"Machine Learning Methods For Predicting Cardiovascular Risk Elements"

Devjot Singh
Department of Computer Science & Engineering
Chandigarh University
Mohali, Punjab-140413, INDIA
devjotsingh50@gmail.com

Rounika Acharya
Department of Computer Science & Engineering
Chandigarh University
Mohali, Punjab-140413, INDIA
rounikaacharya@gmail.com

Aditya Aman
Department of Computer Science & Engineering
Chandigarh University
Mohali, Punjab-140413, INDIA
aditya12601aman@gmail.com

Dipti Mahato
Department of Computer Science & Engineering
Chandigarh University
Mohali, Punjab-140413, INDIA
diptimahato910@gmail.com

Mohd Shahid
Department of Computer Science & Engineering
Chandigarh University
Mohali, Punjab-140413, INDIA
md01shahid01@gmail.com

Bhanu Devi
Department of Computer Science & Engineering
Chandigarh University
Mohali, Punjab-140413, INDIA
bhanu.e12251@cumail.in

***Abstract*— Disease of heart is a crucial medical condition. This needs timely and accurate intervention for successful treatment outcomes. It's equally vital to recognize symptoms of heart disease early. This can significantly enhance health outcomes. It also can prevent severe complications. In this research paper we delve into use of Machine Learning (ML) techniques. The focus is early detection of heart disease symptoms. The study concentrates on creation of predictive models. The goal is to assess risk of developing heart disease in an individual. The required analysis is a mix of physiological and clinical parameters. They originate from health data spanning past and present. Leveraging ML is the target. The aim is to bolster early diagnosis. The paper also aims for more effective prevention against heart disease. In this research we made use of one dataset. The dataset is from Kaggle. It was used to evaluate accuracy of different machine learning algorithms. The best accuracy achieved is 86.578%. The dataset's parameters are Age, Sex, Is Smoking, Cigarettes Per Day, BP Medicine, Prevalent Stroke. Furthermore, there is Prevalent Hypertension, Diabetes, Total Cholesterol. Followed by Systolic BP, Diastolic BP, BMI, Heart Rate and Glucose.**

Keywords*— Heart Disease, Prediction Model, Random Forest Classifier, AdaBoost Classifier, Gradient Boosting Classifier, Calibrated Classifier, Logistic Regression, Decision Tree Classifier, K-Nearest Neighbors (KNN).

I. INTRODUCTION

Heart disease holds its place as a leading cause of mortality worldwide. It stands as a notable challenge to healthcare systems. Such systems emphasize the need for early detection and prompt intervention. Initial symptoms of this disease can be elusive. It is of great importance to spot and tackle these concerns early. This must happen before the symptoms grow into more serious conditions. Life-threatening conditions like cardiac arrest. The prevalence of heart disease is seeing an alarming increase. Younger populations are bearing the brunt of it more and more. A disconcerting trend emerges in recent data. Data from EMRI 108 emergency services. There is a 28% surge in heart-related emergencies in 2023 compared to last year. By September 25, 46,155 emergency calls for cardiac issues had been recorded. That is an average of 173 calls per day or 7 calls every hour. Lifestyle factors are implicated in this surge. Factors like consuming unhealthy food widespread lack of physical activity and the increase in central obesity. Doctor Gajendra Dubey is an interventional cardiologist at the UN Mehta Institute. He underlines a key point. Modern lifestyle dramatically escalates risks of heart disease.

To combat the rising menace of heart disease Healthcare field is increasingly adopting sophisticated technologies such as Artificial Intelligence And Machine Learning (AI/ML). .These technologies proffer weighty Tools for early detection and prevention.. They Make use of vast quantities of data.. The aim is to define trends and anticipate potential cardiac troubles Before they turn fatal. AI/ML models can Evaluate Intricate datasets. These datasets include patient medical documents lifestyle factors ,and genetic information . They offer precise forecasts of heart disease risk. The Blending of AIML in heart disease prediction signifies a major step in healthcare. It introduces a chance to alter our approach towards prevention and therapy. Through allowing early intervention technologies can decrease the burden of heart disease on persons. This technology also helps in lessening the stress on healthcare systems at large. It improves patient results and rescues life .This research paper explores the use of AI/ML in predicting heart disease It also discusses methodologies challenges and potential future directions .The field is Rapidly growing.

II. RELATED WORK

Author of [1] contrasts diverse machine learning classifiers. These include Random Forest Logistic Regression, Support Vector Machines Naive Bayes, Decision Tree and k-Nearest Neighbors. The examination employed four datasets from Kaggle. The Highest accuracy rate of 82.35% Was attained with the heart disease set. 7459% accuracy was achieved with the heart disease 2020 set .686 precision was Accomplished with the Framingham dataset. Author of [2] Showcases effectiveness of Convolutional Neural Networks (CNNs). These are for classifying medical data. These networks were trained on a heart disease dataset. Both single and multilayer networks were employed. The CNN managed a 99% accuracy. In contrast, other classifiers were surpassed. The k-Nearest Neighbors (KNN) tool displayed the least accurate performance. Paper [3] substantiates an inventive machine learning (ML) score This ML score combines heart rate variability (HRV) data. The Purpose is to Prioritize extremely unwell patients. The setting is the emergency department. It was Tested against the Modified Early Warning Score (MEWS). The Performance of the novel ML score was contrasted against MEWS. The ML score predicted cardiac

Arrest within 72 hours, with an Area Under the Receiver Operating Characteristic (AUROC) of 0.781. Moreover, for in hospital death the same score had an AUROC of 0.741. These numbers Were superior to MEWS which showed AUROCs of 0.680 and 0.693 respectively. The ML score outperformed MEWS performance Metrics in both instances. Paper [4] leveraged the Medical Information Mart for Intensive Care IV (MIMIC-IV) database. Paper used it for crafting a predictive nomogram. The Goal was to predict in-hospital mortality among cardiac arrest patients in the ICU. The nomogram found a high area under the curve (AUC) of 0.7912. This came with a broad Net Benefit Threshold range. It also boasted significant net benefit. These Factors combine to make the tool an invaluable Asset for clinical decision-making. Research piece [5] brings DeepCARS™ to the fore. This is an AI-based tool. It leverages deep learning and vital sign data. DeepCARS™ outperformed traditional early warning scores. MEWS News and SPTTS fell short. In predicting in-hospital cardiac arrest (IHCA) and unplanned ICU Transfers (UIT) DeepCARS™ was superior. DeepCARS™ weaved timely alarms. This led to swift Rapid Response Team (RRT) interventions. Proving superior Prediction performance in Clinical settings. DeepCARS™ is An AI tool based on deep learning. It uses patient's vital Signs to predict in-hospital cardiac arrests IHCA and unplanned ICU transfers (UIT). This tool Outperforms MEWS NEWS, and SPTTS. It Provides timely alarms leading to RRT interventions. Superior prediction performance is demonstrated in Real Clinical settings. Paper [6] improved prediction performance by focusing on specific input parameters (e.g., maximum SBP, minimum SBP) and excluding others (e.g., sex, DBP, AST). Correlation analysis identified attributes influencing cardiac arrest, with machine learning and deep learning algorithms, including decision tree, random forest, logistic regression, long short-term memory (LSTM), gated recurrent unit (GRU), and the LSTM–GRU hybrid model. The LSTM model achieved a positive predictive value of 85.92% and sensitivity of 89.70%. Paper [7] introduces the Feasible Artificial Intelligence with Simple Trajectories for Predicting Adverse Catastrophic Events (FAST-PACE) solution, which predicts cardiac arrest or acute respiratory failure 1 to 6 hours in advance. FAST-PACE outperforms traditional warning scores like MEWS and NEWS, achieving an area under the receiver operating characteristic curve of 0.886 for cardiac arrest and 0.869 for respiratory failure 6 hours prior to events. It also demonstrated superior prediction performance compared to MEWS and NEWS. Paper [8] describes a retrospective cohort study analyzing data from 52,131 patients admitted to two hospitals between June 2010 and July 2017. A recurrent neural network, trained on data from June 2010 to January 2017 and tested on data from February to July 2017, demonstrated high sensitivity and a low false-alarm rate in detecting cardiac arrest. Originally in [9] A study was conducted. Five strategies were devised that predicted heart disease. All these techniques Were tested on an in-built dataset contained in the research. The techniques included were the following: Naive Bayes, k-Nearest Neighbor (KNN), Decision Tree, Artificial Neural Network (ANN), and Random Forest. This study uncovered that Naive Bayes Achieved the Highest Accuracy level. It Reached 88%. In The paper [10] the focus is on creating a heart disease prediction system. This system Is meant to evaluate the probability of a patient Being diagnosed with heart disease. The prediction is based on the patient's medical history. In this Process different machine Learning methods were used. This included logistic Regression among others. K-Nearest Neighbors (KNN) was Also used. The goal was to classify and forecast heart disease. The overall approach had a goal. That goal was to refine the precision of forecasting. KNN and logistic regression model were seen as notable. They Achieved results That demonstrated potential. This potential was for Enhanced risk valuations. It was Compared To the older methods like Naive Bayes.

III. ML Classification Techniques

After evaluating 22 different machine learning models, we found that four techniques stand out with the highest accuracy for our heart disease prediction model. These top-performing methods are decision trees, support vector machines, neural networks, and random forests.

A. Logistic Regression

Logistic Regression [6] is a statistical technique which is used for binary classification tasks, which estimates the probability of a binary result based on various input features. Its simplicity and interpretability make it well-suited for datasets such as CHDD, where the connection between predictors and the outcome is likely linear. Furthermore, logistic regression can be augmented with regularization methods like L1 (lasso) or L2 (ridge) to enhance model performance and mitigate overfitting, especially when confronted with a significant number of features.

B. CatBoost

CatBoost Is a gradient boosting technique. Like others, it is developed to Manage categorical data effectively It manages this without Extensive preprocessing. In A sequence it constructs decision trees and each new tree addresses the errors made by its predecessor trees. This method Is advantageous for the CHDD dataset. It allows the model to Grasp intricate Relationships among features which Enhances predictive accuracy.

CatBoost is noted for resilience against overfitting. It is also exceptional for working efficiently with large datasets.

C. Random Forest

Random Forest [6] is a machine learning Technique is Random Forest. During the training process, it Constructs several decision trees. Then it merges their outputs. This enhances prediction accuracy and stability.

It's particularly well Suited for the CHDD dataset. This is because it effectively manages large voluminous, high-dimensional datasets. Another advantage is it's less prone to overfitting. By averaging the output of individual trees, Random Forest minimizes variance. It boosts the model's ability for generalization.

D. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is as A classification system. It recognizes linear mix of features. . It has a capability to separate between various categories well. This method is especially useful for datasets like CHDD as these types of datasets have separable Classes that Require dimensionality reduction. It's all about improving the space between different classes' means. We aim to minimize variance within each class as because this will enhance model's ability of prediction. The model can accurately classify instances even when a large Number of predictors are used.

E. Calibrated Classifier

Calibrated classifier is used to sharpen probability estimates from base classifier. It ensures predicted probabilities align better with the actual likelihood of each class. This method is of particular benefit for the CHDD dataset. This is true especially when there are issues with data imbalance.

The classifier is calibrated for models like SVM or Random Forest. It delivers more dependable probability outputs. These outputs form an essential part of making well-informed clinical decisions. The decisions are based on anticipated risks.

IV. METHODOLOGY

Here is a section that discusses methodology. This is related to predicting heart disease. The focus is on using Artificial Intelligence and Machine Learning. Abbreviated as AI/ML. Research adopts a structured procedure. It starts with collection of data. Then data preprocessing follows. Next comes feature selection. After that model development occurs. The research continues with an evaluation stage. The final stage is performance analysis.

A. Data Collection

The dataset is sourced from Kaggle[16] which has been used in our research. The 15 features of the dataset are as follows:

S. No.	ATTRIBUTES		
	Attribute	***Desc.***	***Mean Value***
1	age	In years	49.542
2	sex	Male, Female	0.432
3	Is_smoking	Yes, No	0.497
4	cigsPerDay	No. of cigarettes per day	9.069
5	BPMeds	Yes, No	0.029
6	prevalentStroke	“0” for negative and “1” for positive	0.006
7	prevalentHYP	“0” for negative and “1” for positive	0.315
8	diabetes	Yes, No	0.025
9	totChol	“0” for negative and “1” for positive	237.074
10	sysBP	Systolic Blood Pressure	132.601
11	diaBP	Diastolic Blood Pressure	82.883
12	BMI	Body Mass Index	25.794
13	heartRate	No. of contractions of heart per minute	75.977
14	glucose	Blood Glucose level	82.086
15	Target	Output Class	

In this study, a dataset containing 3,389 instances has been utilized, As the dataset is relatively small it could limit generalizability and increase the risk of overfitting. The dataset comprises several attributes, and Table I presents the mean values of each. Missing data is present in the dataset under the attributes like glucose, BMI, CigsPerDay and appropriate preprocessing techniques, such as statistical imputation, have been applied to handle these missing values effectively.

The dataset is categorized into two classes: Class 1 indicates "tested positive for the disease," while Class 0 signifies "tested negative for the disease." For model training and evaluation, the dataset has been split, with 80% designated as training data and the remaining 20% as testing data.

B. Data Preprocessing

Data preprocessing is a crucial step in ensuring the dataset is suitable for building robust and reliable models. In this study, different preprocessing methods are used to manage missing data and normalize the dataset and prepare it for model training.

1. **Handling Missing Data**: The dataset presented a challenge that is the missing values which were evident in some attributes. To address these missing values, statistical imputation techniques were used. In case of numerical columns, missing data was not left vacant and was replaced by the median value. This was the median value of the relevant attribute. The method has a clear purpose. It ensures that extreme values do not alter the distribution. This helps to maintain integrity of dataset. By avoiding skewness, robustness is preserved.
2. **Encoding Categorical Variables**: Included In The dataset were categorical variables .The variables Included "sex" and the values for "sex" were 'F' and 'M'. Another variable Was "is_smoking" and the 'is_smoking' values were 'YES' And 'NO'. These variables are encoded as binary features.

   For example ,'F' is encoded as 0 and 'M' is Encoded as 1 and 'NO' is encoded as 0, similarly 'YES' is encoded as 1.
3. **Feature Scaling**: The dataset has attributes which have varying ranges. For example, blood pressure, Cholesterol, Glucose levels are some examples. So, to handle this feature scaling is needed. Numerical attributes are scaled using Standard Scaler. This method normalizes data. It first subtracts the mean value and then it adjusts to unit variance. Scaling process is crucial to ensure smooth operations of algorithms. It is especially important for algorithms using distance calculations. Logistic regression, support vector machines and neural networks are some examples which require this.
4. **Splitting the Dataset**: After preprocessing, the dataset is divided two parts one is used for training and other one is used for testing purposes. Division is done in such a way that 80% of the dataset (2,711 instances) is used for training and the remaining 20% (678 instances) is used for testing. This splitting allows the model to be trained on a large portion of the data while being evaluated on a separate set of unseen examples to test its ability of prediction.

The processed dataset is then ready for building model and evaluation.

C. Building Model

Google Colab platform is used for building and evaluating the prediction model. It is an open-source environment which provides its users a GPU support for machine learning and deep learning purposes. It allows users to integrate with

Python libraries such as Scikit-learn and TensorFlow, providing a user-friendly interface for model creation.

Standard data mining tasks such as classification, feature scaling, and performance evaluation are done using libraries like Scikit-learn's. The study includes testing various models like Random Forest, Support Vector Machines (SVM), Logistic Regression, Gradient Boosting, and Neural Networks and some others. The ability and performance of all these models are compared using multiple accuracy measures to find the model with best accuracy., which include:

1. **Precision (Positive Predictive Value)**: Precision is measured to get the proportion of correctly predicted positive values out of all values predicted as positive by the prediction model. It is calculated as:

   Precision = (Number of correctly predicted positives)/(Number of true positives + False predicted positives).

2. **Recall**: Recall represents the proportion of correctly predicted positive instances out of all actual positive instances. It is given by:

   Recall= (True positive values)/(True positive values + False negative values).

3. **Accuracy**: Accuracy is calculated to an idea of overall effectiveness of the prediction model. It is measured by taking the proportion of total correct predictions .These include both Positive and negative outcomes. This proportion is out of total number of instances. The formula for accuracy is:

   Accuracy = (True Positives + True Negatives) / (Total Number of Instances)

4. **F1 Score**: F1 score calculated by taking harmonic mean of Precision and recall. This provides a single metric which balances Both precision and recall values and it is helpful in the cases of imbalanced datasets. It is calculated as:

   F1 score= 2*(Precision*Recall)/(Precision + Recall)

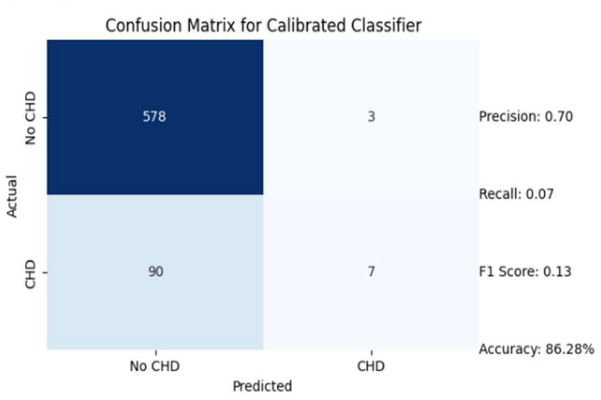

Fig. 1. Confusion Matrix for Calibrated Classifier

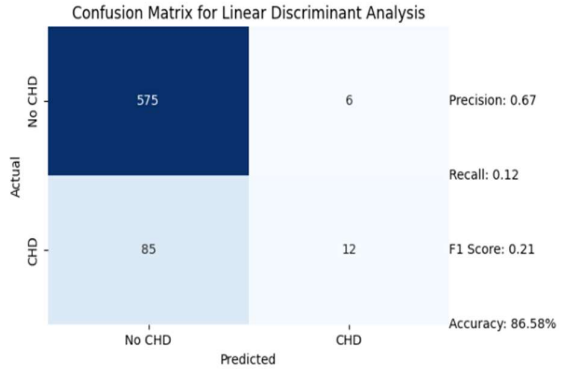

Fig. 2. Confusion Matrix for Linear Discriminant Analysis

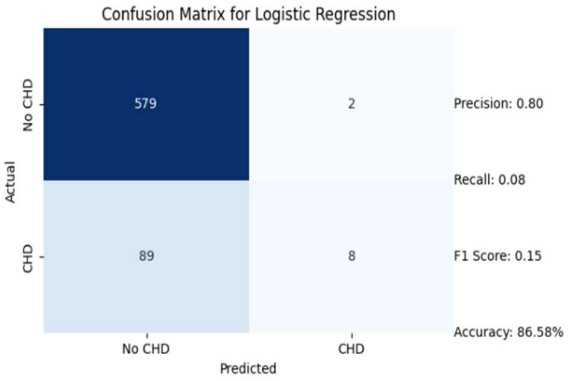

Fig. 3. Confusion Matrix for Logistic Regression

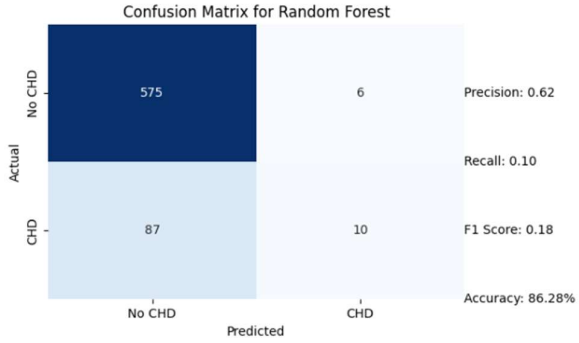

Fig. 4. Confusion Matrix for Random Forest

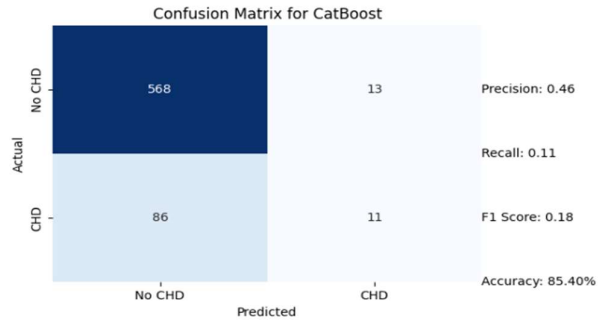

Fig. 5. Confusion Matrix for CatBoost

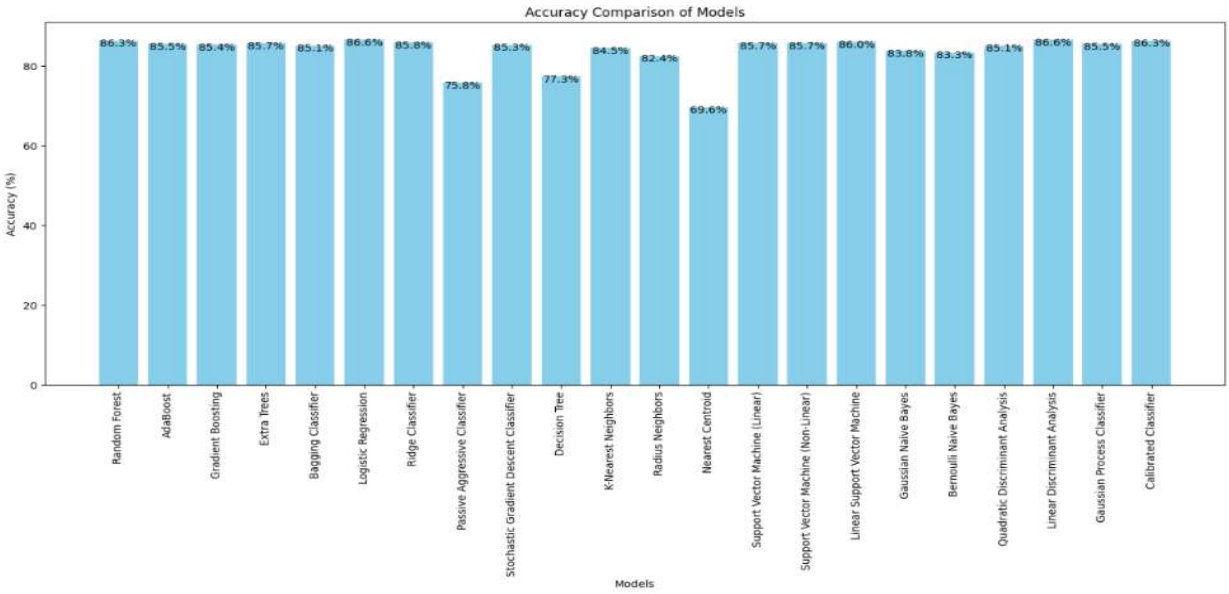

Fig. 8. Bar Graph Showcasing Accuracies of all models used.

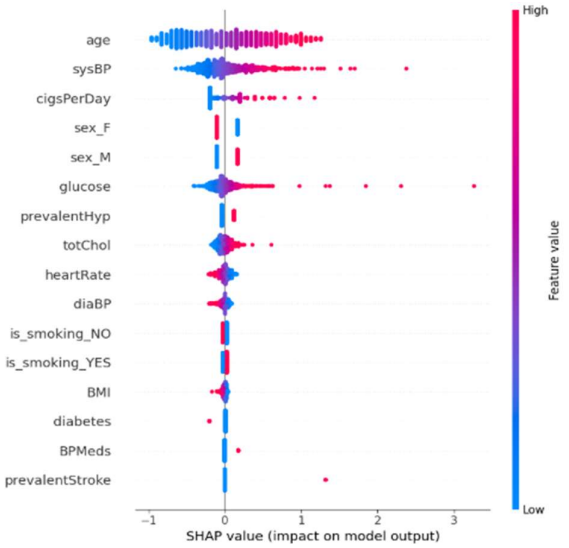

Fig. 6. SHAP Summary Plot for Feature Importance in Logistic Regression

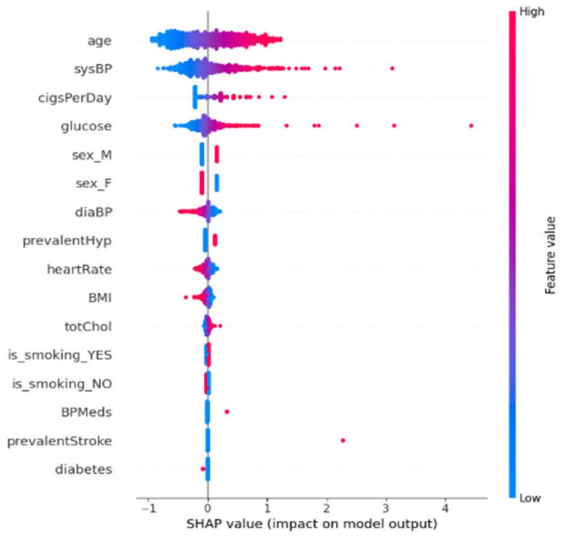

Fig. 7. SHAP Summary Plot for Feature Importance in Linear Discriminant

For model evaluation we have used a 10-fold cross-validation technique. This technique splits dataset into 10 equal parts out of these 10 the model is trained on 9 of these parts and then it is tested on the remaining part. This process is repeated until each of the 10 parts is used for testing to ensure better effectiveness of the model. Using this technique the chance of overfitting is reduced and also It improves model's capability to evaluate new, unseen data.

By employing these techniques and accuracy measures, a thorough comparison of the models is conducted to identify the best-performing algorithm for the given dataset.

V. Results

Among the 22 models evaluated, the Logistic Regression model emerged as the best performer with an accuracy of 86.58%. It was closely followed by Linear Discriminant Analysis, which achieved an accuracy of 86.28%. CatBoost, while also performing well, achieved a slightly lower accuracy of 85.40%. Although CatBoost did not match the top two models in accuracy, it remained a strong contender overall. Random Forest and Calibrated Classifier also demonstrated robust performance, each attaining an accuracy of 86.28%, which is 0.30% lower than Logistic Regression. Some Models like Nearest Centroid and Passive Aggressive Classifier underperformed with accuracies 69.62% and 75.81% as Nearest Centroid assumes that classes possess a single centroid, causing it to be ineffective with complicated decision boundaries and imbalanced data and Aggressive Classifier is high noise sensitive and designed for online learning but perform poor on structured datasets. These results indicate that Logistic Regression and Linear Discriminant Analysis were the most effective in this evaluation, with only minor differences in their performance metrics compared to CatBoost and other strong performers.

TABLE II.

Models	*Precision*	*Recall*	*Accuracy*
Logistic Regression	0.80	0..08	86.58%
Linear Discriminant Analysis	0.67	0.12	86.58%
Calibrated Classifier	0.70	0.07	86.28%
CatBoost	0.46	0.11	85.40%
Random Forest	0.62	0.10	86.28%

VI. Conclusion

The primary focus of our study was to explore diverse machine learning algorithms to predict heart disease. We utilized attributes such as age, sex, smoking habits and other vital health metrics. This constitutes the core of our research. The dataset included 3,389 instances. We divided them into 80% for training, 20% for testing. We assessed the performance of some established machine learning models which included CatBoost Logistic Regression, Random Forest, Linear Discriminant Analysis and Calibrated Classifier. After evaluation we found that Logistic Regression and Linear Discriminant Analysis had the highest accuracy. They performed at 86.58%. The accuracy of Random Forest and Calibrated Classifier was slightly lower at 86.28%. CatBoost achieved an accuracy of 85.40%. This study underscores the effectiveness of Logistic Regression and Linear Discriminant Analysis for a certain dataset. The effectiveness of different algorithms can change depending on varied datasets. Increasing the amount of training data might be beneficial and doing this could improve accuracy yet this process would increase computational demands. Given our current findings we recommend Logistic Regression and Linear Discriminant Analysis. These models are recommended for high accuracy and computational efficiency.

References

[1] K. Pal, S. Panwar and D. Choudhury, "A Pragmatic Approach of Heart and Liver Disease Prediction using Machine Learning Classifiers," 2024 International Conference on Emerging Systems and Intelligent Computing (ESIC), Bhubaneswar, India, 2024, pp. 728-734, doi: 10.1109/ESIC60604.2024.10481536.

[2] K. VARKALA, "Heart Disease Prediction System Using Convolutional Neural Networks," Oct. 2022, doi: https://doi.org/10.21203/rs.3.rs-2009078/v2.

[3] M. E. H. Ong et al., "Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score," Critical Care, vol. 16, no. 3, p. R108, 2012, doi: https://doi.org/10.1186/cc11396.

[4] Pietro Arina et al., "Prediction of Complications and Prognostication in Perioperative Medicine: A Systematic Review and PROBAST Assessment of Machine Learning Tools," Anesthesiology, vol. 140, no. 1,pp.85101,Nov.2023,doi:https://doi.org/10.1097/aln.0000000000004764.

[5] K.-J. Cho et al., "Prospective, multicenter validation of the deep learning-based cardiac arrest risk management system for predicting in-hospital cardiac arrest or unplanned intensive care unit transfer in patients admitted to general wards," Critical care, vol. 27, no. 1, Sep. 2023, doi: https://doi.org/10.1186/s13054-023-04609-0.

[6] M. Chae, H.-W. Gil, N.-J. Cho, and H. Lee, "Machine Learning-Based Cardiac Arrest Prediction for Early Warning System," Mathematics, vol. 10, no. 12, p. 2049, Jun. 2022, doi: https://doi.org/10.3390/math10122049.

[7] J. Kim, M. Chae, H.-J. Chang, Y.-A. Kim, and E. Park, "Predicting Cardiac Arrest and Respiratory Failure Using Feasible Artificial Intelligence with Simple Trajectories of Patient Data," Journal of Clinical Medicine, vol. 8, no. 9, p. 1336, Aug. 2019, doi: https://doi.org/10.3390/jcm8091336.

[8] J. Kwon, Y. Lee, Y. Lee, S. Lee, and J. Park, "An Algorithm Based on Deep Learning for Predicting In-Hospital Cardiac Arrest," Journal of the American Heart Association, vol. 7, no. 13, Jul. 2018, doi: https://doi.org/10.1161/jaha.118.008678.

[9] A. Nayab, "Heart Disease Prediction," Feb. 09, 2021. https://www.researchgate.net/publication/349140147_Heart_Disease_Prediction

[10] 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, p. 012072, Jan. 2021, doi: https://doi.org/10.1088/1757-899x/1022/1/012072.

[11] Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin, "CatBoost: unbiased boosting with categorical features," *arXiv (Cornell University)*, Jun. 2017, doi: https://doi.org/10.48550/arxiv.1706.09516.

[12] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting," *The Journal of Educational Research*, vol. 96, no. 1, pp. 3–14, Sep. 2002, doi: https://doi.org/10.1080/00220670209598786.

[13] S. J. Rigatti, "Random Forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, Jan. 2017, Available: https://meridian.allenpress.com/jim/article/47/1/31/131479/Random-Forest

[14] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI Communications*, vol. 30, no. 2, pp. 169–190, May 2017, doi: https://doi.org/10.3233/aic-170729.

[15] Christofel Ganteng, "Cardiovascular Study Dataset," *Kaggle.com*, 2020. https://www.kaggle.com/datasets/christofel04/cardiovascular-study-dataset-predict-heart-disea?select=train.csv (accessed Sep. 11, 2024).