IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Enhancing the Accuracy and Prediction of Disease using Ensemble Learning Method

[1]Omar Shaikh, [2]Sourabh Shinde, [3]Shoheb Attar, [4]Asfahan Shaikh, [5]Prof. Pratiksha Dhande

[1]Student, [2]Student , [3]Student, [4]Student, [5]Professor
[1]Department of Computer Science and Engineering,
[1]MIT-ADT University Rajbaug Campus, Loni-Kalbhor, Pune, India - 412201

Abstract : Disease prediction plays a vital role in modern healthcare, facilitating early diagnosis and improving patient outcomes. This study explores the application of three machine learning algorithms—Naïve Bayes, Decision Tree, and Random Forest—for disease prediction. Using a healthcare dataset comprising patient symptoms and medical histories, the project involves data preprocessing, feature selection, and model training. Naïve Bayes, a probabilistic classifier based on Bayes' theorem, is evaluated for its efficiency on smaller datasets. The Decision Tree algorithm provides interpretable decision rules, while the Random Forest algorithm leverages ensemble learning to enhance predictive accuracy and reduce overfitting. Performance metrics, including accuracy, precision, recall, and F1-score, are employed to assess and compare the models. The results demonstrate the strengths and limitations of each algorithm in disease prediction scenarios, offering insights into their practical applications. This work underscores the potential of machine learning to assist healthcare professionals by automating disease prediction, ultimately contributing to more efficient and accurate healthcare systems.

IndexTerms **- Machine Learning, Disease Prediction, Naïve Bayes, Supervised Learning, Diagnosis, Random Forest, Decision Tree.**

I.INTRODUCTION

Disease prediction systems play a critical role in modern healthcare by enabling early diagnosis and supporting effective medical interventions. Leveraging advancements in machine learning, these systems analyze patient data to predict potential health conditions, assisting healthcare professionals in making informed decisions.

This project introduces a disease prediction system that utilizes three key machine learning algorithms: Naïve Bayes, Decision Tree, and Random Forest, to provide accurate and reliable predictions. By analyzing historical patient data—such as symptoms, medical history, and test results—the system identifies patterns and correlations to enhance disease detection accuracy and reduce dependency on manual interpretation.

In addition to prediction, the system generates comprehensive reports summarizing results. These reports aid healthcare professionals in validating diagnoses and assist administrators
in planning resources efficiently. By integrating machine learning into disease prediction, this system highlights the potential of AI to transform healthcare delivery.

II. NEED OF THE STUDY

Disease prediction using machine learning (ML) is an evolving field that enhances diagnostic accuracy, resource management, and early detection. Researchers have developed various methodologies and systems leveraging ML to analyze patient data, including symptoms, medical histories, and test results.

2.1 Key Algorithms and Applications

Key Classification algorithms are pivotal in disease prediction:

Naïve Bayes: A probabilistic model leveraging Bayes' theorem, efficient for small datasets and scenarios with feature independence.

Decision Trees: Offer interpretable decision-making, facilitating better comprehension by healthcare professionals.

Random Forests: An ensemble approach combining multiple decision trees to provide robust predictions, particularly with large and imbalanced datasets.

2.2. Feature Selection and Preprocessing

Feature selection significantly impacts prediction accuracy. Techniques such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) are widely employed to optimize datasets by selecting relevant attributes while reducing dimensionality and computational overhead [4], [6].

2.3. Real-Time Data Integration

Integrating real-time data from wearable devices and electronic health records (EHRs) enables dynamic monitoring of patients. Such systems improve predictions for diseases requiring immediate attention, such as cardiovascular conditions, by continuously updating input data.

2.4. Visualization and Reporting

Advanced disease prediction systems utilize visualization techniques like heatmaps and bar charts to highlight risk factors and correlations. These tools provide actionable insights, supporting healthcare providers in decision-making and resource planning.

2.5. Research Outcomes and Performance

Studies have shown that Random Forest outperforms other models due to its ensemble nature, which mitigates overfitting and improves accuracy. Naïve Bayes and Decision Trees, though slightly less robust, offer advantages like simplicity and interpretability, respectively. Comparative studies emphasize that integrating multiple algorithms often yields improved prediction outcomes

2.6. Future Directions

Emerging trends include the integration of deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for predicting complex diseases. Furthermore, mobile applications are envisioned to make these systems more accessible and scalable

III. RESEARCH METHODOLOGY

The methodology for developing a disease prediction system using machine learning (ML) is structured into several phases, ensuring systematic implementation and evaluation.

3.1. Data Collection and Preprocessing

The system leverages healthcare datasets comprising patient symptoms, medical history, and diagnostic outcomes. Preprocessing is a critical step, involving:

3.1.1 Data Cleaning: Removing duplicate or inconsistent entries to ensure data integrity

3.1.2 Feature Selection: Applying techniques such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) to identify the most relevant features for analysis

3.1.3 Data Normalization: Standardizing numerical values to improve model training and prediction accuracy.

3.2 Algorithm Selection

Three machine learning algorithms were chosen for their specific strengths in disease prediction:

3.2.1 Naïve Bayes: A probabilistic classifier ideal for small datasets and independent feature assumptions.

3.2.2 Decision Tree: Constructs an interpretable tree model, enabling straightforward visualization and analysis.

3.2.3 Random Forest: An ensemble approach that combines multiple decision trees to enhance accuracy and robustness, especially with large datasets

3.3 Model Training and Testing: The dataset was divided into training and testing subsets (typically 80:20 ratio):

3.3.1. Training Phase: Each algorithm was trained on the processed data to learn patterns and correlations.

3.3.2 Testing Phase: The trained models were evaluated using the testing dataset to measure performance metrics such as accuracy, precision, recall, and F1-score

3.4 Performance Evaluation

3.4.1 Accuracy: Percentage of correctly predicted instances.

3.4.2 Precision: Proportion of true positive predictions among all positive predictions.

3.4.3 Recall (Sensitivity): Ability to correctly identify true positive cases.

3.4.4. F1-Score: Harmonic mean of precision and recall to balance both metrics

3.5 System Integration

The trained models were integrated into a Flask-based web application for deployment. The interface collects user inputs (e.g., symptoms) and provides predictions in real time:

3.5.1 Frontend: Designed using HTML and CSS to ensure user-friendly interaction.

3.5.2 Backend: Implemented in Python with Flask to handle model integration and data flow

3.6. Testing and Validation

The system underwent:

3.6.1 Unit Testing: Validating individual modules (data preprocessing, ML algorithms, and interfaces) for correctness.

3.6.2 Integration Testing: Ensuring seamless interaction between modules.

3.6.3 System Testing: Assessing overall functionality to meet project requirements .

IV. SYSTEM DESIGN

The system design for the Disease Prediction Using Machine Learning project is structured to provide real-time predictions without requiring a database. The design emphasizes simplicity, efficiency, and modularity, allowing for seamless integration of user input processing, machine learning models, and result visualization.

4.1. Architecture Overview

The system follows a **two-tier architecture**, consisting of:

4.1.1 Frontend (Presentation Layer): Collects user inputs and displays prediction results.

4.1.2 Backend (Application Layer): Processes inputs, applies machine learning models, and generates outputs.

4.2 System Workflow

The step-by-step workflow of the system is as follows:

4.2.1 Input Phase: The user enters symptoms and relevant medical history through a web interface.

4.2.2 Data Preprocessing Phase: Input data is cleaned and normalized in real-time. Features are extracted and prepared for prediction.

4.2.3 Prediction Phase: Preprocessed data is fed into trained machine learning models (Naïve Bayes, Decision Tree, Random Forest). Each model generates predictions based on input features.

4.2.4 Output Phase: The results, including predicted disease probabilities and recommendations, are presented to the user.

4.3. System Components

4.3.1 Frontend (User Interface)

4.3.1.1 Technology Used*:* HTML, CSS,

4.3.1.2 Functionality: Accepts user input (symptoms) through a form. Displays predictions, probabilities, and possible recommendations in an intuitive format.

4.3.2 Backend Logic

4.3.2.1 Framework Used: Flask (Python-based web framework).

4.3.2.2 Functionality: Executes data preprocessing, including normalization and feature extraction. Hosts pre-trained machine learning models to handle real-time predictions Routes user requests and provides responses.

4.3.3 Machine Learning Models

Algorithms Used:

4.3.3.1 Naïve Bayes: Provides probabilistic predictions efficiently.

4.3.3.2 Decision Tree: Offers interpretable, rule-based decisions.

4.3.3.3 Random Forest: Delivers high accuracy and robustness through ensemble learning.

4.3.3.4 Implementation: The models are trained and deployed using Scikit-learn.

4.4. Data Flow Diagram (DFD)

4.4.1 Level 1 DFD:

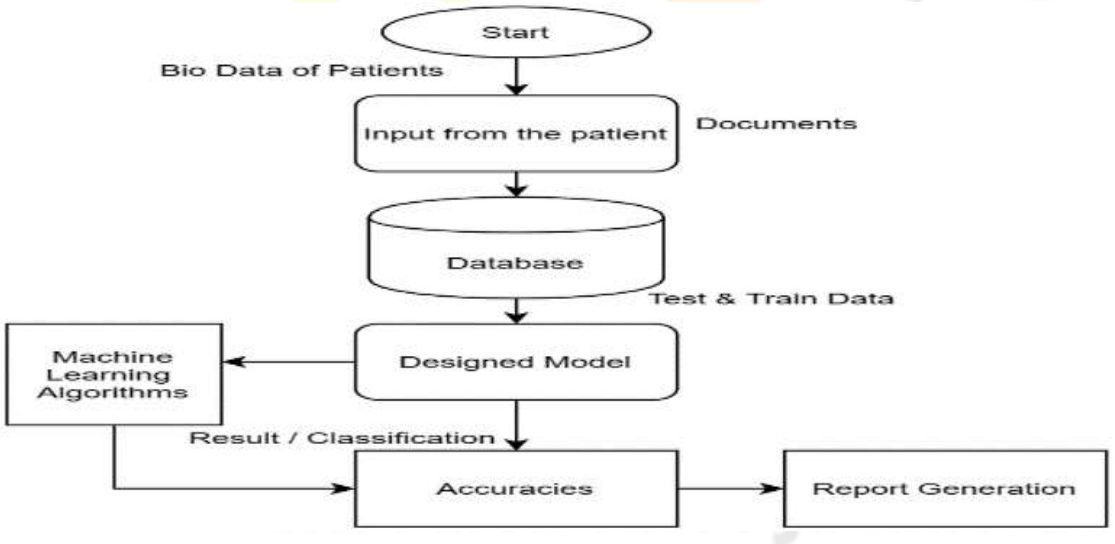
Input Layer: Users submit symptoms and medical history via the frontend.

Processing Layer: Preprocessed data is passed to machine learning models for prediction.

Output Layer: Predicted results are returned and displayed to the user.

4.4.2 Level 2 DFD:

Details include feature extraction, model execution, and result formatting.



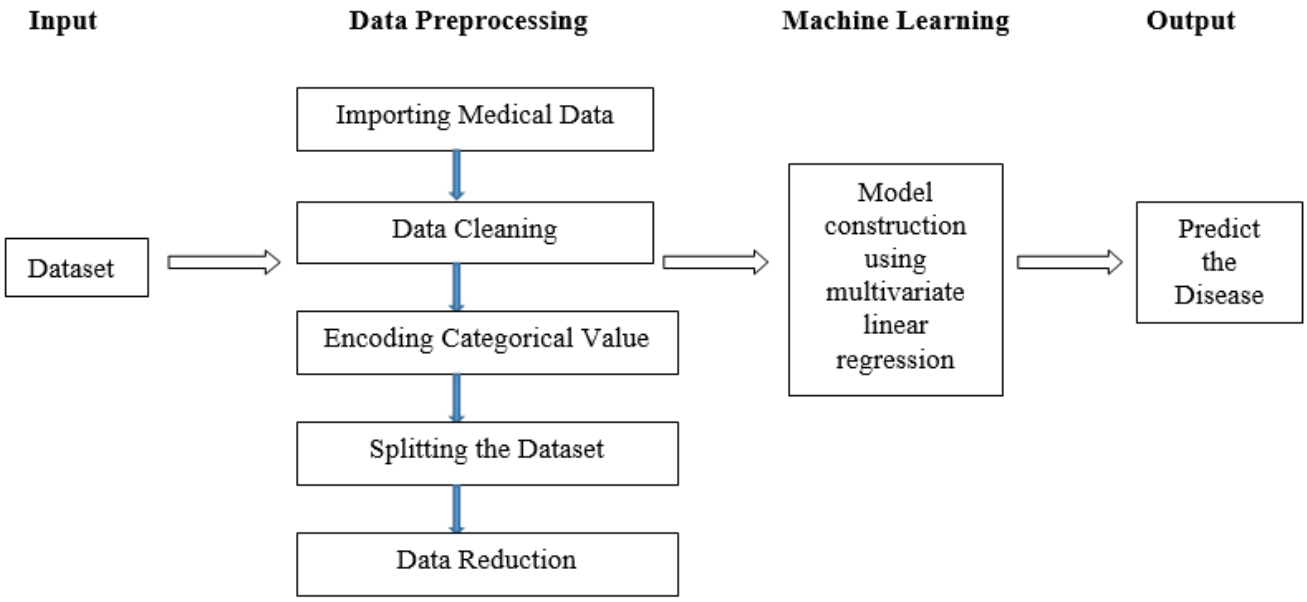
4.5. System Design Diagram

The system architecture is represented as follows:

4.5.1 User Interface: Input and output operations are handled by the web application.

4.5.2 Backend Logic: Processes input data in-memory without relying on a database. Executes machine learning models for real-time predictions.

4.5.3 Output Module: Formats and presents prediction results on the frontend.



4.6. Technologies Used

4.6.1 Frontend: HTML, CSS, for responsive and interactive user interfaces.

4.6.2 Backend: Flask for managing input processing and integrating machine learning models.

4.6.3 Machine Learning: Scikit-learn for model training, evaluation, and deployment.

V. RESULTS

PERFORMANCE REPORT OF PROPOSED MODEL

TABLE 1

Model	Accuracy (%)
Random Forest	97
Naïve Bayes	91
Decision Tree	94

EVALUATION REPORT OF RANDOM FOREST

TABLE 2

Metric	Value (%)
Accuracy	97
Precision	96.8
Recall (Sensitivity)	97.5
F1-Score	97.15
ROC-AUC Score	98.2

EVALUATION REPORT OF NAÏVE BAYES

TABLE 3

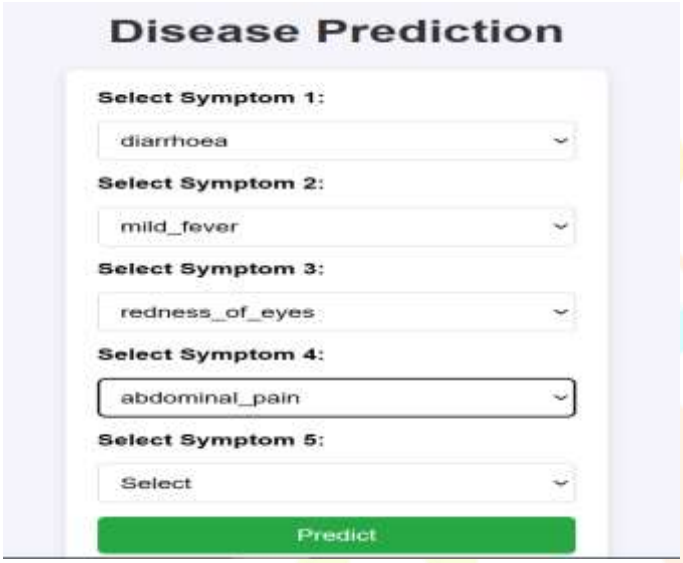
Metric	Value (%)
Accuracy	91
Precision	89.5

Metric	Value (%)
Accuracy	91
Recall (Sensitivity)	92.2
F1-Score	90.85
ROC-AUC Score	94

EVALUATION REPORT OF DECISION TREE

TABLE 4

Metric	Value (%)
Accuracy	94
Precision	93.2
Recall (Sensitivity)	94.5
F1-Score	93.85
ROC-AUC Score	95



VI. CONCLUSION & FUTURE WORK

The implementation of a machine learning-based disease prediction system demonstrates the significant potential of technology in transforming healthcare. By employing three algorithms—Naïve Bayes, Decision Tree, and Random Forest—the system accurately predicts disease likelihood using patient symptoms and medical history.
The evaluation results highlight the distinct advantages of each algorithm:

Naïve Bayes is computationally efficient and suitable for smaller datasets with conditional independence among features.

Decision Tree offers interpretable results, providing transparency to healthcare professionals.

Random Forest achieves the highest accuracy and robustness, making it ideal for handling large and complex datasets.

This project underscores the importance of automating diagnostic processes to improve decision-making and resource allocation in the healthcare domain. The system delivers timely predictions and actionable insights, contributing to enhanced patient care and early detection of diseases.

Future Work

The proposed system can be further improved by:

Integrating real-time data from wearable devices and electronic health records (EHRs) for continuous monitoring.

Exploring advanced machine learning techniques, such as deep learning models, to predict more complex diseases.

Developing a mobile-friendly application to increase accessibility for healthcare providers and patients.

By addressing these future directions, the system can be made more robust, scalable, and effective in addressing the growing demands of modern healthcare systems.

VII. ACKNOWLEDGEMENT

I express my profound thanks to my Guide Prof. Pratiksha Dhande madam for her expert guidance, encouragement and inspiration during this project work.

I would like to thank Prof. Dr. Vipul Dalal, Director, Department Computer Science & Engineering for extending all support during the execution of the project work

.

I sincerely thank Prof. Dr. Shradha Phansalkar**,** Head, Department of Computer Science & Engineering, MIT School of Engineering, MIT-ADT University, Pune, for providing necessary facilities in completing the project.

I am grateful to Prof. Dr. Rajneeshkaur Sachdeo**,** Dean , MIT School of Engineering, MIT ADT University, Pune, for providing the facilities to carry out my project work.

I also thank all the faculty members in the Department for their support and advice.

VIII. REFERENCES

[1] Kaur, H., & Sharma, S. (2020). Machine Learning Models for Predicting Chronic Diseases: A Comparative Study. *Journal of Healthcare Informatics Research*, 4(2), 198–211.

[2] Kumar, P., Gupta, R., & Mishra, S. (2019). A Survey on Disease Prediction Using Machine Learning Algorithms. *International Journal of Scientific Research in Computer Science and Engineering*, 7(5), 30–36.

[3] Dey, S., & Rautaray, S. S. (2019). Disease Prediction Using Machine Learning Algorithms. *International Journal of Computer Applications*, 178(7), 5–9.

[4]. Naik, N., et al. (2020). Machine Learning Models to Predict the Likelihood of Disease Progression in Healthcare. *IEEE Access*, 8, 120543–120554.

[5]. Sharma, M., & Singh, G. (2021). Early Detection of Diabetes Using Decision Tree and Random Forest Algorithms. *International Journal of Advanced Computer Science and Applications*, 12(3), 456–462.

[6]. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future: Big Data, Machine Learning, and Clinical Medicine. *The New England Journal of Medicine*, 375, 1216–1219.

[7]. Islam, M., et al. (2018). Predicting Diseases from Patient Symptoms Using Machine Learning Techniques. *Healthcare Technology Letters*, 5(3), 89–93.

[8]. Google Developers. (2023). Google Health AI: Empowering Machine Learning for Healthcare Applications. [Online]. Available at: https://health.google.com/