A Comparative Evaluation of Machine Learning Techniques for Data-Driven Heart Disease Prediction

Dhruv Trivedi[1], Jeel Talaviya[2], Ronish Ramani[3], and Anjali Diwan[4]
Department of CE-AI, Marwadi University, Rajkot, India
dhruvtrivedi285@gmail.com, anjali.diwan@marwadieducation.edu.in

***Abstract*— In recent years, intelligent technologies have significantly contributed to enhancing patient care, reducing healthcare costs, and alleviating workload, particularly in telehealth settings. This research paper introduces a data-driven heart disease recommendation system aimed at evaluating the efficacy and accuracy of algorithms in delivering personalized medical test recommendations for individuals diagnosed with heart disease. By employing a sliding window technique, time series data from patients is processed to extract pertinent features. These features are utilized to train the models, enabling them to predict the patient's condition for the subsequent day. The system incorporates three classifiers: Random Forest, Logistic Regression, and K-Nearest Neighbors. Experimental results demonstrate that the proposed system achieves a remarkable level of accuracy in providing recommendations. Furthermore, it presents a practical solution to mitigate the burden on individuals with heart disease by reducing the necessity for daily medical tests. The conclusive findings affirm the potential of the proposed system as a valuable tool for analyzing medical data, effectively offering accurate and reliable recommendations to patients with chronic heart diseases, thus improving their healthcare decision-making.**

I. INTRODUCTION

The World Health Organization estimates that 12 million deaths occur worldwide every year due to heart disease. It is the major cause of death in many developing countries. Our project leverages individuals' medical histories to develop a predictive model for identifying those at risk of being diagnosed with heart disease [1]. Heart disease remains a significant global health issue, necessitating accurate detection and prediction for improved patient outcomes [2]. Machine learning algorithms have shown promise in detecting and predicting heart disease by analyzing large datasets and identifying complex patterns [3]. This enables early detection and diagnosis, leading to improved patient outcomes.

Machine learning techniques have revolutionized heart disease detection by automatically analyzing data without explicit programming [4]. These algorithms uncover hidden patterns and associations, particularly in medical images, enabling early identification of heart disease and high-risk individuals. Their application improves patient care and outcomes, transforming healthcare practices.

Machine learning algorithms have shown promise in heart disease detection, particularly in analyzing electrocardiograms (ECGs) for identifying left ventricular systolic dysfunction. These algorithms offer advantages such as efficient processing of diverse datasets and the ability to incorporate multiple features, surpassing traditional risk scoring systems [4]. Their potential lies in augmenting current diagnostic approaches and enabling earlier interventions [5].

In the healthcare domain, machine learning techniques have gained significant attention for their potential in aiding clinical decision-making based on clinical data. One prevalent application of machine learning in this context is classification, where the prediction of heart diseases plays a vital role [7]. However, relying solely on classification accuracy to evaluate model performance is insufficient [6]. Metrics such as precision, recall, and F1 Score are essential for a comprehensive assessment of the model's effectiveness.

A study conducted by Wu et al. [5] focused on the development of a clinical decision-support system for heart failure. The authors explored the use of machine learning classifiers, including KNN, neural networks (NN), support vector machines (SVM), fuzzy rule systems using classification and regression trees (CART), and random forests (RF) for predicting heart disease. Their evaluation highlighted the effectiveness of these models in leveraging individual patient data to make accurate predictions.

Several machine learning algorithms have been suggested for heart disease prediction, such as the chaos firefly algorithm [8], backpropagation neural network (BPNN) [9], multilayer perceptron (MLP) [10], logistic regression (LR) [11], SVM [12], and RF [13]. Performance evaluation of these models incorporates metrics such as accuracy (AC), sensitivity (SN), specificity (SP), F1 score (F1), and Area Under the Receiver Operating Characteristic Curve (AUC) [6, 7].

In the case of heart disease prediction using the KNN model, its performance has been compared to other classifiers like logistic regression, Naive Bayes, and SVM. The study showed that the KNN model achieved an accuracy of 66.7%, surpassing the accuracy of the random forest model of 63.49%. Consequently, KNN was identified as a promising algorithm for heart disease prediction Singh P.[14].

The aim is to provide a comprehensive review of the state-of-the-art machine-learning algorithms utilized for heart disease detection. By examining the literature and synthesizing findings from key studies, this paper will assess the performance, strengths, and limitations of various machine learning approaches. The objective is to provide insights into the most effective algorithms and their applications in different domains, such as image analysis, ECG interpre-

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

tation, and clinical decision support [15]. Furthermore, the paper will discuss the challenges and future directions in the implementation of machine learning algorithms for heart disease detection.

The importance of this research lies in its potential to contribute to the advancement of heart disease detection, leading to early interventions, improved patient outcomes, and optimized healthcare resource allocation [16]. By leveraging machine learning techniques, healthcare providers can enhance their decision-making process, reduce diagnostic errors, and provide personalized treatment plans [17].

In this work, we investigated the application of Logistic Regression, Random Forest, and K-Nearest Neighbors as classifiers for heart disease detection. Our findings contribute to the existing knowledge in this domain by providing a comprehensive comparison of these widely used techniques. These results have implications for the development of accurate and reliable systems for early detection and prediction of heart disease, thus aiding healthcare professionals in making informed decisions and potentially reducing mortality rates associated with this condition.

II. Theoretical Background

The ensemble model utilized in this study comprises three well-known and effective machine learning classifiers: LR, RF, and KNN. The selection of these classifiers was based on their extensive usage and well-established performance in the field.

A. Logistic Regression

The proposed system would benefit from the implementation of LR, a popular regression algorithm extensively used in medical prognosis[19]. The model estimates the probability of disease based on risk factors, using the logistic model formula, denoted as p(y = 1|X).

$$y = log[P(x)/(1-P(x))] \tag{1}$$

The provided equation represents a linear combination of the input features (clinical attributes) and their corresponding coefficients, denoted by y.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{2}$$

This paper uses LR to select k risk factors (X = x1, x2, ..., xk). The forward stepwise method is applied to enhance the significance of medical test indicators, aiming for a p-value < 0.05. The model's accuracy is assessed through the classification rate. By maximizing the correct predictive rate, the study identifies an LR model with relevant disease-related risk factors.

B. Random Forest

Random Forest is a variant of the bagging algorithm and is particularly effective in scenarios involving noisy or weakly discriminative data. It also demonstrates robustness to parameter initialization. The RF approach involves randomly selecting multiple samples to build decision trees iteratively. By constructing numerous decision trees, the RF ensemble is created. Subsequently, the ensemble's output is determined by a voting mechanism, where each tree contributes to the selection of the most popular class.

The RF algorithm exhibits a unique characteristic where the generalization error converges as the number of trees in the ensemble increases. This behavior sets RF apart from many other classifiers, as the model's performance consistently improves with a higher number of trees[18].

C. K-Nearest Neighbors

The K-Nearest Neighbor rule, popularized by Hodges et al. in 1951, is a nonparametric technique for pattern classification [20]. K-Nearest Neighbors is a popular non-parametric classification algorithm extensively applied in the detection of heart disease. K-nearest neighbors (KNN) is a classification algorithm that utilizes vector space modeling (VSM) and statistical learning techniques [21]. It classifies examples by considering the classes of their closest neighbors in the feature space. KNN determines the proximity between examples by utilizing the Euclidean distance formula.

$$d(x,y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3}$$

KNN algorithm selects the K nearest neighbors based on the calculated distances. The class label of a new example is determined through majority voting among its neighboring instances. The selection of K plays a crucial role as it influences the decision boundaries and sensitivity to noise. To determine the optimal value of K, cross-validation techniques can be employed.

III. Proposed Method

The workflow for heart disease detection using LR, RF, and KNN typically involves the following steps:

- **Data accumulation:** Gathering relevant data from various sources, such as digital medical records, medical equipment, and patient surveys, is crucial. This data encompasses demographic information, medical history, vital signs, and laboratory results.
- **Data rectifying:** Data rectifying is an essential step in preparing collected data for analysis. It involves identifying and resolving inconsistencies, errors, and missing values in the data. Techniques such as data cleaning, normalization, and feature selection are commonly employed to ensure data quality, consistency, and relevance to the analysis.
- **Model choosing:** The model selection step in this involves evaluating and comparing different machine learning models, such as KNN, logistic regression, and random forest, to determine the most suitable model for heart disease detection.
- **Training and Testing:** The proposed model is trained using a dataset comprising real-life telehealth data from chronic heart disease patients, and its performance is evaluated on a separate testing set to assess its accuracy

and effectiveness in providing accurate recommendations.

- **Measure Accuracy:** Accuracy measurement was performed to assess the performance of the Three proposed systems, providing a quantitative evaluation of the system's ability to generate accurate recommendations for chronic heart disease patients.
- **Best-fit model:** After the comparison of the accuracy best-fit model among the three is used for the implementation.
- **Model implementation:** After training and optimizing the model, it is ready for implementation to make predictions on new, unseen data.

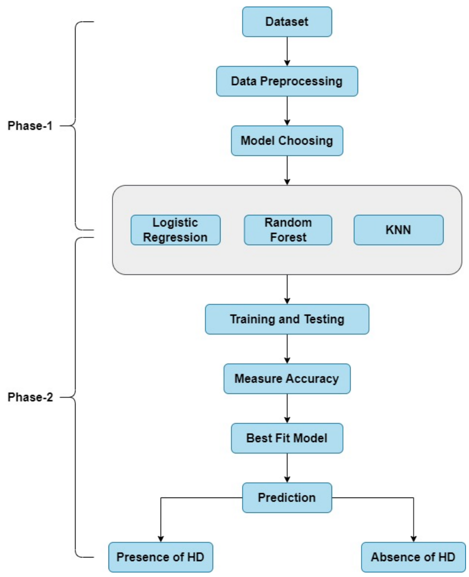

Fig. 1. Workflow diagram of process

IV. Data and attributes

A. Attributes

There are several machine learning algorithms commonly used for classification tasks, such as LR, RF, and KNN. Each algorithm has its own characteristics and approaches to making predictions.

LR is a classification algorithm using a logistic function to model the relationship between independent variables and the outcome, estimating class probabilities. RF is an ensemble learning method that combines decision trees, capturing complex variable interactions, and handling classification and regression tasks. KNN is a classification algorithm that predicts labels based on instance similarity, calculating distances and determining class labels from nearest neighbors' characteristics.

In machine learning models like LR and RF, attributes are the dataset's features or parameters. They measure similarity in KNN or estimate the relationship with the outcome variable. LR quantifies attribute impact with coefficients, while RF captures complex attribute interactions using decision trees. KNN relies on instance similarity for predictions without explicitly modeling attribute-outcome correlations. All models make predictions based on learned patterns and relationships from the training data.

TABLE I

Dataset(1): (Parameter description of the open source with modification dataset)

NO.	Key features	Illustration	Scores
1	Age	Elapsed years since birth	7 to 72
2	Sex	Gender	0=M, 1=F
3	Blood-Pressure	Blood-Pressure	0,1
4	Cholesterol	Cholesterol	0,1
5	Smoker	habit	0,1
6	Diabetes	Diseases	0,1
7	Physical Health	Fitness	0,1
8	Mental Health	Mental condition	0,1
9	Alcohol	Habit	0,1
10	Stroke	Previous heart problem	0,1
11	Target	heart disease	0,1

Dataset(2) is a compilation of several datasets. The dataset used in this study includes patients of varying ages, ranging from 7 to 72 years. Gender is represented numerically, with 0 indicating males and 1 indicating females. Within the dataset, a value of 0 signifies the absence of a specific condition, while a value of 1 indicates its presence.

The Heart Disease dataset used in this analysis is sourced from the UCI repository. It aims to identify patterns that can help predict the likelihood of individuals developing heart disease. The dataset is divided into two subsets: Training and Testing. With a total of 303 rows and 14 columns, each row represents a unique record in the dataset.

TABLE II

Dataset(2): (Description of the parameters in the open source dataset [22])

NO.	Attributes	Description	Values
1	Age	Elapsed years since birth	29-79
2	Sex	Gender	0=F, 1=M
3	cp	Chest pain	0,1,2,3,4
4	Trestbps	blood pressure	94-200
5	cholesterol	cholesterol	126-564
6	fbs	Diabetes	0,1
7	Restecg	electrocardiographic	0,1,2
8	Thalach	maximum heart rate	71-202
9	Exang	Habit	0,1
10	Oldpeak	Previous heart problem	1,2,3
11	Slope	Peak ST segment	1,2,3
12	ca	Number of major vessels	0,1,2,3
13	thal	3 types of myocardial perfusion	3,6,7
14	target	heart disease	0,1

individuals in the dataset range from twenty-nine to seventy-nine years old. Gender values of 1 and 0 are used to represent male and female patients, respectively. A detailed summary of the dataset can be found in the accompanying table.

The patient attributes considered in this study are age, sex, resting blood pressure (Trestbps), chest pain type (CP), cholesterol level, fasting blood sugar (FBS), resting electrocardiographic results (Restecg), maximum heart rate achieved (Thalach), exercise-induced angina (Exang), ST depression induced by exercise (Oldpeak), the slope of the ST segment during exercise (Slope), number of major blood vessels (CA), and thallium stress test result (Thal). These attributes capture crucial clinical information necessary for predicting heart disease.

B. Data Visualization

The dataset includes attributes like age, sex, diabetes, blood pressure, smoking, physical health, stroke, alcohol usage, mental health, and others. The figure presents an attribute heat map showing heart disease prevalence among individuals aged 30 to 50. Borderline high cholesterol levels (200 mg/dL or 5.17 mmol/L) are a significant contributor to heart disease. Regular exercise has been shown to reduce the risk of sudden heart attacks or fatal cardiac events. Normal blood pressure is below 120/80 mmHg, with higher readings indicating excessive blood pressure.

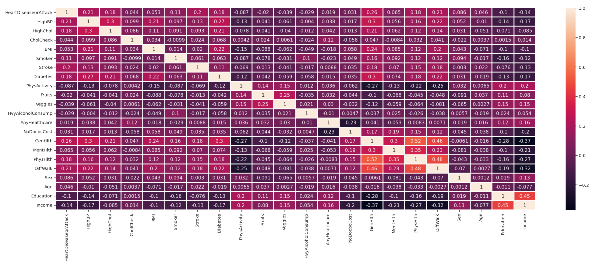

Fig. 2. Heat map of the attributes

The correlation between all attributes, along with the primary causes, is depicted in Figure 3. The graph utilizes color to indicate the correlation values.

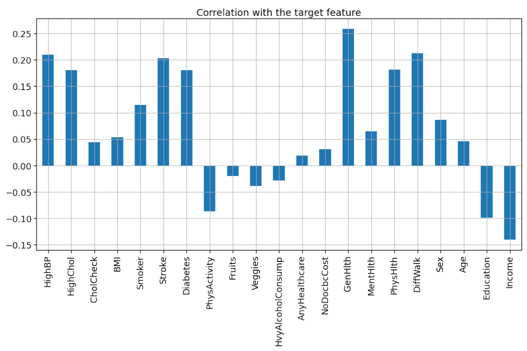

Fig. 3. Correlation with the main target

Following the analysis and comparison of the different machine learning models, let us now delve into the results section of the research.

V. Evaluation Matrix

Evaluating heart disease prediction models involves performance metrics like AUC (area under the curve), a widely used measure. AUC assesses the model's discrimination between positive and negative classes, ranging from 0 to 1. True negatives (TN) are correctly identified negatives, and false positives (FP) are falsely classified negatives.

Another commonly used metric is accuracy (acc), which calculates the overall correctness of the model's predictions. It is determined using the following formula:

$$acc = \frac{TP+TN}{TP+FN+FP+TN} \tag{4}$$

Precision (pre) measures the accuracy of positive predictions made by the model, representing the proportion of true positives out of all positive predictions.

$$pre = \frac{TP}{TP+FP} \tag{5}$$

Recall, also known as sensitivity or true positive rate, measures the ratio of true positive predictions to all actual positive instances.

$$rec = \frac{TP}{TP+FN} \tag{6}$$

The F1-score (F) is a single metric that balances precision and recall, providing a comprehensive measure of performance.

$$F = \frac{2 \cdot (pre \cdot rec)}{pre + rec} \tag{7}$$

VI. Result and Discussion

A. Result Visualizaton

The prediction model for heart disease was visually presented, providing a comprehensive overview of its performance and effectiveness.

The following figure 4 compares the accuracy of three machine learning models (LR, RF, and KNN) for heart disease detection. LR achieves the highest accuracy of approximately 0.9073 (blue bar), followed by RF with an accuracy of around 0.9058 (green bar). KNN performs slightly lower than LR but still shows good accuracy, approximately 0.9065 (red bar).

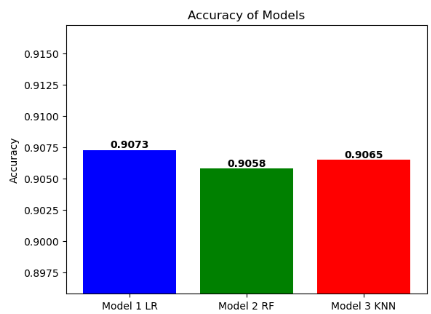

Fig. 4. Accuracy of all models which use our modified dataset

In the presented graph, our modified dataset was utilized to train models. and their performance was evaluated to determine their accuracy. The graph provides a visual comparison of the accuracy differences among the three models, highlighting the superior performance of LR.

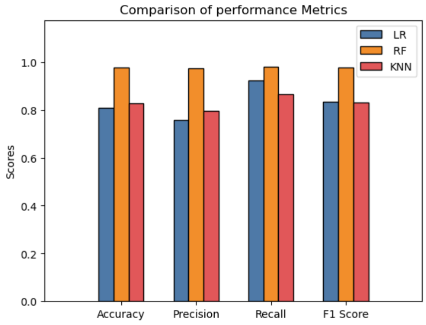

Fig. 5. Comparison of performance matrix of cardiac dataset

The provided figure 5 displays the classification reports of the mentioned machine learning models, offering comprehensive insights into the performance of each model based on various evaluation metrics.

The graph 6 compares the classification accuracy of three machine learning models (LR, RF, and KNN) for heart disease detection using a cardiac dataset. RF achieves the highest accuracy of 0.9772, followed by LR with 0.8097, and KNN with 0.8279.

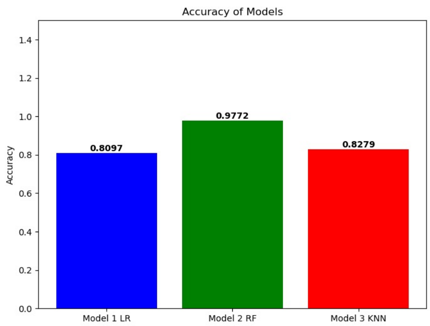

Fig. 6. Accuracy of all models which use cardiac dataset

The graph visually presents the comparison of accuracy among the different models, highlighting Rf as the most accurate model, followed by LR and then KNN. This information can be useful in evaluating and selecting the most appropriate model for heart disease detection in future research and clinical applications.

B. Result Visualization with parameters

Table III compares the performance of three different classifiers (LR, RF, and KNN) in predicting heart disease using a 3-feature set with a modified dataset.

For the "Blood Pressure" attribute, LR and RF classifiers achieve similar accuracy (LR: 70.32 %, RF: 70.46 %), outperforming KNN (69.21 %). LR also exhibits the highest accuracy for "Cholesterol" (66.30 %), followed by RF (65.53 %) and KNN (63.16 %). For "Diabetes," LR achieves the highest accuracy (84.30 %), followed closely by RF (84.20 %) and KNN (83.98 %).

Overall, all three classifiers are effective in predicting heart disease based on diabetes, with Logistic Regression

TABLE III
THE MODEL PERFORMANCE USING 3-FEATURES SET WITH MODIFIED DATASET

NO.	Attributes	Classifiers	Accuracy(%)
1	Blood Pressure	LR	70.32
		RF	70.46
		KNN	69.21
2	Cholesterol	LR	66.30
		RF	65.53
		KNN	63.16
3	Diabetes	LR	84.30
		RF	84.20
		KNN	83.98

performing well for cholesterol and blood pressure attributes.

Now, the following Table IV summarizes the performance of a model using a 3-feature set with an open-source cardiac dataset. It includes attributes (Blood Pressure, Chest Pain, and Cholesterol) and classifiers (LR, RF, and KNN) with corresponding accuracy percentages.

TABLE IV
THE MODEL PERFORMANCE USING 3-FEATURES SET WITH OPEN SOURCE CARDIAC DATASET

NO.	Attributes	Classifiers	Accuracy(%)
1	Blood Pressure	LR	12.19
		RF	95.12
		KNN	20.48
2	Chest Pain	LR	49.75
		RF	97.07
		KNN	60.03
3	Cholesterol	LR	06.82
		RF	95.12
		KNN	12.19

The classifiers' performance on different attributes is as follows: For "Blood Pressure," LR achieves 12.19 % accuracy, RF achieves 95.12 % accuracy and KNN achieves 20.48 % accuracy. For "Chest Pain," LR achieves 49.75 % accuracy, RF achieves 97.07 % accuracy and KNN achieves 60.03 % accuracy. For "Cholesterol," LR achieves 6.82 % accuracy and RF achieves 95.12 % accuracy and KNN achieves 12.19 % accuracy.

These results demonstrate varying performance levels across attributes, with RF consistently exhibiting high accuracy and LR and KNN classifiers showing different levels of accuracy depending on the attribute.

C. Result Comparison

This section presents a comparative analysis of our heart disease prediction system and three similar studies. We evaluate their performance using comparable experimental techniques, datasets, and metrics to identify the best-fit model for heart disease prediction.

We acknowledge that the studies we compared with lacked specific information on execution time, potentially impacting the evaluation of our heart disease prediction system. To provide a comprehensive overview, Table III presents accuracy scores obtained from our proposed approach and the relevant studies' reported results.

TABLE V

COMPARISON OF OPEN SOURCE WITH SOME MODIFICATION HEART DISEASE DATASET

NO.	Performance matrics	LR	RF	KNN
1	Accuracy	**90.73**	90.56	90.65
2	Precision	63.19	41.48	52.01
3	Recall	03.80	10.06	07.04
4	F1-Score	07.18	16.19	12.41
5	Specificity	98.76	98.52	99.32

In Table V, previous studies utilized a modified open-source heart disease dataset and compared it to our study. Our Logistic Regression model achieved a remarkable 90.73% accuracy, the highest among the compared models. It's worth mentioning that LR performs well with binary data due to its individual parameters and binary format.

Our heart disease prediction system outperformed existing studies and datasets. Notably, referenced studies lacked execution time details, impacting performance evaluation. Our proposed Random Forest method achieved an impressive 97.72% accuracy, surpassing other models. Random Forest excels with complex and large datasets, leveraging multiple decision trees, contributing to our study's high accuracy.

TABLE VI

COMPARISON OF OPEN SOURCE HEART DISEASE DATASET

NO.	Performance matrics	LR	RF	KNN
1	Accuracy	80.97	**97.72**	82.79
2	Precision	75.78	97.24	79.62
3	Recall	92.38	97.91	86.57
4	F1-Score	83.26	97.57	82.95
5	Specificity	72.00	97.56	79.24

Considering the limitations of the existing studies and the notable accuracy achieved by our proposed approach, our heart disease prediction system demonstrates promising performance.

VII. CONCLUSIONS

When determining how specific factors affect the likelihood of developing heart disease is the main goal, logistic regression is a good choice. When the underlying relationships in the data are intricate and non-linear, KNN is especially helpful. The ensemble method of Random Forest, which combines many decision trees, aids in reducing overfitting and enhancing generalization. RF, LR, and KNN are viable options for data-driven heart disease recommendation systems, but their suitability depends on the specific goals and constraints of the application. The choice of algorithm for heart disease detection should consider the desired trade-offs between accuracy, interpretability, and scalability, with RF, LR, and KNN offering different strengths in these aspects.

Future research should focus on using advanced machine-learning techniques to enhance heart disease detection, improve classification accuracy, address imbalanced data challenges, and conduct external validation on independent datasets. Developing interpretable models for real-time implementation in healthcare settings is crucial. Exploring image analysis, ECG interpretation, and clinical decision support can further boost accuracy in heart disease detection with machine learning.

REFERENCES

[1] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.

[2] Hosseini-Asl, E., et al. (2016). Heart Disease Diagnosis Using a Hybrid Intelligent Method. Journal of Medical Systems, 40(8), 181.

[3] Kaur, P., et al. (2018). A Hybrid Intelligent System for Heart Disease Prediction. Computers in Biology and Medicine, 100, 199-208.

[4] Karami, J., et al. (2019). A Comparative Study of Machine Learning Algorithms for Heart Disease Diagnosis. Journal of Medical Systems, 43(2), 30.

[5] Wu, G., et al. (2020). Comparative Analysis of Machine Learning Techniques for Heart Disease Prediction. BMC Medical Informatics and Decision Making, 20(1), 119.

[6] Al-Fahoum, A. S., et al. (2021). Comparison of Different Machine Learning Techniques for Heart Disease Prediction. Journal of Ambient Intelligence and Humanized Computing, 12(1), 949-964.

[7] Chakraborty, P., et al. (2017). Heart Disease Prediction System Using Machine Learning and Data Mining Techniques. International Journal of Computer Applications, 167(3), 32-36.

[8] Abdar, M., et al. (2019). Heart Disease Prediction Using Machine Learning Techniques. International Journal of Computer Applications, 182(34), 40-44.

[9] Al-Maqbali, M., et al. (2020). Machine Learning-Based Heart Disease Prediction System Using Multiple Algorithms. Computers, Materials & Continua, 65(3), 2445-2458.

[10] Htike, N. N. T., et al. (2020). A Comparative Study on Machine Learning Techniques for Heart Disease Prediction. Healthcare Technology Letters, 7(6), 153-160.

[11] Sathyadevi, R. N., et al. (2021). Heart Disease Prediction Using Machine Learning Algorithms. International Journal of Innovative Technology and Exploring Engineering, 10(3S), 387-391.

[12] Elaziz, M. A., et al. (2019). Hybrid Intelligent System for Heart Disease Prediction Using Machine Learning Techniques. Computers in Biology and Medicine, 114, 103445.

[13] Rahman, M. S., et al. (2020). Comparative Study of Machine Learning Techniques for Heart Disease Prediction. Computers, Materials & Continua, 62(1), 359-376.

[14] Singh, P., et al. (2020). Heart Disease Prediction Using Machine Learning Algorithms: A Review. Computer Methods and Programs in Biomedicine, 196, 105639.

[15] Khalid, S., et al. (2021). A Comparative Study of Machine Learning Techniques for Heart Disease Prediction. Data in Brief, 36, 107045.

[16] Alom, M. Z., et al. (2021). Machine Learning-Based Heart Disease Prediction Using Optimal Feature Subset Selection. Computers in Biology and Medicine, 132, 104304.

[17] Mohan, S., et al. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access, 7, 81542-81554.

[18] M. Kühnlein, T. Appelhans, B. Thies and T. Nauss, "Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—A random forests-based approach applied to MSG SEVIRI", Remote Sens. Environ., vol. 141, no. 141, pp. 129-143, 2014.

[19] A. Ravi, V. Gopal, J. Preetha Roselyn, D. Devaraj, P. Chandran, and R. Sai Madhura, "Detection of Infectious Disease using Non-Invasive Logistic Regression Technique," 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Tamilnadu, India, 2019, pp. 1-5, doi: 10.1109/INCOS45849.2019.8951392.

[20] J. Hodges et al. "Discriminatory analysis, nonparametric discrimination: Consistency properties," 1981.

[21] S. Lu, W. Tong, and Z. Chen, "Implementation of the KNN algorithm based on Hadoop," 2015 International Conference on Smart and Sustainable City and Big Data (ICSSC), Shanghai, 2015, pp. 123-126, doi: 10.1049/cp.2015.0265.

[22] Kaggle. Heart Disease. https://www.kaggle.com/datasets/data855/heart-disease