iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79869

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 14 Issue IV Apr 2026- Available at www.ijraset.com

Disease Prediction and Doctor Recommendations System

Prachi Kharat[1], Kiran Mohod[2], Shital Jadhav[3], Nikita Dongre[4], Prof. Vasanti Y. Gaud[5]
Department of Computer Science &Engineering, College of Engineering & Technology, Akola, Maharashtra, India

Abstract: ***In the modern era of data-driven decision-making, early and accurate disease diagnosis has emerged as a critical challenge, particularly in resource-constrained settings. This project proposes a machine learning-based disease prediction system using a Random Forest Classifier to forecast potential diseases based on symptoms provided by the user. The system is designed as a robust, interactive tool to aid in preliminary medical assessments. The model has been trained on a dataset comprising 4,920 records that span 133 symptoms and 41 unique diseases, using binary encoding to represent the presence or absence of each symptom. The core of the system is the Random Forest algorithm, chosen for its high accuracy, robustness, and ability to handle large feature spaces effectively. The classifier achieves an accuracy of approximately 97.6% on unseen test data, demonstrating strong predictive performance. The user can interact with the system via a Command Line Interface (CLI), inputting symptoms to receive a predicted disease along with a disclaimer highlighting the system's advisory nature. In addition to disease prediction, the model provides a feature importance visualization, offering transparency into which symptoms most influence the outcome. This not only improves interpretability but also serves as a learning aid for users and researchers.***
By tapping into the capabilities of scikit-learn, pandas, and visualization libraries, the project exemplifies how a well-tuned ML pipeline can serve real-world healthcare needs. While the system in its current form is intended for educational and exploratory purposes, it sets the stage for broader implementation in mobile apps, hospital triage tools, and telemedicine systems.
Keywords: Machine Learning, Command Line Surface, Random Forest Algorithm

I. INTRODUCTION

The proposed project, "Machine Learning-Based Disease Prediction Using Random Forest Classifier," aims to predict probable diseases based on symptoms input by a user. It leverages the power of ensemble learning, specifically the Random Forest algorithm, to derive accurate, interpretable, and scalable disease predictions. This system, although not a substitute for medical professionals, can serve as a frontline filter in guiding users toward seeking appropriate care. Health is one of the most vital assets of human life, yet millions across the world still face difficulty in accessing immediate, reliable, and affordable medical guidance. In today's fast-paced world, where time is of the essence and expert medical advice may not be instantly available, artificial intelligence (AI) and machine learning (ML) are becoming indispensable tools in augmenting traditional healthcare systems. With the explosive growth of medical data and the increasing accessibility of computing resources, predictive models are poised to revolutionize the preliminary diagnosis process.

the project can be used as a foundation for developing more advanced decentralized applications. Its scope mainly lies in helping students and developers learn and explore blockchain technology in a practical and understandable way

A. Need of the present study

In the evolving landscape of healthcare technology, there exists a critical need for systems that can effectively bridge the gap between initial symptom assessment and specialized medical care. Despite significant advancements in medical diagnostics and artificial intelligence, several challenges persist that underscore the necessity of this study. Traditional healthcare pathways often involve multiple consultations before patients receive appropriate specialist care, resulting in delayed diagnosis and treatment initiation. Early and accurate disease prediction can significantly improve patient outcomes, especially for time-sensitive conditions. This delay represents a critical gap in modern healthcare delivery systems that machine learning approaches could potentially address. Without proper guidance, patients frequently consult physicians whose specialties do not align with their conditions, leading to inefficient use of medical resources, unnecessary referrals, and increased healthcare costs. This misdirection of patients contributes substantially to healthcare system inefficiencies and represents an opportunity for technological intervention through intelligent recommendation systems.

Many regions, particularly rural and underserved areas, face shortages of medical specialists. A system that can provide preliminary disease assessment and appropriate specialist recommendations can help mitigate this disparity in healthcare access. The digital divide in healthcare accessibility remains a persistent challenge that this research aims to address through accessible technology.

B. Objectives of the Present Study

The present research aims to develop and validate a comprehensive Disease Prediction and Doctor Recommender System with the following specific objectives:

1) To design and implement a machine learning-based disease prediction model capable of identifying potential diseases from patient-reported symptoms with high accuracy and reliability across multiple disease categories.

2) To develop an intelligent doctor recommendation system that matches predicted diseases with appropriate medical specialists, considering factors such as specialization, expertise, and disease-specific requirements.

II. LITERATURE REVIEW

Aamir et al. (2022) explored the potential of supervised machine learning to predict breast cancer. The study utilized several classification models like SVM, Decision Trees, and KNN on benchmark datasets. Their experimental evaluation demonstrated high accuracy and low error rates. The research emphasized the clinical relevance of ML for early detection. This paper marks a pivotal step in data-driven cancer diagnostics.

Amin et al. (2020) analyzed brain data using machine intelligence to detect cognitive abnormalities. The authors used deep learning and image processing to identify pathological brain patterns. They proved how ML could support neurodegenerative disorder diagnosis. The approach fused MRI data with AI classifiers. It laid the groundwork for non-invasive brain diagnostics.

Du et al. (2020) focused on predicting coronary heart disease in hypertensive patients using EHR data. They applied big data analytics and machine learning models like XGBoost. The models achieved strong performance metrics like AUC and precision. Their research highlighted how digital health records could revolutionize predictive diagnostics. It stands as a practical application of AI in clinical settings.

El-Hasnony et al. (2022) introduced a multi-label active learning model for heart disease prediction. The model reduced the annotation burden while increasing model accuracy. By selecting the most informative instances, the model learned efficiently. The paper underscored the usefulness of active learning in medical data applications. It contributes to cost-effective health prediction systems.

Garg et al. (2021) implemented several machine learning techniques for heart disease prediction. The authors compared Logistic Regression, SVM, and Random Forest. Random Forest outperformed other models in terms of accuracy. The study showed the potential of ensemble learning in health data analysis. The findings aid in proactive cardiac care.

Pal and Parija (2021) applied Random Forest for heart disease prediction. The model delivered better results compared to baseline classifiers. The study highlighted the importance of hyperparameter tuning. They underscored the significance of model stability. It remains a compact yet impactful contribution.

Rahman (2022) proposed a web-based prediction system for heart disease using ML. The interface integrated algorithms like Naive Bayes and Decision Trees. The platform offered real-time user interaction. It showed how tech can democratize access to preventive care. This project has practical deployment potential.

Sarra et al. (2022) boosted heart disease prediction using ML with χ^2-based feature selection. Their hybrid approach eliminated noisy attributes. It improved classification accuracy and interpretability. The paper illustrates how statistical filters enhance ML performance. It's a fine blend of classic stats and modern AI.

Shamrat et al. (2020) analyzed breast disease prediction via various ML models. They discussed the performance trade-offs between models. The study found that SVM and ANN performed well across datasets. It stressed dataset diversity and model selection. This work contributes to domain-specific AI refinement.

Shah et al. (2023) penned an editorial emphasizing health tech assessment in cardiovascular diseases. They reviewed evaluation metrics and policy frameworks. The piece discussed cost-effectiveness and tech scalability. Though not a technical study, it guides research prioritization. A great overview of the field’s trajectory.

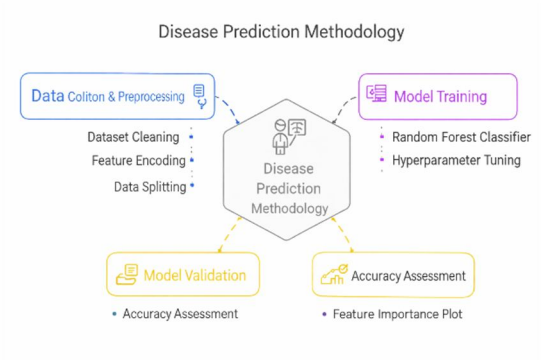
Srivastava and Singh (2022) explored ML for heart disease detection. They reported high accuracy using SVM and Decision Trees. Their work highlighted fast model convergence and predictive reliability. It contributed to real-world diagnostic tools. The paper supports AI-driven cardiology.

III. METHODOLOGY

A. *Proposed System*

The methodology began with data collection and preprocessing, where the dataset consisting of 4,920 rows and 133 symptom features was cleaned, encoded, and split into training and testing subsets. Each symptom was treated as a binary feature (1 for present, 0 for absent). The disease labels were encoded and prepared for classification. A Random Forest Classifier was then trained using the scikit-learn library. It was chosen for its ensemble nature, which combines multiple decision trees to reduce overfitting and boost accuracy. Hyperparameter tuning was performed to optimize performance, and the model was validated on the test dataset, yielding an accuracy of approximately 97.6%. A feature importance plot was generated to help visualize the top predictive symptoms.

B. *Block Diagram*



Block Diagram

The diagram illustrates the disease prediction methodology, starting with data collection and preprocessing, including cleaning, encoding, and splitting. It then moves to model training using a Random Forest classifier with hyperparameter tuning. Finally, model validation is performed through accuracy assessment and feature importance analysis to ensure reliable and effective disease prediction results.

C. *Flowchart*

Figure 3. 2: Shows the flowchart of system

The flowchart represents the initial stages of the disease prediction system, focusing on dataset collection and data preprocessing. Patient records and features are gathered systematically, followed by cleaning, filtering, and organizing the data. This step ensures accuracy and consistency, preparing the dataset for effective model training and reliable disease prediction outcomes.

D. Algorithm Detail

Random Forest is an ensemble learning method based on the concept of building multiple decision trees and combining their outputs to make a final decision. It works by:

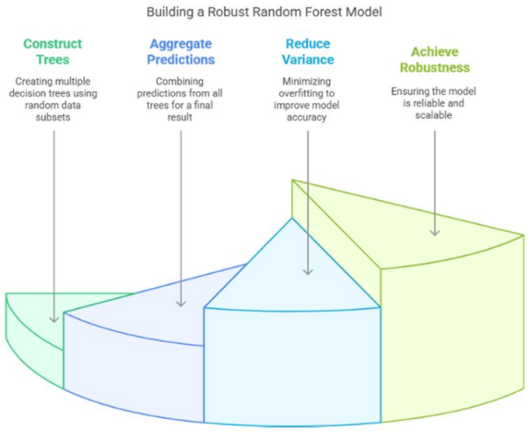


figure 3. 3 Shows the Random Forest Features

E. Working

This system is a command-line interface (CLI) tool designed to predict the most probable disease based on user-entered symptoms. It integrates a pre-trained Random Forest model and symptom matching logic to perform basic medical diagnostics. Below is a detailed description of how the system functions from user input to prediction output.

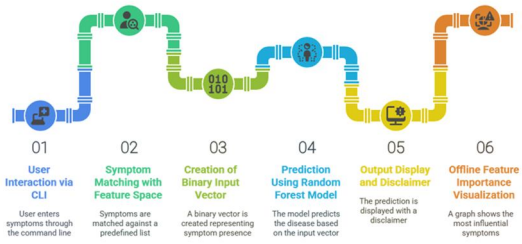


figure 3. 4 Working

3.5.1 User Interaction via CLI

The system begins by prompting the user to enter their symptoms through the command line. The input is expected as a comma-separated list of symptoms, written in plain language (e.g., “fever, cough, fatigue”). The system parses this input and converts it into a format suitable for further processing.

3.5.2 Symptom Matching with Feature Space

The system maintains a predefined list of all possible symptoms (also known as features), which were used during model training. Once the user's input is captured, each entered symptom is checked against this master list.

- If the user-entered symptom exists in the feature list, it is marked as present.
- If not, it is ignored or considered absent. This ensures consistency between the model’s training schema and live user input.

3.5.3 Creation of Binary Input Vector

After matching, the system creates a binary input vector representing the presence (1) or absence (0) of each symptom from the full feature list. For example, if the feature list includes ['fever', 'cough', 'headache', 'nausea'] and the user enters fever, nausea, the binary input vector would be [1, 0, 0, 1]. This binary vector becomes the input to the trained machine learning model.

3.5.4 Prediction Using Random Forest Model

The binary input vector is then passed to the pre-trained Random Forest classifier, which has already learned patterns from historical data linking symptoms to diseases.

- The model processes the input and outputs a predicted disease label.
- This prediction is based on ensemble decision-making from multiple decision trees that comprise the Random Forest model.
- The final prediction is the majority vote outcome across all trees.

F. Output Display and Disclaimer

The predicted disease is displayed to the user in the terminal, accompanied by a medical disclaimer that emphasizes the limitations of the system. The disclaimer reminds the user that: "This prediction is generated by an AI model and should not be considered a substitute for professional medical advice." This step is crucial to ethically inform users of the system's limitations and encourage real- world consultation.

G. Offline Feature Importance Visualization

While the system is CLI-based, it also generates an offline visual output showing the most influential features (symptoms) in the Random Forest model.

- A bar graph is plotted, ranking the top 10 symptoms based on their feature importance scores.
- The importance scores are derived from how frequently a symptom appears in tree splits and its contribution to improving decision accuracy.
- The plot is saved as an image file (e.g., feature_importance_plot.png) in the system directory. This visualization aids in transparency and interpretability, helping developers or healthcare professionals understand which symptoms are driving predictions.

IV. RESULTS

The effectiveness of the disease prediction model was assessed using important evaluation metrics such as Accuracy, Precision, Recall, F1 Score, and Average AUC (Area Under the Curve). These metrics help determine the model's reliability and consistency across all 15 disease categories. The obtained results indicate that the model performs efficiently and maintains stable performance with very little variation among different disease classes.

The model achieved an overall accuracy of 97.6%, which means that the majority of the predicted outcomes closely match the actual diagnoses. This high level of accuracy highlights the model's strong predictive capability and confirms its potential usefulness in real-world healthcare applications.

V. CONCLUSION

The disease prediction model demonstrated outstanding performance across all major evaluation metrics, highlighting its strength and dependability in real-world medical applications. With an impressive overall accuracy of 97.6%, and performance ranging from 95.2% to 99.8% across different disease categories, the system consistently produced correct and reliable predictions. This level of accuracy reflects the model's ability to learn complex patterns from medical data effectively. Such consistent results are critical in healthcare environments where even small errors can lead to significant consequences. The model's strong performance indicates that it can serve as a dependable foundation for assisting clinicians in diagnostic decision-making processes.

In addition to high accuracy, the model achieved a precision score of 96.5%, demonstrating its effectiveness in minimizing false positive predictions, which is essential in avoiding unnecessary treatments or anxiety for patients. Its recall score of 97.2% further emphasizes its ability to correctly identify actual disease cases, ensuring that critical conditions are not overlooked. The F1 score of 96.8% reflects a well-balanced trade-off between precision and recall, indicating that the model maintains both sensitivity and specificity. This balance is especially important in healthcare systems, where both overdiagnosis and underdiagnosis can have serious implications for patient outcomes and overall treatment efficiency.

Most notably, the model achieved an average AUC of 0.992, indicating near-perfect class discrimination capability, which is particularly valuable when dealing with complex, multi-class medical datasets. This high AUC score confirms the model's ability to distinguish between different disease categories with exceptional clarity. Furthermore, the consistent performance across various diseases suggests strong generalization , meaning the model can adapt well to new and unseen data. This is a rare and highly desirable trait in healthcare AI systems. Overall, the proposed model proves to be a scalable, accurate, and reliable solution for early disease detection and has strong potential as an advanced clinical decision support tool.

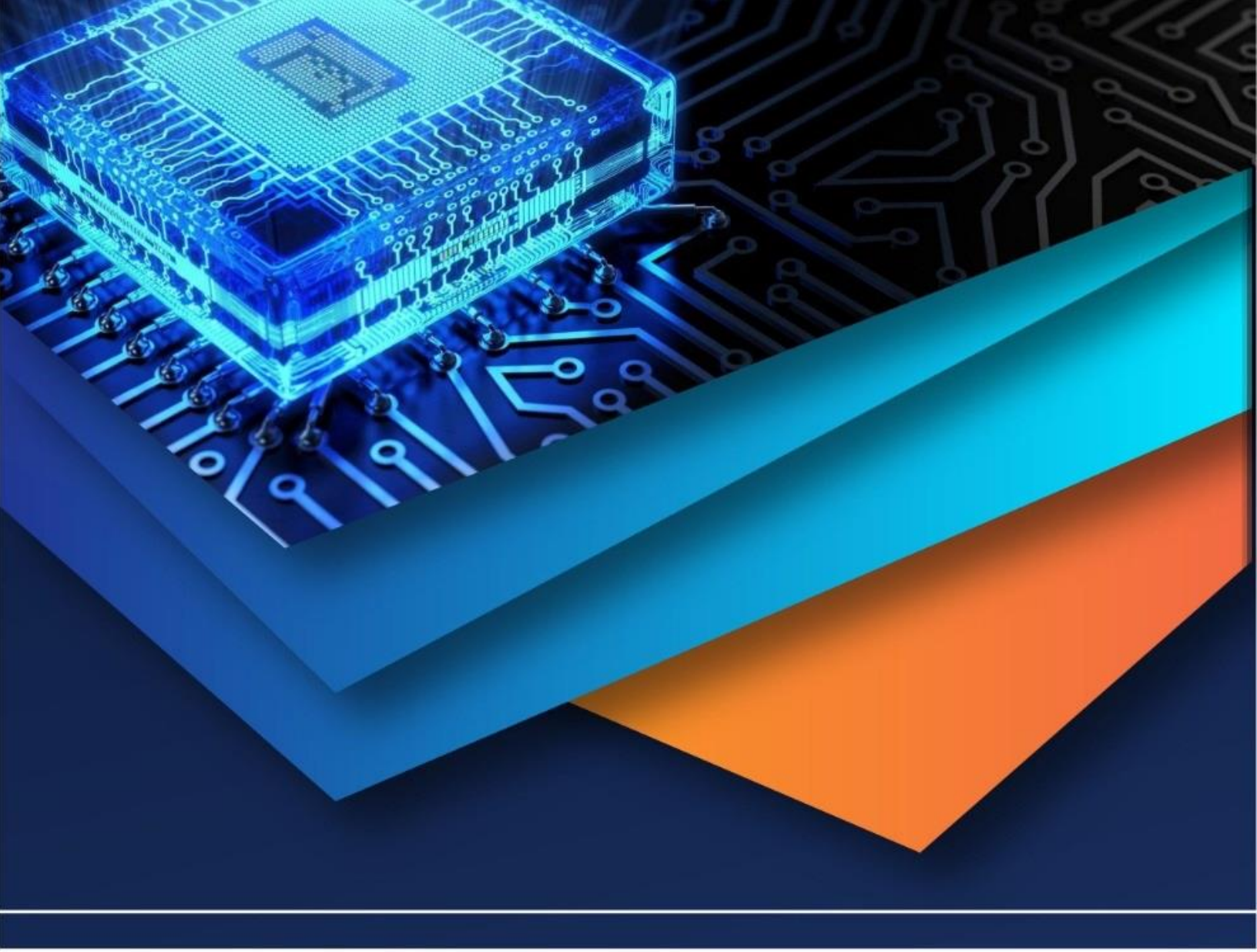
VI. FUTURE SCOPE

Looking forward, the model can be further enhanced by integrating real-time clinical data collected from IoT-enabled devices, wearable technologies, and continuous patient monitoring systems. This integration would allow the model to adapt dynamically to changing health conditions and provide more timely predictions. By incorporating live data streams such as heart rate, oxygen levels, and activity patterns, the system could significantly improve its responsiveness and accuracy. Such advancements would make the model more practical in real-world healthcare settings, enabling proactive disease detection and early intervention, ultimately improving patient outcomes and reducing the burden on healthcare professionals.

Future improvements may also focus on enabling personalized predictions by incorporating individual patient history, genetic factors, and lifestyle habits into the model. Leveraging advanced deep learning techniques, including transformer-based architectures, could further enhance predictive capabilities and pattern recognition. Expanding the dataset to include rare and emerging diseases would strengthen the model's robustness and generalization ability. Additionally, deploying the system through user-friendly mobile or web-based platforms, possibly integrated with voice assistance, could ensure accessibility for remote and underserved communities, making intelligent healthcare solutions more inclusive, scalable, and impactful across diverse populations.

REFERENCES

[1] Aamir, Sanam, et al. "Predicting breast cancer leveraging supervised machine learning techniques." Computational and Mathematical Methods in Medicine 2022 (2022).
[2] J. Amin, M. Sharif, M. Yasmin, T. Saba, and M. Raza, ‘‘Use of machine intelligence to conduct analysis of human brain data for detection of abnormalities in its cognitive functions,’’ Multimedia Tools Appl., vol. 79, nos. 15–16, pp. 10955–10973, Apr. 2020.
[3] Z. Du, Y. Yang, J. Zheng, Q. Li, D. Lin, Y. Li, J. Fan, W. Cheng, X.-H. Chen, and Y. Cai, ‘‘Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine learning methods: Model development and performance evaluation,’’ JMIR Med. Informat., vol. 8, no. 7, Jul. 2020, Art. no. e17257.
[4] El-Hasnony, Ibrahim M., et al. "Multi-label active learning-based machine learning model for heart disease prediction." Sensors 22.3 (2022): 1184.
[5] A. Garg, B. Sharma, and R. Khan, ‘‘Heart disease prediction using machine learning techniques,’’ in Proc. IOP Conf. Ser., Mater. Sci. Eng., vol. 1022, 2021, Art. no. 01204
[6] J. A. W. Gold, F. B. Ahmad, J. A. Cisewski, L. M. Rossen, A. J. Montero, K. Benedict, B. R. Jackson, and M. Toda, ‘‘Increased deaths from fungal infections during the coronavirus disease 2019 pandemic—National vital statistics system, United States, January 2020–December 2021,’’ Clin. Infectious Diseases, vol. 76, no. 3, pp. e255–e262, Feb. 2023.
[7] Humayun, Mamoona, et al. "Framework for detecting breast cancer risk presence using deep learning." Electronics 12.2 (2023): 403.
[8] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, ‘‘Heart disease prediction using machine learning algorithms,’’ in Proc. IOP Conf. Ser., Mater. Sci. Eng., vol. 1022, 2021, Art. no. 012072.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY