IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

Predictive Analytics and Diagnosis for Diseases: A Machine Learning Approach

Prof. Rahul Samant, Prachi Sapkal, Vaishnavi Barsavade, Shravani Rakate
Department of Information Technology Engineering
NBN Sinhgad Technical Institutes Campus, Pune, India

Abstract: *Healthcare systems face significant challenges in early disease detection, misdiagnosis, and accessibility, particularly in resource-limited areas. This project proposes a disease prediction system utilizing Multinomial Naive Bayes (Multinomial NB) and Random Forest Classifier to address these issues. The system analyses patient symptoms and medical data to predict diseases and provide appropriate remedies, promoting timely interventions and better patient outcomes. The Multinomial Naive Bayes algorithm efficiently handles categorical symptom data through probabilistic modelling , making it well-suited for symptom-based inputs. The Random Forest Classifier, employing ensemble learning, enhances the accuracy and robustness of predictions, especially in handling complex datasets and missing data. Together, these algorithms form a reliable and scalable solution designed for real-world healthcare applications. The system follows a structured methodology: collecting and preprocessing medical datasets, implementing algorithms for disease prediction, and developing a user-friendly interface for patients and healthcare professionals. This interface facilitates symptom input, disease prediction, and remedy suggestions. Model evaluation metrics like accuracy and precision ensure the system's performance and scalability. Motivated by the need for accessible healthcare, this project aims to reduce diagnostic errors, improve decision-making, and support healthcare providers with data-driven tools. Future extensions include integrating wearable devices for real-time monitoring, telemedicine platforms for remote consultations, and advanced AI techniques like deep learning to enhance prediction capabilities. By leveraging machine learning, this system addresses critical gaps in healthcare, offering a scalable and efficient solution that enhances diagnostic accuracy, supports early detection, and improves global healthcare accessibility..*

Keywords: Disease prediction, Multinomial Naive Bayes, Random Forest Classifier, early detection, healthcare accessibility, machine learning, probabilistic modelling , ensemble learning, diagnostic accuracy, patient care, symptom analysis, scalable system, real-world healthcare, telemedicine, wearable devices integration

I. INTRODUCTION

The advancement of machine learning and artificial intelligence has opened new avenues for revolutionizing the healthcare industry, particularly in early disease detection and diagnostic accuracy. This project proposes an advanced disease prediction system utilizing Multinomial Naive Bayes (Multinomial NB) and Random Forest Classifier, two robust machine learning algorithms, to address critical challenges in modern healthcare. The system is designed to analyze patient-provided symptom data and associated medical information to predict the likelihood of specific diseases while offering suitable remedies. By focusing on scalability, accuracy, and usability, this solution aims to bridge the gap between medical expertise and accessible, data-driven healthcare services. At the core of the system lies the Multinomial Naive Bayes algorithm, chosen for its efficiency in handling categorical data. This algorithm employs probabilistic modelling to process symptom-based inputs, making it an ideal choice for healthcare datasets characterized by discrete and non-continuous features. In tandem, the Random Forest Classifier, a powerful ensemble learning technique, enhances the system's predictive performance by combining the outputs of multiple decision trees.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

This combination ensures that the system achieves a high degree of accuracy, robustness against noisy data ,and resilience in handling incomplete datasets—a common issue in real-world medical data. The development pipeline includes a rigorous data preprocessing phase, which involves cleaning, encoding, and managing missing values in large medical datasets. This step ensures that the algorithms operate on high-quality input data, thereby optimizing predictive performance. The system also features a user-friendly graphical interface, enabling healthcare professionals and patients to input symptoms and receive detailed predictions with remedy suggestions seamlessly. This project aligns with the broader goal of leveraging machine learning to reduce diagnostic errors, facilitate timely medical interventions, and make advanced healthcare tools accessible across diverse geographical and socioeconomic contexts. Future expansion plans include incorporating wearable device data, telemedicine integration, and advanced AI models like deep neural networks to address complex, multi modal medical datasets. Through these efforts, the project strives to redefine how machine learning can transform modern healthcare.

II. LITERATURE SURVEY

The authors of this study titled "Disease Related Research Using Machine Learning" published in the International Research Journal of Engineering and Technology (IRJET) are Raj H. Chauhan, Daksh N. Naik, Rinal A. Harpa TI, Sagakumar J. Patel, along with Mr. AD Prajapati, focuses on the importance of disease prediction through machine learning. Their system uses a predictive model that uses discrete decision trees to calculate the probability of a disease based on symptoms provided by the user. The significance of this work lies in its applications in the medical field, where early diagnosis and nursing are crucial. This work demonstrates how machine learning and its training and testing phase provide a powerful platform for improving healthcare and improving patient diagnoses, given that it is revolutionizing medicine. The aim of the research is to create a user-friendly system that can predict diseases without having to visit a doctor in person. The system, called “AI Therapist,” incorporates machine learning to improve predictive accuracy and analyze structured and unstructured data. The authors are affiliated with the Department of Computer Engineering, R.N.G. Patel Institute of Technology in Gujarat, India, and highlight the limitations of the medical model and advocate for its adoption in medical decision-making. From machine learning to rapid data analysis, technology is enabling doctors to make informed decisions that will improve patient care. This article provides an overview of its sources and working algorithms, with a special focus on matching comparisons as well as other different decision trees. The results show that random forest matching outperforms other algorithms, demonstrating its ability to encourage people to focus on their health and to accurately predict and even reduce illnesses based on symptoms. The impact extends to the brain. Health issues such as depression and anxiety

III. DESIGN AND ANALYSIS MODEL

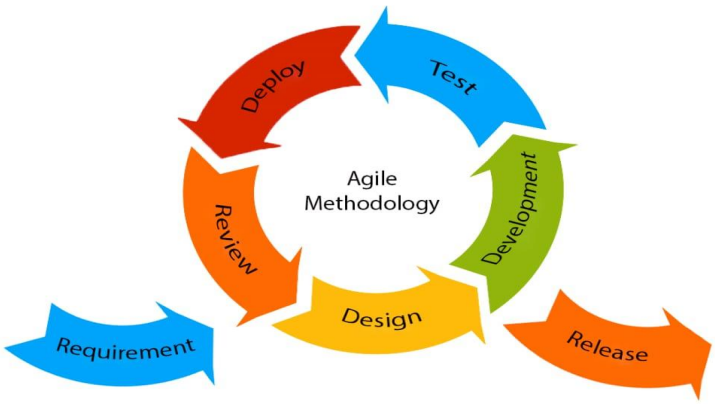


Fig 3.1 Agile methodology

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 10, May 2025 Impact Factor: 7.67

Agile methods have recently become widely accepted in the software world. However, this approach may not work for all products. It is a real approach to software development. It encourages collaboration and cross-training. These features can be developed and demonstrated quickly. Used to edit or change rules. Provides a semi-functional solution at the beginning. Has minimum requirements and easy-to-use information. Also allows products to be developed and delivered in a master planning environment.

IV. METHODOLOGY

The development of the Healthcare Diagnosis and Prediction System follows a structured and technically rigorous approach to ensure its robustness, scalability, and real-world applicability. The methodology is divided into key stages, each addressing critical aspects of the system's design, implementation, and evaluation.

Problem Identification and Requirement Analysis

- Healthcare Challenges: Identify gaps in existing diagnostic systems, such as limited access to resources, high misdiagnosis rates, and delayed treatments
- Stakeholder Needs: Understand the requirements of healthcare professionals and patients, focusing on usability, scalability, and accuracy .
- Objective Definition: Define technical objectives, such as integrating machine learning algorithms and ensuring interoperability with healthcare data formats.

Data Collection and Preprocessing

- Data Acquisition: Gather diverse medical datasets, including patient symptoms, disease diagnoses, and treatment records, ensuring comprehensive coverage of diseases and conditions.
- Data Cleaning: Address missing values, inconsistent entries, and anomalies using techniques like imputation and outlier detection.
- Feature Engineering: Encode categorical data (e.g., symptoms) into numerical formats using methods like one-hot encoding or label encoding. Generate meaningful features that enhance model interpretability.
- Data Normalization: Standardize or normalize numerical features to ensure compatibility with machine learning algorithms.

Algorithm Implementation

- Multinomial Naïve Bayes (MultinomialNB): Designed for categorical data, this algorithm applies probabilistic modelling to classify symptoms into likely disease categories. O It calculates posterior probabilities using Bayes' theorem, assuming feature independence to simplify computations.
- Random Forest Classifier: o A robust ensemble learning algorithm that builds multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting. Handles missing data and complex interactions among features, ensuring resilience against noisy datasets

System Development

- Backend: Implement data processing and machine learning components using Python libraries such as NumPy, Pandas, and Scikit-learn.
- Database Integration: Store patient data, prediction history, and remedy suggestions in a relational database, such as MySQL, ensuring secure and efficient data management .
- Frontend: Develop a user-friendly graphical interface using Flask or Django, allowing healthcare professionals and patients to interact with the system for symptom input and disease prediction.

Model Evaluation and Optimization

- Evaluation Metrics: Use precision, recall, F1 score, and accuracy to evaluate model performance on test datasets.
- Hyperparameter Tuning: Optimize parameters such as the number of trees in the Random Forest and Laplace smoothing in MultinomialNB to enhance performance.
- Cross-Validation: Validate model performance on multiple dataset splits to ensure generalization across unseen data.

Iterative Development and Feedback Integration

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

• Adopt an Iterative SDLC Model, building the system incrementally .
• Conduct real-world testing with domain experts and refine the system based on feedback regarding usability, accuracy, and relevance

Deployment and Scalability
• Local Deployment: Begin with local deployment to evaluate performance under controlled conditions.
• Cloud Integration: Migrate the system to cloud platforms for broader accessibility, enabling remote access for healthcare providers and patients.
• Real-Time Processing: Enable real-time data processing capabilities, integrating wearable devices and IoT sensors for dynamic symptom monitoring.
Future Enhancements
• Incorporate advanced AI models like deep neural networks for handling more complex medical datasets
• Expand functionality to include telemedicine integration, multi-disease prediction, and personalized treatment recommendations.

V. ARCHITECTURAL DIAGRAM

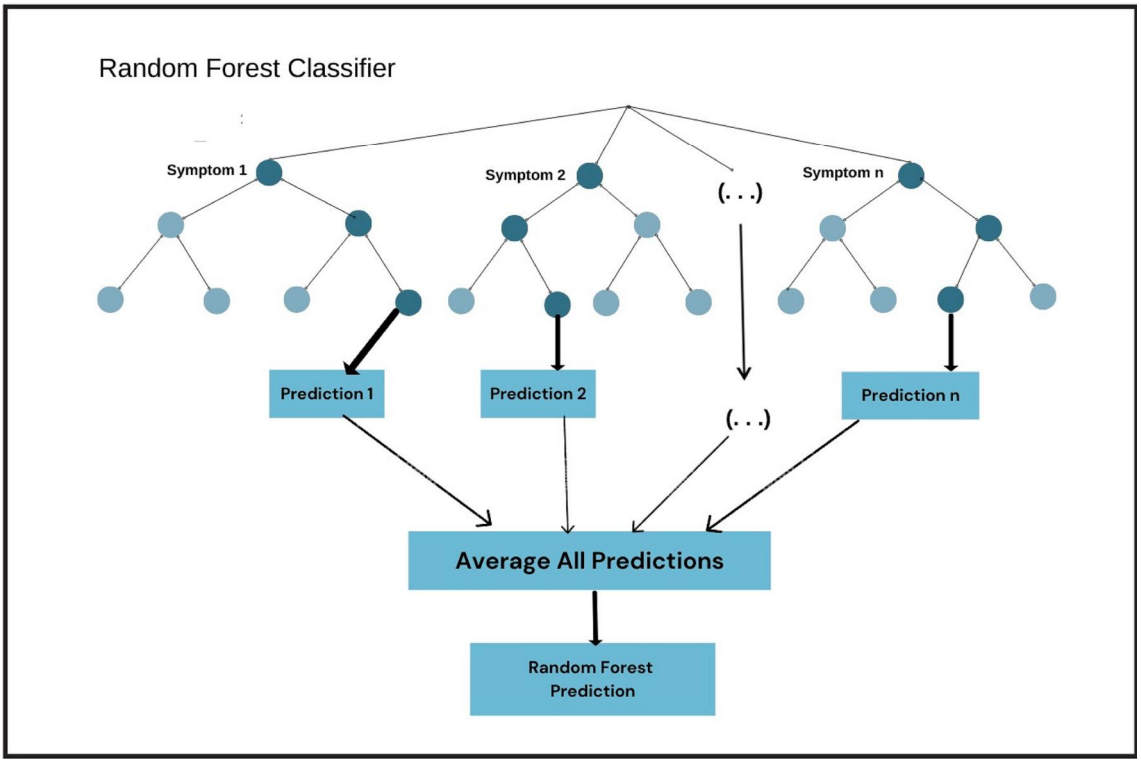


Fig 4.1 Random Forest Classifier

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

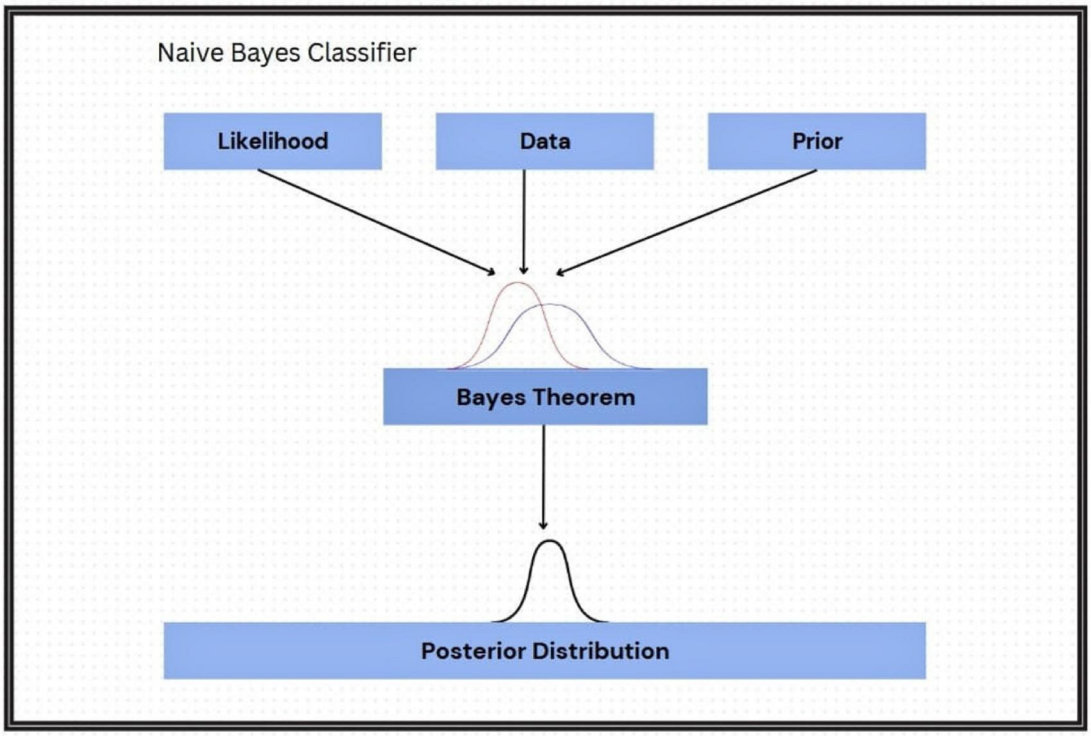


Fig 4.2 Naïve Bayes Classifier

The Random Forest Classifier enhances the system's accuracy and robustness by combining multiple decision trees, managing complex datasets, and effectively handling missing data.

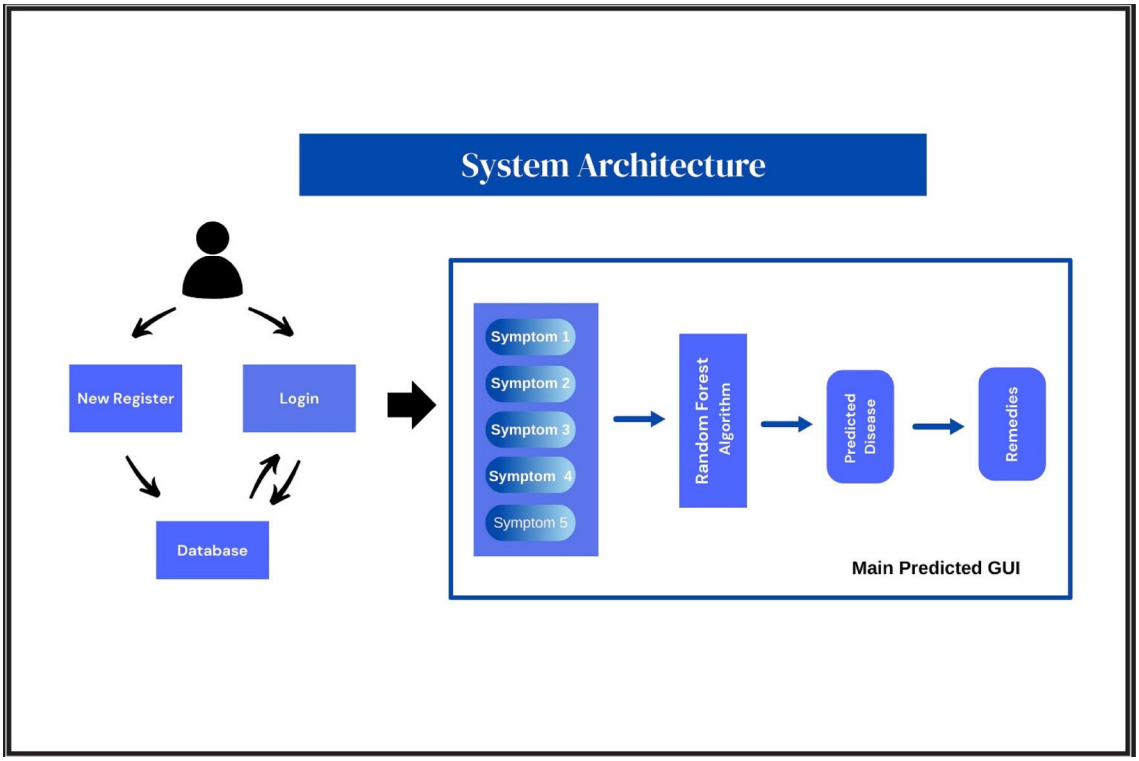
The Naive Bayes Algorithm is employed for its effectiveness in handling categorical data and probabilistic modeling, making it ideal for symptom-based inputs.

V. APPLICATIONS

The use of machine learning in disease prediction is changing the healthcare industry. Leveraging the power of algorithms and data analysis, machine learning is revolutionizing early disease detection, risk assessment, and personalized medicine. It plays a key role in the analysis of clinical data, diagnosis, and genetic information, allowing doctors to detect diseases at the earliest, allowing patients to take risks and self-correct according to their needs. The technology also increases the accuracy and efficiency of medical image analysis, aids drug discovery, and contributes to public health assessment by predicting infectious diseases. Machine learning also helps remotely monitor patients and fraud, and supports epidemiological studies to identify health conditions and track infection. Overall, the use of machine learning in disease prediction creates a future of better healthcare, efficiency and personalization, improved patient outcomes, and managed public health.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

VI. SYSTEM ARCHITECTURE



VIII. CONCLUSION

The Healthcare Diagnosis and Prediction System, powered by Multinomial Naive Bayes and Random Forest Classifier, addresses some of the most pressing challenges in modern healthcare—early disease detection, diagnostic accuracy, and accessibility. By leveraging machine learning techniques, the project offers a transformative approach to improving patient care and outcomes, aligning seamlessly with its underlying motivations.The system fulfills the need for early disease detection by accurately analyzing patient symptoms and medical data, enabling timely interventions. Early diagnosis is a critical factor in reducing the progression of diseases and improving survival rates, particularly in cases where delayed treatments can have severe consequences. This directly supports the motivation of enhancing patient outcomes while reducing the burden on healthcare providers. The use of machine learning algorithms mitigates diagnostic errors, a significant issue in conventional healthcare systems. The probabilistic modeling capabilities of Multinomial Naive Bayes efficiently process symptom-based inputs, while the ensemble learning power of the Random Forest Classifier ensures robustness and precision in handling real-world, noisy datasets. This results in a highly reliable system that minimizes errors and builds confidence among healthcare professionals and patients. By addressing healthcare accessibility, the project serves regions with limited medical resources. The system's scalability ensures its adaptability across different healthcare environments, from local clinics to large hospitals. Its ability to integrate with future technologies, such as wearable devices and telemedicine platforms, extends its reach to underserved populations, fulfilling the motivation to democratize healthcare services. Furthermore, the project promotes technology-driven innovation in healthcare, showcasing how machine learning can revolutionize diagnostics. It provides a cost-effective, scalable, and user-friendly solution, significantly reducing the financial and operational challenges faced by healthcare providers.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

IX. ACKNOWLEDGMENT

It gives us immense pleasure and satisfaction to present our project, "Predictive Analysis and Diagnosis for Diseases : A Machine leaning approach " developed as a part of our academic journey. We express our heartfelt gratitude to Prof. Rahul Samant for his continuous support, insightful feedback, and valuable guidance throughout the course of this project. His expertise in the field and timely encouragement played a crucial role in shaping the direction and successful execution of our work. Our deepest appreciation goes out to our friends, peers, and family members for their unwavering encouragement, moral support, and belief in our potential during all phases of this endeavour. Their presence and motivation were instrumental in helping us stay focused and committed to the successful completion of this project.

REFERENCES

[1]. Chauhan, R. H., Naik, D. N., Halpati, R. A., Patel, S. J., & Prajapati, A. D. (2020).

[2]. Disease prediction using machine learning. This system utilized the decision tree classifier algorithm to predict diseases based on symptoms and aimed to offer better diagnostic accuracy along with motivational content.

[3]. Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., Warang, M., & Mehendale, N. (2020).

[4]. Disease prediction from various symptoms using machine learning. The study used a Weighted KNN model which achieved 93.5% accuracy using symptom, age, and gender data for the prediction of over 230 diseases.

[5]. Nahian, J. A., Kaisar, A., Masum, M., Abujar, S., & Mia, M. J. (2022).

[6]. Identification and prediction of chronic diseases using machine learning approach. The proposed model was compared against Naïve Bayes, decision tree, and logistic regression algorithms. It achieved an accuracy of 95%, using CNN for feature extraction and KNN for distance calculation.

[7]. Arya, A., Sudhanshu, S., & Agarwal, R. (2023).

[8]. Disease prediction based on symptoms. Four different algorithms were used to build the system, achieving up to 95% accuracy and showing high potential for medical applications in the future.

[9]. Suresh, H., & Guttag, J. V. (2021).

[10]. A Framework for Understanding Unintended Consequences of Machine Learning. Communications of the ACM, 64(4), 62–71.