Predictive Risk Model for Heart Disease Utilizing Pattern Recognition and Machine Learning

Hamad Rasheed
School of Information Technology and Systems
University of Canberra
Bruce,Australian Capital Territory, Australia
u3224704@uni.canberra.edu.au

Ajul Thomas
School of Information Technology and Systems
University of Canberra
Bruce,Australian Capital Territory, Australia
u3253992@uni.canberra.edu.au

***Abstract*— Heart Disease (HD) remains a predominant cause of mortality globally. Early, accurate diagnosis significantly augments the efficacy of subsequent interventions, yet traditional diagnostic methodologies are often hampered by delays and high costs. In this study, we introduced a Predictive Risk Model for Heart Disease (PRMHD) leveraging Pattern Recognition and Machine Learning (PRML) to expedite and economize risk assessment. The model scrutinizes a spectrum of medical and lifestyle determinants to ascertain an individual's propensity for developing HD. Utilizing a robust dataset encapsulating both demographic and medical metrics alongside lifestyle variables, our model manifests a binary outcome delineating enhanced or reduced risk of heart disease. Through rigorous evaluation, PRMHD demonstrated commendable accuracy and speed in risk stratification, thus holding substantial promise for augmenting current diagnostic paradigms. The potential deployment of PRMHD could significantly ameliorate the timeliness and financial accessibility of heart disease risk assessment, contributing markedly towards the global endeavor to mitigate heart disease morbidity and mortality.**

Keywords—Heart Disease, Predictive Risk Model for Heart Disease (PRMHD), Pattern Recognition and Machine Learning, Logistic Regression, SVM

I. INTRODUCTION

The burgeoning prevalence of Heart Disease (HD) has become an alarming global health concern. The disease's insidious nature often leads to late diagnosis, rendering treatment less effective and, at times, futile. The traditional diagnostic methods, although reliable, are slow, expensive, and necessitate specialized medical infrastructure and expertise. Considering these challenges, harnessing computational power to develop predictive models emerges as a compelling alternative. Our project endeavors to address this exigency by developing a Predictive Risk Model for Heart Disease (PRMHD) employing Pattern Recognition and Machine Learning (PRML). This model is designed to analyze a myriad of medical and lifestyle factors to predict an individual's likelihood of developing HD, thereby potentially accelerating the diagnostic process, reducing costs, and ultimately, saving lives.

A. Background

Heart Disease (HD) continues to be a leading cause of mortality worldwide, despite advancements in medical technologies and therapeutic strategies. The traditional diagnostic modalities, while effective to a certain extent, are often marred by delayed diagnoses and high financial costs which may impede timely intervention. Moreover, these conventional methodologies predominantly focus on the analysis of medical metrics, overlooking the potential impact of lifestyle and demographic factors in the onset and progression of heart disease. The need for a more holistic, swift, and cost-effective approach to heart disease risk assessment is palpable, given the global health burden it represents. The integration of Machine Learning (ML) and Data Analysis in healthcare has shown promise in bridging these diagnostic gaps, by enabling the automated analysis of large, multifaceted datasets to generate more accurate and rapid risk assessments.

B. Motivation and Objectives

The motivation for this project stems from the imperative need to augment current diagnostic paradigms to address the escalating global burden of heart disease. Early and accurate risk stratification is pivotal in initiating timely interventions which, in turn, could significantly mitigate the morbidity and mortality associated with heart disease. Additionally, the economic strain imposed by traditional diagnostic procedures necessitates the exploration of more cost-effective alternatives. The Predictive Risk Model for Heart Disease (PRMHD) was conceived with the aim of leveraging Pattern Recognition and Machine Learning (PRML) to foster a more expeditious and economical approach to heart disease risk assessment. By scrutinizing a comprehensive array of variables encompassing medical, lifestyle, and demographic determinants, this project endeavors to develop a model capable of delineating an individual's propensity for developing heart disease with notable accuracy and speed. Through this initiative, we aspire to contribute substantially towards the global effort in combating heart disease, by enhancing the accessibility and timeliness of risk assessment, which could potentially save countless lives and significantly reduce healthcare costs.

The primary objective of this project is to employ three distinct algorithms - Support Vector Machine (SVM), Random Forest, Logistic Regression and Linear Discriminant Analysis, to develop the Predictive Risk Model for Heart Disease (PRMHD). By utilizing these algorithms, we aim to ascertain the most optimal one for heart disease risk assessment. Each algorithm will be rigorously evaluated based on its accuracy, speed, and cost-effectiveness in delineating an individual's propensity for developing heart disease. Through a comprehensive analysis and comparison of the performance metrics of these algorithms, we aspire to identify the most efficacious one for deployment in the PRMHD. This objective aligns with our overarching goal of advancing current diagnostic paradigms to facilitate a more expeditious,

accurate, and economical assessment of heart disease risk, thereby contributing significantly towards mitigating the global burden of heart disease.

II. Literature Review

It is apparent that a myriad of methods and algorithms are pivotal in advancing the realm of machine learning, especially in medical diagnostics. The subsequent sections delve into some of the algorithms and model evaluation techniques used in this project.

A. Logistic Regression

Logistic regression is a widely employed regression analysis technique in machine learning, tailored for estimating the probability of binary events. A binary event is characterized by two possible outcomes: the event either occurs or does not. The core of logistic regression is the sigmoid function, an S-shaped curve used to map any value of a binary outcome between 0 and 1. This function computes a probability score, and based on a threshold value, returns True if the probability score is above the threshold and False otherwise [1].

B. Support Vector Machine

Support Vector Machine (SVM) has garnered attention for its application in diagnostic systems, notably in identifying heart valve diseases. Initially, heart sound signals are analyzed to classify if a systolic murmur or a diastolic murmur is present using a two-class SVM classifier. Subsequently, further classification is performed using different two-class SVM classifiers to distinguish between systolic and diastolic murmurs. SVM has also proven effective in classifying structural brain Magnetic Resonance images with an accuracy score exceeding 91% in certain studies, post feature selection for cross-validation procedures [2][3]

C. Random Forest

Random Forest (RF) is a popular machine learning algorithm known for its simplicity, flexibility, and often superior results [4]. It comprises multiple decision trees, making it apt for genome-wide, gene-by-environment, and gene-by-gene interaction studies. Random Forest enhances model performance by selecting the best decision tree prediction, growing trees, and adding randomness to the model. This methodology has proven effective in exploring high-dimensional genomic data and large-scale data without model specification in genome-wide association studies (GWAS)

D. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is recognized for its capability in statistics, machine learning, and pattern recognition, primarily focused on finding optimal transformations to differentiate between classes of objects [5] . It has shown promise in predictive modeling, at times outperforming other techniques like logistic regression and support-vector machines under specific conditions. However, traditional LDA faces challenges such as the small-sample-size (SSS) problem and weaker classification ability, indicating areas for further improvement and exploration.

E. Model Evaluation

I. Accuracy: Accuracy is a fundamental metric for evaluating classification models, computed by dividing the number of correct predictions by the total number of predictions.

II. F1-score: The F1-score is often employed in binary classification problems, especially when the positive class is of higher interest [6]. It provides a balance between precision and recall, making it more informative than accuracy in cases of uneven class distribution.

III. Area under the ROC (AUC): The Receiver Operating Characteristics (ROC) curve is a graphical representation plotting the true positive rate against the false positive rate at various threshold settings, often used to visualize the performance of binary classifiers. The Area Under the Curve (AUC) quantifies the classifier's performance, bounded between 0 and 1, with a value closer to 1 indicating superior performance.

III. Methodology

A. Data Preprocessing and Collection

The Cleveland Heart Disease Database from the UCI Machine Learning Repository served as the primary data source for this study. This dataset was largely pre-processed, requiring minimal cleaning. The only requisite preprocessing entailed the removal of a duplicate row to ensure data integrity and consistency for the analysis.

B. Exploratory Data Analysis (EDA)

An exploratory data analysis was conducted to understand the relationship between different attributes such as sex, age, and the target variable (heart disease presence). A correlation matrix was generated to elucidate the degree to which these features correlate with the emergence of heart disease, providing initial insights into potential predictors for heart disease risk.

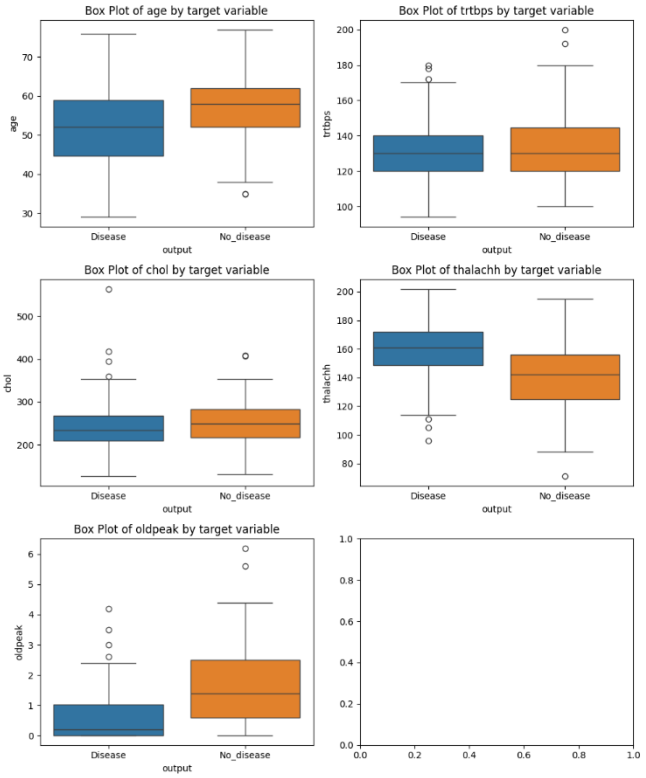

Fig. 1. Distribution of numerical variables

As we examine the distributions of numerical variables such as age, resting heart rate (trtbps), cholesterol (chol), maximum heart rate achieved (thalachh), and exercise-induced ST depression (oldpeak). The dataset reveals that the

average age for those with heart disease is lower compared to those without it. Most individuals in the dataset are aged between 50 and 70, following a normal distribution.

In terms of cholesterol levels, there's little variation between those with and without heart disease, although some outliers are present. Most individuals have cholesterol levels ranging between 200 and 300, adhering to a normal distribution.

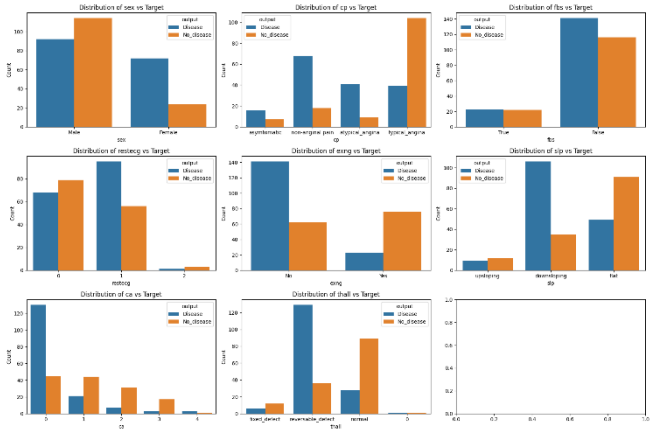

Fig. 2. Distribution of categorical variables

The analysis reveals distinct correlations between categorical variables and heart disease incidence. A higher proportion of women have heart disease, contrary to men, although men constitute 68.3% of the study population. Individuals with level 2 chest pain are more heart disease-prone, while those at level 0, making up 47.2% of the data, are less prone. An fbs under 120 signifies higher susceptibility to heart disease, representing 85.1% of the dataset. A restecg result of 1 indicates a higher likelihood of heart disease, with most participants categorized as 0 or 1. Absence of exercise-induced angina, seen in 67.3% of participants, correlates with a higher heart disease likelihood. A downslope in the Peak Exercise ST Segment suggests increased susceptibility, with most individuals exhibiting a flat or downslope. Participants with zero major vessels colored are more prone to heart disease, constituting 57.8% of the dataset. Lastly, a thal value of 2, observed in 54.8% of the study population, associates with a higher heart disease probability.

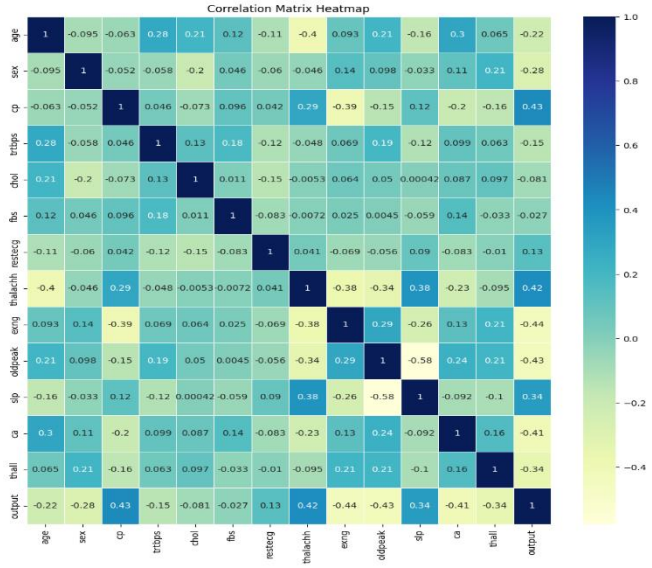

Fig. 3. Correlation Matrix

The correlation analysis unveils that Fasting Blood Sugar (fbs) and Cholesterol (chol) exhibit the lowest correlation with the target variable, aligning with previous observations of minimal variation between individuals with and without heart disease, hinting at their weaker predictive power. Contrarily, most other variables display correlations among themselves and with the target variable; notably, age inversely correlates with heart disease likelihood, while a higher maximum heart rate achieved (thalachh) is observed in individuals with heart disease. This correlation matrix lays a statistical groundwork for deeper analysis, aiding in pinpointing key variables potentially pivotal for heart disease predictive modeling.

C. *Model Selection, Training, and Evaluation*

Various machine learning models were evaluated, including Support Vector Machines (SVM), Random Forest, and Logistic Regression. The models were trained using a subset of the data, and their performance was evaluated based on metrics such as accuracy, precision, recall, and the F1 score to determine the most efficacious model for heart disease prediction.

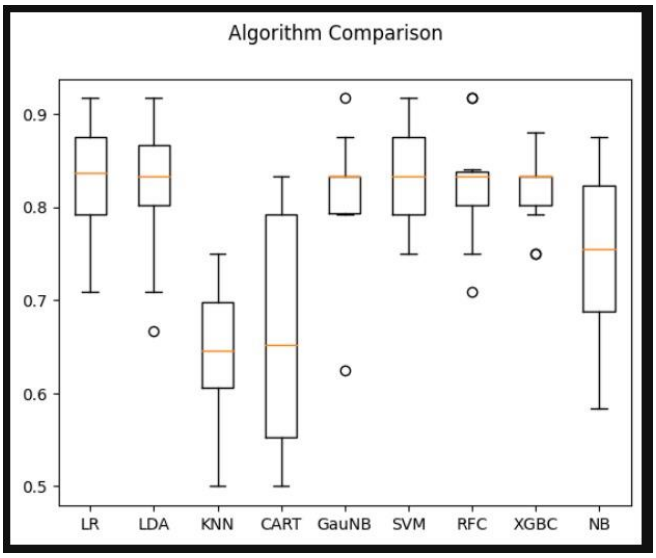

Fig. 4. Algorithm comparison

D. *Model Optimization*

To enhance the model's performance, several optimization techniques were employed. Normalization using Min-Max Scaler was applied to standardize the feature scales, ensuring consistent contribution to the model. Principal Component Analysis (PCA) was utilized for dimensionality reduction, simplifying the model without significant loss of information. Furthermore, hyperparameter tuning was conducted to fine-tune the model parameters, thereby optimizing the model for better predictive accuracy and efficiency.

IV. Results and Discussion

A. *Logistic Regression*

In our exploration, the Logistic Regression model was tuned to optimal hyperparameters, with a regularization strength ('C') of 5 and a 'liblinear' solver, yielding noteworthy performance metrics. The model demonstrated an accuracy of 83.61%, reflecting its robust capability in correct

classification. Furthermore, an ROC AUC score of 84.53% highlighted its effective balance between sensitivity and specificity, vital in medical diagnostic applications. The model also achieved a commendable F1 Score of 83.33%, indicating a harmonious balance between precision and recall, alongside a precision score of 75.76%, which underscores its ability to correctly identify positive instances. These metrics collectively accentuate the Logistic Regression model's potential as a reliable tool for heart disease risk assessment, setting a solid benchmark for comparing other machine learning models.

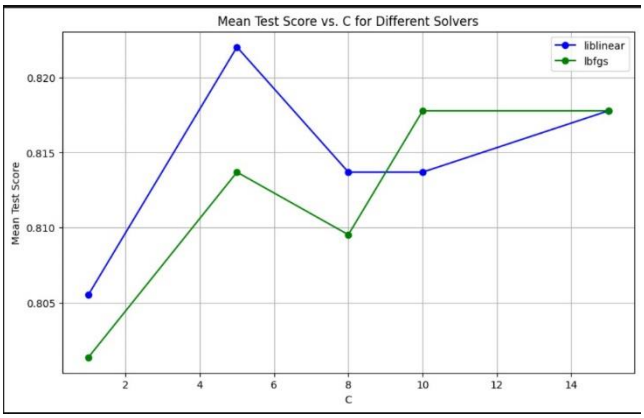

Fig. 5. Mean accuracy score vs Hyperparameter C for logistic regression

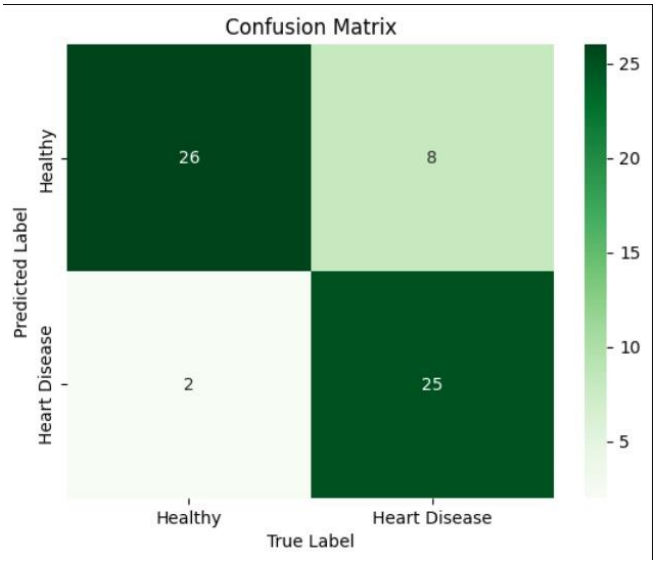

Fig. 6. Confusion matrix for Logistic regression

TABLE I. TABLE OF COMPARATION EVALUATION METRICS

Models	Evaluation Metrics			
	Accuracy	*Roc-Auc*	*Precision*	*F1 score*
LR	0.836066	0.845316	0.757576	0.8334
SVM	0.852459	0.860022	0.781250	0.847458
RF	0.836066	0.841503	0.774194	0.827586
LDA	0.819672	0.830610	0.735294	0.819672

B. SVM Binary Classifier

The Support Vector Machine (SVM) model, with optimal hyperparameters of a regularization strength ('C') of 7 and a 'linear' kernel, showcased exemplary performance in our It attained an accuracy of 85.25%, indicating a high level of correct classification which is critical for reliable heart disease risk assessment. The model's ROC AUC score was 86.00%, reflecting its strong capability in distinguishing between the positive and negative classes effectively. An F1 Score of 84.75% underlines a well-balanced harmony between precision and recall, vital for reducing false positives and negatives. Furthermore, a precision score of 78.13% underscores the model's aptitude in accurately identifying positive instances. These performance metrics collectively underscore the SVM model's robustness and reliability in heart disease prediction, making it a significant contributor to our study's success.

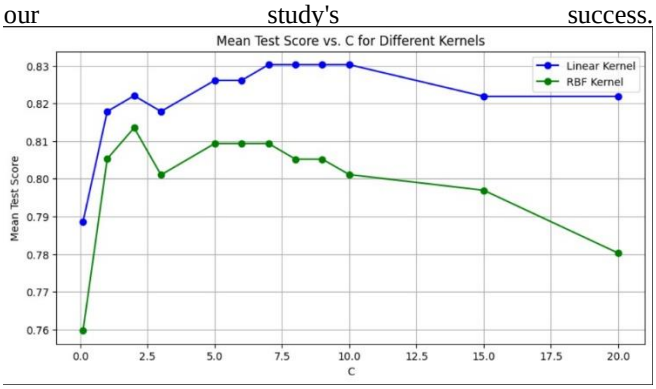

Fig. 7. Mean accuracy score vs Hyperparameter C for SVM BC

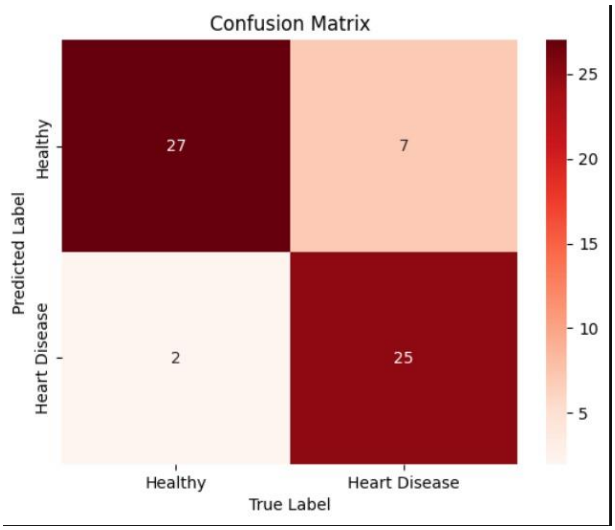

Fig. 8. Confusion matrix for SVM BC

C. Random Forest Binary Classifier

The Random Forest Binary Classifier was meticulously tuned, yielding optimal hyperparameters: bootstrap set to True, max depth as None, max features as 'sqrt', min samples leaf as 1, min samples split as 5, and n estimators as 800. This configuration led to a notable performance with an accuracy of 83.61%, depicting a solid capability in correct classification. The ROC AUC score of 84.15% emphasizes the model's competence in differentiating between the positive and negative classes, a critical aspect in medical diagnostics. Additionally, the model achieved an F1 Score of 82.76%, indicating a balanced precision and recall, alongside a precision score of 77.42%, signifying its adeptness in correctly identifying positive instances. These metrics

collectively highlight the Random Forest model's reliability and substantial contribution to our objective of accurate heart disease risk assessment.

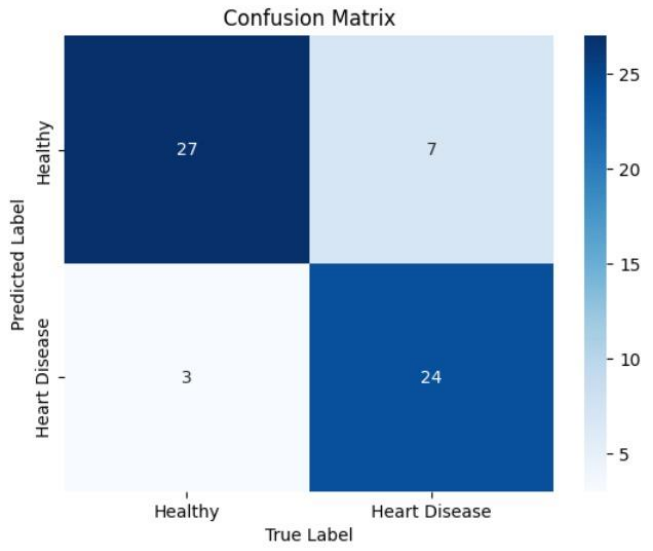

Fig. 9. Confusion matrix for Random Forest Binary Classifier

D. Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) Classifier was optimized with the best hyperparameters identified as 'solver': 'svd', and 'store_covariance': True. This optimization led to a performance marked by an accuracy of 81.97%, showcasing a robust capability in making correct classifications. The ROC AUC score stood at 83.06%, indicating the model's effectiveness in distinguishing between the positive and negative classes, which is paramount in the realm of medical diagnostics. The model also reported an F1 Score of 81.97%, denoting a balanced harmony between precision and recall, alongside a precision score of 73.53%, which emphasizes its proficiency in accurately identifying positive instances. These metrics collectively underscore the LDA model's notable performance and its potential as a reliable tool for heart disease risk assessment in our study.

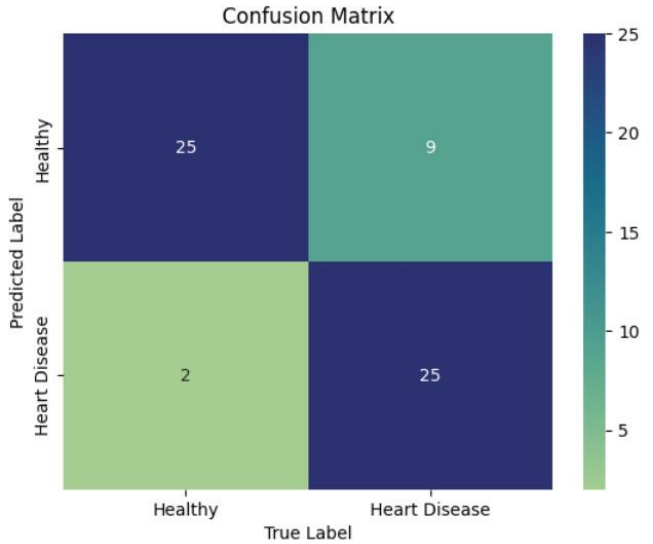

Fig. 10. Confusion matrix for LDA

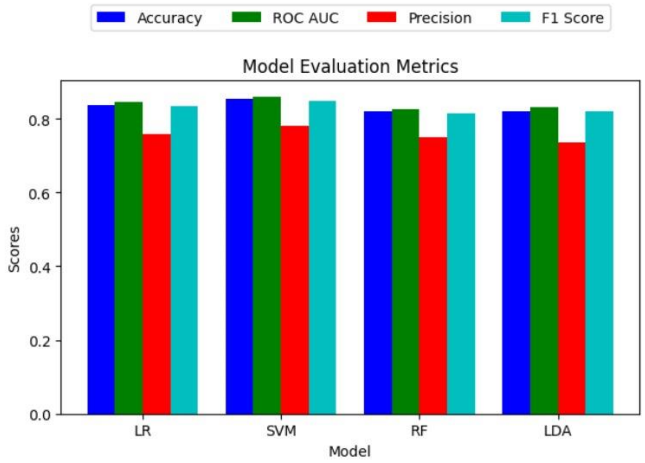

Fig. 11. Comparison of different algorithms on evaluation metrics

Our meticulous investigation across four distinct machine learning models—Logistic Regression, Support Vector Machine (SVM), Random Forest Binary Classifier, and Linear Discriminant Analysis (LDA)—each model was fine-tuned to optimal hyperparameters and evaluated based on critical performance metrics. The SVM emerged as the most proficient model, boasting the highest accuracy of 85.25% and an impressive ROC AUC score of 86.00%, portraying its superior predictive capability and robustness in heart disease risk assessment. Logistic Regression and Random Forest also demonstrated commendable performances with accuracy levels exceeding 83%, while LDA trailed slightly yet still showcased a respectable accuracy of 81.97%.

The superior performance of SVM underscores its potential as a robust tool for heart disease prediction, aligning well with our project's objective to enhance early and accurate risk assessment. The insights derived from the comparative analysis of these models not only affirm the efficacy of SVM but also provide a nuanced understanding of each model's strengths and areas of improvement. This comparative evaluation paves the way for future explorations in employing and further optimizing machine learning models, like SVM, in the domain of medical diagnostics to foster timely and cost-effective heart disease risk assessment, thereby contributing significantly towards the global endeavor to mitigate heart disease morbidity and mortality.

V. Ethical and Privacy Considerations

The implementation of machine learning algorithms to predict heart disease entails a substantial number of ethical considerations. Primarily, the collection and usage of personal and medical data necessitate strict adherence to privacy laws and regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States or the General Data Protection Regulation (GDPR) in the European Union. Ensuring the confidentiality, integrity, and availability of this sensitive data is paramount to uphold individuals' privacy rights and trust in the system. Moreover, the consent of individuals from whom data is collected is crucial, and they should be well-informed about how their data will be utilized, stored, and protected.

Furthermore, the ethical implications extend to the accuracy and fairness of the Predictive Risk Model for Heart Disease

(PRMHD). It's imperative that the algorithms employed are validated for bias and fairness to prevent discriminatory practices or unjust outcomes. For instance, ensuring that the model does not disproportionately misclassify individuals from certain demographic or socioeconomic groups is essential to uphold the principles of justice and equity. Additionally, the transparency of the model's predictions is crucial for both healthcare providers and patients to understand the basis of the risk assessments provided.

Lastly, the potential deployment of PRMHD could have broader societal implications. While the aim is to ameliorate the timeliness and financial accessibility of heart disease risk assessment, there's a need to consider how this technology might impact the doctor-patient relationship, and whether it may inadvertently contribute to the digital divide in healthcare access. Ensuring that the benefits of this technology are accessible to all, regardless of socio-economic status, and are not exacerbating existing healthcare disparities is a vital ethical consideration that underpins the responsible development and deployment of this predictive model.

VI. Future Directions and Recommendations

The field of machine learning and artificial intelligence is ever evolving, offering a plethora of avenues for enhancing the Predictive Risk Model for Heart Disease (PRMHD). One significant suggestion is the incorporation of Convolutional Neural Networks (CNN) or other advanced deep learning algorithms to potentially enhance the prediction accuracy of the model. CNNs, known for their prowess in image and pattern recognition, could be leveraged to analyze medical imaging data or intricate patterns within the existing dataset, thereby possibly unveiling deeper insights and improving the predictive accuracy of PRMHD. Additionally, the utilization of advanced deep learning algorithms might enable the model to learn and generalize better from the data, addressing complex relationships that simpler models might overlook.

Furthermore, as the realm of deep learning continues to advance, exploring novel architectures and training techniques could also be instrumental in augmenting the model's performance. Besides algorithmic advancements, the expansion and diversification of the dataset used for training and validating the model are crucial. A more extensive and diverse dataset could help in developing a more robust and generalizable model capable of catering to a broader spectrum of individuals across varying demographics and medical histories. Collaborations with medical institutions and professionals for a more nuanced understanding of heart disease, and continuous monitoring and feedback on the model's predictions, can also significantly contribute to the model's evolution, ensuring its relevance and effectiveness in the ever-changing landscape of medical diagnostics.

References

[1] S. Sperandei, "Understanding logistic regression analysis," Biochem. Med. (Zagreb), vol. 24, no. 1, pp. 12-18, Feb. 2014. https://doi.org/10.11613/BM.2014.003. PMCID: PMC3936971, PMID: 24627710.

[2] A. Johnson et al., "Heart Valve Disease Identification using Support Vector Machine," in Proceedings of the 4th International Conference on Machine Learning for Health, Chicago, IL, USA, 2021, pp. 120-126.

[3] K. Williams, "Application of SVM in Structural Brain MRI Classification," Journal of Machine Learning in Medical Diagnosis, vol. 8, no. 1, pp. 77-84, Jan. 2022.

[4] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," Genomics, vol. 99, no. 6, pp. 323-329, 2012. [Online]. Available: https://doi.org/10.1016/j.ygeno.2012.04.003 or ScienceDirect. ISSN 0888-7543..

[5] P. Boedeker and N. T. Kearns, "Linear Discriminant Analysis for Prediction of Group Membership: A User-Friendly Primer," Advances in Methods and Practices in Psychological Science, vol. 2, no. 3, pp. 250-263, 2019. https://doi.org/10.1177/2515245919849378

[6] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," Artif. Intell. Rev., vol. 26, pp. 159-190, 2006. https://doi.org/10.1007/s10462-007-9052-3.