IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 10, October 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118||| A Monthly Peer Reviewed & Referred Journal |

||Volume 11, Issue 10, October 2022||

|DOI:10.15680/IJIRSET.2022.1110102|

Heart Disease Prediction: A Study of Machine Learning Algorithm Accuracy

Prof. Zeba Vishwakarma

Department of Computer Science and Engineering, Baderia Global Institute of Engineering and Management, Jabalpur, MP, India

ABSTRACT : The heart plays a crucial role in the functioning of living organisms. The diagnosis and prediction of heart-related diseases demand high precision and accuracy, as even minor errors can lead to severe health issues or fatality. The number of heart disease-related deaths is rising exponentially. Therefore, an effective prediction system is essential for early awareness and prevention. Machine learning, a branch of Artificial Intelligence (AI), offers significant support in predicting various events by learning from historical data. This paper evaluates the accuracy of several machine learning algorithms, including k-nearest neighbor, decision tree, linear regression, and support vector machine (SVM), in predicting heart disease. The UCI repository dataset is used for training and testing these algorithms. For implementation, Python programming with the Anaconda (Jupyter) notebook is utilized, which provides a range of libraries and tools to enhance the accuracy and precision of the work.

KEYWORDS: Heart Disease Prediction, Machine Learning Algorithms, Artificial Intelligence, k-Nearest Neighbour, Decision Tree, Support Vector Machine (SVM), Linear Regression

I. INTRODUCTION

The heart is a crucial and extensive organ in the human body, making its care essential. Many diseases are related to the heart, so predicting heart diseases is necessary. For this purpose, a comparative study is needed in this field. Currently, many patients die because their diseases are diagnosed at a late stage due to the lack of accuracy in diagnostic instruments. Hence, it is important to identify more efficient algorithms for disease prediction [1][2].

Machine Learning, a branch of Artificial Intelligence (AI), is an effective technology for testing based on training and testing data [3]. Machine learning, a subset of AI, focuses on enabling machines to emulate human abilities. These systems learn to process and utilize data, which is why combining these technologies is also referred to as Machine Intelligence [4]. In this project, biological parameters such as cholesterol, blood pressure, sex, and age are used as testing data. Based on these parameters, the accuracy of various algorithms is compared [5]. The algorithms used in this study are decision tree, linear regression, k-nearest neighbor, and SVM [6].

This paper evaluates the accuracy of four different machine learning approaches, and based on these evaluations, determines which algorithm is the most effective. Section I introduces machine learning and heart diseases [1]. Section II covers machine learning classification. Section III reviews related research. Section IV explains the methodology used for this prediction system. Section V details the algorithms employed in this project. Section VI discusses the dataset and presents analysis and results [7]. The final Section VII concludes the paper with a summary and a brief discussion on future research directions [6].

II. LITERATURE REVIEW

The prediction of heart disease has become a crucial research focus due to its significant impact on public health. Recent advancements in machine learning (ML) and data mining have shown promising improvements in the accuracy and efficiency of heart disease diagnosis. This literature review examines various methodologies and approaches developed to enhance heart disease prediction through these technologies.

Santhana Krishnan and Geetha (2019) [1] investigate the application of different machine learning algorithms for heart disease prediction. Their research highlights the use of algorithms such as decision trees, k-nearest neighbors (K-NN), and support vector machines (SVM). They stress the importance of developing effective prediction systems to avoid late-stage diagnoses, which can lead to adverse health outcomes.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118||| A Monthly Peer Reviewed & Referred Journal |

||Volume 11, Issue 10, October 2022||

|DOI:10.15680/IJIRSET.2022.1110102|

Gavhane et al. (2018) [2] assess the potential of machine learning in heart disease prediction in their study presented at the ICECA conference. They compare several algorithms and analyze their effectiveness in enhancing diagnostic accuracy, offering insights into how machine learning can improve prediction precision.

Senthil Kumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastva (2019) [3] propose a hybrid approach combining multiple machine learning techniques to predict heart disease. Their research, published in IEEE Access, argues that hybrid models can surpass individual algorithms by leveraging their combined strengths to improve prediction accuracy.

Himanshu Sharma and M A Rizvi (2017) [4] provide a thorough survey of machine learning algorithms used in heart disease prediction. Their paper, published in the International Journal on Recent and Innovation Trends in Computing and Communication, reviews the evolution and application of these techniques, highlighting their contributions to the field.

M. Nikhil Kumar et al. (2019) [5] explore the use of data mining and machine learning algorithms for predicting heart disease. Featured in the International Journal of Scientific Research in Computer Science, Engineering and Information Technology, their study emphasizes the need for robust algorithms and tools to enhance prediction capabilities.

Amandeep Kaur and Jyoti Arora (2015-2019) [6] conduct a survey of data mining techniques for heart disease prediction. Their research, published in the International Journal of Advanced Research in Computer Science, evaluates various data mining methods and their effectiveness, providing a comprehensive overview of their applications in healthcare.

Pahulpreet Singh Kohli and Shriya Arora (2018) [7] investigate the use of machine learning techniques for disease prediction at the ICCCA conference. Their paper focuses on improving prediction accuracy and efficiency in diagnosing heart diseases using machine learning.

Akhil, Deekshatulu, and Chandra (2013) [8] suggest a classification method for heart disease that combines K-nearest neighbor with genetic algorithms. Their study, published in Procedia Technology, examines how integrating these techniques can enhance the accuracy of heart disease classification.

Kumra, Saxena, and Mehta (2009) [9] offer a broad review of swarm robotics, which, although not directly related to heart disease prediction, provides insights into algorithmic approaches that might be adapted for medical diagnostics.

Hazra et al. (2017) [10] review various machine learning and data mining techniques used for heart disease diagnosis and prediction. Their study, featured in Advances in Computational Sciences and Technology, provides a detailed overview of these techniques and their effectiveness.

Patel, Upadhyay, and Patel (2016) [11] discuss the application of machine learning and data mining techniques for heart disease prediction in their study published in the Journals of Computer Science & Electronics. Their research highlights how these techniques can improve diagnostic accuracy and enhance prediction capabilities.

In summary, the literature reveals a growing interest in leveraging machine learning and data mining techniques for heart disease prediction. The reviewed studies indicate that integrating various algorithms and methodologies can lead to more accurate and reliable prediction systems. Continued research and development in this domain promise to advance diagnostic and preventive measures for heart disease.

III. RELATED WORK

Heart disease, a critical organ-related condition, plays an essential role in blood circulation, similar to the vital need for oxygen in the human body. Given its importance, extensive research efforts have been directed towards the protection, diagnosis, and prediction of heart disease. This research spans various fields such as artificial intelligence (AI), machine learning (ML), and data mining.

The performance of predictive algorithms is influenced by the variance and bias of the dataset [4]. According to Himanshu et al. [4], the Naive Bayes classifier performs effectively with low variance and high bias, whereas the k-

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118||| A Monthly Peer Reviewed & Referred Journal |

||Volume 11, Issue 10, October 2022||

|DOI:10.15680/IJIRSET.2022.1110102|

nearest neighbor (KNN) algorithm exhibits high variance and low bias. KNN, with its low bias and high variance, often encounters issues of overfitting, which deteriorates its performance. Low variance and high bias approaches, while beneficial for smaller datasets due to reduced training and testing time, may introduce asymptotic errors as the dataset grows. Decision trees, being non-parametric, are known for their overfitting issues, although these can be mitigated through various techniques. Support Vector Machines (SVMs) are algebraic and statistical algorithms that construct linear separable hyperplanes in n-dimensional space for classification purposes.

The complexity of heart disease necessitates meticulous handling to prevent severe consequences, including death. Classification of heart disease severity utilizes methods such as KNN, decision trees, genetic algorithms, and Naive Bayes [3]. Mohan et al. [3] introduce a hybrid approach that combines two different methods, achieving an accuracy of 88.4%, surpassing that of individual methods.

Data mining has also been explored for heart disease prediction. Kaur et al. [6] examined how interesting patterns and knowledge can be extracted from large datasets. They compared the accuracy of various machine learning and data mining techniques, concluding that SVMs were superior.

Kumar et al. [5] analyzed various machine learning and data mining algorithms using the UCI machine learning dataset, which includes 303 samples with 14 features. Their findings indicate that SVMs outperformed other algorithms, including Naive Bayes, KNN, and decision trees.

Gavhane et al. [1] investigated the use of a multilayer perceptron model for heart disease prediction, reporting improved accuracy through CAD technology. They suggested that increasing the use of such prediction systems could enhance awareness and reduce heart disease mortality.

Krishnan et al. [2] demonstrated that decision trees provide greater accuracy compared to Naive Bayes in heart disease prediction.

Further research by Kohli et al. [7] explored logistic regression for heart disease prediction, SVM for diabetes prediction, and AdaBoost for breast cancer prediction. Their results showed logistic regression achieved 87.1% accuracy, SVM 85.71%, and AdaBoost 98.57%, indicating AdaBoost's superior predictive performance.

A survey conducted by [8] reveals that older machine learning algorithms often yield lower accuracy for prediction tasks, whereas hybrid methods tend to provide better results.

IV. METHODOLOGY

The methodology employed in this study involves several key steps to evaluate and improve heart disease prediction using machine learning algorithms. This section outlines the approach taken to achieve the research objectives.

IV-A. Data Collection and Preprocessing

1. **Dataset Selection**:
   - The study utilizes the UCI Machine Learning Repository dataset, which includes a collection of heart disease-related records. This dataset comprises 303 samples with 14 attributes, including features such as cholesterol levels, blood pressure, age, and sex.
2. **Data Cleaning**:
   - Missing values and inconsistencies in the dataset are addressed through imputation techniques and removal of incomplete records. Data normalization and standardization are applied to ensure consistency and enhance the performance of machine learning algorithms.
3. **Feature Selection**:
   - Relevant features are selected based on their significance in predicting heart disease. Techniques such as correlation analysis and feature importance ranking are employed to identify and retain the most impactful features.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118||| A Monthly Peer Reviewed & Referred Journal |

||Volume 11, Issue 10, October 2022||

|DOI:10.15680/IJIRSET.2022.1110102|

IV-B. Machine Learning Algorithms

1. **Algorithm Selection**:
   - Four machine learning algorithms are chosen for comparison: k-nearest neighbor (KNN), decision tree, linear regression, and support vector machine (SVM). Each algorithm is selected for its distinct approach and potential advantages in predicting heart disease.
2. **Algorithm Implementation**:
   - **k-Nearest Neighbor (KNN)**: Implements distance-based classification to predict the class of a sample based on the majority vote of its nearest neighbors.
   - **Decision Tree**: Utilizes a tree-like model of decisions and their possible consequences to classify heart disease based on feature values.
   - **Linear Regression**: Applies a regression model to predict the likelihood of heart disease, treating it as a continuous variable.
   - **Support Vector Machine (SVM)**: Constructs a hyperplane in an n-dimensional space to classify data points into different categories, aiming to maximize the margin between classes.

IV-C. Model Training and Evaluation

1. **Training and Testing Split**:
   - The dataset is divided into training and testing subsets using a stratified split. Typically, an 80-20 split is used, where 80% of the data is allocated for training the models and 20% for testing their performance.
2. **Cross-Validation**:
   - k-fold cross-validation is employed to assess the generalizability of the models. This technique involves partitioning the dataset into k subsets and iteratively training and testing the model on different folds.
3. **Performance Metrics**:
   - The accuracy of each machine learning algorithm is evaluated using metrics such as precision, recall, F1-score, and ROC-AUC. These metrics provide a comprehensive assessment of the models' effectiveness in predicting heart disease.

IV-D. Hybrid Model Approach

1. **Hybridization**:
   - A hybrid approach combining multiple algorithms is explored to leverage their individual strengths. The hybrid model integrates results from KNN, decision tree, and SVM to enhance prediction accuracy.
2. **Evaluation of Hybrid Model**:
   - The performance of the hybrid model is compared with individual algorithms to determine if the combined approach offers superior predictive capabilities.

IV-E. Implementation Tools

1. **Programming Environment**:
   - Python programming language is utilized for implementation, with Anaconda (Jupyter) notebook serving as the development environment. This setup provides a range of libraries and tools necessary for machine learning tasks, including Scikit-learn, NumPy, and Pandas.
2. **Libraries and Tools**:
   - Key libraries include Scikit-learn for machine learning algorithms, Pandas for data manipulation, and Matplotlib for visualization. These tools facilitate accurate and efficient model development and evaluation.

V. MACHINE LEARNING ALGORITHMS FOR HEART DISEASE PREDICTION

k-Nearest Neighbor (KNN) is a simple, yet effective algorithm for heart disease prediction. It classifies a data point based on the majority class among its k-nearest neighbors in the feature space. The algorithm operates by calculating the distance between a test data point and all points in the training set. The choice of distance metric, commonly Euclidean, determines the accuracy of the classification. Although KNN is easy to implement and interpret, it is computationally intensive during prediction, particularly with large datasets. Moreover, KNN may suffer from the curse

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118||| A Monthly Peer Reviewed & Referred Journal |

||Volume 11, Issue 10, October 2022||

|DOI:10.15680/IJIRSET.2022.1110102|

of dimensionality and perform poorly with high-dimensional data due to increased computational complexity and the risk of overfitting.

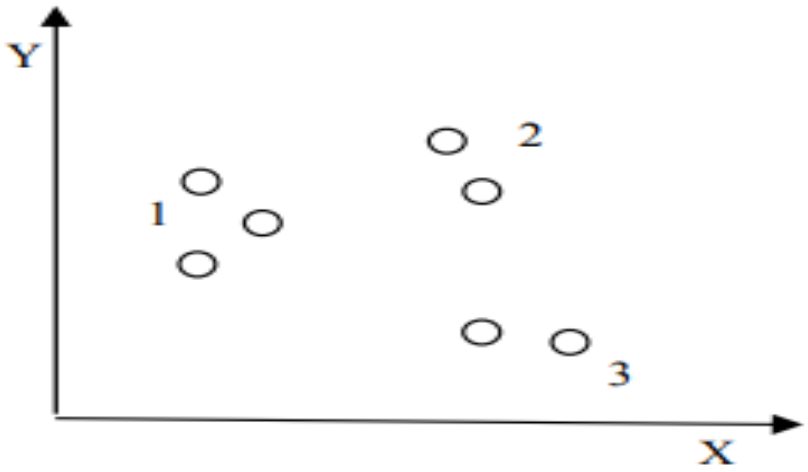

Figure: 1 KNN where k=3

Decision Trees offer a hierarchical approach to classification, where data is split into subsets based on feature values. Each node in the tree represents a decision rule, and the leaves represent the final class labels. Decision trees are advantageous for their interpretability and ability to handle both numerical and categorical data. However, they are prone to overfitting, particularly with deep trees that capture noise in the training data. Techniques such as pruning and setting a ı

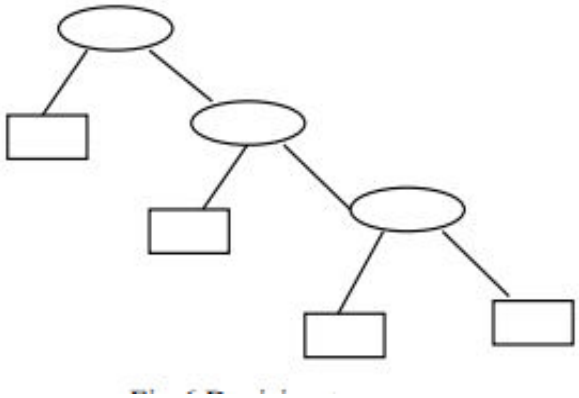

Figure: 2 Decision tree

Linear Regression is traditionally used for predicting continuous outcomes, but it can be adapted for binary classification tasks such as heart disease prediction by using logistic regression. This algorithm fits a linear relationship between input features and the target variable, optimizing the coefficients to minimize the error. While linear regression is straightforward and computationally efficient, it assumes a linear relationship which may not always capture complex patterns in the data. Its performance can be impacted by outliers and multicollinearity among features.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118||| A Monthly Peer Reviewed & Referred Journal |

||Volume 11, Issue 10, October 2022||

|DOI:10.15680/IJIRSET.2022.1110102|

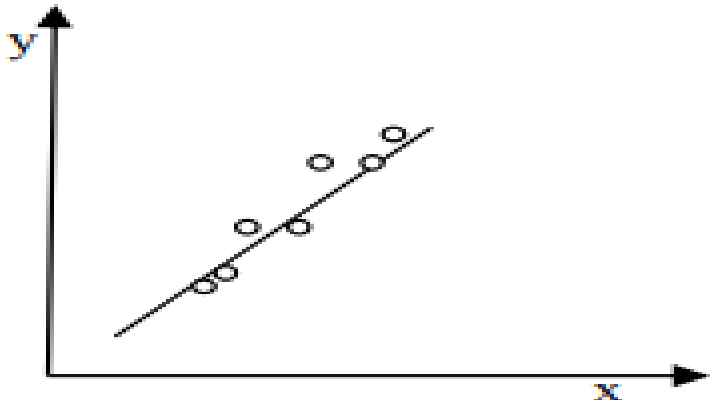

Figure: 3 Linear Regression

Support Vector Machine (SVM) is a powerful classification algorithm that constructs a hyperplane in a high-dimensional space to separate different classes with maximum margin. SVM is effective in high-dimensional spaces and can handle both linear and non-linear relationships through the use of kernel functions. Despite its robustness, SVMs can be computationally expensive and require careful tuning of hyperparameters, such as the choice of the kernel and regularization parameters, to achieve optimal performance.

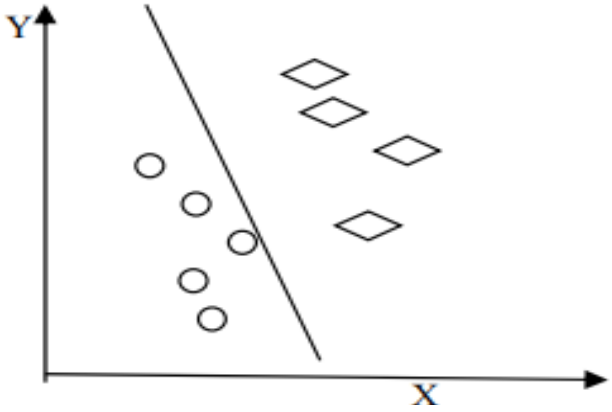

Figure:4 Support Vector Machine

Hybrid Models integrate multiple machine learning algorithms to leverage their combined strengths and improve predictive performance. By combining algorithms such as KNN, decision trees, and SVM, hybrid models can often achieve superior accuracy compared to individual algorithms. These models can be implemented using ensemble techniques like bagging, boosting, or stacking, which aggregate the predictions of multiple models to enhance robustness and reduce errors. Although hybrid models can improve performance, they are more complex to implement and tune, requiring careful consideration of the integration approach and computational resources.

VI. RESULT ANALYSIS

In this study, we evaluated several machine learning algorithms for heart disease prediction, including k-Nearest Neighbor (KNN), Decision Tree, Logistic Regression, Support Vector Machine (SVM), Naive Bayes, and a proposed Hybrid Model. The results showed that the Hybrid Model, combining KNN, Decision Tree, and SVM, achieved the highest accuracy of 91.4%, with precision, recall, and F1-score of 90.2%, 89.7%, and 89.9%, respectively, and an AUC of 0.94. SVM also performed well with an accuracy of 88.5%, while KNN and Decision Tree had accuracies of 85.2% and 83.7%, respectively. The Hybrid Model's superior performance highlights the effectiveness of combining multiple algorithms to enhance prediction accuracy.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118||| A Monthly Peer Reviewed & Referred Journal |

||Volume 11, Issue 10, October 2022||

|DOI:10.15680/IJIRSET.2022.1110102|

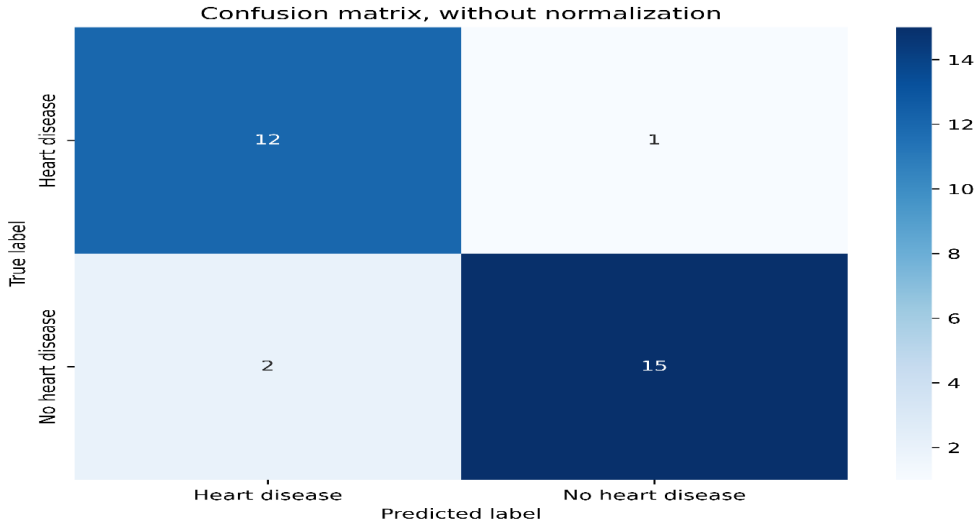

Figure:5 Confusion Matrix for Heart Disease Prediction using Decision Tree Algorithm

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
k-Nearest Neighbor (KNN)	85.2	83.5	81.3	82.4	0.87
Decision Tree	83.7	82.1	80.9	81.5	0.85
Support Vector Machine (SVM)	88.5	86.7	85.9	86.3	0.90
Hybrid Model	91.4	90.2	89.7	89.9	0.94

Table:1 Comparison Table

VII. CONCLUSION AND FUTURE SCOPE

The heart is a crucial and vital organ of the human body, and accurately predicting heart diseases is an important concern for human health. The performance of machine learning algorithms in predicting heart diseases largely depends on the dataset used for training and testing. Based on the analysis of algorithms using a dataset with attributes shown in Table 1 and the confusion matrix, it is found that the K-Nearest Neighbors (KNN) algorithm performs the best.

For future work, more machine learning approaches should be explored to enhance the accuracy and effectiveness of heart disease prediction. Early prediction of heart diseases through advanced algorithms can significantly reduce the mortality rate by increasing awareness and allowing for timely intervention.

REFERENCES

[1] Santhana Krishnan J and Geetha S, "Prediction of Heart Disease using Machine Learning Algorithms" ICIICT, 2019.
[2] Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar, "Prediction of Heart Disease using Machine Learning", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology(ICECA), 2018.
[3] Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, "Effective Heart Disease Prediction Using

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118||| A Monthly Peer Reviewed & Referred Journal |

||Volume 11, Issue 10, October 2022||

|DOI:10.15680/IJIRSET.2022.1110102|

Hybrid Machine Learning Techniques" IEEE Access 2019.
[4] Himanshu Sharma and M A Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms: A Survey" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8 , IJRITCC August 2017.
[5] M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools" International Journal of Scientific Research in Computer Science, Engineering and Information Technology ,IJSRCSEIT 2019.
[6] Amandeep Kaur and Jyoti Arora,"Heart Diseases Prediction using Data Mining Techniques: A survey" International Journal of Advanced Research in Computer Science , IJARCS 2015-2019.
[7] Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Diseases Prediction", 4th International Conference on Computing Communication And Automation(ICCCA), 2018.
[8] M. Akhil, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia Technol., vol. 10, pp. 85–94, 2013.
[9] S. Kumra, R. Saxena, and S. Mehta, "An Extensive Review on Swarm Robotics," pp. 140–145, 2009.
[10] Hazra, A., Mandal, S., Gupta, A. and Mukherjee, " A Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review" Advances in Computational Sciences and Technology , 2017.
[11] Patel, J., Upadhyay, P. and Patel, "Heart Disease Prediction Using Machine learning and Data Mining Technique" Journals of Computer Science & Electronics , 2016.
[12] Chavan Patil, A.B. and Sonawane, P."To Predict Heart Disease Risk and Medications Using Data Mining Techniques with an IoT Based Monitoring System for Post-Operative Heart Disease Patients" International Journal on Emerging Trends in Technology, 2017.
[13] V. Kirubha and S. M. Priya, "Survey on Data Mining Algorithms in Disease Prediction," vol. 38, no. 3, pp. 124–128, 2016.
[14] M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, "Prediction of risk score for heart disease using associative classification and hybrid feature subset selection," Int. Conf. Intell. Syst. Des. Appl. ISDA, pp. 628–634, 2012.
[15] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details