leading to increased healthcare costs and reduced quality of life [2]. Chronic heart conditions, including vascular disease, coronary artery disease (CAD), and arrhythmias, are common in modern lifestyles, causing blockages or narrowing of blood vessels that may lead to heart attacks or chest discomfort. Early detection is crucial for timely management, which can help reduce the impact of cardiovascular diseases (CVD), including symptoms like shortness of breath, chest tightness, and irregular heartbeats.

CVDs are the leading cause of death globally, contributing significantly to healthcare costs and premature deaths. HF, affecting millions worldwide, has a considerable financial burden, with costs expected to rise as the global population grows [3]. CVDs are responsible for a large proportion of premature deaths and are often preventable by addressing behavioral and environmental risk factors such as poor diet, obesity, smoking, and physical inactivity [4]. Early detection of CVDs is critical for initiating timely interventions, which can reduce the disease's impact. The financial strain caused by HF is substantial, with projections indicating an increase in incidence by 2030 [5]. Recognizing early warning signs of deteriorating heart health is essential to prevent heart attacks and other complications. The healthcare industry is generating vast amounts of data, but current methods have yet to fully utilize this information to identify early symptoms. Traditional prediction models for

HF often rely on demographic and clinical factors but demonstrate limited accuracy. Misdiagnosis and incorrect treatment can harm patients and damage the reputation of healthcare institutions. Additionally, healthcare systems face challenges due to a shortage of professionals, particularly cardiologists. Integrating advanced models with medical information systems could provide significant benefits, although clinical approaches for predicting HF still face accuracy limitations and high costs. HF is a major public health concern, impacting millions globally and creating a considerable economic burden. Research shows that the lifetime risk of HF has increased to 24%, implying that approximately 1 in 4 people will experience HF in their lifetime [6]. In the United States alone, HF accounted for over 10% of total cardiovascular healthcare expenditures in 2012, totalling more than $31 billion. Around 550,000 new cases are diagnosed annually [7]. The major risk factors leading to HF are age, hypertension, diabetes, obesity, lifestyle, etc. As of now the traditional methods in diagnosing HF is Electrocardiogram, Blood Tests, Echocardiogram and Medical History and Physical Exam.

This study addresses the research gap related to the high mortality rate associated with HF by employing machine learning algorithms to enable early-stage prediction. Timely prediction is essential for effective intervention, which can substantially improve patient outcomes and lessen the strain on healthcare systems. This study takes a small step towards improving the accuracy of diagnoses based on patient's medical histories, with the goal of saving the lives of HF patients. As the disease becomes more widespread, it is essential to develop new testing methods that are accurate, quick, and efficient. To enhance predictive capabilities, this study explored the potential of machine learning techniques. Various studies have employed different machine learning algorithms to predict CVD, such as Naive Bayes, Random Forest, Gradient Boosting, and Logistic Regression. By incorporating a broader spectrum of patient data, the model aimed to improve risk stratification. The results indicate that the Random Forest model effectively predicts HF, outperforming traditional methods in terms of F1 score, sensitivity and specificity. Many researchers have used data mining techniques and machine learning algorithms to enhance accuracy. Machine learning methods are effective at forecasting diseases based on medical data. Given the severity of the disease, these techniques and algorithms assist in assessing various factors that can aid in accurately diagnosing a patient even with limited data. A framework for predicting CVD can help healthcare professionals assess heart health based on clinical data provided by patients within the system. However, additional research is needed to confirm these results in larger and more diverse populations. The contributions of this study are presented as follows

- Create a predictive model using machine learning algorithms to assess the likelihood of HF.
- Analyze the correlation between age groups and their associations with HF, heart disease, and smoking habits.
- Tackles the critical issue of data imbalance in HF identification by applying effective machine learning techniques.
- Identify and select the most effective classification model for predicting HF.
- Apply the chosen model to make predictions on previously unseen data.
- Measure the model's performance using evaluation criteria such as the F1 score, Sensitivity and Specificity.

2. Literature review

This study report focuses on the prediction of HF. In today's world, HF has become a complex issue. As heart disease remains a significant global health concern, it has prompted extensive research into the use of machine learning algorithms for predictive modeling. Hoque *et al* (2024) have presented promising findings using their SVM-based approach for predicting heart disease. Their study highlights the accuracy of SVM in identifying cardiovascular conditions [8]. Duraisamy *et al* (2024) [9] also employed SVM, emphasizing its effectiveness in modeling cardiac diseases. In 2024 [10], Victor *et al* explored the use of various machine learning techniques for non-invasive HF diagnosis. Their research led to the development of currently implemented non-invasive diagnostic methods that are essential for early detection and treatment. Qadri *et al* (2023) [11] emphasized the importance of feature engineering techniques in enhancing the identification of heart disease. M M Mamun analyzed the UCI HF dataset, which includes relevant medical information from 299 HF patients. By applying various machine learning methods-such as Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), XGBoost, LightGBM, Random Forest (RF), K-Nearest Neighbors (KNN), and Bagging-the author achieved improved prediction outcomes [12]. In this study[13], the authors proposed a machine learning-based approach to identify key features for heart disease prediction. The objective of their work is to design and implement a feature-ranking system using machine learning techniques. As a result, ML plays a vital role in saving lives, supporting physicians, enabling deeper research for valuable insights, assisting in complex

decision-making, and helping businesses develop innovative solutions to achieve critical goals. In this study [14], the authors developed a graphical user interface (web application) along with an automated system for detecting HF using multiple machine learning models. The publicly available HFCR dataset from Kaggle was utilized to evaluate the performance of four selected classification models: Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbors (KNN), and Random Forest (RF). Hyperparameter tuning was performed using GridSearchCV. The findings highlight that advancements in machine learning techniques have significantly improved the ability to predict and diagnose cardiac diseases, a task considered highly challenging by the medical community.

Zhang *et al* (2024) [15] faced challenges with ECG models not capturing all relevant cardiac features, particularly in atypical HF cases. To mitigate this, they introduced a multi-lead ECG approach and optimized the model for computational efficiency, ensuring its real-time application in clinical settings. Sonia *et al* (2024) [16] pointed out that some features used in their model might not be easily available in clinical settings, and the model's applicability could be limited due to a homogeneous patient group. They focused on feature selection to prioritize easily obtainable data and tested the model on various hospital datasets to improve its generalizability. Choi *et al* (2024) [17] encountered challenges with their model's performance for patients with comorbid conditions or different stages of HF. To address this, they used data augmentation to increase the variety of ECG features and incorporated multi-center data to improve model robustness. Across these studies, general strategies such as data imputation, feature selection, cross-validation, and model interpretability have been employed to mitigate limitations, ensuring that the models are both practical and reliable for real-world clinical applications. Jialin *et al* (2022) [18] study addressed key limitations such as the lack of interpretability in traditional heart failure prediction models, the retrospective nature of the data, and reliance on a single data source. To overcome these, the authors used SHAP to make the XGBoost model interpretable, applied robust validation techniques on a large patient cohort, and acknowledged the need for external validation to improve generalizability. These efforts enhance the clinical applicability and reliability of their predictive model. Jingxian *et al* (2025) [19] faced limitations including class imbalance in the dataset, lack of external validation, and limited interpretability of complex models like XGBoost. To address these, the authors used SMOTE to balance the dataset, applied five-fold cross-validation for robust model evaluation, and employed SHAP to enhance interpretability by highlighting key predictive features. These measures improved the model's fairness, reliability, and clinical relevance. Sona *et al* (2024) [20] study highlights limitations such as the oversimplification of heart failure classification using only LVEF, variability in circadian ECG features due to external factors, and biases from its retrospective design. To address these, the authors incorporated circadian ECG features and machine learning techniques to improve classification accuracy and capture more complex physiological patterns. They also acknowledged these limitations and recommended future prospective studies with broader diagnostic parameters for validation. Mohammed Khalid *et al* (2025) [21] faced limitations such as a limited dataset size, potential overfitting, and the absence of external validation, which could affect the model's generalizability. To address these, the authors used robust algorithms like Random Forest and SVM, applied relevant feature selection, and recommended future research with larger, more diverse datasets and external validation. These steps aim to improve model reliability and applicability across broader clinical settings. Srinivas and Vempathy's (2024) [22] research identified key limitations, including reliance on a single dataset without external validation, insufficient diversity in patient data, and difficulty interpreting complex models like Random Forest. To mitigate these issues, the authors compared several machine learning algorithms and employed comprehensive performance metrics for evaluation. They also emphasized the need for future studies to use broader datasets and external validation to enhance the reliability and clinical applicability of the models. Hosea and Mulapnen (2025) [23] faced challenges such as class imbalance, overfitting due to a small dataset, and a lack of external validation, which could affect the generalizability of the results. To address these, the authors applied SMOTE for data balancing, used ANOVA and PCA for feature selection and dimensionality reduction, and optimized model performance through hyperparameter tuning and cross-validation. Furthermore, Alexander *et al* (2025) [24] encountered several challenges, including a limited dataset of 122 patients, average model accuracy with a ROC-AUC of 0.657, and low sensitivity, which increased the risk of undetected hospitalizations. To overcome these, the authors adopted transparent models like LASSO-regularized logistic regression and RuleFit to improve clinical interpretability. They also used fivefold cross-validation and selected basic, frequently recorded clinical features to build more reliable predictive models.

3. Proposed methodolgy

The overall framework of the proposed model, as illustrated in figure 1, consists of multiple critical stages to ensure a streamlined approach to HF prediction. Beginning with dataset acquisition, the framework sources a robust dataset that encompasses a wide range of clinical variables pertinent to HF risk factors. This is followed by

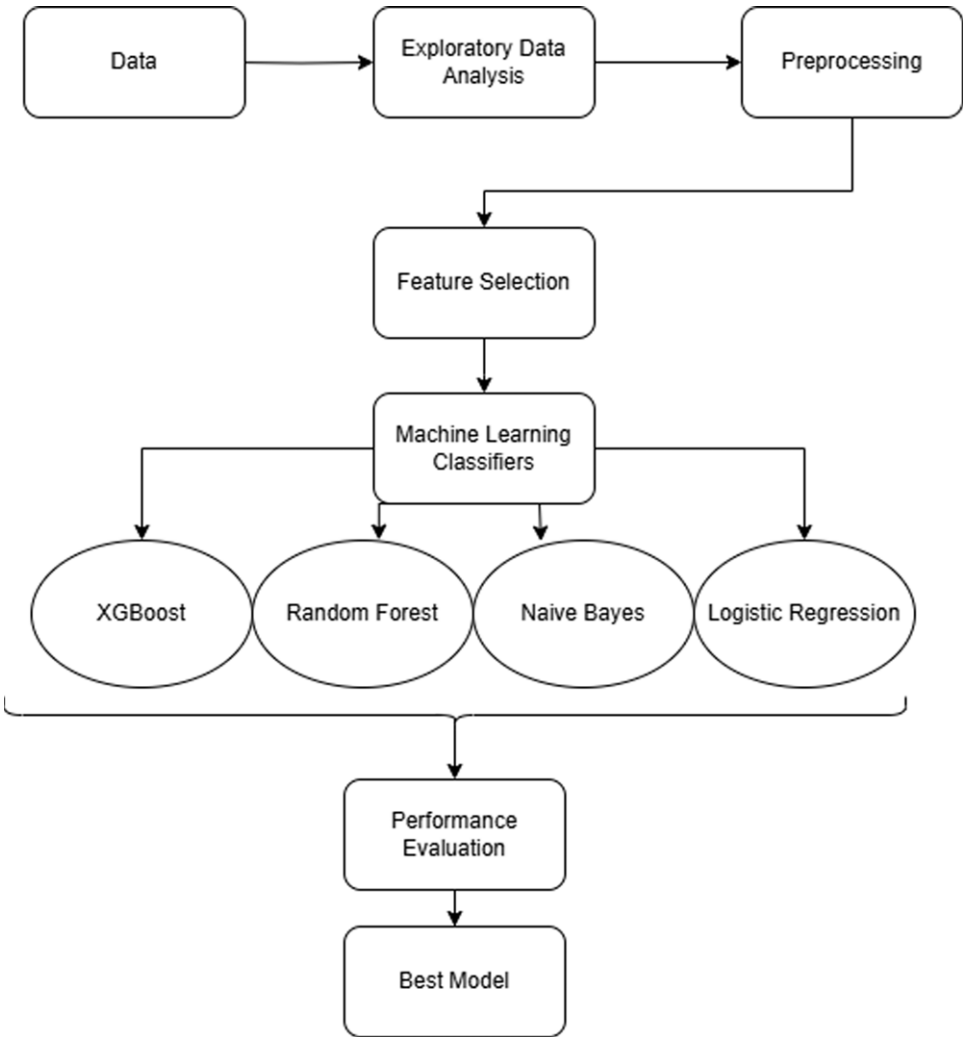

Figure 1. Framework outline for the proposed model and intervention strategies.

exploratory data analysis (EDA), where the data is examined to uncover initial insights, detect patterns, and identify any irregularities, such as outliers or missing values, which could impact model performance.

Subsequent data preprocessing prepares the dataset for modeling by addressing these inconsistencies and encoding categorical data to maintain data integrity. Data visualization provides a clearer view of the relationships between variables, highlighting important correlations and distributions that guide the next step, feature selection. By selecting the most relevant features, the model avoids unnecessary complexity and enhances prediction accuracy. Afterwards, The HF classification phase applies machine learning algorithms, including XGBoost, Random Forest, Naive Bayes, and Logistic Regression, to classify individuals' risk, with the model tailored to effectively address HF detection. The performance evaluation employs metrics such as F1-score, Sensitivity, Specificity, and a confusion matrix to measure the model's predictive capabilities. These metrics reflect the model's performance exclusively on the test data, ensuring an unbiased evaluation. External testing was conducted after selecting the best model, using a separate dataset that was not involved in training or testing the model. This organized framework aims to provide an accurate, data-driven tool for the early detection and intervention of HF while ensuring that the model generalizes effectively across varied patient profiles.

3.1. Data collection

The dataset used in this study is sourced from the Centers for Disease Control and Prevention (CDC) [25] and is a key part of the Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is a comprehensive national survey program that conducts annual telephone interviews to gather data on the health status and risk factors of U.S. residents. The CDC, a leading national institution dedicated to advancing public health through disease prevention, environmental health, and health promotion, oversees this initiative. Founded in 1984 with involvement from 15 states, the BRFSS has since grown to include data collection across all 50 states, the District

Table 1. Dataset features.

Sr. No	Features	Percentage of Population
1	HeartDisease	91% with No HeartDisease and 9% with HeartDisease
2	BMI	Average BMI of the population is 28.32
3	Smoking	59% Non-Smokers and 41% Smokers
4	AlcoholDrinking	93% is non-alcoholic and 7% is alcoholic
5	PhysicalHealth	70.85% have poor, 6.10% have excellent and 23.05% have in between.
6	MentalHealth	64.23% have bad, 5.43% have excellent and 30.34% have in between.
7	DiffWalking	86% have difficulty to walk while 14% don't have
8	Sex	52% Female and 48% male
9	AgeCategory	45.17% is aged between 60 and 80, 42.88% between 30 and 60 and 11.95% between 18 and 29.
10	Race	76.67% white, 8.58% Hispanic, 7.17% Black, 3.41% Other, 2.52% Asian and 1.65% American Indian
11	Diabetic	86.44% non-diabetic, 13.56% diabetic
12	PhysicalActivity	78% engages in physical activities and 22% does not
13	GenHealth	35.6% have Very good, 29.12% Good,20.9% have excellent, 10.84% have fair and 3.54% poor
14	SleepTime	Average Sleeptime is 7 hours
15	Asthma	87% have asthma and 13% does not
16	KidneyDisease	96% does not have Kidney disease and 4% have it
17	SkinCancer	91% with No skin cancer and 9% with skin cancer
18	Stroke	96% No stroke and 4% have stroke

of Columbia, and three U.S. territories. This extensive survey system performs over 400,000 adult interviews annually, making it the largest and most consistently conducted health survey globally.

The specific dataset employed for this analysis is titled 'Indicators of Heart Disease (2022 UPDATE)' [26]. The dataset is publicly accessible to everyone and is accessible through Kaggle, the premier data science community renowned for its advanced tools and resources that support data science endeavors. In the course of examining this dataset, we identified a range of factors (variables) that exhibit both direct and indirect influences on HF. In light of these observations, we proceeded to meticulously select the most relevant variables for subsequent analysis to ensure the focus on the most impactful factors.

3.2. Dataset

Our dataset encompasses medical records for 319,794 patients, featuring a range of attributes including 'HeartDisease', 'BMI', 'Smoking', 'AlcoholDrinking', 'PhysicalHealth', 'MentalHealth', 'DiffWalking', 'Sex', 'AgeCategory', 'Race', 'Diabetic', 'PhysicalActivity', 'GenHealth', 'SleepTime', 'Asthma', 'KidneyDisease', 'SkinCancer' and 'Stroke'. This dataset is openly accessible on Kaggle, a platform renowned for its extensive collection of data science resources.

In this dataset, 'Stroke' serves as the target variable, while the remaining attributes act as predictor variables. The majority of these attributes are of object type, with 'BMI' and 'SleepTime' being classified as int64 and float64 types, respectively. Given the open nature of this dataset, it is freely available for analysis. For this study, we will leverage these features to develop predictive models for stroke using various machine-learning algorithms. Table 1 presents a comprehensive summary of all the features and percentage of population.

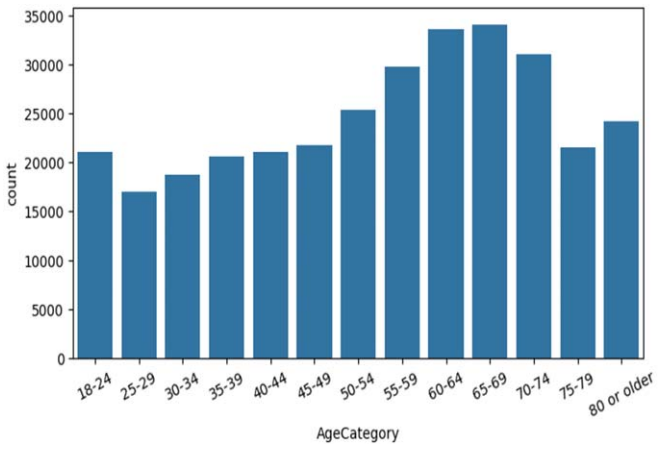

Figure 2. Age distribution.

3.3. Exploratory data analysis

Conducting EDA is indispensable when beginning the process of analyzing datasets, aimed at comprehensively summarizing and visualizing the key attributes of a dataset. The primary objective of EDA is to thoroughly understand the underlying structure of the data, discern patterns, identify anomalies, and formulate hypotheses. The process typically begins with data acquisition, followed by an initial examination through descriptive statistics and graphical representations. Subsequently, data cleaning is performed to address issues such as missing values and outliers. This is followed by an in-depth analysis to extract more nuanced insights, culminating in the presentation of findings to effectively communicate the results. In this study, we have utilized different visualization methods to gain a more comprehensive insight into the data.

3.3.1. Distribution of age and sleeptime

We examined the distribution of age and sleep patterns, recognizing the significant roles both factors play in contributing to HF risk. Age is a well-established determinant, as the risk of cardiovascular conditions generally increases with advancing age. Similarly, sleep quality and duration are critical factors influencing cardiovascular health. Both insufficient and excessive sleep can elevate stroke risk due to their impact on physiological functions, such as blood pressure regulation and inflammation. Adults generally need a minimum of seven hours of sleep per night [27]. Over one in three American adults report failing to meet the recommended sleep duration. Although short-term sleep deprivation may be manageable, prolonged lack of sleep can result in significant health problems and worsen existing conditions [28].

The visual representation highlights notable variations in participant age ranges and typical sleep durations across the dataset. The age distribution chart in figure 2 reveals that the highest number of individuals falls within the 65–69 age range, with the x-axis representing age categories and the y-axis representing the count. As illustrated in sleep time distribution chart in figure 3, individuals aged over 80 report the highest average sleep duration, whereas those in the 30–39 age group report the lowest. In this chart, the x-axis denotes age, and the y-axis indicates sleep duration in hours.

3.3.2. Distribution of heart disease and stroke

Heart disease and stroke impact individuals across various age groups, with their prevalence and severity increasing markedly with age. In younger adults (under 45), these conditions are relatively uncommon and are often linked to genetic predispositions, lifestyle choices, or specific underlying health conditions. As individuals reach middle age (45–65), lifestyle factors such as unhealthy diets, physical inactivity, and smoking become significant contributors to the risk of heart disease and stroke.

In older adults (65+), the risk is substantially higher, primarily due to the cumulative effects of aging on the cardiovascular system and other age-related health conditions, such as diabetes or atrial fibrillation. This trend reflects the natural progression of cardiovascular risk factors, including hypertension and cholesterol buildup, that intensify with age. An age-related increase in the prevalence of both heart disease and stroke is evident, as shown in figures 4 and 5, respectively.

The bar chart in figure 4 illustrates the distribution of heart disease across different age groups. The data shows that individuals aged 60–64 have the lowest incidence of heart disease, while those aged 80 and older experience the highest incidence. In contrast, figure 5 presents the distribution of stroke across age groups,

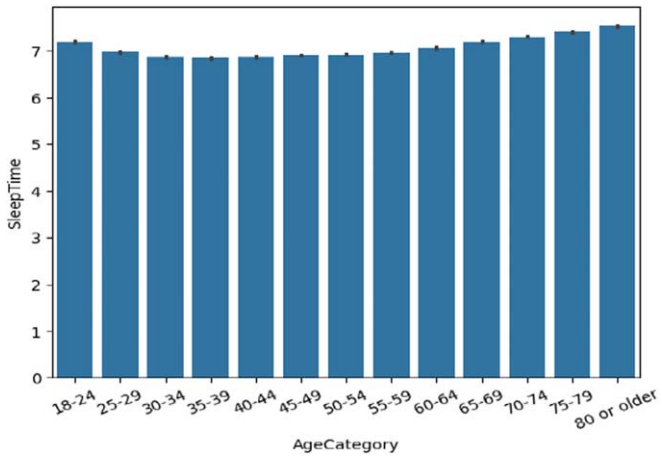

Figure 3. Distribution of sleep time.

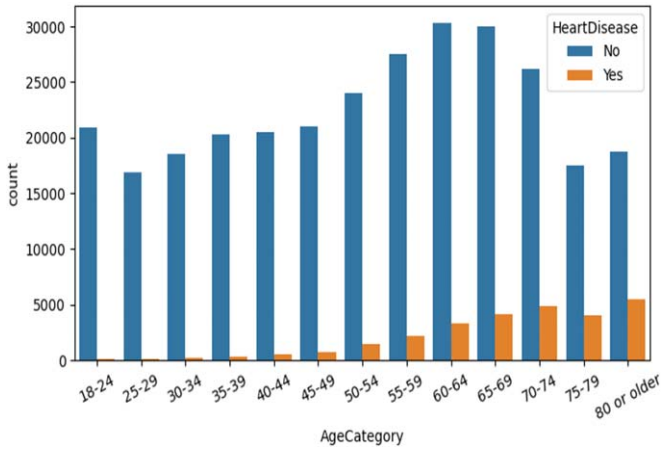

Figure 4. Distribution of heart disease.

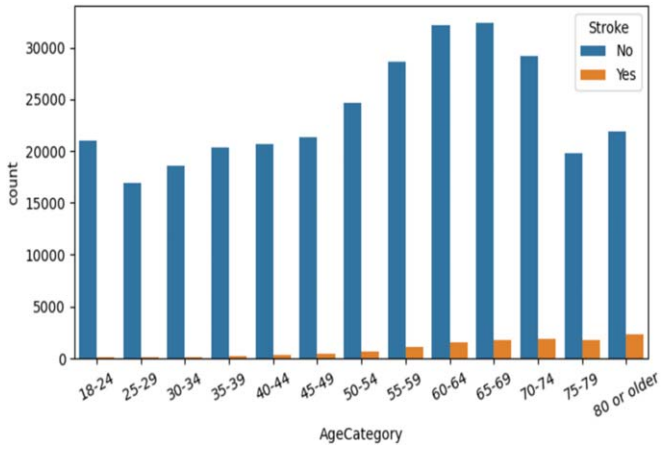

Figure 5. Distribution of HF.

revealing that individuals aged 25–29 have the lowest stroke rates, whereas the highest stroke incidence is observed in individuals aged 80 or older. These insights underscore the variability in heart disease and stroke prevalence across age groups, highlighting specific age-related risk patterns.

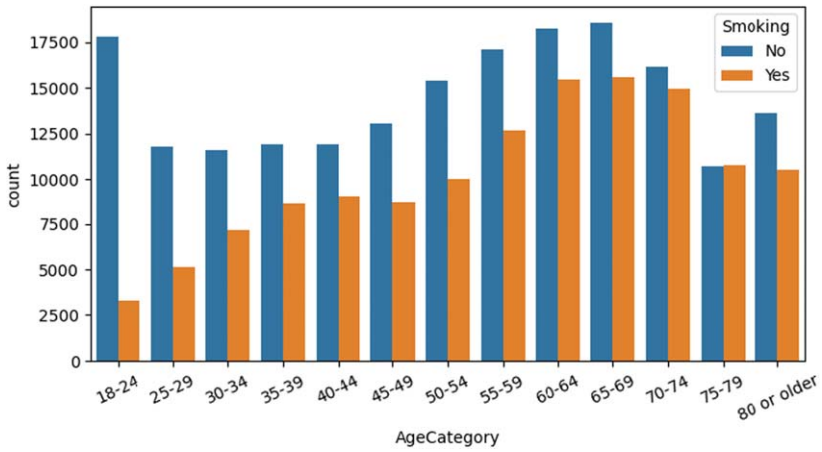

Figure 6. Distribution of smokers.

3.3.3. Distribution of smokers

Cigarette smoking is a leading contributor to CVD and is responsible for one in every four CVD-related deaths [29]. The American Heart Association (AHA) reports that smoking-related illnesses result in the deaths of more than 440,000 people each year in the United States. Many new smokers are young, including children and teenagers. Smokers face a much greater risk of developing lung conditions like lung cancer and emphysema, along with heart disease and stroke [30].

The highest representation of both smokers and non-smokers falls within the 65–69 age group, according to the data in figure 6. This observation highlights a significant concentration of smoking habits within this age range, providing insights into smoking prevalence among older adults.

3.4. Data pre-processing

Data preprocessing is an essential stage in the data analysis workflow, aimed at converting raw data into a structured and usable form. This process encompasses multiple tasks: cleansing data by addressing missing values and eliminating duplicates, transforming categorical variables into numerical representations and eliminating imbalance in the data. It also involves detecting and addressing outliers, ensuring data types are correct, etc. Managing missing values is an essential component of data preprocessing, as it greatly influences the reliability and precision of data analysis. Missing data can occur for multiple reasons, such as inaccuracies in data collection processes or inherent limitations of measurement instruments. If left unaddressed, missing values can lead to biased results, reduced statistical power, and unreliable model predictions. To mitigate these issues, it is essential to implement appropriate strategies for filling missing values. Removing duplicate values from a dataset is essential for ensuring data integrity and accuracy. Duplicates can distort analysis by inflating the significance of certain data points, leading to biased results and inefficient model performance. By eliminating duplicates, we maintain a consistent dataset that accurately represents unique observations, which enhances the reliability of statistical analyses and predictive models. Converting categorical variables into numerical formats is an essential step in preparing data for analysis and machine learning applications. Most statistical algorithms and machine learning models require numerical inputs to perform computations and make predictions effectively. Categorical variables, which represent qualitative attributes need to be converted into numerical values to be processed by these algorithms. Encoding transforms categories into a format that reflects the underlying structure of the data, allowing models to interpret and utilize these variables in their calculations.

On our dataset, we have checked for missing values and any kind of anomalies or outliers and our dataset was clean from outliers and missing values. We also checked for duplicate values and found duplicated values which were handled by keeping only the last occurrence. The next step involved transforming data values into the appropriate data types. To achieve this, we applied LabelEncoder and OrdinalEncoder for data transformation. The columns ‘HeartDisease’, ‘Smoking’, ‘AlcoholDrinking’, ‘Stroke’, ‘DiffWalking’, ‘PhysicalActivity’, ‘Asthma’, and ‘KidneyDisease’ are binary variables and were encoded using Label Encoding. The ‘AgeCategory’ column, being ordinal in nature, was encoded using an Ordinal Encoder to preserve the inherent order of the categories. Since most algorithms work with numeric data, encoding is essential for handling categorical variables. We have done manual label mapping to convert categorical values in the ‘Diabetic’ column into binary numerical format, where diabetic-related conditions are assigned ‘1’ and non-diabetic or borderline cases are assigned ‘0‘. Then, for ease, we have sorted the dataset in ascending order of age category. Following this, we examined the class distribution and discovered that our dataset was highly imbalanced. The dataset consists of

only 3.77% data which belongs to the positive class (stroke cases), while the remaining 96.23% represents the negative class (non-stroke cases). This significant class imbalance poses challenges for traditional machine learning models, as they tend to favor the majority class. To mitigate the significant class imbalance, we utilized the Hybrid Sampling (Random Oversampling + Random Undersampling). We have oversampled the minority class to increase its count and undersampled the majority class to decrease its count. Following these resampling techniques, the dataset was adjusted equal representation of stroke and non-stroke cases. If the class imbalance were not addressed, the model would tend to favor the majority class, resulting in low recall for stroke cases. Given the critical medical consequences of undetected stroke cases, balancing the dataset was essential to enhance the model's effectiveness in accurately identifying strokes. Consequently, data preprocessing constitutes a pivotal phase in the exploration of data and the application of machine learning methodologies.

3.5. Performance evaluation metrics

Four classification algorithms were applied to the dataset to identify the best-performing model by comparing F1 score, Sensitivity and Specificity. The algorithms evaluated include XGBoost, Naïve Bayes, Random Forest, and Logistic Regression. The performance of these algorithms was evaluated using various metrics with a particular focus on sensitivity due to the critical importance of minimizing false negatives in stroke detection. A concise summary of their performance assessment is presented in this subsection. A confusion matrix was generated to compute the sensitivity, specificity, and F1 score for each algorithm. The following formulas were utilized to determine these performance metrics [31].

$$F1\text{-}score = \frac{2TP}{2TP + FP + FN} \tag{1}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

Here, TP and TN denote true positives and true negatives, respectively, while FP and FN represent false positives and false negatives, respectively. The F1 score is the harmonic mean of the precision and recall. Its maximal value (of 1) is obtained when both precision and recall are 1, and its minimal value (of 0) is obtained whenever one of them is 0 [31].Sensitivity refers to the proportion of actual positives that the classifier correctly identifies, reflecting the number of positive instances accurately predicted by the model [32]. Specificity is the ability of the classifier to correctly distinguish negative outcomes [32]. A high sensitivity indicates that the model correctly identifies the majority of actual positive cases, reducing the likelihood of missing them. On the other hand, low sensitivity means the model overlooks many TPs, leading to an increased number of FNs. A high specificity indicates that the model generates fewer FP predictions. Conversely, low specificity means the model mistakenly classifies negative cases as positive, resulting in false alarms Sensitivity and specificity often present a trade-off. While it's important to optimize both, the prioritization of one over the other may depend on the specific application.

3.6. Feature selection

Feature selection is a data preparation technique used in machine learning to select key features that contribute the most to the prediction task, thereby enhancing model performance. It concentrates on a subset of important variables, reducing data dimensionality. This approach enables more efficient computation, faster training times, and lowers the risk of overfitting. It also enhances the model's ability to generalize by eliminating noise and redundant information, thus improving accuracy and predictive power. Moreover, choosing the appropriate features can streamline the model, making it more understandable and interpretable, while also offering insights into the underlying data relationships. Effective feature selection is vital for creating robust and efficient machine learning models that deliver reliable and actionable outcomes.

In this study, we performed a correlation analysis to examine the relationships between various features and the target variable, aiming to identify factors that may contribute to HF risk. By analyzing the strength and direction of these correlations, we sought to understand how individual characteristics and health indicators are associated with stroke likelihood. To visualize these relationships, we employed a correlation heatmap, as shown in figure 7. The heatmap provides an intuitive view of the correlations between multiple variables within the dataset, utilizing color gradients to represent correlation coefficients. Colors indicate the strength (ranging from −1 to 1) and direction (positive or negative) of each correlation, making it easy to identify strongly correlated variable pairs and gain insights into the dataset's underlying structure. This visual approach aids in uncovering patterns and potential predictive relationships that could inform stroke risk assessment.

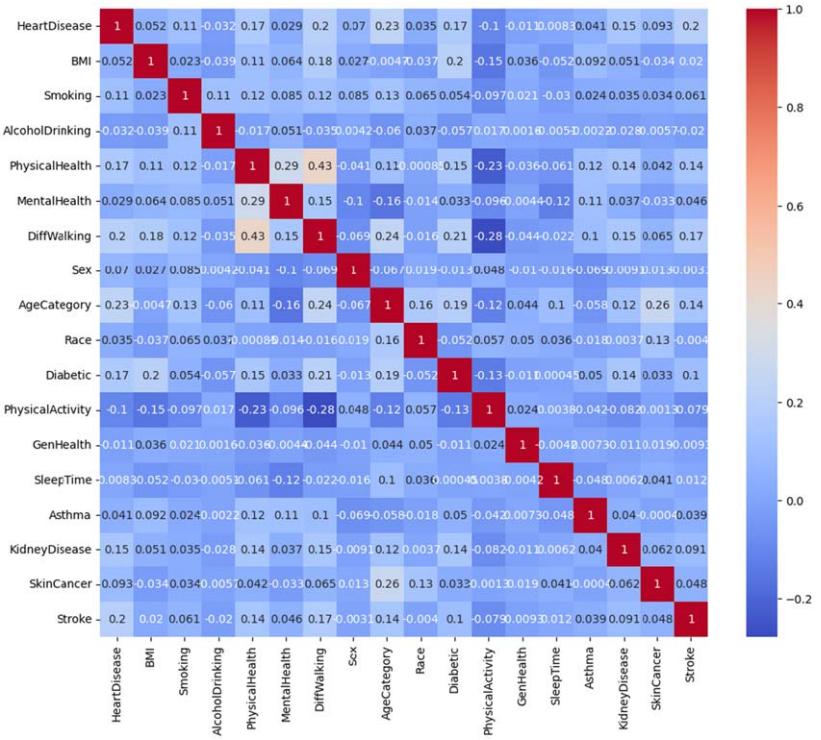

Figure 7. Correlation heatmap.

The correlation heatmap analysis indicates that features such as 'Heart Disease', 'Diff Walking', 'Diabetes', and 'Kidney Disease' exhibit a strong correlation with the target variable 'Stroke' while 'PhysicalActivity', 'GenHealth', 'MentalHealth' and 'AlchoholDrinking' exhibit a weak correlation. The correlation heatmap analysis reveals several variables that demonstrate a noteworthy association with stroke incidence. Among them, AgeCategory exhibits the strongest positive correlation, signifying that advancing age is a predominant risk factor for stroke. This is closely followed by DiffWalking, HeartDisease, GenHealth, and PhysicalHealth, highlighting that individuals experiencing mobility impairments, compromised general health, or underlying cardiovascular conditions are more susceptible to stroke. Variables such as KidneyDisease, SkinCancer, and Smoking display moderate correlations, suggesting a contributory but less pronounced influence. Conversely, PhysicalActivity demonstrates a negative correlation, indicating a potential protective role against stroke. Other factors, including Sex, Race, BMI, AlcoholDrinking, and MentalHealth, exhibit relatively weak linear associations with stroke risk. These insights are instrumental in guiding the feature selection process for predictive modeling and underscore the critical interplay between age, health status, and lifestyle factors in stroke vulnerability assessment. Our dataset contains the medical records of 319794 patients with features such as 'HeartDisease', 'BMI', 'Smoking', 'AlcoholDrinking', 'PhysicalHealth', 'MentalHealth', 'DiffWalking', 'Sex', 'AgeCategory', 'Race', 'Diabetic', 'PhysicalActivity', 'GenHealth', 'SleepTime', 'Asthma', 'KidneyDisease', 'SkinCancer' and 'Stroke'. Based on the correlation between variables and their clinical relevance, certain features were excluded from the analysis 'Race', 'SkinCancer','GenHealth', 'Sex', 'PhysicalHealth', 'MentalHealth' are the features which we have excluded here. The final set of features was selected based on insights from the correlation heatmap and the clinical relevancy of the variables. These features are presented in table 2. These features were retained as predictor variables, with 'Stroke' designated as the target variable.

3.7. Machine learning models

Machine learning, a subset of artificial intelligence, uses algorithms trained on datasets to create models. These models allow machines to carry out tasks that were once performed by humans, including image classification, data analysis, and forecasting price changes. Essentially, machine learning empowers computers to learn from data and generalize this knowledge to carry out tasks without specific programming. Machine Learning can be further categorized into Supervised, Unsupervised, Semi-Supervised, and Reinforcement Learning. since we are utilizing labeled data for classification, our approach falls under supervised learning. We are partitioning the dataset into two equal subsets, with 80% allocated for training and 20% reserved for testing. Some of the common algorithms that are used in this study are XGBoost, Random Forest, Naive Bayes, and Logistic Regression. These algorithms are used for classification problems.

Table 2. Selected features.

Sr. No	Features	Description
1	HeartDisease	Indicates if the individual is diagnosed with heart disease.
2	BMI	Representing body fat based on height and weight.
3	Smoking	Whether the individual is a smoker or not.
4	AlcoholDrinking	Indicates if the individual consumes alcohol.
5	DiffWalking	Whether the individual experiences difficulty walking
6	AgeCategory	Categorized age group of the individual
7	Diabetic	Indicates the diabetic condition
8	PhysicalActivity	Denotes whether the individual engages in physical activity
9	SleepTime	Denotes sleeptime in hours
10	Asthma	Indicates wheather the patient is suffering from asthma
11	KidneyDisease	Indicates if the individual has kidney-related illnesses.

3.7.1. XGBoost

XGBoost, or 'Extreme Gradient Boosting,' represents a sophisticated and high-efficiency gradient boosting framework engineered for scalable and computationally intensive machine learning tasks. Utilizing an ensemble learning methodology, XGBoost synthesizes multiple base models to yield a highly accurate, cohesive predictive model, optimizing predictive power and resilience across diverse data scenarios. Esteemed for its capacity to achieve benchmark performance across tasks such as classification and regression, XGBoost demonstrates a remarkable aptitude for managing datasets with missing values, effectively mitigating preprocessing requirements that would otherwise be essential. Moreover, its inherent parallel processing capability facilitates accelerated model training, rendering it advantageous for time-sensitive and large-scale data applications. The model is trained using distinct training and testing datasets, with rigorous and uniform evaluation metrics applied throughout, ensuring empirical consistency and robustness. These attributes, coupled with XGBoost's adeptness at navigating complex data architectures, have solidified its reputation as an indispensable tool for data scientists and researchers who require both methodological rigor and elevated performance in their machine learning endeavors.

The XGBoost model has been fine-tuned with a carefully selected set of hyperparameters to improve the accuracy of HF prediction. The number of boosting iterations is set to 200 (n_estimators=200), providing the model with more chances to learn intricate patterns within the data. The maximum depth of the trees is increased to 5 (max_depth=5), allowing the model to capture more complex relationships while still considering the risk of overfitting. A lower learning rate of 0.05 facilitates a smoother learning process, promoting stability and better generalization. To enhance regularization and reduce model variance, both subsample and colsample_bytree are configured at 0.8, ensuring that each tree is trained on a random subset of data samples and features. Since class imbalance has already been addressed during preprocessing, scale_pos_weight is set to 1, treating classes as balanced. The classification threshold is lowered to 0.4 to improve the model's ability to detect true stroke cases, increasing sensitivity while managing the trade-off with FPs. Overall, these adjustments aim to strengthen the model's performance and reliability in clinical risk prediction.

3.7.2. Random forest

Random forests, or random decision forests, are an ensemble learning approach widely used for classification, regression, and various predictive tasks. This method constructs multiple decision trees during training, where, in classification tasks, the output is determined by the most commonly chosen class, and in regression tasks, the average prediction from the trees is returned. Building multiple trees helps reduce the overfitting tendency seen in individual decision trees, enhancing the model's generalization ability. The initial algorithm for random decision forests was developed by Tin Kam Ho in 1995, applying the 'random subspace' method to implement the concept of stochastic discrimination in classification, initially proposed by Eugene Kleinberg. Leo Breiman and Adele Cutler expanded on this work, combining Breiman's bagging technique with random feature selection, which was first introduced by Ho and later independently by Amit and Geman. This combination created a robust ensemble of decision trees with controlled variance, and Breiman and Cutler registered 'Random Forests' as a trademark in 2006, now owned by Minitab, Inc. The algorithm is well-regarded for its

versatility and effectiveness in classification and regression tasks, often requiring minimal tuning to achieve strong results. It is especially appreciated for its capacity to manage various data types and offer insights into feature importance, making it a highly effective tool in the machine learning domain.

The Random Forest model was initially implemented using its default hyperparameters to establish a baseline for performance evaluation. Owing to its ensemble nature, which leverages bootstrapping and feature randomness, Random Forest is inherently robust and capable of delivering reliable results without immediate hyperparameter tuning. In this configuration, we employed the default number of decision trees, which is 100, to maintain consistency with the standard implementation. Using the default setup allows for an unbiased assessment of the model's predictive capability in its standard form. This approach facilitates fair comparisons with future tuned versions, highlighting the actual impact of optimization. Also, default settings promote model simplicity and interpretability at the initial evaluation stage.

3.7.3. Naive bayes

The Naïve Bayes algorithm is a supervised learning technique grounded in Bayes' Theorem, primarily utilized for solving classification problems. It is particularly effective in text classification tasks, especially when handling high-dimensional datasets. As one of the most straightforward yet efficient classification algorithms, Naïve Bayes facilitates the development of rapid machine learning models capable of making prompt predictions. Being a probabilistic classifier, it derives its predictions based on the likelihood of an object belonging to a particular class. Notably, Naïve Bayes excels in both binary and multi-class classification problems, often outperforming other algorithms in multi-class scenarios. Its ease of implementation and computational efficiency make it a preferred tool in areas such as spam filtering, sentiment analysis, and document categorization. The algorithm operates under the assumption of conditional independence between features given the class label, simplifying the computational process, although this assumption may not always align with real-world dependencies. Variants of the Naïve Bayes classifier, including Gaussian, Multinomial, and Bernoulli, are tailored for different data types, contributing to its versatility and widespread use.

The Naïve Bayes model is configured for HF prediction with default parameters, utilizing probabilistic classification. This probabilistic approach assumes feature independence and is particularly effective for high-dimensional data with relatively low training complexity. To enhance the sensitivity of the model in identifying positive cases, a custom classification threshold of 0.4 was applied to the predicted probabilities, instead of the standard 0.5. This adjustment shifts the decision boundary to favor recall, ensuring a higher likelihood of correctly identifying stroke cases, which is critical in healthcare applications. The model's performance was evaluated using F1 score, sensitivity and specificity to provide a balanced view of its predictive capability.

Bayes Theorem:- Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. This is a minor extension of conditional probability. This can be viewed as a formula for evaluating a certain kind of inverse probability

Bayes' theorem is mathematically expressed by the following equation:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{4}$$

Where:

- $P(A|B)$ is the posterior probability of event A given that B has occurred.
- $P(B|A)$ is the likelihood of event B given that A has occurred.
- $P(A)$ is the prior probability of event A.
- $P(B)$ is the marginal probability of event B.

3.7.4. Logistic regression

Logistic regression is a foundational algorithm in machine learning, specifically suited for binary classification tasks where the objective is to predict one of two possible outcomes. Unlike linear regression, which forecasts continuous values, logistic regression estimates the probability that an input belongs to a particular class. It achieves this by applying the logistic (or sigmoid) function to a linear combination of the input features, yielding a probability score between 0 and 1. This probability can be converted into a binary class label using an appropriate threshold. Due to its simplicity, interpretability, and computational efficiency, logistic regression is widely used in applications such as spam detection, medical diagnostics, and credit risk assessment. Moreover, the model's coefficients offer valuable insights into the influence of individual features, enhancing understanding of the relationships within the dataset.

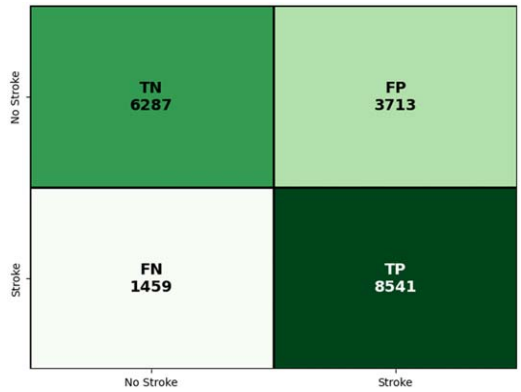

Figure 8. XGBoost confusion matrix on test data.

The Logistic Regression model was fine-tuned with specific hyperparameters to improve its predictive performance for HF detection. The 'class_weight='balanced' parameter was employed to address class imbalance by assigning weights inversely proportional to class frequencies, thus giving more importance to the minority class. A regularization strength of 'C=0.1' was selected to prevent overfitting by applying stronger L2 regularization, encouraging the model to generalize better on unseen data. The 'max_iter' was set to 1000 to ensure convergence during training, especially with imbalanced or complex datasets. Furthermore, the prediction threshold was manually adjusted to 0.4, deviating from the default 0.5 to improve the model's sensitivity-ensuring more TP cases of HF are identified, even at the cost of an increase in FPs. This tuning strategy reflects a deliberate focus on enhancing recall while maintaining overall model robustness.

3.8. Results and discussion

3.8.1. Results based on confusion matrix

The research started with a comprehensive review of features linked to mortality prediction across different classification algorithms. Four machine learning models were employed to predict the risk of HF in hospitalized patients. The dataset, obtained from Kaggle, includes patient records. With 11 variables from this dataset, it is possible to diagnose a patient's HF. The study employed four machine learning models: XGBoost, Random Forest, Naive Bayes and Logistic Regression. This section presents the results of performance levels of these models based on the confusion matrix. Typically, the model with the lowest FNs is considered as the best. A confusion matrix is a performance assessment tool commonly utilized in machine learning to illustrate the classification accuracy of a model. The matrix includes four key components: TP, TN, FP, and FN. We have evaluated every models performance using confusion matrix and the results are presented in this section.

The performance of the XGBoost model is evaluated using a confusion matrix, which provides insight into the model's classification accuracy. The confusion matrix, shown in figure 8, highlights the distribution of TPs, TNs, FPs and FNs.

The confusion matrix indicates that the model successfully classifies 6,287 non-stroke cases (TNs) and 8,541 stroke cases (TPs), while misclassifying 1,459 stroke cases as non-stroke (FNs) and 3,713 non-stroke cases as stroke (FPs). This outcome highlights the model's strong capability in detecting actual stroke cases, which is critical in scenarios where early diagnosis can significantly impact patient outcomes. The relatively low number of FNs demonstrates the model's effectiveness in minimizing missed diagnoses, which is essential for stroke prediction. Also, the reduction of FPs supports efficient resource allocation in clinical follow-up. Overall, the model exhibits a reliable balance between sensitivity and specificity, making it a valuable tool in medical decision-making processes. Subsequently, the performance of the Random Forest model is evaluated, with the corresponding confusion matrix presented in figure 9.

The confusion matrix demonstrates that the model accurately identifies 8,038 non-stroke cases (TNs) and 9,834 stroke cases (TPs), while only misclassifying 166 stroke cases as non-stroke (FNs) and 1,962 non-stroke cases as stroke (FPs). This performance reflects the model's exceptional ability to detect stroke cases with high sensitivity, which is crucial in medical applications where early and accurate identification can significantly influence treatment outcomes. The remarkably low number of FNs highlights its effectiveness in minimizing missed diagnoses, aligning well with the primary objective of stroke prediction. Furthermore, the relatively low FP rate ensures a balanced specificity, reducing unnecessary clinical interventions. Overall, this model demonstrates the most robust and reliable performance among all evaluated, making it highly suitable for

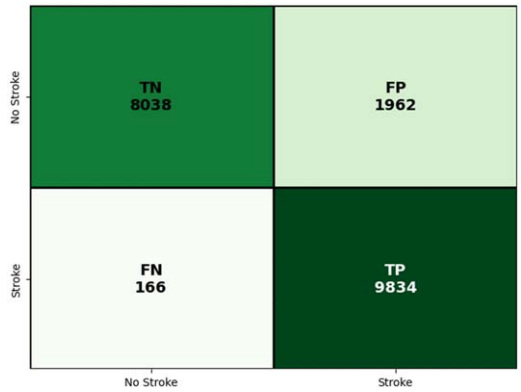

Figure 9. Random forest confusion matrix on test data.

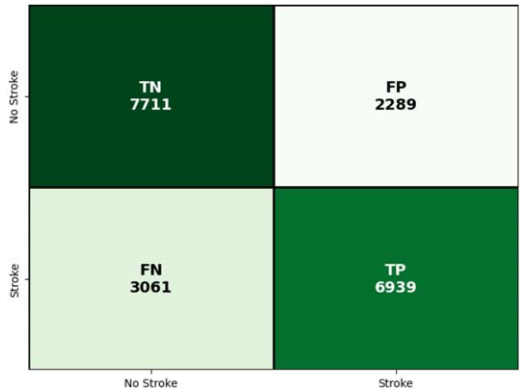

Figure 10. Naive bayes confusion matrix on test data.

deployment in real-world healthcare scenarios. Following this, the confusion matrix for the Naive Bayes model was generated and is presented in figure 10.

The confusion matrix indicates that the model accurately identifies 7,711 non-stroke cases(TNs) and 6,939 stroke cases as (TPs). It misclassifies 3,061 stroke cases as non-stroke (FNs) and 2,289 non-stroke cases as stroke (FPs). This distribution shows that the model is capable of correctly predicting a substantial number of both stroke and non-stroke cases, with higher precision in detecting non-stroke cases. However, the presence of 3,061 FNs suggests that a notable number of stroke cases are being missed, while the 2,289 FPs reflect instances where non-stroke cases were incorrectly flagged. Overall, the model demonstrates a balanced performance with moderate sensitivity and specificity. Lastly, the confusion matrix corresponding to the Logistic Regression model is presented in figure 11.

The confusion matrix shows that the model correctly classifies 6,369 non-stroke cases as (TNs) and 8,344 stroke cases as (TPs). It misclassifies 1,656 stroke cases as non-stroke (FNs) and 3,631 non-stroke cases as stroke (FPs). This indicates that the model is effective in detecting stroke cases, with a relatively high number of TPs. Although some non-stroke cases are incorrectly flagged and a portion of stroke cases are missed, the model maintains a strong ability to distinguish between the two classes, making it suitable for tasks where capturing stroke cases accurately is a priority.

The overall analysis of the confusion matrices reveals that the Random Forest model proves to be highly effective for stroke prediction. It correctly identifies 9,834 stroke cases (TPs) and 8,038 non-stroke cases (TNs), while misclassifying only 166 stroke cases as non-stroke (FNs) and 1,962 non-stroke cases as stroke (FPs). This minimal count of FNs is especially important in a medical context, as missing a stroke diagnosis can have critical consequences. The model's ability to accurately detect a high number of true stroke cases, along with a low rate of misclassification, highlights its robustness and reliability. These results make Random Forest a strong candidate for deployment in real-world clinical environments, particularly in scenarios where early and accurate stroke detection is essential.

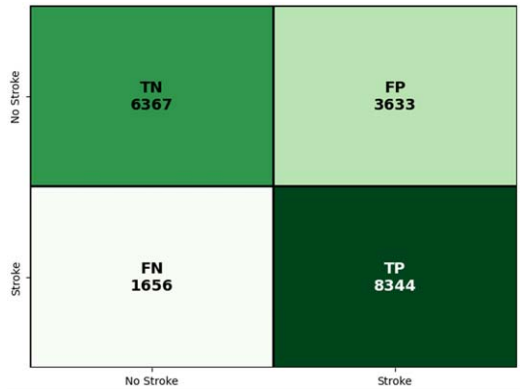

Figure 11. Logistic regression confusion matrix on test data.

Table 3. Comparison of F1 score, sensitivity and specificity of test data on all models.

Sr. No	Model	F1 Score	Sensitivity	Specificity
1	XGBoost	0.76	0.85	0.62
2	Random Forest	0.90	0.98	0.80
3	Naive Bayes	0.72	0.69	0.77
4	Logistic Regression	0.75	0.83	0.63

3.8.2. Results based on performance metrics

A comparison of the performance metrics across all models is in table 3:

Random Forest emerged as the top-performing model, achieving an impressive F1 score of 0.90, a high sensitivity of 0.98, and a specificity of 0.80, making it the most suitable choice for stroke prediction. While the model shows a moderately lower specificity, this is an acceptable trade-off given the critical importance of minimizing FNs. In the medical context, particularly for stroke prediction, failing to identify a TP can lead to severe consequences. By prioritizing sensitivity, the model effectively ensures that high-risk individuals are correctly flagged for further medical evaluation, even if it results in a higher number of FPs. Among the evaluated models, Random Forest demonstrated the best overall performance for stroke prediction. This indicates its strong ability to correctly identify stroke cases while maintaining a reasonable balance with specificity, making it the most reliable model for early detection and clinical decision-making. XGBoost also showed promising results offering a good trade-off between sensitivity and specificity. Logistic Regression and Naïve Bayes models exhibited moderate performance, with lower sensitivity and F1-scores compared to Random Forest, suggesting they are less effective in minimizing FNs. Overall, Random Forest stands out as the most suitable model given the critical importance of identifying as many true stroke cases as possible. The Performance comparison plot for the four different techniques is displayed in figure 12.

3.8.3. Experimentation

The experimentation section of this research explores the predictive capability of the HF prediction model by testing it across five distinct cases, each representing different patient health profiles. These cases are designed to evaluate the model's performance in various scenarios, providing a thorough understanding of its strengths and limitations. In Case 1, the model is tested with data from a healthy individual, characterized by the absence of any risk factors associated with HF. This test serves as a baseline, assessing the model's ability to accurately identify low-risk profiles and predict a 'non-failure' outcome. It is essential to verify that the model does not incorrectly categorize healthy individuals as high-risk, ensuring its accuracy in such cases.

Case 2 involves testing the model with data from an individual exhibiting multiple risk factors typically linked to HF. This case is critical for evaluating the model's sensitivity to significant health indicators, allowing for the assessment of its ability to correctly identify individuals at high risk of HF and predict a 'failure' outcome.

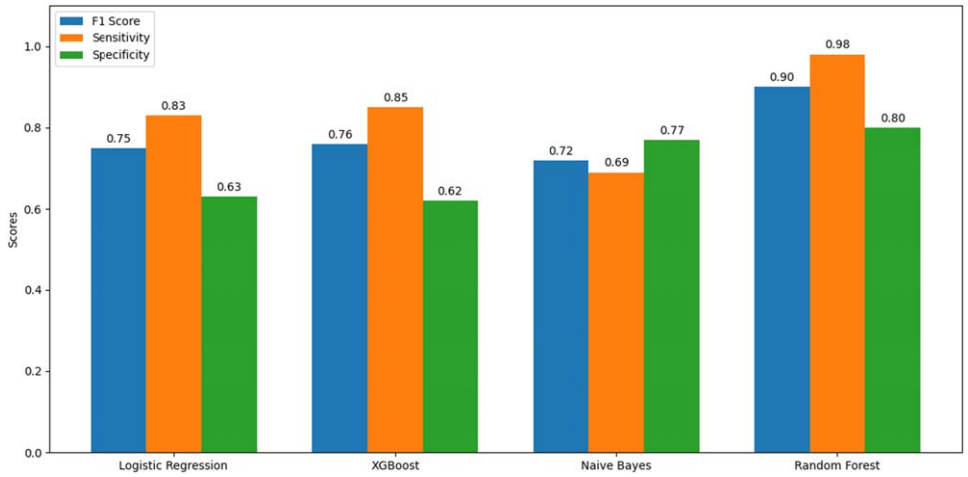

Figure 12. Performance Comparison of various models on Test data.

Finally, Cases 3–5 involve the use of random patient data, which includes a mix of varying health conditions, risk factors, and demographic variations. These cases provide insight into the model's generalization ability, testing its performance on diverse patient profiles. This helps ensure that the model remains effective across a wide range of scenarios, not just specific, extreme cases.

By employing this structured approach, the experimentation section provides a comprehensive evaluation of the model's predictive reliability, verifying its ability to distinguish between healthy and high-risk individuals while maintaining reliable performance in a variety of patient conditions.

In assessing the likelihood of an individual experiencing a HF, the following attributes are utilized:

- **Heart Disease** An indicator variable for heart disease, where 1 signifies a positive diagnosis and 0 indicates no heart disease.
- **Smoking** Binary indicator of smoking status, where 1 denotes that the individual smokes and 0 denotes that the individual does not smoke.
- **Alcohol Drinking** Binary indicator of alcohol consumption, where 1 denotes that the individual consumes alcohol and 0 denotes that the individual does not consume alcohol.
- **Difficulty Walking** Binary indicator of difficulty in walking, where 1 signifies the presence of difficulty and 0 signifies no difficulty.
- **Diabetic** A binary variable representing diabetes status: 1 indicates the individual is diabetic, while 0 indicates they are not.
- **Physical Activity** Binary indicator of physical activity engagement, where 1 indicates active participation and 0 indicates no participation.
- **Asthma** Binary indicator of asthma status, where 1 denotes the presence of asthma and 0 denotes its absence.
- **Kidney Disease** An indicator flag for kidney disease, where 1 represents a positive case and 0 represents no disease.

Additional attributes include:

- **Age Category** Categorical variable denoting the age range of the individual, encoded as follows:
  - 0: 18–24 years
  - 1: 25–29 years
  - 2: 30–34 years
  - 3: 35–39 years
  - 4: 40–44 years

Table 4. (a) Healthy person data, (b) Model predictions.

(a)		(b)		
Features	Value		Models	Predictions
HeartDisease	0		XGBoost	0
BMI	16.60		Random Forest	0
Smoking	0	ML Models →	Naive Bayes	0
AlcoholDrinking	0		Logistic Regression	0
DiffWalking	0			
AgeCategory	4			
Diabetic	0			
PhysicalActivity	1			
SleepTime	8			
Asthma	0			
KidneyDisease	0			

- 5: 45–49 years
- 6: 50–54 years
- 7: 55–59 years
- 8: 60–64 years
- 9: 65–69 years
- 10: 70–74 years
- 11: 75–79 years
- 12: 80 years or older

- **BMI** A feature expressing the Body Mass Index of an individual in continuous numeric form.
- **Sleeptime** A continuous variable representing the number of hours of sleep the individual receives.

These attributes should be meticulously recorded and formatted to ensure accurate input and analysis within the predictive model. To thoroughly evaluate the model's performance, the following cases were designed to test its ability to accurately predict HF outcomes across different patient profiles.

3.8.4. Case 1

In the first case, We are providing the model with data corresponding to a healthy individual. For this context, a healthy body mass index (BMI) is defined as being within the range of 18.5 to 25 kg/m^2. A BMI below 18.5 kg/m^2 is categorized as underweight, while a BMI between 25 kg/m^2 and 29.9 kg/m^2 is classified as overweight. A BMI of 30 kg/m^2 or above is classified as obese. To represent a typical healthy BMI, we use the mean value of 21.5 kg/m^2. In addition, the average sleep duration for a healthy person is approximately 8 hours per night. Individuals in good health typically avoid smoking and alcohol consumption and are not affected by persistent conditions like heart disease, asthma, or kidney disease. The prediction phase is now underway, utilizing the trained models.

Table 4(a) represents the attributes of a healthy person while table 4(b) represents the predictions made by the model on the healthy person data. In the conducted analysis, all predictive models uniformly classified the risk of HF as 'No' for the given data. We will focus on the predictions made by our best-performing model to determine whether a patient is at risk of experiencing HF.

3.8.5. Case 2

We are inputting data corresponding to an individual with health concerns into the model. For this context, an unhealthy BMI is defined as either below 18.5 kg/m^2, which is categorized as underweight, or above 25 kg/m^2, where a BMI between 25 kg/m^2 and 29.9 kg/m^2 is considered overweight, and a BMI of 30 kg/m^2 or above is classified as obese. Additionally, unhealthy individuals may have an average sleep duration that deviates from the typical 8 hours per night. Their lifestyle may include habits like smoking and alcohol consumption, and they could be affected by persistent health conditions like heart disease, asthma, or kidney disease.

Table 5. (a) Unhealthy person data, (b) Model predictions.

(a)		(b)		
Features	Value		Models	Predictions
HeartDisease	0		XGBoost	1
BMI	22.60		Random Forest	1
Smoking	1	ML Models →	Naive Bayes	1
AlcoholDrinking	0		Logistic Regression	1
DiffWalking	1			
AgeCategory	8			
Diabetic	0			
PhysicalActivity	1			
SleepTime	4			
Asthma	0			
KidneyDisease	0			

Table 6. (a) Random Patient data, (b) Model predictions.

(a)		(b)		
Features	Value		Models	Predictions
HeartDisease	0		XGBoost	0
BMI	25		Random Forest	0
Smoking	1	ML Models →	Naive Bayes	0
AlcoholDrinking	1		Logistic Regression	0
DiffWalking	0			
AgeCategory	4			
Diabetic	1			
PhysicalActivity	1			
SleepTime	7			
Asthma	0			
KidneyDisease	0			

Table 5(a) presents the data for an individual assessed as having a high risk of HF. Table 5(b) provides the corresponding predictions made by the predictive models based on this data. The analysis revealed that all predictive models consistently identified a high risk of HF for the patient. This uniform prediction indicates a strong agreement across the models regarding the elevated risk associated with the patient's profile. The Random Forest model's prediction indicates that the patient is at risk of developing HF.

3.8.6. Case 3

In this instance, we are inputting data from a randomly selected patient into the predictive model to assess the likelihood of a HF. This assessment seeks to identify whether the patient is at risk of HF based on the given attributes.

Table 6(a) displays the data for a randomly selected patient, which has been used as input for the predictive model. Table 6(b) illustrates the predictions generated by the model based on the data provided in table 6(a). The models consistently predicted a 'No' for the risk of a HF. This result indicates that the model assessed the patient as not being at high risk for a HF, based on the provided data attributes. Based on the prediction generated by Random Forest model we have determined that the patient is not at risk of having HF.

3.8.7. Case 4

In this case, data from a randomly chosen patient is being fed into the predictive model to evaluate the probability of a HF. The purpose of this assessment is to ascertain whether the patient is at risk of a HF based on the given attributes.

Table 7(a) displays the data for a randomly selected patient, which has been used as input for the predictive model. Table 7(b) illustrates the predictions generated by the model based on the data provided in table 7(a). In the evaluation, two models–XGBoost and Logistic Regression–consistently predicted a risk of HF for the patient. Conversely, the Naive Bayes model and Random Forest model assessed the patient as having no risk of HF. Based

Table 7. (a) Random Patient data, (b) Model predictions.

(a)		(b)		
Features	Value		Models	Predictions
HeartDisease	0		XGBoost	1
BMI	20		Random Forest	0
Smoking	0	ML Models →	Naive Bayes	0
AlcoholDrinking	0		Logistic Regression	1
DiffWalking	0			
AgeCategory	12			
Diabetic	1			
PhysicalActivity	1			
SleepTime	8			
Asthma	0			
KidneyDisease	0			

Table 8. (a) Random Patient data, (b) Model predictions.

(a)		(b)		
Features	Value		Models	Predictions
HeartDisease	1		XGBoost	1
BMI	37		Random Forest	1
Smoking	0	ML Models →	Naive Bayes	1
AlcoholDrinking	0		Logistic Regression	1
DiffWalking	1			
AgeCategory	5			
Diabetic	1			
PhysicalActivity	0			
SleepTime	15			
Asthma	1			
KidneyDisease	0			

on the predictions generated by the Random Forest model, we have determined that the patient is at no risk for a HF.

3.8.8. Case 5

In this scenario, data from a randomly selected patient is entered into the predictive model to estimate the likelihood of a HF. The purpose of this evaluation is to assess whether the patient is at risk of HF based on the given attributes.

Table 8(a) displays the data for a randomly selected patient, which has been used as input for the predictive model. Table 8(b) illustrates the predictions generated by the model based on the data provided in table 8(a). In the evaluation, every models consistently predicted a risk of HF for the patient. Based on the predictions generated by the Random Forest model, we have determined that the patient have risk for a HF.

3.9. Conclusion

The rise in CVD cases necessitates advanced systems for early detection and accurate prognosis. This study investigates how well machine learning algorithms can predict mortality risk in HF patients. Specifically, it assesses the performance of Random Forest-a robust classification technique-against other widely-used methods. The findings indicate that Random Forest achieves superior performance in mortality prediction, demonstrating its potential utility for clinicians in assessing patient risk and guiding treatment decisions. This research advances the field by showcasing how machine learning can enhance diagnostic results and offering key insights for better patient care in cardiology. There are multiple approaches available to enhance the effectiveness of a machine learning model. Firstly, improving data quality through meticulous cleaning, normalization, and feature engineering is essential, as these steps ensure that the data accurately reflects the underlying patterns. Increasing the size of the training dataset can also significantly boost model performance, as a larger dataset generally provides more comprehensive information, leading to better predictions. Additionally, selecting only the most relevant features is essential, as reducing dimensionality helps eliminate noise and

irrelevant information, thereby enhancing the model's performance. Adopting these approaches can bring about substantial improvements in both model precision and predictive performance. One of the limitations of this study is the increased number of FPs and FNs. This can be addressed by incorporating additional data from the CDC, which may enhance the model's performance by improving its ability to accurately classify TPs and TNs, while reducing both FPs and FNs. As part of future work, We plan to integrate advanced deep learning architectures, such as recurrent neural networks and transformers, to effectively capture temporal dynamics in patient data. We also aim to incorporate real-time data from wearable devices to enable continuous health monitoring and improve early detection of HF. Additionally, We intend to deploy the work as a web-based application to enhance accessibility and facilitate its application in real-world healthcare settings.

Acknowledgments

The researcher extends sincere gratitude to Phuket Rajabhat University for their support.

Data availability statement

The data that support the findings of this study are openly available at the following [33] URL: http://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data.

ORCID iDs

Vishnumon S https://orcid.org/0009-0004-4402-6664
Sayooj Aby Jose https://orcid.org/0000-0003-4437-1623
Anuwat Jirawattanapanit https://orcid.org/0000-0002-6319-0214

References

[1] *National Library of Medicine. Heart Failure Epidemiology and Outcome Statistics* (PMID: 37797885) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10864030/
[2] Shams P, Malik A and Chhabra L 2025 *Heart Failure (Congestive Heart Failure) (Treasure Island (FL): StatPearls Publishing) (PMID: 28613623)* https://pubmed.ncbi.nlm.nih.gov/28613623/
[3] *Global Public Health Burden Of Heart Failure. (PMCID: PMC5494150)* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5494150/
[4] Salim S *et al* 2020 Heart disease and stroke statistics-2020 update: a report from the american heart association *J Am Heart Assoc* **141** 139–596
[5] (American Heart Association) 2017 *Heart Failure Projected To Increase Dramatically, According To New Statistics* https://www.heart.org/en/news/2018/07/19/heart-failure-projected-to-increase-dramatically-according-to-new-statistics
[6] *Heart Failure Epidemiology and Outcome Statistics: A Report of the Heart Failure Society of America (PMCID: PMC10864030)* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10864030/
[7] Dariush M *et al* 2016 Heart disease and stroke statistics-2016 update: a report from the american heart association *J Am Heart Assoc* **133** 38–360
[8] Hoque R, Billah M, Debnath A, Hossain S S and Sharif N B 2024 Heart disease prediction using SVM *Int. J. Sci. Res. Arch* **11** 412–20
[9] Duraisamy B, Sunku R, Selvaraj K, Pilla V V R and Sanikala M 2023 Heart disease prediction using support vector machine multidiscip *Sci. J.* **6** 1–6
[10] Victor O A, Chen Y and Ding X 2024 Non-invasive heart failure evaluation using machine learning algorithms *J. Sens.* **24** 2248
[11] Qadri A M, Raza A, Munir K and Almutairi M S 2023 *Effective Feature Engineering Technique For Heart Disease Prediction With Machine Learning IEEE Access* **11** 56214–24
[12] Mamun M M, Farjana A, Mamun M A, Ahammed M S and Rahman M M 2022 Heart failure survival prediction using machine learning algorithm: am i safe from heart failure? *IEEE World AI IoT Congress (AIIoT)* 194–200
[13] Pushpavathi T P, Kumari S and Kubra N K 2021 Heart failure prediction by feature ranking analysis in machine learning **2021** *6th International Conference on Inventive Computation Technologies (ICICT)* 915–23
[14] Arunaggiri Pandian K, Sai Kumar T S, Dhandare S P and Aara S T 2021 Development and deployment of a machine learning model for automatic heart failure prediction *Asian Conference on Innovation in Technology (ASIANCON)* **1** 1–6
[15] Zhang Y, Wang X, Liu Q and Li M 2024 Large language model-informed ecg dual attention network for heart failure risk prediction *IEEE TBD* **11** 1–13
[16] Sonia J, Masoud A, Ehsan R and Azadeh-Fard N 2024 A Machine learning model to predict heart failure readmission: toward optimal feature set front *Artif. Intell.* **7**
[17] Choi E, Lee J, Kim Y, Han S and Park J 2024 Artificial Intelligence–based electrocardiographic biomarker for outcome prediction in patients with acute heart failure: prospective cohort study *J. Med. Internet Res* **26** 52139
[18] Jili L, Siru L, Yundi H, Lingfeng Z, Yujia M and Jialin L 2022 Predicting mortality in intensive care unit patients with heartfailure using an interpretable machine learning model: retrospective cohort study *J. Med. Internet Res.* **24**
[19] Jing-xian W, Chang-ping L, Zhuang C, Yan L, Yu-hang W, Zhou Y, Yin L and Jing G 2025 Machine learning algorithms to predict heart failure with preserved ejection fraction among patients with premature myocardial infarction front *Cardiovasc. Med.* **12**
[20] Sona A, Leontios A, Ahsan K, Cesare S, Stergios S, Petros A, Ioannis D, Konstantinos G and Konstantinos T 2024 Prediction of heart failure patients with distinct left ventricular ejection fraction levels using circadian ecg features and machine learning *PLoS One* **19**

[21] Mohammed Khalid H, Sharayar W and Adamu A 2024 Examining mortality risk prediction using machine learning in heart failure patients *IJPCC* **11** 81–7
[22] Srinivas K and Vempathy P 2024 Comparative analysis of heart failure prediction using machine learning models *IJ-ICT* **13** 297
[23] Hosea I and Mulapnen K 2025 Death events from heart failure prediction using machine learning approach *IJDST* **11** 1–10
[24] Alexander A P X, Simon L C, Morten H J and Ole H 2025 Prediction of 14-day hospitalization risk in chronic heart failure patients, using interpretable machine learning methods *Health Technol* **15**
[25] *Centre for Disease Control and Prevention (CDC)* https://www.cdc.gov/index.html
[26] Kamil P 2022 *Indicators of Heart Disease*UPDATE) https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/code
[27] Paula A and Paivi P 2007 Sleep deprivation: impact on cognitive performance *Neuropsychiatr Dis Treat* **3** 553–67 (PMID: 19300585)
[28] Liu Y, Wheaton A G, Chapman D P, Cunningham T J, Lu H and Croft J B 2014 Prevalence of healthy sleep duration among adults–united states *MMWR* **65** 137–41
[29] David L *et al* 2010 *How Tobacco Smoke Causes Disease: The Biology And Behavioral Basis For Smoking-Attributable Disease: A Report Of the Surgeon General* (Centers for Disease Control and Prevention)
[30] *Smoking and Cardio Vascular Disease. Johns Hopkins Medicine* https://www.hopkinsmedicine.org/health/conditions-and-diseases/smoking-and-cardiovascular-disease
[31] Shalev-Shwartz S and Ben-David S 2014 *Understanding Machine Learning: From Theory To Algorithms* (Cambridge University Press) (https://doi.org/10.1017/CBO9781107298019)
[32] Sayed A, Raihan A, Hasin R, Setu C, Shariful I and Touhid B 2021 Machine learning to reveal an astute risk predictive framework for gynecologic cancer and its impact on women psychology: bangladeshi perspective *BMC Bioinform* **22** 213
[33] 2022 Indicators of Heart Disease (2022 UPDATE) *Kaggle* www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data