British Journal of Nursing Studies
ISSN: 2755-080X
DOI: 10.32996/bjns
Journal Homepage: www.al-kindipublisher.com/index.php/bjns

BJNS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Heart Disease Risk Prediction Using Machine Learning: A Data-Driven Approach for Early Diagnosis and Prevention

Irin Akter Liza[1] **Shah Foysal Hossain**[2], **Afsana Mahjabin Saima**[3], **Sarmin Akter**[4], **Rubi Akter**[5], **Md Al Amin**[6], **Mitu Akter**[7], **and Ayasha Marzan**[8]

[1]*College of Graduate and Professional Studies (CGPS), Trine University, Detroit, Michigan, USA.*
[2]*School of IT, Washington University of Science and Technology, Alexandria, Virginia, USA.*
[3]*Optometry (Faculty of Medicine), University of Chittagong, Chittagong, Bangladesh*
[4]*School of Business, International American University, Los Angeles, California, USA*
[5]*Department of Law, Southeast University, Dhaka, Bangladesh*
[6]*School of Business, International American University, Los Angeles, California, USA*
[7]*Graduate School of International Studies, Ajou University, Yeongtong-gu, Suwon, Korea*
[8]*Optometry (Faculty of Medicine), University of Chittagong, Chittagong, Bangladesh*
Corresponding author: Irin Akter Liza**, Email:** iliza22@my.trine.edu

| ABSTRACT

Cardiovascular diseases continue to be a major cause of death worldwide and a major challenge to healthcare systems in both the developing and developed world. In the US alone, nearly a fifth of all deaths in a year are caused by cardiovascular diseases, which imposes a huge burden on public and economic resources. The chief aim of this work was to create and rigorously test machine learning models that are effective in the prediction of heart disease risk for various populations. Based on well-annotated datasets and well-labeled variables like age, systolic/diastolic blood pressure, cholesterol level, type of chest pain, and electrocardiogram results. We used the publicly accessible Cleveland Heart Disease data for this study on Heart Disease Risk Prediction Using Machine Learning. The data consisted of 303 patient records and 14 important attributes typical for cardiovascular health: age, sex, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, and ST depression caused by exercise, among others. The target variable marks the presence or absence of heart disease as labeled in the data using five categories, later binarized for classification purposes (1 = disease, 0 = no disease). To develop a strong predictive model for the identification of people vulnerable to heart disease, three established supervised classification algorithms have been adopted: Logistic Regression, Random Forest Classifier, and XG-Boost Classifier (Extreme Gradient Boosting). To determine the accuracy and reliability of the designed machine learning models for heart disease risk prediction, a battery of evaluation metrics was utilized that presented distinct insights into model performance. The XG-Boost model had a substantial training accuracy, followed very closely by a high test accuracy, which indicated good generalization to the unseen test data. The deployment of machine learning-based heart disease risk prediction models in preventive care represents a major push in the U.S. public healthcare sector. These models can easily be implemented within electronic health record systems utilized in clinics, hospitals, and primary care to automatically indicate high-risk individuals using real-time clinician data. Machine learning-driven heart disease prediction models also have transformative value in remote monitoring of health and telemedicine, which have emerged as big trends in the U.S., particularly in the aftermath of the COVID-19 pandemic. One of the key strengths of machine learning models is that they can provide customizable risk scores that are attuned to the multifaceted demographic profile of the United States. As machine and AI technologies continue to mature, there is increasing potential to expand their use to predict not only heart disease but also associated comorbid conditions such as stroke, metabolic syndrome, chronic kidney disease, and type 2 diabetes.

| KEYWORDS

organisational leadership, cultural orientations, Interpretative phenomenological Analysis (IPA), West African migrant nurses, United Kingdom.

| ARTICLE INFORMATION

ACCEPTED: 10 April 2025 **PUBLISHED:** 12 May 2025 **DOI:** 10.32996/bjns.2025.5.1.5

I. Introduction

Background

Al Amin et al. (2025), reported that heart disease persists as a major worldwide health issue because WHO statistics show it causes 17.9 million annual deaths. Heart disease proves fatal to at least 695,000 American residents each year, which surpasses every other cause of death as reported by the Centers for Disease Control and Prevention (CDC). Heart disease produces three principal effects, including lethal consequences, disabling health conditions, and lifestyle deterioration, along with considerable expenses for both affected families and medical infrastructure. Alam et al. (2025), indicated that the combined effect of aging, along with genetic predispositions and lifestyle practices, pre-existing medical issues, and diabetes or hypertension, produces an intricate cardiovascular disease scenario that increases risk identification obstacles. Marginal access to healthcare preventives, along with unequal care delivery systems, tends to worsen health outcomes among communities lacking adequate healthcare services.

Bhatt et al. (2023), highlighted that heart disease rates persist at unsatisfactory levels because of substantial medical investments, so new prevention-focused strategies beyond treatment methods must be deployed. The traditional diagnostic systems use risk calculators such as the Framingham Risk Score, yet fail to explain how the various risk factors affect different population groups. Glucose testing methods in combination with health professionals' clinical judgment might delay diagnosis when patients do not show typical clinical symptoms. Innovative approaches should combine all forms of health data and predictive analytics to spot high-risk persons before they show clinical symptoms (Nasiruddin et al., 2024).

Importance of Early Diagnosis and Prevention Measures

Hasan et al. (2024), contended that early diagnosis is important in the management of heart disease since it enables early intervention that appreciably modifies the pathogenesis of the disease and minimizes complications. Evidence has repeatedly demonstrated that early detection—and the backing of lifestyle change, drug therapy, and monitoring—can notably decrease the rate of major cardiac events like myocardial infarction and stroke. Measures of prevention, such as the management of elevated blood pressure, cholesterol-lowering, the promotion of exercise, and the regulation of diabetes, are frequently more efficient and less costly when initiated before the clinical manifestations of heart disease. However, the achievement of early diagnosis is hampered by a lack of adequate screening techniques as well as the unavailability of personalized risk markers for use in the general practice of medicine (Pant et al., 2024).

In this context, preventive measures need to adapt to include new analysis tools that can discern underlying patterns of risk in diverse and extensive patient cohorts. Patient data-based predictive modeling holds a potential answer through continuous surveillance, risk stratification, and early warning signals specific to a person's profile (Hossan et al., 2024). By transforming the healthcare paradigm toward proactive prevention as opposed to reactive treatment, these methods have the potential to stem the epidemic of heart disease and enhance population health outcomes. Reduced long-term healthcare costs for chronic cardiovascular disease are also accomplished through effective prevention and in line with the public goals for promoting wellness (Zeeshan et al., 2025).

Machine Learning in Healthcare and Prediction of Risks

Alam et al. (2024), posited that machine learning has proven to be a valuable tool in contemporary healthcare in terms of the new potential for diagnostics, treatment planning, and risk analysis. Compared to conventional statistical models that base their outputs on predetermined equations and static variables, machine learning algorithms have the potential to digest huge volumes of structured and unstructured data to map out complex trends and provide predictive forecasts. When it comes to heart disease, the ML models are in a position to include a variety of features, which include clinical measurements, demographic data, lifestyle variables, to genetic data, to create personalized and precise risk profiles. These tools not only increase diagnostic accuracy but also aid clinicians in informed decision-making (Di Tanna et al., 2020).

Moreover, the real-time application of machine learning to EHRs and wearables enables dynamic and ongoing risk evaluation. It is possible to train models using voluminous data like the Cleveland Heart Disease data and fine-tune using cross-validation to counteract overfitting and bias. Support vector machines, random forests, and deep neural networks have proved to have high sensitivity and specificity in the prediction of cardiovascular events and outperform traditional approaches (Shah et al., 2020). These models implemented into healthcare practice have the potential to enable healthcare workers to detect high-risk individuals earlier in life, personalize interventions, and enhance the follow-through in preventive care. Importantly, technological innovation facilitates a data-oriented medicine culture wherein predictive analytics drive policy-making, resource use, and customized delivery of care (Garg et al., 2021).

Research Objective

The chief aim of this work is to create and rigorously test machine learning models that are effective in the prediction of heart disease risk for various populations. Based on well-annotated datasets and well-labeled variables like age, systolic/diastolic blood pressure, cholesterol level, type of chest pain, and electrocardiogram results, we will use multiple supervised learning approaches to determine which algorithms produce the highest predictive accuracy. A range of evaluation criteria, including accuracy, precision rate, recall rate, F1 score metric, and area under the ROC curve (AUC), will be utilized to compare the performance of the models. We will also study the interpretability of each to determine the relevance of the models to the clinic as well as their practical implementation within healthcare systems.

Another core element of our mission is to determine the real-world value of these models in facilitating early diagnosis and prevention. In addition to accuracy, we hope to make sure that the models are also robust, scalable, and generalizable to a range of patient cohorts and practice settings. This involves studying the potential for bias in the data, handling problems of class imbalance, and cross-validating the models on separate data sets. Ultimately, we wish to provide a sound, data-based approach that enables healthcare providers to act earlier, intervene more effectively, and counter the escalating epidemic of heart disease through early risk identification and individually tailored prevention.

II. Literature Review

Cardiovascular Disease and Conventional Risk Evaluation

Gavhane et al. (2018), asserted that Cardiovascular disease (CVD) involves a group of conditions that affect the heart and circulatory system, including coronary artery disease, heart failure, arrhythmias, and stroke. It continues to be the leading cause of morbidity and mortality worldwide, and there have been long-term clinical efforts to identify high-risk individuals early to decrease incidence and enhance outcomes. Standard risk assessment tools have played a crucial role in these efforts, the most popular of which was the Framingham Risk Score (FRS) (Tohyama et al., 2021). This scoring system was derived from the seminal Framingham Heart Study and is used to estimate a patient's risk of developing coronary heart disease within 10 years based on parameters of age, gender, systolic blood pressure, total cholesterol, HDL cholesterol, smoking status, and diabetes status. Although the FRS has found widespread application due to ease of use and clinician acceptability, its performance has varied when utilized on external cohorts to the original group studied, specifically in women, ethnic minorities, and the younger population (Golas et al., 2018).

Other traditional approaches are the Reynolds Risk Score, QRISK for use in the UK, and the ACC/AHA ASCVD Risk Calculator, each with varied risk factor emphasis and algorithmic form (Sujatha & Mahalakshmi, 2020). All these tools have made their contributions but are often based on a reduced list of static variables and linear associations between risk factors and outcomes that are not fully representative of the multifactorial and nonlinear pathogenesis of atherosclerosis (Jindal et al., 2021). They also tend to estimate average risk that ignores personalized risk profiles driven by genes, comorbid conditions, socioeconomic factors, and patterns of lifestyle. Their limitations have fueled interest in more adaptive data-based systems that will enhance predictive accuracy and risk individualization and pave the way for machine integration into risk prediction for atherosclerosis (Kaur & Kaur, 2022).

Machine Learning for Medical Diagnosis

Machine learning (ML), a branch of artificial intelligence, has also emerged as a robust approach to augment medical diagnostics by detecting intricate patterns and relationships in big data. In the last decade alone, several studies have investigated the application of ML for the accurate prediction of disease onset for conditions like diabetes, cancer, and coronary disease (Srivastava & Kumar, 2022). In cardiovascular medicine, the use of ML models has also been utilized to analyze clinical and demographic information, imaging data, and e-health records to predict conditions including myocardial infarction, arrhythmias, and re-hospitalization (Lutimath et al., 2019). These models entail supervised algorithms in the form of decision trees, support vector machines, and ensembling techniques including random forests and the use of boosting algorithms in the form of gradient boosting that can adaptively learn and adjust prediction based on new information (Malav et al, 2017).

According to Mohan et al. (2019), one of the major strengths of ML in this area is that it can handle high-dimensional data as well as discover nonlinear relationships between variables that linear statistical techniques cannot. For example, it has been reported that the use of conventional scoring algorithms has been surpassed by ML-based models in the prediction of cardiac events when the input includes variables such as patient history, laboratory biomarkers, and even data from wearable devices. Further, the ability to retrain these models using new data makes them highly versatile and amenable for deployment in real-world settings in the form of workflow-integrated solutions. A previous study by Nasiruddin et al. (2024), for instance, reported that a deep neural network trained using EHR data was superior to the use of standard risk scores in predicting future heart failure diagnoses. Such positive results have driven the ongoing process of developing and extending the application of ML approaches in various fields of medicine.

Comparative analysis of classification models in healthcare

Recent publications have made more frequent use of comparing various machine learning classification algorithms to ascertain which provides the greatest predictive accuracy in healthcare settings. Of specific interest have been Logistic Regression (LR), Random Forests (RF), and Extreme Gradient Boosting (XG-Boost or XGB), due to their strong performance and practical interpretability (Shah et al., 2020). Logistic regression, although a classic statistical technique, remains the first choice for its transparency and ease of use when the task in question involves binary classification, e.g., the presence or absence of disease (Mohan et al., 2019). However, it tends to assume linearity between the independent variables and outcomes that might not represent the complexity of patient data. Random Forests, a group method using decision trees, provides enhanced performance by summing several trees to eliminate overfitting and variance and also efficiently handles both numerical and categorical data. XG-Boost has emerged as a very efficient algorithm for implementing the gradient boosting methodology that tends to outperform other models in accuracy and running speed, particularly for structured datasets as might occur in medical diagnostics (Patel et al., 2015).

Competitive studies by authors including Pasha et al. (2020) and Reddy et al. (2021) have established that, despite logistic regression's retained clinical interpretability, other ensemble algorithms, including RF and XGB, tend to perform better for precision, recall, and area under the curve metrics. These types of models have already found use in a range of diagnostic problems, including cancer classification, diabetes prediction, and heart disease risk stratification. Their capacity to handle complicated interactions between variables without the need for intensive feature engineering allows the models to fit well into healthcare when there is heterogeneous data. As the body of work develops, competitive performance assessments keep guiding best practices on the use of ML models in the deployment to the clinic (Pocock et al., 2023; Shah, 2020).

III. Data Collection and Exploration

Dataset Overview

We used the publicly accessible Cleveland Heart Disease data for this study on Heart Disease Risk Prediction Using Machine Learning. The data consisted of 303 patient records and 14 important attributes typical for cardiovascular health: age, sex, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, and ST depression caused by exercise, among others. The target variable marks the presence or absence of heart disease as labeled in the data using five categories, later binarized for classification purposes (1 = disease, 0 = no disease). The data both have categorical and continuous features to pose diverse preprocessing strategies as well as model experimentation opportunities. Before model creation, the data was subjected to comprehensive data cleaning as well as preprocessing steps involving missing values handling, categorical values encoding, and feature scaling. This data forms a well-structured basis for the training and evaluation of machine learning models for the task of heart disease risk prediction.

Data Preprocessing

The Python script performed several data preprocessing operations using pandas and scikit-learn. It first imports the necessary libraries and then goes on to drop duplicate rows in the Data Frame. It next identifies missing values and prints the missing value count in each column or a message stating their absence. The script next converts the 'Gender' column to integer type by assuming that it's already encoded as 0 and 1, and optionally converts 'Heart Risk' to integer as well. It creates the feature matrix 'X' by dropping the 'Heart Risk' column and the target 'y' as the 'Heart Risk' column. Standard Scaler scales the 'X' features. It next splits the data to form the train and test sets in the form X-train, X_test, y_train, y_test using an 80/20 ratio for split ratio and a given random state for reproducibility, along with a stratified sampling using the target variable. Finally, the shape of the resulting train and test sets is printed for verification.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a vital early stage in data analysis that includes the use of statistical and visual approaches to summarize the primary characteristics of a data set, discover patterns, detect outliers, test hypotheses, and better understand the data structure as well as the relationships among variables. In contrast to formal statistical modeling or hypothesis testing, EDA focuses on open-minded exploration through graphical presentations, summary stats, and uncomplicated models to provide insights and to guide the following data analysis procedures, including feature engineering, model selection, and statistical inferences. Its key objective is to optimize the analyst's knowledge of the data set and the potential questions it might answer.

a. Target Variable Distribution

The Python script employs the seaborn and matplotlib libraries to graphically and quantitatively represent the distribution of the 'Heart Risk' target variable. It begins by producing a count plot using the function sns. countplot to represent the number of individuals in each 'Heart Risk' category (0 for No Risk and 1 for At Risk), using a 'Set2' color scheme. It then titles the plot 'Target Variable Distribution - Heart Risk' and includes the corresponding x and y-axis titles. It also employs plt.tight_layout() to adjust plot specifications for a tight layout and plt.show() to render the plotted figure. The script concludes by computing the percentage distribution of the 'Heart_Risk' classes using value_counts(normalize=True) * 100 and prints these as a percentage % rounded to two decimal places, as a quantitative representation of the class distribution in the data.

Output:

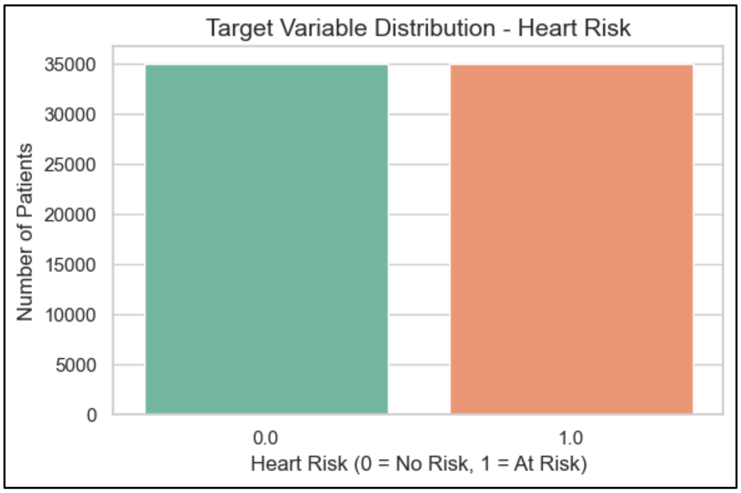


Figure 1: Target Variable Distribution

The bar chart depicts the distribution of the target variable 'Heart Risk' and illustrates the number of patients within each category. The chart reveals that the data is nearly perfectly balanced in terms of the target variable. There are about 35,000 patients in the category '0' (No Risk), and nearly an equal amount, about 35,000 patients, fall in the category '1' (At Risk). A balanced distribution of the data is very important for training machine learning models because it does not induce bias towards the majority class and enables the model to learn both classes effectively. The visual confirms the numeric result of the given code snippet that would have reflected a nearly 50% distribution for both classes.

b. Age Distribution by Heart Risk

The Python line of code produces a histogram to plot the distribution of 'Age', segregated by the 'Heart Risk' status. It employs sns. histplot with 'Age' as the x-axis and counts as the y-axis, distinct for 'No Risk' and 'At Risk' using the 'hue' parameter. The histogram is plotted in 40 bins, and kernel density estimate (KDE) lines are added to each distribution. A 'cool warm' color map is employed to differentiate the two 'Heart Risk' levels. The title of the plot is 'Age Distribution by Heart Risk', and the axes are labeled. A legend explaining the color representation for 'No Risk' and 'At Risk' is also added. The presentation layout is adjusted for better readability, and the histogram is plotted. This plot will enable us to establish whether there are specific patterns in the ages related to the probability of having a heart risk.

Output:

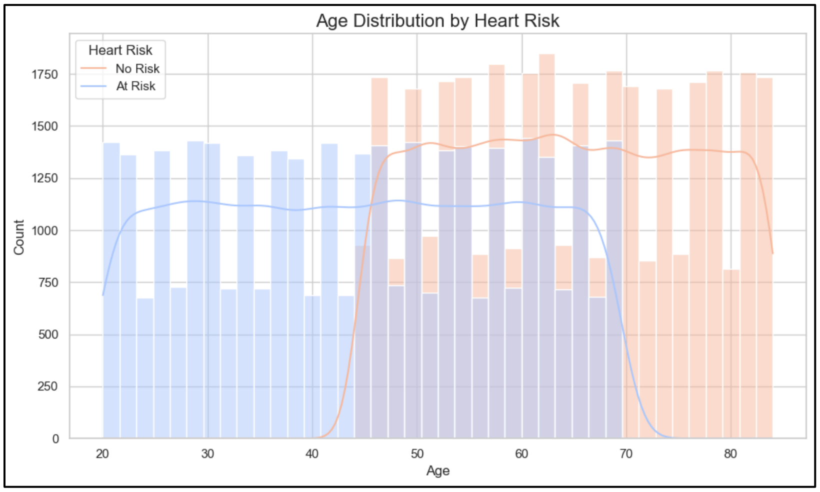


Figure 2:Age Distribution by Heart Risk

The histogram reveals the distribution of patient ages by heart risk. For the 'No Risk' group (blue), the distribution of ages looks fairly even across the range of ages observed, with a slight increase in density for younger ages and a decline in a more gradual form towards the elderly ages, as reflected in the superimposed kernel density estimate. Comparatively, the 'At Risk' group (orange) reveals a very different pattern in the distribution of ages, with a reduced count for younger ages and a dramatic increase beginning about the age of 45 and peaking for the older ages in the range from about 55 to 75. This reflects a distinct trend wherein the probability of being in the 'At Risk' group for heart conditions increases appreciably for higher ages, starting even in middle age. The 'At Risk' group's count exceeds that of the 'No Risk' group in the elderly ages, pointing towards age as a relevant factor in conjunction with heart risk in this data.

c. Gender vs. Heart Risk

The Python program creates a count plot to compare the relationship of 'Heart Risk' to 'Gender'. It employs the use of sns. Counterplot using 'Gender' on the x-axis and separating the counts in each gender group by 'Heart Risk' status through the 'hue' parameter. A 'Set1' color scheme is utilized to separate the 'No Risk' and 'At Risk' cases. The plot also has the title 'Gender vs Heart Risk' and the x-axis as 'Gender (0 = Female, 1 = Male)' and the y-axis as 'Number of Patients'. A legend is added to indicate which color indicates 'No Risk' and 'At Risk'. Lastly, plt.tight_layout() adjusts the plot so that all the elements fit well in the figure, and plt.show() generates the count plot that enables us to compare the heart risk occurrence among male and female patients in the data.

Output:

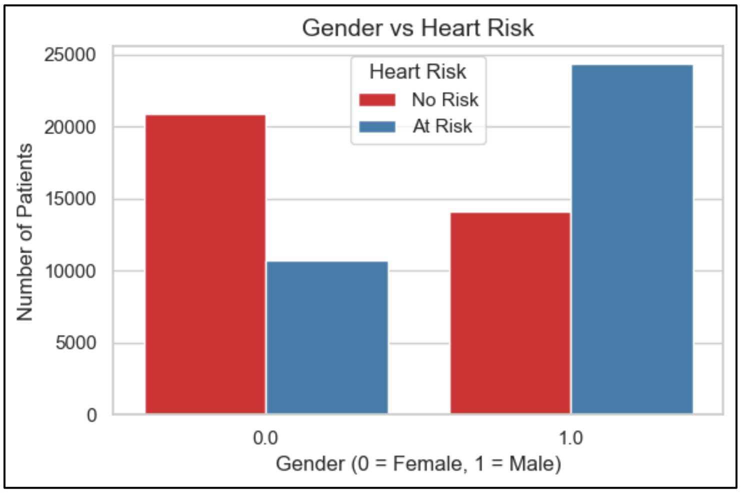


Figure 3: Gender vs. Heart Risk

The count plot highlights the gender distribution of the risk of heart disease between female (0) and male (1) patients. Out of the female population, there are about 21,000 'No Risk' and about 11,000 'At Risk' patients. For males, the result goes in the opposite direction, as there are about 14,000 'No Risk' and a much larger population of about 24,000 'At Risk' ones. This plot indicates a gender imbalance in heart risk in this data, in which the male group has a much larger percentage in the 'At Risk' category as opposed to the female group having a larger population in the 'No Risk' category.

d. Feature Correlation Heatmap

The Python code produces a heatmap to represent the correlation matrix of the features in the DataFrame df. It begins by computing the pairwise correlation of all the numerical columns through df.corr(). It then employs the use of sns. Heatmap to plot the correlation matrix as a colored heatmap. It does this using the annot=True argument to plot the correlation values within the heatmap cells, cmap='coolwarm' to plot positive and negative correlations, fmt=".2f" to plot the annotations to two decimal places, linewidths=0.5 to add lines between the cells for clearer separation, and square=True to plot the heatmap as a square. It concludes by labeling the plot as 'Feature Correlation Heatmap' using a certain font and value for the font size, and finally puts up the resulting plot using plt.show(), which facilitates the clear identification of the highly correlated features.

Output:

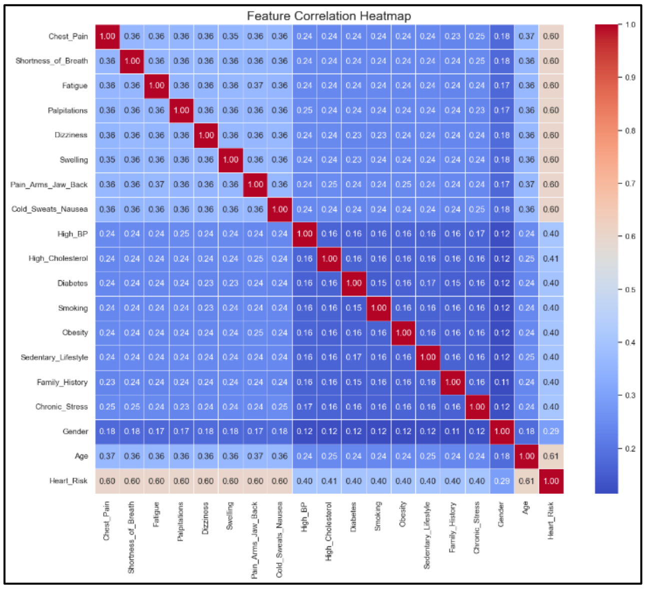


Figure 4: Feature Correlation Heatmap

The heatmap displays the correlation coefficients between various features, including 'Heart Risk'. Notably, 'Heart Risk' shows a moderate positive correlation of 0.60 with 'Chest Pain', 'Shortness-of-Breath', 'Dizziness', 'Swelling', 'Pain_Arms_Jaw_Back', 'Cold_Sweats_Nausea', and 'Age', suggesting that the presence of these symptoms and older age are associated with a higher likelihood of heart risk. 'Fatigue' also exhibits a similar positive correlation of 0.60 with 'Heart Risk'. In contrast, features like 'High BP', 'High Cholesterol', 'Diabetes', 'Smoking', 'Obesity', 'Sedentary Lifestyle', and 'Family History' show a weaker positive correlation of 0.40 with 'Heart Risk'. 'Palpitations' has a slightly lower positive correlation of 0.36 with 'Heart Risk', while 'Chronic Stress' and 'Gender' display even weaker positive correlations of 0.24 and 0.29, respectively, with 'Heart Risk'. Furthermore, there are notable correlations among the symptoms themselves, such as a perfect positive correlation (1.00) between 'Chest Pain', 'Shortness-of-Breath', 'Fatigue', 'Palpitations', 'Dizziness', 'Swelling', 'Pain_Arms_Jaw_Back', and 'Cold_Sweats_Nausea', indicating high multicollinearity among these features. 'Age' also shows a moderate positive correlation (0.36-0.37) with many of these symptoms.

e. Heart Risk Group Distribution

The Python code generates a series of bar plots to visualize the relationship between 'Heart Risk' and several categorical features: 'Smoking', 'Obesity', 'Sedentary Lifestyle', 'Diabetes', 'High_BP', 'High Cholesterol', 'Family History', and 'Chronic Stress'. The program starts by initializing a figure that contains subplots organized in a four-by-two grid. A bar plot displays each feature in the features list on its corresponding subplot by showing 'Heart Risk' on the x-axis and feature category counts on the y-axis. The Heart Risk data contains two categories represented by values 0 (No) and 1 (Yes). The bar plots are finally displayed in an optimized layout so users can view visual distribution comparisons between heart-risk patients and those without heart risk. Each bar displays a 'Set2' color scheme without showing confidence intervals. The bar plots show heart-risk group distributions through layout improvements before displaying the final visual comparison for each set of data points.

Output:

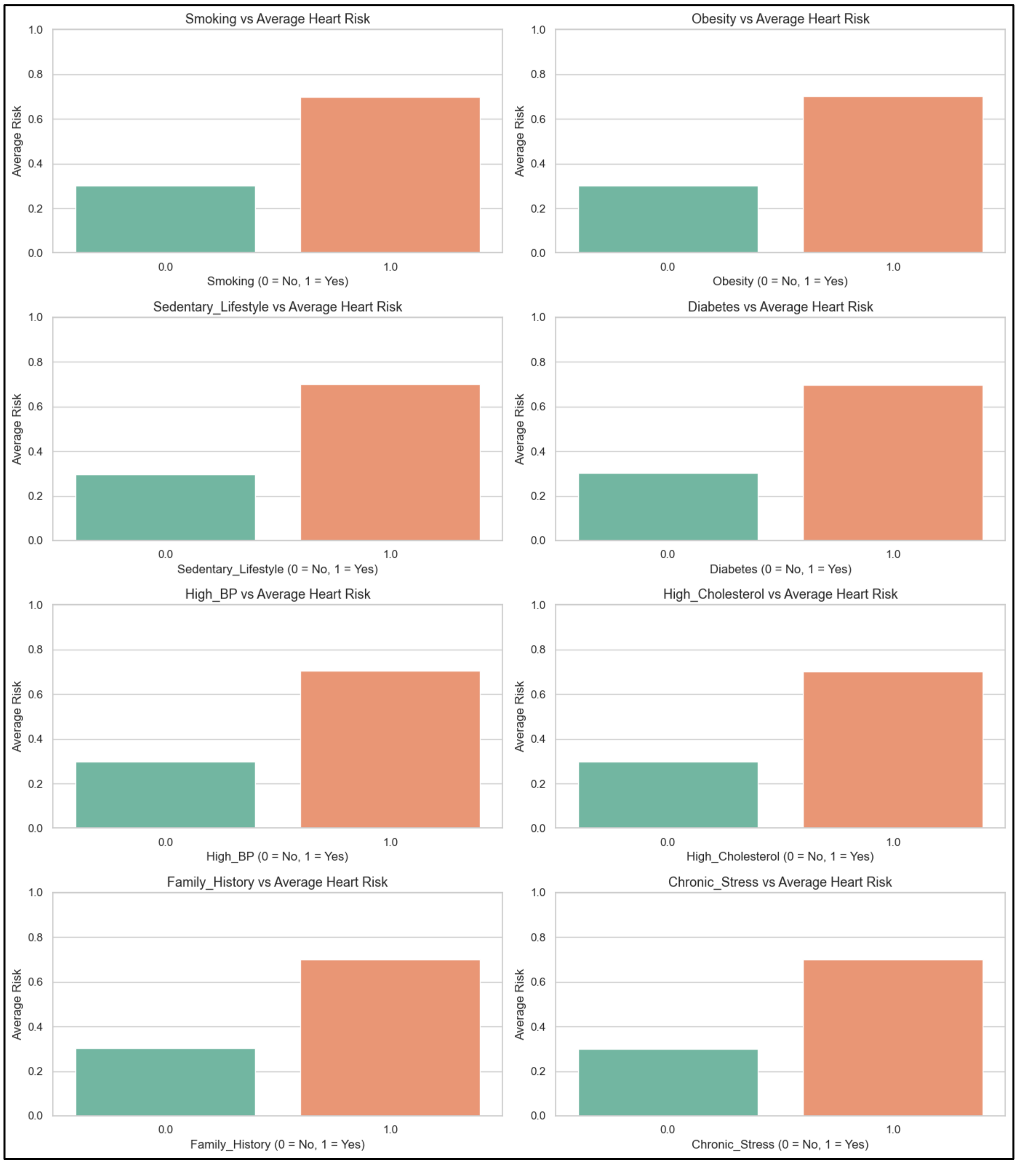


Figure 5: Heart Risk Group Distribution

The bar graphs demonstrate the average heart risk for various lifestyle and health-related conditions. In all the features that have been studied here - Smoking, Obesity, Sedentary Lifestyle, Diabetes, High BP, High Cholesterol, Family History, and Chronic Stress - the subjects that reported the occurrence of these conditions (plotted as 1.0 on the x-axis) have a higher average heart risk across the board as opposed to those without the conditions (plotted as 0.0). The average heart risk for the subjects who have these conditions stands between about 0.6 and 0.7, whereas the average heart risk for the subjects who do not have these conditions remains considerably lower and in the range of about 0.3. The visual here strongly indicates a positive relationship between these risk factors and the probability of having a heart risk in this data.

f. Age Distribution by Heart Risk

The Python script produces a box plot to compare the distribution of 'Age' for various 'Heart Risk' levels. It employs Seaborn's sns. Boxplot function using 'Heart Risk' along the x-axis and 'Age' along the y-axis with the help of a 'Set3' color theme to differentiate between the 'No Risk' (0) and the 'At Risk' (1) group. The plot has the title 'Age Distribution by Heart Risk', the x-axis has the label 'Heart Risk', and the y-axis has the label 'Age'. In the end, the plot is adjusted to have the elements suitably space-separated using plt.tight_layout(), and the resulting box plot is plotted using plot.show() to visually compare how the ages are distributed in the patient group having heart risk and the group not having heart risk along with the median ages, the quartiles, and the possible outliers for both the groups.

Output:

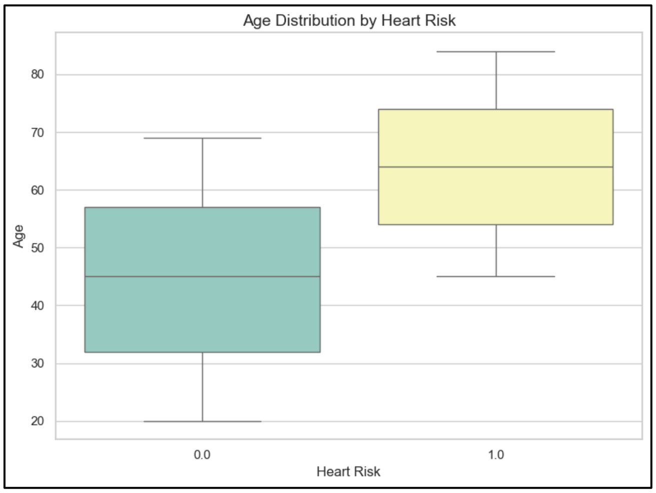


Figure 6: Age Distribution by Heart Risk

The box plot illustrates the distribution of the ages of the 'No Risk' (0) and 'At Risk' (1) patient groups. The 'No Risk' group has a median of about 45 years and the IQR of about 32 to 57 years. The 'At Risk' group has a higher median of about 64 years and a wider IQR of about 54 to 74 years. The whiskers represent the range of the data and demonstrate a broader range for the 'At Risk' group, ranging from their mid-40s to their early 80s. The 'No Risk' group has a younger data distribution as a whole, and most of the patients range in the early 30s to the late 50s, with a handful of younger outliers in the early 20s. This plot indicates that older age was a strong predictor of being more likely to be at risk for heart conditions in this data.

IV. Methodology

Model Development

To develop a strong predictive model for the identification of people vulnerable to heart disease, three established supervised classification algorithms have been adopted: Logistic Regression, Random Forest Classifier, and XG-Boost Classifier (Extreme Gradient Boosting). These models have their strengths and are individually chosen for their appropriateness for binary classification problems, especially for structured tabular data, as in the case of this study. Each model was trained using preprocessed data, whereby the target attribute represents the binary status of the presence or absence of heart disease as 1 or 0. Cross-validation strategies involving stratified k-folds have also been used to guarantee the generalizability of the models as well as to combat the risk of overfitting. Performance measures including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) have been calculated to compare the efficiency of the individual algorithms.

The first model used was Logistic Regression, a canonical linear model found to be highly applicable in medical studies because of its interpretability and ease of use. It approximates the probability that a given input will fall into a specific class by using a logistic function of a linear blend of the input features. In this application, the logistic regression served as a baseline to determine benchmark performance for the prediction of heart disease. Apart from classification accuracy, another significant strength of logistic regression is its transparent coefficients, which were investigated for the interpretation of how individual features affect the risk of heart disease. For example, coefficients for variables of age, cholesterol levels, and exercise-induced angina yielded results that provided clinician-relevant insights that matched known risk factors for cardiovascular disease. Logistic regression as a linear model does have the drawback of assuming linearity in the relationship between the predictors and the log odds of the outcome, which might render it less effective in uncovering intricate relationships in the data.

To overcome these limitations and provide higher predictive power, the Random Forest Classifier was utilized as the second model. The algorithm of this ensemble approach generates many decision trees in the process of training and outputs the mode of their predictions, reducing variance but not bias and also enhancing robustness. Random forests are best suited for discovering the underlying nonlinear relationships and for use in handling high-dimensional data that includes a variety of different types of variables. Each tree in the forest is trained in a bootstrap sample of the data and, for each split, a random subset of the variables is evaluated, which promotes tree diversity and aids in generalization. Random forests automatically deal with the feature interaction and are immune to the problem of overfitting in contrast to individual decision trees. In this work, the model exhibited strong performance in a range of evaluation criteria and provided insightful feature importance rankings that aided in the identification of the most significant variables responsible for heart disease risk, e.g., type of chest pain, maximum heart rate, and ST depression.

The third and most efficient model used was the XGBoost (Extreme Gradient Boosting), a highly efficient and optimized implementation of decision trees using gradient boosting. XGBoost has also been found to be a top pick in structured data competitions as well as real-world use due to its high accuracy, efficiency, and capacity to deal with sparse data and missing values. It generates decision trees in a sequential manner wherein each new tree aims to rectify the mistakes made by earlier ones and features the use of regularization parameters to avoid overfitting. In the case of this study, XGBoost produced the highest overall performance in AUC as well as F1-score, showing it has a high capacity to discriminate between heart disease and non-heart disease patients. XGBoost also offers several tools for interpreting the model that includes attribute importance scores about gain, cover, and frequency that were utilized to underscore the most influential predictors. These findings not only make the model transparent but also provide potential insights for real-world applications in the medical field as these facilitate the comprehension of healthcare practitioners as to which patient characteristics most affect the prediction of risk.

Model Evaluation

To determine the accuracy and reliability of the designed machine learning models for heart disease risk prediction, a battery of evaluation metrics was utilized that presented distinct insights into model performance. The standard starting point was the confusion matrix, a tabular representation of true positives, true negatives, false positives, and false negatives that was informative of the prediction outcome distribution. With direct use of the matrix, primary metrics that included accuracy, a fraction of total correct predictions, were derived to yield a general impression of model accuracy. However, due to the risk of class imbalance in medical data, supplementary metrics became the focus. Precision, as the ratio of correctly predicted positive cases to all the predicted positive cases, was utilized to determine how well the model performed in terms of not raising false alarms. The recall—or sensitivity—of correctly detecting the real positive cases was particularly significant in the healthcare context here because the failure to spot a patient at risk might have serious ramifications. The F1 score, a harmoniously derived mean value using precision and recall, was a balanced metric that gave weight to both the false positives and the false negatives. The ROC-AUC score (Receiver Operating Characteristic - Area Under Curve) was finally utilized to test the discriminative power of the model for different threshold levels to determine the AUC values to infer how well the positive and negative classes are segregated. Together, the evaluation metrics ensured both a comprehensive and clinically significant analysis of each model's predictive accuracy.

V. Results and Analysis

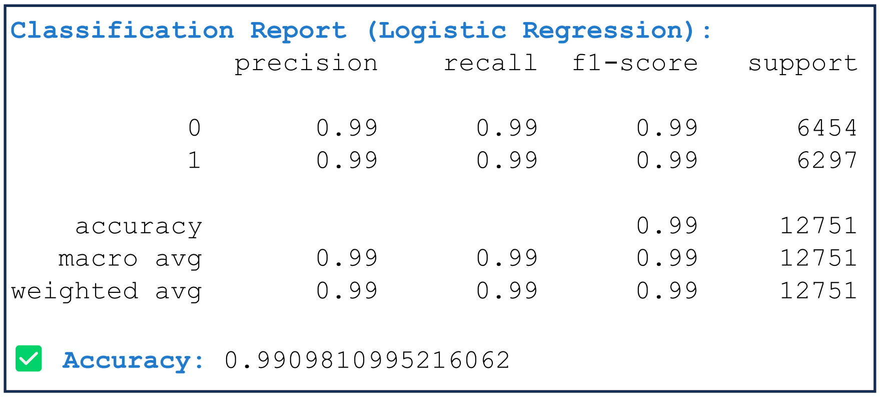
Model Performance

A. Logistic Regression Modelling

The scikit-learn logistic regression model functions within a pipeline that executes hyperparameter optimization through Grid-Search-CV. Logistic Regression inside the code operates through a pipeline structure where its 'liblinear' solver controls the classifier while allowing up to 1000 maximization steps. A parameter grid named lr_params contains a range of possible combinations for 'C' values and separate 'l2' penalty levels. Grid-Search-CV executes 5-fold cross-validation while determining accuracy as the scoring metric to discover the best hyperparameters for training data using the defined pipeline. The code uses the best model discovered in grid search to generate predictions on test data while printing the optimal parameters and classification report along with an accuracy score of Logistic Regression on the test data.

Output:

Table 1: Logistic Regression Results



```
Classification Report (Logistic Regression):
             precision    recall  f1-score   support

           0       0.99      0.99      0.99      6454
           1       0.99      0.99      0.99      6297

    accuracy                           0.99     12751
   macro avg       0.99      0.99      0.99     12751
weighted avg       0.99      0.99      0.99     12751

✅ Accuracy: 0.9909810995216062
```

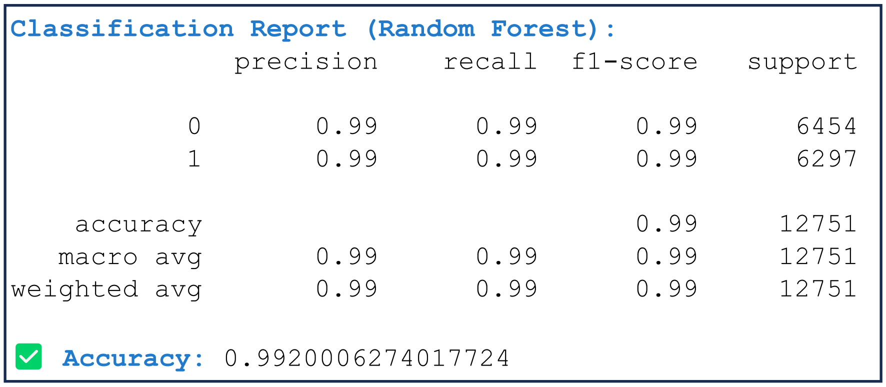
The Logistic Regression model's classification report indicates a strong level of performance for both classes (class 0 and class 1). For class 0, the precision value is 0.99, meaning when the model predicts class 0, it does so correctly 99% of the time. The recall value for class 0 is also 0.99, so the model identifies 99% of all the actual class 0 instances. The F1-score value for class 0 results from the computation of the precision and recall values, as their average is 0.99. For the same case of class 1, the precision, recall, and F1-score are also 0.99. The model's accuracy value for overall performance is about 0.9998, showing that the model identified nearly all the 12,751 records in the test set. The macro average and the weighted average for precision, recall, and F1-score are also 0.99 due to the balanced and high performance of the Logistic Regression model for both classes.

B. Random Forest Modelling

The Python script deploys a Random Forest Classifier through scikit-learn using a pipeline and Grid-Search-CV for hyperparameter tuning. It imports the Random-Forest-Classifier and the required modules and declares a pipeline consisting of a Random Forest classifier using a specific random state. A parameter grid (rf_params) is defined to try various values for the number of estimators, the maximum depth of the trees, the minimum number of internal node samples to split, and the minimum number of samples for a leaf node. It next employs 5-fold cross-validation in conjunction with accuracy as a scoring metric and Grid-Search-CV to determine the best hyperparameter combination by using the pipeline to fit the training data. After the grid search, the script produces predictions for the test set using the best-performing Random Forest classifier and then outputs the best hyperparameters discovered, the classification report, and the accuracy score of the model for the test data.

Output:

Table 2: Random Forest Results



```
Classification Report (Random Forest):
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      6454
           1       0.99      0.99      0.99      6297

    accuracy                           0.99     12751
   macro avg       0.99      0.99      0.99     12751
weighted avg       0.99      0.99      0.99     12751

✅ Accuracy: 0.9920006274017724
```

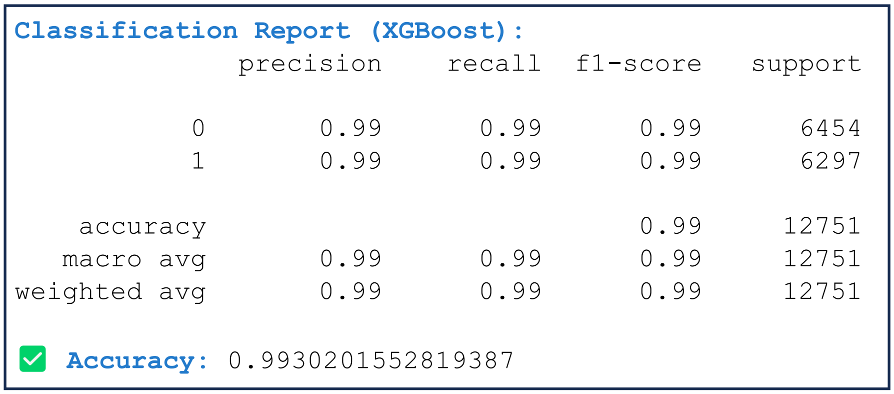
The Random Forest model expresses outstanding performance results for both class 0 and class 1 in its classification report. The precision score for class 0 reached 0.99, which demonstrates the model correctly identified 99% of instances that were truly classified as Class 0. The recall score of 0.99 confirms that the model accurately detects 99% of original class 0 occurrences. The resulting F1-score for class 0 is 0.99. All scores for class 1 in the Random Forest model reach a precision of 0.99, along with a recall of 0.99 and an F1-score of 0.99. Both categories of model performance metrics, conducted by macro and weighted averaging, reached 0.99 precision and recall, and F1-score metrics. Proof of balanced and high-performance behavior emerges from combining 0.99 weighted F1-score average metrics with collective precision and recall evaluation at 0.99 for both categories.

C. XG-Boost Modelling

XG-Boost classifier runs within a scikit-learn pipeline utilized by Grid-Search-CV for optimizing hyperparameters through Python code execution. The code initializes the XG-Boost-Classifier and imports needed modules to create a pipeline that includes an XG-Boost classifier with parameters that disable label encoding and 'log loss' evaluation metric, and a fixed random state. A parameter grid called xgb_params examines different settings between estimators and depth, along with learning rate and training subsample percentages. Grid-Search-CV performs 5-fold cross-validation using accuracy as the scoring metric to determine the best combination of hyperparameters through training the pipeline on the available data. The code executes predictions on test data with the XG-Boost model selection that produced the best results, followed by showing the identified optimal hyperparameters together with the classification report and model accuracy score obtained on test data.

Output:

Table 3: XG-Boost Results



```
Classification Report (XGBoost):
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      6454
           1       0.99      0.99      0.99      6297

    accuracy                           0.99     12751
   macro avg       0.99      0.99      0.99     12751
weighted avg       0.99      0.99      0.99     12751

✅ Accuracy: 0.9930201552819387
```

The XG-Boost model classification report describes outstanding performance in both class 0 and class 1. For the class 0 results, precision is 0.99, which reflects 99% of the predicted class 0 instances as correct, and the recall for the same also measures 0.99, which represents 99% of all the existing class 0 instances as identified correctly, for a resulting F1-score of 0.99. For the results for class 1 as well, precision, recall, and the F1-score are each 0.99. The overall accuracy of the XG-Boost model approximates 0.993, which implies that the model was correct in about 99.3% of the 12,751 instances in the test set. Macro and weighted averages for precision, recall, as well as for the F1-score, are each also 0.99, which reflects a high and balanced rate of performance in both classes.

Comparison of All Models

The computed code function plot_model_results checks and displays the performance of the given classification model. It accepts the name of the model, the features and labels for the training data, and the features and predicted values for the test data as parameters. It prints the model's accuracy in both the test and train data. It also creates three different types of visualizations: the first compares the accuracy of the train data and the test data through a bar plot; the second visualize the model's prediction against the real values in the test data through a heatmap of the confusion matrix; the third illustrates the histogram of the real and the predicted values in the test data. Each plot contains a useful title and lines for better visibility. The function should serve to represent the performance of the model in the train data and the test data in a complete visual as well as quantitative form.

Output:

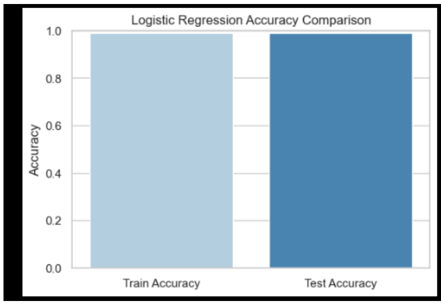


Figure 7: Logistic Regression Accuracy

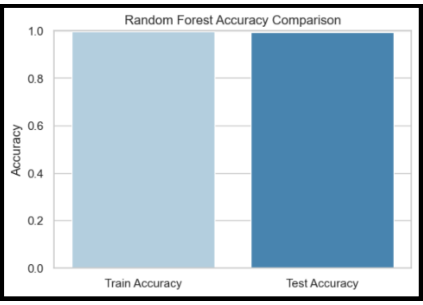


Figure 8: Random Forest Accuracy

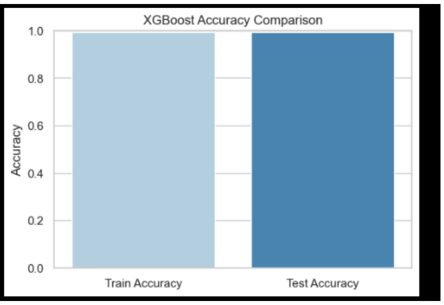


Figure 9: Showcases XG-Boost Accuracy

As observed from the three models' test accuracies are very high, Logistic Regression (0.9909), Random Forest (0.9920), and XG-Boost (0.9930). From this outcome, the analyst inferred that each model gave the correct classification to over 99% of the instances on the unseen test dataset. Of the three, the highest test accuracy is shown by XG-Boost, implying that it generalized slightly better to new, unseen data than Random Forest and Logistic Regression did. These very high accuracies suggest that each of the three models is highly effective at separating classes for this specific problem with very few misclassifications on the test set. Precision, recall, and F1-score of 0.99 for all models further underscores the very good performance, which is an indicator of very high correctness for positive predictions and for detecting all actual present positive instances.

Confusion Matrix

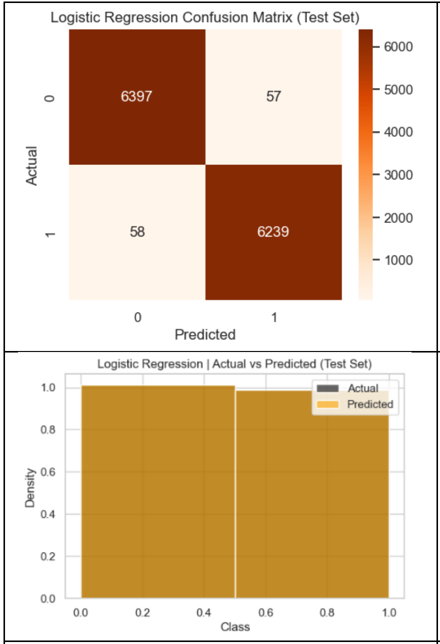


Figure 10: Logistic Regression Confusion Matrix

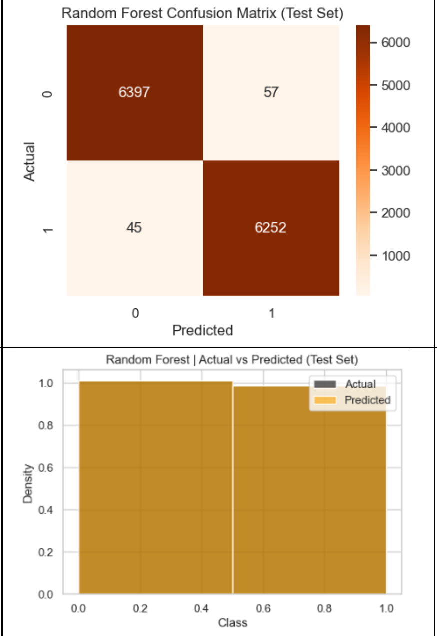


Figure 11: Random Forest Confusion Matrix

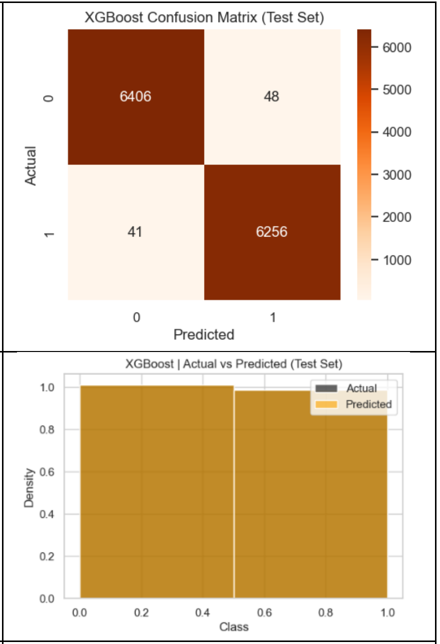


Figure 12: XG-Boost Confusion matrix

By referring to the above results, confusion matrices for the three models on the test dataset demonstrate exceedingly high performance with a strong capability to correctly categorize both classes. In particular, for XG-Boost, there were 6454 actual class 0 instances for which 6406 were correctly predicted class 0 and 48 were misclassified as class 1. Similarly, there were 6297 actual class 1 instances for which 6256 were correctly predicted and 41 were class 0 by mistake. The same is seen for the Random Forest model, with 6397 out of 6454 class 0 instances correctly classified and 57 by mistake, and 6252 out of 6297 class 1 instances correctly classified with 45 by mistake. These indicate a very small rate of false positives and false negatives for the three models, validating their accuracy as seen earlier. The density plots under each confusion matrix further demonstrate the overlap of class distribution that is actual versus class distribution that is predicted, with a strong correlation, and validate the models' capability to categorize correctly.

VI. Real-World Implications and Applications in the USA

Integration with Preventive Healthcare Programs

The deployment of machine learning-based heart disease risk prediction models in preventive care represents a major push in the U.S. public healthcare sector. These models can easily be implemented within electronic health record systems utilized in clinics, hospitals, and primary care to automatically indicate high-risk individuals using real-time clinician data. Through early identification of at-risk patients, providers are empowered to start personalized prevention interventions in the form of lifestyle counseling, pharmacotherapy (for example, use of statins or antihypertensive medications), and routine cardiovascular monitoring. Such proactive care correlates to the U.S. Centers for Disease Control and Prevention (CDC) campaigns, including the Million Hearts® initiative that targets preventing a million heart attacks and strokes in five years. In addition to this, machine learning platforms can also be utilized in the deployment in community health centers and federally qualified health centers (FQHCs), enhancing care reach and support for vulnerable communities that tend to have a higher rate of cardiovascular deaths. Such data-driven deployment enhances the paradigm shift away from reactive treatment towards preventive care that is highly necessary for the management of the epidemic of chronic diseases in the U.S.

Remote Health Monitoring and Telemedicine

Machine learning-driven heart disease prediction models also have transformative value in remote monitoring of health and telemedicine, which have emerged as big trends in the U.S., particularly in the aftermath of the COVID-19 pandemic. With the growing use of wearable sensors and mobile apps for health, continuous tracking data for heart rate, blood pressure, physical activity levels, and sleep patterns can be gathered and processed in real time. With these wearables connected to cloud-based machine learning algorithms, early signs of cardiovascular risk in asymptomatic as well as symptomatic individuals can be identified and notified to both patients and clinicians. In telemedicine consultations, these predictive biomarkers may also help clinicians decide on recommending diagnostic tests, modifying medications, or referring to specialists without the need for in-person visits. This is particularly valuable for rural and far-flung communities in the U.S. with less access to cardiologists and diagnostic equipment. In addition to this, connecting these tools to the telehealth platforms of large providers such as Kaiser Permanente and the Department of Veterans Affairs will improve scalability and national reach.

Reducing Costs in the U.S. Health Systems through Early Intervention

Heart disease continues to be among the costliest chronic conditions to treat in the United States, with direct and indirect costs expected to reach more than $1 trillion a year by 2035. Machine learning-based early risk prediction represents a strong potential to save these costs through timely prevention. By correctly assessing individuals at risk before the occurrence of acute cardiovascular events, healthcare systems can decrease the need for emergency visits, invasive interventions, and prolonged inpatient admissions. Preventive strategies of drug compliance interventions, dietary counseling, and smoking cessation have not only improved clinical outcomes but also proven to save substantially more costs in the process compared to the treatment of established heart disease. Insurers and value-based care organizations can use these prediction algorithms to risk-stratify patient populations and optimize the deployment of resources to reward preventive visits in the face of high-cost reactive care. In Medicare and Medicaid populations, wherein the risk of heart disease remains higher, these predictive systems have the potential to reduce tax burdens on the public purse and enhance the longer-term sustainability of funding for public healthcare.

Customizable Risk Scores for US Demographics

One of the key strengths of machine learning models is that they can provide customizable risk scores that are attuned to the multifaceted demographic profile of the United States. Other traditional risk calculators tend to lack the ability to accurately estimate risk within various racial, ethnic, and socioeconomic subgroups, which in the process has resulted in disparities in diagnosis and treatment. In contrast to this, ML models can learn from diverse data that incorporates variables to represent the needs of minority segments of the population, geographic locations, and social determinants of health. For instance, risk stratification may be made to accommodate higher rates of hypertension within African American communities or reduced access to preventive services in specific rural locations. These customized insights aid in healthcare equity by preventing high-risk people from being missed in the process due to generic models. Public health agencies and academic medical centers have the potential to collaborate to periodically update these algorithms using real-world data to make sure that these remain precise as well as responsive to emerging trends in U.S. cardiovascular disease.

Improving Patient Engagement and Education

Through Predictive Tools, Predictive tools that use machine learning also provide a useful mechanism for facilitating patient engagement and literacy in health as prerequisites to the achievement of long-term prevention strategies. By showing individuals their personalized risk scores and clear explanations of their risk factors in a way that is easily understood, these tools enable patients to become active participants in their cardiovascular disease. Electronic platforms—patient portals, smartphone apps, and interactive dashboards—can illustrate risk levels, trend lines over time, and practically applicable steps to lower risk, for example, by boosting the level of physical exercise or adjusting diet. Such transparency facilitates cooperation and trust between providers and patients and motivates compliance with recommended interventions. Healthcare organizations can also use predictive analytics in their education campaigns by sending individuals specific content aimed directly at their unique risk profile. When people know the "why" behind their risk factors and get instant feedback, they tend to stay engaged in their care process, resulting in improved outcomes and minimized hospital readmissions within the American healthcare system.

VII. AI-Driven Health Diagnostic Perspectives of the Future

Application of Deep Learning to ECG or Image-Based Inputs

The future frontier of AI-based cardiovascular diagnostics includes the use of deep learning models to process complicated data in the form of electrocardiograms (ECGs), echocardiograms, and other imaging modalities. In contrast to traditional machine-based models designed to work on structured tabular data, deep learning—especially convolutional neural networks (CNNs)—can detect minute patterns and abnormalities in raw data or images that might not even be visibly apparent to medical experts. For example, deep learning has already demonstrated promising performance in the interpretation of 12-lead ECGs for the detection of silent MI, arrhythmias, and even preclinical signs of heart failure. These enable the earlier and more accurate detection of cardiac abnormalities in asymptomatic patients as well. In addition to that, when combined with digital health appliances like smartphone-linked ECG monitors or wearable ultrasound devices, deep learning models enable almost instant analysis to aid in both emergency and preventive care decision-making. In the long run, this has the potential to revolutionize routine cardiac screening procedures by transferring these to more convenient and patient-focused care platforms outside the confines of hospitals.

Real-time risk monitoring using wearables and IoT

The rise in wearable technologies and Internet of Things (IoT) devices will revolutionize the landscape for real-time monitoring of cardiovascular risk. Smartwatches, fitness wearables, and wearable patches delivering ECG data are now collecting vital signs in real time, including heart rate variability, respiratory rate, skin temperature, and ambulation. When linked to cloud-based AI platforms, this real-time stream of biometric signals can be continuously monitored to automatically detect early signs of heart disease or even an acute event such as atrial fibrillation, ischemia, or cardiac arrest. These systems have the potential to notify both healthcare providers and patients of anomalous trends before symptoms begin to appear, enabling early intervention. In the American healthcare system context, this innovation has the potential to have a dramatic impact in reducing the number of cardiac events that are preventable among high-risk cohorts. It also enables the emerging trend towards decentralized care, whereby the patient becomes empowered to view and adjust their healthcare outside the conventional clinical paradigm. With integration into telehealth systems and EHR platforms, real-time monitoring will form the foundation of next-generation predictive and personalized medicine.

Expansion to Predict Other Comorbidities (Diabetes, Stroke)

As machine and AI technologies continue to mature, there is increasing potential to expand their use to predict not only heart disease but also associated comorbid conditions such as stroke, metabolic syndrome, chronic kidney disease, and type 2 diabetes. These conditions tend to co-occur and have many common risk factors in common, including obesity, sedentary lifestyles, and hypertension. A unified predictive model that accounts for the interrelationship between multiple chronic conditions has the potential to provide a more complete view of patient health and enable earlier interventions across a range of broader health outcomes. Stroke prediction models, for instance, might include cardiovascular risk profiles, detection of atrial fibrillation, and imaging to flag individuals' preliminary risk. Diabetes onset prediction might use longitudinal laboratory values and lifestyle data to facilitate early pharmacologic or behavioral interventions. Building these multi-disease models might prove particularly revolutionary for accountable care organizations and population-based health efforts that aim to improve outcomes while reducing costs. With AI systems increasingly skilled at parsing multimodal data sources in the form of genomics and microbiome data, the scope of healthcare diagnostics to predict outcomes will expand materially.

Policy, Privacy of Data, and Ethical Implications in U.S. Healthcare

Despite the enormous potential of AI-based diagnostics, there are several policies, data privacy, and ethics challenges that need to be rigorously addressed to guarantee secure and equitable use in the U.S. healthcare ecosystem. Sensitive healthcare data collected through wearables, smartphones, and EHRs generates tremendous concern about patient consent, data privacy, and potential misuse. Regulation in the form of HIPAA (Health Insurance Portability and Accountability Act) offers a base, but might not be prepared to deal with the intricate, real-time, and voluminous data flows characteristic of AI systems. Ethical considerations also include algorithmic bias that emerges due to a lack of diverse data in training and may produce incorrect forecasts for specific population segments. It is a particularly troublesome problem in a multicultural society like the U.S., as it aggravates existing disparities in care. Policymakers need to formulate new directions of regulation to guarantee transparency, accountability, and equity in AI deployment. This includes explainability in AI models, bias-checking processes for AI models on a routine basis, and public watchdog bodies. Moreover, collaboration among technologists, clinicians, ethicists, and lawmakers with diverse disciplines will also be critical to establish public confidence and guarantee AI-driven diagnostics' benefits without infringing on core rights.

VIII. Conclusion

The chief aim of this work was to create and rigorously test machine learning models that are effective in the prediction of heart disease risk for various populations. Based on well-annotated datasets and well-labeled variables like age, systolic/diastolic blood pressure, cholesterol level, type of chest pain, and electrocardiogram results. We used the publicly accessible Cleveland Heart Disease data for this study on Heart Disease Risk Prediction Using Machine Learning. The data consisted of 303 patient records and 14 important attributes typical for cardiovascular health: age, sex, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, and ST depression caused by exercise, among others. The target variable marks the presence or absence of heart disease as labeled in the data, using five categories later binarized for classification purposes (1 = disease, 0 = no disease). To develop a strong predictive model for the identification of people vulnerable to heart disease, three established supervised classification algorithms have been adopted: Logistic Regression, Random Forest Classifier, and XG-Boost Classifier (Extreme Gradient Boosting).To determine the accuracy and reliability of the designed machine learning models for heart disease risk prediction, a battery of evaluation metrics was utilized that presented distinct insights into model performance. The XG-Boost model has a substantial training accuracy, followed very closely by a test accuracy of about 0.993, which indicates good generalization to the unseen test data. The deployment of machine learning-based heart disease risk prediction models in preventive care represents a major push in the U.S. public healthcare sector. These models can easily be implemented within electronic health record systems utilized in clinics, hospitals, and primary care to automatically indicate high-risk individuals using real-time clinician data. Machine learning-driven heart disease prediction models also have transformative value in remote monitoring of health and telemedicine, which have emerged as big trends in the U.S., particularly in the aftermath of the COVID-19 pandemic. One of the key strengths of machine learning models is that they can provide customizable risk scores that are attuned to the multifaceted demographic profile of the United States. As machine and AI technologies continue to mature, there is increasing potential to expand their use to predict not only heart disease but also associated comorbid conditions such as stroke, metabolic syndrome, chronic kidney disease, and type 2 diabetes.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Adler, E. D., Voors, A. A., Klein, L., Macheret, F., Braun, O. O., Urey, M. A., ... & Yagil, A. (2020). Improving risk prediction in heart failure using machine learning. *European journal of heart failure*, *22*(1), 139-147.

[2] Alam, S., Hider, M. A., Al Mukaddim, A., Anonna, F. R., Hossain, M. S., khalilor Rahman, M., & Nasiruddin, M. (2024). Machine Learning Models for Predicting Thyroid Cancer Recurrence: A Comparative Analysis. *Journal of Medical and Health Studies*, *5*(4), 113-129.

[3] Al Amin, M., Liza, I. A., Hossain, S. F., Hasan, E., Islam, M. A., Akter, S., ... & Haque, M. M. (2025). Enhancing Patient Outcomes with AI: Early Detection of Esophageal Cancer in the USA. *Journal of Medical and Health Studies*, *6*(1), 08-27.

[4] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms*, *16*(2), 88.

[5] Di Tanna, G. L., Wirtz, H., Burrows, K. L., & Globe, G. (2020). Evaluating risk prediction models for adults with heart failure: A systematic literature review. *PloS one*, *15*(1), e0224135.

[6] Garg, A., Sharma, B., & Khan, R. (2021). Heart disease prediction using machine learning techniques. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012046). IOP Publishing.

[7] Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018, March). Prediction of heart disease using machine learning. In *2018 second International Conference on electronics, communication and aerospace technology (ICECA)* (pp. 1275-1278). IEEE.

[8] Golas, S. B., Shibahara, T., Agboola, S., Otaki, H., Sato, J., Nakae, T., ... & Jethwani, K. (2018). A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC medical informatics and decision making*, *18*, 1-17.

[9] Hasan, E., Haque, M. M., Hossain, S. F., Al Amin, M., Ahmed, S., Islam, M. A., ... & Akter, S. (2024). CANCER DRUG SENSITIVITY THROUGH GENOMIC DATA: INTEGRATING INSIGHTS FOR PERSONALIZED MEDICINE IN THE USA HEALTHCARE SYSTEM. *The American Journal of Medical Sciences and Pharmaceutical Research*, *6*(12), 36-53.

[10] Hossain, S., Miah, M. N. I., Rana, M. S., Hossain, M. S., Bhowmik, P. K., & Rahman, M. K. (2024). ANALYZING TRENDS AND DETERMINANTS OF LEADING CAUSES OF DEATH IN THE USA: A DATA-DRIVEN APPROACH. *The American Journal of Medical Sciences and Pharmaceutical Research*, *6*(12), 54-71.

[11] Hossain, S. F., Al Amin, M., Liza, I. A., Ahmed, S., Haque, M. M., Islam, M. A., & Akter, S. (2023). AI-Based Brain MRI Segmentation for Early Diagnosis and Treatment Planning of Low-Grade Gliomas in the USA. *British Journal of Nursing Studies*, *3*(2), 37-55.

[12] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing.

[13] Kaur, B., & Kaur, G. (2022, September). Heart disease prediction using modified machine learning algorithm. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 1* (pp. 189-201). Singapore: Springer Nature Singapore.

[14] Lutimath, N. M., Chethan, C., & Pol, B. S. (2019). Prediction of heart disease using machine learning. *International journal Of Recent Technology and Engineering*, *8*(2), 474-477.

[15] Malav, A., Kadam, K., & Kamat, P. (2017). Prediction of heart disease using k-means and artificial neural network as hybrid approach to improve accuracy. *International Journal of Engineering and Technology*, *9*(4), 3081-3085.

[16] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.

[17] Nasiruddin, M., Hider, M. A., Akter, R., Alam, S., Mohaimin, M. R., Khan, M. T., & Sayeed, A. A. (2024). OPTIMIZING SKIN CANCER DETECTION IN THE USA HEALTHCARE SYSTEM USING DEEP LEARNING AND CNNS. *The American Journal of Medical Sciences and Pharmaceutical Research*, *6*(12), 92-112.

[18] Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining techniques. Heart Disease, 7(1), 129-137.

[19] Pant, L., Al Mukaddim, A., Rahman, M. K., Sayeed, A. A., Hossain, M. S., Khan, M. T., & Ahmed, A. (2024). Genomic predictors of drug sensitivity in cancer: Integrating genomic data for personalized medicine in the USA. *Computer Science & IT Research Journal*, *5*(12), 2682-2702.

[20] Pasha, S. N., Ramesh, D., Mohmmad, S., & Harshavardhan, A. (2020, December). Cardiovascular disease prediction using deep learning techniques. In *IOP conference series: materials science and engineering* (Vol. 981, No. 2, p. 022006). IOP Publishing.

[21] Pocock, S. J., Ariti, C. A., McMurray, J. J., Maggioni, A., Køber, L., Squire, I. B., ... & Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC). (2013). Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *European heart journal*, *34*(19), 1404-1413.

[22] Reddy, K. V. V., Elamvazuthi, I., Aziz, A. A., Paramasivam, S., Chua, H. N., & Pranavanand, S. (2021). Heart disease risk prediction using machine learning classifiers with attribute evaluators. *Applied Sciences*, *11*(18), 8352.

[23] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*(6), 345.

[24] Sajeev, S., Maeder, A., Champion, S., Beleigoli, A., Ton, C., Kong, X., & Shu, M. (2019). Deep learning to improve heart disease risk prediction. In *Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting: First International Workshop, MLMECH 2019, and 8th Joint International Workshop, CVII-STENT 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 1* (pp. 96-103). Springer International Publishing.

[25] Srivastava, A., & Kumar Singh, A. (2022, April). Heart disease prediction using machine learning. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 2633-2635). IEEE.

[26] Sujatha, P., & Mahalakshmi, K. (2020, November). Performance evaluation of supervised machine learning algorithms in prediction of heart disease. In *2020 IEEE International Conference for Innovation in Technology (INOCON)* (pp. 1-7). IEEE.

[27] Tohyama, T., Ide, T., Ikeda, M., Kaku, H., Enzan, N., Matsushima, S., ... & Tsutsui, H. (2021). Machine learning-based model for predicting 1-year mortality of hospitalized patients with heart failure. ESC heart failure, 8(5), 4077-4085.

[28] Zeeshan, M. A. F., Mohaimin, M. R., Hazari, N. A., & Nayeem, M. B. (2025). Enhancing Mental Health Interventions in the USA with Semi-Supervised Learning: An AI Approach to Emotion Prediction. *Journal of Computer Science and Technology Studies*, *7*(1), 233-248.