WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Machine Learning-Based Prediction and Diagnosis of Heart Disease: A Data-Driven Approach for Early Detection

Subhasis Misra[1], Divya Sharma[2]

[1,2] Department of Computer Science and Engineering,
University of Engineering and Management, Jaipur, India

ABSTRACT:

Cardiovascular Diseases (CVDs) have become the leading cause of death globally over recent decades, posing a significant threat to life not only in India but worldwide. Therefore, there is an urgent requirement for a dependable, precise, and practical system to timely diagnose these diseases for appropriate treatment. Data analytics plays a crucial role in making predictions from extensive information, assisting medical facilities in forecasting various illnesses. A significant amount of data related to patients is gathered on a monthly basis. This stored information can serve as a resource for forecasting the potential emergence of future diseases. This research paper discusses different characteristics associated with heart disease and the models derived from supervised learning techniques like Naïve Bayes, decision trees, K-nearest neighbors, and the random forest algorithm. It utilizes a previously established dataset from the Cleveland database located in the UCI repository, which pertains to individuals with heart disease. The dataset consists of 303 instances and 76 attributes, out of which only 14 attributes are utilized for testing, as they are crucial for evaluating the performance of different algorithms. This research paper seeks to assess the likelihood of patients developing heart disease. The findings indicate that the K-nearest neighbor algorithm achieves the highest accuracy score. The primary aim of this research project is to utilize machine learning algorithms to predict whether a patient has heart disease.

Keywords: Neural Network, Machine Learning, Supervised learning, Support vector machine, Random forest.

Introduction:

The heart is a crucial organ in the human body, responsible for pumping blood throughout our system. If it does not operate properly, the brain and other vital organs will cease to function, leading to death within minutes. Lifestyle changes, work-related stress, and poor dietary choices have contributed to the rising rates of various heart diseases. Heart-related ailments have become one of the leading causes of death worldwide. The World Health Organization reports that these diseases are accountable for 17.7 million deaths annually, representing 31% of all global fatalities. In India, heart diseases have also emerged as the top cause of death. In 2016, heart diseases claimed the lives of 1.7 million Indians, according to the Global Burden of Disease Report published on September 15, 2017.

Identifying heart disease can be challenging due to various risk factors such as diabetes, hypertension, high cholesterol levels, irregular pulse rates, and several others. Different approaches in data mining and neural networks have been utilized to assess the severity of heart disease in individuals. The assessment of the severity of diseases utilizes techniques such as the K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic Algorithm (GA), and Naive Bayes (NB) [11], [13]. Given the intricate nature of heart disease, it is important to handle it with care; failure to do so may compromise heart function or lead to early mortality. Medical science and data mining perspectives are employed to uncover various types of metabolic syndromes. The role of data mining with classification is crucial in predicting heart disease and conducting data analysis.

We have also observed the application of decision trees in forecasting the outcomes of events associated with heart disease [1]. Various techniques have been utilized for knowledge abstraction through established data mining methods to predict heart disease. In this study, multiple analyses have been performed to create a prediction model, employing not only individual techniques but also by integrating two or more methods. These combined new approaches are often referred to as hybrid methods [14]. We present neural networks that use time series of heart rate data. This approach utilizes several clinical records for prediction, including Left bundle branch block (LBBB), Right bundle branch block (RBBB), Atrial fibrillation (AFIB), Normal Sinus Rhythm (NSR), Sinus bradycardia (SBR), Atrial flutter (AFL), Premature Ventricular Contraction (PVC), and Second-degree block (BII) to ascertain the patient's condition concerning heart disease. The dataset with a radial basis function network (RBFN) is employed for classification, where 70% of the data is allocated for training and the remaining 30% is used for classification [4], [15].

These datasets, which can be too complex for human understanding, can be effectively examined through various machine learning methods. Consequently, these algorithms have recently proven to be very beneficial in accurately predicting the presence or absence of heart-related diseases. The integration of information technology in the healthcare sector is steadily increasing to support doctors in their decision-making processes. It assists

medical professionals in managing diseases, prescribing medications, and uncovering patterns and connections within diagnostic data. Current methods for predicting cardiovascular risk fall short in identifying many individuals who could gain from preventive treatment, while others undergo unnecessary interventions. Machine learning presents a chance to enhance accuracy by leveraging the intricate interactions among risk factors. We evaluated whether machine learning can enhance the prediction of cardiovascular risk.

Literature Survey:

ChalaBeyene et al. [1] suggested a methodology for forecasting and analyzing the incidence of heart disease using data mining techniques. The main objective is to predict instances of heart disease for timely automated diagnosis in a brief period. This proposed approach is particularly important for healthcare organizations that deal with professionals who may lack extensive knowledge and expertise. It incorporates various medical factors, including blood sugar, heart rate, age, and sex, among others, to determine whether an individual has heart disease. The dataset analyses are performed using WEKA software. Senthilkumar Mohan et al. [2] executed a hybrid machine learning approach for predicting heart disease. They utilized the Cleveland dataset for their research. The initial phase is data preprocessing, where tuples with missing values are excluded from the dataset. They also chose not to include the attributes of age and sex, as the authors believe these are personal information that does not influence the prediction. The remaining eleven attributes are considered essential since they contain critical clinical data. They developed a Novel Hybrid Random Forest Linear Method (HRFLM), which integrates Random Forest (RF) with Linear Method (LM). The HRFLM algorithm incorporates four distinct algorithms. The first algorithm focuses on partitioning the input dataset, which employs a decision tree that is executed for each sample. Upon identifying the feature space, the dataset is divided into leaf nodes. The output of this first algorithm is a partitioned dataset. In the second algorithm, rules are applied to the dataset, with the output being the classification of data according to those rules. The third algorithm extracts features using a Less Error Classifier, which works on determining the minimum and maximum error rates from the classifier. The output here consists of features with classified attributes. In the fourth algorithm, they utilize a Classifier that employs a hybrid method based on error rates derived from the Extracted Features. Ultimately, they compared the results from applying HRFLM against other classification algorithms, such as decision trees and support vector machines. The results showed that RF and LM produced superior outcomes compared to the others; thus, both algorithms were integrated to form the unique HRFLM algorithm. The authors recommend further enhancements in accuracy through the combination of various machine learning algorithms. Ali, Liaqat, et al. [3] proposed a system comprising two models based on linear Support Vector Machine (SVM). The first model is referred to as L1 regularized, while the second is named L2 regularized. The first model is employed to eliminate unnecessary features by setting the coefficients of those features to zero. The second model focuses on prediction, which is conducted in this segment. To enhance both models, they introduced a hybrid grid search algorithm, optimizing the two models using metrics such as accuracy, sensitivity, specificity, Matthews correlation coefficient, ROC chart, and area under the curve. They used the Cleveland dataset, with data split into 70% for training and 30% for testing, employing holdout validation. Two experiments were conducted, with each experiment testing various values of C1, C2, and k—where C1 is the hyper-parameter of the L1 regularized model, C2 is the hyper-parameter of the L2 regularized model, and k denotes the size of the selected feature subset. The initial experiment consisted of the L1-linear SVM model combined with the L2-linear SVM model, resulting in a peak testing accuracy of 91.11% and a training accuracy of 83.06%. The second experiment utilized the L1-linear SVM model cascaded with the L2-linear SVM model incorporating an RBF kernel.

This research demonstrates a peak testing accuracy of 92.22% and a training accuracy of 85.02%. They have achieved a 3.3% enhancement in accuracy compared to traditional SVM models. Singh, Yeshvendra K. et al[4] explore various supervised machine learning techniques, including Random Forest, Support Vector Machine, Logistic Regression, Linear Regression, and Decision Tree, utilizing 3-fold, 5-fold, and 10-fold cross-validation methods. They utilized the Cleveland dataset, comprising 303 tuples, though some of them have missing attributes. During data preprocessing, they removed the six tuples with missing values, resulting in 297 remaining tuples, which were subsequently split into a training set of 70% and a testing set of 30%. The first algorithm implemented was Linear Regression, where they established the relationship between one attribute and others that can be separated linearly. Essentially, the classification is performed using a group of attributes designated for binary classification. Their best outcome was noted in the 10-fold validation, yielding an accuracy of 83.82%. Logistic regression classification employs a sigmoid function. This algorithm, used in predicting heart disease, achieved maximum accuracy with both 3- and 5-fold cross-validation, reaching 83.83%.

The Support Vector Machine functions as the classification approach in supervised machine learning, with classification carried out via a hyperplane. The best accuracy achieved by SVM during 3-fold cross-validation was 83.16%. For the Decision Tree, the researchers tested different amounts of splits and leaf nodes to find the best accuracy. By using a setup of 36 splits and 5 leaf nodes, they reached a top accuracy of 78.12%. When applying cross-validation, the decision tree achieved 79.54% accuracy with 5-fold validation. The Random Forest algorithm, when applied to a nonlinear dataset, yields superior results compared to the decision tree. Random forest consists of a collection of decision trees generated from distinct root nodes. In this ensemble of decision trees, a voting process is conducted to determine the classification based on the tree that receives the highest number of votes. The authors experimented with various numbers of splits, trees per observation, and folds for cross-validation. The random forest achieved an accuracy of 85.81% with 20 splits, 75 trees, and 10 folds.

Data set:

Cardiovascular diseases (CVDs) rank as the leading global cause of death, resulting in approximately 17.9 million fatalities annually, which constitutes 31% of all deaths worldwide. Heart failure is a common condition associated with cardiovascular diseases, and this dataset comprises 12 attributes that can help forecast mortality related to heart failure. Many cardiovascular diseases can be avoided by tackling behavioral risk factors such as tobacco consumption, poor dietary habits and obesity, lack of physical activity, and excessive alcohol use through community-wide initiatives. Individuals with

cardiovascular disease or those at heightened cardiovascular risk (due to one or more risk factors like hypertension, diabetes, hyperglycemia, or existing conditions) require early detection and management, where a machine learning model could be highly beneficial.

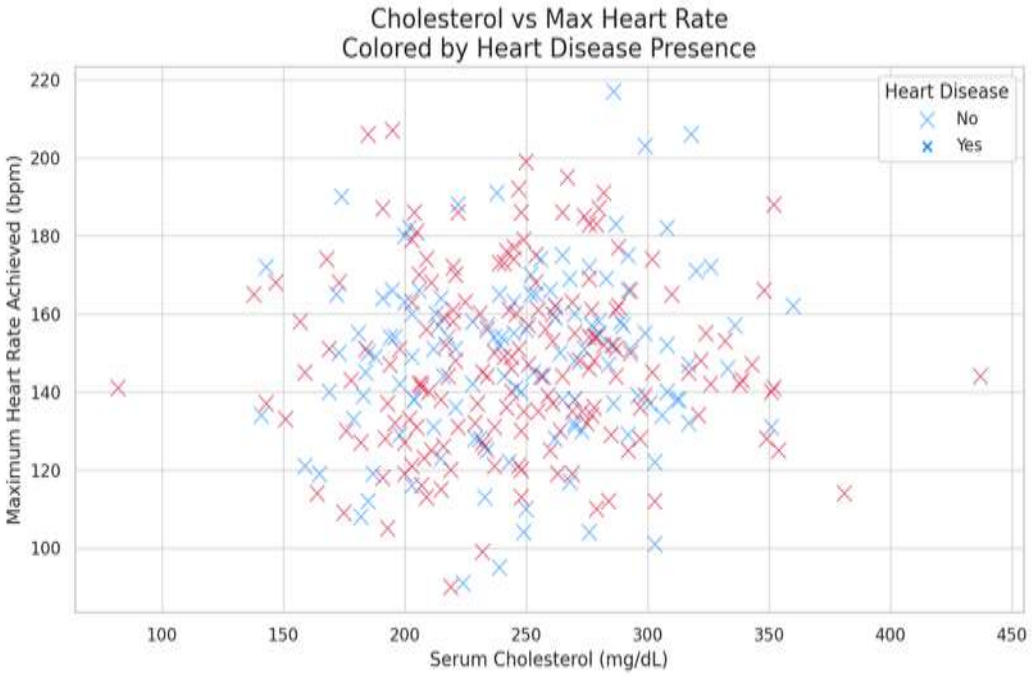

Fig 1: Cholesterol vs Max Heart rate colored by Heart Disease presence.

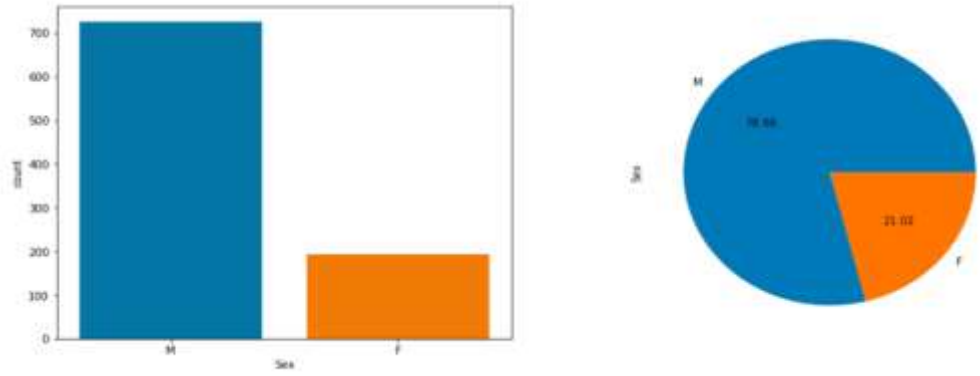

Fig 2: Relation of male and female for heart disease.

Methodology:

The primary goal of this research is to equip healthcare providers with a tool for the early diagnosis of heart issues. This will facilitate timely and effective treatment for patients, thereby preventing severe outcomes. In this research, we compared the efficacy of various machine learning models utilizing the Jellyfish algorithm and feature selection techniques for predicting heart disease, aiming to identify the most effective machine learning model. Figure 6 illustrates a summary of the proposed methodology. As depicted in Figure 6, the Jellyfish algorithm, introduced in 2021, was first applied to the dataset to extract the most relevant features. The Jellyfish algorithm seeks optimal solutions to a variety of optimization challenges by mimicking the intelligent behavior of jellyfish. This algorithm is capable of avoiding local minima and achieving the global minimum more quickly than other optimization methods. Its simplicity, minimal parameter requirements, and adaptability have made the Jellyfish algorithm popular worldwide. Due to these benefits, it was chosen for this research to identify the best features from the dataset. The Jellyfish algorithm is crucial for feature selection, and this study utilizes a binary version of it. This algorithm begins with a population, which represents a set of potential solutions containing the most significant features. In each iteration of the algorithm, the most effective features are chosen for transfer to the next stage, leading to the optimal solution for the features. After generating a new dataset consisting of the selected features, this dataset was utilized to train four distinct classifiers: ANN, DT, Adaboost, and SVM. The performances of the machine learning models that resulted from the training were evaluated and compared using metrics such as Accuracy, Sensitivity, Specificity, and Area Under the Curve, with the top-performing model being chosen. A 10-fold cross-validation approach was employed during both the training and testing phases of the ML algorithms. The chosen model demonstrated excellent performance in differentiating and classifying new data samples into two categories: no disease and diseased.

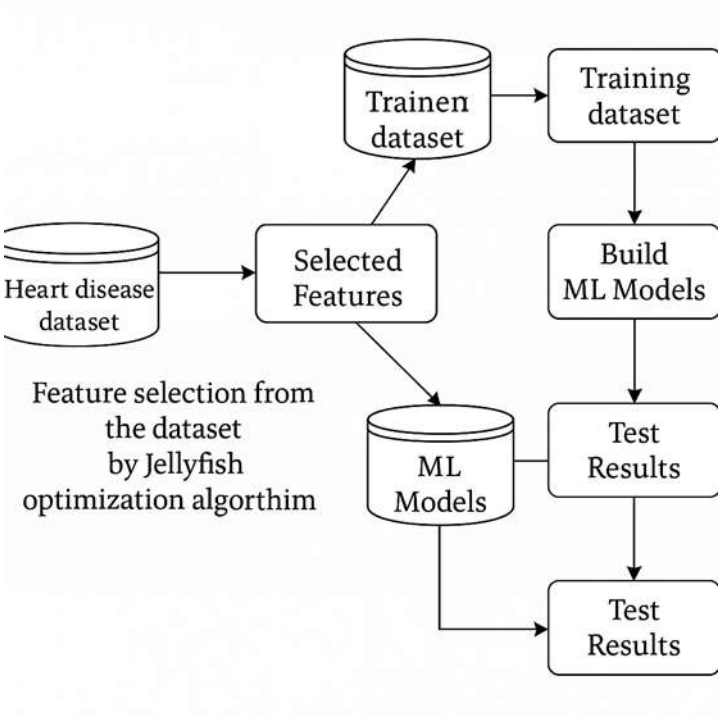

Fig 3: Flowchart of the proposed approach for heart disease prediction.

The Basic of python:

Python is a popular, interpreted, general-purpose, high-level programming language that was created by Guido van Rossum in 1991. It is a straightforward language characterized by English-like syntax, which contributes to its widespread adoption. Compared to other programming languages, it requires fewer lines of code to achieve similar functionality. Python is primarily known for its speed and efficient integration with various systems. The language is employed for creating both web applications and complex scientific applications. Additionally, Python is effective for data analysis, and many libraries are available to facilitate data visualization through those resources.

The basic data science and machine learning:

Data science integrates mathematics and statistics, specialized programming skills, advanced analytics, artificial intelligence (AI), and machine learning along with domain expertise to reveal actionable insights that lie within an organization's data. These insights can inform decision-making and strategic development. Machine learning (ML) is a form of artificial intelligence (AI) that enables software applications to enhance their accuracy in predicting outcomes without requiring explicit programming. Machine learning algorithms utilize historical data as input to forecast new output values.

Confusion matrix:

In the domain of machine learning, particularly in the context of statistical classification, a confusion matrix is widely recognized as an error matrix. This specific tabular format allows for the visualization of how well an algorithm is performing. Typically, it is used in supervised learning, while in unsupervised learning, it is frequently referred to as a matching matrix. Each row in the matrix represents the instances classified as a predicted class, whereas each column indicates the instances that belong to an actual class, or the arrangement could be reversed.

Accuracy:

In a classification task, the outcome achieved for any class is regarded as correct when assessing the experiment's accuracy. Performance metrics such as accuracy, sensitivity, specificity, and precision are used to evaluate effectiveness. Accuracy is determined by comparing the number of correctly predicted instances to the total number of input samples. This metric is most effective when there is an equal distribution of samples across each class.

Correlation:

A correlation matrix is a table that displays the correlation coefficients between different variables. This matrix shows the correlation for every possible pair of values in a dataset. It acts as a powerful tool for summarizing large datasets and aids in recognizing and visualizing patterns within the data. Each cell in the table contains the correlation coefficient. Additionally, the correlation matrix is frequently used in conjunction with other types of statistical analysis. For instance, it can be valuable when examining multiple linear regression models. It is essential to understand that these models consist of several independent variables. In the realm of multiple linear regression, the correlation matrix reveals the correlation coefficients among the independent variables involved in the model.

Result And Discussion:

We utilize several libraries[5] available in Python to carry out this project. Following our experiments, the Random Forest algorithm yields the highest test accuracy at 0.85%. Its superior performance can be attributed to its flexibility concerning the dataset's properties. Naive Bayes assumes that features are independent of each other. Logistic regression relies on features being linearly separable. SVM mandates that the parameters be precisely configured, while neural networks necessitate a complex and extensive dataset. Although we achieved an accuracy of 88.04%, it remains insufficient as it does not eliminate the possibility of incorrect diagnoses. To enhance accuracy, we aim to acquire more datasets since 920 instances are inadequate for optimal performance. In the future, we plan to explore predictions for various diseases, such as lung cancer, by employing image detection techniques. This will complicate the dataset, allowing us to implement a convolutional neural network for more precise predictions.

Algorithm	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Naïve Bayes	82.5	80.2	81.0	80.6
Decision Tree	84.1	83.5	82.0	82.7
K-Nearest Neighbor	87.8	86.9	85.4	86.1
Random Forest	86.3	85.2	84.7	84.9

Conclusion

To forecast the likelihood of patients having heart disease, a confusion matrix (data set) was developed, where 1 represents patients with heart disease, and 0 signifies those without it. A confusion matrix provides insights into the actual and predicted classifications made by a classification system. The details within the matrix are analyzed to assess the performance of such systems. The confusion matrix consists of four components: *TP (true positive): The count of records identified as true while being genuinely true.*FP (false positive): The count of records identified as true when they were actually false.*FN (false negative): The count of records identified as false when they were genuinely true.*TN (true negative): The count of records identified as false when they were indeed false. The comprehensive approach of an effective heart disease prediction system (EHDPS) involves three key steps: 1. Data gathering, 2. Data preparation, 3. Data classification. The data is obtained from a reputable data set comprising 920 records. The twelve parameters, such as age, sex, chest pain type (CP), and cholesterol (chol), along with their respective domain values, are taken into account to predict the likelihood of heart disease, as presented in the dataset.

References:

1. Sapkal, S., Jadhav, S., Mallikarjun, P., Shamim, R., Islam, A. U., & Bamane, K. (2024, June). Innovative Healthcare Advancements: Harnessing Artificial and Human Intelligence for Bionic Solutions. In *2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0* (pp. 1-5). IEEE.
2. Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), 43-48.
3. Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*(6), 345.
4. Shamim, R., & Farhaoui, Y. (2023, November). Enhancing cloud-based machine learning models with federated learning techniques. In *The International Conference on Artificial Intelligence and Smart Environment* (pp. 594-606). Cham: Springer Nature Switzerland.
5. Shamim, R., & Lahby, M. (2023). Automated detection and analysis of cyberbullying behavior using machine learning. In *Combatting Cyberbullying in Digital Media with Artificial Intelligence* (pp. 116-136). Chapman and Hall/CRC.
6. Raman, R., Shamim, R., Akram, S. V., Thakur, L., Pillai, B. G., & Ponnusamy, R. (2023, January). Classification and contrast of supervised machine learning algorithms. In *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)* (pp. 629-633). IEEE.
7. Shamim, R. (2022). Machine learning's algorithm profoundly impacts predicting the share market stock's price. *IJFMR-International Journal For Multidisciplinary Research*, *4*(5), 1-9.
8. Shamim, R., & Bentalha, B. (2023). Blockchain-enabled machine learning framework for demand forecasting in supply chain management. In *Integrating Intelligence and Sustainability in Supply Chains* (pp. 28-48). IGI Global.
9. Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing.

10. Shamim, R., Alfurhood, B. S., Agarwal, T., & Mallik, B. B. (2024). YOLOv8 for Anomaly Detection in Surveillance Videos: Advanced Techniques for Identifying and Mitigating Abnormal Events. *Mathematical Modeling for Computer Applications*, 317-349.
11. Arshad, M., Farhaoui, Y., & Shamim, R. (2024, April). Optimizing Hyperparameters for Fraud Detection: A Comparative Analysis of Machine Learning Algorithms. In *The International Workshop on Big Data and Business Intelligence* (pp. 218-228). Cham: Springer Nature Switzerland.
12. Ticku, A., Tripathy, N., Mishra, P. K., Sinha, A., Jadon, S., Raj, A., ... & Shamim, R. (2023, July). Vader protocol based sentiment analysis technique using LSTM for Stock Trend prediction. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-7). IEEE.
13. Shamim, R., Mallik, B. B., & Agarwal, T. (2024). Advancements in Enhancing Car Object Detection in Complex and Adverse Environmental Conditions Through Deep Learning Techniques. *Mathematical Modeling for Computer Applications*, 29-60.
14. Shamim, R., Alfurhood, B. S., & Mallik, B. B. (2024). Refining Medical Text Query Responses: Tailoring Hugging Face's BERT Model for Precise and Swift Medical Question Answering. *Mathematical Modeling for Computer Applications*, 241-262.
15. Shamim, R., & Agarwal, T. (2024). Optimizing Crop Yield Prediction Using Machine Learning Algorithms. *Smart Agritech: Robotics, AI, and Internet of Things (IoT) in Agriculture*, 443-487.
16. Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms*, *16*(2), 88.
17. Marimuthu, M., Abinaya, M., Hariesh, K. S., Madhankumar, K., & Pavithra, V. (2018). A review on heart disease prediction using machine learning and data analytics approach. *International Journal of Computer Applications*, *181*(18), 20-25.
18. Shamim, R., Farhaoui, Y., & Arshad, M. (2024, April). Genomic Insights Revealed: Multiclass DNA Sequence Classification Using Optimized Naive Bayes Modeling. In *The International Workshop on Big Data and Business Intelligence* (pp. 210-221). Cham: Springer Nature Switzerland.
19. Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2016). Heart disease prediction system using data mining techniques and intelligent fuzzy approach: a review. *international journal of computer applications*, *136*(2), 43-51.
20. Shamim, R., & Farhaoui, Y. (2024, April). An In-depth Comparative Study: YOLOv3 vs. Faster R-CNN for Object Detection in Computer Vision. In *The International Workshop on Big Data and Business Intelligence* (pp. 266-277). Cham: Springer Nature Switzerland.
21. Shamim, Rejuwan, Badria Sulaiman Alfurhood, and Biswadip Basu Mallik. "Refining Medical Text Query Responses: Tailoring Hugging Face's BERT Model for Precise and Swift Medical Question Answering." *Mathematical Modeling for Computer Applications* (2024): 241-262.
22. Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. *Heart Disease*, *7*(1), 129-137.
23. Shamim, R., & Shaikh, A. (2024). Harnessing the power of hugging face's multilingual transformers: unravelling the code-mixed named entity recognition enigma. *International Journal of Intelligent Engineering Informatics*, *12*(3), 353-376.
24. Kundu, P. J. A. S. R., Bagaria, R. S. M. K., & Punia, Y. S. R. P. (2023). AI Driven False Data Injection Attack Recognition Approach for Cyber-Physical Systems in Smart Cities. *Journal of Smart Internet of Things (JSIoT)*, *2023*(02), 13-32.
25. Shamim, Rejuwan, Yousef Farhaoui, and Md Arshad. "Genomic Insights Revealed: Multiclass DNA Sequence Classification Using Optimized Naive Bayes Modeling." In *The International Workshop on Big Data and Business Intelligence*, pp. 210-221. Cham: Springer Nature Switzerland, 2024.
26. Kundu, P. J. A. S. R., Bagaria, R. S. M. K., & Punia, Y. S. R. P. (2023). AI Driven False Data Injection Attack Recognition Approach for Cyber-Physical Systems in Smart Cities. *Journal of Smart Internet of Things (JSIoT)*, *2023*(02), 13-32.
27. Katarya, R., & Meena, S. K. (2021). Machine learning techniques for heart disease prediction: a comparative study and analysis. *Health and Technology*, *11*(1), 87-97.
28. Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, *1*(8), 1-4.