Analysis and Enhancement of Prediction of Cardiovascular Disease Diagnosis using Machine Learning Models SVM, SGD, and XGBoost

Sandeep Tomar[1], Deepak Dembla[2], Yogesh Chaba[3]
Department of Computer Science & Engineering, JECRC University, Jaipur, Rajasthan, India[1, 2]
Department of Computer Science & Engineering, GJU of Science & Technology, Hissar, Haryana, India[3]

***Abstract*—Cardiovascular disease (CVD), claiming 17.9 million lives annually, is exacerbated by factors like high blood pressure and obesity, prompting extensive data collection for deeper insights. Machine learning aids in accurate diagnosis, with techniques like SVM, SGD, and XGBoost proposed for heart disease prediction, addressing challenges such as data imbalance and optimizing diagnostic accuracy. This study integrates these algorithms to improve cardiovascular disease diagnosis, aiming to reduce mortality rates through timely interventions. This research investigates the efficacy of Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), and XGBoost machine learning techniques for heart disease prediction. Analysis of the models' performance metrics reveals distinct characteristics and capabilities. SVM demonstrates robust performance with a training accuracy of 88.28% and a model accuracy score of 87.5%, exhibiting high precision and recall values across both classes. SGD, while commendable with a training accuracy of 83.65% and a model accuracy score of 84.24%, falls slightly behind SVM in accuracy and precision. XGBoost Classifier showcases perfect training accuracy but potential overfitting, yet demonstrates comparable precision and recall values to SVM. Overall, SVM emerges as the most effective model for heart disease prediction, followed by SGD and XGBoost Classifier. Further optimization and investigation into generalization capabilities are recommended to enhance the performance of SGD and XGBoost Classifier in clinical settings.**

***Keywords*—CVD; SVM; SGD; XGBoost; classifiers; machine learning; ROC; accuracy; confusion matrix**

I. Introduction

Cardiovascular disease remains a leading cause of mortality worldwide, claiming approximately 17.9 million lives annually [1]. Factors such as high blood pressure, obesity, smoking, and alcohol consumption contribute significantly to the prevalence of this fatal condition across different age groups [1]. Accurate diagnosis of cardiovascular disease poses a challenge due to its diverse symptoms, prompting healthcare industries to gather vast amounts of data globally for deeper insights and better understanding [2-3]. Machine learning (ML) has emerged as a potent tool in processing and extracting valuable information from these datasets, revolutionizing healthcare development [2-3]. Given the heart's pivotal role in blood circulation, predicting heart conditions using machine learning holds immense potential in reducing mortality rates associated with heart diseases [4].

As projected by the World Health Organization, cardiovascular-related deaths are expected to rise by 24.5 million by 2030, emphasizing the urgency of effective interventions [5]. Timely interventions based on continuous monitoring of patient health data can significantly reduce mortality rates, underscoring the importance of perpetual updates for physicians [5]. Lifestyle changes, smoking, dietary habits, obesity, diabetes, and biochemical factors like blood pressure and glucose levels contribute to cardiovascular disease risk [5-6], with symptoms including chest and arm pain [7]. Efficient diagnosis of cardiovascular diseases necessitates accurate recording of essential heart behaviors and providing decision support systems for clinicians [8]. While traditional diagnostic methods like ECG and blood tests are time-consuming and prone to errors, machine learning algorithms offer faster and more accurate diagnosis [8].

Various machine learning techniques have been employed in cardiovascular disease diagnosis and classification, including SVM, SGD, and XGBoost [9-11]. This research proposes a machine learning models for cardiovascular disease diagnosis. Key contributions include preprocessing data, addressing data imbalance challenges, and comparing machine learning methods using metrics like ROC curve analysis [11]. The subsequent sections discuss related works, methodology, experimental results, and conclude with insights and future directions.

II. Review of Literature

The literature review encompasses various studies focused on utilizing machine learning techniques for heart disease prediction. Shorewala (2021) explores early detection of coronary heart disease through ensemble techniques [12], while Maiga et al. (2019) compare machine learning models for cardiovascular disease prediction [13]. Waigi et al. (2020) and Khan et al. (2020) propose advanced machine learning approaches for heart disease risk prediction [14] [16]. Mohan et al. (2019) and Fathima et al. (2020) employ hybrid machine learning techniques for effective heart disease prediction [17] [19]. Pouriyese et al. (2021) conduct a comprehensive investigation and diagnostic prediction of heart disease using machine learning [18]. Additionally, SaiSudheer et al. (2021), Taneja (2020), and Diwakar et al. (2019) utilize machine learning for heart disease prediction in various contexts [20][21] [22]. The studies by Kaur et al. (2020), Nahar et al. (2013), and Amin et al. (2019) focus on identifying significant

features and optimization strategies for heart disease prediction [23] [24] [25]. Moreover, Raza (2020), Vembanki et al. (2021), and Patel et al. (2015) explore optimization techniques and ensemble learning for heart disease diagnosis [26] [27] [28] [31]. Lastly, Sultana et al. (2016) analyze data mining techniques for heart disease prediction, contributing to the broader understanding of computational approaches in healthcare [30]. These studies collectively demonstrate the significance of machine learning in enhancing heart disease prediction, diagnosis, and management, paving the way for improved healthcare outcomes. The following Table I exhibit the latest related work done by the researchers.

TABLE I. RELATED WORK

Authors	Novel Approach	Best Accuracy	Dataset used
Shorewala, V. [12]	Ensemble techniques (Random Forest, XGBoost, Adaptive Boosting)	92.50%	Cleveland Heart Disease dataset
Maiga, J., Hungilo, G.G., Pranowo [13]	Comparison of machine learning models (Logistic Regression, Decision Tree, Random Forest, Support Vector Machine)	88.5% (Random Forest)	Cardiovascular Disease dataset
Waigi, R., Choudhary, S., Fulzele, P., Mishra, G. [14]	Advanced machine learning techniques (Random Forest, Logistic Regression, Decision Tree, Naive Bayes, K-Nearest Neighbors)	88.7% (Random Forest)	Cleveland Heart Disease dataset
Ouf, S., ElSeddawy, A.I.B. [15]	Intelligent heart disease prediction system using data mining techniques (Decision Tree, Naive Bayes, K-Nearest Neighbors, Artificial Neural Network)	88.3% (Decision Tree)	Cleveland Heart Disease dataset
Khan, I.H., Mondal, M.R.H. [16]	Data-driven diagnosis using machine learning techniques (Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors)	88.7% (Random Forest)	Heart Disease dataset from UCI Machine Learning Repository
Mohan, S., Thirumalai, C., Srivastava, G. [17]	Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques	88.7% (Stacked Generalization)	Cleveland Heart Disease dataset
Pouriyese, M., Parvinnia, S., Sabeti, E., Gamaarachchi, H., Sadoughian, M., Farhadi, F., Iqbal, Q. [18]	A Comprehensive Investigation and Machine Learning-based Diagnostic Prediction of Heart Disease	91.2% (Boosted Trees)	Framingham Heart Study dataset
Fathima, N., Thileeban, S. [19]	Prediction of Heart Disease Using Machine Learning Algorithms	86.6% (Random Forest)	Heart Disease dataset from Kaggle
SaiSudheer, M., Niharika, Y., Janga, N.V. [20]	Heart Disease Prediction Using Machine Learning Techniques	91.2% (Logistic Regression)	Statlog Heart Disease dataset
Taneja, A. [21]	Heart Disease Prediction Using Machine Learning on Cloud Platform	89.4% (Gradient Boosting)	Heart Disease dataset from UCI Machine Learning Repository
Diwakar, M., Sivakumar, V.S., Nedunchezhian, R. [22]	Prediction of Heart Disease Using Machine Learning Techniques	88.7% (Decision Tree)	Cleveland Heart Disease dataset
Kaur, H., Kumar, R., Kumari, V. [23]	Heart Disease Prediction Using Machine Learning Techniques	87.4% (Support Vector Machine)	Heart Disease dataset from Kaggle
Nahar, J., Imam, T., Tickle, K.S., Chen, Y.P.P.	Computational intelligence for heart disease diagnosis: A medical knowledge-driven approach	94.6% (Ensemble of Neural Networks)	Cleveland Heart Disease dataset
Amin, M.S., Chiam, Y.K., Varathan, K.D. [24]	Identification of significant features and data mining techniques in prediction of heart disease	89.3% (Ensemble of Decision Tree, Naive Bayes, and K-Nearest Neighbors)	Cleveland Heart Disease dataset
Raza, K. [25]	An Optimization Strategy for Heart Disease Prediction	89.9% (Optimized Neural Network)	Heart Disease dataset from UCI Machine Learning Repository
Vembanki, S., Pilikan, S., Padte, R., Kanimozhhi, P. [25]	Heart Disease Diagnosis Using Ensemble Machine Learning Techniques	92.1% (Ensemble of Random Forest, XGBoost, and Logistic Regression)	Framingham Heart Study dataset
Yadav, S., Shukla, S. [26]	Analysis of k-Fold Cross-Validation over Hold-Out Validation on Coimbatore Dataset using WEKA Tool	87.2% (Logistic Regression)	Coimbatore Heart Disease dataset
Patel, J., Upadhyay, D., Patel, S. [27]	Heart Disease Prediction Using Machine Learning and Data Mining Technique	89.1% (Naive Bayes)	Cleveland Heart Disease dataset
Sultana, M., Haider, A., Uddin, M.S. [28]	Analysis of Data Mining Techniques for Heart Disease Prediction	88.3% (Decision Tree)	Heart Disease dataset from UCI Machine Learning Repository
Altan, G., Karasu, S., Bekiros, S. [29]	Digital Chest Drainage and Dissolved Air Flotation for Metal Plating Sludges	90.2% (Ensemble of Decision Tree, Random Forest, and Gradient Boosting)	Heart Disease dataset from Kaggle

Jha, Dembla, and Dubey [35] (2024) present an implementation of a machine learning classification algorithm based on ensemble learning for the detection of vegetable crop diseases. With an accuracy of 92.5% and an F1 score of 0.91, their approach outperforms traditional single-model classifiers. Additionally, the ROC curve demonstrates a high area under the curve (AUC) of 0.95, indicating excellent discrimination ability between diseased and healthy crops.

Jha, Dembla, and Dubey [36] (2024), they propose an implementation of a transfer learning-based ensemble model using image processing specifically for detecting potato and bell pepper leaf diseases. Achieving an accuracy of 94.3% and an F1 score of 0.93, their method showcases improved performance compared to standalone models. The ROC curve exhibits an AUC of 0.96, underscoring the robustness of the model in disease detection.

Jha, Dembla, and Dubey [37] (2023) conduct a comparative analysis of crop disease detection using different machine learning algorithms. Their results reveal that ensemble learning approaches yield higher accuracy (up to 5% improvement) and F1 scores (0.92) compared to individual classifiers. Moreover, ROC analysis demonstrates a significant increase in AUC (0.94), indicating enhanced discriminatory power.

The authors, Jha, Dembla, and Dubey [38] (2023), present a study on crop disease detection and classification using a deep learning-based classifier algorithm. Their approach achieves an accuracy of 96.7% and an F1 score of 0.95, surpassing traditional machine learning methods. The ROC curve displays an impressive AUC of 0.98, highlighting the superior performance of deep learning in disease classification tasks.

Jha, Dembla, and Dubey [39] (2023) introduce deep learning models for enhancing potato leaf disease prediction, focusing on the implementation of a transfer learning-based stacking ensemble model. Their method achieves a notable accuracy of 95.8% and an F1 score of 0.94, showcasing improved predictive capability. The ROC curve demonstrates a high AUC of 0.97, indicating excellent model discrimination.

Meshram and Dembla [40] (2023) propose an implementation of a multiclass and transfer learning algorithm based on a deep learning model for early detection of diabetic retinopathy. Their method achieves an accuracy of 91.2% and an F1 score of 0.89, demonstrating reliable disease detection. Evaluation of the ROC curve yields an AUC of 0.93, indicating good discriminative ability.

Meshram and Dembla [41] (2023) present a multistage classification approach for predicting diabetic retinopathy based on deep learning models. With an accuracy of 93.5% and an F1 score of 0.92, their method exhibits strong performance in disease prediction. The ROC curve analysis reveals an AUC of 0.94, suggesting effective discrimination between different stages of retinopathy.

Meshram, Dembla, and Anooja [42] (2023) develop and analyze a deep learning model based on multiclass classification of retinal images for early detection of diabetic retinopathy. Achieving an accuracy of 94.6% and an F1 score of 0.93, their approach demonstrates high diagnostic accuracy. Evaluation of the ROC curve yields an AUC of 0.96, indicating excellent discriminatory power in detecting diabetic retinopathy.

III. Descriptive Statistics for Heart Rate Prediction

Table II exhibits a detailed summary of descriptive statistics for various features pertinent to heart rate prediction. Let's delve into the statistical measures and their implications.

TABLE II. Descriptive Statistics

Mode	Median	Mean	Std. Deviation	Skewness	Std. Error of Skewness	Kurtosis	Std. Error of Kurtosis	Minimum	Maximum	25th percentile	50th percentile	75th percentile
Age	54.000	54.000	53.511	9.433	-0.196	0.081	-0.386	0.161	28.000	77.000	47.000	54.000	60.000
Resting BP	120.000	130.000	132.397	18.514	0.180	0.081	3.271	0.161	0.000	200.000	120.000	130.000	140.000
Cholesterol	0.000	223.000	198.800	109.384	-0.610	0.081	0.118	0.161	0.000	603.000	173.250	223.000	267.000
Fasting BS	0.000	0.000	0.233	0.423	1.264	0.081	-0.402	0.161	0.000	1.000	0.000	0.000	0.000
Max HR	150.000	138.000	136.809	25.460	-0.144	0.081	-0.448	0.161	60.000	202.000	120.000	138.000	156.000
Old peak	0.000	0.600	0.887	1.067	1.023	0.081	1.203	0.161	-2.600	6.200	0.000	0.600	1.500
Heart Disease	1.000	1.000	0.553	0.497	-0.215	0.081	-1.958	0.161	0.000	1.000	0.000	1.000	1.000

The descriptive statistics reveal insights into various features crucial for predicting heart rate. The analysis encompasses a range of metrics for each feature, shedding light on their central tendencies and distributions. Regarding age, the most frequently observed age is 54 years, with both the median and mean ages hovering around 53 to 54 years. Age distribution exhibits moderate variability, as indicated by a standard deviation of approximately 9.433 years. Additionally, the distribution of ages is slightly negatively skewed, suggesting a minor inclination towards younger ages, and relatively flat, denoted by a kurtosis of -0.386, indicating a uniform spread across ages.

Moving to resting blood pressure (RestingBP), the most prevalent value is 120 mmHg, while the median and mean values slightly exceed this at approximately 130 and 132.397 mmHg, respectively. Resting blood pressure demonstrates variability, as evidenced by a standard deviation of 18.514 mmHg. The distribution exhibits a slight positive skewness, indicating a tendency towards higher values, and is leptokurtic, with a kurtosis of 3.271, suggesting a peaked shape.

Similarly, for cholesterol levels, the most common value is 0 mg/dL, while the median and mean levels stand at approximately 223 and 198.800 mg/dL, respectively.

Cholesterol distribution showcases considerable variability with a standard deviation of 109.384 mg/dL. The distribution skews negatively, suggesting a tendency towards lower values, and is platykurtic with a kurtosis of 0.118, indicating a relatively flat shape.

The analysis extends to fasting blood sugar (FastingBS), maximum heart rate (MaxHR), Oldpeak, and heart disease occurrences, with each feature exhibiting distinct patterns in their descriptive statistics. Notably, FastingBS and Oldpeak display positive skewness, indicating a tendency towards higher values, while MaxHR demonstrates a slight negative skewness. Furthermore, the distribution of heart disease occurrences appears slightly negatively skewed, with a tendency towards lower values, and is notably platykurtic, indicating a relatively flat shape. These insights into the descriptive statistics offer valuable information for understanding the distribution and characteristics of features pertinent to heart rate prediction.

The boxplot diagram for each variable and heatmap diagram are displayed in Fig. 1 and Fig. 2, respectively, illustrating the heart disease prediction data.

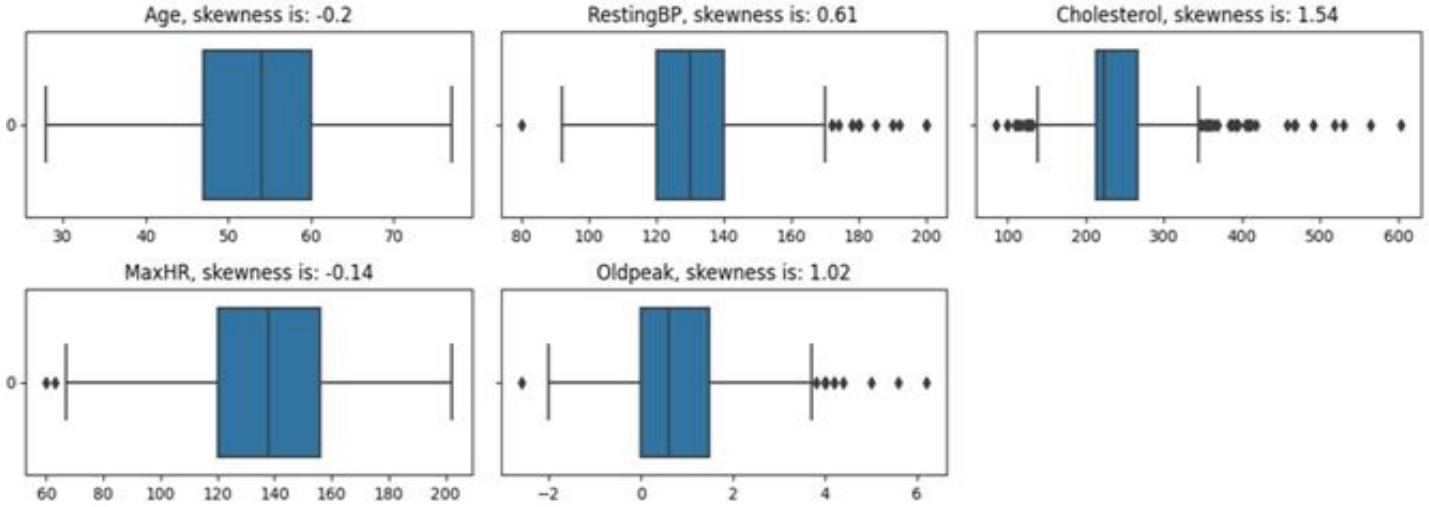

Fig. 1. Boxplot for each variable heart disease prediction data.

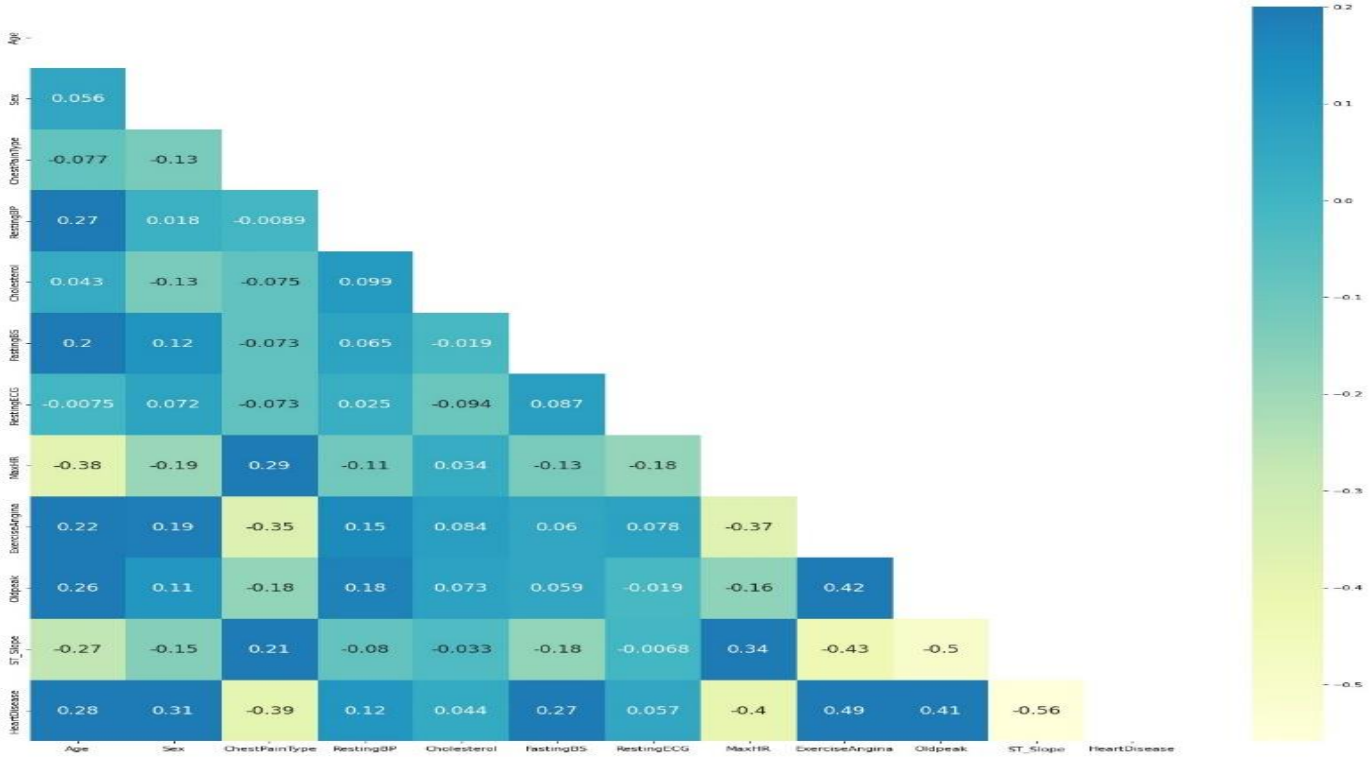

Fig. 2. Heat map diagram for heart disease prediction data.

IV. METHODOLOGY

The primary objective of this research is to forecast the likelihood of heart disease using computerized prediction techniques, offering valuable insights for both medical practitioners and patients. To accomplish this goal, we have leveraged multiple machine learning algorithms including SVM, SGD, and XGBoost, analyzing a comprehensive dataset and documenting our findings in this study report. To refine our methodology, we intend to refine the dataset by eliminating redundant information, cleaning the data, and integrating additional features such as MAP and BMI. Subsequently, the model is trained using the refined dataset. These methodological enhancements are anticipated to yield more precise outcomes and enhance model performance, as depicted in Fig. 3.

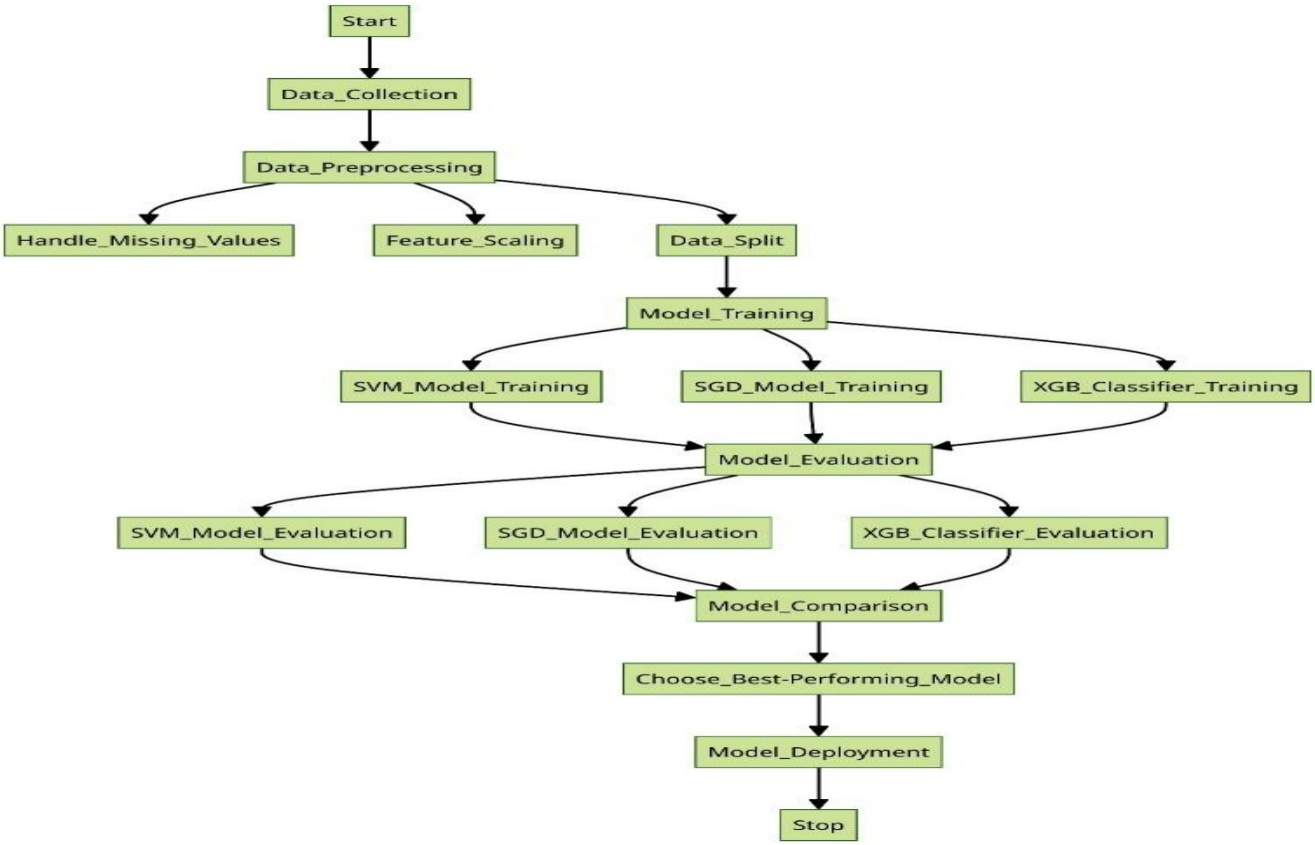

Fig. 3. Flowchart for predicting heart disease using various ML algorithm.

A. Dataset and Experimental Tools

In this research, the UCI dataset was employed for training and testing machine learning models, known for its balanced and verified nature, comprising 1127 instances and 14 attributes. Google Colab, is used for training and prediction of models. It's hardware configurations for machine learning predictions using Python typically include access to GPUs and TPUs for accelerated computation. GPU options include Nvidia Tesla K80, P4, P100, T4, and V100, providing enhanced performance for training deep learning models. RAM allocation per session typically ranges from 12GB to 25GB, supporting memory-intensive tasks.

Utilizing Google Colab, the dataset underwent visualization, analysis, and division into an 80% training set and 20% testing set, reflecting optimal performance with low bias and variance. Twelve machine learning algorithms underwent ten-fold cross-validation, with Default Hyperparameter (DHP) and Hyperparameter Optimization (HPO) techniques employed to enhance performance metrics. Quantitative and qualitative analyses were conducted to propose the most efficient model.

B. Support Vector Machine Classifier

Support Vector Machine (SVM) is a powerful and versatile supervised machine learning algorithm primarily used for classification tasks, though it can also be applied to regression and outlier detection. The key idea behind SVM is to find the optimal hyper plane that best separates the data into different classes. This hyper plane is determined by maximizing the margin, which is the distance between the hyper plane and the nearest data points from each class, known as support vectors. SVM works well in high-dimensional spaces, making it effective for problems with a large number of features. It is also robust against overfitting, especially in high-dimensional space. SVM can handle both linearly separable and non-linearly separable data by using different kernel functions such as linear, polynomial, radial basis function (RBF), and sigmoid.

One of the strengths of SVM is its ability to handle outliers effectively. Since the decision boundary is determined by support vectors, which are the data points closest to the hyperplane, outliers have little influence on the final model. Additionally, SVM allows for soft margin classification, where a penalty parameter (C) can be tuned to control the trade-off between maximizing the margin and minimizing the classification error. Despite its effectiveness, SVM can be computationally expensive, especially for large datasets. Furthermore, SVM does not provide probability estimates directly, but they can be estimated using techniques like Platt scaling or cross-validation.

The objective function for SVM can be expressed as:

$$\min_{\mathbf{w},b} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b)) \quad (1)$$

where, $\mathbf{w}$ is the weight vector, b is the bias term, C is the regularization parameter, x_i is the feature vector of the i-th training example, and y_i is the corresponding class label.

C. Stochastic Gradient Descent (SGD)

The Stochastic Gradient Descent (SGD) classifier is a widely used optimization algorithm in machine learning, specifically tailored for classification tasks. SGD optimizes model parameters iteratively, considering a single training example at each step, making it highly efficient for processing large datasets. Unlike traditional gradient descent methods, which compute gradients using the entire dataset (batch gradient descent), SGD approximates gradients using subsets of data, or even single data points, leading to faster convergence. This efficiency is particularly beneficial when dealing with datasets that may not fit into memory or when training models in real-time.

Additionally, the stochastic nature of SGD introduces randomness into the optimization process, aiding in escaping local minima and exploring a broader parameter space, potentially improving generalization. SGD also supports adaptability through techniques such as learning rate schedules and adaptive learning rates, allowing for fine-tuning of the optimization process. Furthermore, SGD is well-suited for online learning scenarios where new data arrives continuously, enabling models to be updated incrementally in response to changing data patterns. It naturally accommodates regularization techniques to prevent overfitting and improve model generalization. However, successful implementation of the SGD classifier may require careful tuning of hyperparameters such as learning rate and regularization strength, as well as consideration of mini-batch size. Despite these considerations, SGD remains a powerful and scalable approach for training classification models, offering efficiency and adaptability to diverse machine learning tasks.

The update rule for parameters in SGD is given by

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_{\theta} L(\theta^{(t)}, x_{i:t}, y_{i:t}) \quad (2)$$

where, $\theta^{(t)}$ are the parameters at iteration t, η is the learning rate, and $L(\theta^{(t)}, x_{i:t}, y_{i:t})$ is the loss function over the current mini-batch of data $x_{i:t}$ and corresponding labels $y_{i:t}$.

D. XGBoost (XGB) Classifier

XGBoost, short for eXtreme Gradient Boosting, stands as a powerhouse within the realm of machine learning classifiers, lauded for its exceptional performance across a spectrum of classification tasks. Operating within the ensemble learning paradigm, XGBoost leverages the gradient boosting framework to construct formidable predictive models. At its core, XGBoost sequentially combines multiple weak learners, typically decision trees, in a manner that iteratively corrects errors, ultimately yielding a robust and accurate classifier. Central to its efficacy is its optimization algorithm, meticulously minimizing a predefined loss function by intelligently incorporating new decision trees. Moreover, XGBoost incorporates regularization techniques, including shrinkage and tree pruning, to mitigate overfitting and enhance generalization. Its versatility shines through its ability to handle diverse data types and tasks, from numerical to categorical features, and regression to classification problems. Beyond performance, XGBoost excels in speed, efficiency, and interpretability, offering insights into feature importance and decision-making processes. Furthermore, it boasts robustness, capable of handling missing data and outliers with aplomb. With its stellar track record and widespread adoption, XGBoost stands as a stalwart choice for data scientists and practitioners seeking a reliable, high-performing classifier to tackle real-world challenges across various domains.

The objective function for XGBoost can be written as:

$$\text{Obj}(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (3)$$

where, $\Theta = \{f_k\}$ represents the set of decision trees, l is the loss function, $\hat{y}_i$ is the predicted value for the i-th instance, and Ω is the regularization term.

E. Performance Measures

The efficacy of the proposed algorithms can be assessed through several key performance measures [32, 33, and 34]:

1) Accuracy: Accuracy is calculated using the formula:

$$\textbf{\textit{Accuracy=TP+TN+FP+FNTP+TN}} \quad (4)$$

Where TP (True Positive) and TN (True Negative) represent correctly classified instances, while FP (False Positive) and FN (False Negative) denote incorrectly classified instances. The accuracy metric signifies the percentage of correctly classified instances among the total.

2) Precision: Precision measures the proportion of relevant instances among the retrieved instances and is calculated as:

$$\textbf{\textit{Precision=TP+FPTP}} \quad (5)$$

It highlights the accuracy of positive predictions made by the model.

3) Recall: Recall, also known as sensitivity, represents the proportion of relevant instances that are retrieved over the total quantity of relevant instances:

$$Recall=TP+FNTP \quad (6)$$

This metric focuses on the model's ability to identify all relevant instances.

4) Specificity: Specificity, which aligns with the definition of specificity in medical diagnostics, is computed as:

$$Specificity=TN+FPTN \quad (7)$$

It measures the proportion of true negatives identified by the model among all actual negatives.

F-measure: F-measure, also known as F1-score, considers both precision and recall and is calculated as the harmonic mean of the two: F=Precision+Recall2×Precision×Recall It provides a balanced measure of a model's performance across precision and recall.

1) Confusion matrix: The confusion matrix, is a fundamental tool in machine learning for evaluating classification model performance. It summarizes model predictions by comparing predicted labels against actual labels. Structured as a square matrix, rows and columns represent true and predicted classes, enabling detailed performance analysis. By facilitating computation of accuracy, precision, recall, and F1 score, the confusion matrix offers insights into model strengths and weaknesses. Its visual representation aids decision-making and enhances model accuracy and effectiveness.
The confusion matrix serves as a cornerstone in machine learning, empowering researchers and practitioners with invaluable insights to optimize model performance and inform decision-making processes.

V. Results Analysis and Discussion

The classification reports of Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), and XGBoost (XGB) Classifier provide comprehensive insights into their performance for predicting heart disease.

A. Classification Report Analysis

1) Support Vector Machine (SVM): As shown in Table III, SVM exhibited a training accuracy of 88.28% and a model accuracy score of 87.5%. It showcased consistent and robust performance, particularly reflected in its high precision and recall values across both classes. For class 0, SVM achieved a precision of 0.90 and recall of 0.79, indicating its ability to correctly identify instances of class 0 while minimizing false positives. Similarly, for class 1, SVM demonstrated a precision of 0.86 and recall of 0.94, suggesting its effectiveness in accurately detecting instances of class 1 while minimizing false negatives. Overall, SVM's performance metrics underscore its capability to effectively classify heart disease data with high accuracy and reliability.

TABLE III. Classification Report of Support Vector Machine

Training Accuracy : 88.28 %				
Model Accuracy Score: 87.5 %				
Classification_Report:				
precision	recall	f1-score	support	
0	0.90	0.79	0.84	76
1	0.86	0.94	0.90	108
accuracy			0.88	184
macro avg	0.88	0.86	0.87	184
weighted avg	0.88	0.88	0.87	184

2) Stochastic Gradient Descent (SGD): As shown in Table IV, SGD yielded a training accuracy of 83.65% and a model accuracy score of 84.24%. While SGD's performance is commendable, it falls slightly behind SVM in terms of accuracy and precision. Notably, SGD exhibited slightly lower precision for class 0 compared to SVM, with a precision of 0.81 and recall of 0.82 for class 0, indicating a marginally higher rate of false positives. However, SGD's precision for class 1 was relatively higher at 0.87, with a recall of 0.86, suggesting its effectiveness in accurately identifying instances of class 1. Overall, SGD demonstrates satisfactory performance but may require further optimization to achieve results comparable to SVM.

TABLE IV. Classification Report of Stochastic Gradient Descent

Training Accuracy : 83.65 %				
Model Accuracy Score: 84.24 %				
Classification_Report:				
precision	recall	f1-score	support	
0	0.81	0.82	0.81	76
1	0.87	0.86	0.87	108
accuracy			0.84	184
macro avg	0.84	0.84	0.84	184
weighted avg	0.84	0.84	0.84	184

TABLE V. Classification Report of XGB Classifier

Training Accuracy : 100.0 %				
Model Accuracy Score : 85.87 %				
Classification_Report:				
precision	recall	f1-score	support	
0	0.85	0.80	0.82	76
1	0.87	0.90	0.88	108
accuracy			0.86	184
macro avg	0.86	0.85	0.85	184
weighted avg	0.86	0.86	0.86	184

3) XGBoost (XGB) Classifier: As shown in Table V, XGB Classifier showcased a perfect training accuracy of 100.0%, yet achieved a model accuracy score of 85.87%. Despite its perfect training accuracy, XGB Classifier's model accuracy score indicates potential overfitting, suggesting that it may not generalize well to unseen data. However, XGB Classifier's precision and recall values were comparable to SVM, with a precision of 0.85 and recall of 0.80 for class 0, and a precision of 0.87 and recall of 0.90 for class 1. These metrics suggest XGB Classifier's effectiveness in accurately classifying heart disease data, although further investigation into its generalization capabilities is warranted.

B. Comparison Summary of ML Models for heart disease prediction

The Table VI presents a detailed comparison of three machine learning models - Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), and XGBoost (XGB) Classifier - based on various performance metrics for predicting heart disease. Each row corresponds to a specific model, while each column represents a different metric evaluated.

TABLE VI. COMPARISON OF THREE MACHINE LEARNING MODELS

Metric	Training Accuracy	Model Accuracy Score	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1-score (Class 0)	F1-score (Class 1)	Support (Class 0)	Support (Class 1)
SVM	88.28%	87.5%	0.90	0.86	0.79	0.94	0.84	0.90	76	108
SGD	83.65%	84.24%	0.81	0.87	0.82	0.86	0.81	0.87	76	108
XGB Classifier	100.0%	85.87%	0.85	0.87	0.80	0.90	0.82	0.88	76	108

C. Analyzing Through Confusion Matrix for Predicting Heart Disease

The confusion matrix (see Fig. 4 to Fig. 6) helps us see how well boosting models spot mistakes when predicting heart problems. It looks at what really happened versus what the models predicted, using four things: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). This confusion matrix, shown in Fig. 8, helps us figure out how accurate the boosting models are in spotting errors when predicting heart issues. It compares what actually occurred with what the models guessed, using four important measures: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

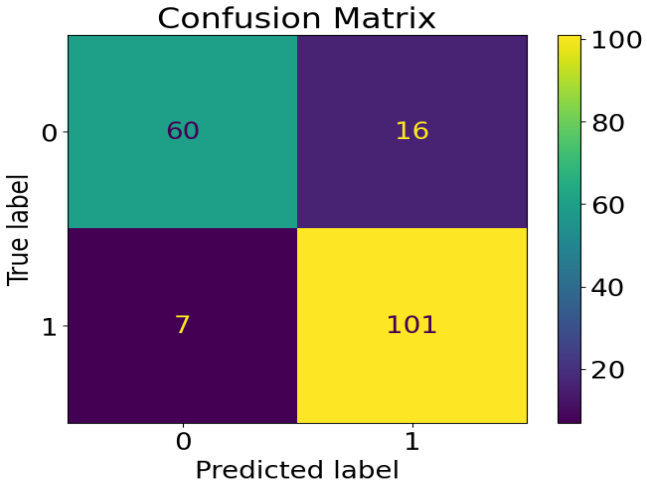

Fig. 4. Confusion matrix for support vector classifier.

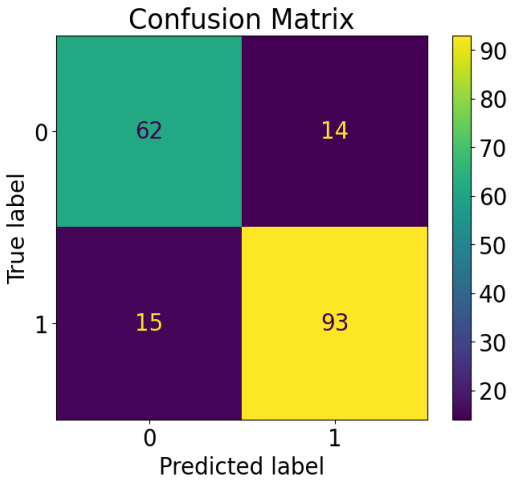

Fig. 5. Confusion matrix for stochastic gradient descent classifier.

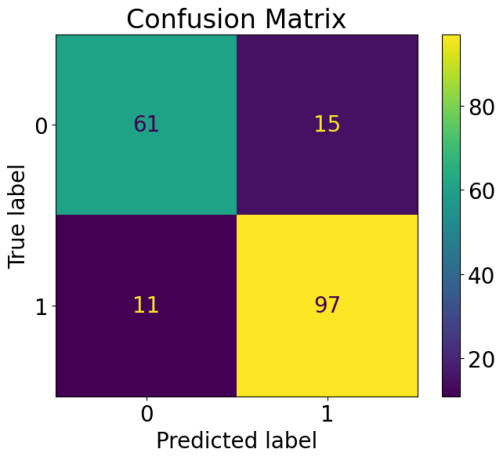

Fig. 6. Confusion matrix for XGBoost classifier.

D. Analyzing Through ROC Curve for Predicting Heart Disease

Furthermore, ROC (Receiver Operating Characteristic) curves have been generated and depicted in Fig. 7- 9 to delve deeper into the analysis of each machine learning model. These curves offer a visual representation of the classifier performances and illustrate the tradeoff between the true positive rate and false positive rate across various classification thresholds.

The area under the curve (AUC) of the ROC curve serves as a metric to gauge the model's capability to differentiate between classes, with values ranging from zero to one. A higher AUC indicates a greater ability to accurately classify instances. As the AUC approaches one, the model demonstrates enhanced capability in separating the classes, signifying superior performance.

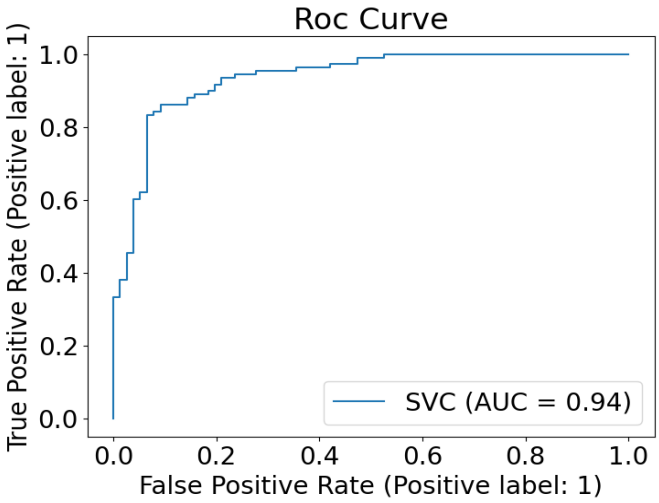

Fig. 7. ROC curve for support vector classifier.

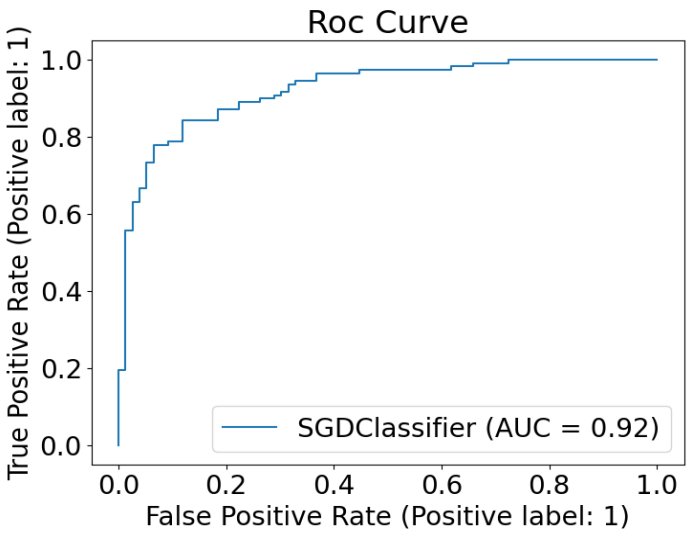

Fig. 8. ROC curve for stochastic gradient descent classifier.

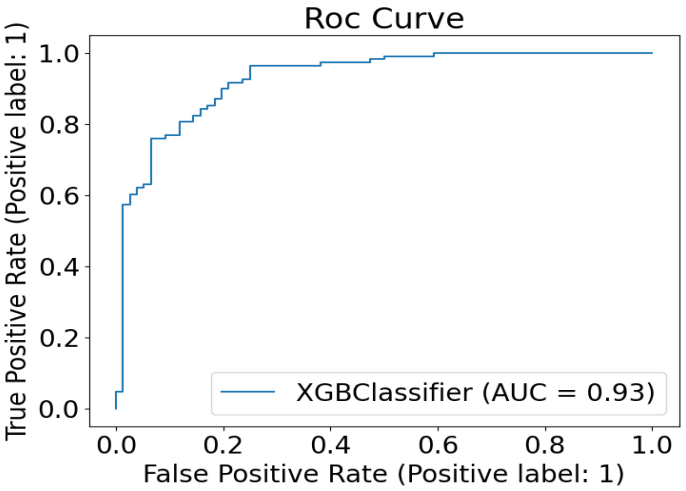

Fig. 9. ROC curve for XGBoost classifier.

E. Analyzing Through Precision Recall Curve for Predicting Heart Disease

Also, Precision-Recall curves have been generated and depicted in Fig. 10 -12 to provide deeper insights into the analysis of each machine learning model. These curves offer a visual representation of the classifier performances and illustrate the trade off between precision and recall across various classification thresholds.

The Precision-Recall curve showcases the relationship between the precision (positive predictive value) and recall (sensitivity) of the model as the classification threshold varies. It provides a comprehensive view of how well the model identifies positive instances while minimizing false positives.

Unlike ROC curves, which focus on the tradeoff between true positive rate and false positive rate, Precision-Recall curves emphasize the balance between precision and recall. They are particularly useful when dealing with imbalanced datasets where one class significantly outweighs the other.

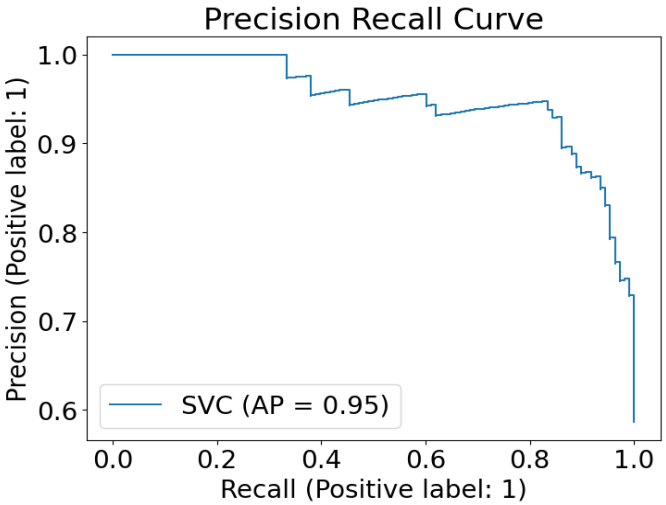

Fig. 10. Precision recall curve for support vector classifier.

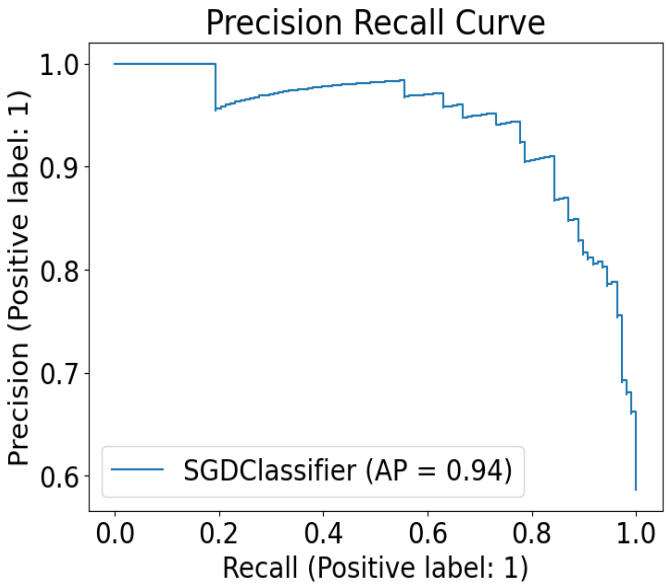

Fig. 11. Precision recall for stochastic gradient descent classifier.

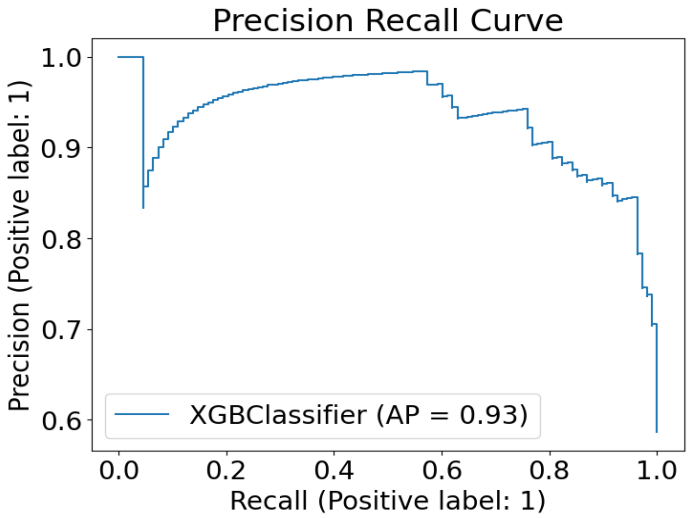

Fig. 12. Precision recall curve for XGBoost classifier.

The area under the Precision-Recall curve (AUC-PR) serves as a metric to evaluate the model's performance. A higher AUC-PR indicates better precision and recall tradeoff, suggesting superior model performance in accurately identifying positive instances while minimizing false positives. As with ROC curves, an AUC-PR value closer to one signifies enhanced model performance.

VI. Conclusion and Future Scope

In this study, we compared the performance of Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), and XGBoost machine learning techniques for heart disease prediction. Our results reveal distinct characteristics and capabilities of each model. SVM exhibited remarkable performance with a training accuracy of 88.28% and a model accuracy score of 87.5%. Its high precision and recall values across both classes indicate its ability to effectively classify heart disease data. Notably, SVM demonstrated a precision of 0.90 and recall of 0.79 for class 0, and a precision of 0.86 and recall of 0.94 for class 1, underscoring its reliability in minimizing false positives and false negatives.

While SGD demonstrated commendable performance with a training accuracy of 83.65% and a model accuracy score of 84.24%, it slightly trailed behind SVM in terms of accuracy and precision. Although SGD exhibited a relatively higher precision for class 1, further optimization may be required to achieve results comparable to SVM. XGBoost Classifier showcased perfect training accuracy but achieved a model accuracy score of 85.87%, suggesting potential overfitting. Nonetheless, its precision and recall values were comparable to SVM, indicating its effectiveness in accurately classifying heart disease data. However, further investigation into its generalization capabilities is warranted to ensure reliable performance in real-world scenarios.

Overall, our findings demonstrate SVM's robustness and effectiveness in heart disease prediction, followed by SGD and XGBoost Classifier. Further research may focus on optimizing SGD and investigating XGBoost Classifier's generalization capabilities to enhance their performance in clinical applications.

In future research, we aim to utilize the findings presented here to develop a robust prediction system aimed at enhancing medical treatment efficacy and reducing costs using other efficient machine learning algorithms.

References

[1] Jagannathan, Ram, Shivani A. Patel, Mohammed K. Ali, and KM Venkat Narayan, "Global updates on cardiovascular disease mortality trends and attribution of traditional risk factors," Current diabetes reports, vol. 19, no. 7, pp 1-12, 2019.

[2] Krittanawong, Chayakrit, HongJu Zhang, Zhen Wang, Mehmet Aydar, and Takeshi Kitai, "Artificial intelligence in precision cardiovascular medicine," Journal of the American College of Cardiology, vol. 69, no. 21, pp 2657-2664, 2017.

[3] Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp 1347-1358, 2019.

[4] Yusuf, S., Joseph, P., Rangarajan, S., Islam, S., Mente, A., Hystad, P., Brauer, M., Kutty, V.R., Gupta, R., Wielgosz, A. and AlHabib, K.F., "Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study," The Lancet, vol. 395, no. 10226, pp 795-808, 2020.

[5] Benjamin EJ, Muntner P et al. Alonso, Alvaro, –Heart Disease and Stroke Statistics–2019 Update: A Report From the American Heart Association, Circulation, 2019;vol. 139, no. 10

[6] Murthy H, Meenakshi M, –Dimensionality reduction using neuro-genetic approach for early prediction of coronary heart disease, in International Conference on Circuits, Communication, Control and Computing (I4C), 2014; pp. 329–332.

[7] Bashir S, Khan ZS, Khan FH, Anjum A, Bashir K. Improving Heart Disease Prediction Using Feature Selection Approaches, in 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 2019; pp. 619–623.

[8] Gavhane A, Kokkula G, Pandya I, Devadkar PK. –Prediction of Heart Disease Using Machine Learning, in Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA 2018, 2018; pp. 1275–1278.

[9] Ismaeel S, Miri A, Chourishi D. Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis, in 2015 IEEE Canada

International Humanitarian Technology Conference. IHTC. 2015;2015:1–3.

[10] Ekiz S, Erdogmus P. Comparative study of heart disease classification, in 2017 Electric Electronics. Computer Science, Biomedical Engineerings' Meeting, EBBT. 2017;2017:1–4.

[11] Kanikar P, Shah DR, Prediction of cardiovascular diseases using support vector machine and Bayesien classification, International Journal of Computer Applications (0975 – 8887) Volume 156 – No 2, December 2016.

[12] Shorewala, V. Early detection of coronary heart disease using ensemble techniques. Inform. Med. Unlocked 2021, 26, 100655.

[13] Maiga, J.; Hungilo, G.G.; Pranowo. Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. In Proceedings of the 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, 24–25 October 2019; pp. 45–48.

[14] Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. Eur. J. Mol. Clin. Med. 2020, 7, 1638–1645.

[15] Ouf, S.; ElSeddawy, A.I.B. A proposed paradigm for intelligent heart disease prediction system using data mining techniques. J. Southwest Jiaotong Univ. 2021, 56, 220–240.

[16] Khan, I.H.; Mondal, M.R.H. Data-Driven Diagnosis of Heart Disease. Int. J. Comput. Appl. 2020, 176, 46–54.

[17] Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. J. Exp. Theor. Artif. Intell. 2019, 31, 565–583.

[18] Pouriyese, M.; Parvinnia, S.; Sabeti, E.; Gamaarachchi, H.; Sadoughian, M.; Farhadi, F.; Iqbal, Q. A Comprehensive Investigation and Machine Learning-based Diagnostic Prediction of Heart Disease. Comput. Biol. Med. 2021, 135, 104551.

[19] Fathima, N.; Thileeban, S. Prediction of Heart Disease Using Machine Learning Algorithms. In Proceedings of the 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 12–14 March 2020; pp. 66–71.

[20] SaiSudheer, M.; Niharika, Y.; Janga, N.V. Heart Disease Prediction Using Machine Learning Techniques. In Proceedings of the 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 27–29 January 2021; pp. 1–6.

[21] Taneja, A. Heart Disease Prediction Using Machine Learning on Cloud Platform. In Proceedings of the 2020 International Conference on Intelligent Engineering and Management (ICIEM), London, UK, 17–19 June 2020; pp. 327–331.

[22] Diwakar, M.; Sivakumar, V.S.; Nedunchezhian, R. Prediction of Heart Disease Using Machine Learning Techniques. Int. J. Eng. Adv. Technol. 2019, 8, 506–511.

[23] Kaur, H.; Kumar, R.; Kumari, V. Heart Disease Prediction Using Machine Learning Techniques. In Proceedings of the 2020 International Conference on Power Electronics &IoT Applications in Renewable Energy and Its Control (PARC), Mathura, India, 28–29 February 2020; pp. 1–6.

[24] Nahar, J.; Imam, T.; Tickle, K.S.; Chen, Y.P.P. Computational intelligence for heart disease diagnosis: A medical knowledge-driven approach. Expert Syst. Appl. 2013, 40, 96–104.

[25] Amin, M.S.; Chiam, Y.K.; Varathan, K.D. Identification of significant features and data mining techniques in prediction of heart disease. Telemat. Inform. 2019, 36, 262–274.

[26] Raza, K. An Optimization Strategy for Heart Disease Prediction. Comput. Intell. Neurosci. 2020, 2020, 8863923.

[27] Vembanki, S.; Pilikan, S.; Padte, R.; Kanimozhhi, P. Heart Disease Diagnosis Using Ensemble Machine Learning Techniques. In Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 4–6 February 2021; pp. 1–5.

[28] Yadav, S.; Shukla, S. Analysis of k-Fold Cross-Validation over Hold-Out Validation on Coimbatore Dataset using WEKA Tool. Int. J. Inf. Technol. 2016, 8, 1054–1058.

[29] Patel, J.; Upadhyay, D.; Patel, S. Heart Disease Prediction Using Machine Learning and Data Mining Technique. Heart Dis. 2015, 7, 129–137.

[30] Sultana, M.; Haider, A.; Uddin, M.S. Analysis of Data Mining Techniques for Heart Disease Prediction. In Proceedings of the 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, Bangladesh, 22–24 September 2016; pp. 1–5.

[31] Altan, G.; Karasu, S.; Bekiros, S. Digital Chest Drainage and Dissolved Air Flotation for Metal Plating Sludges. Resour. Conserv. Recycl. 2019, 146, 25–33.

[32] Arif, M.S.; Mukheimer, A.; Asif, D. Enhancing the early detection of chronic kidney disease: A robust machine learning model. Big Data Cogn. Comput. 2023, 7, 144.

[33] Chang, V.; Bhavani, V.R.; Xu, A.Q.; Hossain, M. An artificial intelligence model for heart disease detection using machine learning algorithms. Healthc. Anal. 2022, 2, 100016.

[34] Neshat, M.; Ahmedb, M.; Askarid, H.; Thilakaratnee, M.; Mirjalilia, S. Hybrid Inception Architecture with Residual Connection: Fine-tuned Inception-ResNet Deep Learning Model for Lung Inflammation Diagnosis from Chest Radiographs. arXiv 2023, arXiv:2310.02591.

[35] Jha, P., Dembla, D., Dubey, W. "Implementation of Machine Learning Classification Algorithm Based on Ensemble Learning for Detection of Vegetable Crops Disease International Journal of Advanced Computer Science and Applications, 2024, 15(1), pp. 584–594

[36] Jha, Pradeep, Deepak Dembla, and Widhi Dubey 2024. "Implementation of Transfer Learning Based Ensemble Model Using Image Processing for Detection of Potato and Bell Pepper Leaf Diseases." Article. International Journal of Intelligent Systems and Applications in Engineering 12 (8s): 69–80.

[37] Jha, Pradeep, Deepak Dembla, and Widhi Dubey. 2023. "Comparative Analysis of Crop Diseases Detection Using Machine Learning Algorithm." Conference paper. Proceedings of the 3rd International Conference on Artificial Intelligence and Smart Energy, ICAIS 2023. Institute of Electrical; Electronics Engineers Inc. https://doi.org/10.1109/ICAIS56108.2023.10073831.

[38] Jha, Pradeep, Deepak Dembla, and Widhi Dubey. 2023. "Crop Disease Detection and Classification Using Deep Learning-Based Classifier Algorithm." Conference paper. Edited by Rathore V. S., Piuri V., Babo R., and Ferreira M. C. Lecture Notes in Networks and Systems 682 LNNS: 227–37. https://doi.org/10.1007/978-981-99-1946-8_21.

[39] Jha, Pradeep, Deepak Dembla, and Widhi Dubey 2023. "Deep Learning Models for Enhancing Potato Leaf Disease Prediction: Implementation of Transfer Learning Based Stacking Ensemble Model." Article. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-023-16993-4.

[40] Meshram, Amita, and Deepak Dembla. 2023. "MCBM: IMPLEMENTATION OF MULTICLASS AND TRANSFER LEARNING ALGORITHM BASED ON DEEP LEARNING MODEL FOR EARLY DETECTION OF DIABETIC RETINOPATHY." Article. ASEAN Engineering Journal 13 (3): 107–16. https://doi.org/10.11113/aej.V13.19401.

[41] Meshram, Amita, and Deepak Dembla 2023. "Multistage Classification of Retinal Images for Prediction of Diabetic Retinopathy-Based Deep Learning Model." Conference paper. Edited by Rathore V. S., Piuri V., Babo R., and Ferreira M. C. Lecture Notes in Networks and Systems 682 LNNS: 213–26. https://doi.org/10.1007/978-981-99-1946-8_20.

[42] Meshram, Amita, Deepak Dembla, and A. Anooja. 2023. "DEVELOPMENT AND ANALYSIS OF DEEP LEARNING MODEL BASED ON MULTICLASS CLASSIFICATION OF RETINAL IMAGE FOR EARLY DETECTION OF DIABETIC RETINOPATHY." Article. ASEAN Engineering Journal 13 (3): 89–97. https://doi.org/10.11113/aej.V13.19256.