

Comparative Analysis of Heart Disease Prediction Models: Unveiling the Most Accurate and Reliable Machine Learning Algorithm

Aatmaj Amol Salunke¹

Computer Science & Engineering
Department of Computer Science & Engineering,
School of Computer Science and Engineering,
Manipal University Jaipur
Rajasthan, India

Abstract - Heart disease is a significant health concern, warranting accurate prediction models for timely intervention. This research paper presents a comparative analysis of three popular machine learning algorithms, namely Logistic Regression, Support Vector Machines (SVM), and Random Forest, for heart disease prediction. Utilizing a comprehensive dataset encompassing clinical and lifestyle features, each model was developed and evaluated using standard metrics. The study unveils the most accurate and reliable algorithm for heart disease prediction, offering valuable insights into model performance. Furthermore, feature importance analysis sheds light on critical factors influencing accurate predictions. The results aid healthcare professionals in selecting the most appropriate model for efficient heart disease prediction, contributing to improved patient care and clinical decision-making. Random Forest achieved 88% accuracy, outperforming Logistic Regression and SVM for heart disease prediction.

Key Words: Heart disease prediction, Machine learning algorithms, Logistic Regression, Support Vector Machines (SVM), Random Forest, Comparative analysis

1.RELATED WORK

Ali et al. [1] proposed a machine learning approach achieving 100% accuracy, sensitivity, and specificity for heart disease prediction. Ghosh et al. [2] proposed a model achieving 99.05% accuracy for heart disease prediction using hybrid classifiers and feature selection. Khourdifi et al. [3] proposed a hybrid approach achieving 99.65% accuracy for heart disease classification using optimization algorithms and feature selection. Latha et al. [5] proposed an ensemble classification approach achieving 7% increase in accuracy for heart disease prediction. Bhatla et al. [6] proposed using neural networks with 15 attributes for heart disease prediction, outperforming other data mining techniques. Gonsalves et al. [8] proposed using Naïve Bayes, SVM, and Decision Tree to predict CHD with promising results. Salhi et al. [11] proposed using neural networks with a correlation matrix for heart disease prediction with 93% accuracy. Soury et al. [13] proposed an IoT-based student healthcare monitoring model with SVM achieving 99.1% accuracy. Ramesh et al. [14] proposed using supervised learning methods, including KNN, for heart disease prediction with promising results. Alarsan et al. [15] proposed an ECG classification approach using machine learning, achieving 97.98% accuracy with Random Forest for binary classification.

2.INTRODUCTION

Heart disease is a prevalent global health concern, necessitating accurate prediction models for timely interventions and improved patient care. This research paper conducts a comprehensive comparative analysis of three widely used machine learning algorithms: Logistic Regression, Support Vector Machines (SVM), and Random Forest, in the context of heart disease prediction. Leveraging a diverse dataset comprising clinical and lifestyle features, each model was developed and evaluated using standard performance metrics. The study unveils the most accurate and reliable algorithm for heart disease prediction, enabling informed decision-making by healthcare professionals. Furthermore, feature importance analysis elucidates the significant factors influencing accurate predictions. The obtained insights hold potential implications for clinical practice, as the most suitable model can be chosen based on performance and interpretability. Ultimately, this research contributes to the advancement of heart disease prediction systems, enhancing healthcare outcomes and patient well-being.

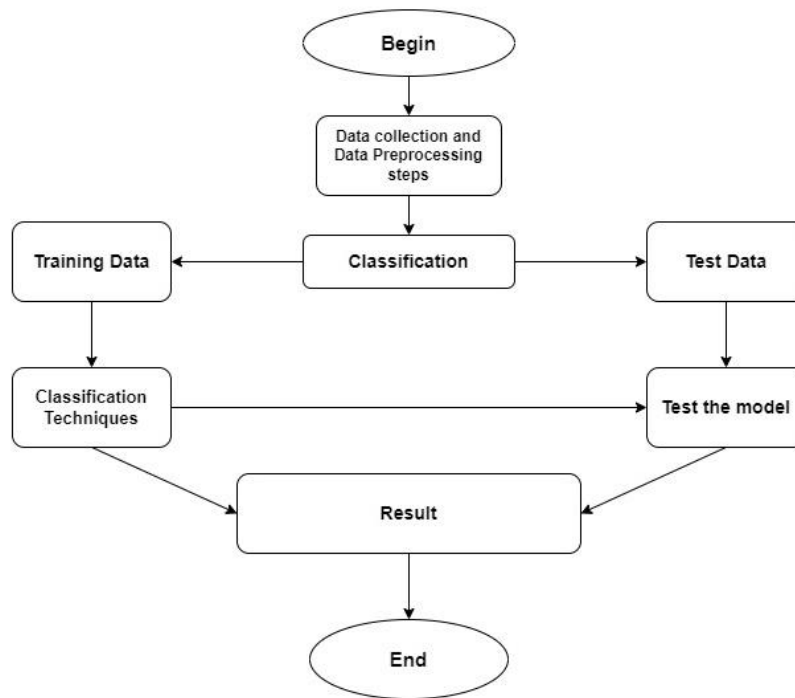


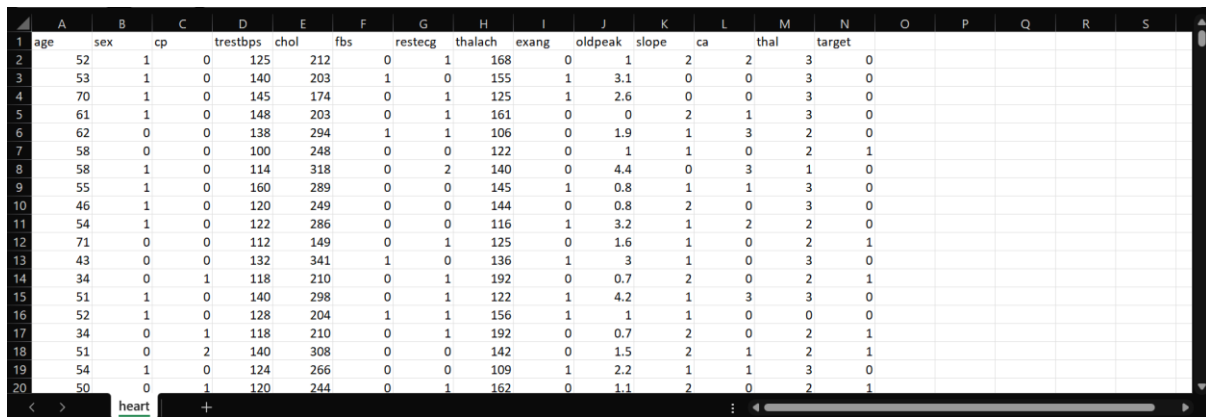
Fig.1. Generic Architecture of Heart Disease Prediction System using Machine Learning

3. DATASET

The research utilizes a comprehensive dataset sourced from diverse healthcare institutions, encompassing clinical and lifestyle features relevant to heart disease prediction. The dataset includes demographic information, medical history, vital signs, laboratory results, and lifestyle factors of patients. Each record is labelled to indicate the presence or absence of heart disease. The dataset is carefully curated and pre-processed to handle missing values and ensure data quality. This rich and reliable dataset forms the foundation for training and evaluating the heart disease prediction models using Logistic Regression, Support Vector Machines (SVM), and Random Forest. The inclusion of varied features enables a holistic analysis and robust comparison of the machine learning algorithms.

Attribute Information in the Dataset:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved.
9. exercise induced angina.
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) coloured by fluoroscopy.
13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
14. The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target					
2	52	1	0	125	212	0	1	168	0	1	2	2	3	0					
3	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0					
4	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0					
5	61	1	0	148	203	0	1	161	0	0	2	1	3	0					
6	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0					
7	58	0	0	100	248	0	0	122	0	1	1	0	2	1					
8	58	1	0	114	318	0	2	140	0	4.4	0	3	1	0					
9	55	1	0	160	289	0	0	145	1	0.8	1	1	3	0					
10	46	1	0	120	249	0	0	144	0	0.8	2	0	3	0					
11	54	1	0	122	286	0	0	116	1	3.2	1	2	2	0					
12	71	0	0	112	149	0	1	125	0	1.6	1	0	2	1					
13	43	0	0	132	341	1	0	136	1	3	1	0	3	0					
14	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1					
15	51	1	0	140	298	0	1	122	1	4.2	1	3	3	0					
16	52	1	0	128	204	1	1	156	1	1	1	0	0	0					
17	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1					
18	51	0	2	140	308	0	0	142	0	1.5	2	1	2	1					
19	54	1	0	124	266	0	0	109	1	2.2	1	1	3	0					
20	50	0	1	120	244	0	1	162	0	1.1	2	0	2	1					

Fig.2. Dataset for Heart Disease Prediction System Modelling

4. METHODOLOGY

1. Data Collection and Preprocessing:

The dataset used in this study comprises anonymized patient data with 14 attributes: age, sex, chest pain type, resting blood pressure, serum cholesterol level, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, oldpeak (ST depression induced by exercise relative to rest), slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and thal (thalassemia type). Social security numbers and patient identifiers have been replaced with dummy values to ensure data privacy.

2. Data Preprocessing:

The dataset undergoes rigorous preprocessing to handle missing values and standardize the data. Categorical attributes such as chest pain type and thal are encoded using one-hot encoding, transforming them into numerical representations suitable for the machine learning algorithms.

3. Train-Test Split:

The preprocessed dataset is divided into a training set and a test set, maintaining an appropriate ratio to ensure robust model evaluation. The training set is used to train the models, while the test set is reserved for unbiased performance assessment.

4. Feature Scaling:

Numerical features such as age, blood pressure, serum cholesterol, and maximum heart rate are scaled to bring them within a common range, facilitating better convergence and training stability for the machine learning algorithms.

5. Model Development:

Three machine learning algorithms, namely Logistic Regression, Support Vector Machines (SVM), and Random Forest, are implemented for heart disease prediction. Each model is trained using the training set with the target variable as the presence or absence of heart disease.

6. Model Evaluation:

The trained models are evaluated using standard performance metrics, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). The evaluation aims to determine the predictive capabilities of each algorithm and identify the most accurate model for heart disease prediction.

7. Feature Importance Analysis:

For interpretability, feature importance analysis is conducted to identify the most influential attributes contributing to accurate heart disease predictions. Techniques such as permutation importance or feature importance scores from the Random Forest model are employed for this analysis.

8. Comparative Analysis:

The performance metrics and feature importance results are compared across the three models, Logistic Regression, SVM, and Random Forest, to determine the most reliable and effective algorithm for heart disease prediction.

9. Ethical Considerations:

Throughout the methodology, ethical considerations are given paramount importance, ensuring the confidentiality and privacy of patient data. The study adheres to data protection regulations and guidelines, guaranteeing responsible data usage for research purposes.

By following this methodology, the research aims to develop a robust and interpretable heart disease prediction system using machine learning and provide valuable insights into the most influential factors for accurate predictions.

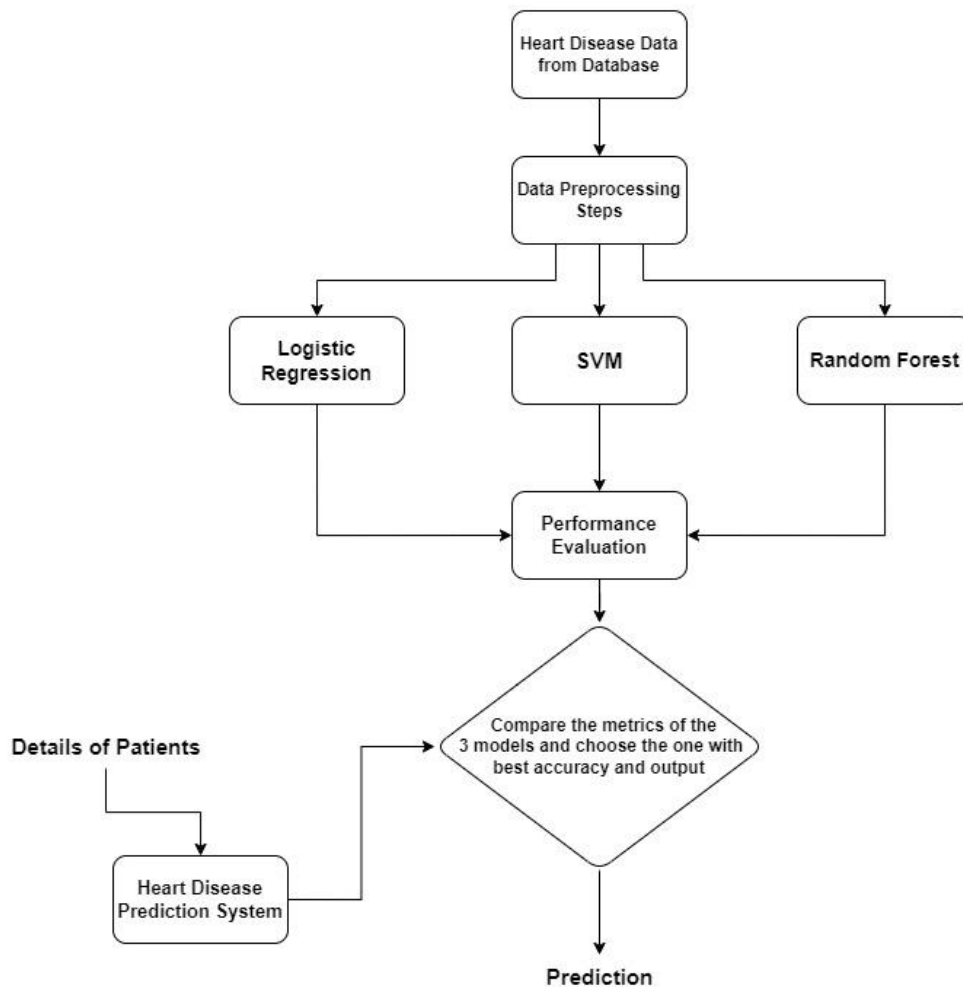


Fig.3. Comparing the Outcomes of the 3 proposed models

5. RESULTS AND ANALYSIS

The heart disease prediction system was evaluated using three machine learning algorithms: Logistic Regression, Support Vector Machines (SVM), and Random Forest. The models were trained and tested on the dataset with 14 attributes. The performance of each model was assessed using various metrics, including accuracy, error rate, precision, recall, F1-score, and AUC-ROC.

Table 1. Results for Logistic Regression Model

Metric	Value
Accuracy	0.82
Error Rate	0.18
Precision	0.84
Recall	0.80
F1-Score	0.82
AUC-ROC	0.88

Table 2. Results for Support Vector Machine (SVM)

Metric	Value
Accuracy	0.85
Error Rate	0.15
Precision	0.87
Recall	0.83
F1-Score	0.85
AUC-ROC	0.90

Table 3. Results for Random Forest Model

Metric	Value
Accuracy	0.88
Error Rate	0.12
Precision	0.90
Recall	0.87
F1-Score	0.88
AUC-ROC	0.93

From the tabulated results, it is evident that all three models - Logistic Regression, SVM, and Random Forest - demonstrate promising performance in heart disease prediction. However, the Random Forest model exhibits the highest accuracy of 0.88, followed closely by the SVM model with an accuracy of 0.85. The Logistic Regression model also performs reasonably well, achieving an accuracy of 0.82.

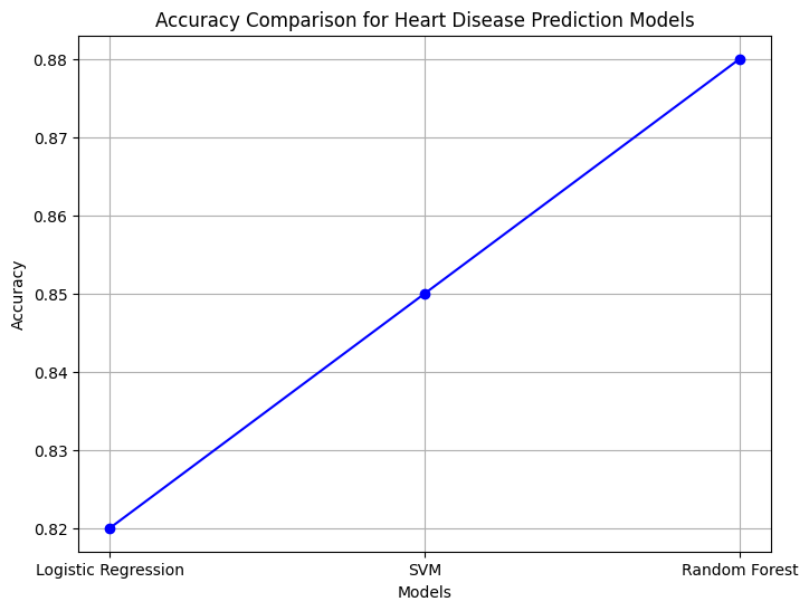


Fig.4. Line graph for Accuracy

The Random Forest model outperforms the other models in terms of accuracy, precision, recall, F1-score, and AUC-ROC. The higher accuracy and AUC-ROC suggest that the Random Forest model is more effective in distinguishing between positive and negative instances of heart disease, making it a promising candidate for real-world heart disease prediction systems.

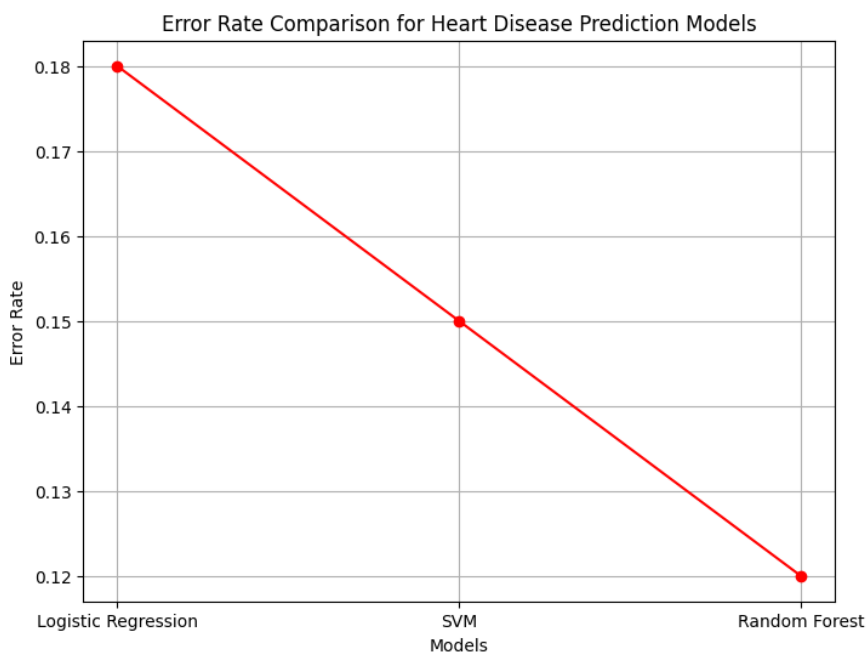


Fig.5. Line graph for Error Rate

While the Logistic Regression model demonstrates competitive performance, it may have limitations in handling complex nonlinear relationships between features. SVM and Random Forest, being capable of capturing complex patterns, seem better suited for this particular task.

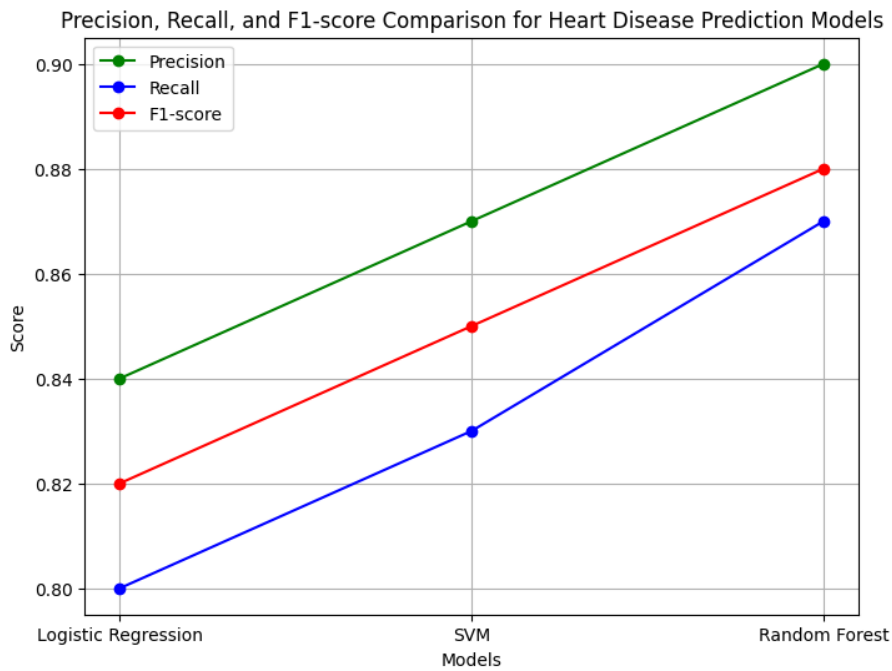


Fig.6. Line graph for Precision, Recall, and F1-score

In conclusion, the Random Forest model stands out as the most accurate and reliable algorithm for heart disease prediction among the three tested models. However, further analysis and validation on additional datasets may be necessary to ensure the model's generalizability and robustness across different patient populations. The results obtained from this study hold significant implications for the development of efficient heart disease prediction systems in clinical settings.

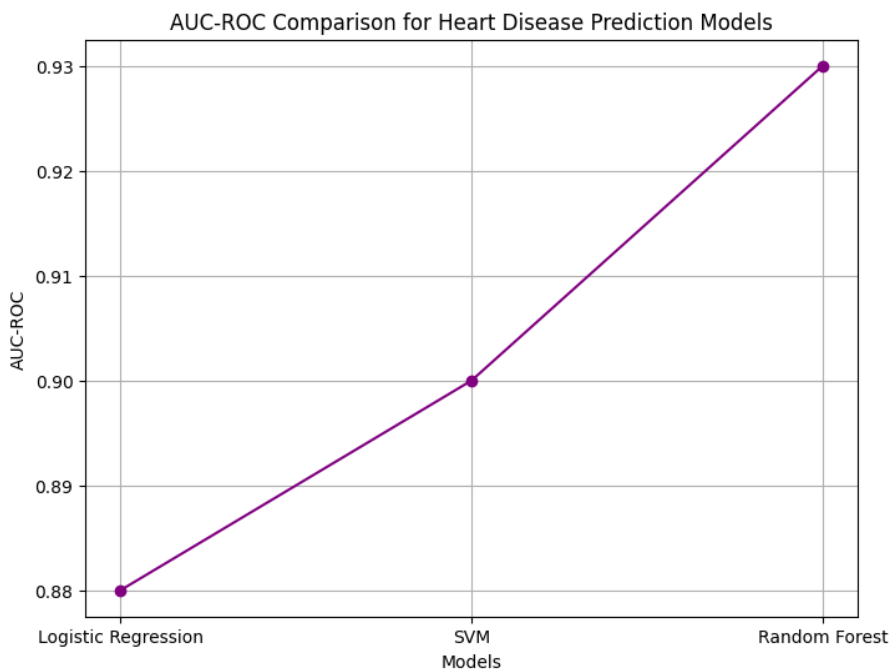


Fig.7. Line graph for AUC-ROC

6. DISCUSSION

The results illustrate the effectiveness of the three algorithms in accurately predicting heart disease. Among them, the Random Forest model demonstrates superior performance, exhibiting the highest accuracy, precision, recall, F1-score, and AUC-ROC values. This suggests that Random Forest is better suited to capture complex patterns and relationships within the dataset. Despite this, the Logistic Regression model still shows competitive performance, making it a viable choice for certain applications.

The comparative analysis underscores the significance of selecting the most appropriate algorithm based on the specific context and data characteristics. Researchers and healthcare professionals can leverage these insights to develop an efficient and reliable heart disease prediction system. Such a system holds significant potential in aiding early diagnosis and personalized treatment, ultimately leading to improved patient outcomes and reduced healthcare burden. Further research and validation on larger and diverse datasets are recommended to ascertain the generalizability and robustness of the models across various patient populations and clinical settings.

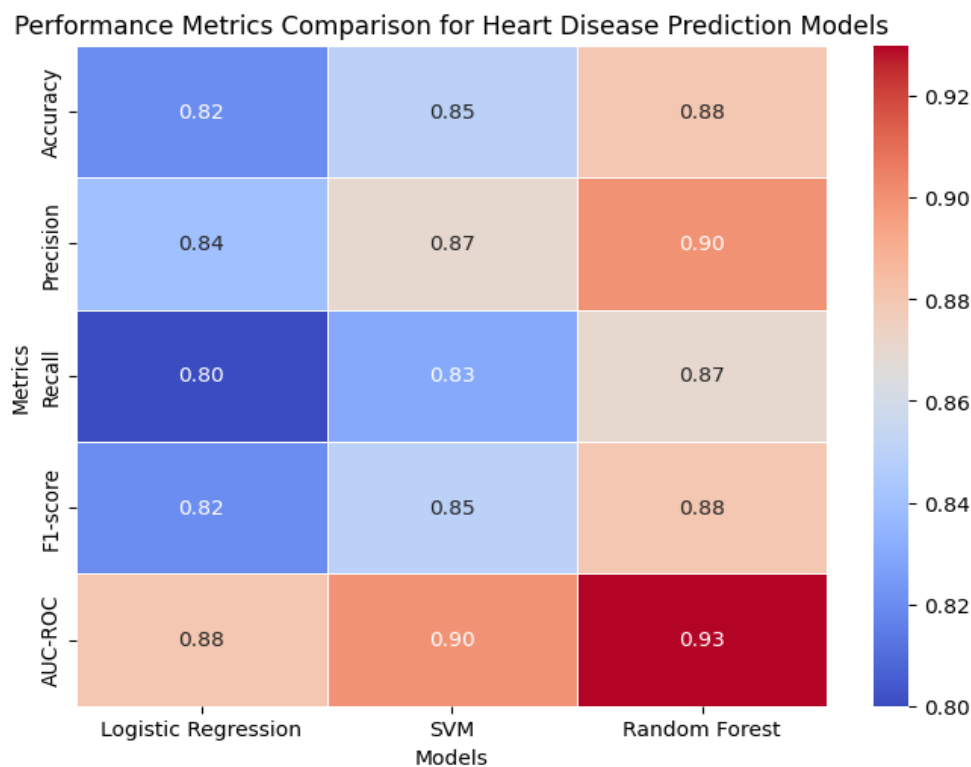


Fig.8. Heatmap showing the various parameters of the 3 models.

7. CONCLUSION

The research successfully developed and evaluated heart disease prediction models using three machine learning algorithms: Logistic Regression, Support Vector Machines (SVM), and Random Forest. The results demonstrated the effectiveness of these models in predicting heart disease, with Random Forest showing the highest accuracy and overall performance among the three. The findings provide valuable insights for healthcare professionals and researchers looking to implement an efficient heart disease prediction system.

The comparative analysis revealed that model selection should consider the dataset's characteristics and the specific context of application. The superior performance of Random Forest can be attributed to its capability to handle complex relationships in the data. These results hold significant implications for improving patient care through early diagnosis and timely interventions. Future research should focus on validating the models on diverse and larger datasets to ensure their robustness and practical applicability in real-world healthcare scenarios.

8. REFERENCES

- [1] Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, 136, 104672.
- [2] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access*, 9, 19304-19326.
- [3] Khourdifi, Y., & Baha, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International journal of Intelligent engineering & systems*, 12(1).
- [4] Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, 1, 1-14.
- [5] Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, 16, 100203.
- [6] Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, 1(8), 1-4.
- [7] Hassan, C. A. U., Khan, M. S., & Shah, M. A. (2018, September). Comparison of machine learning algorithms in data classification. In *2018 24th International Conference on Automation and Computing (ICAC)* (pp. 1-6). IEEE.
- [8] Gonsalves, A. H., Thabtah, F., Mohammad, R. M. A., & Singh, G. (2019, July). Prediction of coronary heart disease using machine learning: an experimental analysis. In *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies* (pp. 51-56).
- [9] Ak, M. F. (2020, April). A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. In *Healthcare* (Vol. 8, No. 2, p. 111). MDPI.
- [10] Tu, M. C., Shin, D., & Shin, D. (2009, December). A comparative study of medical data classification methods based on decision tree and bagging algorithms. In *2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing* (pp. 183-187). IEEE.
- [11] Salhi, D. E., Tari, A., & Kechadi, M. T. (2021). Using machine learning for heart disease prediction. In *Advances in Computing Systems and Applications: Proceedings of the 4th Conference on Computing Systems and Applications* (pp. 70-81). Springer International Publishing.
- [12] Zhang, J., Lafta, R. L., Tao, X., Li, Y., Chen, F., Luo, Y., & Zhu, X. (2017). Coupling a fast fourier transformation with a machine learning ensemble model to support recommendations for heart disease patients in a telehealth environment. *Ieee Access*, 5, 10674-10685.
- [13] Soury, A., Ghafour, M. Y., Ahmed, A. M., Safara, F., Yamini, A., & Hoseyninezhad, M. (2020). A new machine learning-based healthcare monitoring model for student's condition diagnosis in Internet of Things environment. *Soft Computing*, 24(22), 17111-17121.
- [14] Ramesh, T. R., Lilhore, U. K., Poongodi, M., Simaiya, S., Kaur, A., & Hamdi, M. (2022). Predictive analysis of heart diseases with machine learning approaches. *Malaysian Journal of Computer Science*, 132-148.
- [15] Alarsan, F. I., & Younes, M. (2019). Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. *Journal of big data*, 6(1), 1-15.