iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** VI **Month of publication:** June 2024

DOI: https://doi.org/10.22214/ijraset.2024.63088

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue VI June 2024- Available at www.ijraset.com

MediAssist: Self Diagnosing Assistance

K. Vikram Reddy[1], V. Manvitha[2], Harshita[3], Kotha Reethika[4]
[1]*Assistant Professor,* [2, 3, 4]*Students Department of Information Technology, Matrusri Engineering College*

Abstract: ***Despite the availability of advanced technology and easy access to information, many people still rely on traditional methods of seeking medical treatment, such as visiting hospitals and consulting doctors for even minor symptoms. However, this approach can be time-consuming and inefficient, as patients with minor illnesses can take up valuable resources that could be better used to treat more serious cases. As a result, this research proposes a new approach to disease prediction using machine learning. The goal is to develop a predictive model that can accurately identify potential diseases based on a patient's health parameters such as pulse rate, Cholesterol, Blood Pressure etc. Medical records to train and test the model, which demonstrated high accuracy in predicting diseases. This paper uses the machine learning algorithms for predicting chronic disease based on user-provided input data. Its primary goal is to enhance early disease detection and prompt medical intervention using personalized predictions. By training models on relevant datasets, accurate predictions are achieved across various diseases, expanding our understanding and predictive capabilities in healthcare The results of this study suggest that the proposed model could be a useful tool for early diagnosis and treatment of diseases, with the potential to improve healthcare outcomes. Through data-driven insights, this tool aids in reducing medical costs by enabling early intervention and proactive management of medical conditions. Overall, this study highlights the potential of model in healthcare for disease pre-diction and underscores its role in advancing personalized medicine and improving clinical decision-making***

Keywords**:** ***Diseases Prediction, Machine Learning Algorithms-Naive Bayes (NB) classifier, Decision Tree (DT) classifier, K-Nearest Neighbours (KNN) algorithm, XG Boost algorithm.***

I. INTRODUCTION

In recent times, Over the last few years, there has been a significant increase in both the global patient population and the occurrence of various diseases, putting a burden on healthcare systems around the world. Unfortunately, this growth in demand has resulted in higher healthcare costs, driving increasing the cost of medical services in many countries. A doctor's visit is essential to begin therapy for the majority of diseases. However, technological developments and the availability of massive amounts of data give an opportunity to totally revolutionize the diagnostic practice. In today's digital era, data is incredibly valuable, especially in fields like healthcare where it includes patient information. Traditionally, most models focused on predicting one disease at a time, like diabetes, heart disease. There hasn't been a unique system capable of predicting multiple diseases simultaneously. The paper is proposing a new system which will predict multiple diseases such as diabetes, heart disease, breast cancer, kidney and liver disease. This model will also helpful in predicting various disease later on using machine learning algorithms to make these predictions based on user's input. A method for serializing and deserializing Python object performs as a middle step between getting test results and consulting with a doctor. Plus, not everyone will need to consult a doctor right away because user can get valuable insights from the system. Compared to existing systems, our model is more flexible and efficient. This level of precision demonstrates the applicability and promise of our technology for improving medical diagnostics. To improve usability and accessibility, we developed an interactive interface that allows for fluid interaction with the system. This user-friendly design makes it straightforward for patients and healthcare professionals to navigate and submit relevant symptom information, greatly accelerating the diagnosis process. Furthermore, we worked hard to clearly depict and communicate the results of our effort and study the insights derived from this model contribute to early disease prevention strategies, fostering a proactive approach to healthcare that goes beyond reactive responses to symptoms. Overall, our system addresses the limitations of existing models by offering flexibility, efficiency, and is able to detect various disease with just one set of input data. This simplifies the process for organizations analysing patient health reports.

II. LITERATURE REVIEW

There have been numerous studies related to predicting the diseases using different machine learning techniques and algorithms which can be used by medical institutions. This paper reviews some of those studies done in research papers using the techniques and results used by them. Reviews are given below:

©IJRASET: All Rights are Reserved

A. Reviews

MIN CHEN et al, [1] proposed a disease prediction system in his paper where he used machine learning algorithms. In the prediction of disease, he used techniques like CNN-UDRP algorithm, CNN-MDRP algorithm, Naive Bayes, K-Nearest Neighbour, and Decision Tree. This proposed system had an accuracy of 94.8%.

Lambodar Jena et al, [2] focused on risk prediction for chronic diseases by taking advantage of distributed machine learning classifiers and used techniques like Naive Bayes and Multilayer Perceptron. This paper tries to predict Chronic-Kidney-Disease and the accuracy of Naïve Bayes and Multilayer Perceptron is 95% and 99.7% respectively.

Dhomse Kanchan B. et al, [3] studied special disease prediction utilizing principal component analysis using machine learning algorithms involving techniques like Naive Bayes classification, Decision Tree, and Support Vector Machine. The accuracy of this system is 34.89% for Diabetes and 53% for heart disease.

Pahulpreet Singh Kohli et al, [4] suggested disease prediction by using applications and methods of machine learning and used techniques like Logistic Regression, Decision Tree, Support Vector Machine, Random Forest and Adaptive Boosting. This paper focuses on predicting Heart disease, Breast cancer, and Diabetes. The highest accuracies are obtained using Logistic Regression that is 95.71% for Breast cancer, 84.42% for Diabetes, and 87.12% for heart disease.

Deeraj Shetty et al, [5] studied the uses of data mining for diabetes disease prediction by using Naïve Bayes and KNN algorithms. This system predicts diabetes and accuracy obtained by KNN are better than Naïve Bayes.

Senthilkumar Mohan et al, [6] focused on hybrid techniques in machine learning that can be used for effectively predicting heart disease and used algorithms like Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, Neural Network and KNN. The accuracy of this system is 88.47%. Avi Agarwal, Milan Sai [7] focused on heart disease prediction using machine learning by using KNN with an accuracy of 70% Lama A. Alqahtani [8] proposed automated prediction of heart disease using naïve bayes and decision tree gave An accuracy around 60%.

B. A Comparative Study using Various algoriThms in the Literature Review

Table 1: Comparative study using various algorithms in the literature review

Year	Author	Purpose	Techniques/Algorithms Used	Accuracy
2017	MIN CHEN et al, [1]	Proposed a disease prediction system in his paper where he used machine learning algorithms.	CNN-UDRP algorithm, CNN-MDRP algorithm, Naive Bayes, K-Nearest Neighbour, Decision Tree	94.8%
2017	Lambo dar Jena et al, [2]	Focused on risk prediction for chronic kidney diseases by taking advantage of distributed machine learning algorithms.	Naive Bayes	95%
			Multilayer Perceptron	99.7%
2016	Dhom se Kanch an B. et al, [3]	Studied special disease prediction utilizing principal component analysis using machine learning algorithms.	Naive Bayes classification, Decision Tree and Support Vector Machine	Diabetes Disease: 34.89% Heart Disease: 53%

2018	Pahulpreet Singh Kohli et al, [4]	Suggested disease prediction by using applications and methods of machine learning	Logistic Regression	Breast Cancer: 95.71 % Diabetes:84.42 % Heart Disease: 87.12 %
			Decision Tree	Breast Cancer: 94.29 % Diabetes: 74.03 % Heart Disease: 70.97 %
			Random Forest	Breast Cancer: 97.14 % Diabetes: 81.82 % Heart Disease: 77.42 %
			Support Vector Machine	Breast Cancer: 97.14 % Diabetes: 85.71 % Heart Disease: 83.87 %
2017	Deeraj Shetty et al, [5]	Studied the uses of data mining for diabetes disease prediction	Naïve Bayes and KNN	KNN gives better accuracy compared to Naive Bayes.
2019	Senthilkumar Mohan et al, [6]	Focused on hybrid techniques in machine learning that can be used for effectively predicting heart disease	Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, Neural Network and KNN	88.4%
2020	Avi Agarwal, Milan Sai [7]	Focused on heart disease prediction using machine learning	KNN	70%
2020	Lama A . Alqahtani [8]	Automated Prediction of Heart Disease	Naïve Bayes, Decision Tree	60%

III. PROPOSED METHODOLOGY

A methodology tailored for multiple disease prediction using specific machine learning algorithms like K-Nearest Neighbours (KNN), XG Boost, and Decision Trees:

A. Data Collection

Collect comprehensive datasets containing relevant features for each disease. Ensure diversity and quality in the data.

B. Data Preprocessing

Clean and preprocess the data by handling missing values, normalizing, and encoding categorical variables. Split the dataset into training and testing sets.

C. Feature Selection

Identify important features using methods like correlation analysis or feature importance from tree-based models, especially relevant for Decision Trees and XG Boost.

D. Algorithm Selection

Types Of Classification Algorithms

Classification algorithms are categorized as linear and nonlinear algorithms:

1) Linear algorithms

Linear algorithms are ones that can be described mathematically by a linear equation. This indicates a linear relationship between the attributes and the target parameter. The link between height and weight, for example, is linear.

Algorithms that are linear Types:

a) *Logistic Regression:* A statistical technique called logistic regression is used to address binary classification issues. It forecasts the likelihood of an instance belonging to a specific class.

b) *Support Vector Machines (SVM):* SVM is an adaptable technique that may be used for classification as well as regression. It finds the best hyperplane for differentiating several classes or predicts the value of a continuous targeted parameter.

2) Non-Linear Algorithms

Non-linear algorithms, on the other hand, cannot be represented by a linear equation. This means that A non-linear relationship can be seen in the association between the characteristics and the goal variable. For example, the relationship between blood pressure and age is non-linear.

Non-linear algorithm types

a) *Decision Tree Classification:* It is regarded as a highly successful and adaptable classification tool. It is utilized in picture categorization and pattern recognition. Due to its exceptional adaptability, this approach is utilized for classification in highly intricate issues. Furthermore, it demonstrates proficiency in handling challenges involving multiple dimensions. The structure consists of three components, namely the root, nodes, leaves.

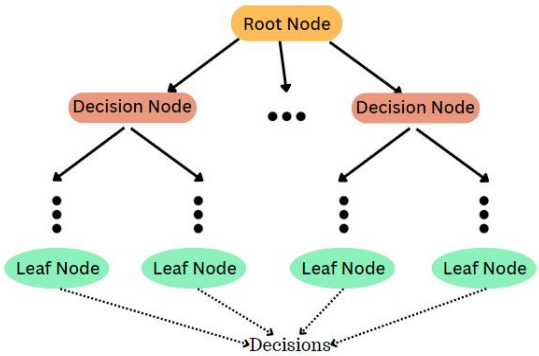


Fig:1 Decision Tree

b) *Nave Bayes:* A nonlinear framework based on the Bayes theorem. The Bayes theorem is an algebraic equation that calculates the likelihood of an event occurring given the likelihood of other events occurring. The likelihood of each characteristic existing in a given class is assumed to be independent of the probability of the existence of the other qualities in the class by Naive Bayes. This assumption is known as the naive Bayes assumption.

c) *Random Forests:* It is a method of supervised learning that can be used to solve regression and classification problems. These algorithm's four crucial steps are as follows:

- It chooses random samples of data from the dataset.
- For each sample dataset chosen, it generates decision trees.
- All potential results will now be tallied and decided upon.
- The final prediction will be determined and provided as the classification outcome.

d) *K-Nearest Neighbor (KNN):* This simple yet successful classification and regression approach uses K-Nearest Neighbor (KNN). It forecasts an instance's class or value using the majority vote or average of its neighboring examples in the feature space.

e) *Linear Vs Non-Linear Algorithm for Diseases Prediction*: Why non-linear algorithm prefers over linear ones for disease prediction? Non-linear algorithms are preferred over linear algorithms for disease prediction via because non- linear algorithms can capture complex relationships and interactions among numerous components that contribute to disease development, allowing for more accurate and nuanced predictions. Linear algorithms, on the other hand, presume a linear relationship between predictors and outcomes, which may be insufficient to represent the complexity of diseases and their risk factors.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

E. Model-Specific Preprocessing

Perform any additional preprocessing specific to each algorithm. For KNN, scaling features might be crucial; for XG Boost, no additional preprocessing might be needed.

F. Model Training

Train each selected model using the training dataset. Fine-tune hyperparameters for optimal performance. For KNN, determine the optimal number of neighbors.

G. Cross-Validation

Implement cross-validation to assess the generalization performance of each model. Adjust parameters accordingly to avoid overfitting.

H. Evaluation Metrics

Evaluate each model using appropriate metrics such as accuracy, precision, recall, and F1-score. Tailor metrics to the specific requirements of disease prediction.

I. Deployment

Deploy the ensemble model or the best-performing individual models in a healthcare setting, ensuring compliance with ethical and privacy standards.

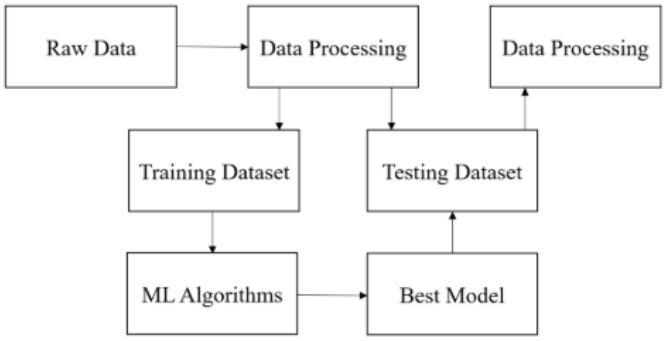


Fig:2 Architecture Model

IV. IMPLEMENTATION

In this project, we utilized several commonly used libraries and environments for database analysis and model building. The project leveraged the following libraries:

1) *Flask:* Flask is a lightweight Python web framework that is easy to set up and use. It supports modular development with Blueprints and integrates seamlessly with various extensions for added functionality. Ideal for small to medium applications, Flask offers flexibility, simplicity, and robust support for templating and routing.
2) *NumPy:* NumPy is a well-known scientific computing library written in Python. It provides you with sophisticated tools for working with multidimensional arrays. NumPy's primary goal is to efficiently handle multidimensional homogenous arrays. It can generate, manipulate, and process arrays with total, mean, standard deviation, max, min, and other functions. The array processing features of NumPy make it ideal for data handling in our project.
3) *Pandas:* Pandas is a popular Python data analysis toolkit that provides enhanced performance through its backend code, which is implemented in C or Python. Pandas introduces data structures called Series and Data Frames, which are used to store and manipulate data. Series represents a one-dimensional array, while Data Frames represent a two-dimensional data structure. These data structures provide efficient storage and enable various operations on the data. Data Frames make it easier to work with attributes and results. In our project, Pandas played a crucial role in data analysis and management. In our project, we used Pandas Data Frames extensively for handling datasets, data manipulation, and preprocessing tasks.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

4) *Scikit-learn (sklearn):* scikit-learn is an open-source Python toolkit that provides a wide range of machine learning algorithms, data preprocessing techniques, cross- validation methods, and visualization tools. It includes support for classification, regression, and clustering tasks, offering algorithms such as support vector machines, random forest classifiers, decision trees, Gaussian naive Bayes, and K-nearest neighbors. In our study, we utilized the built-in classification approaches of scikit-learn, such as decision trees, random forest classifiers, K-nearest neighbors, and as well as naïve Bayes. To analyses the efficacy of our models, we employed scikit-learn's verification cross-validation capabilities and visualization features such as classification reports, confusion matrices, and accuracy scores.
5) *Jupyter Notebook:* Alongside the mentioned libraries, we employed We'll be using Jupyter Notebook as our core platform for development. Jupyter Notebook is open-source software that is free that allows users to create and share notebooks with realtime code, equations, graphics, and text. It supports a wide range of computer languages, including Python, and offers a collaborative environment for activities like data analysis and model construction. We were able to boost our efficiency and collaboration by using Jupyter Notebook.

We used Jupyter Notebook's capabilities to write and execute code, analyze and visualize data, and document our analysis process. Its notebook style allowed us to integrate code cells, markdown cells for written explanations, and visualizations in a single document, making it easier to show and share our work.

V. RESULT

This program provides an automatic diagnostic method based on user input to save time and minimize expenses connected with the first diagnostic process. The program accepts data from the user and properly predicts diseases as output. Analysis of the model accuracies listed below:

Model	Accuracy
Diabetes	92.54%
Heart Disease	98.70%
Breast Cancer	97.66%
Kidney Disease	99.16%
Liver Disease	71.18%

VI. CONCLUSION

In conclusion, employing machine learning algorithms such as Random Forest, KNN, XG Boost, and decision trees for multiple disease prediction has shown promising results. These models contribute to accurate predictions by leveraging various features and patterns in medical data. However, the effectiveness of each algorithm may vary based on the specific characteristics of the dataset and the nature of the diseases under consideration. Integration of diverse algorithms in an ensemble approach could further enhance predictive performance, providing a robust framework for disease prediction and facilitating personalized healthcare solutions. Continuous refinement and validation of these models with updated datasets will be crucial for ensuring their reliability and applicability in real world medical scenarios. However, challenges like interpretability and ethical considerations need to be addressed for seamless integration into the healthcare system. Continued research, collaboration between data scientists and medical professionals, and adherence to privacy regulations will be essential to harness the full potential of machine learning in disease prediction, ushering in an era of more precise and personalized healthcare interventions.

VII. FUTURE WORK

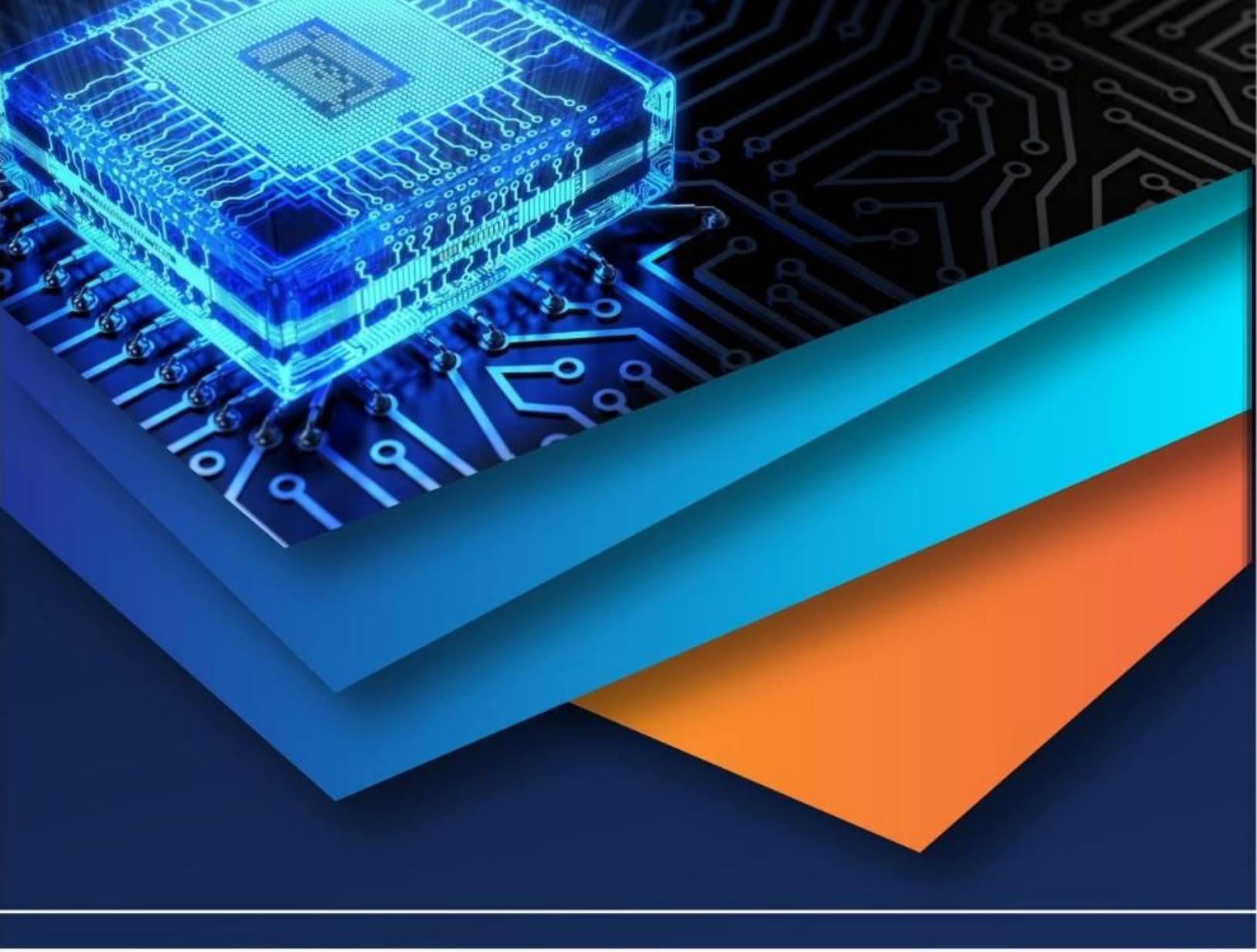
Various algorithms, such as decision trees, random forests, naive bayes, and deep learning models, have been addressed and effectively applied to a wide range of diseases throughout the research. The system also has an intuitive user interface and a variety of visual representations of collected data and findings. In the future, it is important to continue exploring and comparing alternative classification methods in order to enhance disease prediction models. Furthermore, the system could benefit from more comprehensive and diverse datasets to improve accuracy and generalizability. It would also be beneficial to perform comprehensive evaluations and validations of the system using real-world patient data. Furthermore, continued study and collaboration with medical professionals can help refine and expand the system's capabilities, making it an even more trustworthy disease prediction tool.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities" IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.
[2] Sayali Ambekar, Rashmi Phalnikar, "Disease Risk Prediction by Using Convolutional Neural Network" IEEE, 978-1-5386-5257-2/18, 2018.
[3] Naganna Chetty, Kunwar Singh Vaisla and Nagamma Patil, "An Improved Method for Disease Prediction using Fuzzy Approach" IEEE, DOI 10.1109/ICACCE.2015.67, pp. 569-572, 2015
[4] Dhiraj Dahiwade, Gajanan Patle and Ektaa Meshram, "Designing Disease Prediction Model Using Machine Learning Approach" IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4, pp. 1211-1215, 2019.
[5] Lambodar Jena and Ramakrushna Swain, "Chronic Disease Risk Prediction using Distributed Machine Learning Classifiers" IEEE, 978-1-5386-2924-6/17, pp. 170-173, 2017.
[6] Dhomse Kanchan B. and Mahale Kishor M., "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis" IEEE, 978-1-5090-0467-6/16, pp. 5-10, 2016.
[7] Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Disease Prediction" IEEE, 978-1-5386-6947-1/18, pp. 1-4, 2018.
[8] Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil, " Diabetes Disease Prediction Using Data Mining" IEEE, 978-1-5090-3294-5/17, 2017.
[9] Rashmi G Saboji and Prem Kumar Ramesh,"A Scalable Solution for Heart Disease Prediction usingClassification Mining Technique" IEEE, 978-1-5386-1887-5/17, pp. 1780-1785, 2017.
[10] RatiShukla, Vikash Yadav, Parashu Ram Pal and Pankaj Pathak, "Machine Learning Techniques for Detecting and Predicting Breast Cancer" IJITEE, ISSN: 2278-3075, Volume-8, pp. 2658-2662, 2019.
[11] Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" IEEE Access, DOI 10.1109/ACCESS.2019.2923707, pp. 81542-81554, 2019.
[12] Anjan Nikhil Repaka, Sai Deepak Ravikanti and Ramya G Franklin,"Design And Implementing Heart Disease Prediction Using Naives Bayesian" IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8, pp. 292-297, 2019.
[13] Aakash Chauhan, Purushottam Sharma, Vikas Deep and Aditya Jain, "Heart Disease Prediction using Evolutionary Rule Learning" CICT 2018.
[14] Aditi Gavhane, Gouthami Kokkula, Isha Pandya and Kailas Devadkar, "Prediction of Heart Disease Using Machine Learning" IEEE Xplore ISBN: 978-1-5386-0965-1, pp. 1275-1278, 2018.
[15] Ankita Dewan and Meghna Sharma, "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification" IEEE, 978-9-3805-4416-8/15, pp. 704-706, 2015.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)