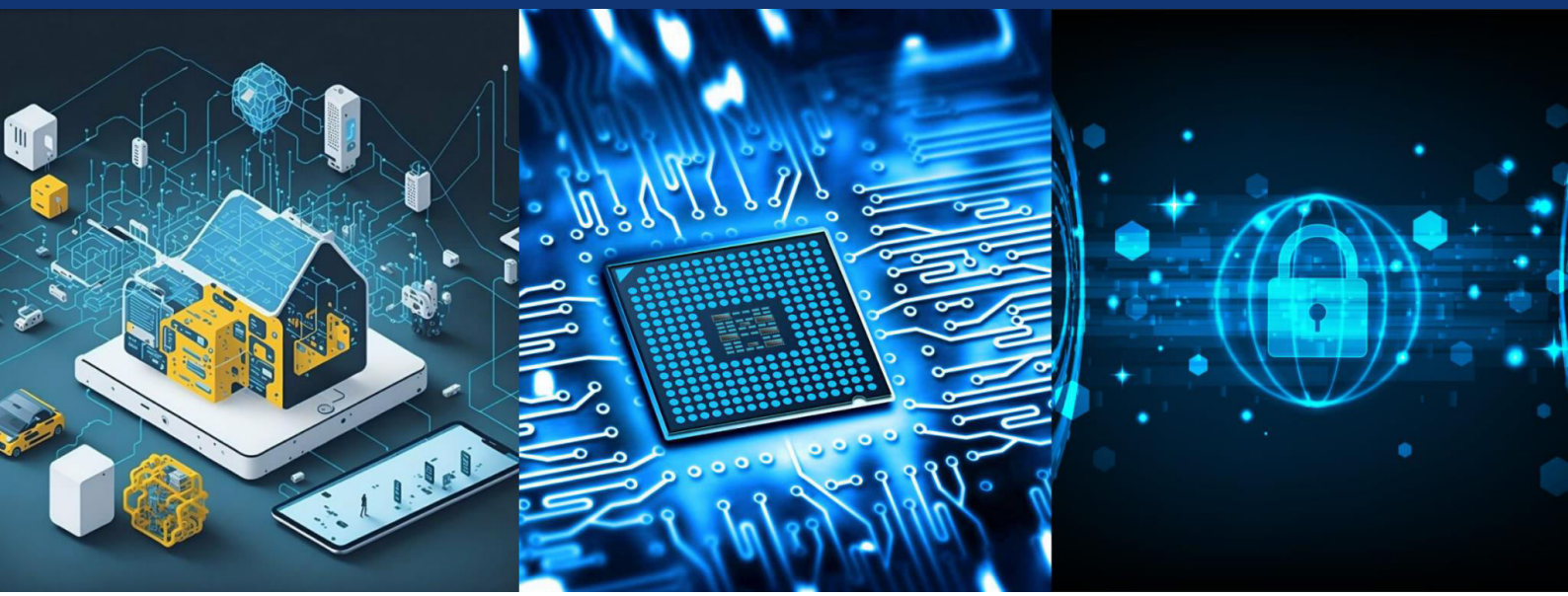
IJIRCCE

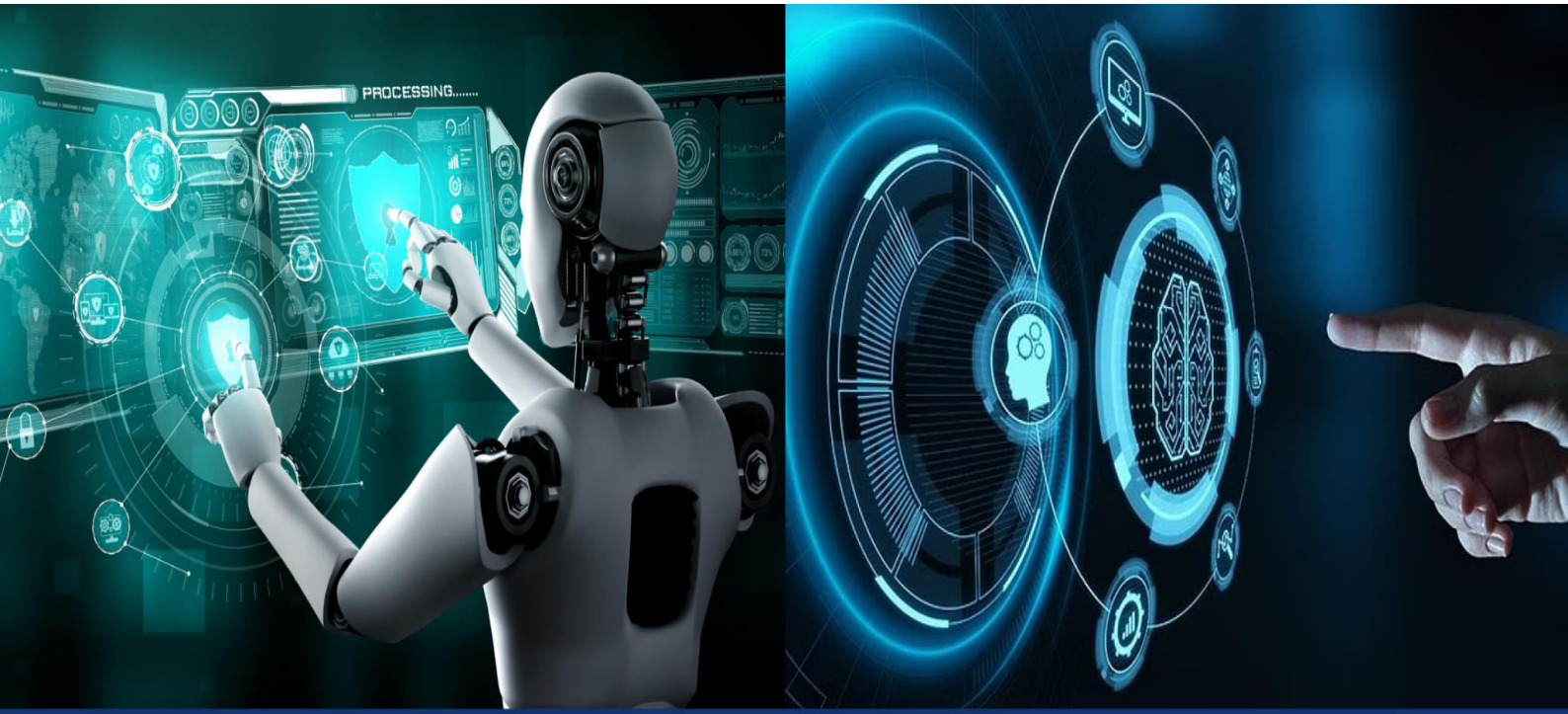
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304142

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Predictive Analytics Model using Machine Learning Techniques for Health Care Diagnosis

D.Jyothi[1], T.Nikhitha[2], P.Srikanth[3], V.Yogesh[4], Dr.V.Sreenivas[5]

U G Students, Dept.of CSE-DS.,SRK Institute of Technology, Enikepadu, Vijayawada, Andhra Pradesh, India[1,2,3,4]

Professor, Dept.of CSE-DS., SRK Institute of Technology, Enikepadu, Vijayawada, Andhra Pradesh, India[5]

ABSTRACT: The Disease Diagnosis Prediction Using Data Analytics project aims to develop an advanced diagnostic system that utilizes machine learning to predict diseases based on patient symptoms, medical history, genetics, and environmental factors. This system supports early detection and personalized treatment, improving patient outcomes by identifying health risks before they become critical. By integrating electronic health records (EHR), medical imaging, genetic data, and lifestyle factors, the model delivers accurate, data-driven predictions to assist healthcare professionals in decision-making. Machine learning algorithms analyze complex datasets to identify patterns associated with disease onset, enabling proactive intervention. The system offers personalized treatment recommendations, enhancing efficiency in health care .Utilizing machine learning techniques such as random forests, support vector machines (SVM), and neural networks, the model automates diagnostics, improves accuracy, reduces costs, and enhances patient care. The integration of machine learning in healthcare enables early diagnosis, risk assessment, and predictive analytics, transforming traditional diagnostics into a more efficient and data-driven approach.

KEYWORDS: Data analytics, disease diagnosis, electronic health records, feature selection, Health care,Medical Imaging, Predictive Modelling.

I. INTRODUCTION

Healthcare diagnostics is a vital aspect of the healthcare system that involves identifying diseases and conditions through various methods, including medical tests, physical examinations, and patient history. Accurate diagnosis is critical in guiding treatment plans, managing patient care, and improving health outcomes. Traditional diagnostic methods, such as manual analysis of symptoms and lab results, can be time-consuming, costly, and prone to human error. As the complexity of diseases increases, healthcare professionals face challenges in quickly identifying diseases and predicting patient outcomes .Machine learning (ML) plays a pivotal role in transforming healthcare diagnostics by enabling computers to analyze complex data patterns and make predictions without explicit programming. In healthcare, machine learning algorithms learn from historical data, improving their ability to diagnose diseases, predict patient outcomes, and assist in decision-making. In this paper ,section1 discusses on Introduction,Section2 on Related Works,Section3 with Back Ground of all algorithms ,Section 4 we explained methodology and our algorithms are Random Forest ,Decision Tree and Logistic Regression,Section5 we compared the results and find out the best algorithm and the last section is concluded with future Works.

II. RELATED WORKS

Machine learning (ML) has become an indispensable tool in healthcare diagnostics and disease prediction. With advancements in computational techniques, data storage, and the proliferation of medical datasets, ML is increasingly used to predict diseases, forecast disease progression, and improve decision-making in healthcare settings. The integration of ML techniques into disease prediction is driven by the need for early diagnosis, timely interventions, and personalized treatment plans.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

s.no	Paper Information	Description	Limitations/Inference
1.	Yogesh Kumar, Apeksha Koul(2023)	This systematic review examined the role of artificial intelligence (AI) in diagnosing diseases such as Alzheimer's, cancer, chronic diseases, heart disease, and tuberculosis	The review examines AI's role in diagnosing various diseases but does not provide a direct comparison between AI-based and traditional diagnostic methods in terms of cost, efficiency, and reliability.
2.	Prof. Kin Young(2023)	This study explored the application of machine learning algorithms in predicting chronic diseases such as diabetes and hypertension	The study focuses on ML algorithms for chronic disease detection but does not account for variations in disease symptoms, data collection inconsistencies, or patient demographics affecting prediction accuracy.
3.	Mohamed Said Ibrahim, Sameh Saber(2023)	This research explored the impact of machine learning and predictive analytics on disease prevention	The study highlights the potential of machine learning in disease prevention but lacks focus on data privacy, model bias, and interpretability. Additionally, challenges in integrating AI with existing healthcare systems and EHRs may hinder real-world implementation.
4.	Mohammed Badawy, Nagy Ramadan(2023)	This survey reviewed 41 papers focusing on machine learning (ML) and deep learning (DL) applications in healthcare prediction across diseases like diabetes, COVID-19, heart, liver, and chronic kidney diseases	The survey reviews multiple papers but does not present a new experimental approach, meaning practical effectiveness and real-world validation of ML/DL models remain uncertain.
5.	K. Purushotam Naidu, V. Lakshmana Rao(2024)	The study proposed a hybrid machine learning model combining multiple algorithms to predict cardiovascular diseases	The study proposes a hybrid model, but it does not specify computational efficiency, model scalability, or potential biases that might arise from integrating multiple algorithms.

6.	Rina S. Patil, Tripti Arjariya, Mohit Gangwar(2023)	This research introduced a hybrid machine learning technique utilizing IoT data for heart disease prediction.	While the research introduces IoT-based hybrid ML techniques, real-world deployment challenges such as data security, transmission delays, and device interoperability are not discussed.
7.	Mihir Walvekar, Shivam Gupta(2023)	The study analyzed the effectiveness of hybrid machine learning algorithms in predicting heart diseases	The study analyzes hybrid ML models for heart disease prediction but may not explore real-time applicability, dataset limitations, or issues with model generalization.

III. BACK GROUND

1.Machine Learning Models

1.1 Random Forest Classifier

The Random Forest classifier is a powerful ensemble learning algorithm used for both classification and regression tasks. It operates by constructing multiple decision trees using a technique called bagging, where each tree is trained on a randomly selected subset of the dataset.

1.2 Decision Tree

A Decision Tree is a supervised learning algorithm used for classification and regression tasks. It splits data based on feature conditions to create a flowchart-like model of decisions. It's easy to interpret but can over fit if not properly pruned.

1.3 Logistic Regression

Logistic Regression is a widely used statistical and machine learning algorithm for binary and multiclass classification tasks. It models the relationship between independent variables and a categorical dependent variable by applying the logistic (sigmoid) function to transform linear combinations of input features into probabilities.

2.Dataset

A Heart Disease dataset with the following attributes are taken for analysis to perform the prediction as shown in Table 1.

Data	Variable	Description
Heart	Glucose	Blood sugar level, often measured in mg/dL.
Heart	Cholesterol	Total cholesterol level in the blood
Heart	Hemoglobin	A protein in red blood cells that carries oxygen
Heart	Platelets	Small blood cells involved in clotting
Heart	White Blood Cells	Cells that help fight infection
Heart	Red Blood Cells	Cells responsible for transporting oxygen
Heart	Hematocrit	The percentage of red blood cells in the blood
Heart	MeanCorpuscular Hemoglobin	The average amount of hemoglobin in a red blood cell, measured in pg.
Heart	MeanCorpuscular Hemoglobin	The average concentration of hemoglobin in red

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304142

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

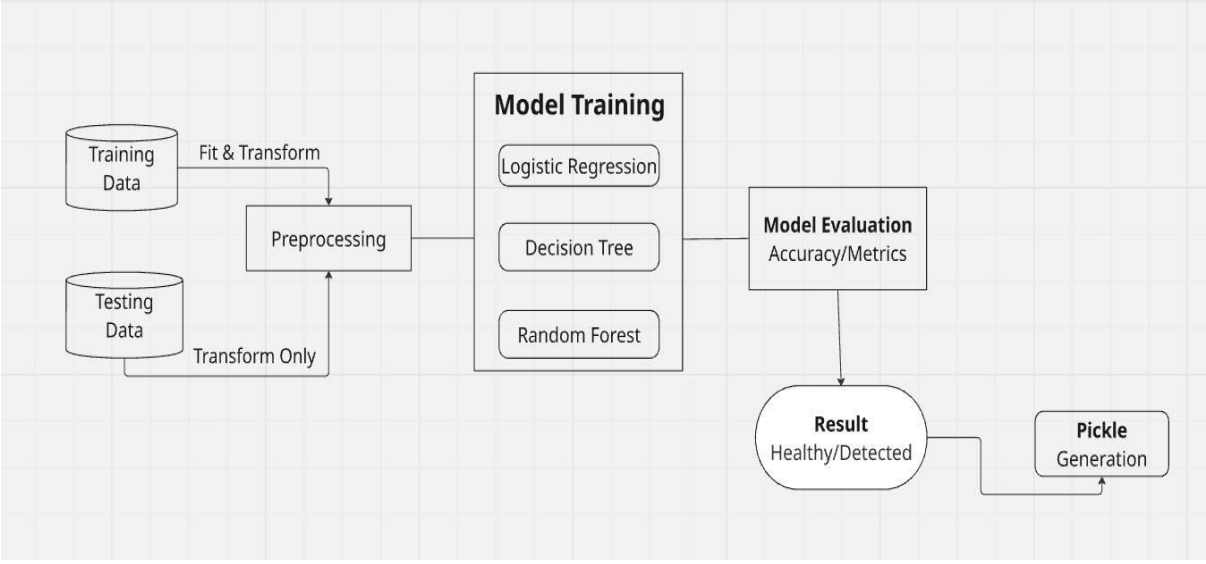
	Concentration	blood cells
Heart	Insulin	A hormone regulating blood sugar.
Heart	BMI	A measure of body fat based on height and weight
Heart	Systolic Blood Pressure	The pressure in arteries when the heart beats
Heart	Diastolic Blood Pressure	The pressure in arteries when the heart rests between beats
Heart	Triglycerids	A type of fat in the blood. High levels are associated with heart disease
Heart	HbA1c	A long-term indicator of blood sugar levels

Table 1. Dataset key attributes and its description

Table1 presents a dataset focused on heart disease dataset containing various medical attributes used for analysis and prediction. It consists of three columns: Data, Variable, and Description. The Data column consistently lists "Heart," indicating that all variables are related to heart health. The Variable column includes key medical indicators such as glucose, cholesterol, Hemoglobin, platelets, white and red blood cells, hematocrit, insulin, BMI, blood pressure, triglycerides, and HbA1c. Each of these variables is accompanied by a Description explaining its significance, such as blood sugar levels, oxygen-carrying capacity, fat levels, and pressure in arteries. This dataset provides essential insights for heart disease analysis and prediction by capturing crucial cardiovascular and metabolic health parameters.

IV. PROPOSED SYSTEM

1.Architechture



The architecture depicted in the diagram represents a health monitoring and alert system based on physiological data. The process begins with the collection of physiological data, such as heart rate, blood pressure, or other vital signs, through an automated acquisition system. This raw data is then passed through a preprocessing stage, where it is cleaned, normalized, and structured for analysis. Following preprocessing, predictive analytics are applied to assess potential health risks or abnormalities. If any critical condition is detected, an alert is generated and sent to both the patient and the hospital, ensuring timely medical intervention. This system facilitates proactive healthcare management by enabling early detection and rapid response to potential medical issues.

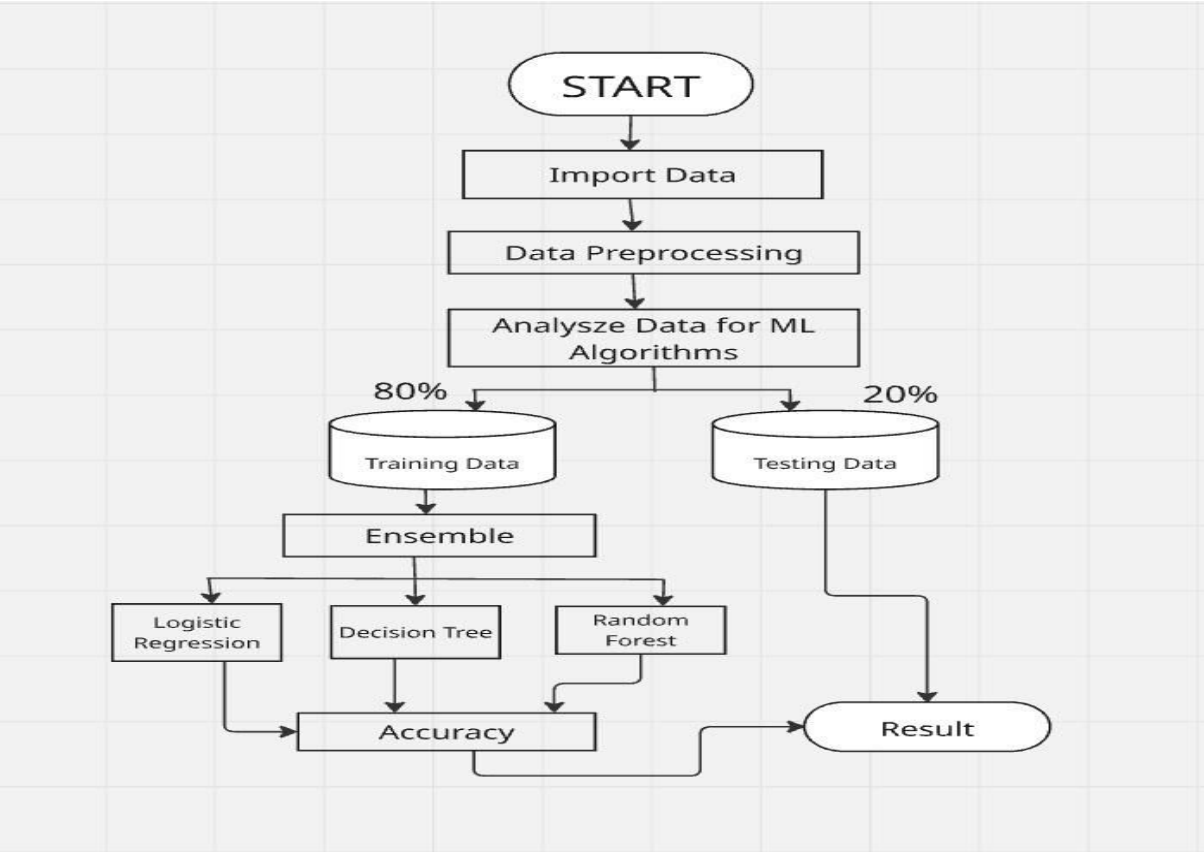
© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304142

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

2.WorkFlow



The flowchart represents a machine learning workflow for data analysis and prediction. It begins with importing and preprocessing the dataset to ensure it is clean and ready for analysis. The data is then analyzed to extract relevant features (). Next, the dataset is split into two parts: 80% for training (T) and 20% for testing (TS). The training data is used to build an ensemble model combining Random Forest (RF), Decision Tree and Logistic Regression. The trained model is then evaluated for accuracy, and the final results are generated based on its performance. This structured approach ensures effective training and testing for accurate predictions.

3.Data Collection and Preprocessing

3.1 Data Source:

The data typically comes from Publicly available health repositories and included anonymized patient information such as demographic, medical history ,symptoms ,and lab results .The data was stored in CSV or Excel Formats and contained binary classification labels(disease/no disease).

3.2 Handling Missing Values:

During data cleaning, missing values were handled using imputation techniques—numerical values were filled with the median or mean, while categorical values used the mode. Outliers were detected and removed using statistical methods like Z-scores and interquartile ranges. Normalization techniques, such as min-max scaling and z-score standardization, were applied to ensure uniform feature scaling. Data transformation involved encoding categorical variables with one-hot or label encoding, while date-related data was transformed into meaningful numerical features.

3.3 Feature Selection and Engineering:

Feature Selection: Feature selection was performed using statistical tests, Recursive Feature Elimination (RFE), and decision tree-based models to retain the most relevant attributes.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304142

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Feature engineering further improved model performance by creating new features, such as BMI from height and weight, categorizing age groups, and incorporating interaction terms like age and smoking status. These preprocessing steps ensured the dataset was clean, structured, and optimized for machine learning predictions.

3.3.1
Splitting the Data set: The data is split into a training set (usually 80%) to train the model and a testing set (usually 20%) to evaluate the model's performance. This ensures that the model is tested on unseen data to check its generalizability.

V. COMPARTIVE ANALYSIS AND RESULTS

1.Comparitive Analysis table

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-score (Class 1)
Logistic Regression	66.87%	0.99	0.67	0.80
Decision Tree	86.63%	1.00	0.86	0.93
Random Forest	96.71%	1.00	0.97	0.98

Fig7: Different models evaluations

The table compares the performance of three machine learning models—Logistic Regression, Decision Tree, and Random Forest—based on their Accuracy, Precision, Recall, and F1-score for Class 1. Logistic Regression shows high precision (0.99) but lower recall (0.67), resulting in a moderate F1-score of 0.80 and the lowest accuracy (66.87%). The Decision Tree improves significantly with 86.63% accuracy, perfect precision (1.00), better recall (0.86), and a stronger F1-score (0.93). Random Forest outperforms both, achieving the highest accuracy (96.71%), perfect precision (1.00), high recall (0.97), and the best F1-score (0.98), making it the most effective model for this classification task.

2.Results:

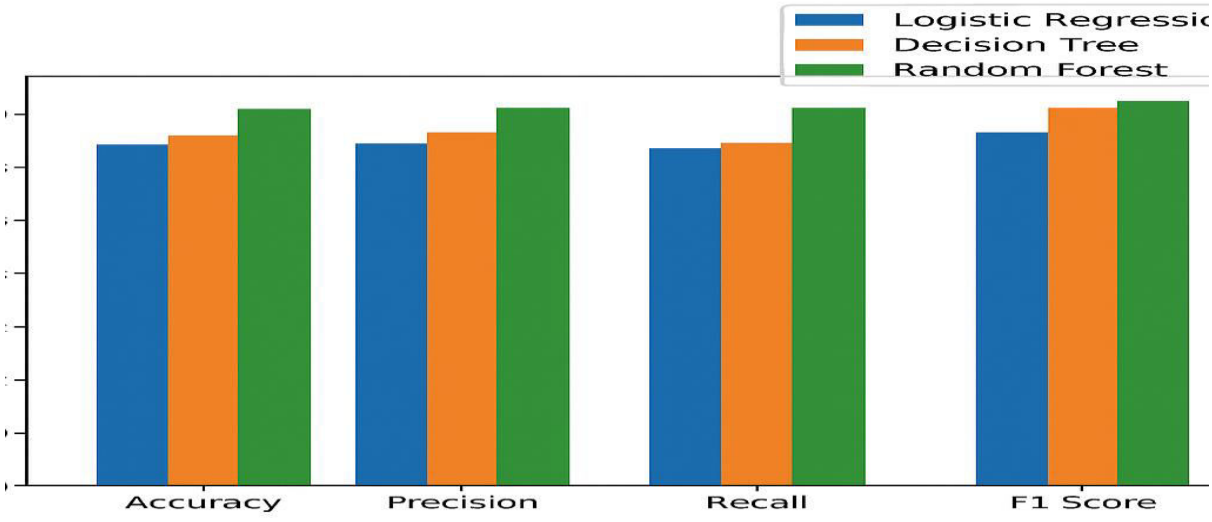


Fig8: Accuracy % for various Classifiers

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304142

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

This bar chart compares the performance of three models—Logistic Regression, Decision Tree, and Random Forest—across four metrics: Accuracy, Precision, Recall, and F1 Score. Each metric is represented by a group of bars, with different colors for each model. Overall, Random Forest (green bars) consistently shows the highest values across all metrics, followed by the Decision Tree (orange), and Logistic Regression (blue), which has the lowest performance.

2.1 APP.py

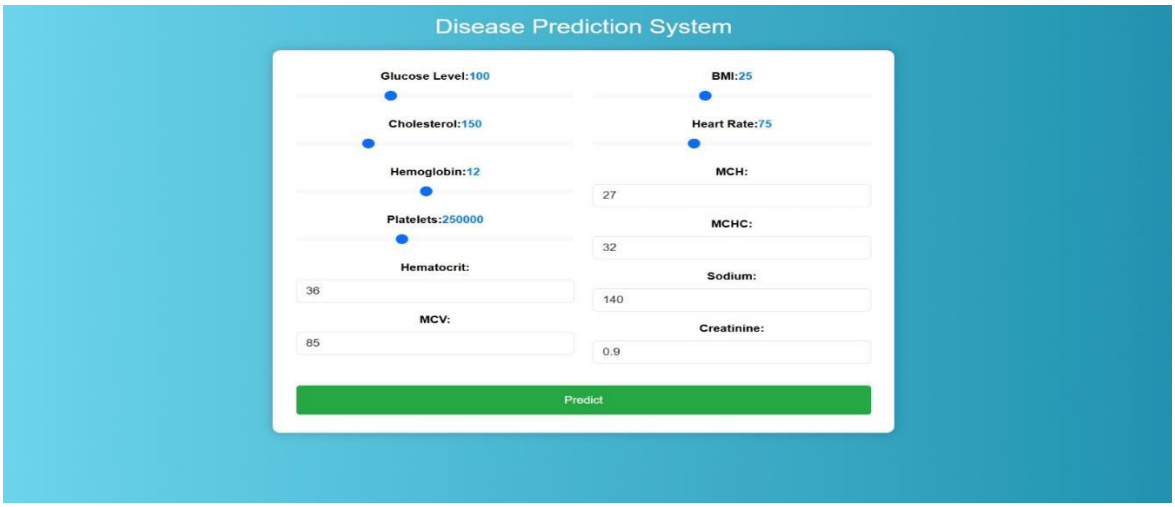


Fig 9: Input Image

Fig9 shows what i/p data we should collect and used for prediction.

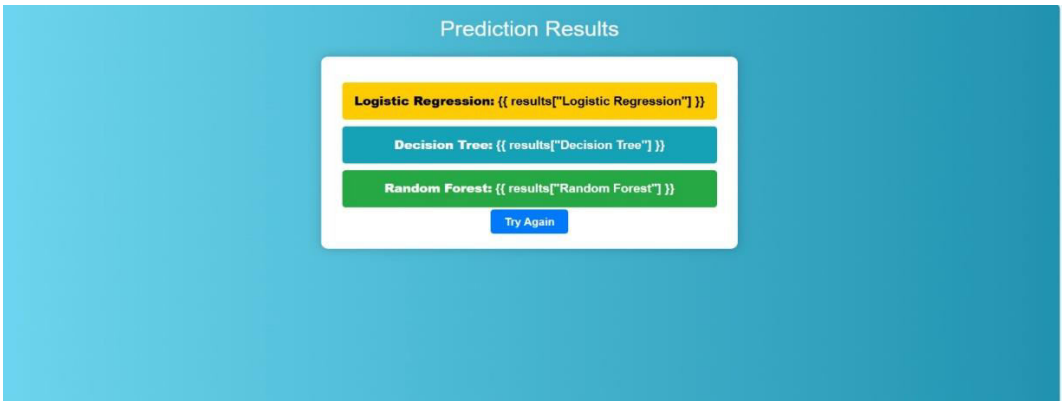


Fig10: Prediction Result

Fig10 Shows the predicted result of give data.

VI. CONCLUSION

The Disease Diagnosis Prediction Using Data Analytics project has the potential to revolutionize healthcare by improving the accuracy and speed of disease diagnosis. Early detection of diseases, especially chronic and life-threatening conditions, can significantly improve patient outcomes by enabling timely treatments. By leveraging advanced machine learning algorithms, big data technologies, and real-time health monitoring, this system provides a scalable and adaptable solution for healthcare providers to deliver personalized, data-driven care.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

1. Jiang, H., Chen, Y., & Liu, Z. (2021). AI-Driven Predictive Analytics in Healthcare. Journal of Biomedical Informatics, 118, 103794.
2. Nguyen, T., Ho, D., & Pham, B. (2020). Deep Learning Applications in Disease Diagnosis. Artificial Intelligence in Medicine, 107, 101910.
3. Patel, J., Shah, S., & Doshi, D. (2019). Predictive Modelling for Chronic Disease Management. IEEE Transactions on Biomedical Engineering, 66(11), 3123-3134.
4. Kumar, R., Gupta, S., & Verma, P. (2022). Machine Learning in Medical Imaging. Computers in Biology and Medicine, 145, 105874.
5. Singh, A., Roy, P., & Chakraborty, S. (2023). Big Data Analytics for Health Prediction. Journal of Medical Systems, 47(2), 24.
6. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. New England Journal of Medicine, 380(14), 1347–1358.
7. Chicco, D., & Jurman, G. (2020). Machine Learning for Healthcare Prediction: A Review. Briefings in Bioinformatics, 21(2), 1-10.
8. Kourou, K., Exarchos, T. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine Learning Applications in Cancer Prognosis and Prediction. Computational and Structural Biotechnology Journal, 13, 8-17
9. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.
10.Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1(1), 1-10.
11.Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246.
12.Dilsizian, S. E., & Siegel, E. L. (2014). Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Current Cardiology Reports, 16(1), 441.
13.Razzak, M. I., Imran, M., & Xu, G. (2019). Big data analytics for preventive medicine. Neural Computing and Applications, 32, 4417–4451.
14.Chen, M., Ma, Y., Li, Y., Wu, D., Zhang, Y., & Youn, C. H. (2017). Wearable 2.0: Enabling human-cloud integration in next generation healthcare systems. IEEE Communications Magazine, 55(1), 54–61.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details