IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier xx.xxxx/ACCESS.2024.xxxxxxx

Multi-Faceted Approach to Cardiovascular Risk Assessment by Utilizing Predictive Machine Learning and Clinical Data in a Unified Web Platform

KHADIZA AKTHER[1], MD. SAIDUR RAHMAN KOHINOOR[2] (Member, IEEE), BUSHRA SIDDIKA PRIYA[1], MD. JAMANER RAHAMAN[1], MD. MAHFUZUR RAHMAN[2, 4], MD. SHAFIULLAH[3, 5] (Senior Member, IEEE)

[1]Department of Computer Science and Engineering, Leading University, Sylhet 3112, Bangladesh
[2]Department of Information and Computer Science, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia
[3]Control & Instrumentation Engineering Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia
[4]Interdisciplinary Research Center for Intelligent Secure Systems (IRC-ISS), King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia
[5]Interdisciplinary Research Center for Sustainable Energy Systems (IRC-SES), King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia

Corresponding author: Md. Saidur Rahman Kohinoor (e-mail: g202391630@kfupm.edu.sa).

ABSTRACT Cardiovascular diseases (CVD) persist as a formidable global health challenge, underscoring the imperative for advanced early detection mechanisms. The evolution of computational methods within healthcare has paved the way for transformative applications of machine learning, offering solutions that enhance diagnostic accuracy and contribute to the SDG-3; Good Health and Well-Being. This study aims to identify an algorithm with consistent performance across diverse datasets and integrate it into a comprehensive and user-centric approach to heart disease prediction. The investigation includes an examination of eight machine learning algorithms, three deep learning algorithms, and four heterogeneous datasets from the Kaggle. The predictive performance of these algorithms is assessed through measures that include Precision, Recall, F1 score, Accuracy, and Area Under the Curve (AUC). A Principal Component Analysis (PCA) feature engineering approach is presented to boost predictive performance. An alternative feature selection method, Lasso, was explored, with PCA emerging as the optimal choice for accuracy in the given datasets. As such, the XGBoost algorithm with PCA achieves an impressive accuracy rate and F1 score of around 99% along with an excellent 97% AUC rate in disease prediction on the other dataset. The selected XGBoost model is integrated into a user-friendly web application, providing a holistic platform for heart disease management. Furthermore, we recommended an RPA, IoTM, and AI-based tailored solution to make our web application more reliable, which we have proven in our study is attainable.

INDEX TERMS SDG-3, Heart Disease Detection, Feature Engineering, Personalized Healthcare Solution, XGBoost.

I. INTRODUCTION

A condition known as heart disease begins when the heart cannot circulate sufficient blood to cover the requirements of the body [1]. The World Heart Federation reported in a recent study that cardiovascular disease (CVD) is responsible for one in three deaths [2]. Annually, 17.5 million people pass away from heart disease [3]. Cardiovascular disease is separated into multiple types, all of which have unique symptoms and treatments. The most prevalent causes of heart disease are overbearing blood pressure, diabetes, obesity, smoking, and other factors [4]. Heart disease treatment is most effective in the beginning stages of the illness, therefore early detection is crucial [5]. The current work desires to create a predictive web application solution that incorporates deep learning and machine learning techniques to assist patients with convenient and effective at-home cardiac testing, allowing them to access information about their heart's status at any time and anywhere. The rapid advancement of ML and DL techniques has facilitated research in many fields, including medicine. Machine learning (ML), which is widely suggested as it can extract more accurate and efficient data from enormous datasets, makes heart disease

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3436020

prediction easier. [6]. Applications of machine learning (ML) reduce hospital mistakes while enhancing health policy, early identification, illness prevention, and preventing fatalities at hospitals. Deep learning outperforms earlier technological approaches in processing complicated data, identifying the key aspects of multi-dimensional data, responding quickly to unstructured input, and developing more accurate classification strategies [7]. Machine learning and deep learning-based AI significantly analyze medical data, identifying correlations and forecasting early patient risk for conditions like heart disease and cancer [8]. Enhancing the effectiveness of remote monitoring requires automation. The medical sector applies robotic process automation, to automate many tasks [9]. The Internet of Things (IoT) associated with healthcare is becoming more vital for tracking and admitting health-related matters. The IoMT connects wearable sensor devices via sophisticated medical gadgets, ensuring continuous monitoring of patient health information and tailored treatment. [10]. Every patient wants to get their heart tested frequently to learn more about its status, but this takes time. Numerous exorbitant diagnostics are carried out to reveal the current condition of the heart, including angiography, ECGs, and chest X-rays. At-home cardiac testing based on certain factors will be more effective and convenient. In the field of deep learning, machine learning, AI, RPA, and IoMT for heart failure detection, this work made the following notable contributions:

- This empirical study systematically evaluates and benchmarks 8 classical machine learning and 3 deep learning models on four different cardiovascular datasets to produce a comprehensive performance baseline.
- Investigated a principal component analysis (PCA)-based feature engineering approach that significantly outperformed traditional feature selection methods, boosting predictive accuracy and measurement matrices for cardiovascular disease risk assessment.
- Identified the XGBoost algorithm coupled with PCA feature engineering as the optimal solution on benchmark datasets.
- Translated the high-performing XGBoost-PCA model into an intuitive, user-friendly web application. It democratizes access to cardiovascular analytics while implementing data handling, visualization, and clinical decision-support capabilities.
- Finally, proposed an end-to-end digital healthcare framework synergizing ML, RPA, IoMT, and AI technologies to architect a holistic solution for personalized, proactive, and pre-emptive cardiovascular disease management.

Additional sections within this article are as follows: Section II contains the related works. In Section III, we examine the machine learning and deep learning techniques used for prediction. The structured methodology of this research is explained in Section IV, while Section V discusses the study's experimental results. Section VI covers the web deployment. Sections VII and VIII present our proposed solution and engage in a detailed discussion. Finally, the research article's conclusion is covered in Section IX.

II. RELATED WORKS

This section analyzes the literature that is pertinent to the research project we have in consideration. Analysis is done on the earlier research that was done to predict cardiovascular disease. A comparative discussion of the relevant study findings and proposed techniques is included.

Cardiovascular disease continues to be a big concern around the entire globe, according to the states that were the focus of earlier investigations. Machine learning (ML) methods have been shown to be incredibly useful forecasters throughout the space of cardiovascular research [11]. Due to the growing amount of bioinformatic data and the wide range of clinical data, DL technology is increasingly being employed in scientific research, particularly in sickness prediction, and is receiving more attention for development [12].

The authors in [13] present a predictive analytics-based strategy to locate the variables that impact heart disease using KNN and DT on the Kaggle dataset. KNN exceeded DT in this investigation. KNN obtained an accuracy of 90.70%, whereas DT accomplished an accuracy of 86.56%. Another study made use of the UCI dataset and four algorithms: LR, DT, SVM, and KNN, to diagnose cardiac disease. Logistic regression reached an average accuracy of 87.5%, which is considered minimal [14].

M. J. Gaikwad et al. constructed five algorithms on the UCI dataset, where the train-test ratio was respectively 75% and 25% over the 303 sample. SVM produced the greatest accuracy, at merely 82% [15]. The study [16] used the Random Forest algorithm to anticipate heart illness. The dataset was taken from Kaggle and has 303 samples and 14 attributes. A 10-fold cross-validation strategy was used to divide the data set into training and testing groups, with 80% for training and 20% for testing. The RF model had 90.6% sensitivity, 82.7% specificity, and 86.9% accuracy.

Previous research [17] used five machine learning methods such as DT, RF, LR, NB, and SVM to successfully estimate CVD patients. The data were collected from Lady Reading Hospital and Khyber Teaching Hospital in Pakistan, with a sample size of 518. The recursive operating characteristic curve (ROC) was used to quantify accuracy. The RF algorithm has the highest accuracy, sensitivity, and recursive operating characteristic curve, with 85.01%, 92.11%, and 87.73%, respectively. Another study by N.Mohan et al. predicts cardiac illness based on medical characteristics using four machine learning algorithms: RF, KNN, LR, and NB. Data was gathered from Kaggle, with the LR algorithm scoring the highest accuracy at 90.2% [18]. A similar inquiry done by C. Dhanamjayulu et al. developed a novel approach that uses machine learning techniques such as SVM, KNN, LR, DT, and NB to reveal crucial characteristics, leading to increased precision in the assumption of cardiovascular sickness. RF obtained 95.5% accuracy with an F-measure of 0.95% [19].

The article [20] presents a heart disease detection method employing KNN, RF, and NB models on the Kaggle cardiovascular dataset. The data was pre-processed using the mean value approach to accommodate missing values, and characteristics were minimized by the info-gain feature selection technique. The Random Forest classifier was the most profitable, with an accuracy value of 95.63%. The author in [21] applied a total of three ML algorithms, namely LR, RF, and KNN, to estimate heart disease. They examined their forecasting potential leveraging the UCI repository's Heart Disease Dataset, with KNN exhibiting the best accuracy of 88.52%. Another examination was carried out by Rajalakshmi et al. Actually, the KNN model once again yielded the best accuracy of all models examined, at 85.71% [22].

A hybrid deep learning model created using RNN and GRU was presented in [23], and the UCI online repository's Cleveland Heart Disease Dataset was used by them. Through the Adam optimizer, the optimal learning rate was chosen. This article addressed the problem of data imbalance using the SMOTe method. Compared to other RNN-based current models, the proposed model has the best accuracy rate, at 98.68%. In [24], the Cleveland Heart Disease Dataset, containing 303 instances and 14 attributes, was used to investigate deep neural networks using a variety of weight initialization approaches and optimization algorithms. Next, the findings are contrasted with the performance of various classifiers, specifically SVM, KNN, and NB. The results showed that DNN's accuracy was 81.9%. The SVM classifier outperforms other models with an accuracy of 86.2%.

A. A. Almulihi et al. improved the accuracy of cardiovascular disease prediction by using a novel ensemble stacking model with a meta-learner (SVM) that integrated two previously trained hybrid CNN-LSTM and CNN-GRU models. Recursive feature elimination (RFE) was employed on two datasets related to heart disease that were obtained from Kaggle. Dataset 1: Heart Disease (attribute 18, sample 319796), and dataset 2: Cleveland Dataset (attribute 14, instance 1026) were the most relevant features that were chosen. They optimized the hybrid models using grid-search and the Bayesian optimizer at 97.17%; the suggested model has the greatest accuracy found in the Cleveland dataset [24]. In another study by [25], DNN was utilized to predict cardiac disease, with the aim of improving its effectiveness and precision in anticipating the probability of a heart attack. They use a few different classification methods, including LR, KNN, SVM, NB, and RF. They use Talos optimization to deploy a deep learning neural network (DNN), and Talos outperforms other optimizations in terms of accuracy at 90.78%.

M. M. Rahman et al. presented an automated web-based prediction approach using a dataset of 1026 instances and 14 attributes from Kaggle. There were eight algorithms used: XgBoost, CNN, LR, SVM, AdaBoost, DT, NB, and RF. The research indicates that the RF and DT methods have 99% accuracy when compared to other algorithms. Less categorization error at 0.0121% was found in the decision tree [26]. The study [27] used SVM, KNN, and LR machine learning algorithms to build a medical test web application to detect serious conditions such as breast cancer, heart disease, and diabetes. They discovered that LR had the highest accuracy for breast cancer at 94.55%, while KNN had the best accuracy for heart disease at 83.84%. A similar study by [28] attempted to construct a web-based multiple illness prediction system employing KNN, SVM, DT, RF, LR, and the Gaussian-NB method, where they used a dataset from Kaggle. They found the best accuracy among the six models for diabetes using RF at 99.5% and for heart disease using SVM at 86.41%. Fahim et al. developed a web-based application to detect cardiovascular illness in people at an early stage using DT, KNN, RF, and XGBoost on a dataset acquired from Kaggle (70,000 instances, 12 features, including a target). Xgboost clearly beats other ML approaches, having an accuracy of 73.72%, implying that performance accuracy is lacking [29].

A study by D. B. Rawat et al. investigates the possibilities of AI in healthcare, focusing on individualized therapy and better illness diagnostics. It emphasizes the importance of addressing issues like data privacy, ethics, and regulation [30]. In [31], they identified the fundamental CPHS functional needs as well as non-functional criteria. Secondly, the study examined a three-layer design for IoT-based systems for healthcare. The paper highlights security vulnerabilities in the Internet of Things architecture, proposes mitigation measures utilizing AI and non-AI methodologies, and gives a framework for developing secure, scalable, and personalized healthcare services. Another study by Shwethaet et al. used robotic process automation (BOT) to monitor and contact 120 cardiac patients at remote locations using the KNN approach. The data is acquired via wearable sensors, presented in Excel, and performed using UiPath, allowing for anomaly detection and automatic data collection [9]. Previous research has shown that several kinds of models continue to deliver varied forecast scores. Hence, decreasing dimensions and feature engineering may boost data picking, which leads to higher predictability. However, the overall studies were not implemented in real-life scenarios to deploy as a personalized healthcare solution that can assist in a user-end nature.

During this work, we focused on finding a suitable model that performs better irrespective of the dataset. As such we incorporated 8 state-of-the-art machine learning and 3 deep learning methodologies, as well as the PCA feature selection technique. In contrast to the prior investigation's progress score, we improved the accuracy and other key evaluation matrices of our proposed study to deploy the model in a web-based solution.

III. BACKGROUND KNOWLEDGE

This section delves into the selected machine learning and deep learning approaches in terms of predicting heart disease. Also, it elucidates the operational principles and fundamental terminology inherent in these models. Our investigation specifically assesses eight advanced machine learning and three deep learning models to enhance the accuracy of heart disease predictions.

A. LOGISTIC REGRESSION

Logistic Regression is a widely used supervised machine learning technique that is utilized for both regression and classification tasks [1]. This approach accurately represents the link between input properties and outcomes, making it appropriate for predicting continuous values and classifying data. The logistic regression approach uses probability to predict the labeling of categorical data [32]. The logistic function maps the linear combination of input features and their weights to a probability in the (0, 1) range. The logistic function is defined as follows:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{1}$$

In Eq. 1, the linear combination z is expressed as: $z = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$. The model forecasts the probability of the positive class using $P(y = 1 \mid x) = \sigma(z)$ and probabilities are translated into binary outcomes by applying a threshold.

B. DECISION TREE

Decision Tree represent a method for supervised classification [33]. It is a hierarchical structure comprising nodes and branches, originating from the root and concluding at the leaf, where internal nodes extend into various branches [34]. The DT structure, with leaf nodes representing final decision outcomes, uses impurity measures and information gain to construct interpretable predictive models. Decision trees possess value in diverse fields by effectively examining input data, recognizing patterns, and rendering well-informed decisions through their hierarchical structures, contributing to the process of making predictions and guiding strategic decisions.

C. RANDOM FOREST

Another supervised machine learning method is the Random Forest [35], also an ensemble technique based on trees, designed to overcome the shortcomings of conventional classification and regression tree approaches [36]. It enhances predictive accuracy by constructing multiple decision trees on bootstrapped datasets with random feature subsets. It is renowned for its robustness and capacity to handle complex, high-dimensional datasets with various features. The algorithm's strength lies in the combination of diverse trees, reducing overfitting and enhancing robustness. It can manage extensive datasets, intricate associations, and the presence of missing values.

For Classification:

$$\hat{y} = \text{argmax}_c \sum_{i=1}^{N_{\text{trees}}} 1(y_i = c) \tag{2}$$

For Regression:

$$\hat{y} = \frac{1}{N_{\text{trees}}} \sum_{i=1}^{N_{\text{trees}}} y_i \tag{3}$$

Where, y_i in Eqs. 2 and 3, is the output of an individual decision tree, N_{trees} is the number of trees and $\hat{y}$ is the final predicted output.

D. SUPPORT VECTOR MACHINE

The Support Vector Machine (SVM) [37] stands as a robust supervised machine learning algorithm applied for tasks involving both classification and regression, categorizing data into distinct classes and forecasting tasks [38]. It aims to identify the optimal hyperplane for separating data points in a high-dimensional space, formulated mathematically using a decision function. The decision function of an SVM is expressed as f(x) = w * x + b, where w is the weight vector, x is the input feature vector, and b is the bias term.

E. NAIVE BAYES

Naive Bayes is predominantly employed for tasks related to classification. The NB approach is probability-based, which means that the classifier predicts according to the likelihood of dataset variables [39]. It assumes feature independence and calculates probabilities for each feature and class during training, which are then utilized for making predictions. Despite its simplicity, Naive Bayes is effective in text classification, spam filtering, and other applications.

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)} \tag{4}$$

where, $P(C|X)$ represents the probability of class C given features X, and $P(X|C)$ denotes the likelihood of observing features X given class C.

F. K-NEAREST NEIGHBORS

K-Nearest Neighbors (KNN) is a supervised machine learning technique utilized for tasks involving classification and regression. [40]. The KNN classification is a widely used algorithm that determines the final classification output by measuring the distance between the test and training samples [41]. The algorithm's flexibility lies in its non-parametric and instance-based nature, allowing it to adapt to various data distributions. The algorithm predicts new data points based on the average value of their k-nearest neighbors using a distance metric like Euclidean distance.

$$y_q = \text{mode}(y_{i_1}, y_{i_2}, \ldots, y_{i_k}) \tag{5}$$

In Eq, 5 [1], where y_q denotes the predicted label and $y_{i_1}, y_{i_2}, \ldots, y_{i_k}$ denotes the indices of the k nearest neighbors.

G. GRADIENT BOOSTING

Gradient Boosting is an ensemble method [42] used for regession and classification tasks. Gradient boosting is a technique for combining the predictions [43] of multiple weak learners, such as decision trees, to produce a strong predictive model. The final model is produced in stages, with each new tree correcting the mistakes of the preceding ones. Its application involves solving in finance for risk assessment,

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3436020

healthcare for disease prediction, and web search ranking for improved relevance. XGBoost, LightGBM, and AdaBoost are popular gradient-boosting implementations.

$$F(x) = \sum_{m=1}^{M} f_m(x)$$

where, $F(x)$is the final prediction and $f_m(x)$ is the prediction of the m-th tree, v is the learning rate.

$$F_m(x) = F_{m-1}(x) + v.h_m(x)$$

H. EXTREME GRADIENT BOOSTING

Extreme Gradient Boost is a machine learning method that employs both sparse and dense matrix inputs, producing outputs in the form of integers for classification or real values for regression [44]. The package includes a linear model solver and a tree learning algorithm, which supports various objective functions like regression, classification, and ranking [45]. XGBoost is a decision tree ensemble that uses weighted combinations of predictors [46]. XGBoost is a weak classifier decision tree that trains a new tree to learn from the previous tree's errors.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{6}$$

In Eq. 6 [1], where $f_k(x_i)$ is the sum of predictions up to the k-th weak learner for the i-th data point.

I. MULTILAYER PERCEPTRON

A Multilayer Perceptron is a type of feedforward neural network composed of interconnected layers of neurons, facilitating the flow of information from input to output. MLP neural networks consist of layers with nodes, each node connects to every node in subsequent layers. They typically have three layers: input, hidden, and output, with linear and nonlinear activation functions [47]. The model is trained on labeled data using backpropagation, with weights and biases adjusted.

$$y = \sigma(W_2\sigma(W_1x + b_1) + b_2) \tag{7}$$

where, W_1 and W_2 are weight matrices, b_1 and b_2 are bias vectors, σ is a non-linear activation function, and x is the input vector.

J. ANN

An Artificial Neural Network (ANN) is a computational model that is derived from the structure and function of biological neural networks. Artificial neural networks (ANN), also known as neural networks (NN), are theoretical or computer models based on biological brain networks [48]. It consists of interconnected nodes organized into layers, including an input layer, one or more hidden layers, and an output layer. The layers of a basic neural network are visible in Figure 1. Each connection between nodes is associated with a weight, and nodes apply activation functions to their inputs. These hidden layers serve to filter out irrelevant input elements that don't significantly influence the output. However, a crucial challenge lies in selecting which neurons to activate and how activation functions make these choices. This process heavily involves the cost function, mathematically expressed as the squared difference between the dataset's actual values and the network's output. ANN is trained using a process called back-propagation to adjust weights and learn from data. During the training phase, the primary objective is to minimize this cost. The movement of data from the input to the output is known as forward propagation. Conversely, when the reverse process occurs, moving from the output back to the input layer, it's called backpropagation [49]. This iterative process of fine-tuning the network helps optimize its performance.

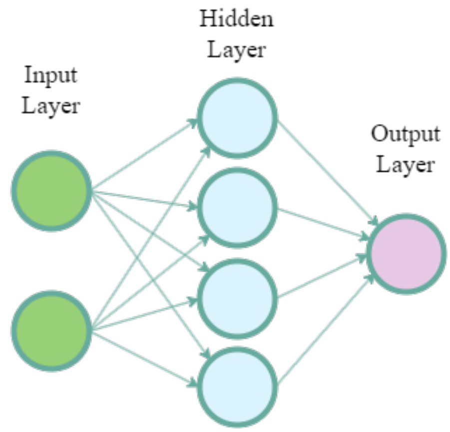


FIGURE 1: Simple Neural Network.

K. DNN

A deep neural network, often known as a feedforward neural network [50]. The deep neural network is a discriminative model that is trainable through the backpropagation algorithm [51]. Deep neural networks use stacked layers with multiple neurons, each combining weight and capacity to magnify input value, each layer can switch on/off, with output from one layer acting as input for the forward direction [52]. Deep neural networks have more hidden layers than artificial neural networks, which have one input and output layer and a maximum of one hidden layer [53]. DNNs use several layers to extract features from high-dimensional and unstructured input for hierarchical comprehension. Because of their ability to capture subtle correlations, they outperform standard models in tasks such as picture and speech recognition, natural language processing, and complex pattern recognition.

IV. STRUCTURED METHODOLOGY

The methodological structure of this study is provided in Figure 2, providing a visual representation of the main actions taken throughout the system's approach. This diagram guides each step of the study's flow, highlighting essential actions for the study's advancement. The paper assesses the effectiveness of our suggested method utilizing Kaggle datasets, offering

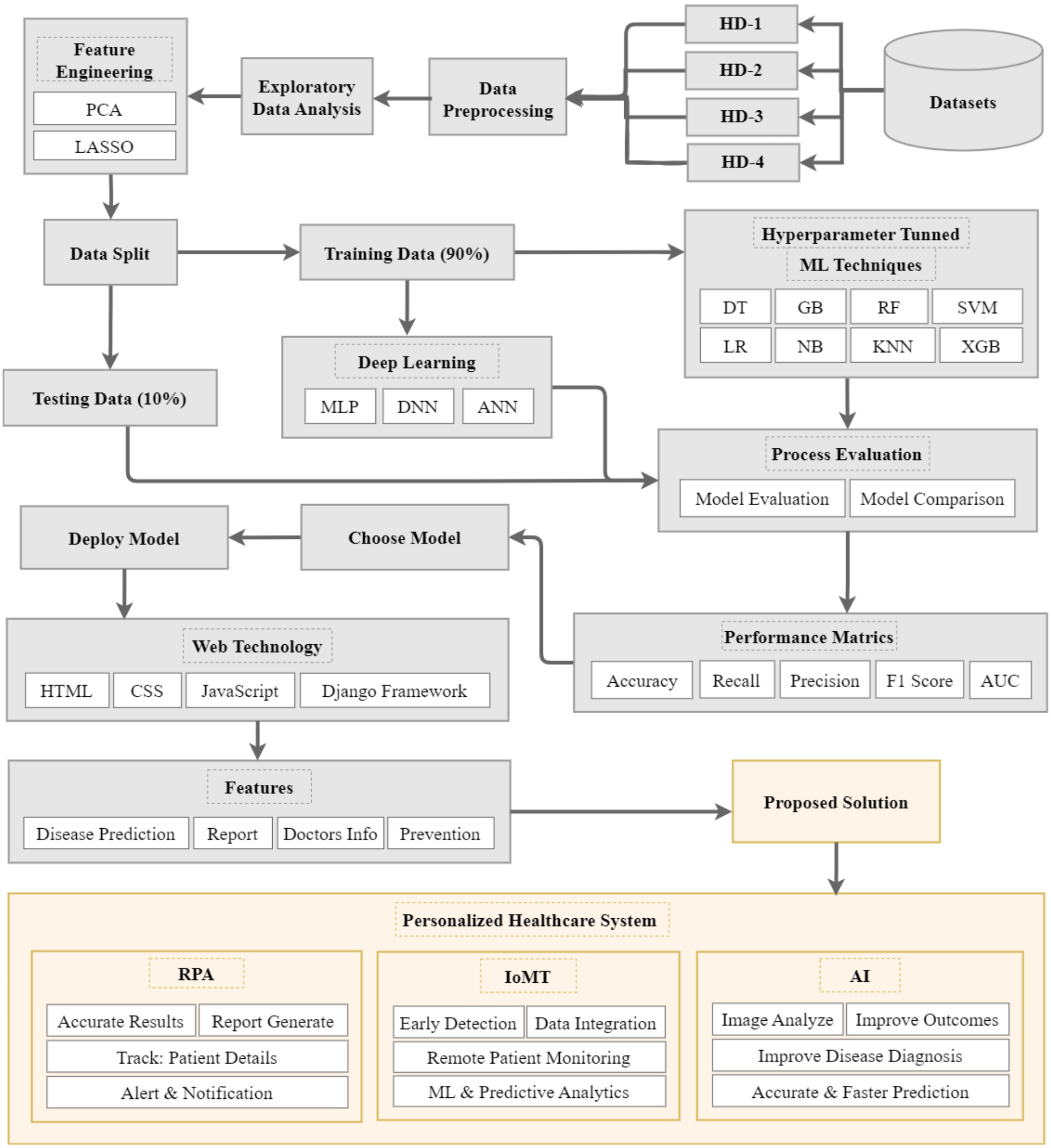


FIGURE 2: **The methodological overview for heart disease prediction.**

sequential stages for a thorough knowledge of its applicability in heart disease prediction. The following steps summarise the proposed methodology's working flow:

Step-1: The initial step is to collect four distinct datasets related to heart disease derived from Kaggle.

Step-2: Data preprocessing steps are taken to manage missing data, remove noise, and scale data using a standard scalar to maintain data quality.

Step-3: The Exploratory Data Analysis is applied to obtain insights into data distribution, correlation, and overall structure of heart disease data.

Step-4: Two feature engineering techniques such as PCA and L1 regularization (Lasso) were applied, among these the PCA technique is proposed to improve machine learning performance scores by detecting and utilizing high-importance features.

Step-5: The datasets are split into training and testing subsets.

Step-6: Machine Learning and Deep Learning models are developed for heart disease prediction, optimizing algorithms based on hyperparameter tuning. The model's performance on a testing dataset is assessed using evaluation matrices,

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3436020

TABLE 1: Descriptions of features in datasets related to the heart.

SL.No	Attributes	Descriptions	Range of Values
1	age	The patient's age expressed in years	29 - 77
2	sex	Sex Indicates gender: 0 denotes female and 1 denotes male.	0, 1
3	cp	Chest pain type	0, 1, 2, 3
4	trestbps	Resting blood pressure (trestbps) measured in mmHg	94 - 200
5	chol	serum cholestoral in mg/dl	126 - 564
6	fbs	fasting blood sugar level of the patient > 120 mg/dL (diabetes) 1 means true; 0 means false.	0, 1
7	restecg	resting electrocardiographic results	0, 1, 2
8	thalach	The maximum heart rate that your heart can reach during strenuous exercise.	71 - 202
9	exang	Angina induced by exercise: 1 = yes; 0 = no.	0, 1
10	oldpeak	ST depression caused by exercise compared to rest.	0 - 6.2
11	slope	ST-T abnormalities, based on the slope of the peak exercise ST segment	1, 2, 3
12	ca	The number of main vessels (0-3) coloured by fluoroscopy.	0, 1, 2, 3
13	thal	Thallium: 0 = null; 1 = fixed defect; 2 = normal; 3 = reversable defect	0, 1, 2, 3
14	target	Outcomes: 0=No chance of heart disease and 1=Chances of heart Disease	0, 1

TABLE 2: Descriptions of features in datasets related to the heart.

SL.No	Attributes	Descriptions	Range of Values
1	age	The patient's age expressed in days	10798 - 23713
2	gender	1 denotes female and 2 denotes male.	1, 2
3	height	The height is specified in centimeters	55-250
4	weight	The weight is specified in kilograms	10-200
5	ap_hi	Systolic blood pressure measured in mmHg	60
6	ap_lo	Diastolic blood pressure measured in mmHg	220
7	cholesterol	Cholesterol levels are categorized as normal, above normal, and well above normal.	1, 2, 3
8	gluc	Glucose levels are categorized as normal, above normal, and well above normal.	1, 2, 3
9	smoke	Whether patient smokes or not	0, 1
10	alco	Alcohol intake	0, 1
11	active	Physical activity	0, 1
12	cardio	The number of main vessels (0-3) coloured by fluoroscopy.	0, 1

which include metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

Step-7: The next step is to build a web application using the Django framework.

Step-8: Finally, the proposed solution integrates a heart disease prediction model into a personalized healthcare system, utilizing AI, RPA, and IoMT for real-time data monitoring and personalized healthcare recommendations.

A. OVERVIEW OF DATASETS

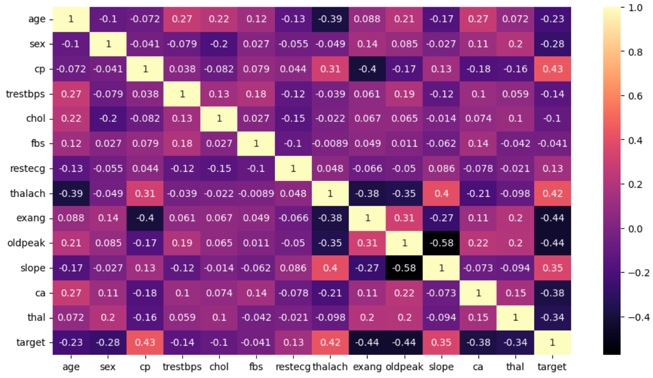
In this research, we utilized four prevalent heart disease datasets sourced from Kaggle: Public Health Dataset [54], Heart Attack Dataset [55], Heart Failure Prediction Dataset [56] and Cardiovascular Disease dataset [57]. The Public Health Dataset and Heart Attack Dataset have 14 feature attributes, while the Heart Failure Prediction Dataset and Cardiovascular Disease Dataset have 12 features, including the 'target' feature. The 'target' features in both datasets denote the patient's heart condition, with 0 indicating no disease and 1 indicating the presence of a disease. The [54], [55], and [56] all have the same features, but the [56] is missing the 'ca' and 'thal' features. Table 1 provides an overview of the feature specifics contained in these datasets and Table 2 provides an overview of the Cardiovascular Disease Dataset with 3 types of input features which are Objective: factual information; Examination: results of medical examination; Subjective: information given by the patient. The [54], [55], [56], and [57] datasets contain 1025, 303, 918, and 70,000 patient records, respectively.

B. DATA PREPROCESSING

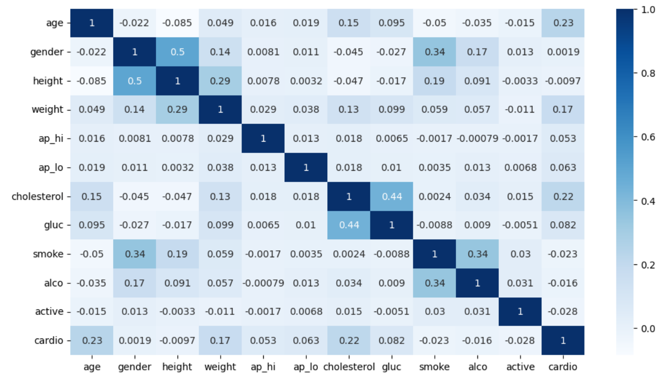
Data preprocessing is a crucial stage in preparing raw datasets for analysis and modeling, aiming to remove noise to enhance model performance and address missing values. Data cleaning and preparation are important for improving machine learning algorithms' precision, and effectiveness, and enabling accurate representation and proper classifier training and testing [58]. In this research, We've used standard scalar to manage missing data and maintain data quality. Also inspected the dataset to identify null values, this has implications for overall performance. Null values are not permitted in the original dataset because the Random Forest technique does not support them. The analysis indicates that the dataset does not contain any null values.

C. EXPLORATORY DATA ANALYSIS

To detect trends in the datasets, we use visualizations such as charts and heatmap graphs using exploratory data analysis. Figure 4, intricately illustrates the distribution of the Public Health Dataset and Heart Attack Dataset based on the gender and target attributes. The Public Health Dataset comprises 1025 patient records, with a detailed breakdown revealing 499 instances of individuals classified as healthy and 526 diagnosed with heart disease. Out of the total, 713 patients are male, with 413 being healthy and 300 having heart disease. On the other hand, there are 312 female patients, among which 86 are healthy, and 226 have heart disease. Let's shift our focus to the Heart Attack Dataset, consisting of 303 patient records. Here, 138 individuals are classified



(a) Common feature datasets



(b) Cardiovascular Disease Dataset

FIGURE 3: The correlation analysis of datasets based on heatmap.

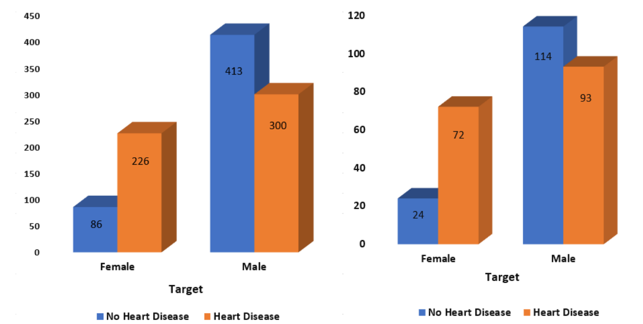


FIGURE 4: The distribution analysis of the Public Health Dataset (left) and Heart Attack Dataset (right).

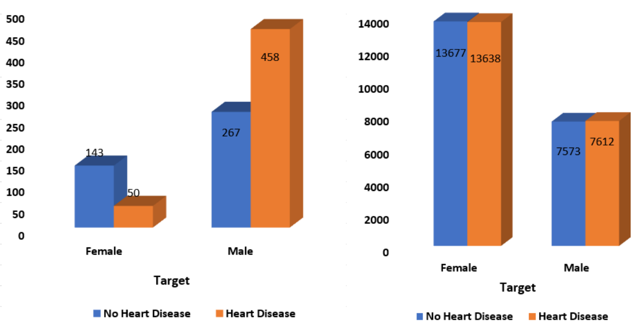


FIGURE 5: The distribution analysis of the Heart Failure Prediction Dataset (left) and Cardiovascular Disease Dataset (right).

as healthy, while 165 have been diagnosed with heart disease. The male cohort, constituting 207 patients, is further categorized into 114 healthy individuals and 93 with heart disease. Simultaneously, the female demographic, totaling 96 patients, is stratified into 24 healthy individuals and 72 with heart disease. Figure 5 depicts a detailed breakdown of the distribution within the Heart Failure Prediction Dataset and Cardiovascular Disease Dataset, categorized by gender and target attributes. Within the Heart Failure Prediction Dataset, among 918 patient records, 410 were identified as healthy, including 267 males and 143 females. Conversely, among 508 patients diagnosed with heart disease, 458 were male and 50 were female. Regarding the Cardiovascular Disease Dataset, we selected 42,500 patient records after addressing the imbalance by removing duplicates from a dataset originally containing 70,000 entries. Of these 15,185 were male, comprising 7573 healthy individuals and 7612 with heart disease. Among the 27,315 female patients, 13,677 were healthy and 13,638 had heart disease. These datasets offer the framework for a full exploratory data study, providing important insights into health patterns across genders.

Exploring the datasets through correlation matrix analysis of Public Health Dataset, Heart Attack Dataset, and Heart Failure Prediction Dataset, as depicted in Figure 3a that reveals a strong positive correlation is observed between chest pain (cp) and maximum heart rate achieved (thalach), while exercise-induced angina (exang) and number of major vessels (ca) exhibit a notable negative correlation. Furthermore, in Figure 3b, the correlation matrix of the Cardiovascular Disease Dataset reveals a positive correlation between age and cholesterol, and a marked negative correlation between glucose (gluc) and physical activity (active). These findings contribute crucial information for subsequent research phases.

D. FEATURE SELECTION TECHNIQUES

In this research endeavor, we confront the challenges of high-dimensional datasets with numerous features, emphasizing the need for a robust feature selection strategy. In this section, we analyzed two feature selection approaches, these are PCA (Principal Component Analysis) and LASSO. Initially, the dataset comprised a total of 14 features. To enhance the PCA mechanism, we innovatively crafted a new feature set, aiming to attain the highest accuracy scores. The PCA feature selection approach was optimized by creating a new dataset with 80% of the total variance in the original data, resulting in the highest accuracy scores. On the other hand, Lasso uses a penalty term to select a subset of features and reduces less significant ones to zero. The method automatically assesses

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3436020

feature importance by tuning the regularization parameter in Lasso regression, ultimately selecting features based on the learned coefficients from the regression process. The study demonstrated that the PCA method exhibited high-performance accuracy in predicting heart disease.

E. DATA SPLITTING

In this research, both datasets related to heart disease are split into training and testing sets for the purpose of assessing the model's performance. The former contains 90% of the data, whereas the latter contains 10% of the data and serves as an independent evaluation of the model's generalization on unseen data. Considering the limited dimensionality of the datasets, this ratio ensures a comprehensive evaluation of the study and increases the dependability of our research findings.

F. EVALUATION MATRICES

The outcomes derived from different models are evaluated using matrices such as accuracy, precision, recall, F1-Score, and AUC for comparison. The equations for accuracy, recall, precision, F1 score, and AUC are provided in Eqs. 8, 9, 10, 11 and 12, respectively. Model evaluation is performed through the utilization of the confusion matrix, yielding four results: TN (True Negative), TP (True Positive), FN (False Negative), and FP (False Positive).

$$Accuracy(Acc) = \frac{tp + tn}{tp + tn + fp + fn} \quad (8)$$

$$Precision = \frac{tp}{tp + fp} \quad (9)$$

$$Recall = \frac{tp}{tp + fn} \quad (10)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (11)$$

Area Under the Curve (AUC): The AUC measures the performance of a binary classification model by analyzing its ROC curve, with higher values indicating better discrimination between positive and negative classes.

$$\text{AUC} \approx \sum_{i=1}^{n-1} \frac{1}{2}(TPR_i + TPR_{i+1}) \times (FPR_{i+1} - FPR_i) \quad (12)$$

V. EXPERIMENTAL RESULTS ANALYSIS

In this part, we delve into the performance evaluation of our study and present key performance indicators including accuracy, precision, recall, F1 scores, and the Area Under the Curve (AUC). Furthermore, we outline the execution of two investigations (PCA and Lasso) carried out on the Public Health Dataset, the Heart Attack Dataset, and the Heart Failure Prediction Dataset. Additionally, we utilized the Cardiovascular Disease Dataset to assess the model's true predictive power.

A. PERFORMANCE EVALUATION WITHOUT FEATURE SELECTION TECHNIQUE.

Table 3 contrasts the outcomes of ML and DL models using the original features of four datasets: the Public Health Dataset, Heart Attack Dataset, Heart Failure Prediction Dataset, and Cardiovascular Disease Dataset. The examination reveals that RF, XGBoost, ANN, DNN, and MLP models have higher accuracy in the Public Health Dataset, and their other performance measures were good on that dataset, although their performance fluctuates in cross-dataset scenarios. But XGBoost from ML and DNN from DL techniques show decent and consistent performance across all four datasets.

Figure 6 displays an accuracy execution comparison analysis of integrated ML and DL models using bar charts. The investigation indicates that DNN and ANN perform optimally with an accuracy of 97.09% and 96.00%, respectively, using the public health dataset. However, they exhibit slightly lower performance in the heart attack dataset, but poor performance in the Heart Failure Dataset. LR and SVM did not perform well in the public health dataset, but they showed good performance in the other two datasets. The gradient boosting model showed the lowest accuracy 77.41% in the heart attack dataset, while its accuracy 88.34% in the public health datasets was insufficient. In the Cardiovascular Disease Dataset, XGBoost and ANN demonstrate slightly better performance compared to other models. However, overall, all models exhibit poor performance relative to other datasets due to the absence of strongly correlated features such as cp, restecg, thalac, and slope, which are critical for predicting heart disease in the other three datasets. Therefore, in this dataset, PCA and LASSO were not employed because of the lack of strongly correlated features. The existing features do not exhibit significant correlations, resulting in ineffective performance. Overall, XGBoost models demonstrate superiority across all datasets.

In Figures 12-15, the XGBoost confusion matrices for four datasets are presented. These matrices assess the performance of the XGBoost model in predicting heart disease across all four datasets by identifying errors or discrepancies. They evaluate the model's predictions against actual outcomes using key metrics: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Statistical metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) are used to analyze different ML classifiers based on these confusion matrices. Figure 9a, 10a and 11a illustrate the ROC curves for all machine learning models used in the Public Health, Heart Attack datasets, and Heart Failure Prediction Dataset, respectively. The greater the area under the curve, the better the result. The score goes from 0 to 1, with a value near 1, suggesting that the algorithm performed well. The graph shows that only RF, MLP, and XGBoost perform well on the public health dataset, with AUC values near one. On the other hand, in heart attack datasets, only MLP outperforms other models with an AUC value of 97.06%. According to the results, it is noticeable

TABLE 3: Comparison of ML and DL models on test data without feature engineering.

Models	Public Health Dataset					Heart Attack Dataset					Heart Failure Prediction Dataset					Cardiovascular Disease Dataset				
	Acc	Prec	Rec	F1	AUC	Acc	Prec	Rec	F1	AUC	Acc	Prec	Rec	F1	AUC	Acc	Prec	Rec	F1	AUC
LR	84.47	92.30	80.00	85.53	93.40	90.32	88.23	93.75	85.53	95.38	88.24	86.27	89.80	88.00	90.70	71.76	69.27	72.91	71.04	78.31
XGBoost	95.15	98.07	92.72	95.32	99.25	87.09	76.47	100.00	86.66	94.54	92.16	88.24	95.74	91.84	95.89	73.76	71.86	74.71	73.25	80.20
RF	95.15	91.22	100.00	95.41	99.92	93.54	100.00	88.23	93.75	95.59	91.18	90.38	92.15	91.26	95.77	72.96	75.74	67.58	71.43	79.62
DT	90.29	85.00	98.07	91.07	96.46	80.64	86.66	76.47	81.25	77.94	85.29	87.50	82.35	84.84	91.56	73.48	75.18	70.12	72.56	79.18
NB	80.58	75.80	90.38	82.45	90.99	83.87	87.50	82.35	84.84	94.96	90.20	90.20	90.20	90.20	94.62	57.62	70.40	26.31	38.30	67.98
SVM	83.49	76.92	96.15	85.47	93.17	90.32	93.75	88.23	90.90	96.22	89.22	88.46	90.20	89.32	90.12	72.40	75.67	66.02	70.52	78.68
KNN	90.29	87.50	84.23	90.74	98.08	83.87	83.33	88.23	85.71	90.97	89.22	91.67	86.27	88.89	91.83	64.14	64.68	62.31	63.47	69.47
GB	88.34	82.25	98.07	89.47	95.93	77.41	91.66	64.70	75.86	86.97	86.27	87.76	84.31	86.00	90.16	73.51	75.19	70.16	7259	79.28
MLP	95.15	98.07	92.72	95.32	99.43	93.54	88.23	100.00	93.75	97.06	87.25	88.24	86.54	87.38	93.46	72.89	72 56	73.05	72.80	79.66
ANN	96.00	95.00	98.00	96.00	99.39	94.00	89.00	1.00	94.00	96.67	87.00	87.00	90.00	88. 00	95.71	74.00	75.00	73.00	74.00	79.78
DNN	97.09	95.00	1.00	97.00	99.47	93.55	89.00	1.00	94.00	95.83	85.87	86.00	88.00	87.00	94.76	73.29	75.00	69.00	72.00	79.64

TABLE 4: The results of comparing ML models on test data were evaluated with PCA technique.

Models	Public Health Dataset					Heart Attack Dataset					Heart Failure Prediction Dataset				
	Acc	Prec	Rec	F1	AUC	Acc	Prec	Rec	F1	AUC	Acc	Prec	Rec	F1	AUC
LR	91.26	94.33	89.28	85.57	96.11	96.77	100.00	94.11	85.98	98.75	90.20	86.27	93.62	89.80	94.39
XGBoost	99.03	100.00	98.15	99.06	100.00	93.55	100.00	88.89	94.12	97.08	91.18	90.20	92.00	91.09	95.19
RF	96.12	96.23	96.23	96.23	99.47	96.77	94.12	100.00	96.97	97.08	90.20	93.62	86.27	89.80	94.73
DT	91.26	95.83	86.79	91.08	97.38	77.42	73.68	87.5	79.99	80.62	86.27	87.76	84.31	86.00	89.91
NB	92.23	95.92	86.79	91.09	95.36	90.32	88.23	8 7.5	79.99	98.33	90.20	95.56	84.31	89.58	94.73
SVM	88.35	84.74	94.34	89.28	95.06	96.77	94.12	1.00	96.97	99.17	89.22	91.67	86.27	88.89	94.23
KNN	92.23	89.47	96.23	92.72	97.49	90.32	88.23	93.75	90.91	98.33	85.29	84.62	86.27	85.44	90.97
GB	90.29	87.71	94.34	90.90	96.83	80.65	72.72	100.00	84.21	95.20	89.22	91.67	86.27	88.89	92.93
MLP	96.12	98.11	94.54	96.29	98.26	96.77	100.00	94.12	96.97	98.75	90.20	86.27	93.62	89.80	94.43

TABLE 5: The results of comparing ML models on test data were evaluated with LASSO technique

Models	Public Health Dataset					Heart Attack Dataset					Heart Failure Prediction Dataset				
	Acc	Prec	Rec	F1	AUC	Acc	Prec	Rec	F1	AUC	Acc	Prec	Rec	F1	AUC
LR	84.47	92.45	80.32	85.96	95.81	93.55	100.00	88.89	94.12	97.08	88.24	88.24	88.24	88.24	91.62
XGBoost	92.23	94.34	90.91	92.59	98.45	90.32	93.75	88.24	90.91	96.25	88.24	86.27	89.79	88.00	95.04
RF	96.12	94.54	98.11	96.30	98.45	93.55	88.89	100.00	94.12	98.75	91.18	90.38	92.16	91.26	94.69
DT	94.17	91.23	98.11	94.54	97.40	90.32	84.21	100.00	91.43	92.71	86.27	91.11	80.39	85.42	94.14
NB	92.23	92.45	98.11	94.54	96.52	96.77	100.00	100.00	91.43	99.58	91.18	90.38	92.16	91.26	95.31
SVM	86.41	83.05	92.45	87.50	95.32	93.55	88.89	100.00	94.12	98.33	90.20	88.68	92.16	90.38	91.16
KNN	90.29	89.09	92.45	90.74	96.04	93.55	88.89	100.00	94.12	98.13	87.25	91.30	82.35	86.60	91.00
GB	91.26	87.93	96.23	91.89	96.98	64.52	59.26	100.00	74.42	98.12	87.25	91.30	82.35	86.60	93.04
MLP	94.17	94.34	94.34	94.34	98.75	87.10	93.75	83.33	88.24	93.33	90.20	88.24	91.84	90.00	94.43

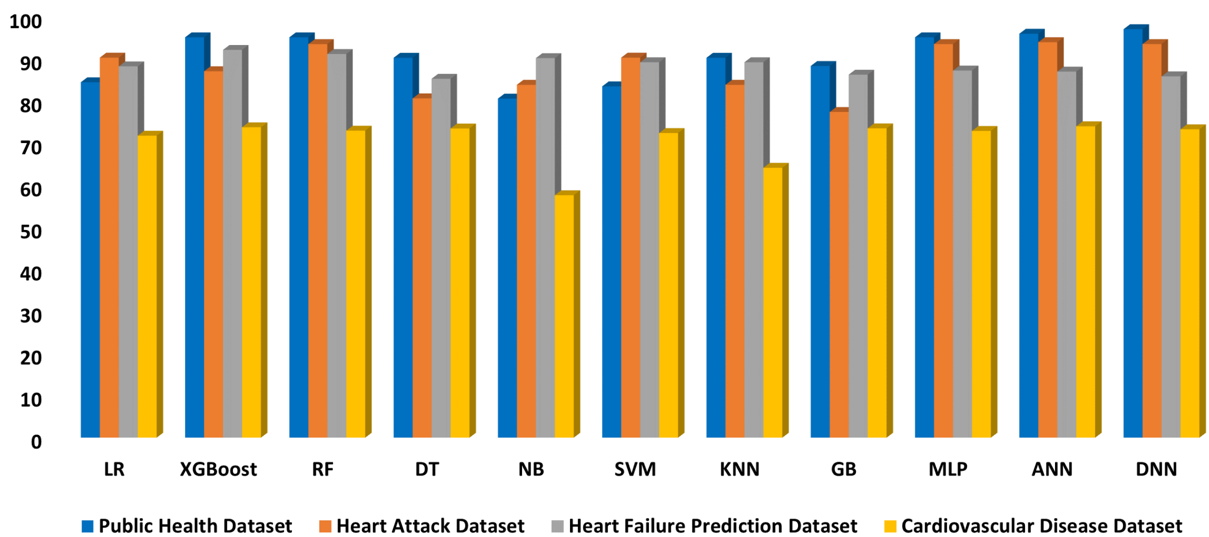


FIGURE 6: **Analyzing the results of ML and DL models through a bar chart comparison without feature engineering.**

that the accuracy of most of the models applied was not impressive when using the original features from both datasets.

Following the approach suggested by Brisimi et al. [59], we performed statistical significance tests comparing the perfor-

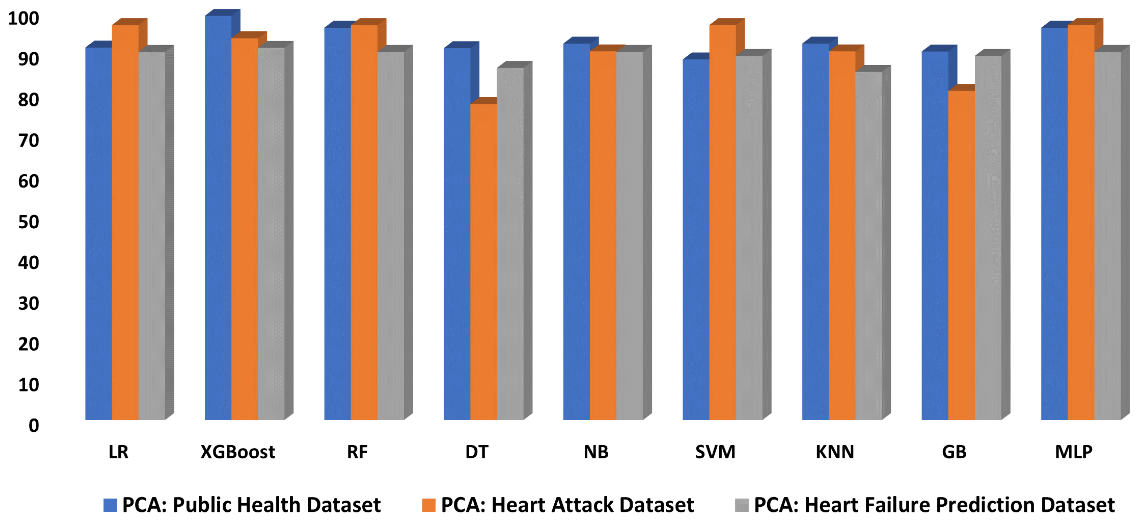


FIGURE 7: **Analyzing the results of ML models applied through a bar chart comparison with PCA technique.**

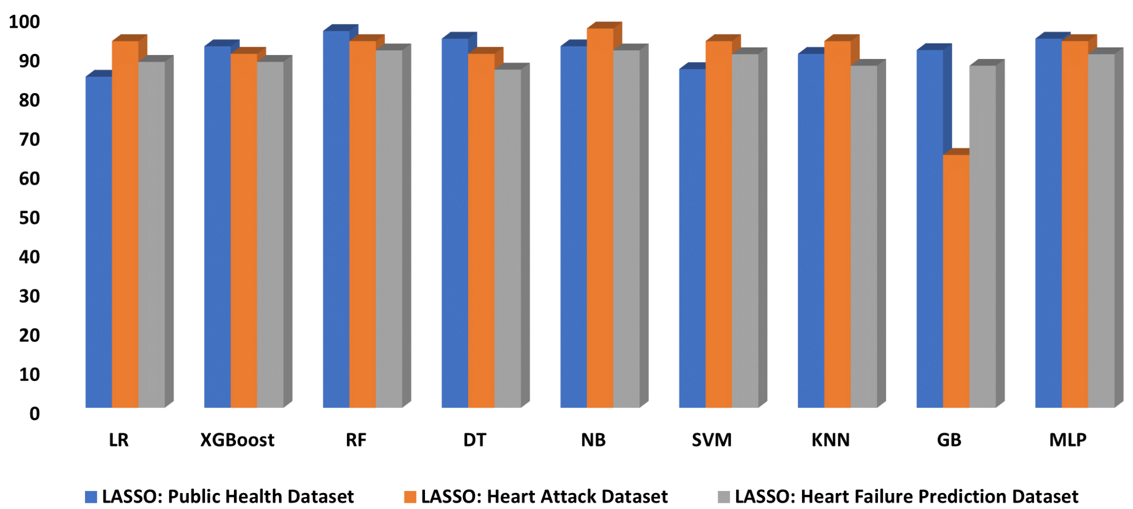


FIGURE 8: **Analyzing the results of ML models applied through a bar chart comparison with LASSO technique.**

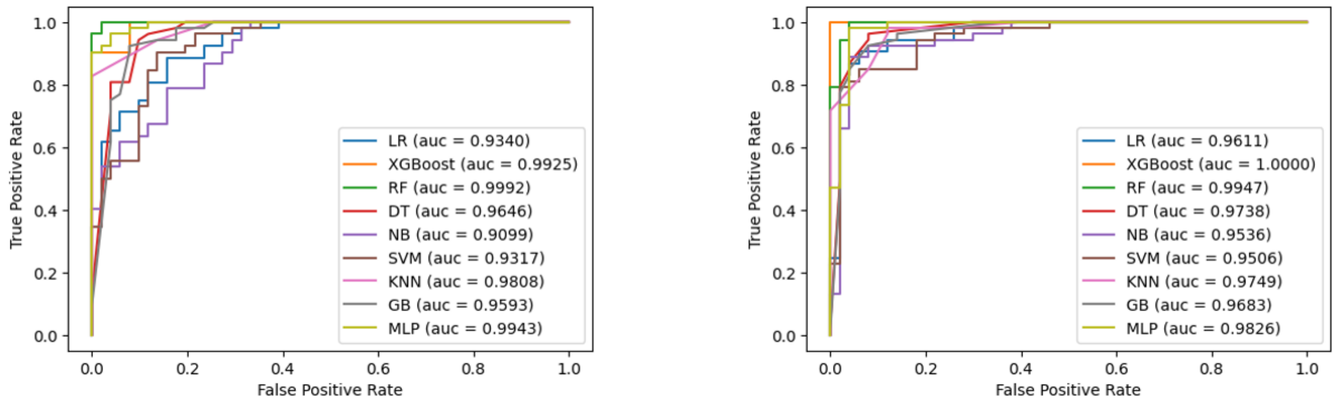
TABLE 6: Model Performance - Public Health Dataset

Model	Mean	Std	S.E	CI	
				Lower CI	Upper CI
LR	0.8446	0.0369	0.0012	0.78	0.91
XGBoost	0.9515	0.0209	0.0007	0.90	0.99
RF	0.9515	0.0209	0.0007	0.90	0.99
DT	0.9030	0.0298	0.0009	0.84	0.96
NB	0.8057	0.0390	0.0012	0.73	0.88
SVM	0.8347	0.0363	0.0011	0.77	0.90
KNN	0.9030	0.0298	0.0009	0.84	0.96
GB	0.8831	0.0317	0.0010	0.82	0.94
MLP	0.9515	0.0209	0.0007	0.90	0.99

TABLE 7: Model Performance - Heart Attack Dataset

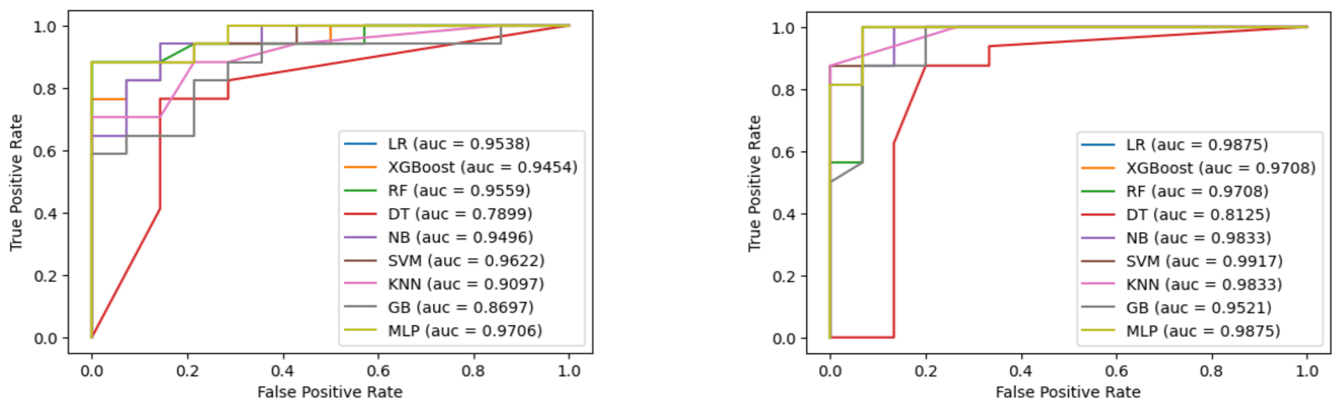
Model	Mean	Std	S.E	CI	
				Lower CI	Upper CI
LR	0.9038	0.0537	0.0017	0.77	1.0
XGBoost	0.8710	0.0583	0.0018	0.74	0.97
RF	0.9355	0.0438	0.0014	0.84	1.0
DT	0.8057	0.0698	0.0022	0.65	0.0.94
NB	0.8384	0.0651	0.0021	0.71	0.97
SVM	0.9038	0.0537	0.0017	0.77	1.0
KNN	0.8384	0.0651	0.0021	0.71	0.97
GB	0.7737	0.0758	0.0024	0.61	0.90
MLP	0.9355	0.0438	0.0014	0.84	1.0

mance differences among the algorithms studied to enhance the reliability of the findings. In Tables 6, 7, 8 and 9, the performance metrics are presented alongside their respective confidence intervals and standard errors. Additionally, boot-



(a) Before PCA (b) After PCA

FIGURE 9: ROC Curve: Comparison of Machine Learning Models on Public Health Dataset



(a) Before PCA (b) After PCA

FIGURE 10: ROC Curve: Comparison of Machine Learning Models on Heart Attack Dataset

TABLE 8: Model Performance - heart failure prediction dataset

Model	Mean	Std	S.E	CI	
				Lower CI	Upper CI
LR	0.8820	0.0313	0.0010	0.82	0.94
XGBoost	0.9223	0.0276	0.0009	0.86	0.97
RF	0.9119	0.0285	0.0009	0.85	0.96
DT	0.8521	0.0350	0.0011	0.78	0.92
NB	0.9017	0.0297	0.0009	0.84	0.96
SVM	0.8926	0.0313	0.0010	0.83	0.95
KNN	0.8926	0.0313	0.0010	0.83	0.95
GB	0.8625	0.0334	0.0011	0.79	0.92
MLP	0.8718	0.0321	0.0010	0.80	0.93

TABLE 9: Model Performance - Cardiovascular Disease Dataset

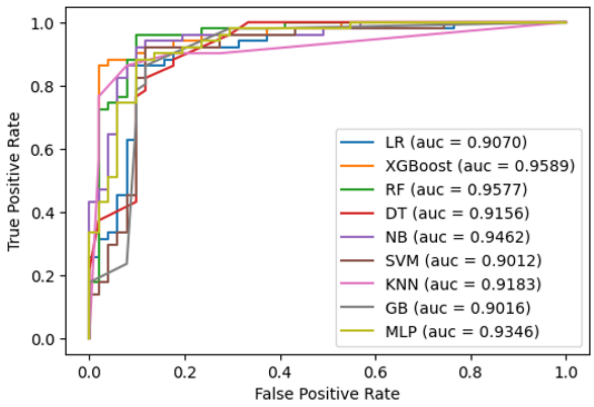
Model	Mean	Std	S.E	CI	
				Lower CI	Upper CI
LR	0.7178	0.0063	0.00023	0.70	0.72
XGBoost	0.7377	0.0070	0.00022	0.72	0.75
RF	0.7293	0.0066	0.00022	0.72	0.74
DT	0.7352	0.0068	0.00021	0.72	0.75
NB	0.5762	0.0077	0.00024	0.56	0.59
SVM	0.7235	0.0067	0.00022	0.71	0.74
KNN	0.6416	0.0067	0.0021	0.63	0.66
GB	0.7352	0.0067	0.00021	0.72	0.75
MLP	0.7290	0.0068	0.00022	0.72	0.74

strapping was employed to investigate the variability of the model's performance on separate held-out test sets. Including mean and standard deviation values for these metrics provides a comprehensive understanding of the variability and consistency. Overall, this helps in understanding the true predictive power of the models, especially given the limited dataset sizes.

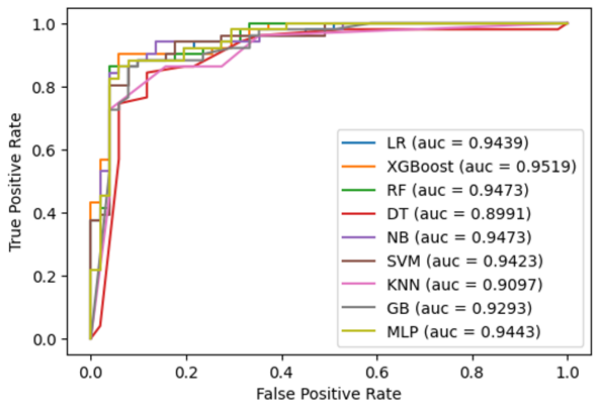
B. PERFORMANCE EVALUATION USING PCA TECHNIQUE

The performance evaluation is conducted by comparing results across applied machine learning models, emphasizing the use of the PCA technique. The comparative analysis of performance metrics is conducted and presented in Table 4. The analysis reveals that XGBoost models achieved 99.03% accuracy, 100% rate of precision and AUC, and recall and

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3436020



(a) Before PCA



(b) After PCA

FIGURE 11: ROC Curve: Comparison of Machine Learning Models on Heart Failure Prediction Dataset

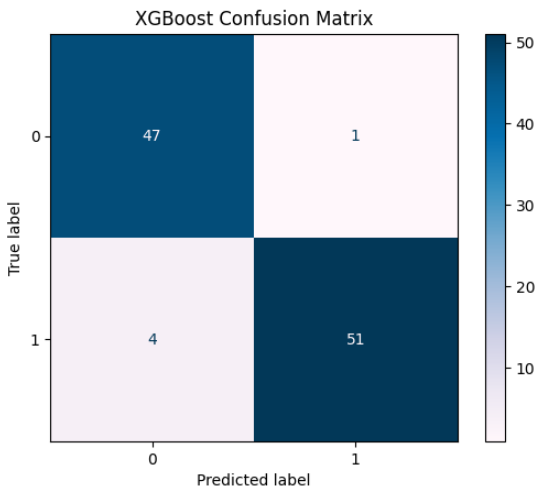


FIGURE 12: Public Health Dataset.

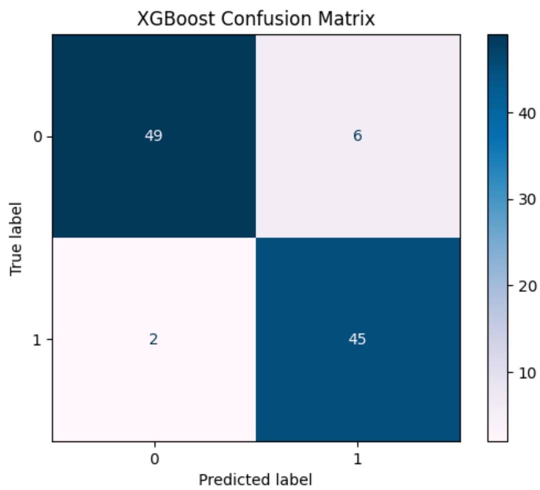


FIGURE 14: Heart Failure Prediction Dataset

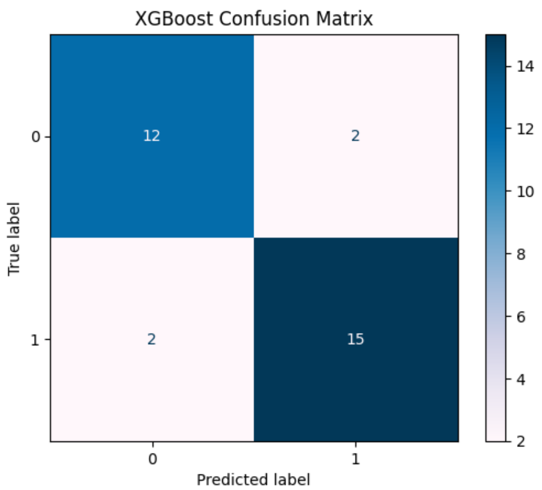


FIGURE 13: Heart Attack Dataset

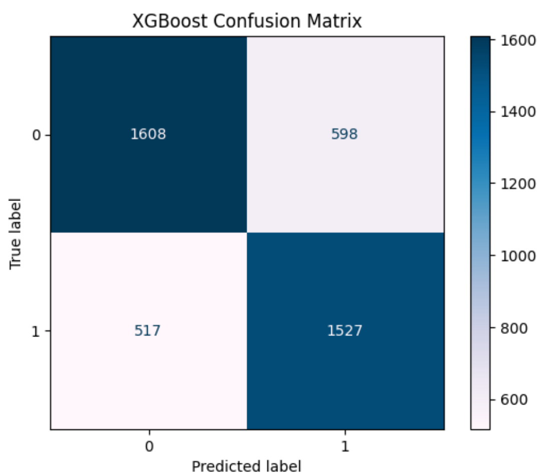


FIGURE 15: Cardiovascular Disease Dataset

f1-score are also impressive in the Public Health Dataset. Within the public health dataset, the RF and MLP methods exhibit competitive accuracy, achieving a notable 96.12%. Simultaneously, the KNN and NB models demonstrate a commendable accuracy rate of 92.23%. Additionally, both LR and DT models perform well, achieving an accuracy rate of 91.26%. The SVM did not perform well, similar to the other models. The comparative analysis of accuracy performance for applied machine learning models, based on a bar chart and utilizing the PCA technique, is visualized in Figure 7. It is evident from the analysis that LR, RF, MLP, and SVM achieved the highest accuracy of 96.77%, and precision, recall, f1-score, and AUC have exhibited excellent performance in the Heart Attack dataset. Both the KNN and NB models showcase a comparable accuracy rate of 90.32%. XGBoost and GB have achieved accuracy levels of 93.55% and 80.65%, respectively. In contrast, DT has demonstrated the poorest performance with an accuracy of 77.42%. On the Heart Failure Prediction Dataset, XGBoost achieved the highest accuracy of 91.18%. The LR, RF, and MLP each had a similar accuracy of 90.20%, while KNN had the lowest accuracy among the models.

The comparison among these models reveals that XGBoost performs well on the Public Health Dataset. The RF, KNN, NB, LR, and MLP demonstrate strong performance across all three datasets, except for KNN, which performed poorly on the Heart Failure Prediction Dataset. However, GB and DT exhibit poor performance in the Heart Attack dataset. Conversely, SVM performs well in this dataset. According to the ROC analysis in Figure 9b, 10b and 11b, the models exhibited distinctive AUC values. Notably, in the Public Health Dataset, XGBoost showcased outstanding performance with a perfect AUC score of 100%. Additionally, RF and MLP achieved commendable AUC values, reaching 99.47% and 98.26%, respectively. The AUC values for the remaining models are also noteworthy. However, in the Heart Attack dataset, SVM achieved the highest AUC value of 99.17%, while Decision Trees displayed poor performance. Also, LR, NB, KNN, and MLP exhibited good performance. On the contrary, within the Heart Disease Prediction Dataset, XGBoost attained the highest accuracy of 95.19%, with the AUC values of the other models also demonstrating notable performance. That means the utilization of the PCA technique has led to a substantial increase in performance results.

C. PERFORMANCE EVALUATION USING LASSO TECHNIQUE

The performance assessment involves a comparison of outcomes across employed machine learning models, with a focus on utilizing the LASSO technique, and Table 5 examines the comparative outcomes of the LASSO technique. The analysis indicates that in the Public Health Dataset, LR and KNN continue to exhibit similar accuracy even after the application of the LASSO technique, maintaining performance comparable to the original features. Moreover, both XGBoost and MLP have experienced a decrease in accuracy. XGBoost, initially at 95.15%, has decreased to 92.23%, while MLP has decreased to 94.17%. Additionally, precision, recall, F1-score, and AUC have also witnessed a proportional decrease. The RF, DT, NB, and GB models have achieved notably high accuracy levels, reaching approximately 96.12%, 94.17%, 92.23%, and 91.26%, respectively. Furthermore, their precision, recall, F1-score, and AUC metrics have also exhibited improvement. In contrast, the SVM model could not achieve a comparable level of accuracy. In the Heart Attack Dataset, LR, RF, SVM, and KNN have achieved an impressive accuracy of 93.55%, with RF, SVM, and KNN models achieving 100% recall, and LR achieving 100% precision. However, LR's recall exhibits comparatively poorer performance than precision, F1-score, and AUC. Both XGBoost and DT perform similarly with an accuracy of 90.32% but NB attains the highest accuracy at 96.77%, with precision and recall both reaching 100%. In contrast, GB and MLP experience a decrease in accuracy, with their accuracy levels being 64.52% and 87.10%, respectively. In the Heart Failure Prediction Dataset, LR and XGBoost show identical accuracy levels. Furthermore, DT, NB, SVM, GB, and MLP have each achieved increased accuracies of approximately 86.27%, 91.18%, 90.20%, 87.25%, and 90.20%, respectively. Notably, these models also demonstrate impressive performance in terms of precision, recall, and f1-score.

Figure 8 displays a comparison of machine learning models' accuracy performance using a LASSO technique, utilizing a bar chart. After applying the LASSO technique in the Public Health Dataset and Heart Attack Dataset datasets, it is observed that LR and SVM have achieved good performance on the Heart Attack Dataset. However, the accuracy of XGBoost is decreasing in one dataset, while it is achieving good accuracy in another dataset, for MLP, the accuracy has decreased in both datasets. The remaining models have achieved fairly consistent accuracy in both datasets. Upon further examination, in the Heart Failure Dataset, XGBoost's accuracy has decreased, LR and RF show identical performance, while the remaining models perform significantly well.

VI. WEB DEPLOYMENT

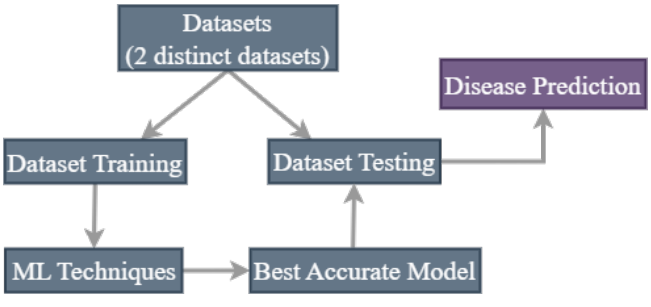


FIGURE 16: System Architecture.

Figure 16 illustrates the process of predicting cardiac problems through multiple machine learning methods, including LR, DT, RF, NB, SVM, XGBoost, K-NN, GB, and MLP.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3436020

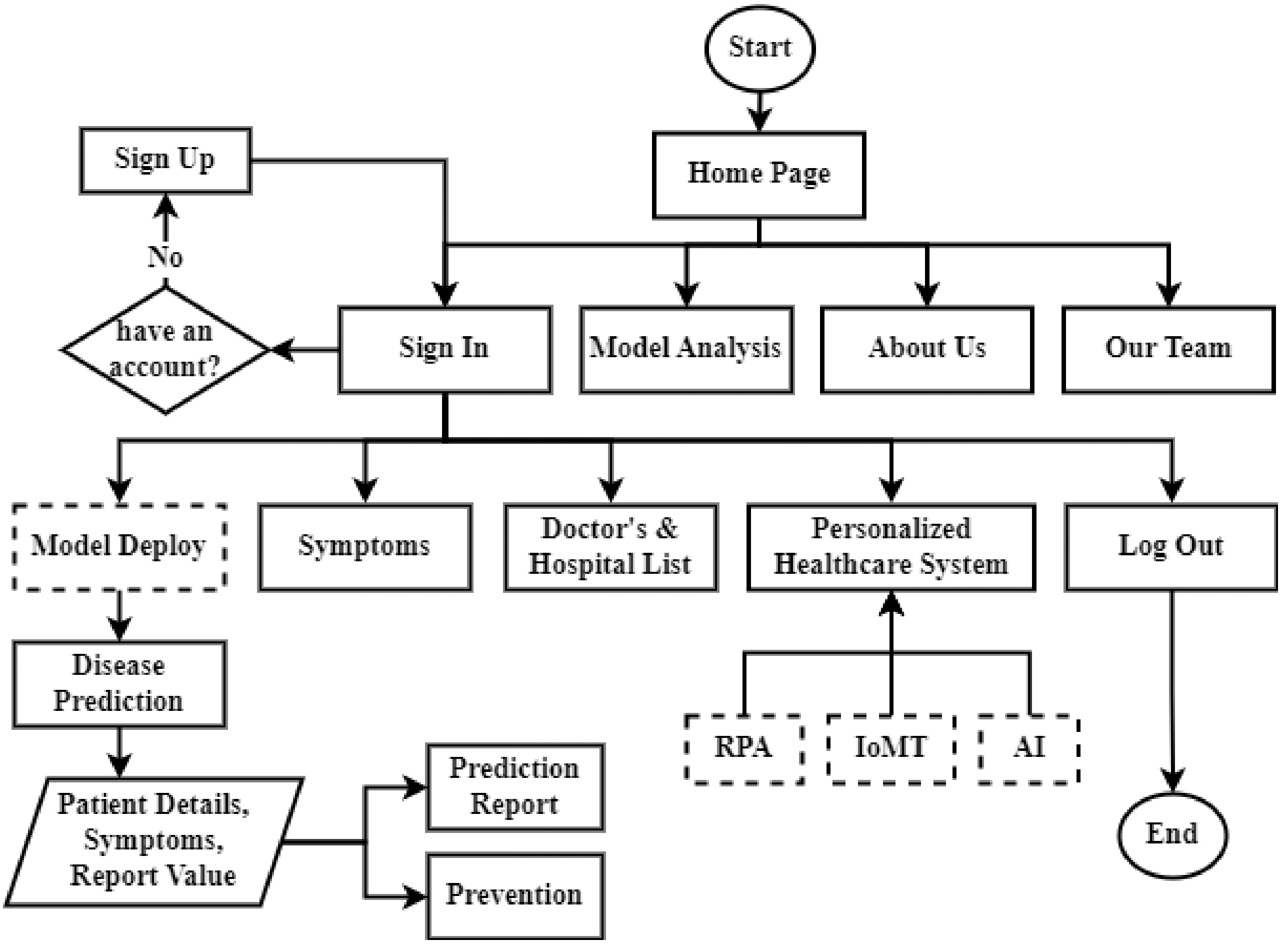


FIGURE 17: **Flowchart of a Web-Based System for Heart Disease Prediction in Healthcare.**

The aim is to enhance communication between patients and doctors, enabling both to pursue their respective objectives effectively. Each algorithm undergoes comprehensive validation, and their respective evaluation matrices are systematically compared to determine the most effective one for prediction. To maximize the accuracy of predicted results, four distinct datasets are employed in the analysis. To facilitate user things, a web application has been developed and deployed the most accurate XGBoost model to assist users in predicting heart diseases by allowing them to input their desired attribute value. Including IoMT, RPA, and AI in personalized healthcare systems, the system has characteristics of predicting heart disease, generating prediction reports, suggesting prevention, and listing hospitals and doctors with heart-related symptoms.

Figure 17 demonstrates all the characteristics of a web application. This comprehensive workflow is organized with a user-centric approach, ensuring transparency and user engagement throughout the entire process. The "Home Page" serves as the entry point, providing an interface for easy navigation. The subsequent steps, namely "Model Analysis," "Our Team," and "About Us," underscore the importance of clarity regarding the research team, methodologies, and the analytical model used. The system operates with a primary focus on predicting heart disease. Users engage by entering their specifying desired attribute values. Subsequently, the system utilizes algorithms to assess this input, facilitating a determination of whether the user is at risk of heart disease. In the event of a positive prediction, the system offers proactive guidance by providing information on healthcare professionals and facilities relevant to the predicted diseases and suggesting tips for prevention and maintenance, empowering users to take control of their heart health. Conversely, if no heart disease is predicted, the system assures the absence of such a condition and quantifies the likelihood with a percentage. Following the prediction, the system generates a comprehensive "Prediction Report," offering users a detailed overview of their health status. The discussion related to personalized healthcare solutions with IoMT, RPA, and AI will take place in section VII, but at the moment, they are not part of our web system deployment.

Overall, this system combines technical sophistication with user-centric design, aiming to enhance medical research by providing an accessible and informative tool for disease prediction and healthcare guidance that can serve in the user-end approach.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3436020

VII. PROPOSED SOLUTION

Expressing the level 1 view in Figures 18, 19 and 20, Our proposed convergence of Internet of Medical Things (IoMT), Robotic Process Automation (RPA), and Artificial Intelligence (AI) ushers in a transformative era for personalized cardiovascular care. This holistic digital healthcare solution goes beyond optimizing existing workflows – it catalyzes a paradigm shift towards proactive, pre-emptive disease management driven by real-world evidence and predictive analytics.

This synergistic framework lies in a robust machine-learning engine powered by the XGBoost algorithm, optimized through PCA-based feature engineering. Rigorously validated across diverse datasets, it demonstrates state-of-the-art performance in cardiovascular risk assessment, with the highest accuracy and precision-recall tradeoff. However, reliable predictions are just the first step.

The IoMT fabric continuously streams multi-modal data from medical-grade wearables and home-monitoring devices, capturing a comprehensive quantified-self portrait. This includes vital signs, activity tracking, sleep patterns, and environmental factors – creating a longitudinal digital phenotype that fuels just-in-time risk scoring and personalized care pathways.

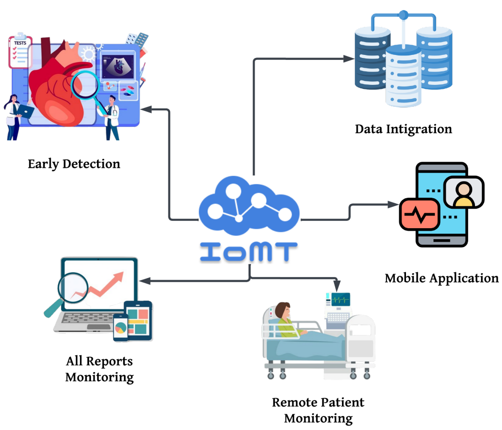


FIGURE 18: IoMT Integration in Healthcare Systems

RPA automates the data ingestion pipeline, seamlessly integrating information flows from electronic health records, IoMT streams, and other sources. Intelligent process orchestration eliminates tedious manual efforts while ensuring relevant data is routed to the predictive engine for real-time risk updates whenever new information is available.

In parallel, the framework's AI core enables cognitive support capabilities. Personalized care plans adhering to evidence-based guidelines are dynamically rendered through intuitive conversational interfaces. Risk stratification models identify high-risk individuals for escalated interventions, while clinical decision support prompts facilitate coordinated care delivery across the health network.

For instance, IoMT sensors could continuously monitor a patient's cardiovascular markers like heart rate, blood pressure, and physical activity levels. This data, automatically ingested via RPA, would periodically update the machine learning model's risk predictions. If an elevated risk is detected, the AI system could prompt the care team with recommended interventions tailored to that individual's profile, while also engaging the patient through their preferred channels with personalized lifestyle guidance.

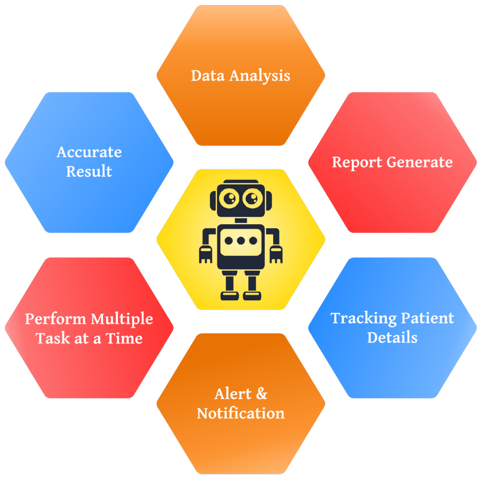


FIGURE 19: Exploring the Integration of Robotic Process Automation (RPA) in Healthcare Systems

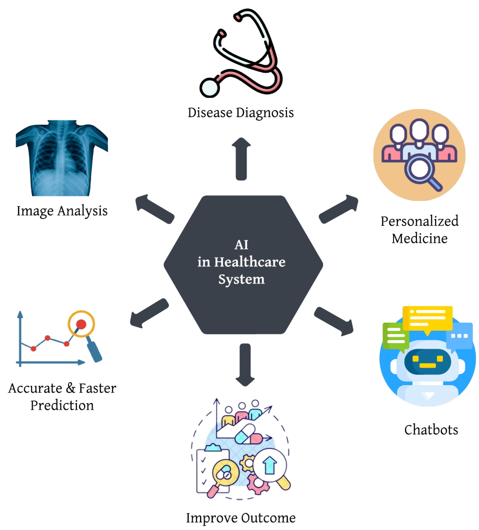


FIGURE 20: AI in Healthcare Systems

This interplay harmonizes predictive risk analytics, preventive monitoring, personalized medicine, patient empowerment, and clinical decision-making – addressing cardiovascular care holistically. Innovative aspects include automating the data value chain, enabling real-time risk scoring, transitioning from reactive to proactive paradigms, and delivering contextualized AI-driven decision support. By synergizing cutting-edge technologies through a human-centered design approach, this pioneering framework provides the crucial links to translate machine learning's potential into tangible impact – reshaping how we detect, manage, and ultimately, prevent cardiovascular disease.

VIII. DISCUSSION

The rigorous empirical evaluation across four datasets provides critical insights into the nuanced performance characteristics of different feature engineering techniques and machine learning models for cardiovascular risk prediction. The application of PCA demonstrated consistently superior accuracy compared to traditional feature selection methods like LASSO. However, our findings also underscore the importance of judiciously selecting the appropriate technique based on the specific dataset characteristics and modeling objectives. While PCA emerged as the optimal choice for boosting XGBoost's performance, the impact on other models like decision trees and gradient boosting was more varied. Certain models exhibited a decline in accuracy on the Heart Attack Dataset when using PCA feature engineering.

Contrastingly, LASSO demonstrated impressive accuracy gains for specific models like XGBoost and MLP on the Public Health Dataset. Nonetheless, its application led to accuracy degradation for the same models on the Heart Attack Dataset, further emphasizing the nuanced interplay between feature engineering strategies, model architectures, and data distributions. These insights contribute to a deeper understanding of the strengths, limitations, and ideal operating conditions for different techniques, guiding practitioners in making informed choices tailored to their specific use cases and data landscapes. It also underscores the importance of rigorous empirical evaluation and benchmarking, as opposed to relying on heuristics or assumptions.

It is imperative to note that our study involved datasets of varying sizes and characteristics, which could potentially influence model performance. Larger datasets may enable more robust feature learning and generalization, while smaller datasets might necessitate more judicious feature engineering. Due to the absence of certain key features in the Cardiovascular Disease Dataset, we opted not to apply feature engineering techniques to this dataset. The missing features significantly impacted its utility for our analysis, leading us to exclude it from the PCA and Lasso investigations. Despite this exclusion, our experiments demonstrated that XGBoost consistently outperformed other models like the remaining datasets. Future research should explore the impact of dataset size, quality, and diversity on the efficacy of different feature engineering and modeling techniques. Furthermore, our findings motivate the investigation of advanced techniques and hybrid approaches that synergistically combine the strengths of multiple methods. Ensemble strategies, joint feature selection and extraction pipelines, and adaptive techniques that dynamically tailor the approach based on data characteristics could further enhance the robustness and generalizability of machine learning models in healthcare applications.

Ultimately, our comprehensive evaluation establishes a robust empirical foundation for leveraging machine learning in cardiovascular risk assessment, while also highlighting the critical importance of principled feature engineering, rigorous benchmarking, and tailoring techniques to specific data landscapes and clinical contexts.

IX. CONCLUSION

This groundbreaking research marks a pivotal milestone in the journey towards intelligent, data-driven cardiovascular care delivery. Through a comprehensive empirical evaluation spanning four distinct real-world datasets – the 1,025-record Public Health dataset, the 303-record Heart Attack dataset, the 918-record Heart Failure Prediction Dataset, and the 42,500-record Cardiovascular Disease Dataset– we have rigorously benchmarked 8 classical machine learning algorithms and 3 deep learning models, thoroughly analyzing their performance under feature engineering techniques like PCA and LASSO regularization.

However, transcending mere predictive analytics, we have seamlessly translated the high-performing XGBoost-PCA model into an intuitive, user-friendly web application developed using the Django framework. This democratizes access to advanced cardiovascular risk assessment capabilities for clinicians and patients alike. Looking ahead, our future endeavors will focus on seamlessly integrating this validated predictive solution into healthcare systems, harmonizing machine learning innovations with clinical workflows. This will drive tangible improvements in cardiovascular care delivery, patient outcomes, and resource optimization through personalized, proactive disease management paradigms.

While our findings establish novel technical benchmarks and practical deployments, they also illuminate exciting avenues for future research. Advanced ensemble techniques, hybrid feature engineering pipelines, and adaptive, data-aware modeling strategies could further elevate performance and generalizability. As we forge ahead, guided by principled empiricism and an unwavering commitment to ethical, responsible innovation, our pioneering work will catalyze the evolution towards truly preventive population health management. The synergistic convergence of biomedical expertise and cutting-edge data science heralds a new frontier in cardiovascular care – where machine intelligence augments human potential, empowering societies worldwide to confront formidable health challenges through the prism of innovation, resilience, and hope.

DATA AND CODE AVAILABILITY

The source code for the heart disease prediction models and related analyses used in this study is available as supplementary materials. This includes the implementation details, datasets, and instructions for reproducing the results presented in this manuscript.

ACKNOWLEDGEMENT

The authors would like to acknowledge the research support provided by the Deanship of Research Oversight and Coordination (DROC) at King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia. Dr. Md Shafiullah would like to express his profound gratitude to the King Abdullah City for Atomic and Renewable Energy (K.A.CARE) for their financial support in accomplishing this work.

REFERENCES

[1] A. M. Qadri, A. Raza, K. Munir, and M. S. Almutairi, "Effective feature engineering technique for heart disease prediction with machine learning," *IEEE Access*, vol. 11, pp. 56 214–56 224, 2023.

[2] A. Abdellatif, H. Abdellatef, J. Kanesan, C.-O. Chow, J. H. Chuah, and H. M. Gheni, "An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods," *IEEE Access*, vol. 10, pp. 79 974–79 985, 2022.

[3] M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, P. Singh *et al.*, "Latest trends on heart disease prediction using machine learning and image fusion," *Materials Today: Proceedings*, vol. 37, pp. 3213–3218, 2021.

[4] P. Rani, R. Kumar, N. M. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 263–275, 2021.

[5] X. Zang, J. Du, and Y. Song, "Early prediction of heart disease via lstm-xgboost," in *Proceedings of the 2023 9th International Conference on Computing and Artificial Intelligence*, 2023, pp. 631–637.

[6] J. Azmi, M. Arif, M. T. Nafis, M. A. Alam, S. Tanweer, and G. Wang, "A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data," *Medical Engineering & Physics*, vol. 105, p. 103825, 2022.

[7] S. Xie, Z. Yu, and Z. Lv, "Multi-disease prediction based on deep learning: A survey." *CMES-Computer Modeling in Engineering & Sciences*, vol. 128, no. 2, 2021.

[8] A. Almulihi, H. Saleh, A. M. Hussien, S. Mostafa, S. El-Sappagh, K. Alnowaiser, A. A. Ali, and M. Refaat Hassan, "Ensemble learning based on hybrid deep learning model for heart disease early prediction," *Diagnostics*, vol. 12, no. 12, p. 3215, 2022.

[9] R. Shwetha and V. Kirubanand, "Remote monitoring of heart patients using robotic process automation (rpa)," in *ITM Web of Conferences*, vol. 37. EDP Sciences, 2021, p. 01002.

[10] M. Umer, T. Aljrees, H. Karamti, A. Ishaq, S. Alsubai, M. Omar, A. K. Bashir, and I. Ashraf, "Heart failure patients monitoring using iot-based remote monitoring system," *Scientific Reports*, vol. 13, no. 1, p. 19213, 2023.

[11] S. Subramani, N. Varshney, M. V. Anand, M. E. M. Soudagar, L. A. Al-Keridis, T. K. Upadhyay, N. Alshammari, M. Saeed, K. Subramanian, K. Anbarasu *et al.*, "Cardiovascular diseases prediction by machine learning incorporation with deep learning," *Frontiers in Medicine*, vol. 10, p. 1150933, 2023.

[12] X. Yu, S. Zhou, H. Zou, Q. Wang, C. Liu, M. Zang, and T. Liu, "Survey of deep learning techniques for disease prediction based on omics data," *Human Gene*, vol. 35, p. 201140, 2023.

[13] G. N. AhmaD, H. Fatima, and S. M. H. Akhter, "Optimal medical diagnosis of human heart disease by k-nearest neighbors and decision trees classifiers algorithms," *Artificial Intelligence and Communication Technologies*, p. 1031–1037, 2023.

[14] M. Azhar and L. M. Gladence, "A machine learning approach for predicting disease in heart using logistic regression," in *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE, 2022, pp. 925–930.

[15] M. J. Gaikwad, P. S. Asole, and L. S. Bitla, "Effective study of machine learning algorithms for heart disease prediction," in *2022 2nd International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*. IEEE, 2022, pp. 1–6.

[16] M. Pal and S. Parija, "Prediction of heart diseases using random forest," in *Journal of Physics: Conference Series*, vol. 1817, no. 1. IOP Publishing, 2021, p. 012009.

[17] A. Khan, M. Qureshi, M. Daniyal, K. Tawiah *et al.*, "A novel study on machine learning algorithm-based cardiovascular disease prediction," *Health & Social Care in the Community*, vol. 2023, 2023.

[18] N. Mohan, V. Jain, and G. Agrawal, "Heart disease prediction using supervised machine learning algorithms," in *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*. IEEE, 2021, pp. 1–3.

[19] C. Dhanamjayulu, G. V. Suraj, M. Nikhil, R. Kaluri, and S. Koppu, "A machine learning algorithm-based iot-based message alert system for predicting coronary heart disease," in *International Conference on Advancements in Smart Computing and Information Security*. Springer, 2022, pp. 362–376.

[20] M. J. N. Nayeem, S. Rana, and M. R. Islam, "Prediction of heart disease using machine learning algorithms," *European Journal of Artificial Intelligence and Machine Learning*, vol. 1, no. 3, pp. 22–26, 2022.

[21] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, p. 012072, jan 2021.

[22] V. Rajalakshmi, D. Sasikala, and A. Kala, "A predictive analysis for heart disease using machine learning," in *Intelligent Computing and Applications: Proceedings of ICICA 2019*. Springer, 2021, pp. 473–479.

[23] S. Krishnan, P. Magalingam, and R. Ibrahim, "Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction." *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 11, no. 6, 2021.

[24] V. Sharma, A. Rasool, and G. Hajela, "Prediction of heart disease using dnn," in *2020 second international conference on inventive research in computing applications (ICIRCA)*. IEEE, 2020, pp. 554–562.

[25] S. Sharma and M. Parmar, "Heart diseases prediction using deep learning neural network model," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 9, no. 3, pp. 2244–2248, 2020.

[26] M. M. Rahman, "A web-based heart disease prediction system using machine learning algorithms," *Network Biology*, vol. 12, no. 2, p. 64, 2022.

[27] I. Mohit, K. S. Kumar, U. A. K. Reddy, and B. S. Kumar, "An approach to detect multiple diseases using machine learning algorithm," in *Journal of Physics: Conference Series*, vol. 2089, no. 1. IOP Publishing, 2021, p. 012009.

[28] L. D. Gopisetti, S. K. L. Kummera, S. R. Pattamsetti, S. Kuna, N. Parsi, and H. P. Kodali, "Multiple disease prediction system using machine learning and streamlit," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE, 2023, pp. 923–931.

[29] K. E. Fahim, H. Yassin, M. H. Amin, P. D. Dewan, and A. Islam, "Detection of cardiovascular disease of patients at an early stage using machine learning algorithms," in *2022 International Conference on Healthcare Engineering (ICHE)*. IEEE, 2022, pp. 1–6.

[30] B. Rawat, Y. Joshi, and A. Kumar, "Ai in healthcare: Opportunities and challenges for personalized medicine and disease diagnosis," in *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2023, pp. 374–379.

[31] N. Taimoor and S. Rehman, "Reliable and resilient ai and iot-based personalised healthcare services: A survey," *IEEE Access*, vol. 10, pp. 535–563, 2021.

[32] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A comparative analysis of logistic regression, random forest and knn models for the text classification," *Augmented Human Research*, vol. 5, pp. 1–16, 2020.

[33] Y. Zhao and Y. Zhang, "Comparison of decision tree methods for finding active objects," *Advances in Space Research*, vol. 41, no. 12, pp. 1955–1959, 2008.

[34] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 5, p. 272, 2012.

[35] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting employee attrition using machine learning approaches," *Applied Sciences*, vol. 12, no. 13, p. 6424, 2022.

[36] M. W. Ahmad, J. Reynolds, and Y. Rezgui, "Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees," *Journal of cleaner production*, vol. 203, pp. 810–821, 2018.

[37] A. Raza, H. U. R. Siddiqui, K. Munir, M. Almutairi, F. Rustam, and I. Ashraf, "Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction," *Plos one*, vol. 17, no. 11, p. e0276525, 2022.

[38] S. Shabani, S. Samadianfard, M. T. Sattari, A. Mosavi, S. Shamshirband, T. Kmet, and A. R. Várkonyi-Kóczy, "Modeling pan evaporation using gaussian process regression k-nearest neighbors random forest and support vector machines; comparative analysis," *Atmosphere*, vol. 11, no. 1, p. 66, 2020.

[39] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, "Improving heart disease prediction using feature selection approaches," in *2019 16th international bhurban conference on applied sciences and technology (IBCAST)*. IEEE, 2019, pp. 619–623.

[40] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: a comparative study and analysis," *Health and Technology*, vol. 11, pp. 87–97, 2021.

[41] N. Ali, D. Neagu, and P. Trundle, "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets," *SN Applied Sciences*, vol. 1, pp. 1–15, 2019.

[42] O. González-Recio, J. Jiménez-Montero, and R. Alenda, "The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets," *Journal of dairy science*, vol. 96, no. 1, pp. 614–624, 2013.

[43] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[44] E. Venkatesan and A. B. Mahindrakar, "Forecasting floods using extreme gradient boosting-a new approach," *International Journal of Civil Engineering and Technology*, vol. 10, no. 2, pp. 1336–1346, 2019.

[45] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[46] S. Dey, Y. Kumar, S. Saha, and S. Basak, "Forecasting to classification: Predicting the direction of stock market price using xtreme gradient boosting," *PESIT South Campus*, pp. 1–10, 2016.

[47] W. H. Delashmit, M. T. Manry *et al.*, "Recent developments in multilayer perceptron neural networks," in *Proceedings of the seventh annual memphis area engineering and science conference, MAESC*, 2005, pp. 1–15.

[48] B. Sujal, J. Nanthini, and M. Reddy, "Web-based heart disease prognosis using neural network and hybrid approach," in *2022 IEEE International Conference on Data Science and Information System (ICDSIS)*. IEEE, 2022, pp. 1–5.

[49] M. Akther, Z. R. Chowdhury, A. Tabassum, and M. S. R. Kohinoor, "A comparative study on different machine learning techniques in diabetes risk assessment," in *2023 3rd International Conference on Intelligent Technologies (CONIT)*, 2023, pp. 1–6.

[50] D. Zhang, Y. Chen, Y. Chen, S. Ye, W. Cai, J. Jiang, Y. Xu, G. Zheng, and M. Chen, "Heart disease prediction based on the embedded feature selection method and deep neural network," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–9, 2021.

[51] T. Epelbaum, "Deep learning: Technical introduction," *arXiv preprint arXiv:1709.01412*, 2017.

[52] M. Ashraf, M. Rizvi, and H. Sharma, "Improved heart disease prediction using deep neural network," *Asian Journal of Computer Science and Technology*, vol. 8, no. 2, pp. 49–54, 2019.

[53] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *2008 IEEE/ACS international conference on computer systems and applications*. IEEE, 2008, pp. 108–115.

[54] Public Health Dataset. Kaggle. [Accessed: Nov. 03, 2023]. [Online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

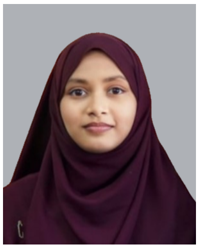
[55] Heart Attack Dataset. Kaggle. [Accessed: Nov. 03, 2023]. [Online]. Available: https://www.kaggle.com/datasets/pritsheta/heart-attack

[56] Heart Failure Prediction Dataset. Kaggle. [Accessed: May. 29, 2024]. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data

[57] Cardiovascular Disease dataset. Kaggle. [Accessed: May. 29, 2024]. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/data

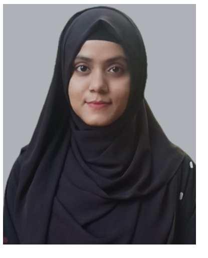
[58] G. A. Ansari, S. S. Bhat, M. D. Ansari, S. Ahmad, J. Nazeer, A. Eljialy *et al.*, "Performance evaluation of machine learning techniques (mlt) for heart disease prediction," *Computational and Mathematical Methods in Medicine*, vol. 2023, 2023.

[59] T. S. Brisimi, T. Xu, T. Wang, W. Dai, W. G. Adams, and I. C. Paschalidis, "Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach," *Proceedings of the IEEE*, vol. 106, no. 4, pp. 690–707, Apr. 2018.



KHADIZA AKTHER undergraduate student from the Department of Computer Science and Engineering, Leading University, Sylhet-3112, Bangladesh, since Jan 2020 to Mar 2024. From 2023 to the present, she is an undergraduate research assistant at InteX Research Lab. Her current research interests include machine learning, deep learning, artificial intelligence, image processing, data science, cloud computing, and software engineering.

MD. SAIDUR RAHMAN KOHINOOR (Member, IEEE) is a Lecturer of Computer Science & Engineering at Leading University, Sylhet, Bangladesh. Currently, he is pursuing an M.Sc. degree at KFUPM, Dhahran, KSA. His research area focused on applied machine learning, specifically in the areas of healthcare and communications, as well as mobile sensing, context-aware computing, intelligent systems, information & knowledge management, and bioinformatics. As the Principal Investigator at InteX Research Lab, he leads a team of 34 research enthusiasts, exploring innovative solutions. Driven by a commitment to excellence, he has received several research grants from reputed organizations including the British Council. Also honored as a Richard E. Merwin Scholar of the IEEE Computer Society, recognizing his community-driven contributions.



BUSHRA SIDDIKA PRIYA undergraduate student from the Department of Computer Science and Engineering, Leading University, Sylhet-3112, Bangladesh, since January 2020 to March 2024. From 2023 to the present, she has been an undergraduate research assistant at the InteX Research Lab. Her current research interests include machine learning, deep learning, the Internet of Things (IoT), artificial intelligence, software engineering, and healthcare technology innovation.

MD. JAMANER RAHAMAN received the B.Sc. and M.Sc. degree in Computer Science and Engineering from the Daffodil International University, Dhaka, Bangladesh, in 2016 and 2018 respectively. He is currently working as a Lecturer in the Department of Computer Science and Engineering at Leading University, Sylhet, Bangladesh since 01 February 2021, along with the Investigator at InteX Research Lab. Previously he worked as a Research Associate in the department of CSE at Daffodil International University and also worked as a Lecturer in the department of CSE at United College of Aviation Science and Management, Dhaka, Bangladesh. His current research interests include machine learning, deep learning, and artificial intelligence.

MD MAHFUZUR RAHMAN is an Assistant Professor in the Department of Information and Computer Science (ICS) at King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia. His current research focuses on Cloud Computing, Cybersecurity, and Machine Learning. He received his Ph.D. degree in Computer Science from the University of Manitoba,Canada.

MD SHAFIULLAH (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical and electronic engineering (EEE) from the Bangladesh University of Engineering and Technology (BUET), Bangladesh, in 2009 and 2013, respectively, and the Ph.D. degree in electrical power and energy systems from the King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, in 2018. He was a Faculty Member of the Department of Electrical and Electronic Engineering, International Islamic University of Chittagong (IIUC), Bangladesh, from 2009 to 2013. He is currently an Assistant Professor with the Control and Instrumentation Department and a Research Affiliate with the Interdisciplinary Research Center for Sustainable Energy Systems (IRC-SES), KFUPM. He received the prestigious King Abdullah City for Atomic and Renewable Energy (K.A.CARE) Fellowship and the Early Career Research Award by the KFUPM, both for the year 2024. His research interests include grid fault diagnosis, grid integration of renewable energy resources, power quality analysis, power system control and stability, evolutionary algorithms, and machine learning techniques.

• • •