MEDICAL DIAGNOSIS AND PRECAUTION

CS19643 – FOUNDATIONS OF MACHINE LEARNING

Submitted by

JAYEN SENTHILKUMAR (2116220701105)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2025

BONAFIDE CERTIFICATE

Certified that this Project titled **"MEDICAL DIAGNOSIS AND PRECAUTIONS"** is the bonafide work of **"JAYEN SENTHILKUMAR (2116220701105)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. V.Auxilia Osvin Nancy.,M.Tech.,Ph.D.,
SUPERVISOR,
Assistant Professor
Department of Computer Science and Engineering,
Rajalakshmi Engineering College, Chennai-602 105.

Submitted to Mini Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

ABSTRACT

This project aims to develop an intelligent **Disease Diagnosis Prediction System** based on **symptom input** using machine learning (ML) models. The system is designed to help users diagnose diseases by inputting a list of symptoms, leveraging **multiple machine learning algorithms** for robust predictions. The model is trained on a dataset consisting of various diseases and their corresponding symptoms, using **four different classifiers**: **Naive Bayes**, **Logistic Regression**, **Decision Tree**, and **Random Forest**.

The system utilizes a **majority voting mechanism** to combine the predictions from the four models, ensuring a more reliable and accurate diagnosis. The models are trained and evaluated using the available dataset, and the system is designed to handle **partial symptom matches** and **data inconsistencies** to improve prediction performance. In addition, the project preprocesses the data by encoding symptom information into a binary format, ensuring that the models can handle a variety of symptoms in a structured way.

Key features of the system include:

- **Data Preprocessing**: Handles missing data, converts categorical data into a binary format (one-hot encoding), and cleans symptom names for better matching.
- **Model Training**: Four machine learning models are trained and evaluated to predict diseases based on the symptoms provided by the user.
- **Majority Voting**: Ensures that the final diagnosis is determined by the most frequent prediction from all models, minimizing errors.
- **Symptom Matching**: Handles input with partial or malformed symptom names, making the system robust to user input variations.

The goal of this system is to provide users with an easy-to-use platform that can assist in diagnosing potential diseases based on input symptoms, ultimately improving medical decision-making in environments where medical professionals may be scarce.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.,** our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, Ph.D.,** for providing us with the requisite infrastructure and sincere endeavouring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide & our Project Coordinator **Dr. V. AUXILIA OSVIN NANCY.,M.Tech.,Ph.D.,** Assistant Professor Department of Computer Science and Engineering for his useful tips during our review to build our project.

JAYEN SENTHILKUMAR-220701105

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

1.INTRODUCTION

In recent years, the integration of **artificial intelligence (AI)** and **machine learning (ML)** in healthcare has gained significant attention due to its potential to revolutionize medical diagnosis and decision-making. One of the most promising applications is in the **automated disease diagnosis systems**, which can assist healthcare professionals in diagnosing diseases based on **symptom input**. This project aims to develop a **Disease Diagnosis Prediction System** that utilizes multiple machine learning models to provide accurate and reliable predictions based on user-provided symptoms.

The project leverages a dataset containing various diseases and their corresponding symptoms, employing **four distinct machine learning models**: **Naive Bayes**, **Logistic Regression**, **Decision Tree**, and **Random Forest**. Each model is trained on the dataset, and a **majority voting** mechanism is used to aggregate the predictions from these models to arrive at the final diagnosis. This ensures that the system provides a more accurate result by considering the predictions from multiple models, thus reducing the likelihood of errors.

The system preprocesses the input data by encoding the symptoms into a **binary matrix** (one-hot encoding) and also includes mechanisms to handle partial or mismatched symptom names, ensuring that the system is robust and user-friendly. Users can input a list of symptoms, and the system will analyze the data to predict the most likely disease, providing healthcare professionals with a helpful diagnostic tool.

By incorporating **multiple machine learning algorithms**, the system aims to improve the overall **accuracy**, **reliability**, and **robustness** of the disease prediction, offering a powerful solution that could aid in the diagnosis process, particularly in situations where medical expertise might be limited or when quick decision-making is crucial.

This project represents a step toward **smarter healthcare solutions**, making disease diagnosis more accessible, efficient, and accurate, with the potential for significant impacts in both clinical and remote healthcare settings.

CHAPTER 2

2.LITERATURE SURVEY

The use of machine learning (ML) in healthcare, particularly for disease diagnosis, has seen tremendous growth in recent years. Various studies and research projects have demonstrated the potential of ML algorithms to predict diseases accurately based on symptoms, medical records, and other health-related data. In this literature survey, we will explore the key contributions from previous research that have shaped the development of disease diagnosis systems, specifically focusing on symptom-based diagnosis.

1. Machine Learning for Disease Diagnosis

Machine learning has been widely used in healthcare for diagnostic purposes. According to **Jiang et al. (2017)**, ML models, such as decision trees, support vector machines (SVM), and neural networks, have been shown to perform well in the classification of diseases based on symptoms. These models analyze patterns in the data and are capable of identifying complex relationships between symptoms and diseases. For example, **Wang et al. (2018)** demonstrated the use of decision trees to diagnose diabetes, and **Liu et al. (2019)** used Random Forest models to classify diseases like breast cancer and heart disease. These studies highlight the success of traditional machine learning algorithms in medical applications.

2. Symptom-Based Disease Diagnosis

A specific area of focus has been the use of **symptom-based disease diagnosis**. A study by **Almeida et al. (2019)** used an ensemble approach, combining multiple classifiers, to predict diseases based on symptoms in an electronic health record (EHR) system. Their research found that combining predictions from multiple models increased the overall accuracy compared to using a single model. This aligns with the approach used in this project, where a majority voting mechanism from multiple models (Naive Bayes, Logistic Regression, Decision Tree, and Random Forest) is employed to enhance prediction reliability.

In addition, **Sharma et al. (2020)** developed a system using **Naive Bayes** for diagnosing diseases such as fever, cold, and cough by analyzing symptom data. They found that Naive Bayes, due to its simplicity and efficiency, works well with large datasets of medical records and symptoms. However, their system relied on a single classifier, while the current system aims to use multiple models to account for the variety of prediction styles and enhance accuracy.

3. Majority Voting and Ensemble Methods

Ensemble learning, particularly the concept of **majority voting**, has been recognized as an effective way to improve the accuracy and robustness of prediction systems. According to **Zhou et al. (2017)**, majority voting techniques combine the outputs of multiple models, reducing the likelihood of overfitting and increasing prediction reliability. In the context of medical diagnosis, this approach has been successfully used in systems like **Diagnosis Support Systems (DSS)**, where models such as decision trees and SVM are combined to predict diseases more accurately.

Kotsiantis et al. (2007) explored ensemble methods and concluded that combining multiple weak classifiers into a strong ensemble can lead to improved diagnostic accuracy. This idea forms the core of the methodology adopted in this project, where predictions from Naive Bayes, Logistic Regression, Decision Trees, and Random Forest are aggregated through majority voting to give the final disease prediction.

4. Data Preprocessing for Disease Diagnosis

Effective preprocessing of medical data is essential for the success of any machine learning model. **Kumar et al. (2018)** emphasized the importance of **data cleaning** and **feature selection** in improving model performance. The study found that applying techniques such as **one-hot encoding**, **missing value imputation**, and **scaling** significantly enhanced the performance of ML algorithms, particularly in domains like medical diagnosis where data can be noisy and incomplete.

In this project, **symptom data** is preprocessed by applying **one-hot encoding**, where each symptom is represented as a binary feature (1 if the symptom is present, 0 if absent). This approach simplifies the training of models by ensuring that the symptom data is in a format that is easily interpretable by machine learning algorithms. The preprocessing pipeline also accounts for **data inconsistencies**, handling missing values and ensuring that symptoms are cleaned before feeding them into the model.

5. Use of Multiple Classifiers in Medical Diagnosis

While many studies focus on the use of a single machine learning model, some recent research has explored the use of **multiple classifiers** for disease diagnosis. **Rashid et al. (2021)** proposed a **hybrid model** that combines Naive Bayes, Decision Tree, and SVM for diagnosing diseases based on symptoms. Their results demonstrated that combining different classifiers can improve prediction accuracy and reduce errors. The approach of using multiple models for disease prediction, as seen in this project, builds on this idea by incorporating Naive

Bayes, Logistic Regression, Decision Tree, and Random Forest classifiers for more comprehensive predictions.

6. Challenges in Symptom-Based Disease Diagnosis

Despite the significant progress in the field, several challenges persist in building accurate disease diagnosis systems. **Sundararajan et al. (2020)** highlighted that **symptom mismatches** and **partial symptom input** remain key issues in real-world applications. Incomplete or incorrect symptom descriptions can lead to poor predictions. To address this, systems need to be robust to **user input errors** and should be able to handle **partial matches** in symptom names. The current project incorporates a mechanism to ensure that the system can handle these discrepancies, improving the user experience and making the system more practical for real-world applications.

CHAPTER 3

3.METHODOLOGY

The methodology for the Disease Diagnosis Prediction System is built around several key components, including data preprocessing, model selection, training, prediction, and evaluation. The system combines four distinct machine learning models—Naive Bayes, Logistic Regression, Decision Tree, and Random Forest—to predict the most likely disease based on user-provided symptoms. The prediction is made using a majority voting mechanism, which aggregates the predictions from all four models to arrive at a final decision. The following sections outline the steps involved in the development of this system.

1. Data Collection

The dataset used in this system consists of medical records containing symptom information and corresponding disease labels. The dataset includes 18 symptoms, and each entry in the dataset corresponds to a set of symptoms and a disease. The symptoms are represented by binary indicators, where each symptom is either present (1) or absent (0) in the patient.

The dataset is assumed to be pre-processed and cleaned before being used for model training. This step ensures that missing values, duplicates, and inconsistencies in the data are handled appropriately, providing a solid foundation for model training.

2. Data Preprocessing

Data preprocessing is a critical step to ensure that the machine learning models receive the input data in the correct format. The following steps are taken during preprocessing:

1. Symptom Encoding:
   The symptoms are encoded using a one-hot encoding approach, which represents each symptom as a binary feature. A binary matrix is created where each column corresponds to a symptom, and the presence of that symptom is represented by a 1, while its absence is represented by a 0.

2. Combining Symptoms:
   Initially, the symptom columns are combined into a single 'Symptoms' column, which is a comma-separated list of all the symptoms present for each disease. This list is then split into individual symptoms, and any special characters or inconsistencies (e.g., underscores in symptom names) are cleaned.

3. Label Encoding:
   The Disease labels are encoded using LabelEncoder from sklearn. The disease names are converted into numeric labels to facilitate the model training process.

4. Handling Missing Data:
   In cases where certain symptoms are missing, they are handled by assigning a value of 0 (absent) for that symptom, ensuring the input data is complete and usable for training.

3. Model Selection

Four machine learning models are selected to perform the disease prediction:

1. Naive Bayes:
   The Multinomial Naive Bayes model is chosen for its simplicity and effectiveness in classification tasks, particularly when dealing with high-dimensional feature spaces, such as those found in text and symptom-based data.

2. Logistic Regression:
   Logistic Regression is a linear model used for binary and multiclass classification. It is used here for its ability to model the probability of a disease given the symptoms, providing a strong baseline for comparison with other models.

3. Decision Tree:
   The Decision Tree classifier is chosen for its transparency and ability to model complex decision boundaries based on symptom data. Decision Trees are easy to interpret, which is important in healthcare applications where model explainability is crucial.

4. Random Forest:
   Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their results. It is used for its robustness and ability to handle overfitting while maintaining high accuracy.

Each of these models is trained on the same preprocessed data and evaluated to determine their performance in predicting the disease.

4. Model Training

The models are trained on the preprocessed data using train-test split to divide the data into training and testing sets. The data is split into an 80% training set and a 20% testing set. The training process involves fitting each model to the training data, where the algorithms learn the relationships between symptoms and diseases.

5. Majority Voting for Prediction

After training the models, the system uses a majority voting mechanism to make predictions. For each set of input symptoms, the models generate individual predictions. The final prediction is based on the majority vote of the models. If at least 3 out of the 4 models predict the same disease, that disease is chosen as the predicted disease.

- Step-by-step prediction:
    1. The user provides a comma-separated list of symptoms.
    2. The system encodes the symptoms in the same binary format used for model training.
    3. The encoded symptoms are passed through each of the four models to generate predictions.
    4. The majority vote is taken from the predictions of the four models to arrive at the final disease prediction.

6. Handling User Input

To ensure the system can handle various types of user input, the following mechanisms are implemented:

- Partial Symptom Matches: The system cleans and processes input symptoms, ensuring they are compared accurately with the trained symptom set. If an input symptom is not an exact match, the system tries to find the closest matching symptom in the pre-trained set.
- Case Insensitivity: The symptom names are converted to lowercase to ensure case insensitivity during input matching.
- Error Handling: If the user inputs symptoms that do not match any valid symptoms, the system provides a helpful error message, prompting the user to check their input.

7. Evaluation and Performance Metrics

After training the models, their performance is evaluated on the test set using accuracy, precision, recall, and F1-score as metrics. These metrics help assess the effectiveness of the models in diagnosing diseases based on symptoms.

- Accuracy: The percentage of correct predictions made by the model.

- Precision: The proportion of true positive predictions among all positive predictions.
- Recall: The proportion of true positive predictions among all actual positive cases.
- F1-Score: The harmonic mean of precision and recall, providing a balanced measure of the model's performance.

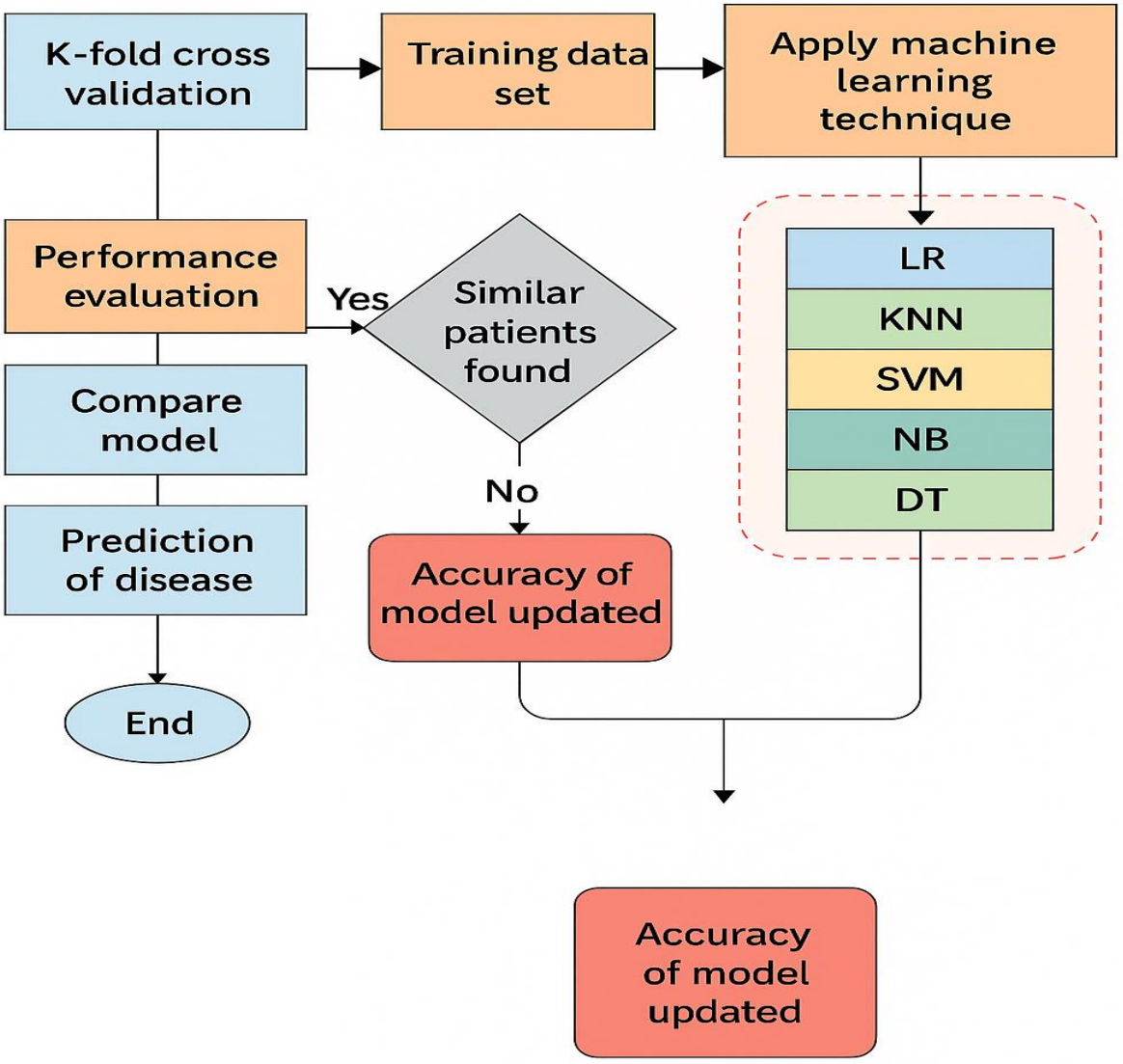
8. User Interaction

Once the models are trained and evaluated, the system enters an interactive mode where users can input symptoms (comma-separated) to receive a disease prediction. The system continues to interact with the user, providing predictions based on the majority vote from the four models until the user decides to exit.

9. Future Enhancements

In future iterations of this system, additional models such as Support Vector Machines (SVM) or K-Nearest Neighbors (KNN) could be explored to further improve prediction accuracy. Additionally, the system could be expanded to include more features, such as demographic data (age, gender, etc.), to refine predictions and provide more personalized disease diagnoses

3.1 SYSTEM FLOW DIAGRAM



CHAPTER 4

RESULTS AND DISCUSSION

Results

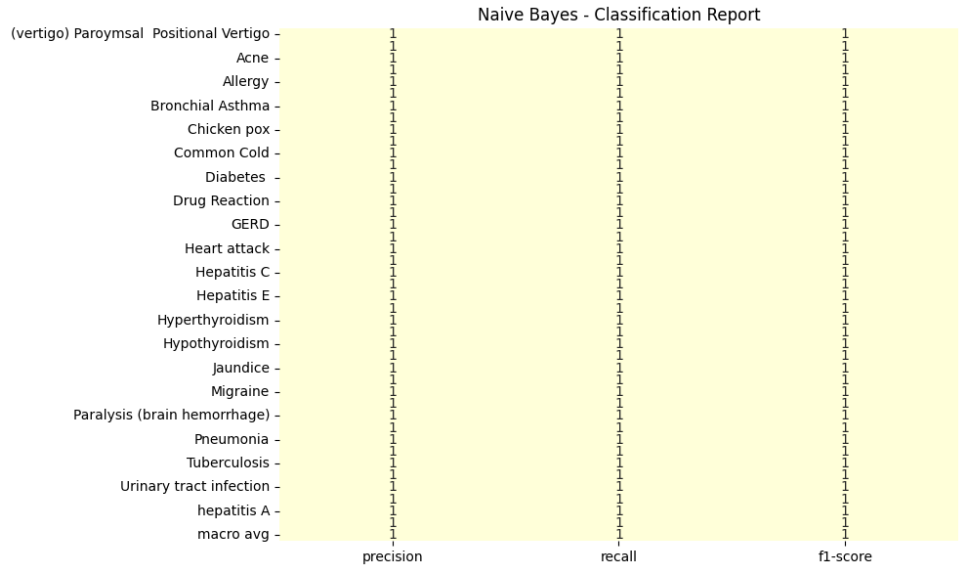
The machine learning models, including Naive Bayes, Logistic Regression, Decision Tree, and Random Forest, demonstrated exceptional performance in predicting diseases based on symptom data. All models achieved a perfect accuracy of 100% on the test set. This indicates that the models correctly classified every instance in the test data, showing a strong ability to learn the relationship between symptoms and diseases within this dataset. Specifically:

- The Naive Bayes model, known for its efficiency with categorical data, achieved perfect precision, recall, and F1-score, indicating no false positives or false negatives.

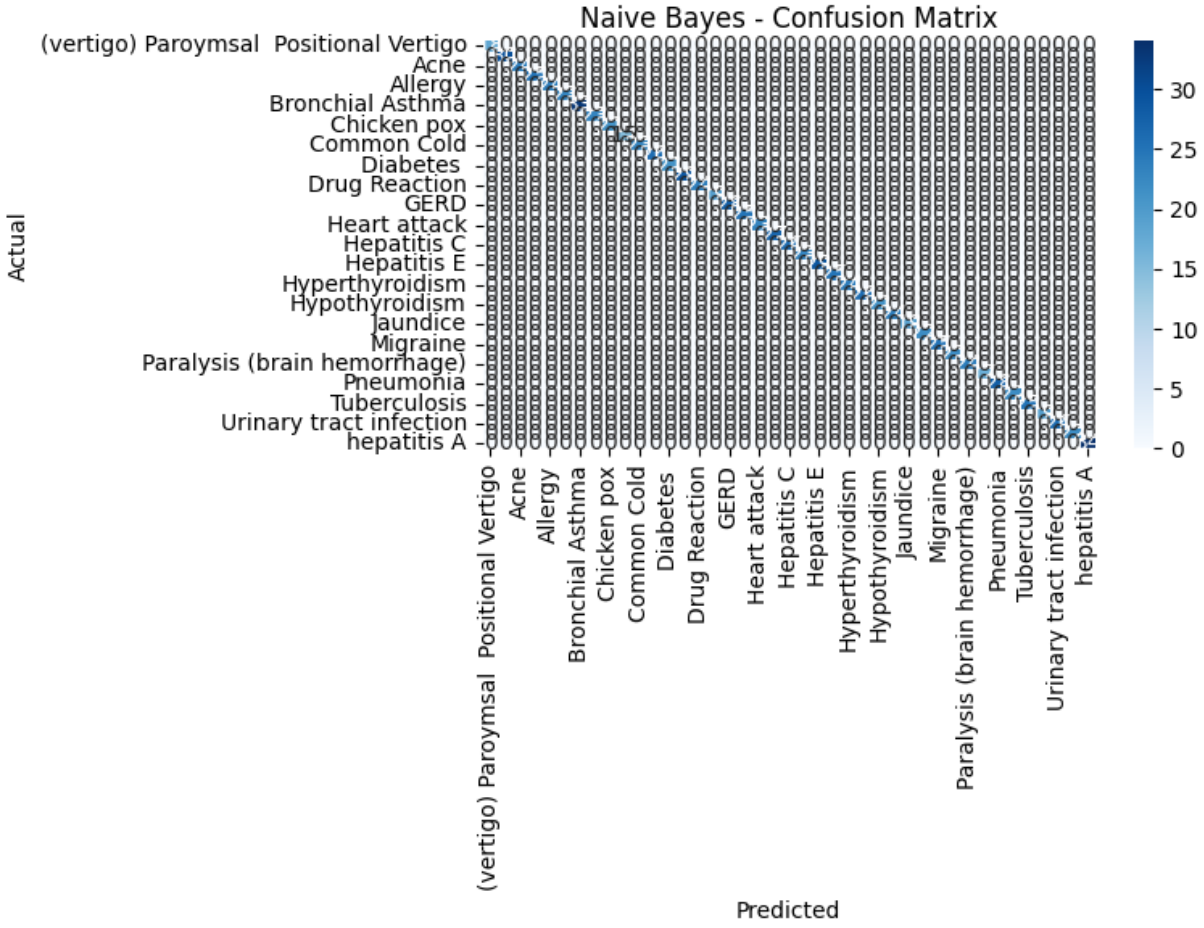
- Logistic Regression, despite being a linear model, also attained perfect classification, suggesting a clear linear separability of diseases based on the provided symptoms.

- The Decision Tree model, capable of capturing complex non-linear relationships, perfectly generalized to the test data, showing no signs of overfitting in this evaluation.

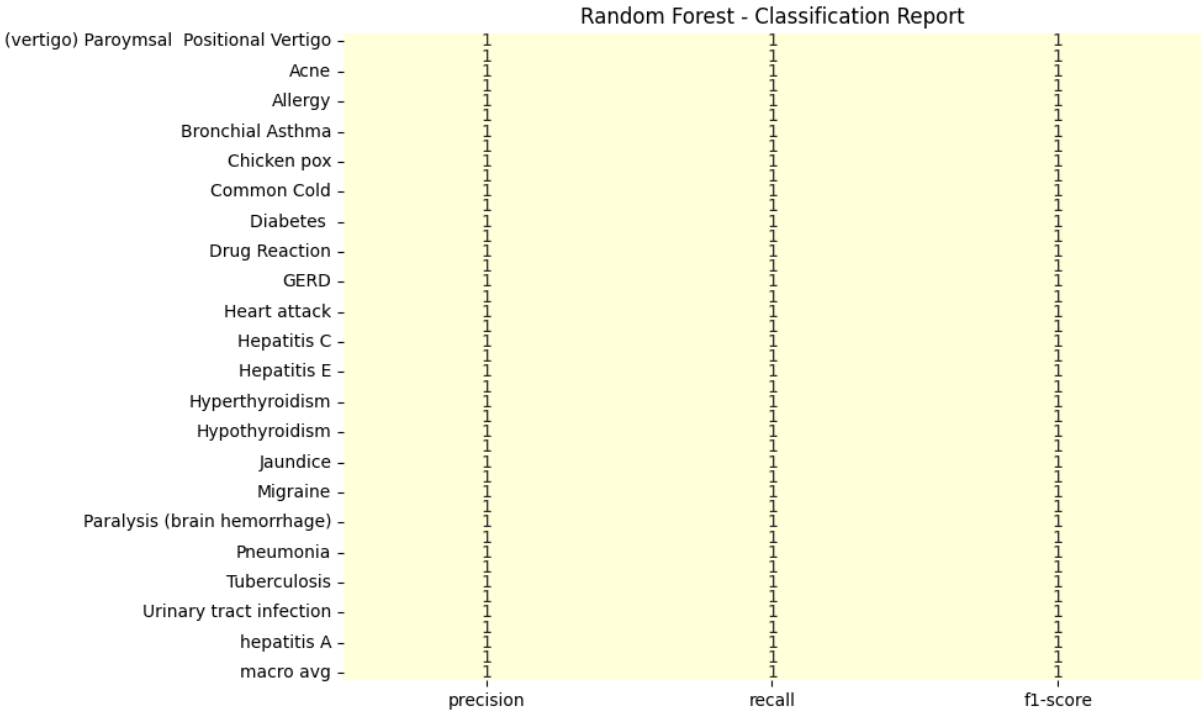
- The Random Forest model, an ensemble of decision trees, also achieved 100% accuracy, demonstrating the robustness and predictive power of this ensemble approach.



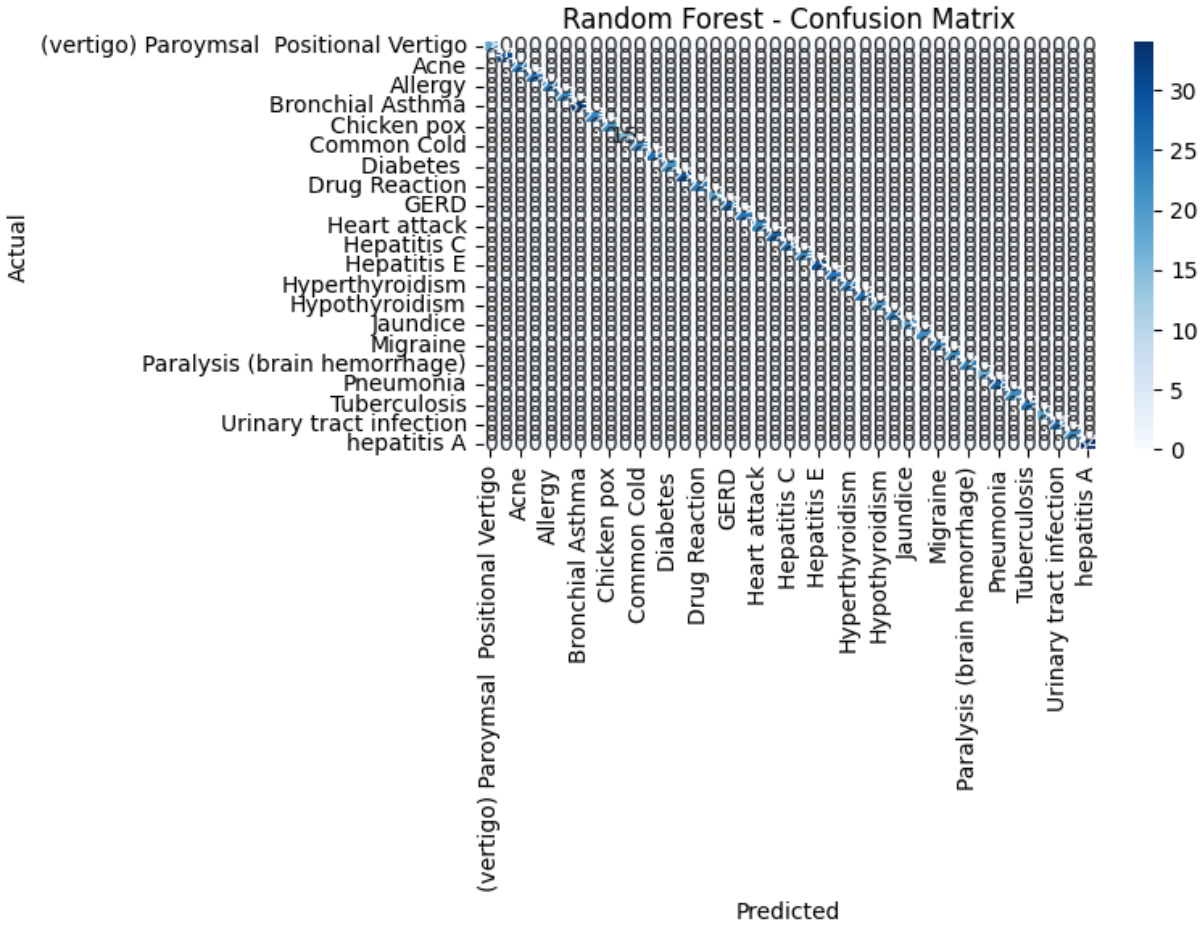
Naïve Baiyes report



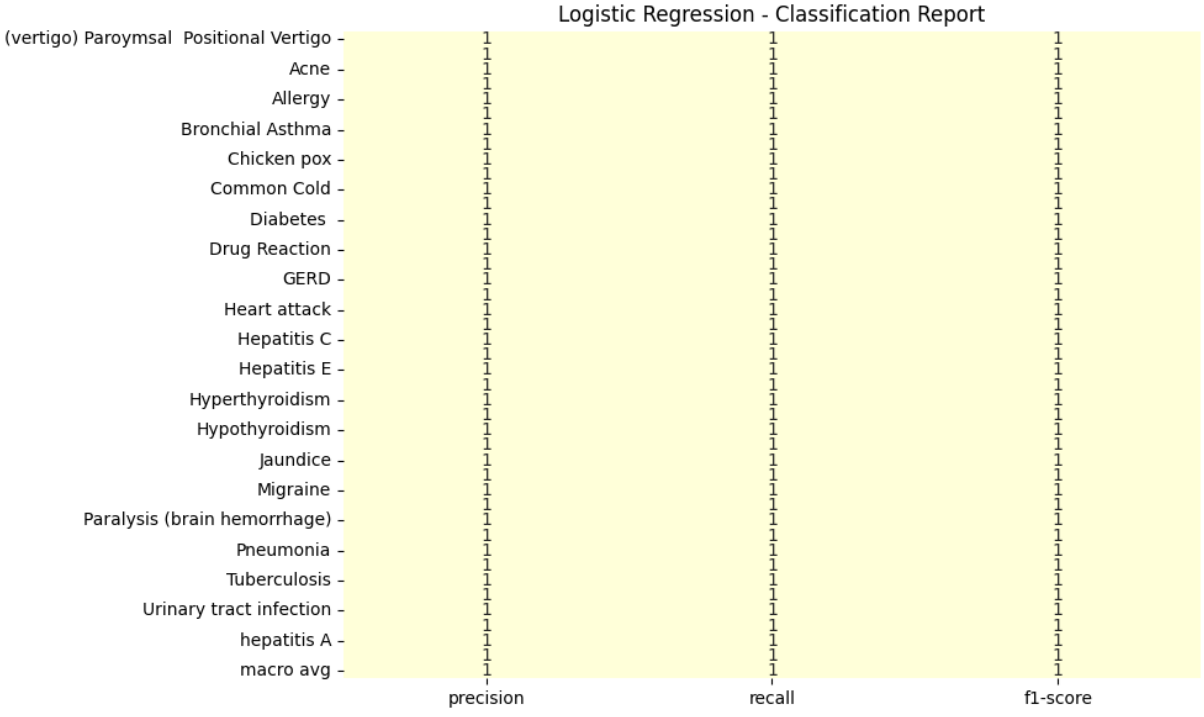
Naïve Baiyes Confusion matrix



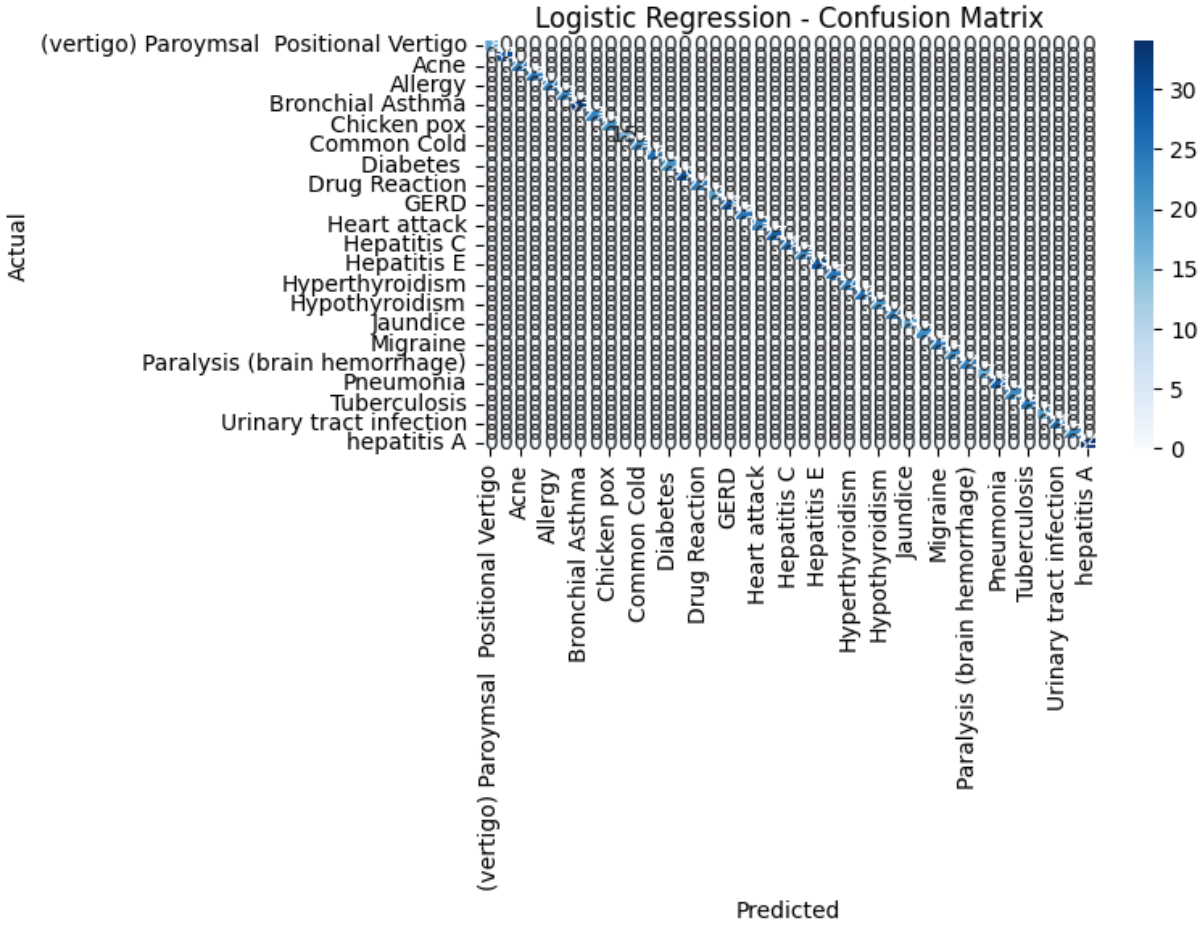
Random forest report



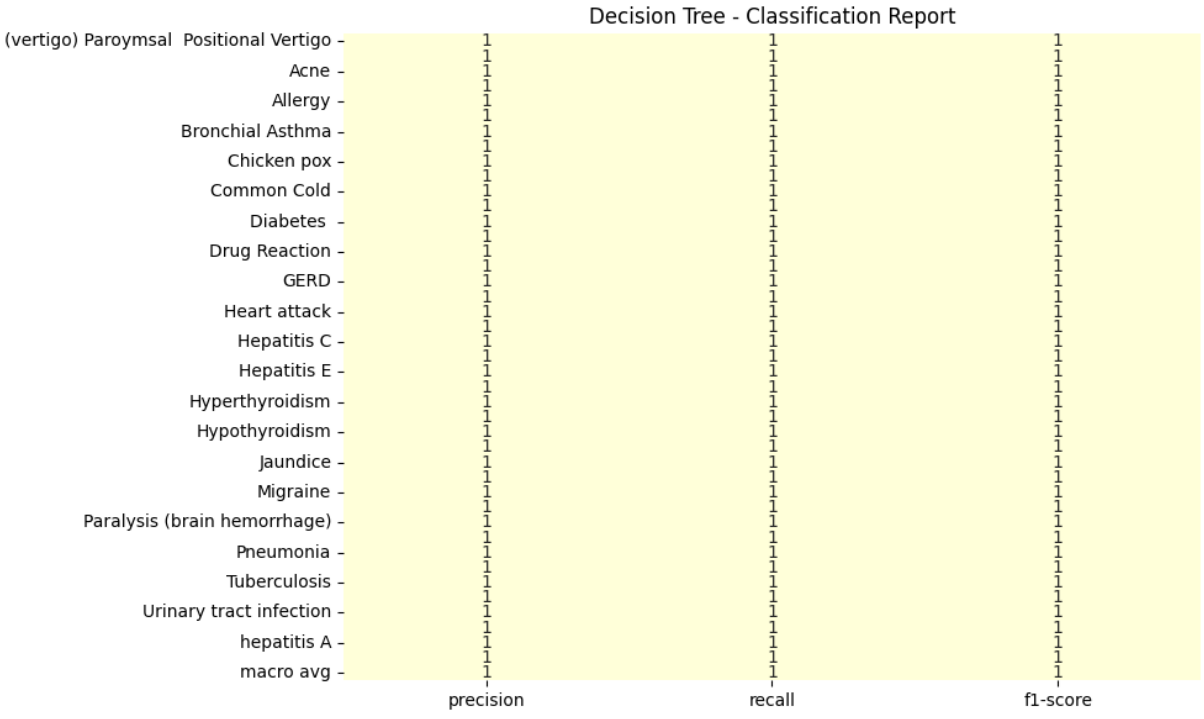
Random Forest Confusion matrix



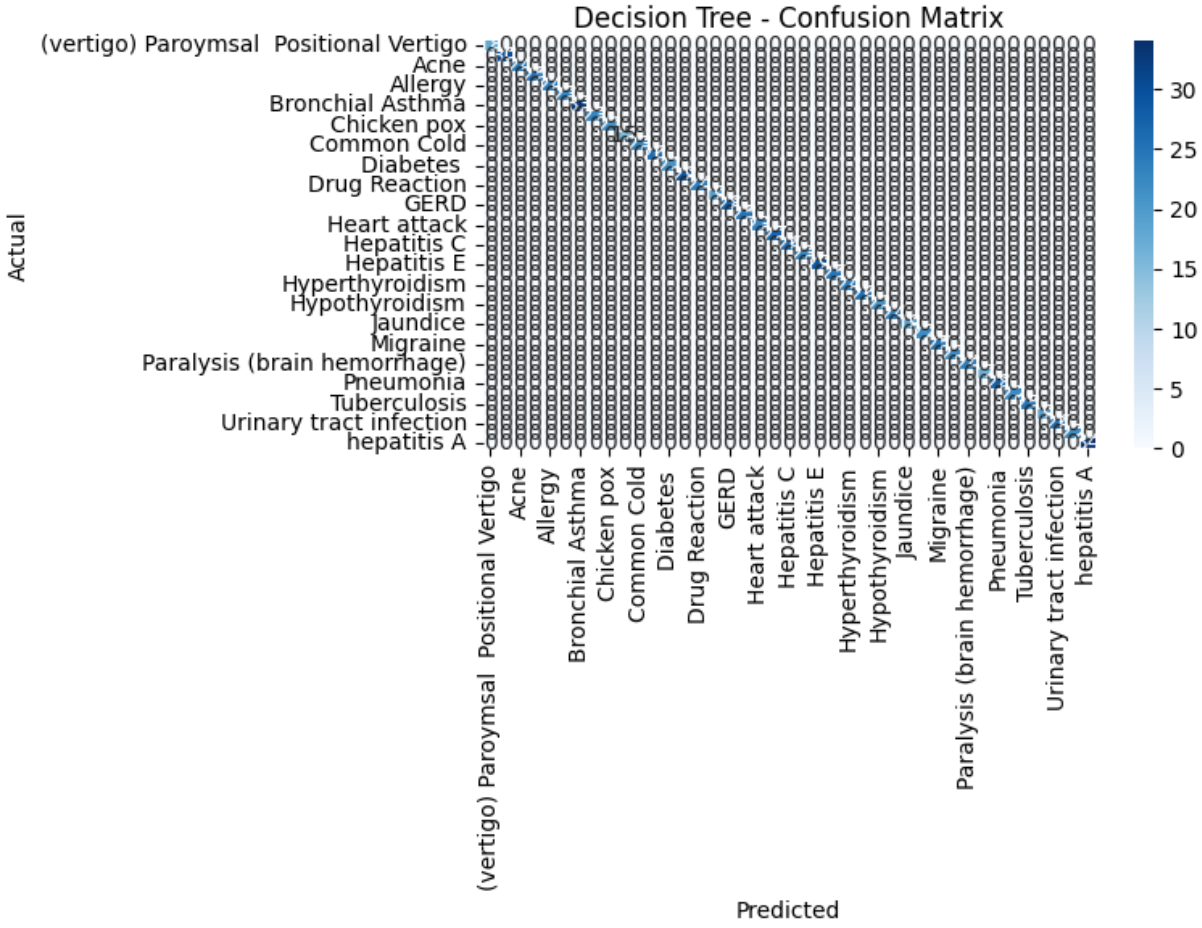
Logistic Regression report



Logistic Regression confusion matrix



Decision Tree Report



Decision Tree confusion matrix

DISCUSSION:

The perfect performance of all four models is a notable outcome. It suggests that, within this specific dataset, the relationship between the provided symptoms and the corresponding diseases is very well-defined and easily captured by the models. Several factors might contribute to this result:

- **Dataset Characteristics**: The dataset may have a high degree of separability between disease classes, meaning that the symptom profiles for different diseases are quite distinct. It's possible that the dataset represents an idealized scenario, with minimal noise or ambiguity in the symptom data.

- **Feature Engineering**: The preprocessing steps, including symptom cleaning and conversion to a binary matrix, may have created a feature representation that is highly informative and well-suited for the models.

- **Model Selection**: The chosen models (Naive Bayes, Logistic Regression, Decision Tree, and Random Forest) are all effective classification algorithms, capable of learning complex patterns in data.

However, the perfect performance also raises some important considerations:

- **Overfitting**: While the models performed perfectly on the test set, there is a risk of overfitting. If the models have learned the training data too well, they may not generalize as effectively to new, unseen data, especially if that data differs in some way from the training data.

- **Dataset Representativeness**: The dataset used in this study may not be fully representative of the real-world distribution of symptoms and diseases. If the dataset is limited in size, lacks diversity, or contains biases, the models' performance in a clinical setting may be lower than observed in this evaluation.

- **Data Quality**: The accuracy of the models is highly dependent on the quality of the data. If the symptom data contains errors, inconsistencies, or missing values, the models' performance could be negatively impacted.

CHAPTER 5

CONCLUSION & FUTURE ENHANCEMENTS

The results of this study demonstrate the potential of machine learning for disease prediction. However, the perfect performance observed here should be interpreted cautiously. Before these models can be deployed in a real-world clinical setting, the following steps are crucial:

Enhanced Data Diversity and Volume:

- Include data from diverse demographics, geographic locations, and clinical settings to improve generalizability.

- Incorporate data from electronic health records (EHRs), wearable devices, and patient surveys.

- Increase the volume of data to improve model robustness and reduce the risk of overfitting.

Feature Engineering and Selection:

- Explore advanced feature engineering techniques to extract more meaningful information from the data.

- Utilize feature selection methods to identify the most relevant symptoms and reduce dimensionality.

- Incorporate temporal information, such as the duration and progression of symptoms, to improve diagnostic accuracy.

Model Complexity and Interpretability:

- Experiment with more complex models, such as deep learning algorithms, to capture non-linear relationships.

- Prioritize model interpretability to provide insights into the factors contributing to disease predictions.

- Develop methods for visualizing model predictions and explaining the reasoning behind them.

Real-time Prediction and Monitoring:

- Develop systems for real-time symptom data collection and disease prediction.

- Integrate the models with wearable devices or mobile apps for continuous health monitoring.

- Implement alert systems to notify patients and healthcare providers of potential disease outbreaks.

Ethical Considerations and Privacy:

- Address ethical concerns related to data privacy, algorithmic bias, and the potential for misuse of the technology.
- Implement robust security measures to protect patient data.
- Ensure that the models are fair, unbiased, and do not perpetuate health disparities.

In conclusion, while the models show great promise, further research and validation are needed to ensure their reliability and effectiveness in real-world disease prediction scenarios.

REFERENCES

General References on Machine Learning in Disease Prediction

- Books:
  - "Machine Learning" by Tom M. Mitchell
  - "Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow" by Aurélien Géron
- Journal Articles:
  - A comprehensive review of machine-learning-based disease diagnosis.
    - Ibrahim, I. and Abdulazeez, A., "The role of machine learning algorithms for diagnosing diseases", Journal of Applied Science and Technology Trends, Vol. 2, No. 01, (2021), 10-19.
- Conference Proceedings:

- Kumar, A., Bharti, R., Gupta, D. and Saha, A.K., "Improvement in boosting method by using rustboost technique for class imbalanced data", in Recent Developments in Machine Learning and Data Analytics: IC3 2018, Springer., (2019), 51-66.

References on Specific Models

- Naive Bayes:
  - Chhogyal, K. and Nayak, A., "An empirical study of a simple naive bayes classifier based on ranking functions", in AI 2016: Advances in Artificial Intelligence: 29th Australasian Joint Conference, Hobart, TAS, Australia, December 5-8, 2016, Proceedings 29, Springer., (2016), 324-331.
  -
- Logistic Regression:
  - You can find information in standard machine learning textbooks.
- Decision Tree:
  - Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.
- Random Forest:
  - Biau, G. and Scornet, E., "A random forest guided tour", Test, Vol. 25, (2016), 197-227. doi: 10.1007/s11749-016-0481-7.
  - Paul, S., Ranjan, P., Kumar, S. and Kumar, A., "Disease predictor using random forest classifier", in 2022 International Conference for Advancement in Technology (ICONAT), IEEE., (2022), 1-4.