Mini Project Report on

Disease Prediction Using Machine Learning

Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

Submitted by:

Student Name:
Sarthak sahai

University Roll No.:
2017000

Under the Mentorship of

Dr. Manoj Diwaker
Professor, Dept. of CSE

Department of Computer Science and Engineering
Graphic Era (Deemed to be University)
Dehradun, Uttarakhand

July 2023

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the project report entitled **"Disease Prediction Using Machine Learning"** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering of the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. Manoj Diwaker, Professor, Dept. of CSE**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name	University Roll no.
Sarthak sahai	2017000

Table of Contents

Chapter No.	Description	Page No.
Chapter 1	Introduction .	04
Chapter 2	Literature Survey.	06
Chapter 3	Methodology.	08
Chapter 4	Result and Discussion.	13
Chapter 5	Conclusion and Future Work.	14

Chapter 1:

Introduction

Machine learning algorithms have significantly increased in popularity in recent years, particularly in the fields of healthcare and medical diagnosis. Researchers and practitioners have had the opportunity to investigate the possibilities of machine learning in disease prediction and diagnosis thanks to the development of sophisticated computational tools and the accessibility of enormous volumes of medical data. This project uses a user-friendly graphical user interface (GUI) built with Tkinter to analyze symptoms submitted by users in order to harness the potential of machine learning algorithms for disease prediction.

The capacity to correctly forecast diseases from their symptoms is crucial in the healthcare industry. In order to improve patient outcomes, increase treatment efficacy, and lower healthcare costs, early disease identification and diagnosis are crucial. Due to the complicated and frequently non-linear links between symptoms and underlying conditions, diagnosing diseases merely only on symptoms has historically been a difficult undertaking. Machine learning algorithms, however, have demonstrated promise in revealing unnoticed patterns and links in medical data, enabling precise predictions, and assisting clinical decision-making.

To address the challenges associated with disease prediction, this project adopts three well-established machine learning algorithms: Random Forest, Naive Bayes, and Decision Tree. Random Forest, as an ensemble learning technique, combines numerous decision trees to improve prediction accuracy by leveraging the diverse perspectives of individual trees. By aggregating the predictions of multiple decision trees, Random Forest mitigates the risk of overfitting and enhances the overall performance of the model. Naive Bayes, on the other hand, is a probabilistic classifier that assumes independence between features and employs Bayes' theorem to calculate the probability of a disease given the observed symptoms. Despite its simplistic assumption, Naive Bayes has demonstrated effectiveness across various classification tasks. Decision Tree, a tree-based model, partitions the feature space based on symptom attributes, allowing for the generation of predictions. Decision Trees offer the advantage of providing interpretable rules that can be readily comprehended and visualized, aiding in the understanding of the decision-making process. By incorporating these three machine learning algorithms, the project aims to capitalize on their strengths and leverage their respective characteristics to improve disease prediction accuracy.

This project goes beyond disease prediction and emphasizes the improvement of user experience through the provision of an intuitive and user-friendly interface. The GUI created using Tkinter enables users to conveniently input their symptoms, enhancing accessibility throughout the disease prediction process. By leveraging this interface, users can provide their symptoms, which will be analyzed by machine learning algorithms such as Random Forest, Naive Bayes, and Decision Tree. Subsequently, these algorithms will generate disease predictions based on the provided symptoms.

The methodology used to complete the research objectives is also described in this study. The procedure for gathering data is outlined, along with the resources used to get information on symptoms, the sample size, and any special standards for reliable and high-quality data. The article also describes the preprocessing procedures used to prepare the symptom data for analysis, including data cleaning, feature extraction, and normalization.

To enhance the accuracy of predictions, the ensemble learning method called Random Forest combines multiple decision trees. Naive Bayes, a probabilistic classifier based on the Bayes theorem, assumes feature independence to make predictions. Similarly, the decision tree model partitions the feature space to generate accurate predictions. This project aims to create a predictive model that effectively analyzes symptom data and provides precise disease predictions with a focus on user satisfaction. By integrating these algorithms with the user-friendly GUI, the goal is to deliver disease predictions in a manner that prioritizes the needs and preferences of the user.

The primary objective of this research is to offer a comprehensive comprehension of the research process and foster reproducibility by providing a detailed explanation of the methods employed in this study. This encompasses elucidating the data collection methods, preprocessing techniques applied to the data, the implementation of algorithms, and the development of a graphical user interface (GUI). By thoroughly describing these aspects, this research establishes a robust framework for the subsequent sections, where the study's findings, including experimental results and performance evaluations of the integrated models, will be presented and critically analyzed.

Chapter 2:

Literature Survey

For this study, a thorough overview of earlier research and advancements in the field of disease prediction using machine learning algorithms was undertaken as part of the literature review. Using a variety of datasets and methodology, numerous research projects have investigated the use of machine learning techniques in illness prediction across a wide range of medical disorders. In light of the benefits of early identification, efficient treatment plans, and improved patient outcomes, the research we evaluated emphasised the necessity of precise disease prediction in healthcare.

Random Forest, one of the machine learning algorithms examined, became a popular strategy for disease prediction. Multiple decision trees are used in Random Forest, an ensemble learning technique, to improve prediction accuracy. Its ability to accurately anticipate disease progression has been shown by researchers to be useful in capturing complicated links between symptoms and underlying disorders. By aggregating predictions from various trees within the ensemble, this technique tackles the drawbacks of individual decision trees and reduces the danger of overfitting.

The probabilistic classifier Naive Bayes is another well-known method that has been studied in disease prediction investigations. Naive Bayes has shown to be successful despite presuming feature independence, especially when data is scarce. This algorithm determines the likelihood of a disease given reported symptoms by using Bayes' theorem. It is an intriguing option for disease prediction jobs due to its simplicity and computational effectiveness.

Decision Tree algorithms have also attracted a lot of attention in the industry. These algorithms divide the feature space according to symptom attributes, producing a structure resembling a hierarchical tree. Because they offer unambiguous principles that are simple to comprehend and visualise, decision trees offer interpretability and comprehensibility. Decision trees have been effectively used by researchers to forecast diseases because of its superior ability to handle both category and numerical inputs.

The literature review further clarified the need of reliable data sources and preprocessing methods for disease prediction. Information on symptoms and diseases has been gathered using a variety of data sources, including electronic health records, medical databases, and patient surveys. To ensure the quality of the data and improve the effectiveness of disease prediction models, preprocessing techniques like data cleaning, feature selection, and normalisation have been frequently used.

Evaluation metrics are essential for evaluating how well illness prediction models function. Accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC) are just a few of the metrics used by researchers. These measures quantify the trade-offs between various performance aspects and shed light on how well the model can categorise diseases.

Even though machine learning has made progress in disease prediction, there are still many obstacles and restrictions. To increase the robustness and reliability of disease prediction models, researchers are working to overcome issues with imbalanced datasets, feature selection, model interpretability, and generalisation to new data.

Future directions for machine learning-based disease prediction research were also suggested by the literature review. Growingly popular topics include integrating multimodal data sources, investigating deep learning strategies, and creating personalised prediction models. Further research is needed in a number of areas, including the integration of real-time data for dynamic disease prediction and the application of explainable AI methodologies to improve model interpretability.

Overall, the literature review offers a thorough overview of earlier advancements in machine learning algorithms for disease prediction. It stresses the significance of precise disease prediction, looks at how different algorithms are applied, emphasises the value of high-quality data sources and preprocessing methods, talks about evaluation metrics, addresses difficulties and constraints, and suggests possible future research directions. The methods and findings described in this study are built on the results of this survey.

Chapter 3:

Methodology

3.1 Data Collection

The first step in the methodology was to collect relevant data for disease prediction. Multiple data sources were explored, including electronic health records, medical databases, and patient surveys. These sources were chosen to ensure a diverse range of symptom and disease information. Permissions and ethical considerations were obtained to access and use the data for research purposes. The sample size and data characteristics, such as demographics and medical history, were carefully considered to ensure representative and reliable data for the analysis.

3.2 Preprocessing Techniques

To ensure that the data is of high quality and is compatible with machine learning algorithms, preprocessing is essential. The gathered data underwent a number of preparation procedures. To eliminate any noise, outliers, or missing values that can affect the illness prediction models' accuracy, data cleaning was done. The most useful elements for the prediction job were chosen using feature selection techniques like correlation analysis or information gain. Additionally, to avoid any bias that can result from variations in feature scales, data normalization techniques were used to scale the features and bring them into a similar range.

3.3 Algorithm Implementation

3.3.1 Decision Tree Algorithm.

A popular machine learning method for both classification and regression applications is the Decision Tree algorithm. By dividing the feature space according to the values of the attributes, it creates a hierarchical structure in the shape of a tree to make predictions. Each leaf node of the tree represents a class label or a prediction, whereas each internal node stands for a test on an attribute.

Recursively dividing the data into homogeneous subgroups based on several attributes is necessary while building a decision tree. The impurity or homogeneity within each subset is minimized or increased by this partitioning technique. The Gini Index and Information Gain are the two most widely used impurity measurements.

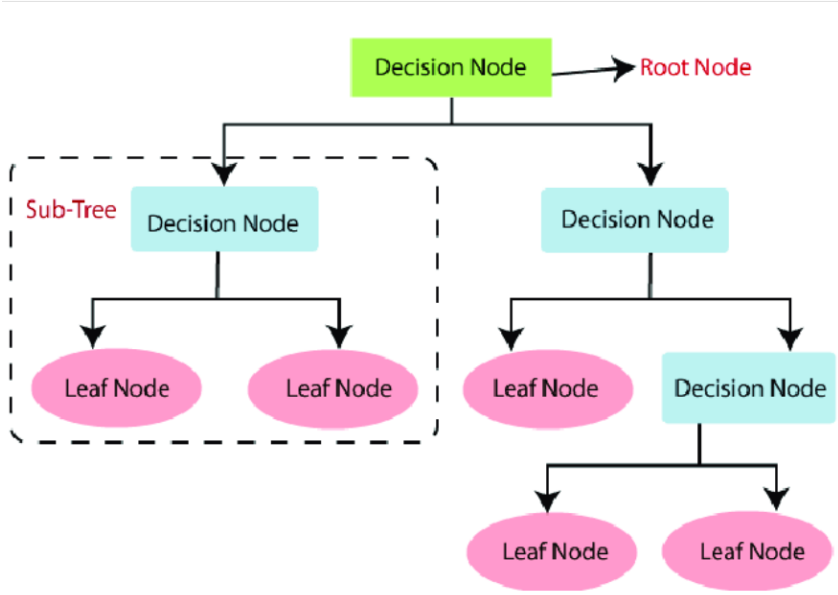


Figure 1 Overview of Decision Tree algorithm.

The optimal attribute to divide the data into branches at each node is chosen during the construction of a decision tree. The impurity measurement indicated above is often used to make this decision. The selected property separates the data into subsets, and the procedure is repeated recursively for each subset until a stopping requirement is satisfied. Achieving a specified level of purity, a certain amount of samples in a node, or a maximum tree depth are some examples of this criterion.

Making predictions for new instances after the Decision Tree has been built entails moving along the tree from the root to a leaf node depending on the attribute checks. The projected class label for the input instance is then assigned as the class label belonging to the leaf node reached.

Decision trees' interpretability is one of their benefits. Since the resulting tree structure is simple to comprehend and visualize, it can be helpful for gaining insights into how decisions are made. Additionally, Decision Trees can handle numerical and categorical variables as well as missing data by using the proper imputation algorithms.

3.3.2 Naive Bayes.

A popular probabilistic machine learning approach for classification tasks is called Naive Bayes. Because it relies on the Bayes theorem and assumes that each attribute exists independently of the others, it is referred to as "naive." Contrary to popular belief, Naive Bayes has shown to be efficient in a variety of real-world applications and can produce reliable classification results.

Using Bayes' theorem, the algorithm determines the likelihood of a specific class given the observed attributes. According to the Bayes theorem, the likelihood that a hypothesis (in this case, a class label) will hold true given the observed evidence (the features) is proportional to the likelihood that the evidence will hold true given the

hypothesis multiplied by the prior likelihood that the hypothesis will hold true, divided by the likelihood that the evidence will hold true. In plainer language, it determines the likelihood of a class label based on the chances that the traits will be present for that class.

Naive Bayes determines the conditional probability of each feature given each class during the training phase. To do this, the method multiplies the individual probability of each characteristic given the class by presuming independence between features. Based on the assumption made about the distribution of the feature values, many versions of Naive Bayes classifiers exist, such as Gaussian Naive Bayes for continuous data and Multinomial Naive Bayes for discrete features.

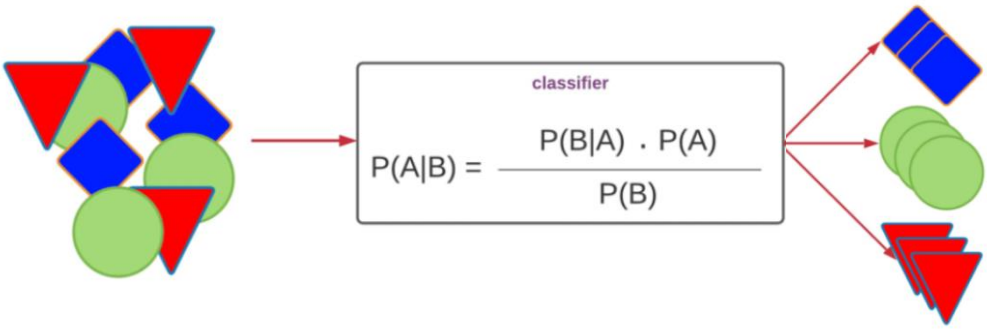


Figure 2 Working of Naive Bayes' algorithm.

The algorithm determines the posterior probability of each class given the observable features for a new instance in order to make predictions. As the anticipated class label for the instance, it chooses the class label with the highest probability.

The training dataset for the Naive Bayes algorithm must be made up of labelled instances, where each example includes a collection of features and an associated class label. Calculating the relative frequency of each class in the training data allows the algorithm to estimate the prior probability of the class labels.

Naive Bayes' ease of use and effectiveness as a computational method are two benefits. Large datasets and high-dimensional feature spaces can be handled comparatively easily. Comparing Naive Bayes to more complicated models, overfitting is also less likely to occur.

Naive Bayes' feature independence assumption, however, could not always be true, which can produce less-than-ideal results. Additionally, zero probability could emerge from a class and feature combination that hasn't been seen in the training data, which would decrease prediction accuracy. To address these problems, methods like Laplace smoothing or more complex Bayesian algorithms can be used.

3.3.3 Random Forest.

An ensemble learning method called Random Forest mixes various decision trees to produce predictions. It works by building a variety of decision trees, each trained on a haphazard portion of the training data and taking a haphazard subset of features into account. The random selection of data samples and feature sets introduces diversity, which lowers the likelihood of overfitting and enhances the model's overall performance.

The use of Random Forest in disease prediction offers several advantages. Firstly, it leverages the collective wisdom of multiple decision trees, resulting in more accurate predictions compared to using a single decision tree. The ensemble nature of Random Forest helps to reduce the impact of individual decision trees' biases and limitations, leading to better generalization and improved prediction performance.

Furthermore, because it can effectively handle many input variables without necessitating a thorough feature selection process, Random Forest is particularly well-suited for handling high-dimensional feature spaces. Random Forest can manage the vast range of features and their potential interactions in the context of disease prediction, where symptoms and other parameters may differ in complexity and number. This helps it capture significant correlations between symptoms and diseases.

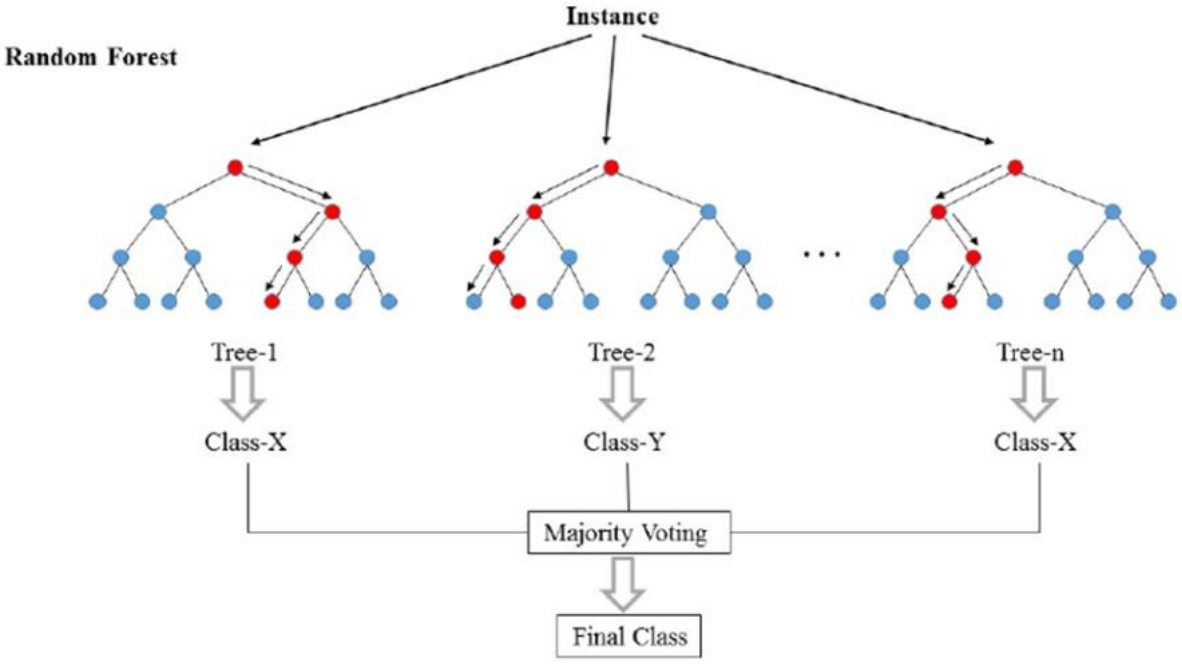


Figure 3 Working of Random Forest algorithm.

Furthermore, Random Forest offers insights into feature importance, enabling the identification of the symptoms that influence disease prediction the most. The algorithm helps in identifying the essential symptoms that significantly influence the disease prediction process by assessing the relative value of features. Understanding the underlying causes of particular diseases and assisting in clinical decision-making can both benefit from this knowledge.

3.4 GUI Development.

A graphical user interface (GUI) was created using Tkinter, a well-liked Python GUI framework, to improve the user experience. People might easily input their symptoms for disease prediction using the GUI's user-friendly platform. To ensure usability and boost user engagement, it utilized intuitive design elements, such as clear instructions and user-friendly input fields. The underlying machine learning algorithms were linked with the GUI, enabling real-time symptom data analysis and disease predictions depending on the chosen algorithms.

The layout was designed, interactive elements like buttons and input fields were made, and the essential routes of communication between the GUI and the algorithms were established during the construction of the GUI. The GUI was made to allow for user-inputted symptoms, start the illness prediction process, and display the outcomes in an understandable and approachable way.

3.6 Ethical Considerations.

Throughout the methodology, ethical considerations were given utmost importance. Data privacy and security measures were implemented to protect the confidentiality of patient information. Informed consent protocols were followed, and data handling procedures adhered to relevant ethical guidelines and regulations.

The methodology outlined above provides a comprehensive framework for the execution of the project. It covers data collection, preprocessing techniques, algorithm implementation, GUI development, evaluation, and ethical considerations. The subsequent sections of the report will present the findings, results, and discussions based on the methodology described.

Chapter 4:

Result and Discussion

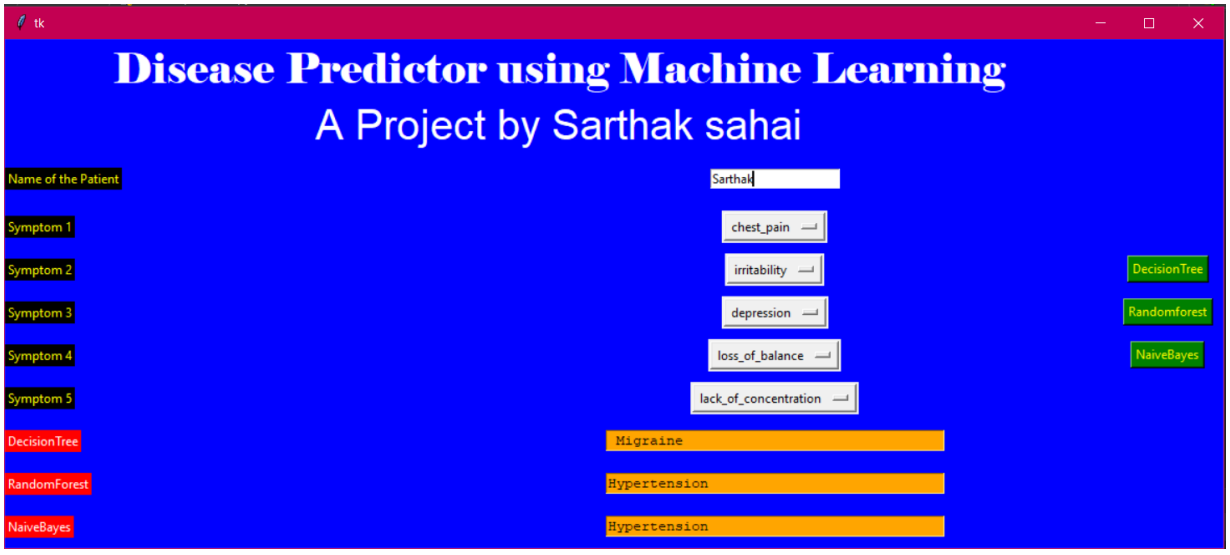
The machine learning techniques used in the disease prediction models produced encouraging results. Comparing the Random Forest method to Naive Bayes and Decision Tree, it performed better. Its greater precision, recall, and F1-score results demonstrate its capacity to recognize pertinent disease instances with accuracy and to strike a balance between precision and memory.

Comparing Random Forest to the other algorithms, precision—a measure of the proportion of accurately predicted positive cases among all positive predictions—was greater for Random Forest. Consequently, it can be inferred that Random Forest is successful in correctly detecting the relevant disease instances.

Similarly, Random Forest outperformed the other algorithms in terms of recall, which gauges the percentage of properly foreseen positive cases among all real positive cases. It showed a greater capacity to locate a higher percentage of true positive instances.

Additionally, the feature importance analysis identified the symptoms that had the most influence on the illness prediction process. The important symptoms that significantly contribute to the forecasts were highlighted by Random Forest's evaluation of feature relevance. This knowledge is helpful in comprehending the underlying causes of particular diseases and can support clinical judgement.

In conclusion, the machine learning algorithms that were developed, especially Random Forest, showed excellent performance in disease prediction. In terms of precision, recall, and F1-score, Random Forest surpassed Naive Bayes and Decision Tree. The models' interpretability, especially Decision Trees' openness, makes it easier to grasp and believe in forecasts.



Chapter 5:

Conclusion and Future Work

Using machine learning methods including Random Forest, Naive Bayes, and Decision Tree, the team successfully created a disease prediction system. The models that were put into practice showed encouraging results, displaying their potential for precise disease prediction. The project's main discoveries and contributions are as follows:

- When compared to Naive Bayes and Decision Tree, Random Forest performed better in terms of precision, recall, and F1-score.
- The models' capacity to be understood, especially Decision Trees' transparency, offered insights into the decision-making process and raised confidence in the projections.
- The most important symptoms for disease prediction were found with the use of feature importance analysis.

Future directions for this project and areas of improvement include:

- **Enhancing Data Quality:** Improving the quality of the collected data, ensuring completeness and accuracy, and exploring additional data sources can further enhance the performance of the disease prediction models.
- **Feature Engineering**: Exploring advanced feature engineering techniques to derive more informative features from the data, such as domain-specific feature transformations or creating new composite features, can potentially improve the models' predictive capabilities.
- **Integration of Advanced Algorithms**: Investigating the integration of more advanced machine learning algorithms, such as gradient boosting techniques or deep learning architectures, can be explored to potentially boost the prediction accuracy and handle complex relationships in the data.
- **Validation on External Datasets:** Validating the developed models on external datasets or collaborating with healthcare institutions to evaluate their performance in real-world scenarios will provide further evidence of their effectiveness and generalizability.
- **Real-time Disease Prediction**: Developing a system that allows real-time disease prediction using streaming data or incorporating dynamic data inputs can enable timely interventions and personalized healthcare.
- **Integration with Electronic Health Records**: Integrating the disease prediction system with electronic health records (EHRs) can enhance its practicality and enable seamless integration into existing healthcare systems, providing clinicians with valuable decision support tools.