International Journal of Advanced Research in Engineering Technology & Science
ISSN: 2349-2819
Impact Factor: 7.10
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Email: editor@ijarets.org Volume-13, Issue-4 April – 2026 www.ijarets.org

Cardiovascular Disease Prediction Using Machine Learning Algorithms

1st Shrey Dave
Department of Information Technology
L. D. College of Engineering
Ahmedabad, India shreykdave2411@gmail.com
2nd Bakul Panchal
Department of Information Technology
L. D. College of Engineering Ahmedabad, India bakul@ldce.ac.in

Abstract**—** *Cardiovascular disease (CVD) remains one of the leading causes of mortality worldwide, emphasizing the need for early prediction to enhance preventive healthcare. This research focuses on individuals aged 30–40 years, investigating the correlation between age, health factors, and lifestyle habits to improve early risk detection. The dataset, after preprocessing tasks such as data cleaning and transformation, undergoes feature selection, resulting in 08 key features. The data is then split into an 80:20 ratio for training and testing. Multiple machine learning algorithms, including Multi-Layer Perceptron (MLP), Random Forest (RF), Decision Tree (DT), and Extreme Gradient Boosting (XGBoost), are employed for classification. Among these, the XGBoost model achieves a superior accuracy of 89.71% and is integrated into a graphical user interface (GUI) for real time CVD risk prediction. Further aiding early diagnosis and prevention strategies. This study contributes to improving CVD risk assessment, particularly for the 30–40 age group, by leveraging machine learning and predictive analytics.*

Index Terms**—***Cardiovascular Disease Prediction (CVD), 31-40 Years Age, Machine Learning, Cardiovascular Disease Classification, Features Selection*

I. INTRODUCTION

Cardiovascular disease (CVD) remains the leading cause of global morbidity and mortality, accounting for over 70% of global mortality. CVDs account for approximately 43% of all deaths, and CVD risk factors include unhealthy diet, smoking, excessive sugar intake, and obesity, according to the Global Burden of Disease study (2017). The economic cost of CVDs between 2010 and 2015 was estimated at USD 3.7 trillion [1]. The World Health Organization (WHO) projects that by 2030, CVD mortality will rise to 23.6 million annually, the vast majority of which will be caused by heart disease and stroke. Diagnostic imaging like electrocardiogram and CT scan, however, remains expensive and inaccessible, particularly to low and middle-income countries, and therefore early detection is the solution [2].

Machine learning (ML) has proven to be an economic and precise method for heart disease prediction, outperforming conventional diagnostic techniques. Recent studies have explored a few ML models, including Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), and XGBoost, for improving early heart disease detection [3]. Even though ML-based models have shown promising results, previous studies have employed small datasets, limiting their generalizability [4]. To address this limitation,

our research employs a publicly available, larger dataset to enhance predictive performance and robustness.

In this study here, we compare the performance of several ML algorithms utilized for heart disease prediction, including data preprocessing techniques like k-modes clustering for better model convergence. Experimental data is collected from Kaggle, and all the calculations are performed in Python using Google Colab. Additionally, strategies like hyperparameter tuning are utilized for enhancing model performance and dataset balancing [5]. Additionally, this study will explore the association of age with primary physiological and lifestyle factors such as height, weight, cholesterol, and lifestyle factors such as smoking and alcohol use. By studying the association of these factors, we anticipate gaining insights into early cardiovascular risk detection and personalized medicine. Weight, after consulting with clinicians, will be assigned to the correlation of these traits such that prediction of the occurrence of cardiovascular disease within the age group of 30 to 40 years becomes more accurate. This process makes the interpretation of ML models in a manner that predictive results align with clinical importance and real-world applicability. Through the application of such methods, this work offers its own contribution to raising the accuracy of early detection, lowering computation expense, and improving ML-based cardiovascular disease prediction to target ages.

The remaining part of this paper is structured as follows: Section 2 presents a literature survey of recent state-of-the-art machine learning-based risk assessment and classification techniques for predicting cardiovascular disease (CVD). Section 3 outlines the methodology, detailing data preprocessing, feature selection methods, and machine learning classifiers used for CVD prediction. Experimental results are discussed in Section 4, where we provide a detailed comparison of the overall model performance against current state-of-the-art techniques. Finally, Section 5 concludes the paper and speculates on potential future work to enhance CVD prediction using novel machine learning methods.

II. RELATED WORK

The recent years have seen unprecedented growth in the use of machine learning (ML) to predict cardiovascular disease (CVD) in the healthcare sector. ML models have shown great promise to enhance early diagnosis and risk prediction with the help of large databases that provide improved prediction results. Various studies have been carried out on ML methods and feature selection techniques for the aim of enhancing CVD prediction models.

TABLE I
OVERVIEW OF RELATED WORK AND ITS COMPARISON

Reference	Methods	Accuracy	Dataset	Limitation
[1]	XGBoost Classifier	97.57 %	1. Cleveland heart disease dataset (CHDD) 2. Private Dataset	Low data quality, like missing values or noise, can adversely affect the learning process of the model as well as its performance.
[2]	Hybrid GA-RBF (Genetic Algorithm - Radial Basis Function)	85.00 %	Cleveland heart disease dataset (CHDD)	Reliance on a small benchmark dataset limits generalizability, emphasizing the need for larger datasets for improved prediction.

[3]	Soft Voting Ensemble Classifier SVE	Dataset 1 : 93.44% Dataset 2 : 95.00%	1. The UC Irvine ML Repository Cleveland dataset 2. The IEEE Dataport heart disease dataset	The measurement is mainly accuracy-based with no regard to other key measures like recall and precision, which may provide a better extensive assessment of model performance.
[4]	Fast Conditional Mutual Information (FCMIM)	92.37 %	Cleveland heart disease dataset (CHDD)	Reliance on the Cleveland dataset limits population variability and affects generalizability.
[5]	k-modes Clustering	With GridSearchCV : 87.28 % , Without GridSearchCV : 86.94 %	Cardio Vascular Disease Dataset (CDD)	Having one dataset restricts the population generalizability of the results.

A research used several ML classifiers, i.e., XGBoost, Naive Bayes, and Support Vector Machine (SVM), to classify heart disease using the Cleveland Heart Disease dataset and a private dataset. The research used sophisticated feature selection methods like ANOVA and chi-square, and the Synthetic Minority Oversampling Technique (SMOTE) to handle class imbalance. The findings showed that XGBoost had the best accuracy of 97.57%, demonstrating its high predictive power in heart disease detection [1].

Yet another study proposed a hybrid ML approach towards coronary disease prediction using a combination of feature selection through Genetic Algorithm (GA) and classification through a Radial Basis Function (RBF) network. The research indicated that through dimensionality reduction, the accuracy of the model was improved from 85.40% to 94.20% depending on the most informative features, highlighting the importance of feature selection in making the ML model ideal for application in medicine [2].

The performance of the ensemble learning was tested in a study with benchmark datasets, e.g., the Cleveland Heart Disease dataset and the IEEE Dataport dataset. The study applied various classifiers like Random Forest, Logistic Regression, and Gradient Boosting, and ultimately demonstrated that a Soft Voting Ensemble Classifier provided higher accuracy (93.44% on the Cleveland dataset and 95% on the IEEE dataset). The study demonstrated the capability of ensemble methods in enhancing CVD prediction performance [3].

An intensive feature selection approach was tested in an-other study where Relief, MRMR, LASSO, and an emerging Fast Conditional Mutual Information (FCMIM) algorithm were utilized. Researchers utilized classifiers such as SVM, Logistic Regression, and Decision Tree, and reported SVM to be the best with maximum accuracy (92.37%). The study emphasized the role of quality feature selection techniques for optimization of cardiovascular disease diagnosis prediction models [4].

In contrast to previous studies with small datasets, one recent study utilized a very large dataset of 70,000 patient records with 12 features including age, gender, and blood pressure. The study suggested k-modes clustering as a pre-processing step and utilized classifiers such as Decision Tree, Random Forest, Multilayer Perceptron (MLP), and XGBoost. The results showed that XGBoost achieved the highest accuracy (87.02% in the absence of cross-validation and 86.87% with cross-validation), followed by MLP. Future directions of research include exploring other clustering methods and improving model interpretability for practical use in real-world clinical applications [5].

These studies collectively highlight the growing significance of ML-based approaches for predicting cardiovascular dis-ease. The mutual reinforcement between ensemble learning approaches, feature selection methods, and extensive datasets is core to the performance enhancement of models and clinical usability.

The primary limitation of the earlier research is that it considered a wide age range without feature selection to identify the most significant predictors. In addition, the studies did not consider a restricted age range explicitly, and therefore the models that were constructed were less effective in predicting the age-specific variables.

Conversely, this research first utilized a cardiovascular dis-ease data set of 11 features. Following the application of feature selection methods, the data set was trimmed down to eight relevant features, thus increasing the accuracy and the applicability of the study. The data set contains 1,790 samples with a particular interest in subjects 30-40 years of age. Additionally, there is a private data set of 500 samples with the same eight selected features to confirm the results. Table 1 is a comparative overview of cardiovascular disease prediction studies on large datasets, once more illustrating the benefit of using a well-characterized dataset to obtain better predictive accuracy.

III. METHODOLOGY

This research aims to predict the probability of cardiovascular disease (CVD) in individuals aged 30 to 40 years using machine learning techniques, which can assist medical professionals in early diagnosis and preventive care. To achieve this objective, we applied various machine learning algorithms to a dataset and analyzed their performance. Our methodology involves preprocessing the data by handling missing values, removing outliers, and normalizing features. We employed the Random Forest algorithm for feature selection to retain the most significant attributes contributing to CVD risk.

Next, we trained multiple models, including Decision Tree, Random Forest, Multilayer Perceptron, and XGBoost, and evaluated their performance using accuracy, precision, recall, and ROC-AUC scores. Based on the results, we identified XGBoost as the best-performing model and integrated it into a graphical user interface (GUI) for real-time CVD prediction. This improved methodology enhances predictive accuracy and model performance, as demonstrated in Figure 1. By leveraging machine learning, our approach provides a data-driven solution for cardiovascular risk assessment, enabling healthcare professionals to make informed decisions and improve early intervention strategies.

A. Data Source

This research utilizes two distinct datasets for cardiovascular disease (CVD) prediction in individuals aged between 30 to 40 years.

- **Dataset 1:** The Cardiovascular Disease Dataset consists of 1,790 patient records, each containing 11 distinct features such as age, gender, systolic blood pressure, diastolic blood pressure, cholesterol levels, glucose levels, smoking habits, alcohol consumption, physical activity levels, height, and weight. This dataset serves as the primary source for model training and evaluation.
- **Dataset 2:** A private dataset, collected through live sampling, consists of 500 patient records with the same 11 features. This dataset provides real-time data, ensuring the model remains updated with recent trends and patterns in CVD occurrences among the target age group.

The target variable, `cardio`, represents the presence of cardiovascular disease, where 1 indicates a patient diagnosed with CVD, and 0 indicates a healthy individual. These datasets collectively enhance the robustness of the predictive model by combining historical records with real-time data, improving its

generalizability and effectiveness in CVD risk assessment.

B. Outlier Removal

Outlier detection and removal is a crucial step in ensuring the reliability and robustness of any machine learning model, particularly in the healthcare domain. In this study, a quantile-based method was applied to eliminate extreme values that could adversely affect model performance. Specifically, for the features `ap_hi` (systolic blood pressure), `ap_lo` (diastolic blood pressure), weight, and height, values falling below the 2.5th percentile and above the 97.5th percentile were identified as outliers and subsequently removed from the dataset. This approach helped in mitigating the influence of data anomalies and measurement errors, thereby improving the quality and stability of the model training process.

C. Feature Selection Using Random Forest Algorithm

A Random Forest-based feature selection technique was employed to identify the most significant predictors of cardiovascular disease (CVD). The dataset included features such as age, height, weight, gender, systolic blood pressure (`ap_hi`), diastolic blood pressure (`ap_lo`), cholesterol level, glucose level, smoking status, alcohol intake, and physical activity. The `id` column was removed as it held no predictive value. The feature set and target variable (`cardio`) were defined, and the dataset was split into training and testing sets using an 80:20 ratio. Numerical features were standardized to ensure consistent scaling. A Random Forest classifier (`RandomForestClassifier` with `random_state=42`) was trained to compute feature importance scores using the Gini impurity criterion. Features with an importance score above the threshold of 0.02 were retained. This method effectively reduced dimensionality, improved model performance, and enhanced the interpretability of the predictive model.

Based on the feature importance scores derived from the Random Forest algorithm, a subset of the most relevant features was selected for cardiovascular disease prediction. The selected features included age, gender, height, weight, systolic blood pressure (`ap_hi`), diastolic blood pressure (`ap_lo`), cholesterol level, and glucose level, as these exhibited higher importance scores in relation to the target variable. In contrast, features such as physical activity (active), smoking status (smoke), and alcohol intake (alco) demonstrated relatively lower importance and were thus excluded from further model training to enhance efficiency and reduce dimensionality.

TABLE II
FEATURE OVERVIEW AND DESCRIPTIONS

Feature	Variable	Min Value	Max Value
Age	Age	10,950	14,600
Height	Height	55	250
Weight	Weight	10	200
Gender	Gender	1 (Female)	2 (Male)
Systolic Blood Pressure	ap_hi	-150	16,020
Diastolic Blood Pressure	ap_lo	-70	11,000
Cholesterol Level	Chol	1 (Low)	3 (High)
Glucose Level	Gluc	1 (Low)	3 (High)

Smoking Habit	Smoke	0 (No)	1 (Yes)
Alcohol Intake	Alco	0 (No)	1 (Yes)
Physical Activity	Active	0 (No)	1 (Yes)
Presence of Cardiovascular Disease	Cardio	0 (No)	1 (Yes)

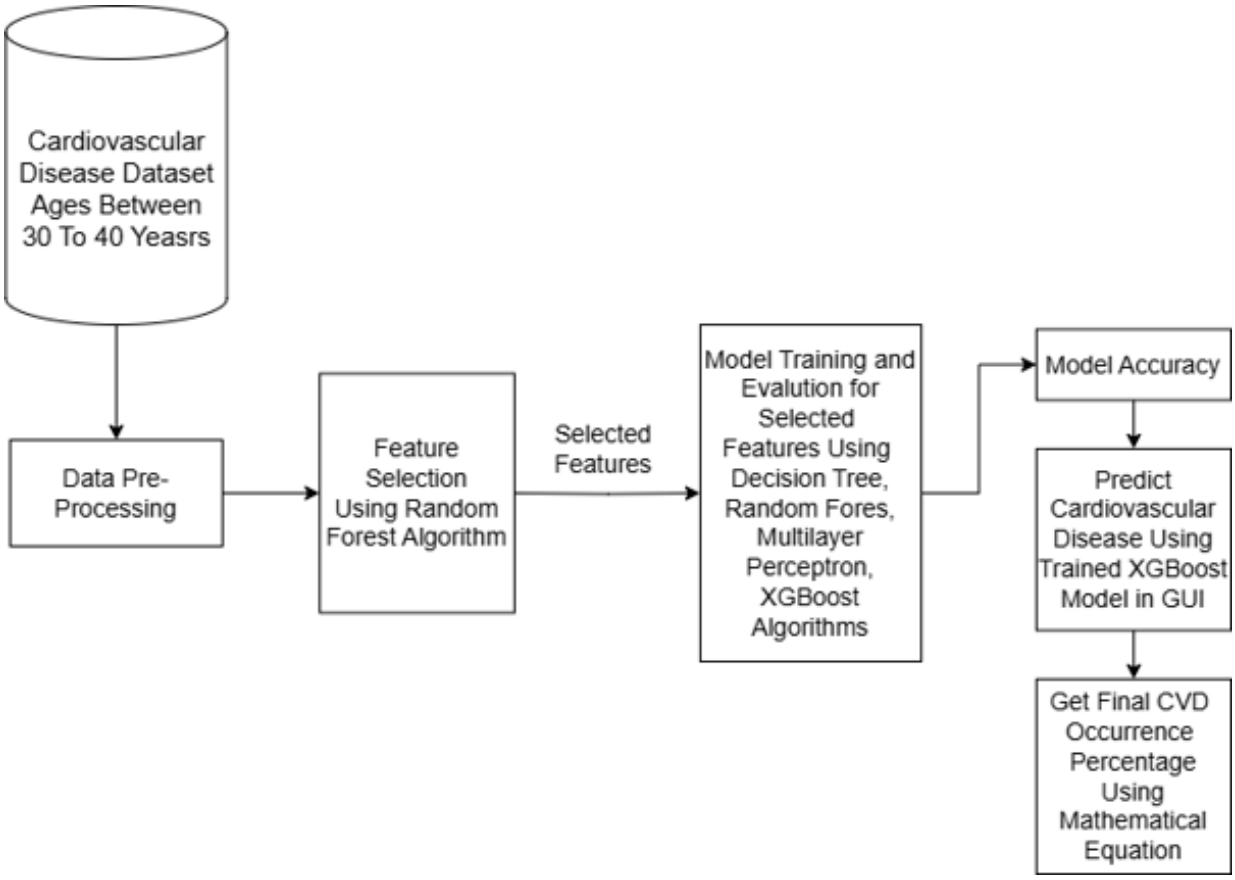

Fig. 1. Flowchart of the Proposed Classification Model

Algorithm 1 Pseudocode: Feature Selection Using Random Forest

1. **Input:** Dataset D with features F and target variable y = cardio
2. **Preprocess Data:**
   a. Remove irrelevant column ('id')
   b. Define feature set $X = D\backslash\{\text{cardio}\}$, and target $y = D[\text{cardio}]$
   c. Split X and y into training and testing sets (80:20)
   d. Standardize features using z-score normalization
3. **Train Model:**

a. Initialize `RandomForestClassifier`

b. Fit the model on training data (X_{train}, y_{train})

4. Feature Selection:

a. Compute feature importance scores from trained model

b. Select features with importance > 0.02

c. Form new dataset $D_{selected}$ with selected features

5. Output: Optimized feature set $D_{selected}$

TABLE III

FEATURE IMPORTANCE SCORES FOR DATASETS 1 AND 2

Features	Importance Score for Dataset 1	Importance Score for Dataset 2
ap_hi	0.277753	0.143004
Age	0.168993	0.090286
weight	0.157546	0.145149
ap_lo	0.142900	0.180172
Height	0.123446	0.050192
Cholesterol	0.045740	0.094150
Gluc	0.022761	0.093588
Gender	0.020417	0.030598
Active	0.018058	0.023558
smoke	0.012332	0.131134
alco	0.010053	0.018168

D. *Feature Engineering and Selection*

In the context of cardiovascular disease (CVD) prediction, the quality and form of input data significantly influence the performance and interpretability of machine learning algorithms. This study implemented a structured feature engineering and selection process that included transforming continuous input variables into categorical features and deriving new meaningful clinical features. These steps are crucial in enhancing both the predictive power and interpretability of classification models used in medical diagnosis.

1) Age Transformation: The original dataset recorded the patients age in days, which was not suitable for direct clinical interpretation. To enhance interpretability and facilitate age-based risk analysis, age was converted from days to years using the following transformation:

$$\text{Age (in years)} = \frac{\text{Age (in days)}}{365}$$

Subsequently, the age variable was discretized into bins spanning five-year intervals between 30 to 40 years to form a new categorical variable `age_bins`. This stratification enabled the examination of CVD prevalence across specific age brackets, with a particular focus on the 30–40 year group, which represents the core of the research gap.

2) *Derivation and Categorization of Body Mass Index (BMI):* Body Mass Index (BMI), a widely recognized measure for evaluating body fat, was calculated using the standard formula:

$$BMI = \frac{\text{Weight (kg)}}{(\text{Height (m)})^2}$$

The resulting continuous BMI values were categorized into clinically defined weight classes based on World Health Organization (WHO) guidelines:

- **Underweight:** BMI < 18.5
- **Normal weight:** 18.5 ≤ BMI < 25
- **Overweight:** 25 ≤ BMI < 30
- **Obese Class I:** 30 ≤ BMI < 35
- **Obese Class II:** 35 ≤ BMI < 40
- **Extreme Obesity:** BMI ≥ 40

Each class was encoded numerically (0 to 5) to create the `bmi_category` feature, allowing algorithms to effectively differentiate patterns associated with various weight-related health conditions.

3) *Calculation of Mean Arterial Pressure (MAP):* To better represent cardiovascular load and blood flow dynamics, the Mean Arterial Pressure (MAP) was computed from systolic (`ap_hi`) and diastolic (`ap_lo`) blood pressure readings using the following formula:

$$MAP = \frac{1}{3}(\text{ap\_hi} + 2 \times \text{ap\_lo})$$

This derived continuous variable was then discretized into deciles within the 70–120 mmHg range, creating a new categorical variable `map_category`. This conversion reduced the influence of outliers and enhanced the interpretability of cardiovascular stress levels within predictive models.

E. *Clustering*

Clustering is an unsupervised machine learning technique that groups instances based on similarity or dissimilarity measures. As the dataset used in this study was pre-processed to consist entirely of categorical variables, the conventional k-means clustering algorithm, which is more suitable for numerical data, would not be effective. To address this, the k-modes clustering algorithm, introduced by Huang (1997), was employed. The k-modes algorithm is an extension of k-means that is capable of handling categorical data by using a simple matching dissimilarity measure and replacing cluster means with modes—the most frequent category values within a cluster.

TABLE IV
MAP RANGE AND CATEGORY

MAP Range (mmHg)	Category
70–80	1
80–90	2
90–100	3
100–110	4
110–120	5

Given that the dataset was transformed into categorical form through techniques such as binning and encoding, k-modes is a suitable method for analyzing patient groupings. To enhance interpretability and leverage known biological differences in cardiovascular disease (CVD) patterns, the dataset was split based on gender into two subsets:

- **Male:** gender = 2
- **Female:** gender = 1

Each gender-specific subset was subjected to independent k-modes clustering to capture gender-specific patterns of cardio-vascular risk. To determine the optimal number of clusters, the elbow method was employed. This involved fitting k-modes models with varying values of k, calculating the corresponding cost (sum of dissimilarities between cluster modes and data points), and plotting these values. The "elbow point" or inflection point in the graph indicates where the marginal gain in clustering quality diminishes with the addition of more clusters. For both male and female datasets, the elbow point was observed at $k = 2$, suggesting that two clusters optimally represent the data distribution within each gender group.

After identifying the optimal clusters:

- Cluster labels were assigned back to the respective male and female datasets.
- The labeled datasets were then merged to form the final gender-wise clustered dataset.
- This combined dataset was saved as a CSV file for further statistical analysis and model development.

This gender-specific clustering approach allows for a deeper understanding of potential sex-based variations in cardiovascular risk factors and enhances the model's capability to offer personalized insights based on demographic segmentation.

F. Modeling

In this research, the clustered dataset was partitioned into two subsets: 80% for training and 20% for testing. Various machine learning classification algorithms were then applied to evaluate their effectiveness in predicting cardiovascular disease. The classifiers employed included the Decision Tree Classifier, Random Forest Classifier, Multilayer Perceptron (MLP), and XGBoost. Each model was trained using the training dataset and subsequently evaluated on the testing dataset. The performance of these classifiers was assessed using standard evaluation metrics, namely Accuracy, Precision, Recall, and F1-Score, to determine the most effective algorithm for the prediction task.

1) ***Decision Tree Classifier (DT):*** The Decision Tree classifier is a supervised learning algorithm that recursively splits the dataset into subsets based on feature values, resulting in a tree-like structure of decisions. It is particularly valued for its interpretability and ability to handle both numerical and categorical data. In this study, we used the `DecisionTreeClassifier` from the Scikit-learn library with `random_state=42` to ensure reproducibility. To enhance performance and control overfitting, a hyperparameter tuning process was carried out using Grid Search, where the parameters explored included:

- `max_depth` with values [5, 10, 15, 20]
- `min_samples_split` with values [2, 5, 10]
- `min_samples_leaf` with values [1, 2, 4]

These parameters govern the complexity of the tree and directly affect its generalization ability.

2) ***Random Forest Classifier (RF)****:* The Random For-est classifier is an ensemble method that constructs multiple decision trees during training and outputs the class that is the mode of the classes predicted by individual trees. It is robust against overfitting and performs well on datasets with both high dimensionality and missing values. The `RandomForestClassifier` was employed with `random_state=42` for consistent results. Hyperparameter tuning was conducted via Grid Search, where the following parameters were optimized:

- `n_estimators` = [50, 100, 200]
- `max_depth` = [5, 10, 15]
- `min_samples_split` = [2, 5, 10]
- `min_samples_leaf` = [1, 2, 4]

These parameters help balance the trade-off between model complexity and computational efficiency, thereby ensuring optimal predictive performance.

3) ***Multilayer Perceptron (MLP)****:* The Multilayer Perceptron is a type of feedforward artificial neural network that uses one or more hidden layers to capture complex, non-linear relationships in the data. It is trained using backpropagation and is well-suited for problems with structured input features. In this research, the `MLPClassifier` from Scikit-learn was used with `random_state=42` to maintain reproducibility. The hyperparameters tuned using Grid Search included:

- `hidden_layer_sizes` with options [(50,), (100,), (50, 50)]
- `activation functions` as ['relu', 'tanh']
- `solver` = ['adam']
- `learning_rate_init` strategies = ['constant', 'adaptive']

TABLE V
FINAL FEATURE SET AFTER FEATURE ENGINEERING

Feature	Variable Name	Value Range / Categories	Description
Gender	gender	1 (male) to 2 (female)	Binary gender representation
Age (Binned)	Age	0 (min) to 6 (max)	Age in 5-year intervals from 30–40 years
Body Mass Index	BMI Class _	0 (min) to 5 (max)	BMI class from underweight to extreme obesity
Mean Arterial Pressure	MAP Class _	0 (min) to 5 (max)	MAP bins in 10 mmHg intervals
Cholesterol Level	Cholesterol	1 (normal) to 3 (high)	Categorical cholesterol levels
Glucose Level	Gluc	1 (normal) to 3 (high)	Categorical glucose levels
Cardiovascular Disease Presence	Cardio	0 (no) to 1 (yes)	Indicates presence or absence of CVD

This configuration enabled exploration of the most effective architecture and learning strategy for the classification task.

4) ***eXtreme Gradient Boosting (XGBoost)****:* XGBoost is an advanced gradient boosting algorithm known for its speed, accuracy, and efficient handling of large-scale data. It builds additive tree models sequentially to minimize a regularized objective function. In this study, the `XGBClassifier` was utilized with predefined parameters including:

- `eval_metric='logloss'`
- `use_label_encoder=False`
- `random_state=42`
- `learning_rate=0.9`
- `max_depth=4`
- `n_estimators=100`

To further optimize its performance, hyperparameter tuning was performed over the following ranges:

- `learning_rate` = [0.01, 0.1, 0.2]
- `max_depth` = [3, 5, 7]
- `n_estimators` = [50, 100, 200]
- `subsample` = [0.8, 1.0]

These parameters were chosen to enhance model generalization, prevent overfitting, and achieve a balance between bias and variance.

IV. RESULTS

This research was conducted using Google Colab and a local system equipped with an Intel(R) Core(TM) i5-4300U CPU @ 1.90GHz (2.49 GHz), 8.00 GB RAM, and Intel(R) HD

Graphics Family, operating on a 64-bit, x64-based architecture. The primary aim of this study was to develop a predictive model for cardiovascular disease (CVD) focusing specifically on individuals aged between 30 to 40 years. Two datasets were used for model training and evaluation. The first dataset contained 1,790 samples, while the second dataset comprised 500 samples. Initially, both datasets had 11 features, but after applying feature selection using the Random Forest algorithm, the number of features was reduced to 8, retaining only the most significant predictors of CVD.

The study utilizedfour machine learning algorithms—Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), and XGBoost classifier. Prior to modeling, data preprocessing steps were performed, which included handling missing values, removing outliers, normalizing the data, and engineering additional features like BMI and Mean Arterial Pressure (MAP). The cleaned data was split into training and testing subsets with an 80/20 ratio. To optimize model performance, GridSearchCV from the scikit-learn library was employed for automated hyper-parameter tuning, leveraging k-fold cross-validation to find the best parameter combinations that maximize accuracy and other scoring metrics. The evaluation was conducted using accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).

In Dataset 1, XGBoost achieved the highest accuracy of 84.96% without cross-validation and 83.23% with cross-validation, with an AUC of 0.80. The MLP model showed improvement in F1-score after cross-validation, rising from 55.55% to 59.12%. All classifiers in this dataset performed above 82% accuracy. In Dataset 2, the models showed even stronger performance. XGBoost once again emerged as the top performer

with 89.71% accuracy, 90.78% precision, 93.50% recall, 92.12% F1-score, and an AUC of 0.95. Random Forest followed closely with an AUC of 0.96 and accuracy of 89.07%. Decision Tree and MLP also showed reliable performance, particularly with improved recall and F1-score values after tuning.

These results highlight the importance of feature selection, age-focused sampling, and cross-validated hyper-parameter tuning in enhancing the accuracy and reliability of CVD prediction models. XGBoost was identified as the most robust model across both datasets, confirming its superiority in capturing complex, non-linear relationships in health-related data. Binary Classified ROC curves and feature correlation heatmap for both dataset are provided for the analysis.

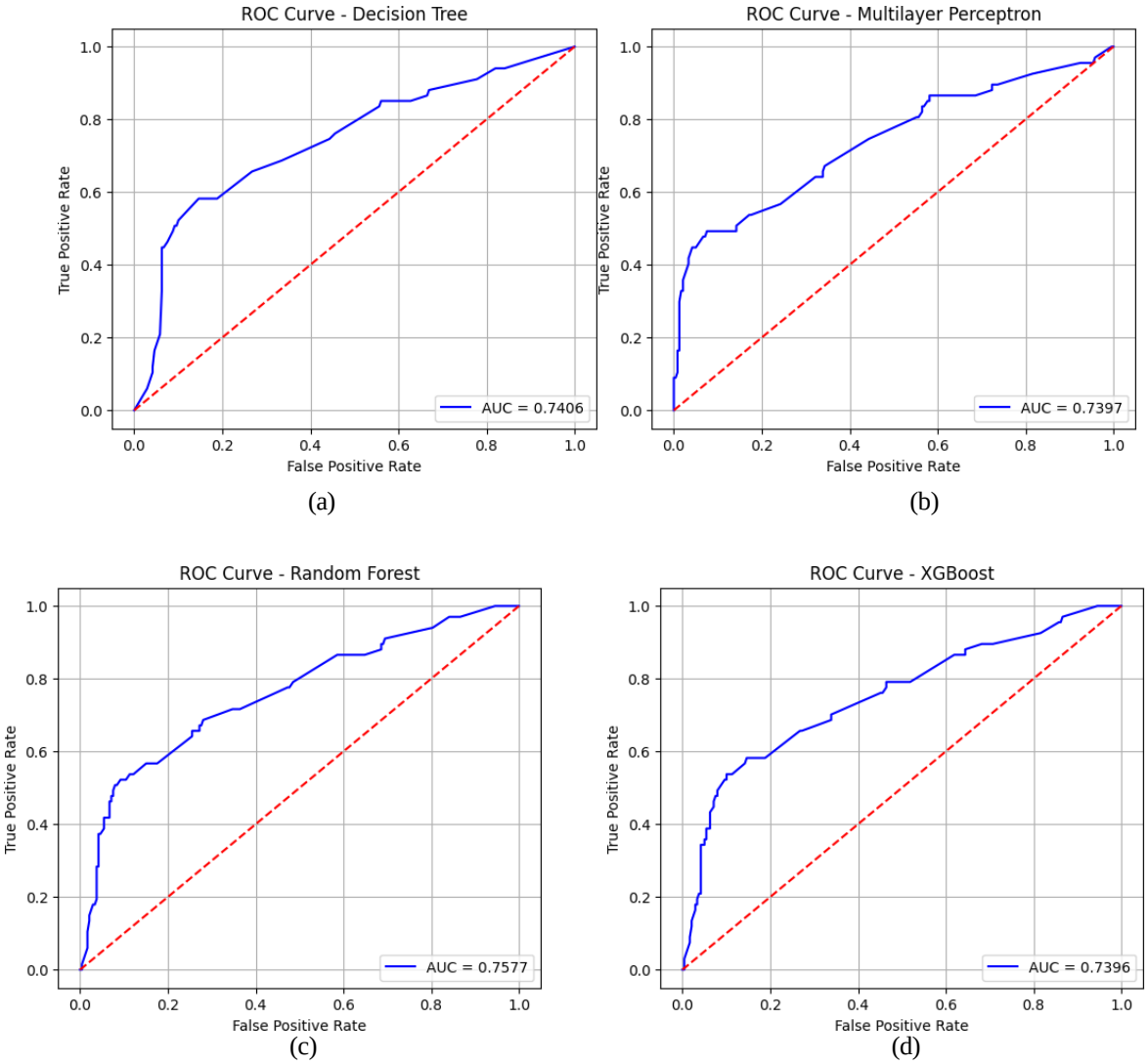

Fig. 2. ROC–area under curve of (a) MLP, (b) RF, (c) DT, and (d) XGB on Dataset 1

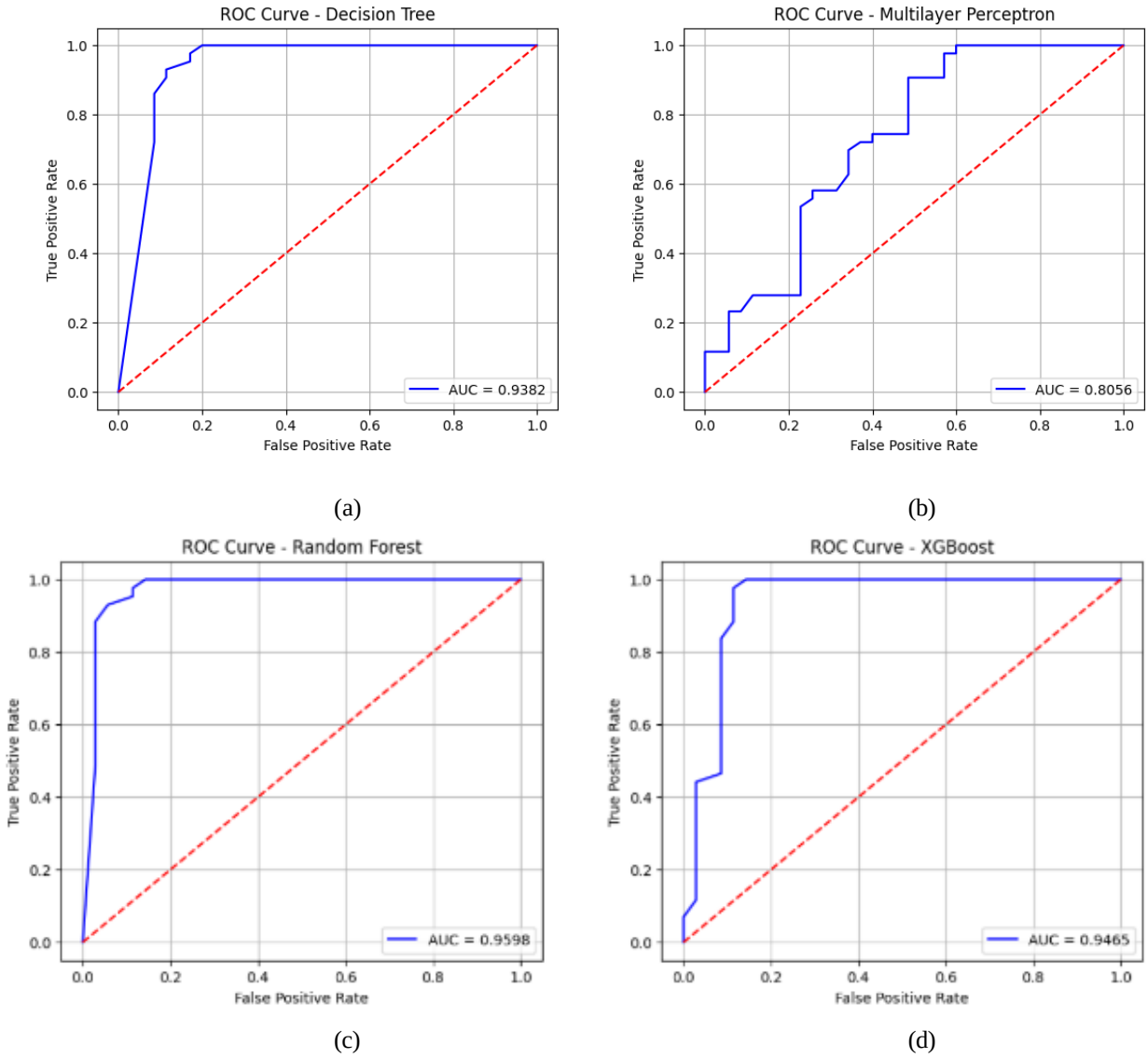

Fig. 3. ROC–area under curve of (a) MLP, (b) RF, (c) DT, and (d) XGB on Dataset 2

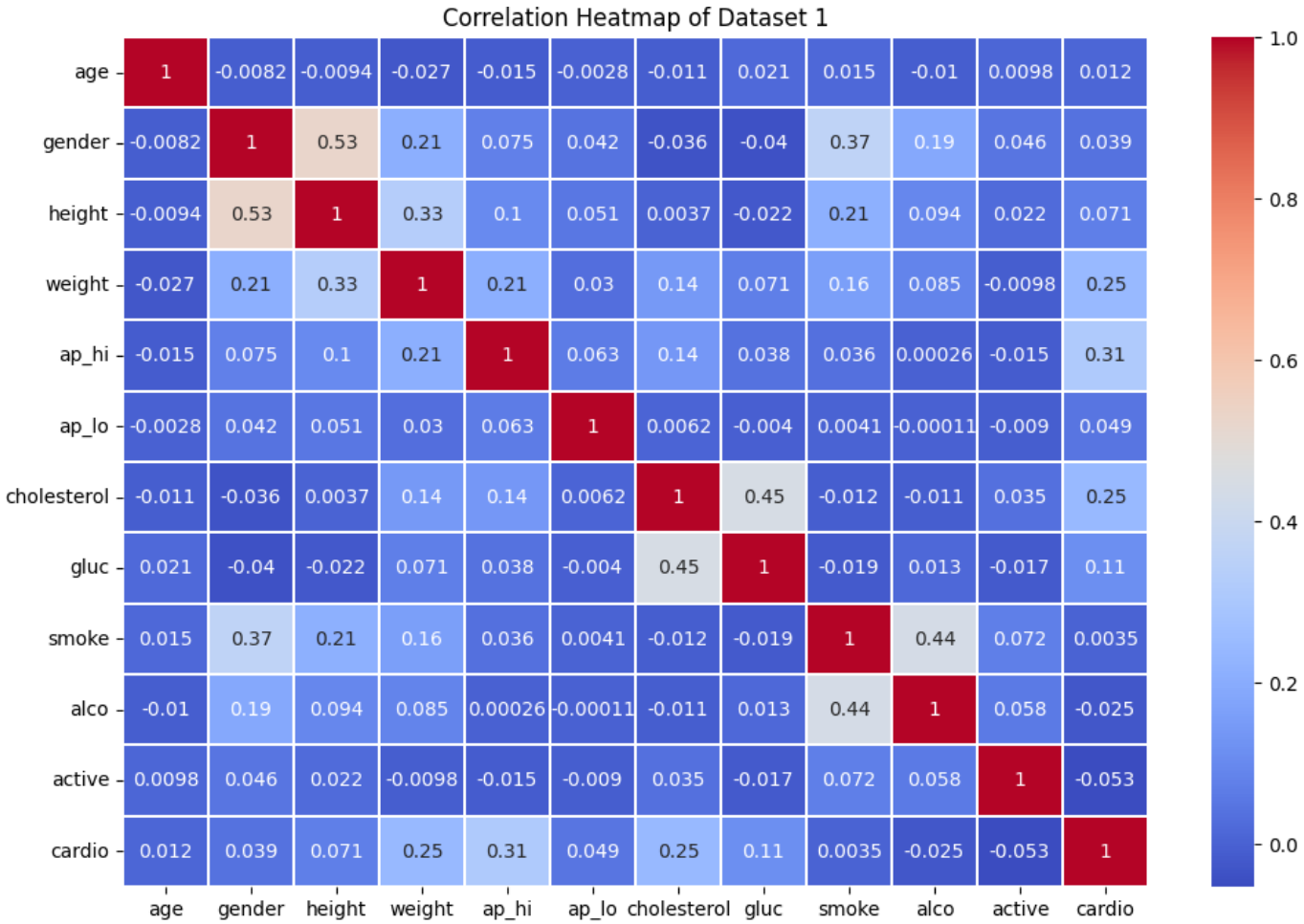

Fig. 4. Correlation Heatmap of Dataset 1

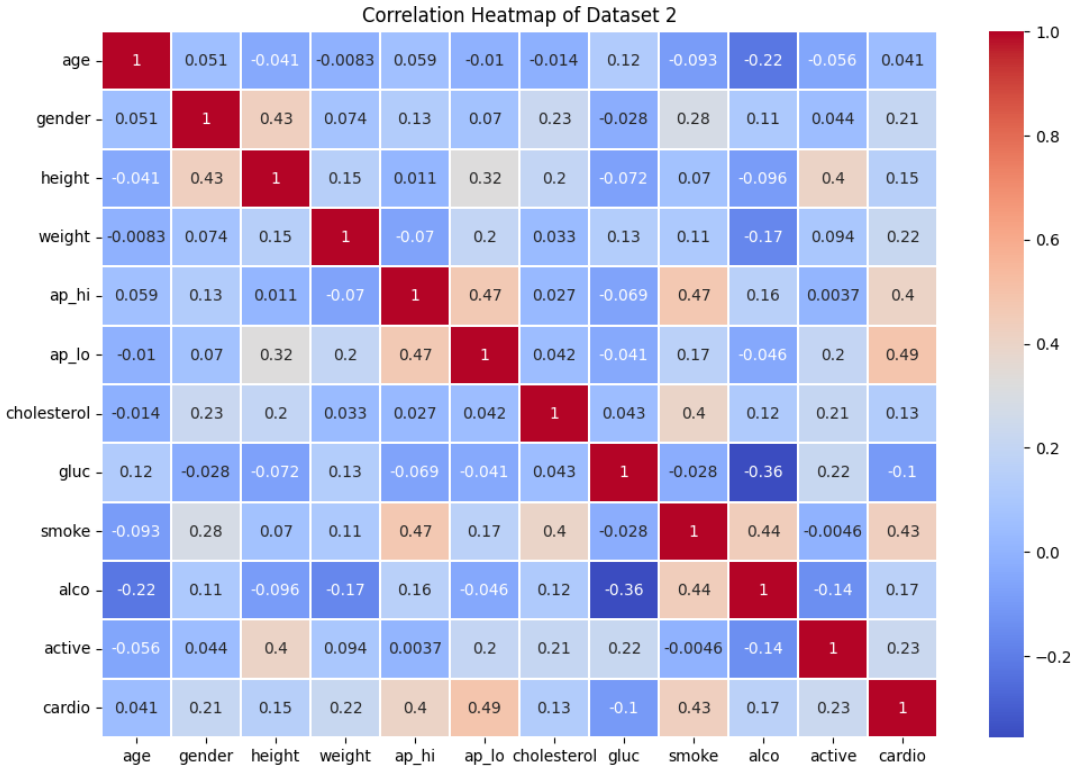

Fig. 5. Correlation Heatmap of Dataset 2

TABLE VI
PERFORMANCE OF ML MODELS WITH AND WITHOUT CROSS-VALIDATION (DATASET 1)

Model	Accuracy (%)		Precision (%)		Recall (%)		F1-Score (%)		AUC
	Without CV	With CV	Without CV	With CV	Without CV	With CV	Without CV	With CV	
DT	83.33	82.74	68.18	69.75	44.77	48.97	54.05	57.54	0.7406
RF	84.64	83.23	75.00	74.03	44.77	45.89	56.07	56.65	0.7577
MLP	84.31	83.15	73.17	70.28	44.77	51.02	55.55	59.12	0.7396
XGB	84.96	83.23	76.92	74.03	44.77	45.89	56.60	56.65	0.7395

TABLE VII
PERFORMANCE OF ML MODELS WITH AND WITHOUT CROSS-VALIDATION (DATASET 2)

Model	Accuracy (%)		Precision (%)		Recall (%)		F1-Score (%)		AUC
	Without CV	With CV	Without CV	With CV	Without CV	With CV	Without CV	With CV	
DT	91.03	86.50	86.00	90.31	100.00	88.50	92.47	89.39	0.9392
RF	91.03	89.07	86.00	89.90	100.00	93.50	92.47	91.67	0.9598
MLP	75.64	71.70	76.09	72.05	81.40	91.50	78.65	80.62	0.8056
XGB	88.46	89.71	82.69	90.78	100.00	93.50	90.53	92.12	0.9465

V. GRAPHICAL USER INTERFACE FOR CARDIOVASCULAR DISEASE PREDICTION

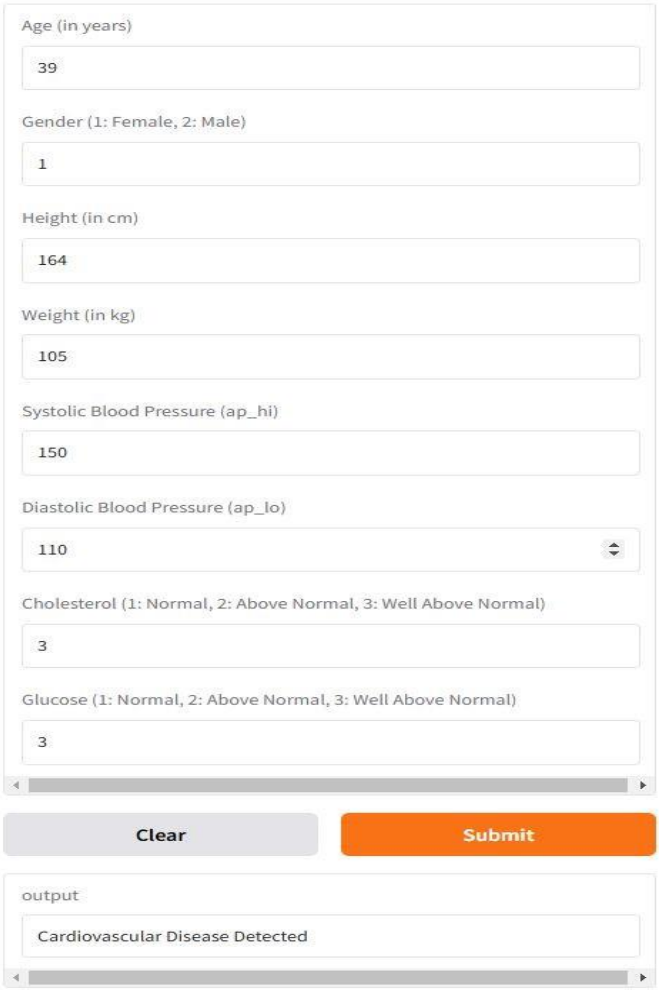

Fig. 6. CVD Prediction in GUI

For enhance the accessibility and usability of the trained XGBoost model for cardiovascular disease (CVD) prediction, a Graphical User Interface (GUI) was developed using the Gradio library. This interface allows users to input basic health parameters such as age, gender, height, weight, systolic and diastolic blood pressure (ap_hi, ap_lo), cholesterol level, and glucose level. These inputs are preprocessed through several steps: age is converted from days to years and binned into 5-year intervals for focused analysis on individuals aged 30 to 40 years; Body Mass Index (BMI) is calculated using standard formula and categorized into six ranges representing different weight categories; and Mean Arterial Pressure (MAP) is computed using the formula $(2 \times ap\_lo + ap\_hi) / 3$ and categorized into meaningful intervals that align with blood pressure severity levels. The processed input is then passed to the XGBoost model, which has been trained on a refined dataset after feature selection and hyperparameter tuning, ensuring optimal performance. The model outputs a binary classification indicating whether a patient is likely to have cardiovascular disease. The result is displayed in a clear and interpretable format—either "Cardiovascular Disease Detected" or "No Cardiovascular Disease"—making it simple for medical practitioners and patients

to understand the outcome. The GUI eliminates the need for technical expertise, making machine learning models more approachable and effective for non-technical users, especially in primary healthcare settings. Furthermore, this tool can be integrated into clinical systems to assist in rapid screening, raise early alerts for at-risk individuals, and support timely medical intervention, ultimately contributing to the reduction of CVD-related morbidity and mortality in the 30–40 age group.

VI. CONCLUSIONS

The primary objective of this research was to predict heart disease by applying various machine learning models on real-world data. Four different algorithms—Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), and XGBoost (XGB)—were evaluated using two distinct datasets, both with and without cross-validation, to assess their accuracy and reliability. In Dataset 1, the XGBoost model delivered the highest accuracy at 84.96% without cross-validation and remained consistent with 83.23% when validated. It also led in precision (76.92%) and F1-score (56.60%), along with a solid AUC of 0.80. On Dataset 2, all models showed enhanced results, with XGBoost again performing strongly—achieving 89.71% accuracy, 90.78% precision, 93.50% recall, and 92.12% F1-score, supported by a high AUC of 0.95. Although Random Forest yielded the highest AUC (0.96), XGBoost consistently demonstrated better balance across all evaluation metrics. These findings confirm XGBoost as the most robust and effective model for heart disease prediction among those tested, providing reliable performance across different datasets.

Despite the promising outcomes, this research has a few limitations that should be acknowledged. The dataset did not include several potentially influential features such as family history, dietary patterns, physical activity levels, medication usage, and genetic or environmental factors, which could further improve the accuracy and generalizability of the model. Moreover, the study focused exclusively on individuals aged 30–40, limiting the scope of applicability to other age groups. The trained models have not yet been clinically validated or tested in real-time medical settings, which may affect their practical usability in healthcare systems.

For future work, expanding the dataset to include more physiological, genetic, and behavioral factors could enhance the robustness and precision of predictive models. Incorporating real-time health data from wearable devices may enable dynamic risk assessment and early warnings. Using deep learning techniques, such as convolutional or recurrent neural networks, could improve performance by capturing complex, non-linear patterns in larger datasets. Collaboration with healthcare professionals for model validation and deployment is essential to ensure clinical relevance and accuracy.

REFERENCES

[1] H. F. El-Sofany, “Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques,” *IEEE Access*, vol. 12, Aug. 2024.

[2] B. P. Doppala, D. Bhattacharyya, M. Chakkravarthy, and T.-h. Kim, ”A Hybrid Machine Learning Approach to Identify Coronary Diseases Using Feature Selection Mechanism on Heart Disease Dataset,” Distributed and Parallel Databases, Springer, Mar. 2021.

[3] N. Chandrasekhar and S. Peddakrishna, “Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization,” Processes, MDPI, vol. 11, Apr. 2023.

[4] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, ”Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare,” IEEE Access, vol. 8, Jun. 2020.

[5] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," Algorithms, MDPI, vol. 16, Feb. 2023.

[6] K. Drozdz˙ et al., "Risk Factors for Cardiovascular Disease in Patients with Metabolic-Associated Fatty Liver Disease: A Machine Learning Approach," Cardiovascular Diabetology, vol. 21, 2022, p. 240.

[7] H. S. N. Murthy and M. Meenakshi, "Dimensionality Reduction Using Neuro-Genetic Approach for Early Prediction of Coronary Heart Disease," in Proc. Int. Conf. Circuits, Communication, Control and Computing, Bangalore, India, Nov. 2014, pp. 329–332.

[8] E. J. Benjamin et al., "Heart Disease and Stroke Statistics—2019 Update: A Report from the American Heart Association," Circulation, vol. 139, 2019, pp. e56–e528.

[9] V. Shorewala, "Early Detection of Coronary Heart Disease Using Ensemble Techniques," Informatics in Medicine Unlocked, vol. 26, 2021,

[10] 100655.

[11] D. Mozaffarian et al., "Heart Disease and Stroke Statistics—2015 Update: A Report from the American Heart Association," Circulation, vol. 131, 2015, pp. e29–e322.

[12] Purushottam, K. Saxena, and R. Sharma, "Efficient Heart Disease Prediction System," Procedia Computer Science, vol. 85, 2016, pp. 962–969.

[13] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, vol. 7, 2019, pp. 81542–81554.

[14] R. Waigi, S. Choudhary, P. Fulzele, and G. Mishra, "Predicting the Risk of Heart Disease Using Advanced Machine Learning Approach," European Journal of Molecular & Clinical Medicine, vol. 7, 2020, pp. 1638–1645.

[15] K. V. and J. Singaraju, "Decision Support System for Congenital Heart Disease Diagnosis Based on Signs and Symptoms Using Neural Networks," International Journal of Computer Applications, vol. 19, 2011, pp. 6–12.

[16] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction Using Machine Learning Techniques," SN Computer Science, vol. 1, 2020, p. 345.

[17] "Cardiovascular Disease, "Wikipedia, accessed on 15 May 2024, https:

[18] //www.wikipedia.com.