iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XII **Month of publication:** December 2023

DOI: https://doi.org/10.22214/ijraset.2023.57309

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 11 Issue XII Dec 2023- Available at www.ijraset.com

Prediction of Disease Based on Symptoms Using Random Forest Classifier

M.. Lakshmi Srija[1], CH . Siva Teja[2] , P . Yagna Deep[3], T. Nandhini[4], CH. Venkata Satya Narayana[5], Mr. N.V. Murali Krishna Raja[6]

[1, 2, 3, 4, 5]*Department of Computer Science and Artificial Intelligence (CAI), Sri Vasavi Engineering College, Tadepalligudem, Andhrapradesh, India*

[6]*Professor, Department Of Computer Science and Engineering (CSE), Sri Vasavi Engineering College, Tadepalligudem, Andhrapradesh, India*

Abstract: ***We address the critical need for early disease detection by focusing on the prediction of diseases based on user-provided symptoms using machine learning (ML). Our primary aim is to create a user-friendly and effective system for early disease detection using Random Forest Classifier. Users can input four symptoms, and behind the scenes, our ML model processes this information to make informed disease predictions. We've leveraged a dataset containing records of 41 diseases and 132 Symptoms to develop and train our model, ensuring its accuracy. The project not only empowers individuals to take control of their health but also promises to enhance healthcare quality and reduce costs, benefiting both patients and the healthcare system. Our approach highlights the transformative potential of computer-assisted healthcare in saving lives and resources.***

I. INTRODUCTION

As healthcare becomes more intertwined with cutting-edge technology, the need for early disease detection takes center stage. Our project, however, takes a broader approach by focusing on the prediction of various diseases based on symptoms, offering a versatile solution to the evolving healthcare challenges. With the power of the Random Forest algorithm, we aim to empower individuals to take control of their health and make well-informed decisions by predicting a range of health-related diseases even before clinical signs manifest. In doing so, we contribute to proactive healthcare initiatives, ensuring individuals can embark on a path to better health with the aid of advanced technology and the foresight of machine learning.

II. LITERATURE SURVEY

Palle Pramod Reddy, Dirisinala Madhu Babu, Hardeep Kumar, and Dr. Shivi Sharma revolves around the development and application of a "Disease Prediction" method in healthcare. Their work centers on utilizing predictive modeling, specifically employing the random forest classifier, to estimate the likelihood of a user having a particular disease based on input symptoms. Through meticulous analysis, their method generates probabilities associated with various diseases, facilitating early detection and diagnosis. This innovative approach harnesses user-provided symptoms to enable a proactive healthcare model, aiming to enhance patient care through timely interventions and personalized disease management strategies.

Dr. C K Gomathy and Mr. A. Rohith Naidu at Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Kanchipuram, introduces a Disease Prediction system employing machine learning methodologies to forecast diseases based on user-reported symptoms. Utilizing a Naïve Bayes classifier alongside other machine learning techniques like linear regression and decision trees, their system calculates the likelihood of specific diseases such as Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis. By tapping into the wealth of biomedical and healthcare data, their innovative approach aims to enhance early disease detection, offering a valuable tool for improving the accuracy of diagnoses and streamlining treatment strategies, ultimately contributing to more efficient patient care and management.

Rinkal Keniya, Aman Khakharia, Vruddhi Shah, Vrushabh Gada, Ruchi Manjalkar, Tirth Thaker, Mahesh Warang, and Ninad Mehendale focuses on creating a robust medical diagnosis system using machine learning algorithms. With a focus on enhancing the accuracy and speed of disease prediction and diagnosis, the project addresses limitations in traditional diagnostic methods, particularly for severe illnesses. Their system incorporates multiple machine learning algorithms and utilizes a comprehensive dataset covering information on more than 230 diseases. By analyzing an individual's symptoms, age, and gender, the system offers predictions for potential diseases. Notably, the weighted KNN algorithm emerged as the most effective during testing, boasting an impressive prediction accuracy of 93.5%. This system serves as an early diagnostic tool, potentially playing a crucial role in timely treatment and interventions, thereby potentially saving lives through prompt medical care.

Kriti Gandhi, Mansi Mittal, Neha Gupta, and Shafali Dhall centers on integrating machine learning into healthcare practices, aiming to significantly enhance patient care standards. Acknowledging the critical need for advanced medical facilities, the project explores the application of various machine learning algorithms, including KNN, Random Forest, and Decision Tree Classifier, within healthcare procedures. Their primary objectives encompass improving patient satisfaction by optimizing treatment processes and highlighting the pivotal role of early disease prediction in healthcare. The project emphasizes the adage "Prevention is better than cure," emphasizing the importance of early detection in halting ailment progression. Of note, the project's focus on KNN and Logistic Regression signifies a specific emphasis on these machine learning techniques within the healthcare sector, underlining their potential for augmenting diagnostic and predictive capabilities in patient care scenarios.

Sneha Grampurohit and Chetan Sagarnal focuses on the application of data mining and machine learning in healthcare and biosciences. Their primary objective involves utilizing these technologies to extract crucial insights from medical data, specifically emphasizing the improvement of data analysis for early disease prediction and enhanced patient care. Through the utilization of machine learning, their project aims to create a system supporting healthcare professionals in early disease prediction and diagnosis. Their methodology involves working with a dataset containing records from 4,920 patients diagnosed with 41 different diseases. They meticulously curate 95 relevant independent variables (symptoms) from a pool of 132 for further analysis and optimization. Employing machine learning algorithms such as Decision Tree, Random Forest, and Naïve Bayes, their research conducts a comparative study to evaluate their effectiveness in disease prediction. The outcomes of this research contribute significantly to advancing healthcare services by enabling early disease detection and ultimately improving patient outcomes through more informed and timely interventions.

III. PROBLEM STATEMENT IN EXISTING SYSTEM

The healthcare sector confronts substantial obstacles when it comes to achieving early and precise disease diagnoses, rectifying diagnostic inaccuracies, handling intricate patient data, and coping with a shortage of medical expertise. There's a significant number of individuals grappling with diseases that could see more effective treatments and better outcomes if they were identified at an earlier stage. To tackle these challenges, the project sets out to build a disease prediction system utilizing machine learning. This system is designed to make accurate predictions about diseases based on symptoms reported by patients and their data. The primary objectives include managing data quality, safeguarding patient information, enhancing the transparency of the model's decision-making process, and building trust with patients. The ultimate goal is to substantially improve early disease detection and the overall management of healthcare, thereby reducing diagnostic inaccuracies and enabling healthcare professionals to provide more precise and prompt care to their patients.

A. Disadvantages in Existing System

1) Data Quality and Quantity: To make accurate predictions, we need a lot of good-quality data. If the data is incomplete, inconsistent, or biased, it can lead to incorrect predictions. Getting the right data is tough but crucial.
2) False Positives and Negatives: Sometimes, the predictions can be wrong. A false positive is when the system predicts a disease that's not really there, causing unnecessary worry. A false negative is when it misses a real disease, delaying treatment. Striking the right balance is tricky.
3) Data Imbalance: Some diseases are rare, so there aren't many cases to learn from. This can make the system biased towards common diseases. It's tough to make it good at predicting both common and rare diseases.
4) Not effective and user-friendly models.

IV. PROPOSED SYSTEM

1) *Data Collection:* We've meticulously curated an extensive dataset, sourced from Kaggle, that encompasses a comprehensive array of diseases along with their associated symptoms. This dataset serves as the cornerstone for training and validating our machine learning model.
2) *Machine Learning Model:* Our predictive engine is fortified with the Random Forest algorithm, a robust ensemble learning technique renowned for its prowess in handling classification tasks. It adeptly processes the copious symptom-disease data, delivering high-precision predictions.
3) *Flask Deployment:* To ensure seamless accessibility, we've chosen the Flask Python framework for deploying our machine learning model. This allows users to effortlessly tap into the disease prediction system through a standard web browser.

4) *User-Friendly Interface:* At the core of our solution is an intuitive, user-friendly interface meticulously crafted using HTML, CSS, and JavaScript. This aesthetically pleasing web application caters to both patients and healthcare providers, enabling them to effortlessly input symptoms, receive accurate disease predictions, and access pertinent information

A. *Advantages In Proposed System*

1) *Early Detection:* By considering a patient's symptoms, our system enables the early detection of diseases, leading to timely intervention and improved outcomes
2) *Support for Healthcare Professionals:* Healthcare providers benefit from this system as it serves as a valuable decision support tool, offering insights and suggestions to aid in diagnosis.
3) *Reducing Misdiagnosis:* With its advanced algorithms, the system can significantly reduce the occurrence of misdiagnosis, thus enhancing patient safety.
4) *Addressing Medical Expert Shortages:* In regions with shortages of medical professionals, this system can augment healthcare delivery by providing accurate disease predictions.
5) *Identifying Rare Diseases:* It excels in recognizing rare and uncommon diseases, which might otherwise be challenging to diagnose.

V. DATASET DESCRIPTION

The dataset for our project , "Disease Prediction Based on Symptoms Using Random Forest Algorithm" is a comprehensive collection encompassing 41 distinct diseases and 132 diverse symptoms. This dataset's richness reflects the complexity of real-world healthcare scenarios, enabling your machine learning model to predict a wide spectrum of health conditions, from common ailments to rarer diseases. Data quality and integrity are emphasized to ensure reliable predictions, and the dataset serves a dual purpose for both model training and testing. Preprocessing may have been applied to handle missing data and ensure data uniformity.

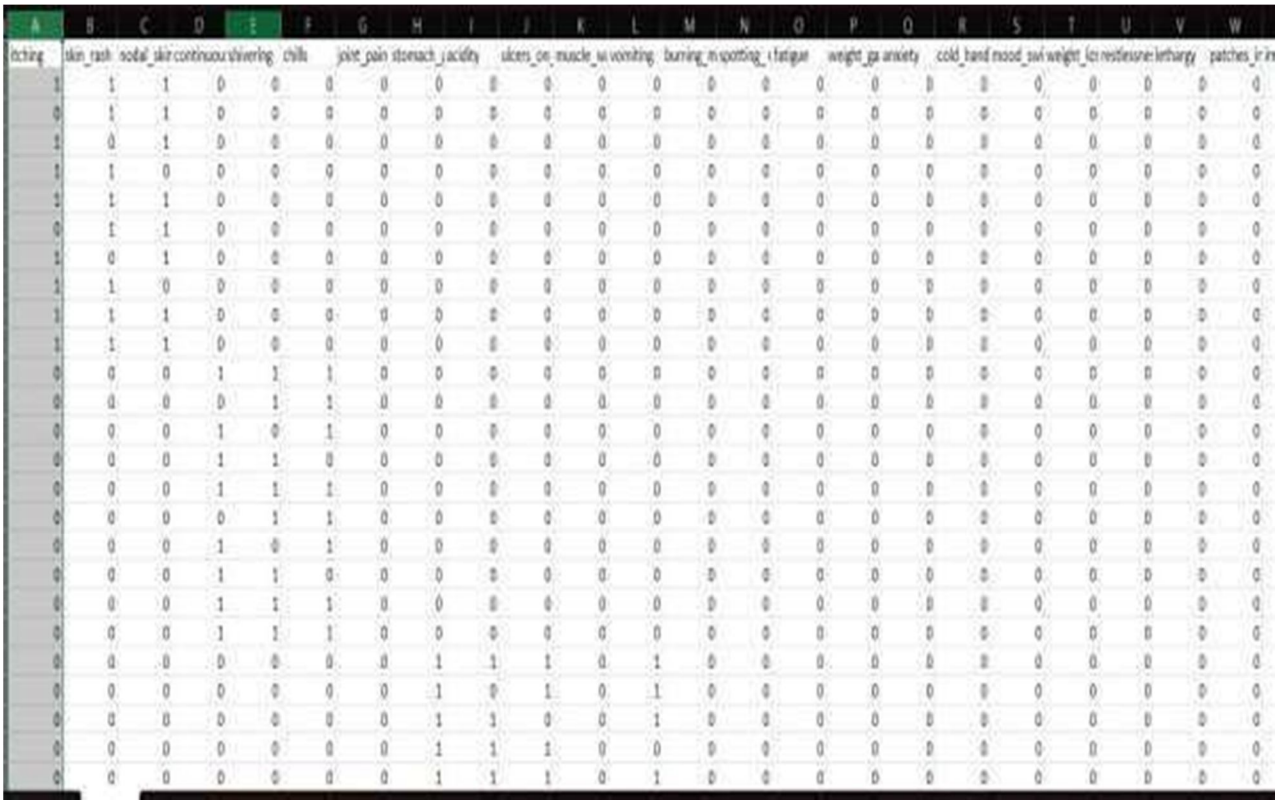


Fig-1 Training Dataset

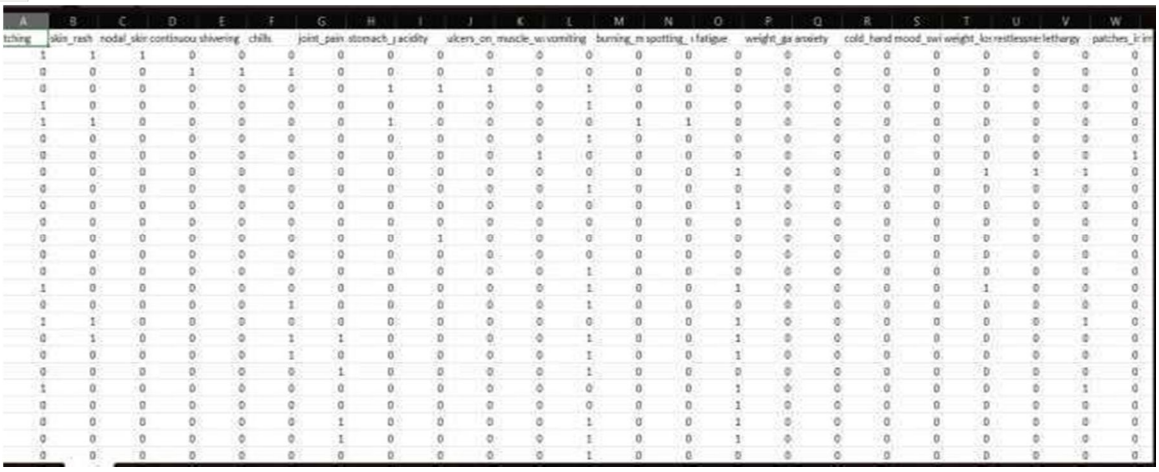


Fig-2 Testing DataSet

VI. METHODOLOGY

The methodology for our project follows a structured approach:

1) *Data Collection:* We start by gathering a diverse dataset containing information on 41 different diseases and 132 symptoms. Ensuring the quality and completeness of the data is paramount for accurate predictions.
2) *Data Preprocessing:* Cleaning and preprocessing the dataset is the next step. This involves handling missing values and normalizing the data to ensure uniformity. Proper data preparation is crucial for effective machine learning model training.
3) *Feature Selection:* Not all symptoms are equally relevant for disease prediction. Therefore, we employ feature selection techniques to identify the most important symptoms that significantly contribute to the accuracy of the model.
4) *Model Selection:* We choose the Random Forest algorithm for disease prediction. Random Forest is renowned for its ability to handle complex datasets and make highly accurate predictions.
5) *Model Training:* The selected Random Forest model is trained using the preprocessed dataset. During this phase, the model learns the patterns and relationships between symptoms and diseases.
6) *User Interface:* We develop a user-friendly interface that enables users to input their symptoms conveniently. The system is designed to process this information effectively and provide predictions promptly.
7) *Prediction and Testing:* We utilize the trained model to make predictions based on the symptoms provided by users. To ensure its reliability and accuracy, we rigorously test the model using a separate testing dataset.
8) *Evaluation Metrics:* Appropriate evaluation metrics, such as accuracy, precision, and recall, are employed to assess the model's performance. This data-driven evaluation helps us refine the model as necessary.
9) *Deployment:* Once the model consistently performs well and meets our defined criteria, we deploy it for real-world use. Users can input their symptoms, and the system will provide disease predictions based on the robust Random Forest algorithm. This systematic methodology ensures that our project is well-structured and capable of delivering accurate disease predictions.

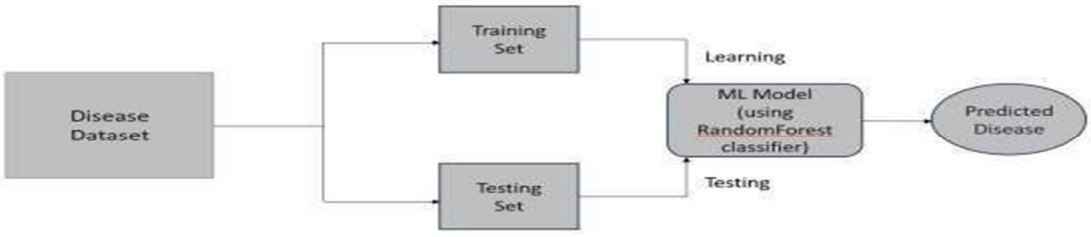


Fig-3 Methadology of Proposed model

VII. SYSTEM ARCHITECTURE

The system architecture for your project on "Disease Prediction Based on Symptoms Using Random Forest Classifier" can be likened to a decision tree. Much like a decision tree, this project employs a straightforward and interpretable machine learning model, the Random Forest Classifier, which is versatile for both classification and regression tasks. It operates by recursively dividing input data into subsets based on symptom values. The architecture consists of nodes, representing decision points; branches, representing potential outcomes; and leaves, representing final predictions. Decision trees, in this context, partition the data based on symptom values to create nodes. The goal is to make splits that best classify the data into disease categories. This process continues until stopping criteria are met. In the case of classification, a leaf node typically represents a predicted disease label. The final prediction in your project is derived by aggregating the predictions of individual decision trees, often accomplished through techniques like majority voting. This architecture ensures an understandable and effective system for disease prediction based on user-reported symptoms using the Random Forest Classifier.

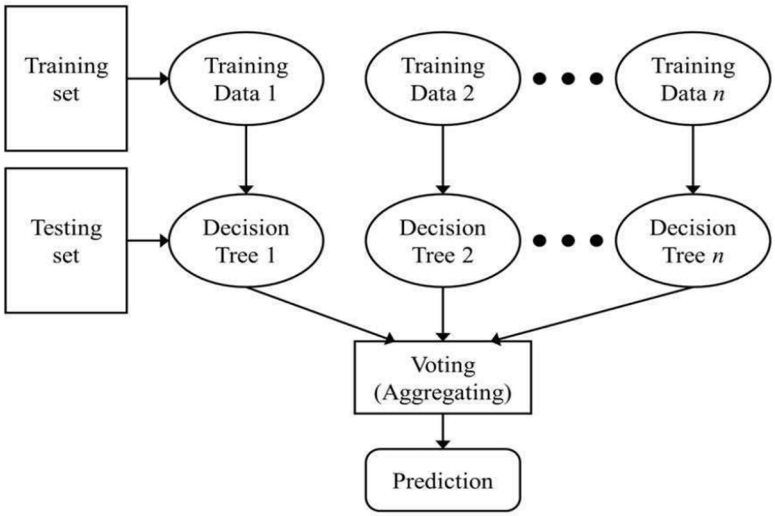


Fig-4 System Architecture

VIII. EXPERIMENTAL RESULTS

The interface of our Disease Prediction System based on symptoms using the Random Forest Classifier presents a user-friendly and intuitive design. It serves as the primary point of interaction between users and the system

1) There is a dedicated section where users can input their symptoms. Users can easily enter up to four specific symptoms that they are experiencing, such as headaches, cough, fatigue, or fever etc.

Fig-5 Interface of the proposed Model

2) The user input section, there is a prominent "Predict" button. Users simply need to click this button after entering their symptoms to initiate the prediction process.

Fig-6 Given 4 Symptoms as Input

3) The main area of the interface is reserved for displaying the prediction results. When users click the "Predict" button, the system processes the input symptoms and promptly provides the predicted disease based on the Random Forest model's analysis. The result is clearly presented in a user-friendly format.

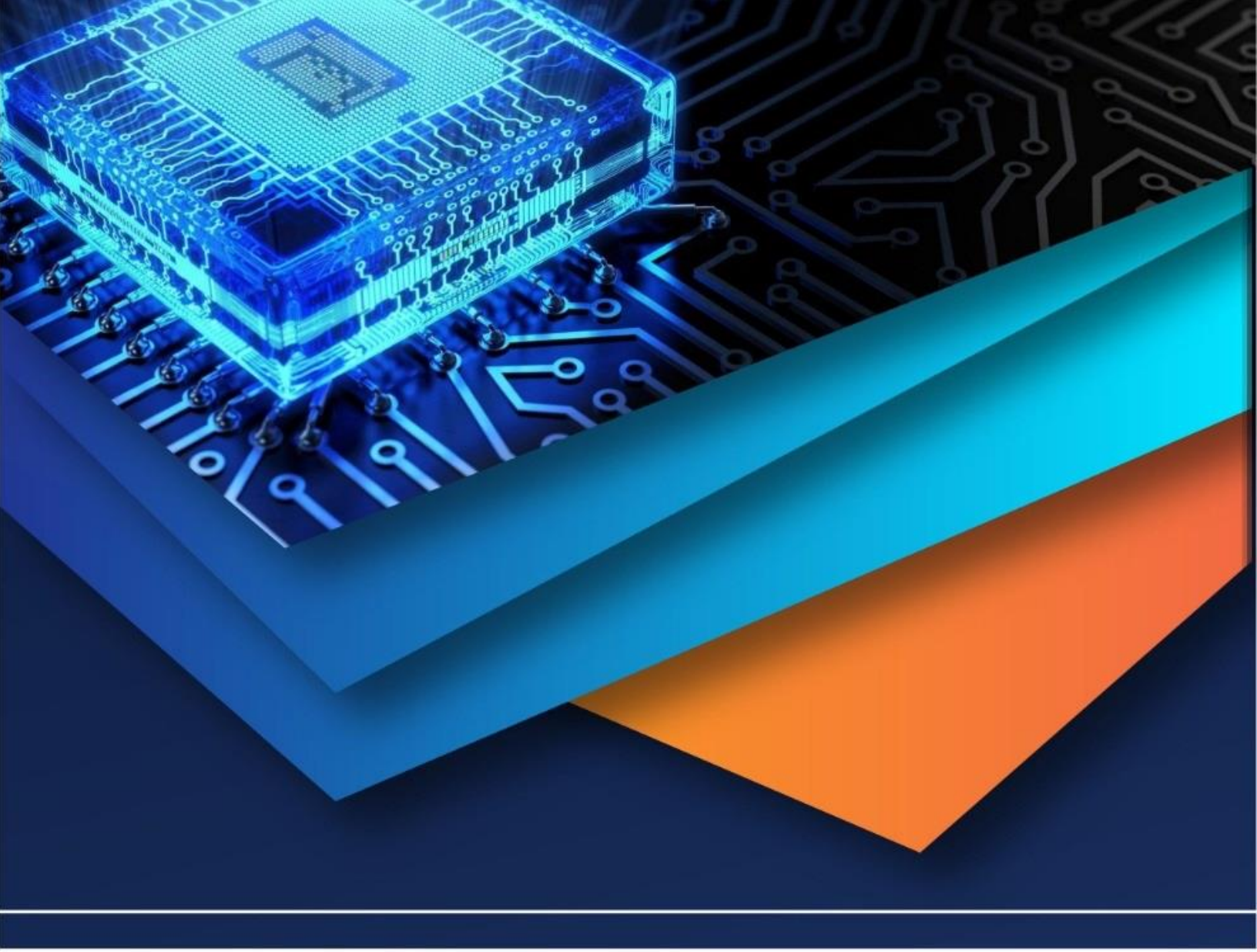
Fig-7 Final Output(Predicted Disease)

IX. CONCLUSION

Our project on "Disease Prediction Based on Symptoms Using Random Forest Classifier" has successfully developed a user-friendly and accurate system for early disease detection. Leveraging machine learning and a comprehensive dataset, we've created a valuable tool that can significantly improve healthcare by empowering individuals and healthcare providers. While challenges and limitations exist, our project represents a crucial step towards more efficient and effective disease prediction, ultimately contributing to better patient outcomes and healthcare quality. This work highlights the potential of technology to make a positive impact on public health and healthcare services.

REFERENCES

[1] Palle Pramod Reddy, Dirisinala Madhu Babu, Hardeep Kumar and Dr.Shivi Sharma ,"Disease Prediction using Machine Learning" in 2021 , DOI:10.2139/ssrn.3167431.
[2] Dr C K Gomathy, Mr. A. Rohith Naidu Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya , Kanchipuram "The prediction of disease using machine learning" in 2021,International Journal of Scientific Research in Engineering and Management (IJSREM), SSN: 2582-3930
[3] Rinkal Keniya , Aman Khakharia ,Vruddhi Shah , Vrushabh Gada ,Ruchi Manjalkar , Tirth Thaker , Mahesh Warang , Ninad Mehendale,"Disease Prediction From Various Symptoms Using Machine Learning" in 2020,SocialScienceResearchNetwork,DOI:10.2139/ssrn.3661426
[4] Kriti Gandhi, Mansi Mittal, Neha Gupta, Shafali Dhall ;" Disease Prediction using Machine Learning" in 2020, International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653
[5] Sneha Grampurohit , Chetan Sagarnal, "Disease Prediction using Machine Learning Algorithms" in 2020 ,DOI:10.1109/INCET49848.2020.9154130



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY