

**DISEASE PREDICTION AND DOCTOR RECOMMENDATION SYSTEM USING  
MACHINE LEARNING APPROACHES: A CASE STUDY IN BANGLADESH**

**BY**

**DARUN KARAS ABIR**

**ID: 201-15-14188**

This Report Presented in Partial Fulfillment of the Requirements for the  
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

**Dr. Md. Fokhray Hossain**

Professor

Department of Computer Science and Engineering  
Daffodil International University

Co-Supervised By

**Md. Aynul Hasan Nahid**

Lecturer

Department of Computer Science and Engineering  
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**JANUARY 2024**

## APPROVAL

This Project titled “**Disease Prediction and Doctor Recommendation System Using Machine Learning Approaches: A Case Study in Bangladesh**”, submitted by Darun Karas Abir, ID No: 201-15-14188 to the Department of Computer Science and Engineering(CSE), Daffodil International University(DIU) has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on January 23, 2024.

### BOARD OF EXAMINERS

**Dr. Md. Zahid Hasan (ZH)**

**Associate Professor**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Chairman**

**Mr. Abdus Sattar (AS)**

**Assistant Professor**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**

**Tapasy Rabeya (TRA)**

**Senior Lecturer**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**

**Dr. Mohammad Shahidur Rahman (DMSR)**

**Professor**

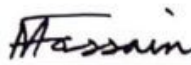
Department of Computer Science and Engineering  
Shahjalal University of Science and Technology

**External Examiner**

## DECLARATION


I hereby declare that, this project has been done by us under the supervision of **Dr. Md. Fokhray Hossain**, Professor, Department of Computer Science and Engineering (CSE), Daffodil International University (DIU). I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree.

**Supervised by:**

 20-1-2024

**Dr. Md. Fokhray Hossain**  
Professor  
Department of CSE  
Daffodil International University

**Co-Supervised by:**

 20/01/2024

**Md. Aynul Hasan Nahid**  
Lecturer  
Department of CSE  
Daffodil International University

**Submitted by:**



**Darun Karas Abir**  
ID: 201-15-14188  
Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

First, I express my heartfelt thanks and gratitude to almighty God for His divine blessing, which makes it possible for me to complete the final year project successfully.

I am grateful and express my profound indebtedness to **Dr. Md. Fokhray Hossain, Professor**, Department of CSE Daffodil International University, Dhaka. Deep knowledge and keen interest of our supervisor in the field of “Machine Learning” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

Additionally, I extend my thanks to **Md. Aynul Hasan Nahid, Lecturer**, Department of CSE, Daffodil International University, for co-supervising this project. His insights, assistance, and collaborative efforts have greatly contributed to the overall success of this endeavor. I am grateful for his support and guidance, which enhanced the quality and depth of our work.

I would like to express my heartfelt gratitude to **Dr. Sheak Rashed Haider Noori, Professor & Head**, Department of CSE, for his kind help in finishing our project and also to other faculty members and the staff of the CSE department of Daffodil International University.

I would like to thank my entire group of course mates at Daffodil International University who took part in this discussion while completing the coursework.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

## ABSTRACT

Despite significant strides in modern technology, a substantial portion of the global population is still not getting proper medical care. This issue is particularly pronounced in developing countries which face a double burden of communicable and non-communicable diseases. On the other hand, limited access to quality healthcare, particularly in remote areas where skilled doctors may not be available, is a significant challenge for patients to receive appropriate treatment. To solve this problem, this project will be able to make the treatment much easier and more accurate. The system is designed to enhance the accuracy, reliability, and efficiency of disease prediction by leveraging the power of machine learning algorithms and reliable datasets. It will also make the work of doctors a lot easier because this project can diagnose possible diseases and give suggestions about what restrictions should be followed at home to get rid of these diseases. Since disease prediction is a very crucial subject where accuracy has to be maximized, ten machine-learning models, including Decision Trees, Random forest, Bagging Classifier, Support Vector Machine and AdaBoostClassifier and hybrid (a combination of several machine-learning) models have been used in this project that capitalize on the strengths of models or data types while mitigating their respective weaknesses.

The fundamental objective of this work is to significantly enhance the user experience of the existing disease prediction system. This is achieved through the design and implementation of a user-friendly interface using a Python framework "Streamlit" that facilitates not only disease prediction but also recommends appropriate doctors and suggests preventative measures.

# TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Board of Examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-7</b>
1.1 Background of Research	1
1.2 Motivation	2
1.3 Problem Statement	2
1.4 Research Questions	3
1.5 Aim of the Research	3
1.6 Proposed Solution	4
1.7 Project Management and Finance	5
1.8 Report Layout	6
1.9 Conclusion	7
<b>CHAPTER 2: LITERATURE REVIEW</b>	<b>8-27</b>
2.1 Introduction	8
2.2 Preliminaries/Terminologies	8
2.3 Literature Review	9
2.4 Comparative Analysis and Summary	18

2.5 Scope of the Problem	26
2.6 Challenges	26
2.7 Conclusion	27
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>28-50</b>
3.1 Introduction	28
3.2 Research Subject and Instrumentation	28
3.3 Workflow	30
3.4 Data Collection Procedure	30
3.5 Statistical Analysis	34
3.6 Data Pre-processing	37
3.7 Proposed Methodology	39
3.8 Implementation Requirements	49
3.9 Conclusion	50
<b>CHAPTER 4: DISEASE PREDICTION AND DOCTOR RECOMMENDATION, SYSTEM ANALYSIS AND DESIGN SPECIFICATION</b>	<b>51-53</b>
4.1 Introduction	51
4.2 Workflow	51
4.3 Proposed Model	52
4.4 Implementation Requirements	52
4.5 Conclusion	53

**CHAPTER 5: DISEASE PREDICTION AND DOCTOR RECOMMENDATION, SYSTEM IMPLEMENTATION AND TESTING 54-67**

5.1 Introduction	54
5.2 Experimental Setup	54
5.3 Development of User Interface	55
5.4 Performance Evaluation and Testing	60
5.5 Conclusion	67

**CHAPTER 6: EXPERIMENTAL RESULT AND DISCUSSION 68-79**

6.1 Introduction	68
6.2 Experimental Setup	68
6.3 Experimental Results & Analysis	69
6.4 Discussion	78
6.5 Conclusion	79

**CHAPTER 7: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY 80-85**

7.1 Introduction	80
7.2 Impact on Society	80
7.3 Impact on Environment	82
7.4 Ethical Aspects	83
7.5 Sustainability Plan	84
7.6 Conclusion	85



<b>CHAPTER 8: CONCLUSION</b>	<b>86-87</b>
8.1 Further Suggested Work	86
8.2 Conclusion	87
<b>REFERENCES</b>	<b>88-91</b>
<b>APPENDICES:</b>	
A. Statistical Analysis	A1
B. Experimental Results & Analysis	B1
C. Development of User interface in Colab	C1
D. Plagiarism	D1

## LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.3.1: General Flow diagram of Methodology	30
Figure 3.3.2: Work Flow diagram of Methodology	30
Figure 3.5.1: Bar chart of disease and number of occurrences in the data set	34
Figure 3.5.2: Pie chart of unique diseases	35
Figure 3.5.3: Count of top 20 symptoms in disease and symptoms data set.	36
Figure 3.5.4: Count specialist for different disease.	36
Figure 3.6.1: Data preprocessing flow diagram	37
Figure 3.6.2: Before and after removing null values	38
Figure 3.6.3: Before and after standard scaling	39
Figure 3.7.1: General Methodology for Machine Learning	39
Figure 3.7.2: Randomly Split the data set	41
Figure 4.2: Work flow diagram for disease prediction and doctor recommendation system.	51
Figure 5.3.1: Local Server containing necessary file and saved model	55
Figure 5.3.2.1: Home Page	57
Figure 5.3.2.2: Disease prediction page	58
Figure 5.3.2.3: Disease prediction, doctor recommendation and others suggestion's page of disease prediction and doctor recommendation system	59
Figure 5.3.2.3: Some extra disease prediction pages (Heart Disease, Diabetes, Perkinsosis)	60
Figures 5.4.1: Figure for showing the system can accurately predict according to test case	63
Figures 5.4.2: Figure for showing the system can accurately predict according to test case.	66
Figure 6.3.1: Confusion matrix of different machine learning models	77
Figure 6.4: Model Accuracy before and after encoding symptoms by its priority number	78

## LIST OF TABLES

<b>TABLES</b>	<b>PAGE NO</b>
Table 2.4: Comparison Between Related Work of Disease Prediction.	19
Table 3.3.1: Sample Dataset of Symptom and Disease	31
Table 3.3.2: Sample Dataset of Disease and Specialist	32
Table 3.3.3: Sample Dataset of Disease and Description	32
Table 3.3.4: Sample Dataset of Disease and Precaution	33
Table 3.3.5: Sample Data set of Disease and Doctor name	33
Table 5.3.1: Test case for testing disease prediction system	61
Table 5.4.2: Test case for testing disease prediction and doctor recommendation system.	64
Table 6.3.1: Accuracy of different ml models when data is preprocessed with approach-1	70
Table 6.3.2: Accuracy of different ml models when data is preprocessed with approach-2	71
Table 6.3.3: Accuracy of different ml models when data is preprocessed with approach-3	72

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction:

Access to quality healthcare is a fundamental human right, but unfortunately, it is not accessible to everyone due to various reasons such as a lack of resources, insufficient medical facilities, and a shortage of skilled medical professionals. Patients are often unable to receive timely and appropriate treatment, which leads to an increased burden of disease on individuals and society as a whole. Traditional medical diagnosis methods rely on clinical experience and knowledge, which can be limited and subject to human error. Therefore, there is a need for an efficient and accurate way of predicting diseases based on symptoms presented by patients, complementing traditional medical diagnosis methods and providing a more data-driven and efficient diagnosis. Disease prediction models have also been developed, which can help identify diseases without physical contact, making them useful in situations such as the COVID-19 pandemic. Since the symptoms of a particular disease have the same effect on humans and almost the same type of precautions are required to get rid of the disease. So it seems reasonable to use the data set in the perspective of Bangladesh.

This system uses machine learning algorithms to identify patterns, diagnose the disease, and then recommend a doctor based on their expertise by processing the data on patient symptoms such as fatigue, vomiting, nausea, headache, chest pain, and other symptoms. Machine Learning has great power to analyze and cope with different diseases so that prediction is more accurate and it is cost-effective in the treatment [1]. Our project's secondary objective is to integrate a user-friendly interface with existing machine learning models using “Streamlit”, a Python web framework. This will allow us to predict diseases in two ways: general and specific. We intend to develop a framework that empowers end users to anticipate common illnesses without needing to visit a doctor and also suggests appropriate medical specialists for consultation.

## **1.2 Motivation:**

In a world where obtaining adequate healthcare is a fundamental right for everyone. But the harsh truth for developing nations, they are far from the ideal. With a large population and limited access to good doctors, especially in remote areas, getting proper medical attention can feel impossible. Traditional ways of diagnosing diseases, often based on the doctor's experience alone, aren't always accurate, leading to unnecessary suffering that leads to other critical issues. My project aims to change this and I'm confident that my dedication and innovative approach will bring this project to fruition, making a lasting impact on the healthcare landscape in developing nations.

By using different machine learning algorithms to find relevant patterns in patient symptoms for a specific disease, I want to make a system that provides faster and more accurate diagnoses, suggests doctors, and recommends better treatments, making healthcare accessible to everyone. This system will even help people in areas where reaching a doctor is difficult, allowing them to anticipate common illnesses and, if needed, find the right specialist for help.

So, the underlying motivation is clear to provide a solution to the challenges faced by patients in accessing quality healthcare, particularly in remote areas where skilled doctors may not be available. Patients are often unable to receive the treatment they require due to a lack of resources or unscrupulous doctors who prioritize their financial gain over patient well-being. In essence, the motivation is to improve the quality of patient care and reduce the burden of disease on individuals and society as a whole.

## **1.3 Problem statement:**

Limited access to quality healthcare, particularly in remote areas where skilled doctors may not be available, is a significant challenge for patients to receive appropriate treatment.

Unscrupulous doctors who prioritize their financial gain over patient well-being and give unnecessary physical tests to patients add to this problem.

As a result, patients face a burden of disease and require an efficient and accurate way of predicting diseases based on symptoms presented, complementing traditional medical

diagnosis methods and providing more data-driven and efficient diagnosis, particularly in remote areas where skilled doctors may not be available.

This user-friendly system not only predicts diseases but also recommends treatment options and preventative measures, empowering patients to take control of their health even in remote locations with limited medical resources. By complementing traditional diagnosis methods with data-driven insights, this project has the potential to significantly improve healthcare outcomes in underserved communities.

#### **1.4 Research questions:**

1. Can machine learning algorithms, particularly hybrid models, accurately predict diseases based on patient symptoms, and how does their performance compare to individual models?
2. How does the integration of a user-friendly interface using the Streamlit framework impact the user experience in disease prediction systems?
3. What is the effectiveness of the system's recommendations for appropriate doctors and preventative measures in improving patient outcomes?
4. To what extent can the system address the challenge of limited access to quality healthcare in remote areas?
5. How does the system's performance vary based on the type and severity of the disease, and can it accurately predict diseases with just a few symptoms provided by patients?

#### **1.5 Aim of the research:**

Addressing the limitations of traditional methods of health care and the people who are deprived of quality healthcare I divide my expected outcome of this system into two categories:

##### **1.5.1 Disease Prediction and Doctor Recommendation:**

- Improved accuracy: By utilizing multiple machine learning algorithms, a hybrid model approach, and prioritizing the symptoms, the system is likely to achieve higher accuracy in disease prediction compared to traditional methods.

- **Reduced uncertainty:** Providing reliable predictions can alleviate anxiety and confusion for patients, enabling them to make informed decisions about their health.
- **Early diagnosis:** Accurate and timely predictions can facilitate early intervention and treatment, potentially leading to better health outcomes and lower healthcare costs.
- **Focus on remote areas:** The system's ability to provide diagnoses without physical contact makes it particularly beneficial for underserved communities with limited access to healthcare professionals.
- **Improved healthcare outcomes:** The combination of accurate diagnosis, timely intervention, and access to specialist care can lead to better patient outcomes and lower disease burden.
- **Reduced healthcare costs:** Early diagnosis and preventative measures can potentially reduce the need for expensive treatments and procedures.

### **1.5.2 User experience:**

- **Enhanced user interaction:** The Streamlit interface allows for a user-friendly and intuitive experience for patients, simplifying symptom input and prediction results.
- **Empowerment and control:** By offering predictions, doctor recommendations, and preventative measures, the system empowers patients to take control of their health and make informed decisions.
- **Increased accessibility:** Accessing healthcare information and advice through a digital platform can remove geographical barriers and provide valuable assistance to those in remote areas.

### **1.6 Proposed solution:**

The proposed solution aims to address the challenges of limited healthcare access by integrating machine learning algorithms and a user-friendly interface. To ensure accurate prediction using the strength of different machine learning models in our system we also integrate a hybrid model approach with multiple machine learning algorithms. To complement traditional diagnosis, bridge the gap in remote healthcare access, and overcome the limitations of existing systems our system prioritizes and analyzes patient

symptoms to improve disease prediction accuracy compared to traditional methods. Focusing on remote areas, and pandemic situations where reaching a doctor may be difficult, our system allows for diagnoses without physical contact through an interactive user interface. We make the interactive interface using the Python Streamlit framework. The provision of our system of reliable disease predictions, doctor recommendations, and preventative measures empowers users to make informed decisions about their health, fostering early intervention and potentially reducing healthcare costs. The Streamlit framework enhances user interaction, simplifying symptom input and prediction interpretation. To ensure reliable system development we collect raw medical data from various sources, undergo data filtration and labeling, and then split the data into training and testing sets. The models are rigorously evaluated to identify the most effective one, and the results are seamlessly integrated into the subsequent module. The workflow encompasses data preprocessing, model training, testing, and evaluation, with continuous refinement to enhance predictive accuracy and reliability, potentially incorporating techniques like cross-validation for robust performance.

## **1.7 Project management and finance:**

### **1.7.1 Planning and scope:**

Clearly define the project's mission: improve healthcare access in underserved communities by providing rapid and accurate diagnoses through machine learning.

Identify key stakeholders: patients, healthcare professionals, government agencies, potential investors.

Break down the project into manageable phases: data collection and cleaning, model development and training, user interface design and implementation, pilot testing, and rollout.

Define success metrics for each phase: accuracy of disease prediction, user satisfaction, adoption rate in target communities.

### **1.7.2 Team and resources:**

Assemble a multidisciplinary team with expertise in machine learning, data science, software development, user experience design, public health, and community engagement. Secure access to high-quality medical data, computing resources, and necessary software licenses.



Establish clear communication channels and responsibilities within the team, utilizing project management tools.

### **1.7.3 Execution and monitoring:**

Establish a detailed project timeline and budget for each phase.

Track progress against the timeline and budget regularly, implementing course corrections as needed.

Conduct risk assessments and develop mitigation strategies.

Ensure quality control throughout the development process with thorough testing and evaluation.

### **1.7.4 Communication and reporting:**

Maintain open and regular communication with stakeholders through progress reports, meetings, and presentations.

Document lessons learned and best practices for future projects.

Develop a communication plan to engage and educate target communities about the project.

### **1.7.5 Finance:**

Explore diverse funding options: grants from government agencies, research institutions, or private foundations; venture capital or angel investors; crowdfunding platforms.

Accurately estimate costs for each project phase, considering personnel, technology, data acquisition, and potential risks.

Allocate budget resources efficiently based on priorities and track expenses meticulously.

Develop a sustainable financial model beyond the initial project phase, exploring potential revenue streams such as subscriptions, partnerships, or data licensing.

## **1.8 Report layout:**

- **Introduction:** Briefly introduce the project, its goals, and its motivation.
- **Background:** Provide context for the problem being addressed. Explain the current challenges in healthcare access and diagnosis, particularly in underserved communities. Briefly mention existing solutions and their limitations.
- **Data Collection:** Describe the process of data acquisition, including data sources, types of data collected, and methods used. Briefly mention any ethical considerations in data collection.

- **Data Preprocessing:** Explain how the collected data was cleaned, prepared, and formatted for analysis and model training. This includes handling missing values, outliers, and ensuring data quality.
- **Research Methodology:** Describe the specific machine learning algorithms and techniques used for disease prediction. Explain the model development and training process, including hyper parameter tuning and evaluation metrics.
- **Experimental Result and Discussion:** Present the findings of the project. This includes the accuracy of disease prediction, performance metrics, and analysis of the results. Discuss the limitations of the model and potential factors influencing the results.
- **Impact on Society and Environment:** Discuss the potential societal and environmental impacts of the project. How will it improve healthcare access for underserved communities? Are there any potential ethical or environmental concerns to consider?
- **Summary, Conclusion, and Future Research:** Briefly summarize the project's key findings and achievements. Draw conclusions based on the results and emphasize the project's significance. Suggest future research directions to improve the model, expand its capabilities, or address additional challenges.
- **References:** List all sources used in the report, including scientific papers, datasets, and any other relevant materials.

### **1.9 Conclusion:**

While limited healthcare access remains a critical challenge, our project proposes a bold solution: leveraging machine learning algorithms and a user-friendly interface to empower individuals and transform healthcare outcomes. Combining early diagnosis, doctor recommendations, and preventative measures, we aim to bridge the gap, especially in underserved areas. With a focus on responsible data practices and seamless integration with existing infrastructure, this project offers the potential to rewrite the narrative, empowering communities, improving health outcomes, and creating a future where quality healthcare is accessible to all. After being successfully implemented, the project holds significant promise in empowering individuals and fostering improved healthcare outcomes, particularly in remote areas.

## CHAPTER 2

### LITERATURE REVIEW

#### 2.1 Introduction:

The landscape of healthcare is undergoing a transformative shift and is fueled by the immense potential of Machine Learning. Artificial intelligence and its subset of machine learning is a blessing, particularly in the realm of disease prediction and doctor recommendation. This chapter enlightens on a comprehensive exploration of the existing literature related to disease prediction. It provides a broad understanding of the diverse approaches, methodologies, and advancements in AI (Machine Learning) powered disease prediction doctor recommendation systems.

As medical data burgeons in volume and complexity, the integration of machine learning algorithms offers unprecedented opportunities to enhance healthcare practices. It also optimizes treatment decisions and empowers individuals with proactive health management. This comprehensive review paints a captivating picture of the immense potential AI (Machine Learning) holds for revolutionizing healthcare. Through this comprehensive study, we acquire numerous benefits and make our system more advanced. Our system can be offered, from early diagnosis and personalized care to streamlined decision-making, empowering both medical professionals and individuals to navigate the healthcare landscape with greater confidence and efficacy.

#### 2.2 Preliminaries/Terminologies:

- **Machine Learning:** A field of computer science that allows computers to learn and improve without being explicitly programmed. This involves algorithms and techniques that enable computers to identify patterns and relationships in data, and then use that knowledge to make predictions or decisions on new data.
- **Hybrid Model:** A machine learning model that combines multiple algorithms or approaches from different branches of machine learning to achieve better performance than any single algorithm could alone. This can be beneficial for tasks where different algorithms excel at different aspects of the problem.

- **Streamlit framework:** An open-source Python framework specifically designed for building and deploying data science applications. It allows developers to create interactive web apps with minimal code, making it ideal for prototyping and showcasing machine learning models. Streamlit handles the backend infrastructure and user interface (UI) development, allowing data scientists to focus on the core logic and data processing.
- **User Interface (UI):** The graphical component of an application that users interact with. It includes elements like buttons, text boxes, charts, and graphs that enable users to input data, receive results, and control the application's functionality. A well-designed UI can make complex machine-learning models more accessible and user-friendly.
- **Pickle Python:** A Python module for serializing and deserializing Python objects. This means it can convert an object like a list, dictionary, or even a custom class into a byte stream for storage or transmission, and then recreate the object from that byte stream later.
- **Joblib:** A Python library for persistent caching and loading of Python objects. It provides functionalities like serialization (similar to pickle), but also offers additional features like memory caching, disk caching, and parallel execution. Joblib aims to be more robust and efficient than Pickle for handling large and complex objects.

### 2.3 Literature review:

[1] This paper proposes a machine learning-based system for disease prediction and doctor recommendation. By analyzing user-input symptoms, algorithms like Naive Bayes and Random Forest predict potential illnesses. Random forests provide the best accuracy of their work around 90.2%. The system then recommends relevant specialists based on location and expertise. This offers benefits like early diagnosis and streamlined doctor-patient matching but faces challenges like data bias, privacy concerns, and limited scope. Future work should focus on integrating clinical knowledge, personalizing models, and addressing ethical considerations for responsible implementation. While promising, this system should complement, not replace, medical professionals' expertise.

[2]This Paper clarifies a prototype for disease prediction and drug recommendation using multiple machine-learning algorithms. Their dataset combines patient symptoms with medical records, offering a richer picture than symptom-only approaches. By comparing Decision Trees, Logistic Regression, Random Forest, and Support Vector Machines, they identified Random Forest as the most accurate predictor. This opens doors for personalized medicine and early intervention. However, data bias and lack of explain ability in algorithms remain challenges. Future work should involve larger, diverse datasets, feature engineering for better model interpretability, and ethical considerations around algorithm fairness and transparency.

[3]Machine learning (ML) is rapidly transforming healthcare, from early cancer detection via blood biomarkers to intelligent data analysis and AI-powered surgery. Smartphone apps and imaging tools powered by ML are revolutionizing diagnosis and treatment, while its tentacles reach beyond healthcare, offering predictive power, efficient asset management, and improved analysis across industries. This evolving technology is saving lives, enabling preemptive diagnosis, and improving patient management. From accelerating drug development to transforming diverse sectors, ML's impact is undeniable, paving the way for a future where it revolutionizes not just healthcare, but the world around us.

[4]In this paper a recent study delves into the potential of collaborative filtering for healthcare, showcasing the "Intelligent Health Recommendation System" (IHRS). Analyzing 10,000 patient ratings across 500 doctors, IHRS leverages big data to predict diagnoses and suggest improved treatments. While promising, the review acknowledges challenges like data security and the evolving nature of healthcare analytics. Future work calls for refined data collection, centralized patient records, and novel methodologies to unlock the transformative potential of data-driven recommendations in advancing patient care.

[5]This review analyzes 48 studies comparing supervised ML algorithms for disease prediction, finding Support Vector Machines (SVM) and Naive Bayes most popular, but Random Forest (RF) most accurate. Researchers can leverage these insights to choose optimal ML algorithms for their disease prediction studies, advancing this promising field. While challenges exist like study heterogeneity and limited scope, future work on meta-

analyses, new algorithms, and standardized metrics can further refine the use of ML for accurate disease prediction.

In this paper, [6] the author presents a "Smart Healthcare Recommendation System" for multidisciplinary diabetes care, revolutionizing personalized medicine. This system delves into the Pima Indian Diabetes Dataset (PIDD), a publicly available but imbalanced dataset containing 768 patient instances, each characterized by 8 numerical features and a binary diabetes diagnosis. By meticulously analyzing this data, the system leverages the synergy of diverse models like CNNs and LSTMs to paint a holistic picture of each patient and predict disease progression with impressive accuracy, exceeding 99.6% in their study. This robust prediction engine then fuels personalized recommendations for interventions, medication adjustments, and lifestyle modifications, empowering patients to actively participate in their health journey. However, the authors acknowledge the limitations of PIDD, including its imbalanced class distribution and potential data biases, which necessitate careful consideration in future refinements of the system.

In this work, [7] the author introduces a novel healthcare system that combines data mining and recommendation systems to predict diseases and recommend suitable doctors and hospitals based on patient reviews. In this research, they use advanced algorithms like Random Forest and K-Nearest Neighbours to predict diseases and recommend personalized healthcare solutions. The approach utilizes electronic medical records and machine learning to enhance patient care and improve doctor-patient matching. Despite promising outcomes, challenges such as data quality limitations and the need for user trust are acknowledged, requiring further research and development. The study underscores the growing role of patient satisfaction in healthcare performance evaluation, emphasizing the importance of data analysis.

This study [8] wanted to find the best way for a medicine recommender system to be both accurate and fast. They tried different models, like ID3 and SVM. ID3 was fast but not very accurate (89%), and SVM was a good balance between accuracy (95%) and speed (0.74 seconds). BP neural network provides the best result about 97%. The researchers built a system with the best model and added a safety check to prevent mistakes. They plan

to improve the system and make it faster for handling lots of data. They wanted a system that gives the right medicine quickly and safely.

The paper [9] proposes a recommendation system called fb-kNN to assist healthcare professionals by suggesting diagnoses and treatments based on a patient's medical profile. Which has computerized health records of 876,084 unique patients from different cities with 57% of the patients being female and 43% being male. Utilizing feature extraction from (Electronic Health Records) EHRs and ECG data, fb-kNN identifies similar patients based on patterns and analyzes their known outcomes. This potentially leads to several benefits: improved diagnosis, personalized treatment, and early intervention. The challenge is to bridge the gap between clinical data and clinicians, requiring the development of a comprehensive big data tool for accurate reporting and intelligent decision support in healthcare, while current recommendation systems primarily focus on disease identification rather than effective feature extraction algorithms. Future work includes validation and testing, clinical workflow integration, explain ability tools, and incorporating additional data sources. The analysis of cases, involving 1056 diagnoses, 845 medicine-related cases, and 910 surgery cases, revealed recommendation accuracies of 93%, 91%, and 95%, respectively. Overall, fb-kNN shows promise in enhancing clinical care but addressing these challenges is crucial to ensure its ethical and successful implementation.

[10] his paper proposes a Disease Diagnosis and Treatment Recommendation System (DDTRS) for medical resource sharing and treatment intelligence in big data and cloud computing environments, utilizing a Density-Peak-based Clustering Analysis algorithm for accurate disease-symptom clustering, defining Disease-Diagnosis and Disease-Treatment association rules, implementing interactive recommendation interfaces, and parallelizing DDTRS on the Apache Spark cloud computing platform for high performance and low latency response based on extensive analysis of large-scale historical inspection datasets. This study finds that for diseases with few treatment stages, traditional classification algorithms, such as C4.5 and RF, exhibit higher accuracy (85.32% to 92.82%), whereas diseases with multiple stages or complex pathogenesis show higher accuracy (up to 88.35%) with clustering algorithms, particularly the DPCA-based disease-symptom

clustering algorithm. Future work should address these issues and rigorously test the system before real-world application.

[11]In this paper provides a comprehensive overview of the application of machine learning in disease prediction, with a specific focus on a Kaggle disease dataset and seven distinct algorithms. In the data set, there were columns like Disease, Symptom1, Symptom2, Symptom3 up to Symptom17. There were a total of 4920 rows and 18 columns in the dataset. The proposed machine learning methods, including Logistic Regression, Random Forest, XGB, KNN, Decision Tree, Naïve Bayes, and SVC, showcase notable benefits in terms of improved accuracy and efficiency, contributing to early disease diagnosis. In this paper, the KNN (K-nearest Neighbors) classification provides the best accuracy which is 94.850949%. However, challenges such as dataset imbalances, null values, interpretability, and computational resource demands are acknowledged, underscoring the importance of addressing these complexities in future research.

[12]This paper proposes a machine learning-based approach for disease prediction using symptoms, achieving an impressive 91.06% accuracy on a dataset of 230 diseases. Their method combines decision trees, Naive Bayes, and random forest algorithms, first identifying candidate diseases via decision trees, then ranking them with Naive Bayes, and finally making the final prediction with random forest. This approach boasts several benefits like speed, scalability, and accuracy, outperforming other machine learning methods like SVMs and neural networks. However, challenges like data quality and symptom ambiguity exist, and drawbacks like interpretability and bias need further work. Future research could focus on improving model interpretability and reducing bias for enhanced clinical applicability.

In this paper [13] Machine Learning, particularly Deep Learning (DL) algorithms are used for their potential in identifying patterns and making predictions for disease detection. ML can tailor diagnoses and treatments to individual patients based on their unique characteristics and assist healthcare professionals in making informed decisions about treatment plans and patient care. However, challenges such as the curse of dimensionality persist, and algorithm-specific variations impact accuracy, exemplified by Support Vector Machines (SVM) showing mixed results. Notably, the J48 algorithm stands out for feature



selection, achieving a commendable accuracy rate of 95.04%. Future research should prioritize refining existing techniques in this rapidly evolving field.

In this paper [14]they propose a hybrid system for chronic disease diagnosis and recommendation, combining decision tree models for prediction with personalized CF-based advice. In this research, RF has demonstrated 99.7% of correctly classified cases. While promising improved accuracy and 24/7 care access. The challenge of this research lies in crafting a dependable hybrid recommender system for Chronic Disease Diagnosis (CDD) that seamlessly merges diverse data mining techniques to ensure accurate predictions and medical advice recommendations for remote patient monitoring. Future work lies in incorporating diverse data, enhancing interpretability, and ensuring ethical considerations for real-world implementation and acceptance. Specific accuracy metrics and details about the dataset are needed for a more thorough evaluation.

In this study, [15]the researchers aim to accurately forecast the onset of chronic diseases based on a high incidence in specific areas. They streamline machine learning methods and utilize data mining to uncover hidden patterns within the large volume of medical data. The proposed system combines unique machine learning and information retrieval (IR) techniques, such as convolutional neural networks (CNNs) and decision trees (DTs), to develop a classification model. The benefits of these techniques include their ability to automatically learn features from raw input, eliminate the need for manual feature engineering, and provide accurate disease predictions. However, the challenge lies in collecting and preparing the data for analysis. The study also mentions the drawbacks of poor information management and the impact it has on data association. In terms of future work, the researchers suggest exploring other data mining methods like random forests and Naive Bayes for disease prediction. The highest accuracy of the models used in the study reached 91.29% with a cross-validation accuracy of 89.1%.

[16]The system recommends disease and suitable medicines and analyzes chemical compositions, offering a promising approach for pharmaceutical development in the face of emerging diseases. In this paper, they use machine learning algorithms for analysis and development. Data cleaning techniques such as outlier detection, are applied to improve the accuracy of AI models. The proposed methodology involves the use of decision trees,

random forests, and naive Bayes classifiers for disease prediction and medicine recommendation. However, challenges remain, such as the time-consuming process of exploring and developing new medicines, and the emergence of new diseases, which pose a burden on the global population. Overfitting is another problem of this work. Future work involves collaboration between pharmaceutical companies, R&D institutions, and medical experts to explore new medicines. The accuracy of the proposed system is determined through testing accuracy, and the results demonstrate the effectiveness of the implemented algorithms and the achieving accuracy of this research is approximately 98%.

[17] The research delves into the exciting potential of data mining and machine learning for revolutionizing healthcare. By analyzing vast medical data with algorithms like KNN and CNN, the proposed system aims to predict diseases with high accuracy (84.5% for CNN), paving the way for personalized drug recommendations and ultimately, improved healthcare decision-making. However, challenges like data quality and potential biases lurk, demanding careful consideration. The research shines a light on the promise of Support Vector Machines, achieving an impressive 99.63% accuracy. Looking ahead, building a user-friendly Android app with deep learning integration and direct hospital data access could further enhance the system's impact on how doctors treat and patients receive care. This research offers a glimpse into a future where data-driven healthcare empowers both medical professionals and individuals, ushering in a new era of personalized medicine.

In this paper, [18] they propose a disease prediction and doctor recommendation system using a combination of Naive Bayes for disease diagnosis and a fuzzy logic approach for doctor recommendation. The system utilizes user-entered symptoms to predict potential diseases and then recommends doctors specializing in those areas based on factors like experience, location, and patient feedback. While the study highlights benefits like increased accessibility to medical information and improved healthcare delivery, it also acknowledges challenges like data accuracy, algorithm bias, and the potential for misdiagnosis due to reliance on symptoms alone. Future work suggestions include integrating medical history and diagnostic tests for better accuracy, as well as addressing ethical concerns surrounding data privacy and transparency. The paper mentions an expected accuracy of over 80% for disease prediction using Naive Bayes.

This research [19] dives deep into diabetes, its types, risk factors, and the potential of analytical tools like MLR and reinforcement learning to predict and manage it. The review delves into type 1, 2, and gestational diabetes, highlighting the prevalence and severity of type 2. It emphasizes lifestyle factors, genetics, and their link to complications like nerve damage and heart disease. Notably, the proposed system uses MLR on patient data to predict diabetes with 83% accuracy, enabling early diagnosis and personalized recommendations. However, data privacy, limited features, and the potential for misdiagnosis pose challenges. Future work aims to integrate reinforcement learning for more adaptable recommendations and address ethical concerns, but further development and rigorous testing are crucial to ensure its safe and effective implementation in healthcare.

In this paper, [20] the author proposes a multi-disciplinary medical treatment decision support system (MDT-DSS) with an intelligent recommendation engine. Utilizing a dataset of patient cases and treatment outcomes, the system leverages machine learning algorithms to suggest optimal treatment plans, considering factors like patient history, diagnosis, and specialist expertise. The benefits include improved treatment quality, reduced decision-making time, and enhanced collaboration among medical professionals. However, challenges remain in data integration, algorithm bias, and user acceptance. Drawbacks include potential overreliance on recommendations and ethical considerations surrounding AI-driven healthcare decisions. Future work involves refining the recommendation engine, incorporating real-time data, and addressing ethical concerns. While accuracy metrics are not explicitly mentioned, the paper highlights the system's effectiveness in achieving high-quality treatment outcomes, suggesting promising potential for clinical adoption.

In this paper [21] enhancing disease diagnosis, researchers have harnessed the capabilities of machine learning to develop a promising predictive system. The approach involves meticulously gathering and preprocessing biomedical and healthcare data, experimenting with various algorithms—including Naïve Bayes, linear regression, decision trees, KNN, Random Forest, and SVM—and ultimately selecting the algorithm that yields the highest accuracy. In this study, Random Forest emerged as the most accurate algorithm, demonstrating a remarkable 98.95% accuracy. Users can interact with the system by simply

inputting their symptoms, and the machine learning model will predict the most probable disease. While this system showcases exceptional potential, it's essential to acknowledge its current limitations, such as its reliance on structured data and a restricted range of predictable diseases. To address these constraints, future research endeavors will concentrate on incorporating unstructured data using natural language processing techniques, broadening the spectrum of predictable diseases, and delving into deep learning approaches for potential accuracy enhancements. This study underscores the transformative possibilities of machine learning in the healthcare domain, paving the way for more precise and personalized patient care.

In their paper, [22]"Human Disease Prediction using Machine Learning Techniques and Real-life Parameters," Gaurav et al. (2023) propose a novel approach to predict diseases using a combination of machine learning and real-life data. They leverage a medical dataset, weighting symptoms based on their diagnostic value and pre-processing the information for optimal model performance. Their system integrates three algorithms: Random Forest, LSTM, and SVM, to analyze patient symptoms and predict potential diseases. Their method achieves an impressive 97% accuracy, effectively leveraging the dataset's characteristics. While acknowledging the potential for further improvement with time-series data integration, the authors highlight the proposed method's superiority compared to its predecessors. Notably, SVM (76% and 90% accuracy) and Logistic Regression (75% accuracy) faced limitations like unsuitability for multi-parameter scenarios and overfitting concerns, respectively. However, limitations such as potential overfitting and dependence on data quality require further investigation. The authors envision enhancing the model by incorporating more real-life parameters and exploring advanced deep-learning techniques. Ultimately, this research paves the way for earlier diagnoses, personalized treatment plans, and a more proactive approach to healthcare through the power of machine learning.

In the realm of healthcare, where precise disease prediction can lead to better outcomes, harnessing the vast amount of medical data is crucial. [23]This paper proposes a disease prediction system that leverages machine learning algorithms to achieve this goal. By collecting and analyzing a comprehensive dataset of diseases and their corresponding

symptoms, the system trains various machine-learning models to identify patterns and relationships. Upon user input of their symptoms, the system employs the best-performing model to predict the most likely illness. While promising, the authors acknowledge limitations such as data quality, limited scope, and potential for false positives and negatives. Drawing upon a meticulously cleaned dataset acquired from a credible source, models were developed for each algorithm. While Multinomial Naive Bayes achieved an accuracy of 92% and Logistic Regression reached 89%, it was the Decision Tree algorithm that demonstrated the highest accuracy, reaching an impressive 97%. To address these, they propose future advancements like expanding the data set, incorporating user-specific factors, and integrating with electronic health records. Ultimately, this system has the potential to empower individuals to be more proactive about their health, while recognizing the need for responsible use and further research to refine its accuracy and functionality.

#### **2.4 Comparative Analysis and Summary:**

I assessed some previous work related to human disease prediction. From this work, I explored how different machine-learning algorithms predict disease based on medical data. I found SVM (Support Vector Machine) was the most common tool they used, but Random Forests were the most accurate (84.5%-98%). A clear picture emerged from all previous work that AI-powered disease prediction and recommendation systems hold immense potential for healthcare offering benefits like early diagnosis, personalized care, and streamlined decision-making. However, challenges like data bias, explainability gaps, and ethical concerns remain. Future work emphasizes richer data, smarter algorithms, and building trust for responsible implementation in healthcare. While some systems already achieve impressive accuracy (e.g., 99.63% with SVM), further development and rigorous testing are crucial to ensure their safety and effectiveness in real-world settings. This exciting field holds the promise to revolutionize healthcare, empowering both medical professionals and individuals. Let's compare the different systems side-by-side. See the table below for more details:

Table 2.4: Comparison Between Related Work of Disease Prediction.

<b>Paper Title</b>	<b>Author</b>	<b>Method</b>	<b>Best Algorithm</b>	<b>Best Accuracy</b>
[1] Disease Prediction and Doctor Recommendation System Using Machine Learning Approaches.	Kumar, A., Sharma, G.K. and Prakash, U.M.	Logistic Regression, Random Forest Classification, Naïve Bayes, KNN	Random Forest Classification	90.2%
[2] An Intelligent Disease Prediction and Drug Recommendation Prototype by Using Multiple Approaches of Machine Learning Algorithms.	Nayak, S.K., Garanayak, M., Swain, S.K., Panda, S.K. and Godavarthi, D.	Naïve Bayes, Extra Tree Classifier, Decision Tree, SVM.	Decision Tree Classifier	89.93%
[3] Significance of machine learning in healthcare: Features, pillars and applications.	Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, Shanay Rab	Machine Learning	Machine Learning	Not Defined

[4] Intelligence-Based Health Recommendation System Using Big Data Analytics.	Sahoo, A.K., Mallik, S., Pradhan, C., Mishra, B.S.P., Barik, R.K. and Das, H.	Decision tree, Artificial neural network (ANN), Bayesian classifier, Logistic Regression, Hybrid Model	Hybrid Model	Not Defined
[5] Comparing different supervised machine learning algorithms for disease prediction	Uddin, S., Khan, A., Hossain, M.E. and Moni, M.A.	Artificial neural network(ANN), Decision tree (DT), K-nearest neighbour (KNN), Logistic regression (LR), Naïve Bayes (NB), Random forest (RF), Support vector machine (SVM)	Random Forest Classifier	93%
[6] A Smart Healthcare Recommendation System for Multidisciplinary Diabetes Patients with Data Fusion Based on Deep Ensemble Learning.	Ihnaini, B., Khan, M.A., Khan, T.A., Abbas, S., Daoud, M.S., Ahmad, M. and Khan, M.A.	LR, Naïve Bayes, Random forest, K-nearest neighbors, Decision tree, Support vector machine, Ensemble Learning (combining multiple classifiers)	Ensemble Learning (combining multiple classifiers)	99.6%

[7] An Approach For Prediction Of Diseases To Suggest Doctors And Hospitals To Patient Based On Recommendation System	Shashidhar, V., Kubear, P.A., Manoj, M. and Jalapreetha, J.	KNN, Random Forest, Naïve Bayes (NB	KNN	96%
[8] An intelligent medicine recommender system framework	Bao, Y. and Jiang, X.	SVM, DT, Neural Network	SVM	95%
[9]Recommendation system using feature extraction and pattern recognition in clinical care systems	Uzair Aslam Bhatti, Mengxing Huang, Di Wu, Yu Zhang, Anum Mehmood & Huirui Han	Fb-KNN, BR-KNN, ML-KNN	Fb-KNN	(91-95)%
[10] A disease diagnosis and treatment recommendation system based on big data mining and cloud computing.	Chen, J., Li, K., Rong, H., Bilal, K., Yang, N. and Li, K.	C4.5, Random Forest, K-means, DCPA	Random Forest	92.82%



[11] Disease Prediction and Treatment Recommendation Using Machine Learning	Mahata, S., Kapadiya, Y.B., Kushwaha, V., Joshi, V. and Farooqui, Y.	LR, RF, XGB (Extreme Gradient Boosting), KNN, Decision Tree classification, Naïve Bayes, SVC (Support Vector Machine).	KNN (K-nearest Neighbors) classification	94.85%
[12] Disease prediction from various symptoms using machine learning	Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., Warang, M. and Mehendale, N.	Naive Bayes, SVM, Random Forest, Decision trees, RUSBoost algorithm, Weighted KNN	Weighted KNN	99.7%
[13] Human Diseases Detection Based On Machine Learning Algorithms: A Review	Nareen O. M. Salim & Adnan Mohsin Abdulazeez	Support Vector Machines, k-nearest neighbors Logistic Regression (LR), Decision trees (DTs), Naive Bayes classification, Deep Learning (DL), Convolutional	J48	95.04%

		Neural Networks (CNNs), Lasso Regression, J48		
[14] Efficient Chronic Disease Diagnosis Prediction and Recommendation System	Hussein, A.S., Omar, W.M., Li, X. and Ati, M.	J48, Decision Stump, REP, RF	Random Forest	99.7%
[15] Human Disease Prediction And Doctor Booking System.	Mr.Joel Roy, Mr. Reeju Koshy, Mr. Roshan Roy, Ms. Anjumol Zachariah	Naïve bytes, Decision tree, Logistic regression, Random forest	Random forest	96%
[16] A Computer-Based Disease Prediction And Medicine Recommendation System Using Machine Learning Approach	Gupta, J.P., Singh, A. and Kumar, R.K.	Decision Tree, Random Forest, Naive Bayes	Naive Bayes	98.12%
[17] Disease Prediction and Medication Advice Using Machine	Pooja Panapana, K. Sri Rakesh Reddy,	Naive Bayes, Gaussian Naive Bayes (GNB), Logistic Regression, K Nearest Neighbour	Support Vector Machine	99.63%

Learning Algorithms	J. Deepika, G. Rushivardhan Babu4, and A. Drakshayani	(KNN), Support Vector Machine (SVM).		
[18] Disease Prediction and Doctor Recommendation System	Dhanashri Gujar, Rashmi Biyani, Tejaswini Bramhane, Snehal Bhosale, Tejaswita P. Vaidya	Naïve Bayes	Naïve Bayes	80%
[19] An Approach for Developing Diabetes Prediction and Recommendation System	Saima Sultana, Abdullah Al Momen, Mohoshi Haque, Mahmudul Hasan Khandaker, Nazmus Sakib	Multiple Linear Regression (MLR) has been used	Multiple Linear Regression (MLR) has been used	83%

[20] A Decision Support System with Intelligent Recommendation for Multi-disciplinary Medical Treatment	KUNWEI SHEN, XIAOSONG CHEN, and SIJI ZHU NENGJUN ZHU, JIAN CAO	k-NN Classifiers and the Rule-based Method	k-NN Classifiers	94%
[21] The Prediction Of Disease Using Machine Learning.	Dr C K Gomathy, Mr. A. Rohith Naidu	Decision Tree, Random Forest, Naïve Bayes, SVM, KNN	Random Forest	98.95%
[22] Human Disease Prediction using Machine Learning Techniques and Real-life Parameters.	K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar*, T. Suryawanshi	Naive Bayes Classifier, Weighted KNN, SVM, Logistic Regression(LR), Random Forest	Random Forest	97%
[23] Multiple Disease Prognostication Based On Symptoms Using Machine Learning Techniques	Kajal Patil, Sakshi Pawar ,Pramita Sandeep and Jyoti Kundli	Naive Bayes, Logistic Regression, Decision Tree	Decision Tree	97%

## **2.5 Scope of the Problem:**

The scope of this problem lies in the limitation of its previous work and lack of access to quality healthcare. The project aims to improve healthcare outcomes for people in developing countries and remote areas by making quality healthcare more accessible and affordable. Traditional medical diagnosis methods rely on clinical experience and knowledge, which can be limited and subject to human error. The paper aims to address this problem by proposing a machine learning-based system for disease prediction and doctor recommendation. The system utilizes algorithms and techniques to analyze patient symptoms and make accurate predictions about potential diseases. By providing early diagnosis and timely intervention, the system can improve healthcare outcomes and reduce the burden of disease on individuals and society as a whole. On the other hand, through this noble work, public health can benefit from outbreak prediction, targeted resource allocation, and personalized health education. Addressing ethical concerns like bias and privacy will be crucial. As AI seamlessly integrates with healthcare systems, we can envision a future where disease prevention and treatment become proactive and personalized, leading to a healthier world.

## **2.6 Challenges:**

Despite the promising potential of this project to bring healthcare closer to those who need it most, several challenges stand in its way. Firstly, the quality and diversity of the data used to train the machine learning models are crucial. Biases within the data can lead to inaccurate predictions, potentially creating new healthcare disparities. Additionally, collecting comprehensive medical data from underserved communities faces limitations, hindering the models' generalizability and effectiveness. Secondly, building trust in a machine-made diagnosis can be difficult for patients, potentially delaying or even neglecting necessary medical attention. This is especially concerning in areas with limited access to healthcare professionals, where overreliance on the system could mask serious underlying conditions. Thirdly, ethical considerations like data privacy, algorithmic bias, and transparency in the decision-making process need careful attention to ensure fair and equitable access to healthcare for all. Finally, technical challenges arise from the need to continuously update the models with new medical information and ensure accurate

functionality even in resource-constrained environments with limited internet and computing power. Integrating the system seamlessly with existing healthcare infrastructure and medical professionals will be crucial for wider adoption and impact. Successfully navigating these challenges will be key to unlocking the project's true potential and achieving its goal of improving healthcare access and outcomes for underserved communities.

## **2.7 Conclusion:**

The studies presented showcase the significant strides made in leveraging advanced machine learning algorithms such as Naive Bayes, Random Forest, Support Vector Machines, and Decision Trees. It revolutionizes medical diagnosis and treatment recommendations. They also use some deep learning methods. Machine learning algorithms, particularly Random Forests and SVM, have demonstrated remarkable accuracy in predicting diseases, reaching up to 99% accuracy. The literature also highlights several challenges that must be addressed for the responsible implementation of these systems. Issues such as data bias, privacy concerns, limited scope, and algorithmic transparency emerge as critical considerations. Future work is consistently emphasized to focus on refining models by incorporating clinical knowledge, personalizing recommendations, and addressing ethical considerations to ensure the safe and effective integration of these technologies into healthcare practices. So, the literature review lays a solid foundation for understanding the landscape of machine learning applications in healthcare, revealing both the tremendous potential and the challenges associated with these advancements.

## CHAPTER 3

### RESEARCH METHODOLOGY

#### 3.1 Introduction:

In addressing the pressing issue of limited access to timely and suitable healthcare, particularly prevalent in underserved communities of developing nations, this research endeavors to contribute to the improvement of healthcare access through the development of a human disease prediction and doctor recommendation system. This introduction provides a contextual backdrop, elucidating the challenges posed by traditional medical diagnosis methods and the urgent need for a more efficient, data-driven approach. The proposed methodology encompasses a comprehensive framework, encompassing research subjects and instrumentation, workflow, data collection procedures, statistical analysis, and data pre-processing. Emphasizing the significance of diverse data sources and careful consideration of ethical aspects, the methodology integrates machine learning algorithms for disease prediction, complemented by a doctor recommendation system. The evaluation process involves rigorous performance assessment and analysis of experimental results. Furthermore, the research endeavors to maximize societal impact by ensuring the developed system is not only accurate but also accessible through a user-friendly web application, thereby aiming to contribute significantly to enhancing healthcare outcomes in developing nations.

#### 3.2 Research Subject and Instrumentation:

##### 3.2.1 Research Subject:

This research paper is dedicated to improving healthcare access in underserved communities, specifically in developing countries where quality healthcare is often scarce. Utilizing machine learning algorithms for disease prediction, the study aims to overcome challenges in timely and accurate diagnoses. Additionally, the research focuses on providing reliable healthcare during pandemics and for isolated populations. The system developed not only predicts diseases but also recommends doctors, offers disease descriptions and provides precautionary measures. A key objective is to ensure a simple

user interface, facilitating usage across various sectors, and addressing the needs of healthcare professionals in identifying and managing unknown diseases. I set a target community for my research based on some factors like disease prevalence, demographics, data availability, technology access, and community needs. Prioritize a specific disease or group of diseases relevant to the chosen population, consider their cultural background and technological literacy for interface design, and ensure your system aligns with local healthcare priorities for maximum impact and scalability. Remember, ethical considerations and community engagement are crucial for responsible implementation and long-term sustainability.

### **3.2.2 Instrumentation:**

This research delves into utilizing machine learning for rapid and accurate diagnoses in underserved communities. I used Python as my main tool. To explore and prepare my data, I leverage various machine learning libraries such as to visualize insights and patterns, I used libraries like Seaborn and Matplotlib. Data cleaning, manipulation, and analysis find a steadfast ally in pandas, while NumPy effortlessly handles complex numerical computations. The versatile sci-kit-learn library empowers me with a diverse array of machine-learning algorithms. In this work I utilized diverse arsenal of algorithms – Random Forests, Decision Trees, Logistic Regression, and more – to analyze patient data gleaned from symptoms and medical histories. Techniques like missing value imputation and feature engineering refine the data for optimal model performance. Accuracy, reliability, and efficiency are scrutinized through various metrics, comparing individual and hybrid models built from these algorithms. Critical symptoms are prioritized via encoding techniques, while ethical concerns like data privacy, security, and potential biases are meticulously addressed to prevent unequal access or discrimination. Google Colab serves as my cloud workspace, Anaconda Navigator managing my Python environment, Spyder and Jupyter Notebook hosting the interactive coding journey. The research further unveils a user-friendly web app, built with the “Streamlit” framework, ensuring broad accessibility across sectors and backgrounds. I have done all this work on my MacBook Air with M1 chip. But to ensure the pervasiveness of my work I coded efficiently that can be able to run smoothly in low configuration device (Core i3) without any complexity.



### 3.3 Workflow:

#### 3.3.1 General Flow diagram of methodology

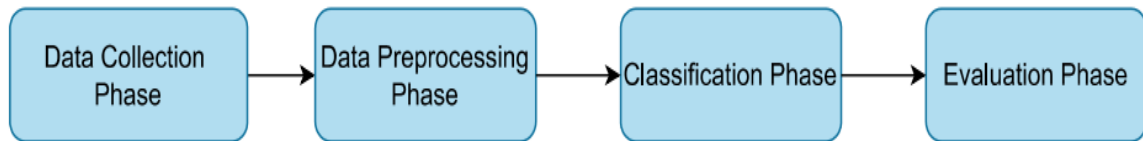


Figure 3.3.1: General Flow diagram of Methodology

#### 3.3.2 Workflow diagram:

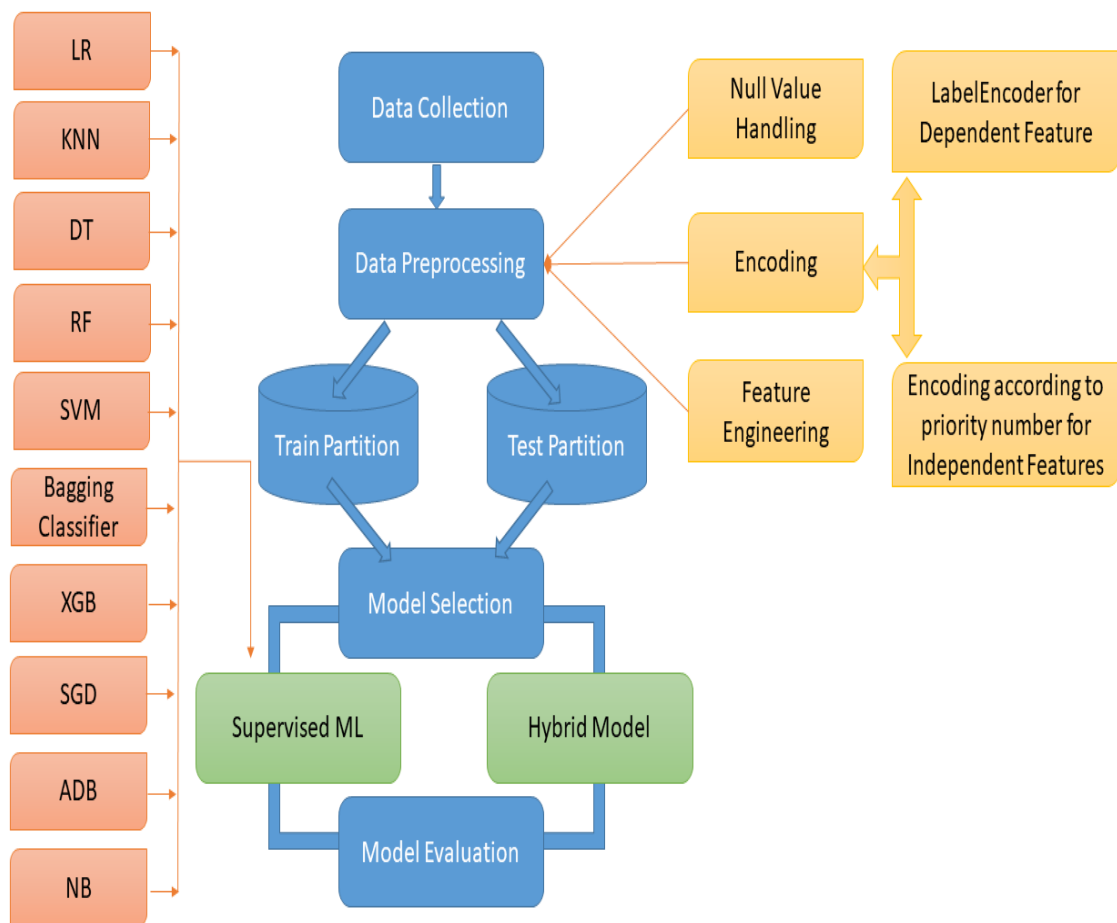


Figure 3.3.2: Work Flow diagram of Methodology

#### 3.4 Data Collection Procedure/Dataset Utilized:

My research draws data from diverse sources, weaving a rich tapestry of medical information. The foundation lies in a knowledge base built from New York Presbyterian

Hospital discharge summaries, offering disease-symptom associations and ranked symptom co-occurrences. That is cited in [24]. I further enrich this with specialized datasets from Kaggle [25] and authentic Bangladeshi internet sources, gathering disease descriptions, and precautions. To suggest the best doctor name for specific diseases in Bangladesh I collect it from "doctorbangladesh" websites cited in [26] This meticulously collected data, boasting 4920 rows and 18 columns (17 symptoms, 1 disease), holds 41 unique diseases and 131 unique symptoms, ready to be explored and analyzed. With this diverse set in hand, I can now delve into exciting research avenues, from visualizing disease-symptom patterns to building predictive models and crafting a doctor recommendation system tailored to the Bangladeshi context.

This study utilizes a supervised learning approach to predict disease based on symptom data. Before applying machine learning algorithms, a data preprocessing and exploratory data analysis (EDA) stage is implemented. This stage cleans the dataset by removing noise and outliers, ensuring the algorithms perform optimally on the processed data. By thoroughly preparing the data before applying the algorithms, this approach aims to achieve accurate disease predictions based on symptom information. Some samples of my data set is given below-

Table 3.3.1: Sample Dataset of Symptom and Disease

Symptom_1	Symptom_2	...	Symptom_18	Disease
itching	skin_rash	...	dischromic_patches	Fungal infection
stomach pain	spotting urination	...	skin rash	Drug Reaction
joint pain	vomiting	...	dark urine	Hepatitis A
high fever	red sore around nose	...	yellow crust ooze	Impetigo

Table 3.3.2: Sample Dataset of Disease and Specialist

<b>Disease</b>	<b>Specialist</b>
Fungal infection	Dermatologist
Allergy	Allergist
GERD	Gastroenterologist
Chronic cholestasis	Hepatologist

Table 3.3.3: Sample Dataset of Disease and Description

Disease	Description
Drug Reaction	An adverse drug reaction (ADR) is an injury caused by taking medication. ADRs may occur following a single dose or prolonged administration of a drug or result from the combination of two or more drugs.
Malaria	An infectious disease caused by protozoan parasites from the Plasmodium family that can be transmitted by the bite of the Anopheles mosquito or by a contaminated needle or transfusion. Falciparum malaria is the most deadly type.

Table 3.3.4: Sample Dataset of Disease and Precaution

<b>Disease</b>	<b>Precaution_1</b>	<b>Precaution_2</b>	<b>Precaution_3</b>	<b>Precaution_4</b>
Drug Reaction	stop irritation	consult nearest hospital	stop taking drug	follow up
Malaria	Consult nearest hospital	avoid oily food	avoid non veg food	keep mosquitos out
Allergy	apply calamine	cover area with bandage	use ice to compress itching	-
Hypothyroidism	reduce stress	exercise	eat healthy	get proper sleep

Table 3.3.5: Sample Data set of Disease and Doctor name

<b>Disease</b>	<b>Doctor</b>
Drug Reaction	Prof. Dr. M. A. Hasanat/Labaid Specialized Hospital   Dhanmondi, House: 1 and, 6 Road No 4, Mirpur Road, Dhanmondi, Dhaka, 1205, Bangladesh
Malaria	Dr. Nikhat Shahla Afsar  MBBS, MD (Internal Medicine) Internal Medicine /Evercare Hospital Dhaka  Plot 81, Block: E, Bashundhara R/A, Dhaka, 1229, Bangladesh
Allergy	Asso. Prof. Lt. Col. Dr. Moyassaque Ahmed

	MBBS, DDV, FCPS (Dermatology), Fellowship in Dermatology Laser Surgery (Skin Allergy & Sex disease (CMH) Dermatologist/Popular Diagnostic Centre Ltd.   Uttara (Unit 2) House # 25, Road # 7, Sector # 4, Jashim Uddin Moar, Uttara, Dhaka, Dhaka, 1230, Bangladesh
--	---

### 3.5 Statistical Analysis:

Our analysis begins with exploring the data structure and distribution. In this research-based project, we use five different data sets. Such as: 1. symptoms and disease data set; 2. disease and specialist data sets; 3. disease and description; 4. disease and its corresponding precautions; 5. disease and doctor name. The first one, which is the symptoms and disease dataset, holds 4920 instances with 18 columns. This data set is collected from New York-Presbyterian Hospital, which is cited in [24]. Which hold 17 categorical symptom columns, and the final column categorizes disease with 41 unique classes. Each disease appears roughly 120 times on average in the predicted class. In this data set, the total number of unique symptoms found is 131. The below bar chart shows the statistical insights of the symptoms and disease data set:

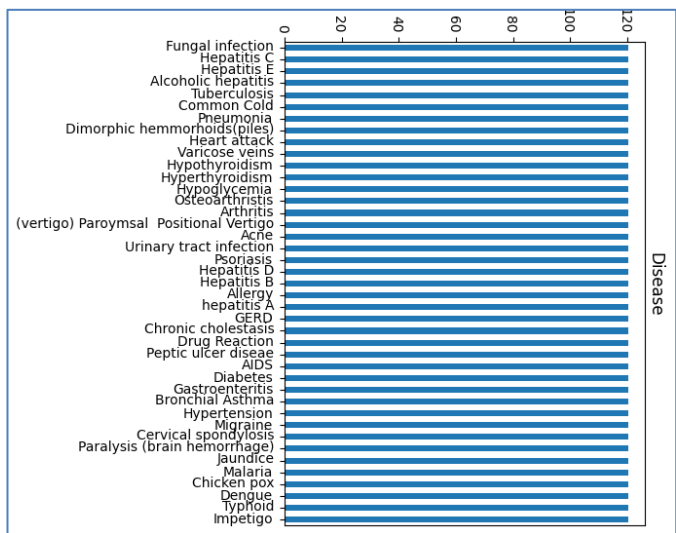


Figure 3.4.1: Bar chart of disease and number of occurrences in the data set.

The symptoms and disease dataset shape are (4920, 18) with 41 unique disease classes. To avoid inconsistency, overfitting, and under-fitting problems with our machine learning model, we made sure each disease had the same number of cases (120). In the below pie chart, each slice represents a different disease. We can see that all the slices are the same size (2.44%, or 8.78 degrees). This helps create a fair test for our model to learn from the data.

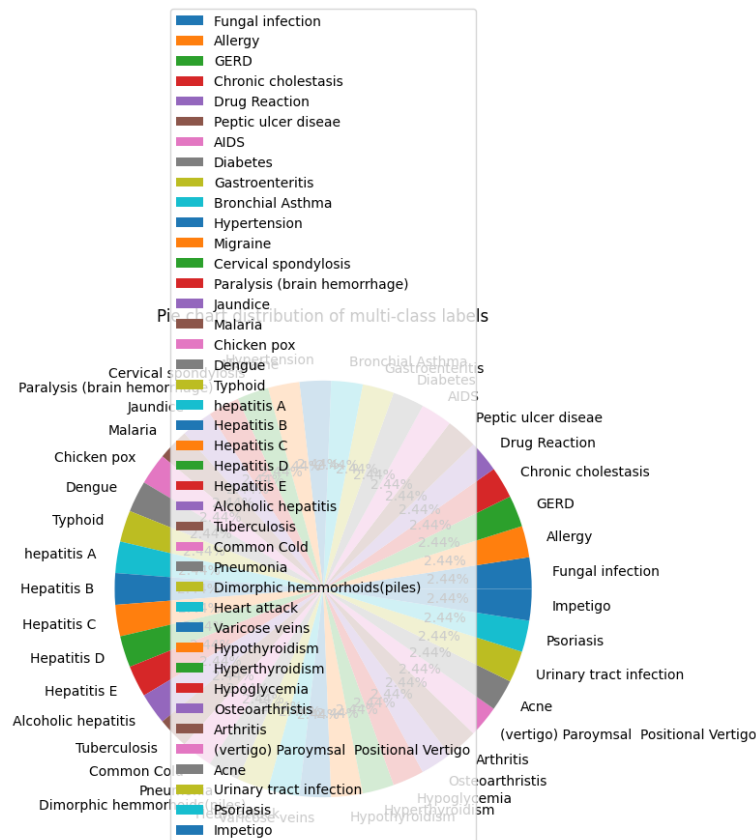


Figure 3.4.2: Pie chart of unique diseases

I previously mentioned that the first data set (symptoms and disease) contains 131 different symptoms listed as categories in a total of 17 symptom columns for each patient. The most common symptoms are fatigue, vomiting, and high fever, each occurring 1932, 1914, and 1362 times, respectively, in patients. Other frequent symptoms include loss of appetite, nausea, headaches, abdominal pain, etc. Less common symptoms like muscle wasting, patches in the throat, and a foul smell of urine still appear in hundreds of patients. Top 20 symptoms in data set is given below bar chart:

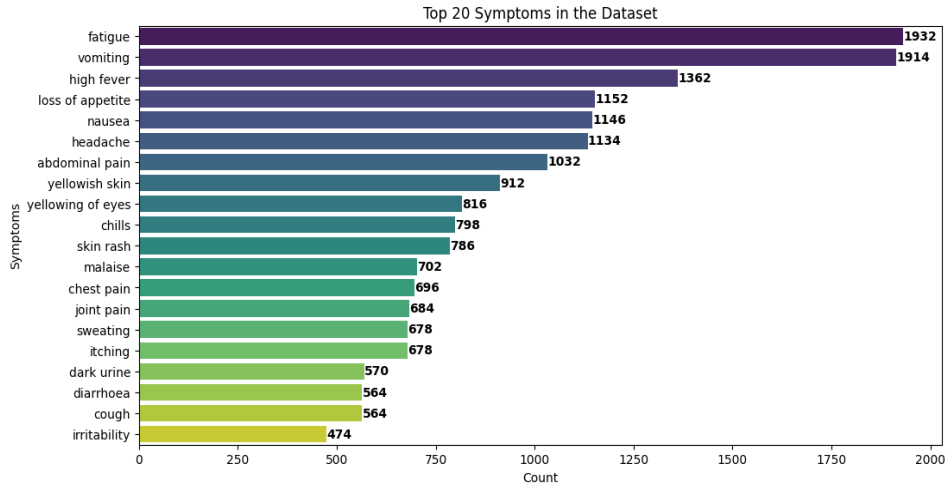


Figure 3.4.3: Count of top 20 symptoms in disease and symptoms data set.

The other data set that was mentioned before has 42 rows and 2 columns, except for the fourth one, the disease and precautions data set, whose shape is (42, 5). The disease and specialist data set contains 19 different specialists for 41 diseases. Hepatologist shows up more times for different diseases. The disease and doctor data set offers insights into doctor allocations and specialties. The disease and description data set allows for word frequency and text similarity analysis.

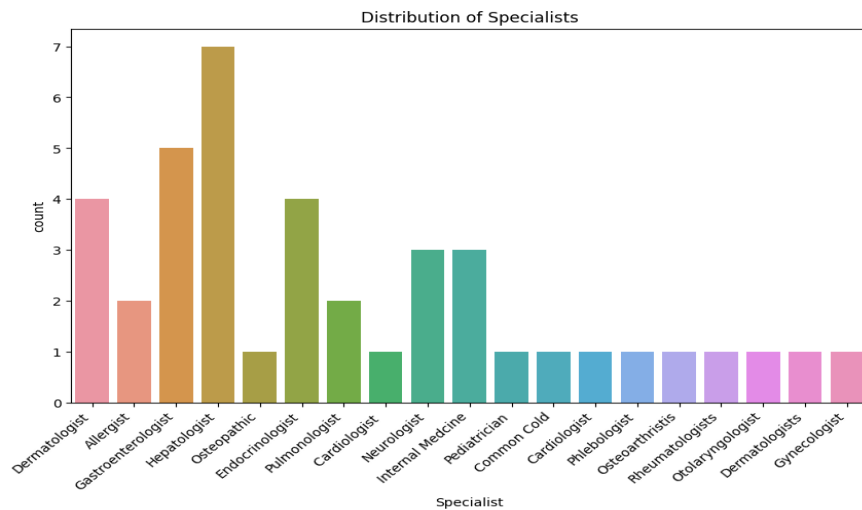


Figure 3.4.4: Count specialist for different disease.

### 3.6 Data Pre-processing:

Pre-processing is a fundamental requirement of Machine Learning models since it gets the data ready to fit in the model for classification techniques to be applied. Down sampling, feature extraction, encoding, feature engineering, and scaling are all parts of the preparation process. The effectiveness of the acquired models is highly contingent on the excellence of the training data. Poor models are inevitable regardless of the classifier inducer used when inaccurate training data is employed.

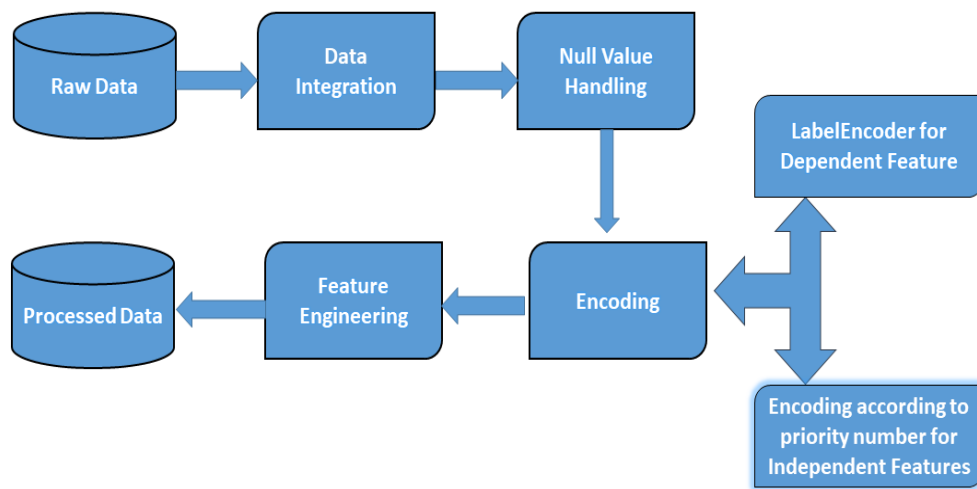
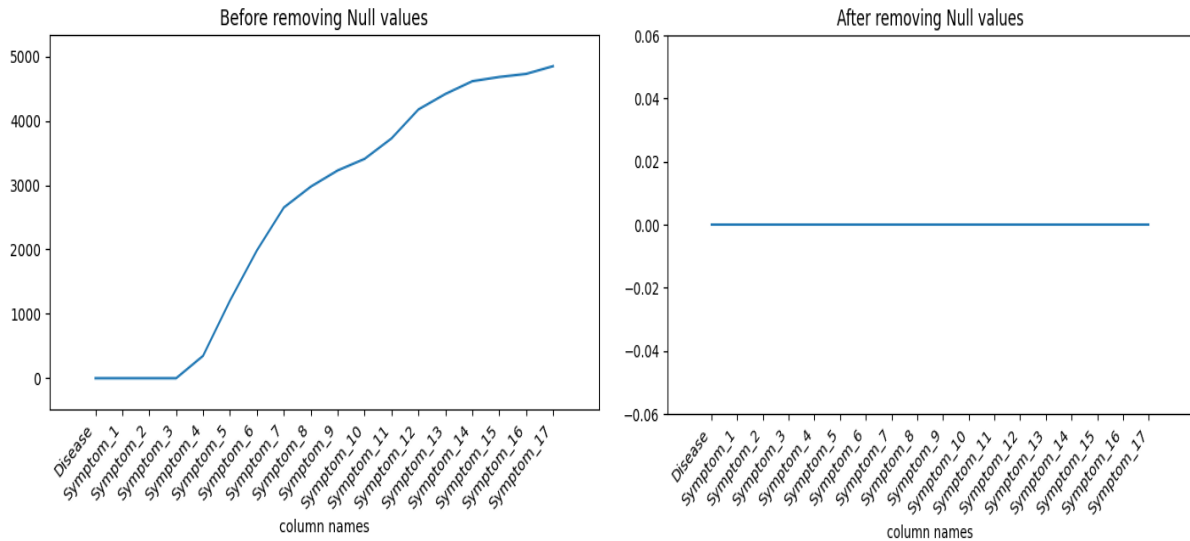


Figure 3.6.1: Data preprocessing flow diagram

Initially, we assess the null values within the dataset, revealing that a significant number of columns contain missing values. These null values have an impact on the accuracy of the machine learning model's performance. So our first requirement is to fill in the null value. But our data set is not like the traditional data that we easily drop null values rows or fill it by using mean or median. The first problem is our data set contains object-type data and the second problem is that it is a medical data set and we use it for predicting disease so we should not fill the null value by traditional object-type data set filling like- which symptom occurrence most of the time in data set fill the null value by the symptom. We have to decide on a different approach as our data set is a sensitive data set in medical related domain. We fill the null value by 0.





. Figure 3.6.2: Before and after removing null values

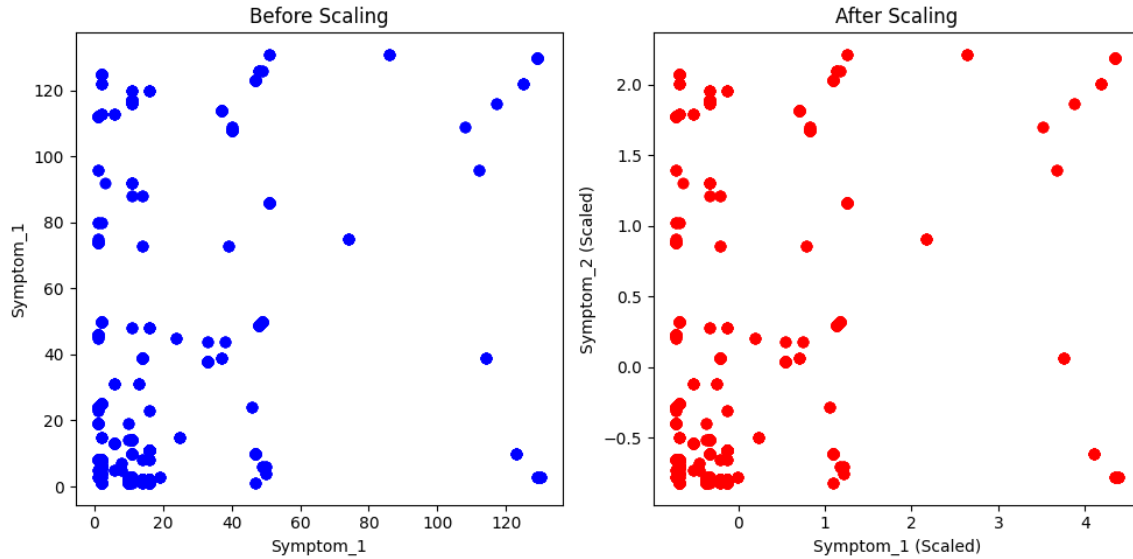
In our "Symptoms and disease" data set it contains all object type data so after handling the null values we encode its data instance. As our independent column has no definite class label so cannot use traditional encoding methods like- Label Encoder, One Hot encoder, or Ordinal encoder. As encoding is mandatory because machines cannot understand any data except numeric we encode our data set. However, we follow two different approaches for encoding- 1. We encode our data using an external data set that contains the symptoms and their corresponding weight in numeric. 2. We first prioritize symptoms data by their total number of occurrences in the data set. Which is the top of the occurrences list we encode it as 1. Then we encode the other 2,3,4... respectively.

For encoding the "Disease" column which is the dependent variable and predicted by the independent variable "Symptom" columns we use Label Encoder. The Label Encoder is incorporated by the scikit-learn library. It is employed for converting categorical (non-numeric) data into numerical labels. This serves as a typical preprocessing procedure in machine learning since numerous algorithms mandate input data to be in numerical format.

Due to the varying scales (0-137) of symptoms in the "Symptoms and Disease" dataset, standard scaling is crucial. This technique transforms [27] features to have a mean of 0 and a standard deviation of 1, ensuring all features contribute equally and boosting model accuracy, training speed, and interpretability. By putting symptoms on common ground,

standard scaling empowers the model to learn effectively from the data and deliver accurate disease predictions.

Figure 3.6.3: Before and after removing scaling



### 3.7 Proposed Methodology:

This research investigated and categorized symptoms to predict specific diseases and for guiding the users a doctor recommendation component is incorporated that suggests the appropriate specialist based on the predicted disease. In this case to accurately predict and obtain a high degree of accuracy it is necessary to use a fresh dataset and relevant attributes extracted from it. Data collecting, data pre-processing, lexical feature engineering, machine learning modelling, and cross-dataset analysis are the main components of the working technique. Our goal is to develop a system that enhances disease prediction accuracy, simplifies healthcare decision support, and guides patients toward relevant medical expertise. To achieve the goal, the system ensures consistency through the four primary phases: Data Collection, Pre-processing, Classification, and Evaluation.



Figure 3.7.1: General Methodology for Machine Learning

Our research-based project can be divided into two phases that is disease prediction and doctor recommendation and user interface (web application). In this chapter, we elaborately discuss phase one. The first phase has also some subcategories-

1. Disease Prediction.
2. Doctor recommendation.
3. Disease description.
4. Precautions for this disease.
5. Doctor name suggestions.

### **3.7.1 Data Collection and Pre-processing:**

In the previous section, we already elaborately discussed data collection and pre-processing. In short, the research employs a diverse dataset sourced from New York Presbyterian Hospital, Kaggle, and Bangladeshi internet platforms, consisting of 4920 rows and 18 columns. Utilizing a supervised learning approach, the study predicts diseases based on symptom data preceded by meticulous data pre-processing and exploratory data analysis. In this research Null values are filled with 0 due to the sensitivity of medical data, and symptom encoding involves both an external dataset that contains the weight of symptoms and prioritization based on occurrences. The "Disease" column is encoded using Label Encoder, and to address varying symptom scales, standard scaling is applied. This comprehensive approach aims to optimize the dataset for accurate disease predictions, incorporating varied sources and innovative pre-processing techniques for robust and interpretable machine learning models.

### **3.7.2 Splitting the data set:**

Ensuring a model that can accurately predict diseases beyond the data it is crucial to train the model first. This is where data splitting comes in, dividing the pre-processed dataset into two distinct sets: training partition and testing partition.

The dataset, consisting of 4920 rows and 18 columns, is partitioned into an 80% training set and a 20% testing set. The training set, comprising 80% of the data, is used to train the machine learning model. During this process, the model learns the patterns and

relationships within the data, adjusting its parameters to make accurate predictions. Training set shape is (3936, 17).

On the other hand, the testing set, representing the remaining 20% of the data, serves as an unseen dataset that the model has not encountered during the training phase. This separation is essential for assessing the model's generalization performance—its ability to make accurate predictions on new, previously unseen data. Testing set shape is (984, 17).

The 80-20 split is [1] a common practice in machine learning to strike a balance between having enough data to train a robust model and having a sufficient portion reserved for evaluation. It helps prevent overfitting, a situation where a model performs well on the training data but fails to generalize to new data.

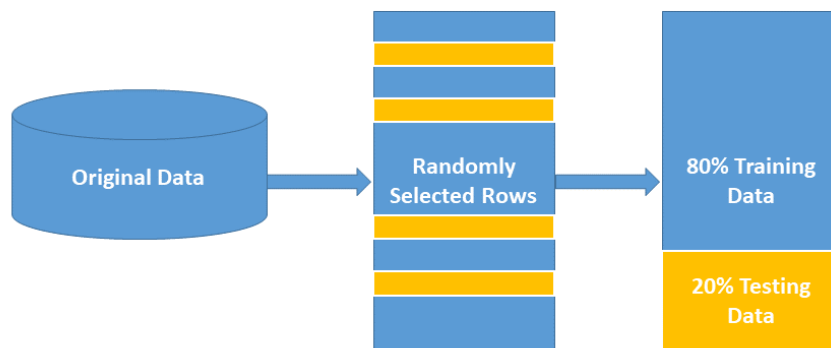


Figure 3.7.2: Randomly Split the data set

### 3.7.3 Classification Phase:

In our project focused on predicting diseases and recommending doctors, we utilized a labelled dataset [28] divided into training and testing sets for supervised learning. Employing ten different classifier algorithms, including Support Vector Machine, Decision Tree, Random Forest, K Nearest Neighbors, Logistic Regression, Naive Bayes, and Extreme Gradient Boosting, we thoroughly explored the performance of each method. The chosen algorithms cover a spectrum of machine-learning techniques to ensure a comprehensive evaluation. Additionally, we incorporated a hybrid model (a combination of machine learning models) that consists of four machine learning models that provide the best accuracy. This holistic approach aims to identify the most effective model for accurate

predictions and doctor recommendations. Model deployment and the outcomes of the evaluation is discussed below.

### 3.7.3.1 Logistic Regression:

Logistic regression is a statistical method used for predicting the probability of an event occurring. It is particularly well-suited for situations where the dependent variable is binary, meaning it has two possible outcomes (usually coded as 0 and 1). However, when dealing with more than two categories, as in the case of our dataset with 41 predicted classes, Multinomial logistic regression is the appropriate choice. This type of logistic regression accommodates scenarios where the dependent variable has three or more unordered categories, making it suitable for classification problems with multiple classes. The primary goal of logistic regression, whether binary or multinomial, is to estimate the relationship between one or more independent variables and the probability of a particular outcome. In the context of our dataset, this would involve understanding the relationship between the independent variables and the 41 predicted classes. With an achieved accuracy of 84.14%, demonstrates the proficiency of logistic regression in classifying.

The following equation represents logistic regression:

$$y = \frac{e^{(b_0 + b_1x)}}{1 + e^{(b_0 + b_1x)}}$$

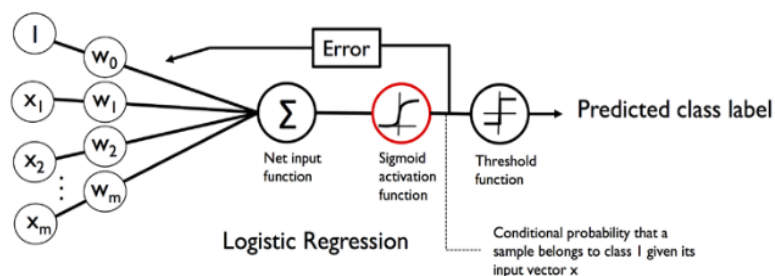
here,

x = input value

y = predicted output

b<sub>0</sub> = bias or intercept term

b<sub>1</sub> = coefficient for input (x)



### 3.7.3.2 KNeighborClassifier:

The KNeighbor technique stands out as a widely employed algorithm for both classification and regression tasks. It measures the distance between the dependent variable (the desired outcome) and independent factors (features) by employing the Euclidean distance formula. The KNeighborsClassifier is a straightforward and useful tool in machine learning for sorting things into categories. The 'k' in KNeighborsClassifier is about deciding how many neighbors to consider when making a decision. It's easy to use, good for smaller datasets, and doesn't require a lot of fancy setup. But be cautious with really big datasets or tricky features, as it might get a bit slow or confusing. The K nearest neighbors, denoted by the letter K in this classifier's name, are determined by the client and represent an integer value. Therefore, as the headline suggests, this classifier uses training dependent just on KNeighborClassifier. The appropriate value of k is determined by the data. This method resulted in an accuracy of 100%.

For choosing neighbor this algorithm calculates distance of the neighbor using these formulas-

$$\text{Euclidean: } \sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

$$\text{Manhattan: } \sum_{i=1}^k |x_i - y_i|$$

$$\text{Minkowski : } (\sum_{i=1}^k (|x_i - y_i|^q))^{1/q}$$

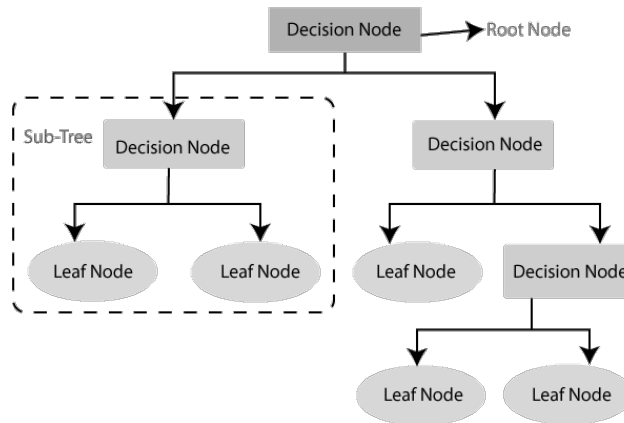
### 3.7.3.3 DecisionTreeClassifier:

In classification, decision trees play a vital role by using a flowchart-like structure. At each node, attribute conditions are tested, guiding the tree's branches. The ultimate leaf node assigns a class label based on the calculated attributes. Decision trees excel in showcasing key factors, revealing interrelationships, and aiding in data exploration for accurate predictions, making them crucial in predictive modelling scenarios.

Decision Tree is useful for predicting diseases based on symptoms because it mimics human decision-making and provides an easily understandable tree-like structure for interpreting the logic behind the predictions. It is integrated into our model and provides the greatest results. By utilizing Attribute Selection Measures (ASM) to divide the records into the best attributes, the Decision Tree Classifier picks the best attribute, creates a

decision node out of it, and divides the dataset into much more manageable portions. Then, until one of the requirements is fulfilled, it recursively repeats these steps for each offspring to start building a tree—

- The identical attribute value should apply to each of the tuples.
- There are no additional attributes left.
- There are currently no further instances.



### 3.7.3.4 RandomForestClassifier:

One of the supervised ML approaches is termed RF, and it consists of a number of random trees. The widely used decision tree approach is extended by the decision tree algorithm RF by mixing several decision trees. This approach aims to lower the inventive decision tree's variance. RF is one of the most often used classifiers since it is simple to use and flexible enough to handle both classification and regression issues. It also provides improved accuracy in our model. The Random Forest (RF) algorithm is a compelling choice for predicting diseases based on symptoms in your dataset. Its ensemble learning approach involves combining the predictions of multiple decision trees, enhancing accuracy and reliability. This versatility allows RF to handle both classification and regression tasks, aligning well with the nature of predicting diseases from symptom data. The algorithm's effectiveness in managing high-dimensional datasets is particularly advantageous when dealing with numerous symptoms. To combat overfitting, RF employs techniques such as bootstrap sampling and random feature selection during tree construction, contributing to a more generalized model. The inherent ability to assess

feature importance is valuable for gaining insights into the influential symptoms in disease prediction.

### **3.7.3.5 Support Vector Machine:**

SVM is a powerhouse in the machine learning world, tackling tasks like text and image classification, spam detection, and even DNA analysis. Its secret weapon? Finding the best-dividing line in high-dimensional data, keeping different classes well away from each other. This flexibility, plus its ability to handle complex relationships and thrive with limited data, makes SVM a go-to choice for a wide range of challenges.

In predicting diseases based on symptoms, the utilization of Support Vector Machines (SVMs) is a judicious choice. SVMs are particularly advantageous in scenarios where the relationship between symptoms and diseases is non-linear, as they can effectively handle such complexities through the use of kernel functions. The algorithm's robustness to outliers is crucial in medical applications, accommodating instances where unusual symptom patterns might exist. Additionally, SVMs excel in high-dimensional data spaces, making them suitable for datasets with numerous symptoms. Their ability to perform well with limited training data is valuable in medical contexts, where acquiring extensive labelled datasets can be challenging. The versatility of SVMs, applicable to both classification and regression tasks, further enhances their suitability for predicting diseases based on symptoms. The algorithm's adherence to the maximum margin principle contributes to robust generalization, a critical aspect of disease prediction and doctor recommendation. If SVMs have demonstrated success in analogous medical prediction tasks, their application becomes even more compelling, albeit with considerations for computational costs and parameter tuning.

### **3.7.3.6 Bagging Classifier:**

Bagging is an ensemble learning technique that trains multiple models on random subsets of data to improve accuracy and reduce overfitting. The final prediction is made by combining the predictions of these individual models, using majority voting for classification and averaging for regression.

Predicting diseases based on symptoms is a critical task in medicine, and choosing the right machine-learning model is crucial for success. In this case, the Bagging Classifier stands out for several compelling reasons.



Firstly, Bagging excels in classification problems, perfectly aligned to assign disease labels based on symptom data. Secondly, medical data can be noisy and complex, leading to high variance in individual models. Bagging's ability to average out this variance makes it ideal for such scenarios. Bagging stabilizes learners by averaging their predictions, leading to more robust and reliable outcomes. Medical data often boasts a multitude of features (symptoms). Bagging's proficiency in handling high-dimensional data prevents overfitting, a common pitfall in such cases. Additionally, improved accuracy is paramount in predicting diseases accurately, and Bagging frequently outperforms individual classifiers in this regard.

So, training multiple base models in parallel, a strength of Bagging, accelerates the model-building process significantly, especially with massive datasets, making it highly practical for real-world applications.

#### **3.7.3.7 eXtreme Gradient Boosting:**

XGBoost, short for eXtreme Gradient Boosting, stands as a titan in the machine learning arena. Its prowess lies in its ability to weave together individual, modest learners (think decision trees) into a tapestry of remarkable accuracy. This ensemble, built layer upon layer with a technique called boosting, continuously refines its predictions, correcting the missteps of its predecessors.

Selecting XGBoost (BXGB) for predicting diseases based on symptoms in the dataset is a judicious choice due to its versatility, efficiently handling a mix of categorical and numeric features. Its scalability proves beneficial for datasets of considerable size, ensuring swift training and improved model performance. Incorporating regularization techniques, such as L1 and L2 regularization, addresses overfitting concerns, crucial in medical data analysis. XGBoost's reputation for accuracy and competitive performance in machine learning competitions underscores its reliability in predicting diseases accurately from symptom data. The built-in feature importance analysis offers valuable insights into the influential symptoms, aiding both model interpretation and guiding future data collection. With its customizability through a wide range of parameters, XGBoost adapts well to the specific characteristics of the dataset, enhancing its suitability for the prediction task, which, falling within the realm of classification, aligns seamlessly with XGBoost's versatile capabilities.

### 3.7.3.8 AdaBoostClassifier:

An AdaBoost classifier started by training classifiers to the original sample and was referred to as a thematic. The same dataset is then fitted using several instances of the classifier, but with the weight of instances that were mistakenly categorized being changed so that fairly late classifiers would focus more on difficult situations.

The equation of AdaBoostClassifier -

$$H(x) = \text{sign} \left( \sum_{t=1}^T \alpha_t h_t(x) \right)$$

### 3.7.3.9 SGD Classifier:

The Stochastic Gradient Descent-Classifer (SGD-Classifer) is an SGD-optimized linear classifier (SVM, logistic regression, etc.). SGD permits minibatch (online/extracurricular) learning. SGD makes sense to utilize in large-scale situations because of how effective it is. Since it is impossible to determine the minimum cost function of the Logistic Regression directly, we attempt to minimize it using the Stochastic Gradient Descent method, commonly referred to as Online Gradient Descent. SGD Classifier should be used instead of SVM or logistic regression if you are unable to maintain the data in RAM. SGD Classifier, however, is still in operation.

### 3.7.3.10 Naive Bayes classifiers:

[29] Naive Bayes classifiers are a group of algorithms that use Bayes' Theorem for classification, assuming features are independent. They're known for being simple, effective, and fast at building and using machine learning models.

The selection of the Naive Bayes (NB) classifier for predicting diseases based on symptoms was motivated by its application to text data, as indicated by the use of the multinomial Naive Bayes variant. This classifier demonstrated high accuracy in the dataset under consideration, effectively anticipating the division of data into distinct groups. The decision to employ NB was further influenced by its simplicity and efficiency, making it an accessible and rapid solution for training and prediction tasks. Notably, the scalability of NB to handle large datasets and high-dimensional features aligns well with the characteristics of the symptom and disease dataset. The assumption of conditional independence between features, though simplistic, did not hinder its effectiveness in this context. Overall, the practical advantages of NB in terms of accuracy, simplicity, and

scalability make it a suitable choice for predicting diseases based on symptoms in the specific dataset analysed.

#### **3.7.3.11 ExtraTreesClassifier:**

An ensemble machine learning system called Extra Trees Classifier combines the forecasts from numerous decision trees. It has a connection to the common Random Forest algorithm. Even though it uses a simpler technique to create the decision trees that are utilized as members of the ensemble, it can frequently produce a performance that is as good as or better than the random forest algorithm. The Extra Trees method produces a significant amount of extremely randomized decision trees from the training dataset. Regression predictions are created by aggregating the results of decision trees, whereas categorization expectations are created via a qualified majority.

The Gaussian Nave Bayes classifier presumes that the data behind each label originates from a simple Distribution function. Scikit-learn provides `sklearn.naive_bayes.GaussianNB` to perform the Gaussian Naive Bayes classification method. The simplest and fastest classification technique known is called naive Bayes, and it excels at handling massive amounts of data. Detect phishing, text categorization, computational linguistics, and decision support systems are just a few of the areas where the Naive Bayes classifier has succeeded. In order to predict outcomes regarding unknown classes, the Bayes theory of probability is utilized. The Bayes Theorem has an impact on Naive Bayes, a straightforward yet powerful stochastic classification paradigm for machine learning.

#### **3.7.4 Evaluation criteria**

To compare the effectiveness of several machine learning classifiers, matrices are utilized. For this study, we compared and evaluated the model's performance using the confusion matrix, accuracy, precision, recall, f1 score, macro average, and weight average. A four-way table of categorization and prediction is called a confusion matrix.

**i) True positive (TP):** predict benign URLs as benign correctly

**ii) False positive (FP):** Predict benign URLs as benign Incorrectly

**iii) True negative (TN):** predict malicious URLs as malicious correctly

**iv) False negative (FN):** predict malicious URLs as malicious incorrectly

Accuracy is one metric for evaluating classification models.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision, or the caliber of a successful prediction generated by the model, is one measure of a machine learning model's effectiveness.

$$\text{precision} = \frac{TP}{TP + FP}$$

The recall is the percentage of accurately forecasted classes among all the positive classes and is computed by-

$$\text{Recall} = \frac{TP}{TP + Fn}$$

F-score aids in gauging recall and precision at the same time and is computed by-

$$F - \text{score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

### **3.8 Implementation Requirements:**

For implementing our work Google Colab is enough. Python libraries needed are given below.

- Python 3.12
- Pandas
- Matplotlib
- Seaborn
- Scikit-Learn
- NumPy
- Spyder
- Jupiter Notebook
- Anaconda
- Streamlit
- Joblib

### **3.9 Conclusion:**

In conclusion, the research methodology adopted for this project presents a robust framework aimed at enhancing healthcare access in underserved communities, particularly in developing nations. Leveraging machine learning algorithms, the methodology focuses on expediting and improving the accuracy of diagnoses, especially crucial during pandemics and for isolated populations. The comprehensive approach encompasses key stages such as research subject identification, instrumentation, workflow development, data collection, statistical analysis, and data pre-processing. By integrating ethical considerations and community engagement, the methodology ensures a responsible and community-centric approach. The incorporation of data scaling, encoding, and feature engineering techniques enhances the precision of disease predictions, while the inclusion of a doctor recommendation component and a user-friendly web application facilitates reliable healthcare decision support. This research methodology not only addresses the limitations of traditional medical diagnosis methods but also underscores the significance of ethical practices, sustainability planning, and societal and environmental impact considerations, thereby contributing to a meaningful improvement in healthcare outcomes in developing nations.

# CHAPTER 4

## DISEASE PREDICTION AND DOCTOR RECOMMENDATION, SYSTEM ANALYSIS AND DESIGN SPECIFICATION

### 4.1 Introduction:

This chapter explores our novel disease prediction and doctor recommendation system designed to improve healthcare access, particularly in resource-limited settings. The system leverages machine learning models trained on diverse medical data to predict diseases based on user-provided symptoms and recommend suitable doctors. It delves into the model's workflow, from data preparation and model selection to implementation requirements, paving the way for a data-driven healthcare experience that empowers individuals to take charge of their health. The chapter not only details the model but also provides a roadmap for implementation, elucidating the essential tools, technologies, and resources required.

### 4.2 Workflow:

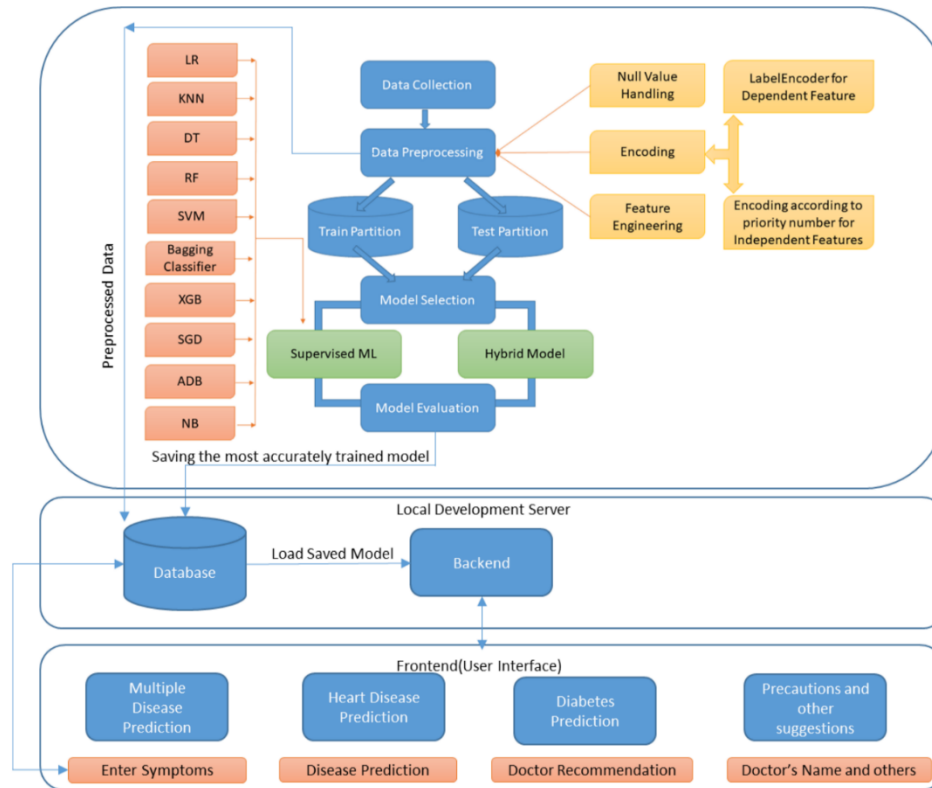


Figure 4.2: Work flow diagram for disease prediction and doctor recommendation system

### 4.3 Proposed Model:

The disease prediction and doctor recommendation model involves a systematic process. Starting with the collection of raw data from various hospitals and trusted internet sources. Covering diseases like Fungal infection, Allergy, GERD, Chronic cholestasis, Drug reactions, etc. This raw data undergoes a filtration and labeling step, followed by a split into training and testing sets (commonly 70/30 or 80/20 ratios), and is stored in an ML database. Three four distinct models that provide high accuracy including- Random Forest, Bagging Classifier, Decision trees, and KNeighborsClassifier - are employed for disease prediction. The models are rigorously evaluated to identify the most effective one, and the results are seamlessly integrated into the subsequent module, ensuring that the optimal model is utilized for further processing or decision-making. This comprehensive workflow encompasses data preprocessing, model training, testing, and evaluation, with a continuous focus on refinement to enhance predictive accuracy and reliability, potentially incorporating techniques like cross-validation for robust performance.

- 2 Random Forest: Utilizing an ensemble of decision trees, Random Forest aims to enhance predictive accuracy and mitigate overfitting.
- 3 Decision Tree: A tree-like model that partitions data into subsets based on specific criteria, aiding in precise disease prediction.
- 4 Bagging Classifier: Employing bootstrap aggregating, this model combines multiple base learners to improve overall accuracy and robustness.
- 5 KNeighborsClassifier: Operating on the principle of proximity, this model classifies diseases based on the majority class of their neighboring instances.

### 4.4 Implementation Requirements:

- **Python Environment:** Set up a Python environment on a laptop or PC, and consider using virtual environments to manage dependencies.
- **IDEs and Notebooks:** Utilize development tools such as Google Colab, Jupyter Notebook, Spyder, or Anaconda Navigator based on preference for the development.
- **Select Machine Learning Model:** Choose a framework like Scikit-learn, TensorFlow, or PyTorch for building the disease prediction model.

- **Train and Save the Model:** Train the machine learning model using platforms like Google Colab or Jupyter Notebook. Save the trained model in a deployment-compatible format (e.g., pickle, joblib).
- **Install Streamlit framework:** Install Streamlit in the Anaconda Navigator environment. In this case, I create a custom environment named machine-learning. Streamlit installation command: **pip install streamlit**
- Also, need to install other features of streamlit based on requirements such as Navbar, Menu bar, Sidebar, etc.
- **pickle5:** Pickle is a [30] module in Python that provides a way to serialize and deserialize objects. Serialization is the process of converting a Python object into a byte stream, and deserialization is the reverse process. Installation Command:
  - **pip install pickle5**
- **Streamlit App Script:** Develop a Streamlit script (e.g., app.py) that loads the trained model and creates the user interface. Implement Streamlit widgets to gather user input for symptoms and display predictions.
- **Compatibility and Specification:** To minimally run Anaconda Navigator, Jupyter Notebook, Spyder, and Streamlit, aim for Windows 7/macOS 10.10 (64-bit), an Intel Core i3/AMD Ryzen 3 processor, 8GB RAM (16GB preferred), an SSD with 50GB free space, and a dedicated graphics card (2GB+ memory) for data-heavy tasks, though not essential for basic use. In this case, I use a MacBook air with m1 chipset.

#### 4.5 Conclusion:

This chapter has laid the groundwork for developing an interactive web app using the Python framework Streamlit, designed to bring our disease prediction and doctor recommendation system to life. By harnessing the power of machine learning and crafting an intuitive user interface, we aim to empower individuals to actively engage with their health, receive personalized predictions, and conveniently connect with qualified healthcare professionals. This system holds the promise to transform healthcare access, particularly in resource-limited settings, enabling early diagnosis, promoting informed decision-making, and ultimately improving healthcare outcomes for all.



## CHAPTER 5

### DISEASE PREDICTION AND DOCTOR RECOMMENDATION, SYSTEM IMPLEMENTATION AND TESTING

#### 5.1 Introduction:

In this chapter unfolds the practical realization of a cutting-edge Disease Prediction and Doctor Recommendation System, an innovative web-based tool aimed at empowering individuals in proactive health management. We embark on a journey through the system's architecture, highlighting its key components and compatibility requirements. Dive into the intricacies of the user interface design, showcasing a visually appealing and user-friendly platform crafted with Python's Streamlit framework and CSS. The chapter also emphasizes the critical aspect of performance evaluation and testing, ensuring the system's robustness and reliability under varying conditions. Join us in unraveling the meticulous development and testing journey, paving the way for a future where technology plays a pivotal role in guiding individuals toward informed healthcare decisions.

#### 5.2 Experimental Setup:

The disease prediction and doctor recommendation system is a web app that is developed using a Python framework "Streamlit".

In this app, I do some customization using "CSS". So first needs to be a web browser that supports CSS and streamlet and python language. like- Google Chrome, Mozilla Firefox, Safari, etc.

The experimental setup for the Disease Prediction and Doctor Recommendation System involves the utilization of a Python environment, specifically within frameworks like Google Colab, Jupyter Notebook, Spyder, or Anaconda Navigator. The machine learning model, constructed using frameworks such as Scikit-learn, TensorFlow, or PyTorch, is trained and saved in a deployment-friendly format like Pickle or joblib. The web application is developed using the Streamlit framework with customizations using CSS for an enhanced user interface. Compatibility is ensured with modern web browsers supporting

CSS, Streamlit, and Python, such as Google Chrome, Mozilla Firefox, and Safari. The system is designed to run optimally on Windows 7/macOS 10.10 (64-bit) with an Intel Core i3/AMD Ryzen 3 processor, 8GB RAM (16GB preferred), an SSD with 50GB free space, and a dedicated graphics card with 2GB+ memory for data-heavy tasks. Thorough testing, documentation, and optional deployment on hosting platforms complete the setup, offering a comprehensive solution for disease prediction and doctor recommendations.

### 5.3 Development of User Interface:

#### 5.3.1 Server:

The server module of the human disease prediction and doctor recommendation system acts as a data hub and prediction engine. At its core, uses a central database that stores preprocessed data sets for the training of machine learning models, doctor information data, precautions data sets, specialist data sets, disease description, and machine learning modes that are already trained by preprocessed data. As an experimental setup, I use my laptop as a local server. In this case, security, scalability, and monitoring round out the module, ensuring data protection, high performance, and smooth operation. In short, the server module is the command center, fueling the entire disease prediction system.

Name	Date Modified	Size	Kind
colab files	6 Dec, 2023 7:30 PM	--	Folder
Multiple disease prediction system - diabetes.ipynb	21 May, 2022 4:38 PM	35 KB	Document
Multiple disease prediction system - heart.ipynb	21 May, 2022 4:38 PM	40 KB	Document
Multiple disease prediction system - Parkinsons.ipynb	21 May, 2022 4:38 PM	49 KB	Document
dataset	8 Dec, 2023 5:13 PM	--	Folder
Extra Data	8 Dec, 2023 5:13 PM	--	Folder
diabetes.csv	8 Feb, 2021 4:03 PM	24 KB	Comm...et (.csv)
heart.csv	25 Mar, 2021 10:28 AM	11 KB	Comm...et (.csv)
parkinsons.csv	5 May, 2021 9:23 AM	41 KB	Comm...et (.csv)
My Data set	8 Dec, 2023 10:38 PM	--	Folder
dataset.csv	24 May, 2020 1:20 AM	632 KB	Comm...et (.csv)
Doctor_Name.csv	8 Dec, 2023 5:50 PM	7 KB	Comm...et (.csv)
link.txt	13 Jul, 2023 11:12 PM	96 bytes	Plain Text
specialist.csv	8 Dec, 2023 10:37 PM	1 KB	Comm...et (.csv)
Specialist.xlsx	18 Jun, 2023 9:50 PM	2 MB	Micros...k (.xlsx)
symptom_Description.csv	24 May, 2020 1:20 AM	11 KB	Comm...et (.csv)
symptom_precaution.csv	24 May, 2020 1:20 AM	3 KB	Comm...et (.csv)
Symptom-severity.csv	24 May, 2020 1:20 AM	2 KB	Comm...et (.csv)
saved models	8 Dec, 2023 11:09 PM	--	Folder
Disease	8 Dec, 2023 11:09 PM	--	Folder
disease_model_bagg.sav	8 Dec, 2023 11:06 PM	876 KB	Document
disease_model_dt.sav	8 Dec, 2023 11:06 PM	78 KB	Document
disease_model_rf.sav	8 Dec, 2023 11:07 PM	9.4 MB	Document
diabetes_model.sav	21 May, 2022 10:16 AM	28 KB	Document
heart_disease_model.sav	21 May, 2022 10:19 AM	1 KB	Document
multiple_model_bagg.sav	7 Dec, 2023 1:08 PM	875 KB	Document
multiple_model_rf.sav	7 Dec, 2023 1:09 PM	9.4 MB	Document
multiple_model.sav	4 Dec, 2023 8:31 AM	80 KB	Document
parkinsons_model.sav	21 May, 2022 10:19 AM	13 KB	Document
background.png	Today 10:30 AM	2.7 MB	PNG image
multiple disease pred.py	Today 5:59 PM	36 KB	Plain Text
streamlit run command.txt	6 Dec, 2023 7:37 PM	103 bytes	Plain Text

Figure 5.3.1: Local Server containing necessary file for disease prediction and doctor recommendation system

### **5.3.2 Front End or User Interface:**

In the [31]front-end module, there are different tabs or we can say different sections. In every section, there are different disease and doctor recommendation modules. For example, there is a disease prediction, disease and doctor recommendation, heart disease, diabetes prediction, etc. tab. The front-end side user can also get advice from doctors The system also provides the disease description and precautions of the doctor. For emergency cases, it also suggests the doctor's name. Front-end design with Python, Streamlit framework, and CSS. The front-end side first requires importing saved files and models from server side using pickle or joblib [32].

#### **5.3.2.1 Home page:**

The homepage serves as a centralized hub with distinct sections such as Disease Prediction, Doctor Recommendation, and Authentication System. Utilizing Streamlit's simplicity, we import saved files and models from the server side for a smooth front-end experience. Interactive elements, coupled with CSS styling, enhance the user-friendly layout, offering features like predictive analytics, personalized doctor recommendations, and expert advice. The design ensures privacy with a robust Authentication System and includes emergency assistance and educational content, creating a dynamic and visually appealing platform for users to navigate their health journey efficiently.

This is the Home page of the Disease Prediction and Doctor Recommendation System. Here we have brief information about Different Disease Prediction. Data Collection, Doctor Recommendation, and Authentication System We also provide some advice on the home page.

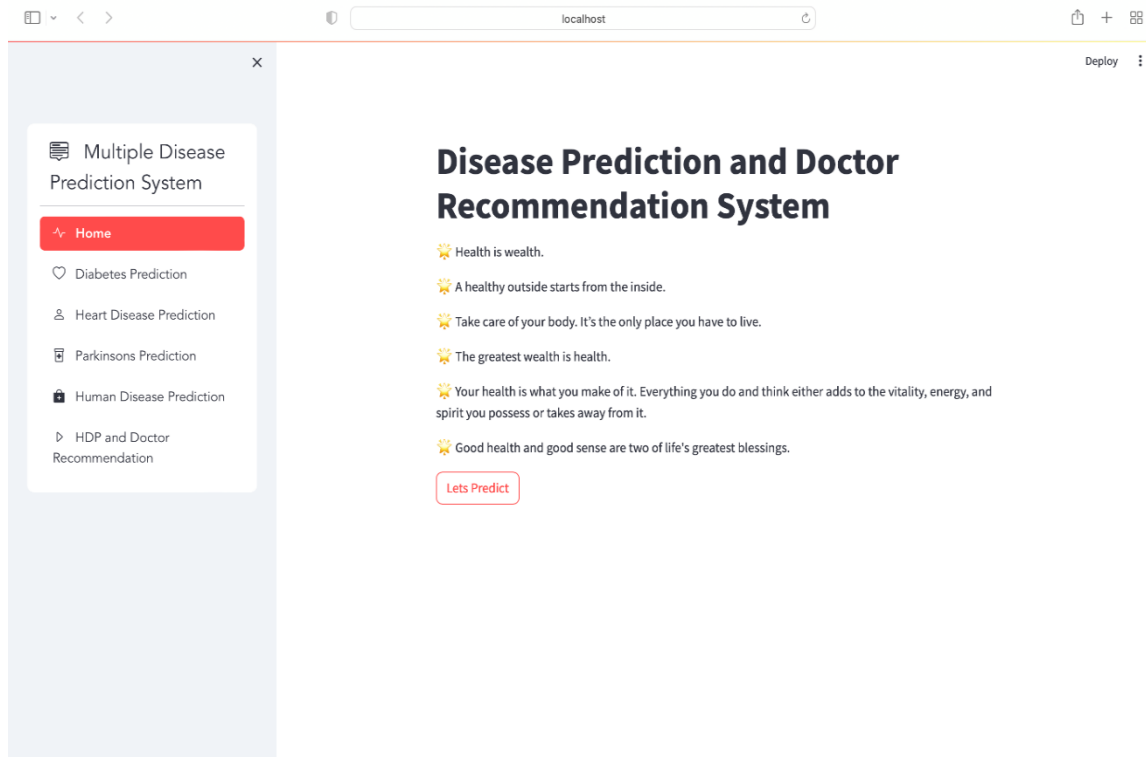


Figure 5.3.2.1: Home Page disease prediction and doctor recommendation system

### 5.3.2.2 Disease Prediction page:

The Disease Prediction page is designed to provide an interactive and accurate health assessment for users. Upon entering the page, users can input their symptoms, and the system enhances user experience by suggesting relevant symptoms dynamically as the user types, based on a pre-existing database. Leveraging three distinct pre-trained machine learning models ensures comprehensive disease prediction, fostering user trust in the system's accuracy. The collaborative use of multiple models allows for a more robust and reliable prediction, maximizing the likelihood of accurate health assessments. Users benefit from a seamless and intuitive experience, ultimately enhancing the effectiveness and reliability of the Disease Prediction page as a valuable tool for health self-assessment.

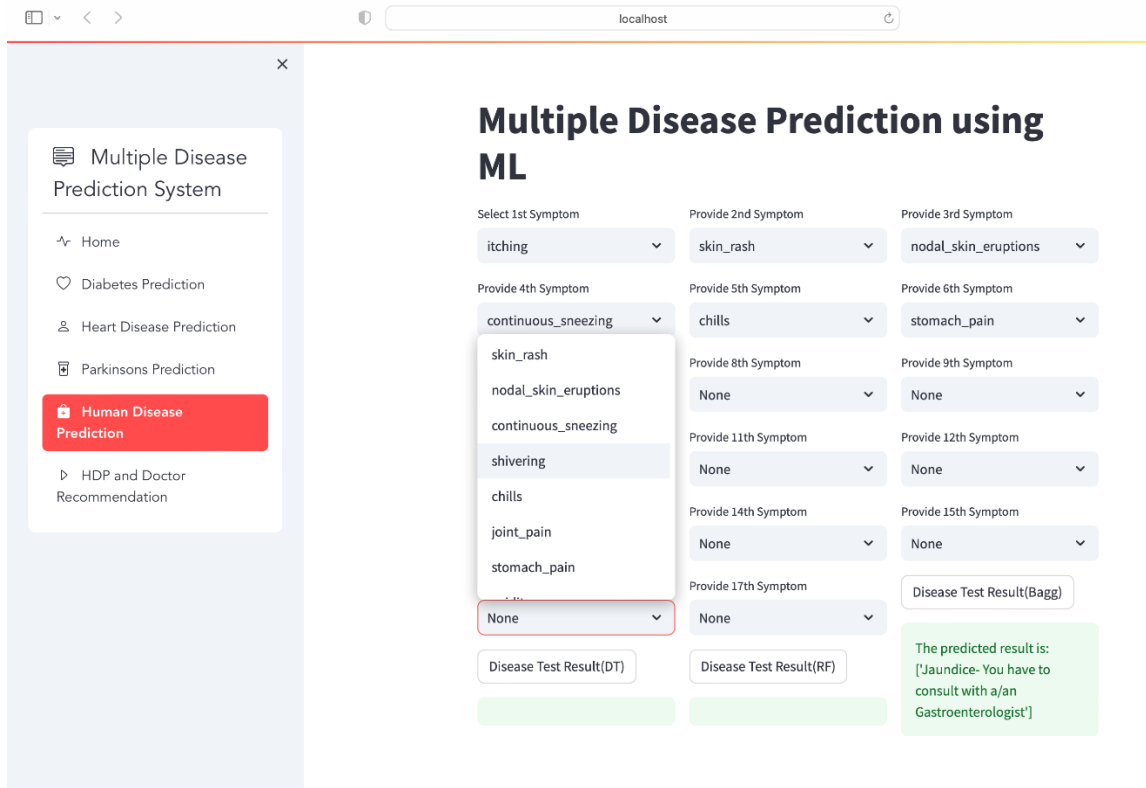


Figure 5.3.2.2: Disease prediction page of disease prediction and doctor recommendation system

### 5.3.2.3 Disease Prediction and Doctor Recommendation page:

The Disease Prediction and Doctor Recommendation page is a user-friendly platform designed to offer an interactive and accurate health assessment. Users can input their symptoms, and the system dynamically suggests relevant symptoms based on a pre-existing database, enhancing the user experience. Utilizing three pre-trained machine learning models ensures comprehensive disease prediction, instilling trust in the system's accuracy. The collaborative use of multiple models enhances the robustness of predictions, providing users with a reliable health self-assessment tool. It is also developed using Python, Streamlit framework, and CSS, allowing users to receive advice from doctors, access disease descriptions, and learn about precautions. In emergency cases, the system suggests Bangladeshi doctors' names with contact information.

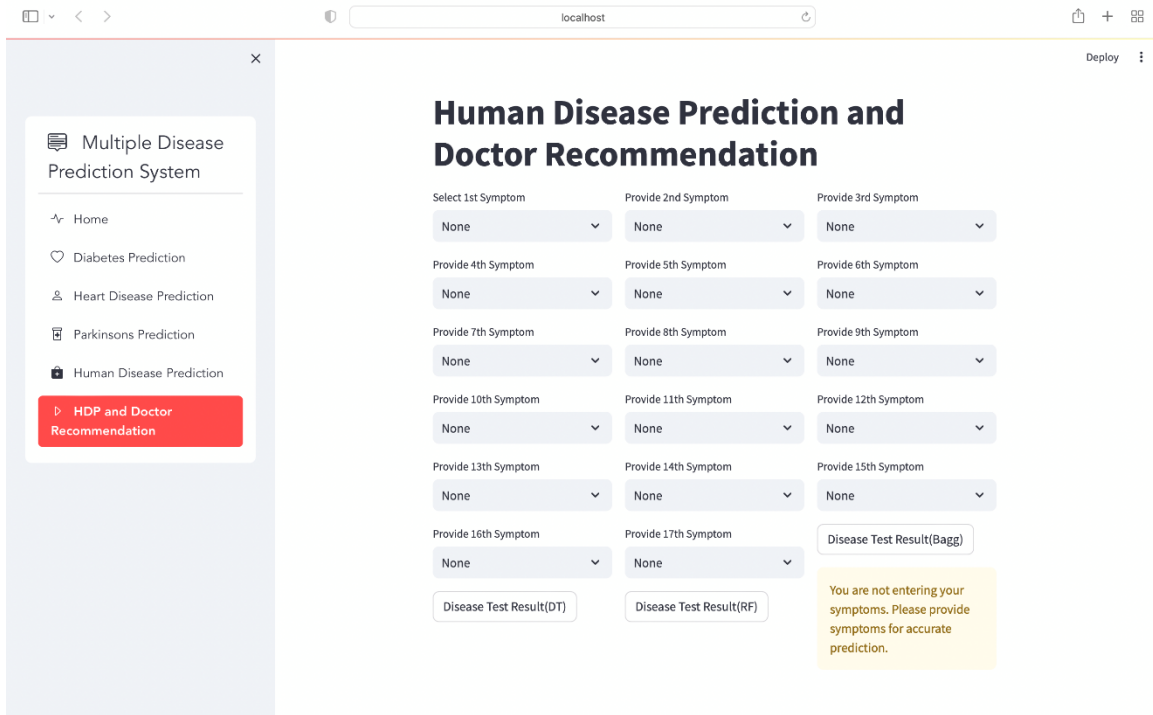
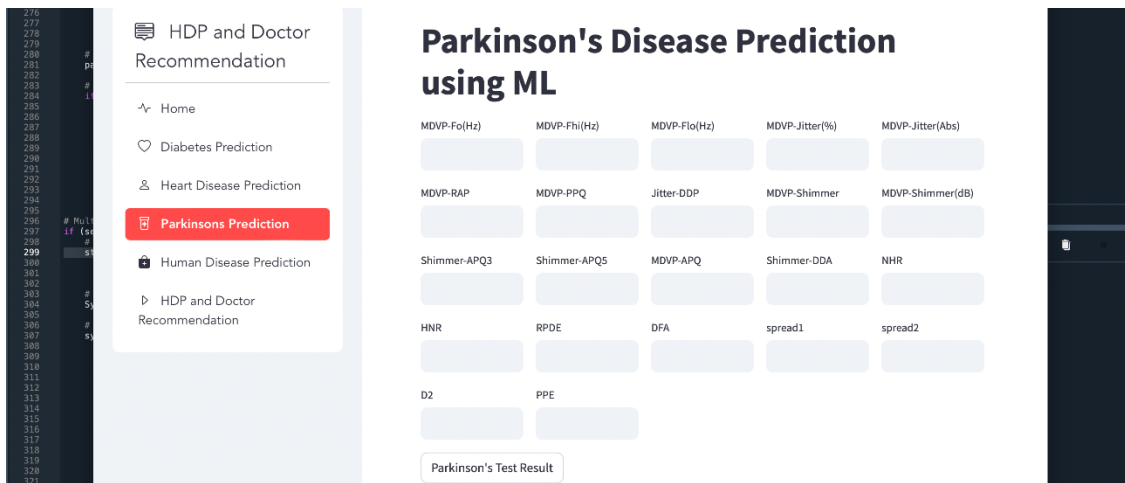


Figure 5.3.2.3: Disease prediction, doctor recommendation and others suggestion's page of disease prediction and doctor recommendation system

### 5.3.2.4: Other page (Heart Disease, Diabetes, Perkinsosis):

These innovative interfaces leverage machine learning to help understand risk for three specific conditions: heart disease, diabetes, and parkinsonism. Simply interact with the user-friendly interface, and the system, powered by advanced algorithms, analyzes your inputs and delivers insightful predictions.



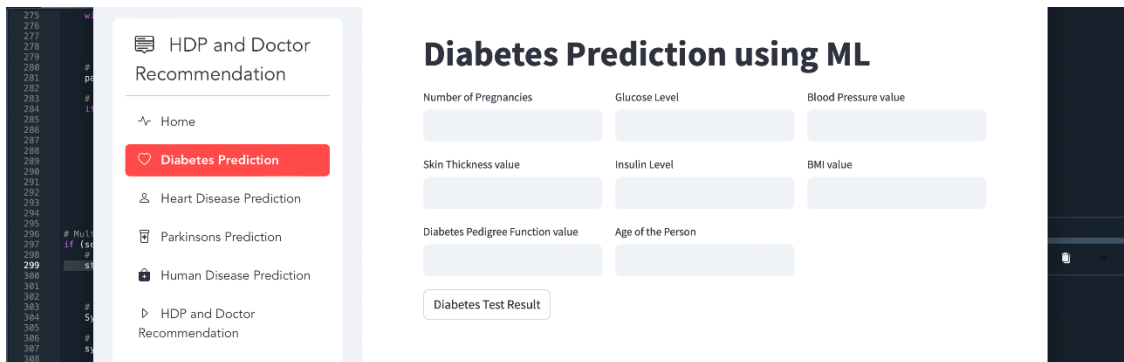
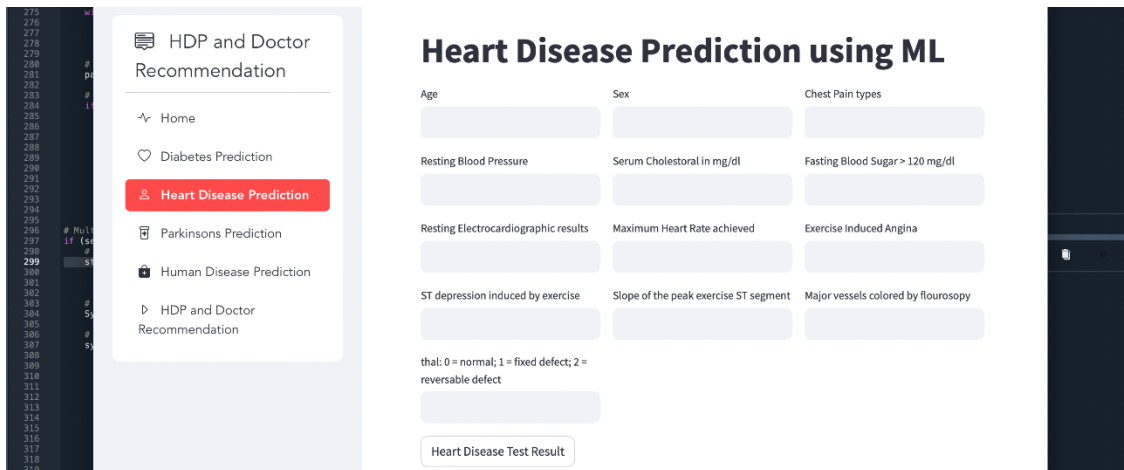


Figure 5.3.2.3: Some extra disease prediction pages (Heart Disease, Diabetes, Perkinsosis)

## 5.4 Performance Evaluation and Testing:

Achieving optimal web application performance necessitates a multifaceted approach to performance testing and evaluation. By conducting load, stress, and UI load tests, among others, one can gauge the application's resilience under varying conditions. Setting clear testing goals, identifying key performance indicators, and utilizing appropriate testing tools are crucial steps in this process. The testing environment must closely mirror the production environment, and meticulous configuration is vital for accurate results. Analyzing performance metrics, such as page speed and response time, provides insights for improvement. The iterative cycle of optimization, reiteration, scaling tests, and

consistent monitoring ensures that the web application remains robust, and adaptive to evolving conditions, and consistently delivers a positive user experience. Different test case and test stats is given below:

#### 5.4.1 Disease Prediction:

Table 5.3.1: Test case for testing disease prediction system

Test Case ID	Test Data	Expected Result	Actual Result	Status
TC_DP_1	Itching, skin rash, nodal skin eruptions, dischromic patches	Fungal infection	Fungal infection	Pass
TC_DP_2	Itching, skin_rash, stomach_pain, spotting_urination	Drug Reaction	Drug Reaction	Pass
TC_DP_3	Fatigue, weight_loss, restlessness, lethargy, irregular_sugar_level, blurred_and_distorted_vision, obesity, excessive_hunger, increased_appetite, polyuria	Diabetes	Diabetes	Pass
TC_DP_4	None	You are not entering your symptoms.	You are not entering your symptoms.	Pass



localhost

## Disease Prediction using ML

Select 1st Symptom: itching

Provide 2nd Symptom: skin\_rash

Provide 3rd Symptom: noda\_skin\_eruptions

Provide 4th Symptom: dischromic\_patches

Provide 5th Symptom: None

Provide 6th Symptom: None

Provide 7th Symptom: None

Provide 8th Symptom: None

Provide 9th Symptom: None

Provide 10th Symptom: None

Provide 11th Symptom: None

Provide 12th Symptom: None

Provide 13th Symptom: None

Provide 14th Symptom: None

Provide 15th Symptom: None

Provide 16th Symptom: None

Provide 17th Symptom: None

Disease Test Result(Bagg)

Disease Test Result(DT)

Disease Test Result(RF)

The predicted result is:  
['Fungal infection- You have to consult with a/an Dermatologist']

localhost

## Disease Prediction using ML

Select 1st Symptom: itching

Provide 2nd Symptom: skin\_rash

Provide 3rd Symptom: stomach\_pain

Provide 4th Symptom: spotting\_urination

Provide 5th Symptom: None

Provide 6th Symptom: None

Provide 7th Symptom: None

Provide 8th Symptom: None

Provide 9th Symptom: None

Provide 10th Symptom: None

Provide 11th Symptom: None

Provide 12th Symptom: None

Provide 13th Symptom: None

Provide 14th Symptom: None

Provide 15th Symptom: None

Provide 16th Symptom: None

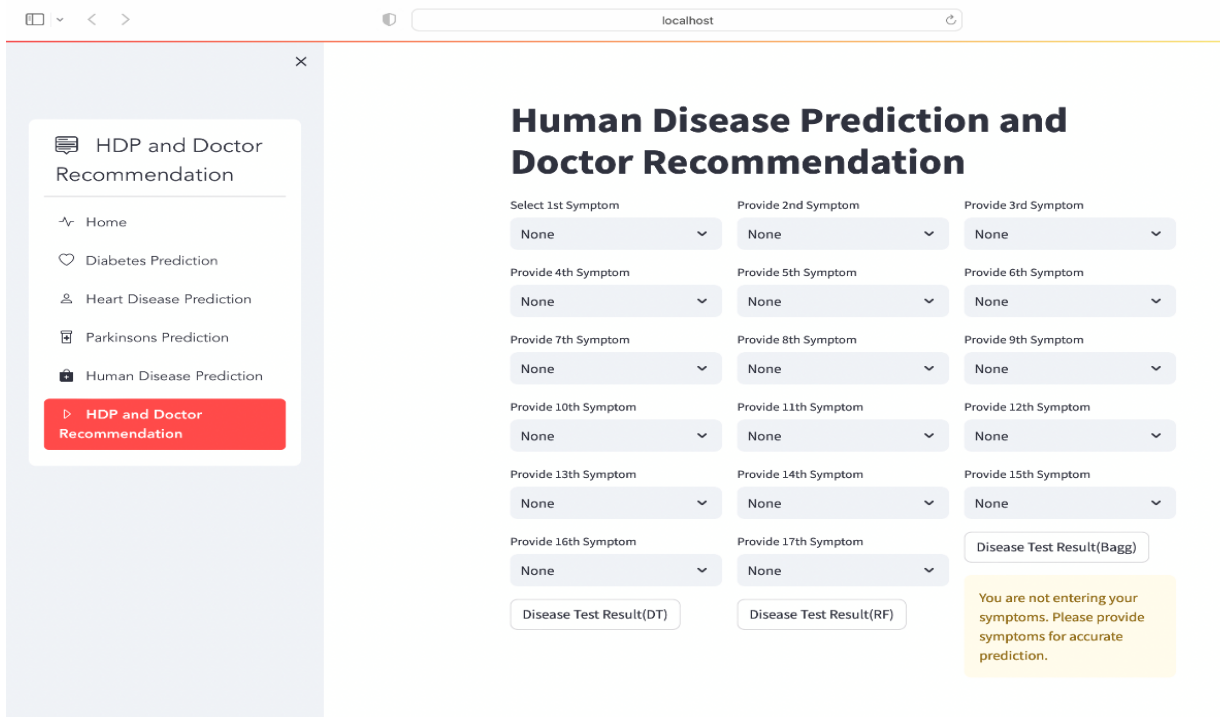
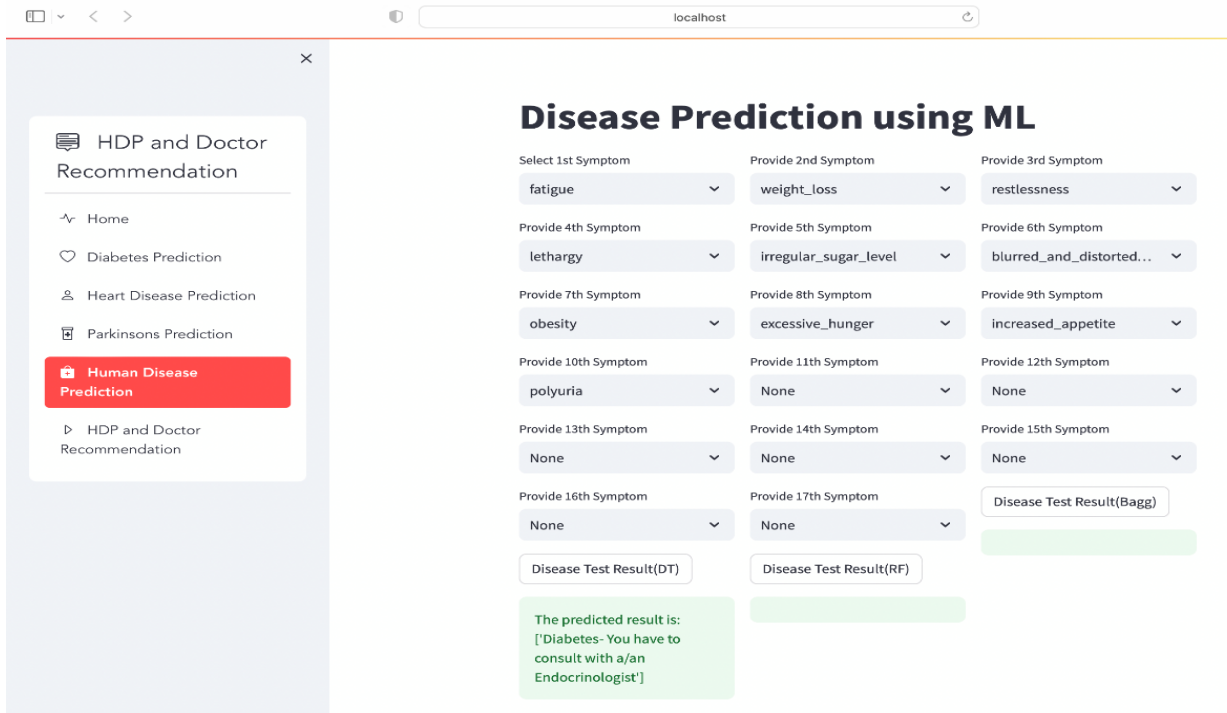
Provide 17th Symptom: None

Disease Test Result(Bagg)

Disease Test Result(DT)

Disease Test Result(RF)

The predicted result is:  
['Drug Reaction- You have to consult with a/an Allergist']



Figures 5.4.1: Figure for showing the system can accurately predict according to test case.

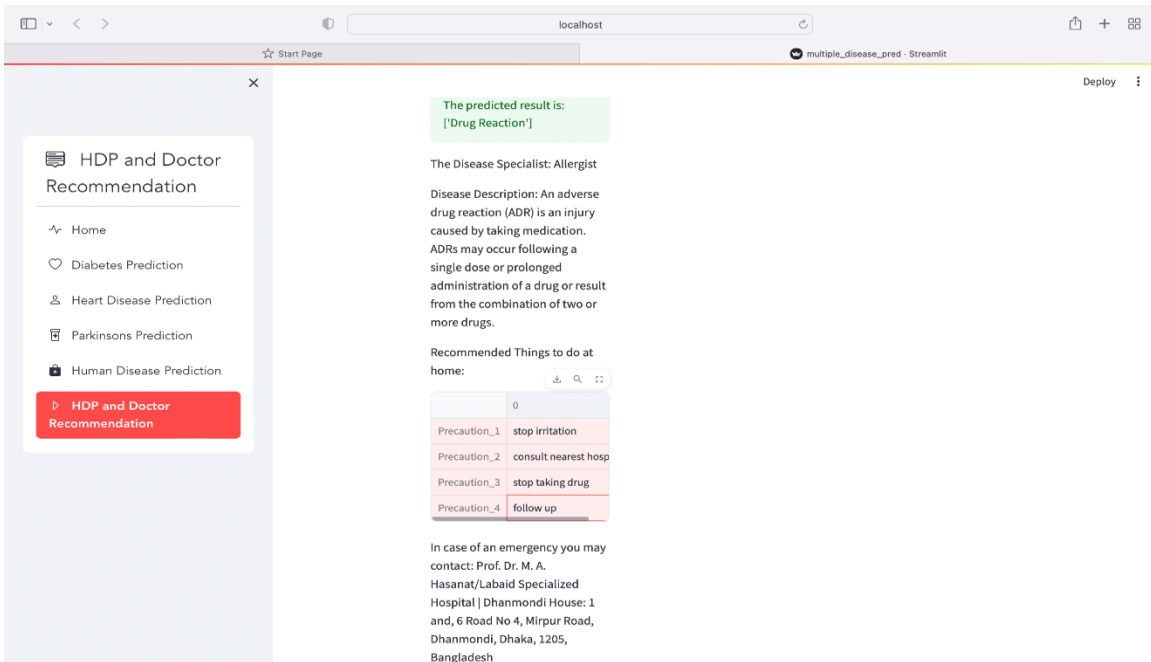
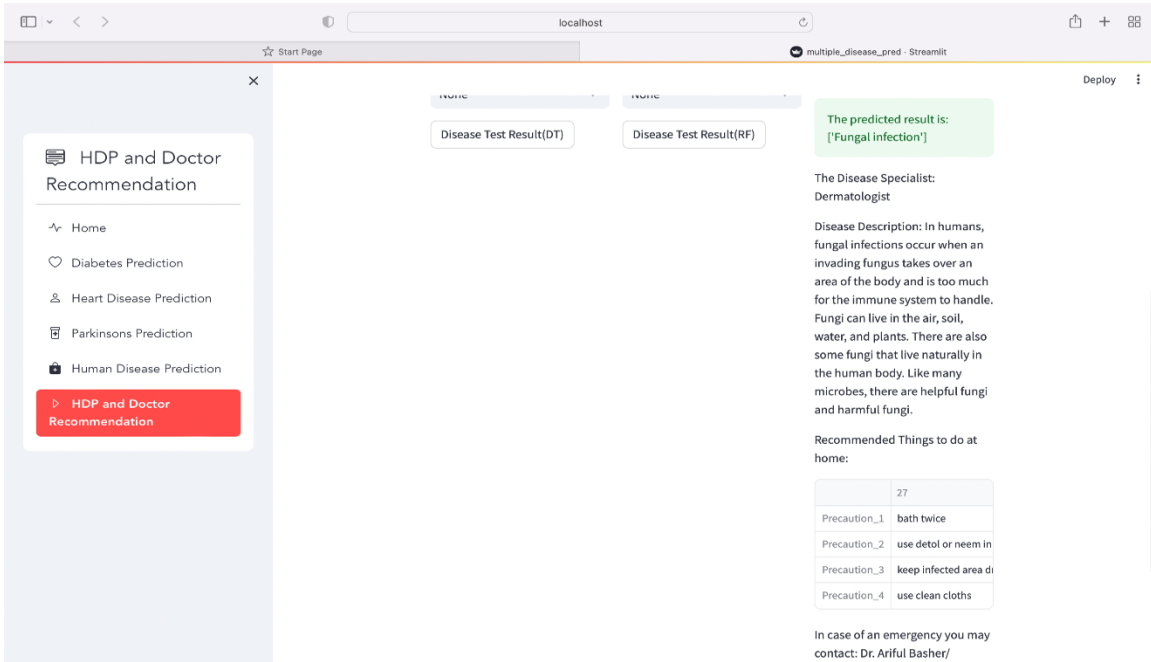
## 5.4.2 Disease Prediction and Doctor Recommendation:

Table 5.4.2: Test case for testing disease prediction and doctor recommendation system.

<b>Test Case ID:</b> TC_DPDR_1
<b>Test Data:</b>
Itching, skin rash, nodal skin eruptions, dischromic patches
<b>Expected Result:</b>
The predicted result is: ['Fungal infection']
The Disease Specialist: Dermatologist
Disease Description: In humans, fungal infections occur when an invading fungus takes over an area of the body and is too much for the immune system to handle. Fungi can live in the air, soil, water, and plants. There are also some fungi that live naturally in the human body. Like many microbes, there are helpful fungi and harmful fungi.
Recommended Things to do at home:
Precaution_1 bath twice
Precaution_2 use detol or neem in bathing water
Precaution_3 keep infected area dry
Precaution_4 use clean cloths
In case of an emergency you may contact: Dr. Ariful Basher/ Locations Bangladesh Specialized Hospital 21 Shyamoli, Mirpur Road, Dhaka, 1207, Bangladesh
<b>Actual Result:</b> Same as expected Result.
<b>Status:</b> Pass

<b>Test Case ID:</b> TC_DPDR_2
<b>Test Data:</b>
Itching, skin_rash, stomach_pain, spotting_ urination.
<b>Expected Result</b>
The predicted result is:
['Drug Reaction']
The Disease Specialist:
Allergist
Disease Description: An adverse drug reaction (ADR) is an injury caused by taking medication. ADRs may occur following a single dose or prolonged administration of a drug or result from the combination of two or more drugs.
Recommended Things to do at home:
Precaution_1 stop irritation
Precaution_2 consult nearest hospital
Precaution_3 stop taking drug
Precaution_4 follow up
In case of an emergency you may contact:
Prof. Dr. M. A. Hasanat/Labaid Specialized Hospital   Dhanmondi House: 1 and, 6 Road No 4, Mirpur Road, Dhanmondi, Dhaka, 1205, Bangladesh
<b>Actual Result:</b> Same as expected result.
<b>Test Status: Pass.</b>

TC_DPDR_3	None	You are not entering your symptoms.	You are not entering your symptoms.	Pass
-----------	------	-------------------------------------	-------------------------------------	------



Figures 5.4.2: Figure for showing the system can accurately predict according to test case.

## **5.5 Conclusion:**

In this chapter on the Disease Prediction and Doctor Recommendation System, we've navigated through the intricate development and testing phases of a robust web application designed to revolutionize health guidance. The system's architecture and components were meticulously explored, shedding light on the crucial role played by Python frameworks and machine learning models in achieving a seamless user experience. The user interface, characterized by an intuitive design and interactive features, reflects the commitment to providing accessible and user-friendly health management tools. Our emphasis on performance evaluation and testing underscores the system's resilience and adaptability under varying conditions, ensuring an optimized experience for end-users. As we close this chapter, the innovative strides taken in merging technology with healthcare present a promising future, where individuals can make informed decisions about their well-being with the aid of advanced and reliable systems.

## CHAPTER 6

### EXPERIMENTAL RESULT AND DISCUSSION

#### 6.1 Introduction:

This chapter discussed the experimental setup and results of using machine learning algorithms for predicting diseases and recommending doctors based on patient symptoms. Utilizing a labeled dataset with 41 diseases and 132 symptoms, machine learning models were implemented and rigorously evaluated. In this case, three distinct data preprocessing approaches were employed. After evaluating we find out four highly effective classification algorithms for disease prediction and doctor recommendation. This chapter also discloses the technological environment for evaluating experimental results. The result and performance of the model are evaluated by accuracy, confusion matrix, cross-validation rate, etc.

#### 6.2 Experimental Setup:

The experimental setup for this project involves the comprehensive utilization of machine-learning algorithms for disease prediction and doctor recommendations based on patient symptoms. For this experiment the raw material is data. As we use a supervised machine learning algorithm we used here a labeled data set that contains 4920 rows of 41 different diseases and 132 unique symptoms. Data are collected from diverse sources including hospitals and the internet. In this research for cleaning, formatting, and preprocessing raw data Python is used, with libraries like pandas, NumPy, sci-kit-learn, seaborn, matplotlib, etc. that facilitate efficient exploration and model development. Various machine learning models, including Random Forests, Decision Trees, and Logistic Regression, Support Vector Machines, Naïve Bayes are employed individually and in hybrid configurations. Hyper parameter tuning and rigorous evaluation metrics ensure optimal model performance. The technological environment includes Google Colab, Anaconda Navigator, Jupiter Notebook, spyder, and Streamlit (A framework that is used for making interactive web apps using Python) framework. During all processes python version 3.11 is used. For version control, I use Git and for collaborative work, Git Hub is used.

### **6.3 Experimental Result and Analysis:**

Within the scope of this investigation, we harnessed a dataset encompassing the medical records of 4920 patients, encompassing information on 131 distinct symptoms and spanning 41 diverse diseases. The focal objective was disease prediction, for which we harnessed the predictive capabilities of 10 distinct machine-learning models in conjunction with a hybrid model. The evaluation of model performance was undertaken through the application of three distinct approaches. For instance, I addressed missing data, standardized and normalized features and encoded categorical variables. We calculated the performance of the machine learning model using three different approaches and preprocessed our trained data in three. For example-

1. In the first case, we encode the data set using an external severity file of symptoms that contain symptoms and their corresponding weight in numeric.
2. In the second case we encode the symptoms using its priority. We find out a priority by its occurrence in the data set.
3. In the third approach we cut off all kinds of redundant data from the data set after preprocessing the data set containing only 12 symptom columns with 128 unique symptoms and 302 rows.



For the first approach, four classification algorithms— Bootstrap aggregating, Decision Tree, Random Forest, and KNeighborsClassifier—were employed, revealing overall excellent performance. Then utilizing these four models we built a hybrid model that also predicts disease with high accuracy. These four models emerged as the most effective, achieving the same accuracy of 99% during training. Subsequently, the model was tested on 41 new patient records, where Random Forest and Bootstrap aggregating exhibited the highest accuracy score of 99%.

Table 6.3.1: Accuracy of different ml models when data is preprocessed with approach 1.

<b>Algorithm Name</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Accuracy</b>
LR	0.91	0.90	0.90	0.90
KN	0.99	0.99	0.99	0.99
DT	1	0.99	0.99	0.99
RF	1	0.99	0.99	0.99
SVM	0.95	0.94	0.94	0.94
Bagg	1	0.99	0.99	0.99
SGD	0.80	0.80	0.78	0.80
XGBoost	1	0.99	0.99	0.99
Hybrid	1	1	1	1
BNB	0.11	0.20	0.12	0.18

If we follow the second approach, we get higher accuracy than all other approaches and it exceeds all existing paper accuracy. And it provides 100% accuracy in some cases. Like the first approach, four machine learning algorithms reach high accuracy including the hybrid model. One of the major causes behind this the data set contains some rows that have the same values in a different order to provide higher accuracy in real-life implementation because the user may not follow any order to provide their symptoms in the system, they can provide their symptoms in any order.

Table 6.3.2: Accuracy of different ml models when data is preprocessed with approach 2.

<b>Algorithm Name</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Accuracy</b>
LR	0.86	0.84	0.84	0.84
KN	1.00	1.00	1.00	1.00
DT	1.00	1.00	1.00	1.00
RF	1.00	1.00	1.00	1.00
SVM	0.95	0.94	0.94	0.94
Bagg	1.00	1.00	1.00	1.00
SGD	0.62	0.66	0.61	0.66
XGBoost	1	0.99	0.99	0.99
Hybrid	1.00	1.00	1.00	1.00
BNB	0.11	0.20	0.12	0.18

The third approach provides lower accuracy than the other approach. But we can divide this approach into two subcategories. First, we drop only columns that contain much null value it does not hamper the accuracy of the model and reaches the highest accuracy. In the second case, we utilize only 302 rows and 12 columns from the data set. It provides a lower but satisfactory accuracy of about 82%.

Table 6.3.3: Accuracy of different ml models when data is preprocessed with approach 3.

<b>Algorithm Name</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Accuracy</b>
LR	0.54	0.60	0.48	0.57
KN	0.49	0.52	0.49	0.62
DT	0.64	0.68	0.68	0.75
RF	0.75	0.74	0.72	0.80
SVM	0.39	0.44	0.40	0.52
Bagg	0.79	0.77	0.75	0.82
SGD	0.37	0.36	0.31	0.48
XGBoost	0.76	0.76	0.72	0.80
Hybrid	0.75	0.73	0.70	0.80
BNB	0.15	0.20	0.12	0.20

Confusion matrix of different machine learning model for approach-2 is given below:













## 6.4 Discussion:

For disease prediction, we can notice that the Bootstrap aggregating (bagging) classifier and Random forest algorithm show the best accuracy in every case. For the bagging classifiers, it shows 99%, 100%, and 82% accuracy respectively. Where the random forest algorithm shows 99%, 100%, and 80% accuracy respectively. Also, decision trees, support vector machines, and hybrid model shows satisfactory results. The hybrid model builds a more robust and versatile model because it is constructed with a combination of different machine-learning models. In our research, we built it only by taking the four best accurate models that take the strengths of other models and overcome individual weaknesses.

The bagging classifier shows the best accuracy in disease prediction due to its ability to reduce variance, boost diversity in models, handle data noise, and tame high-variance algorithms. Random forest algorithms and Decision trees also provide satisfactory results because Decision trees offer interpretability, handle non-linearity, and resist noise, while random forests leverage diverse trees to reduce variance and overfitting, making them a powerful yet interpretable duo for tackling complex tasks. For human disease prediction and doctor recommendation systems accuracy is the major issue because lowly accurate models sometimes lead to wrong prediction.

On the other hand, in our system, a significant accuracy-related concern arises when a person presents symptoms that are associated with multiple diseases. This scenario can lead to confusion within the system, potentially resulting in inaccurate predictions.

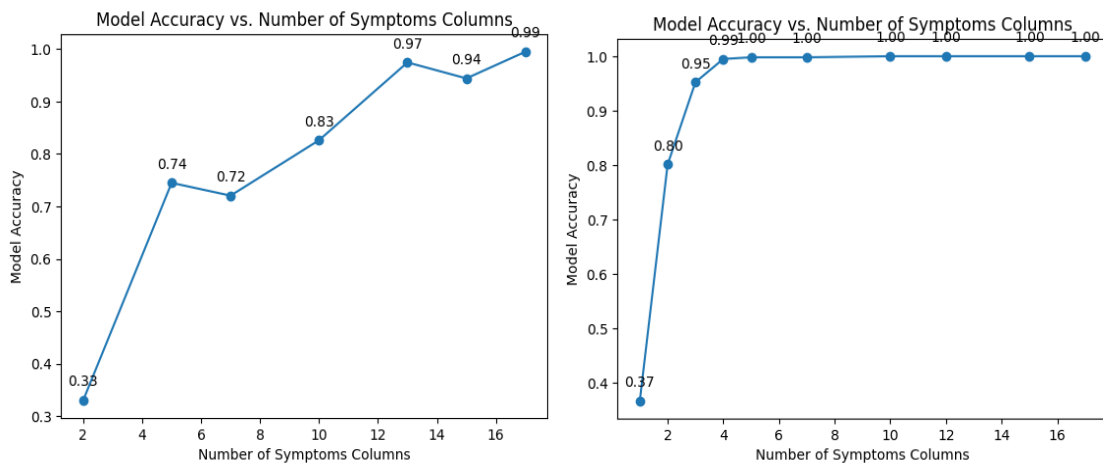


Figure 6.4: Model Accuracy before and after encoding symptoms by its priority number

However, adopting Approach Two (Priority) for data preprocessing can address this issue effectively. By prioritizing specific symptoms, we can achieve higher accuracy by requiring users to input only a subset of key symptoms—around 4 or 5. This streamlined approach is anticipated to provide nearly 99% accuracy, enhancing the reliability of our system.

## **6.5 Conclusion:**

The experimental results indicate the successful application of machine learning algorithms for disease prediction and doctor recommendations based on patient symptoms. Bootstrap aggregating and Random Forest algorithms consistently exhibited the best accuracy in disease prediction, while a hybrid model combining multiple algorithms demonstrated robust performance. From this chapter, it is clear that if we encode our object-type symptom data by setting the priority of the symptoms it provides the best accuracy and overcomes a problem that arises when multiple diseases have symptoms. Overall, the research highlights the potential of machine learning models, particularly when coupled with effective data preprocessing strategies, as valuable tools in the realm of disease prediction and healthcare recommendations. This chapter also addresses a challenge like overlapping symptoms and finds its solutions.

## CHAPTER 7

### IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

#### 7.1 Introduction:

Chapter 7 delves into the multifaceted impact of the disease prediction and doctor recommendation system on society, the environment, and overall sustainability, marking a pivotal moment in the 4th Industrial Revolution with the integration of Artificial Intelligence (AI) and Machine Learning in healthcare. This chapter explores how the system contributes to societal well-being by improving healthcare outcomes, reducing costs, increasing accessibility, and empowering underserved communities. It also delves into the environmental implications, emphasizing its eco-friendly nature through minimized physical interactions, reduced vehicle use, and sustainable practices in pharmaceuticals. Moreover, ethical considerations in smart healthcare are discussed, underscoring the importance of privacy, equity, and transparency. The chapter illuminates the potential benefits this technology holds for enhancing healthcare access, reducing costs, and promoting overall well-being while addressing societal, environmental, and ethical dimensions.

#### 7.2 Impact on Society:

The 4th Industrial Revolution is witnessing a remarkable transformation in healthcare, driven by the power of Artificial Intelligence (AI) and its subfield, Machine Learning. It is breaking the healthcare industry by assisting doctors in diagnosing, finding where the diseases come from, guiding surgery, and predicting if the problem is serious. Also, it helps to identify unknown diseases, early disease predictions, doctor recommendations, give suggestions on what to do at home as first aid, and provide more details about the specific disease. So, the impact of an early disease prediction and doctor recommendation system using machine learning on society can be significant. By accurately predicting diseases based on patient symptoms, individuals can receive timely intervention and treatment, leading to better health outcomes and potentially lower healthcare costs. This can reduce the burden of disease on individuals and society as a whole.

- Improved healthcare outcomes: The project aims to improve healthcare access in underserved communities, which can lead to better patient outcomes and a lower burden of disease on individuals and society as a whole. This is a significant societal impact as it addresses the inequities in healthcare and works towards providing equal access to quality care.
- Reduced healthcare costs: Early diagnosis and preventative measures can potentially reduce the need for expensive treatments and procedures. Our system predicts diseases with high accuracy, reliability, and efficiency based on patient symptoms and it provides an interactive user-friendly open-source UI that leads to providing quality health care at low cost.
- Increased accessibility: The system can address the challenge of limited access to quality healthcare in remote areas. By providing diagnoses without physical contact, the system can improve healthcare access for underserved communities. This can remove geographical barriers and provide valuable assistance to those in remote areas who may have limited access to healthcare professionals.
- Reduced uncertainty: Our system provides reliable disease predictions and doctor recommendations with preventive measures that can lead to eradicating anxiety and confusion for patients, enabling them to make informed decisions about their health. I also provide patient empowerment and overall satisfaction.
- Integration with existing healthcare infrastructure: The integration of the project's system with existing healthcare infrastructure is crucial to its success. The project seeks to explore how the system can be integrated with the current healthcare infrastructure to ensure seamless adoption and use.
- Addressing limited access to quality healthcare in remote areas: The development of disease prediction models that can identify diseases without physical contact, such as during the COVID-19 pandemic, is an important contribution to the project. This enables the system to provide diagnoses remotely, further expanding its reach and impact on healthcare access. The system's ability to provide diagnoses without physical contact makes it particularly beneficial for underserved communities with limited access to healthcare professionals.

- **Impact on underserved communities and patient empowerment:** This project focuses on improving healthcare access in underserved communities, reducing health disparities, and improving health equity. It also empowers patients to take control of their health and make informed decisions. Furthermore, the project connects individuals with the right healthcare providers and suggests preventative measures, the system can contribute to better health management and overall well-being.
- **Potential for future research and development:** Through multidisciplinary collaboration and community engagement, the project aims to bring lasting change to the healthcare landscape in developing nations. It acknowledges the need for future research to improve the model, expand its capabilities, and address additional challenges in healthcare access and diagnosis.

### **7.3 Impact on Environment:**

The project makes a substantial and positive impact on the environment, especially in pandemic situations. By utilizing machine learning for accurate and timely disease diagnoses, the need for physical contact with a doctor is minimized. As a result, it contributes significantly to the containment of infectious diseases. As it predicts disease and also provides preventive measures so in simple cases does not need to visit the doctor which leads to limited use of vehicles and reduced production of carbon dioxide. So, the project's emphasis on minimizing vehicle use aligns seamlessly with environmental sustainability goals, establishing it as an eco-friendly alternative to conventional healthcare practices.

On the other hand, by enabling early diagnosis and intervention, the project actively mitigates the spread of diseases, consequently decreasing the overall demand for healthcare resources. This reduction in demand plays a crucial role in minimizing the environmental impact associated with the production and disposal of medical equipment. The reliance on machine learning algorithms for disease prediction not only enhances healthcare outcomes but also contributes to a more environmentally conscious and sustainable approach to healthcare delivery.

Additionally, the project's focus on early disease prediction results in reduced medication needs, offering several environmental benefits. The diminished demand for medications translates into lower requirements for raw material extraction, leading to the conservation of natural resources and reduced environmental impact in pharmaceutical production.

Low medication needs also reduce packaging waste, such as blister packs and pill bottles. It positively impacts the environment and mitigates plastic pollution. Additionally, it contributes to energy conservation in pharmaceutical manufacturing, reducing energy consumption in synthesis and packaging processes and resulting in a smaller carbon footprint. The project's positive effects extend to the disposal phase, potentially decreasing environmental pollution associated with unused or expired medications

#### **7.4 Ethical Aspects:**

Smart health (AI) care raises ethical concerns related to privacy, data security, and the potential for discrimination against certain groups, such as the elderly or disabled individuals. Additionally, the use of smart health care (AI) technology may lead to ambiguities in laws and regulations, making it difficult to define responsibility in case of wrong treatment or diagnosis. Furthermore, the uneven distribution of medical resources through smart health care platforms may exacerbate existing social inequalities. [33] Smart healthcare lowers costs, improves the quality of care, and benefits all parties involved in the healthcare system. Hospital collaboration is made possible via integrated medical information technology, which enables online appointments and resource sharing. But ethical issues, including discrimination, losing one's job, and disclosing information, must be addressed. Smart healthcare needs to supplement conventional services rather than take their place. To foster public trust, governments ought to set up legislative frameworks that safeguard users and developers who prefer to have government-managed databases. To guarantee that healthcare services continue to evolve positively, a balanced strategy is required. My machine learning-based disease prediction and doctor recommendation system was developed with certain ethical issues in mind. The protection of patient data privacy and security is the most important of these worries, requiring strong safeguards against unauthorized access and breaches. One important factor is algorithmic bias, which necessitates ongoing attempts to reduce differences in forecasts across demographic

groups. Understanding and correcting possible biases in machine learning models requires transparency. The system has to actively support equity and equitable access to healthcare services while taking socioeconomic and geographic differences into account. A patient's autonomy must be respected, and informed permission must be obtained to make sure people understand how their health data is used. Working together with healthcare professionals is crucial to achieving a system that enhances rather than replaces medical judgment by striking a balance between technology and human skill. The system should also be made accessible and reasonably priced, taking into account any differences in people's access to technology. The ethical framework is completed by ongoing oversight, responsibility, and compliance with legal requirements, which guarantee that the system puts patients' needs first and complies with fairness and transparency standards.

### **7.5 Sustainability Plan:**

The disease prediction and doctor recommendation system's sustainability strategy is carefully designed to guarantee its long-term effects. The technological focus is on ongoing development, with frequent updates to machine learning models and strict security protocols. The management of financial sustainability includes cost optimization, collaborations, subscription model exploration, and diversification. Prioritizing user involvement and accessibility on a social level addresses a range of requirements and ethical considerations. In terms of the environment, the strategy emphasizes responsible data handling and energy efficiency. Widespread adoption requires public awareness campaigns, continuous teaching and training, and integration with the current healthcare system. Strong data governance guarantees compliance, while international cooperation expands the system's functionality. The system's long-term effectiveness is attributed to its adaptability to new technology, frequent impact evaluations, and incentive models for healthcare practitioners. The resilience of the system is further strengthened via scalable infrastructure and public-private partnerships. The sustainability plan seeks to ensure the system's prosperity through these factors, having a long-lasting beneficial effect on healthcare across the world.

## **7.6 Conclusion:**

In summary, the disease prediction and doctor recommendation system not only represents a groundbreaking initiative [34]with widespread positive effects on society and the environment but also showcases a commitment to sustainable healthcare delivery. This includes improved healthcare outcomes, cost reduction, and increased accessibility, particularly benefiting underserved communities. The system's eco-friendly practices and reduced environmental footprint underscore its dedication to sustainability. Ethical considerations, such as patient privacy, equity, and transparency, are acknowledged and addressed. The comprehensive sustainability plan outlines strategies for long-term success, incorporating technological advancements, financial optimization, and social awareness. Emphasizing continuous development, user engagement, responsible data handling, and integration with existing healthcare infrastructure, the plan positions the system as a model for future healthcare initiatives prioritizing societal well-being, environmental responsibility, and ethical considerations. The concluding remarks highlight the system's adaptability to advancements, learning from impact evaluations, and [35]creating lasting positive change through public-private partnerships, paving the way for a healthier future.



## CHAPTER 8

### CONCLUSION

#### **8.1 Implication for Further Study:**

Further studies can contribute to the advancement of disease prediction models by expanding data sources and incorporating sophisticated deep-learning methods, aiming to improve precision and generalizability. It is essential to enhance user experience and trust through improved user interfaces, incorporating multi-language support, personalization features, and feedback loops.

Future research should prioritize addressing ethical concerns in healthcare technology, focusing on developing robust data privacy policies, ensuring fair access, and navigating ethical dilemmas associated with advanced predictive models. Connecting the system with real-time doctors can significantly elevate user trust by allowing healthcare professionals to monitor suggestions and integrate them into their work.

While the current system is trained with a structured dataset, future studies should explore the integration of real-time data directly from hospital databases to enhance adaptability and responsiveness to emerging health trends. Additionally, introducing new features such as timely medicine notifications for medication adherence can provide a more comprehensive and personalized approach to healthcare management.

To expand the system's capabilities, further exploration of scanning patient medical reports for tailored suggestions and recommendations for necessary medical tests can be undertaken. A real-time disease prediction system should also focus on identifying new diseases based on ongoing patient medical history, [36] ensuring the system remains at the forefront of healthcare technology for improved patient outcomes and overall efficacy.

## **8.2 Conclusion:**

Human disease prediction and doctor recommendation involve utilizing machine learning algorithms to accurately predict diseases based on patient symptoms, and provide recommendations for appropriate doctors and preventative measures. This approach aims to complement traditional medical diagnosis methods with data-driven insights, facilitating early intervention and treatment, reducing uncertainty for patients, and improving healthcare outcomes.

The importance of human disease prediction and doctor recommendation lies in addressing the challenges posed by limited access to quality healthcare, particularly in underserved communities and remote areas. Traditional medical diagnosis methods can be limited and subject to human error, leading to delays in treatment and increased burden of disease. By implementing a system that accurately predicts diseases based on patient symptoms and provides doctor recommendations, early intervention and treatment can be facilitated, reducing uncertainty for patients and improving healthcare outcomes. This approach has the potential to make healthcare more accessible to everyone, particularly in areas with limited access to healthcare professionals.

The system is implemented by utilizing machine learning techniques to accurately predict diseases based on symptoms presented by patients. The system can work with high accuracy, reliability, and efficiency. To ensure the system's effectiveness and accessibility, in our system we implement an interactive user interface. Within this graphical user interface (GUI), users can input their symptoms, and the system employs its pre-trained machine learning model to identify similarity patterns and make predictions about the potential disease. Additionally, based on the disease the system provides recommendations for appropriate doctors and preventative measures, aiming to facilitate early intervention and treatment for improved healthcare outcomes.

## References:

- [1] A. Kumar, G. K. Sharma and U. Prakash, "Disease Prediction and Doctor Recommendation System using Machine Learning Approaches," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 9, no. VII, pp. 34-44, July 2021.
- [2] S. . K. NAYAK, M. GARANAYAK, S. K. SWAIN, S. . K. PANDA and D. GODAVARTHI, "An Intelligent Disease Prediction and Drug Recommendation Prototype by Using Multiple Approaches of Machine Learning Algorithms," *IEEE Access*, vol. VOLUME 11, 2023 Sep 11.
- [3] M. Javaid, A. Haleem, R. P. Singh, R. Suman and S. Rab , "Significance of machine learning in healthcare: Features, pillars and applications.," *International Journal of Intelligent Networks*, vol. Network 3, p. 58–73, 2022.
- [4] A. K. Sahoo, S. Mallik, C. Pradhan, B. S. B. Shankar, P. M. R. K. B. and H. D. , "Intelligence-based health recommendation system using big data analytics," *Academic Press*, vol. CHAPTER 9, pp. 227-246, 2019.
- [5] S. Uddin, A. Khan, M. E. Hossain and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction.," *BMC Medical Informatics and decision making.*, vol. 19(1), pp. 1-16, 2019.
- [6] B. Ihnaini, M. A. Khan, T. A. Khan, S. Abbas, M. Sh., M. Ahmad and . M. A. Khan, "A smart healthcare recommendation system for multidisciplinary diabetes patients with data fusion based on deep ensemble learning.," *Computational Intelligence and Neuroscience*, vol. Volume 2021, no. 6 September 2021, p. 11 pages, 2021.
- [7] S. V, P. A. Kubeear, M. M, J. J and M. , "AN APPROACH FOR PREDICTION OF DISEASES TO SUGGEST DOCTORS AND HOSPITALS TO PATIENT BASED ON RECOMMENDATION SYSTEM.," *Journal of Contemporary Issues in Business and Government*, vol. 27(3), no. Business and Government, pp. 1591-5198, 3,2021.
- [8] Y. Bao and X. Jiang, "An intelligent medicine recommender system framework.," in *IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, China , 2016, June.
- [9] U. A. Bhatti, M. Huang, D. Wu, Y. Zhang, A. Mehmood and H. Han, "Recommendation system using feature extraction and pattern recognition in clinical care systems.," *Enterprise information systems*, vol. 13(3), pp. 1751-7575, 2018.

- [10] J. Chen, K. Li, H. Rong, K. Bilal, N. Yang and K. Li, "A disease diagnosis and treatment recommendation system based on big data mining and cloud computing.," *Information Sciences*, vol. 435, pp. 124-149, 2018.
- [11] S. Mahata, Y. B. Kapadiya, V. Kushwaha, V. Joshi and Y. Farooqui, "Disease Prediction and Treatment Recommendation Using Machine Learning.," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 11, no. III, pp. 1232-1237, March 2023.
- [12] R. Keniya, A. Khakharia, V. Shah, V. Gada, R. Manjalkar, T. Thaker, M. Warang and N. Mehendale, "Disease prediction from various symptoms using machine learning.," *SSRN 3661426*, 2020.
- [13] N. O. M. Salim and A. M. Abdulazeez, "Human diseases detection based on machine learning algorithms: A review," *International Journal of Science and Business*, vol. 5(2), pp. 102-113, 2021.
- [14] A. S. Hussein, W. M. Omar, X. Li and M. , "Efficient chronic disease diagnosis prediction and recommendation system," in *IEEE EMBS International Conference on Biomedical Engineering and Sciences*, Langkawi, 2012, December..
- [15] M. Roy, R. Koshy and R. Roy, "Human Disease Prediction And Doctor Booking System," *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)*, vol. 11, no. 01, 2023.
- [16] J. P. Gupta, A. Singh and R. K. Kumar, "A computer-based disease prediction and medicine recommendation system using machine learning approach.," *Int J Adv Res Eng Technol (IJARET)*, vol. 12, no. 3, pp. 673-683, 2021.
- [17] P. Panapana, K. S. R. Reddy, J. D. G. R. Babu and A. D. , "DISEASE PREDICTION AND MEDICATION," *DISEASE PREDICTION AND MEDICATION ADVICE USING MACHINE LEARNING ALGORITHMS*, vol. Volume 11, no. Issue 3, pp. 788-801, March 2023.
- [18] D. Gujar, R. Biyani, T. Bramhane, S. Bhosale and T. P. Vaidya, "Disease prediction and doctor recommendation system," *International Research Journal of Engineering and Technology (IRJET)*, vol. 5, no. 03, pp. 3207-3209, 2018.
- [19] S. Sultana, A. M. Al , M. Haque, M. . H. Khandaker and N. Sakib, "An Approach for Developing Diabetes Prediction and Recommendation System," *International Journal of Computer Applications(0975 – 8887)*, vol. 174, pp. 20-28, January 2021.
- [20] N. ZHU, J. CAO, K. SHEN, . X. CHEN and S. ZHU, "A decision support system with intelligent recommendations for multi-disciplinary medical treatment.," *ACM Transactions on*

*Multimedia Computing, Communications, and Applications (TOMM)*, Vols. Vol. 16, No. 1s, Article 33., pp. pp.1-23, March 2020..

- [21] D. C. K. Gomathy and M. A. R. Naidu, "The prediction of disease using machine learning.," *International Journal of Scientific Research in Engineering and Management (IJSREM)*, vol. 5, no. 10, pp. 1-7, 2021.
- [22] K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar and T. Suryawanshi, "Human Disease Prediction Using Machine Learning Techniques and Real-life Parameters.," *International Journal of Engineering*, vol. 36, no. 6, pp. 1092-1098, 2023.
- [23] K. Patil, S. Pawar, P. Sandhyan and J. Kundale, "Multiple Disease Prognostication Based On Symptoms Using Machine Learning Techniques.," in *ITM Web of Conferences 44, 03008*, India, 2022.
- [24] "Disease-Symptom Knowledge Database," [Online]. Available: <https://impact.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html>. [Accessed 28 10 2023].
- [25] "Kaggle," Disease Symptom Prediction, [Online]. Available: <https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset?resource=download>. [Accessed 2023 11 20].
- [26] "DOCTOR BANGLADESH," [Online]. Available: <https://www.doctorbangladesh.com/>.
- [27] "University of Hertfordshire," [Online]. Available: <https://www.herts.ac.uk/international/new-international-students/international-apply-now/submit-your-documents>. [Accessed 1 12 2023].
- [28] "Advanced Computing and Intelligent Technologies," 2022. [Online]. Available: <https://link.springer.com/book/10.1007/978-981-19-2980-9>. [Accessed 2 12 2023].
- [29] "Research Square," Research Square Company, [Online]. Available: <https://www.researchsquare.com/>. [Accessed 3 12 2023].
- [30] "University of Strathclyde," University of Strathclyde, [Online]. Available: <https://www.strath.ac.uk/professionalservices/information/services/libraryhelp/in-depthhelpguides/thesissubmission/>. [Accessed 4 12 2023].
- [31] "Docplayer," [Online]. Available: <https://docplayer.net/3698174-Use-of-internet-and-its-effects-on-our-society.html>. [Accessed 5 12 2023].
- [32] "ijraset," [Online]. Available: <https://www.ijraset.com/>. [Accessed 1 12 2023].

- [33] J. P. Gupta, A. Singh and R. K. Kumar, "A COMPUTER-BASED DISEASE PREDICTION AND MEDICINE RECOMMENDATION SYSTEM USING MACHINE LEARNING APPROACH," *International Journal of Advanced Research in Engineering and Technology (IJARET)*, vol. Volume 12, no. Issue 3, pp. 673-683, March 2021.
- [34] X. Zhou, Y. Li and W. Liang, "CNN-RNN Based Intelligent Recommendation for Online Medical Pre-Diagnosis Support," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 3, pp. 912-021, 2020.
- [35] C. Ju and S. Zhang, "Doctor Recommendation Model for Pre-Diagnosis Online in China: Integrating Ontology Characteristics and Disease Text Mining," in *IEEE 6th International Conference on Big Data Analytics*, China, 2021.
- [36] V. Chang, Y. Cao, T. Li, Y. Shi and P. Baudier, "Smart Healthcare and Ethical Issues," in *International Conference on Finance, Economics, Management and IT Business*, China, January 2019.

# APPENDIX A

## Statistical Analysis

The heatmap illustrates the relationship between the 17 symptoms and the corresponding disease. Each row represents a case, and the colors in the heatmap show how strongly symptoms are related. Darker colors indicate stronger associations, helping to identify patterns where certain symptoms tend to occur together. This visual tool is valuable for predicting diseases based on symptoms. By analyzing the heatmap, we can decide which symptoms are often present simultaneously, aiding in the selection of relevant features for a predictive model.

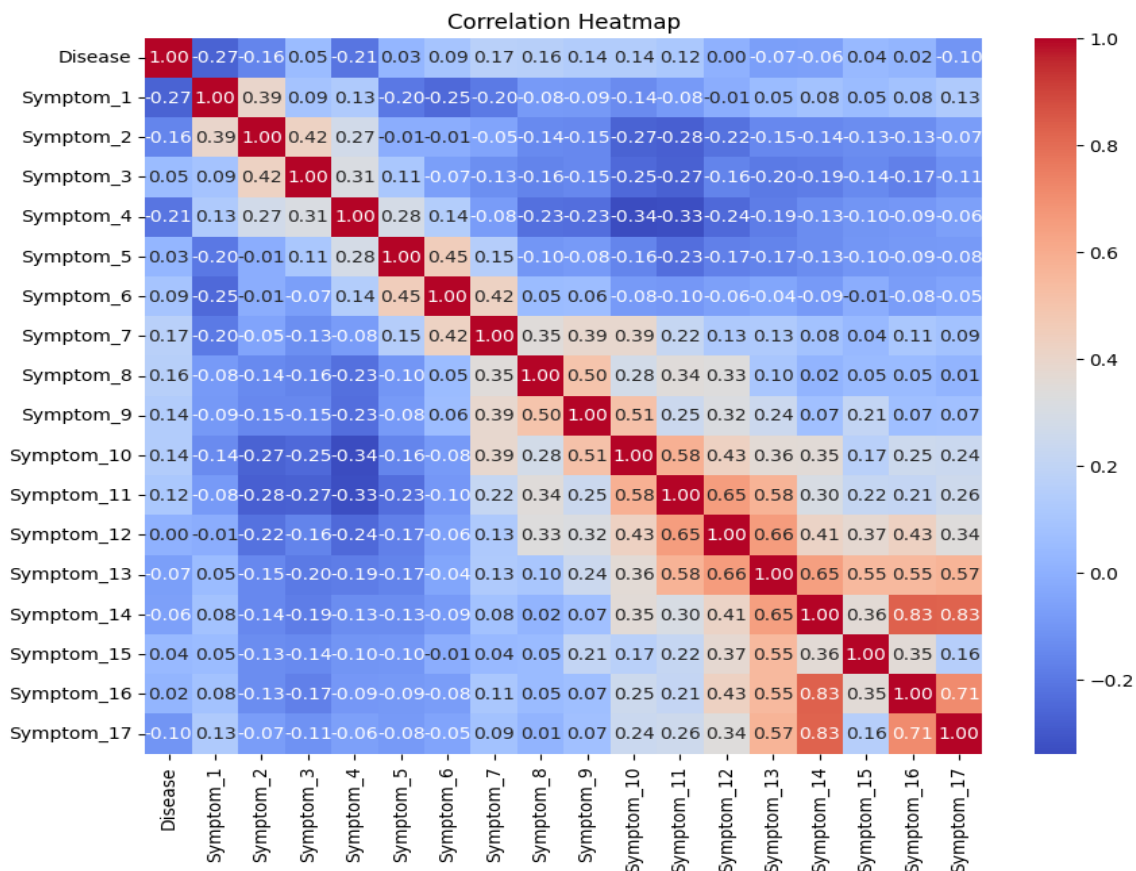


Figure A1: Heat map

The below figure employs Seaborn's pairplot to generate a matrix of scatter plots for pairs of symptoms ('Symptom\_1' to 'Symptom\_5'). The matrix provides a comprehensive view of the relationships between these symptoms, with each scatter plot showcasing the correlation between the two symptoms. The points are colored based on the associated disease, offering insights into how different diseases may exhibit distinct patterns across multiple symptoms. This pairplot is a powerful tool for uncovering multivariate patterns and understanding how combinations of symptoms might be indicative of specific diseases in your dataset.

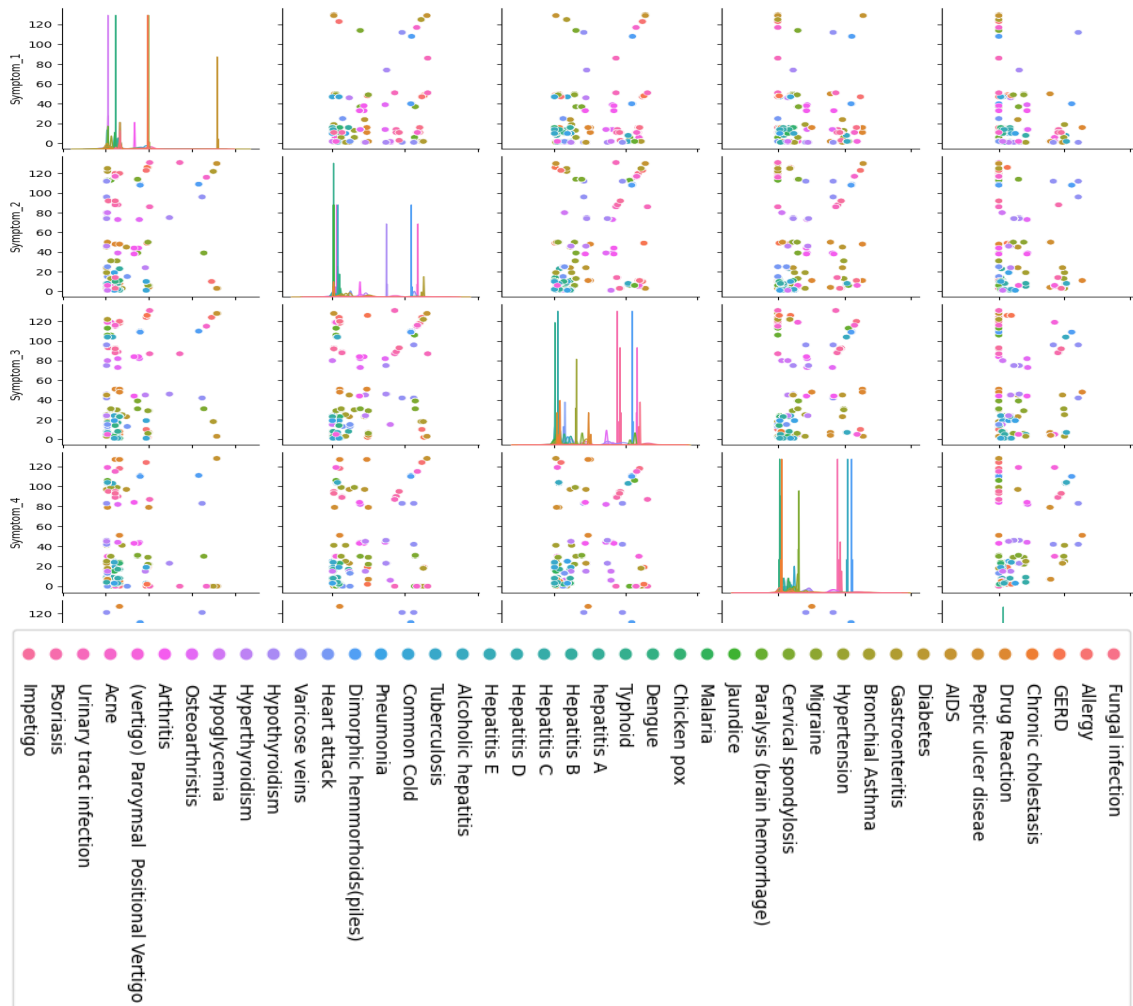


Figure A2: Pair Plot of Different Symptoms

The below figure shows a scatter plot of two specific symptoms, 'Symptom\_1' and 'Symptom\_2,' in the dataset. Each point on the plot represents an observation and the color



of the point corresponds to the associated disease. This visualization allows us to examine the relationship between these two symptoms and observe potential patterns or clusters. The legend, positioned outside the plot area, serves as a key to interpreting the colors and understanding which diseases are represented.

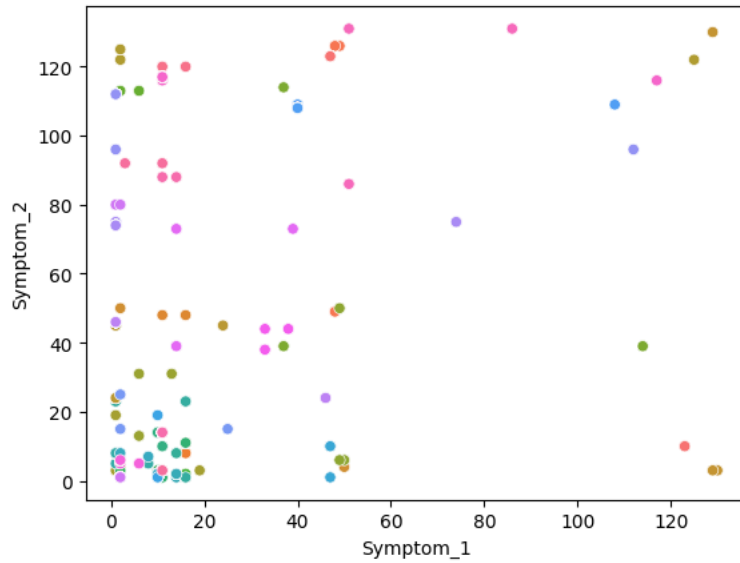


Figure A3: Scatter Plot of Symptom 1 and 2

## Appendix B

### Experimental Results & Analysis:

In our analysis, Random Forest emerged as the superior model for disease prediction, exhibiting higher accuracy compared to Linear Regression. The dataset, comprising 17 symptoms, was utilized to predict a categorical variable encompassing 41 unique disease classes. As it is a multiclass classification task with 41 unique disease classes, we initially binarized the problem to create individual binary classification tasks for each disease. The Random Forest algorithm's superior accuracy suggests its efficacy in handling the complexity of the dataset and capturing nuanced relationships between symptoms and diseases. On the other hand, Linear Regression, being a linear model, demonstrated lower accuracy, potentially indicating that the relationship between symptoms and diseases is better captured by a more complex, non-linear model like Random Forest. These findings underscore the importance of selecting an appropriate algorithm for the specific task at hand, particularly when dealing with a diverse set of disease classes.

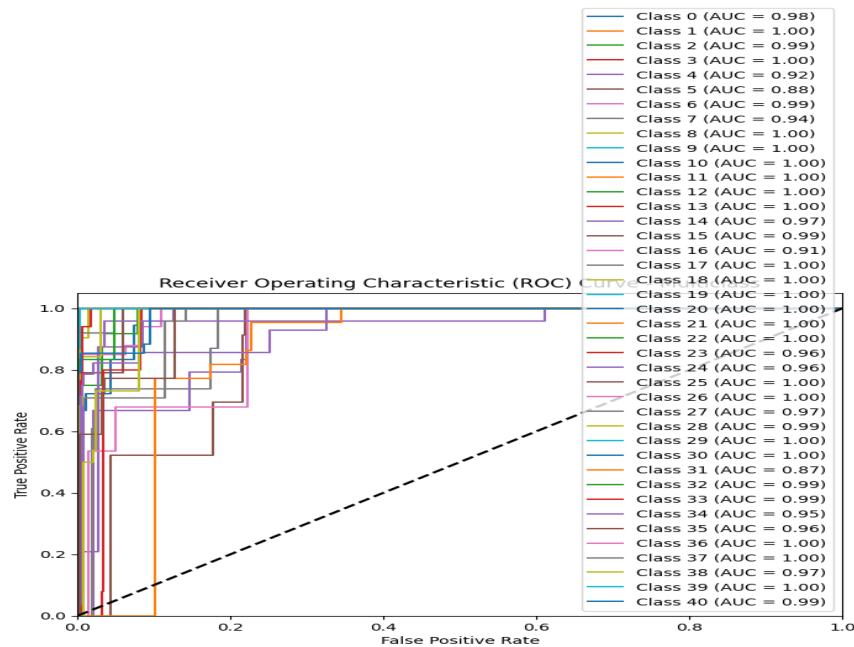


Figure: ROC Curve of Logistic Regression Classifier

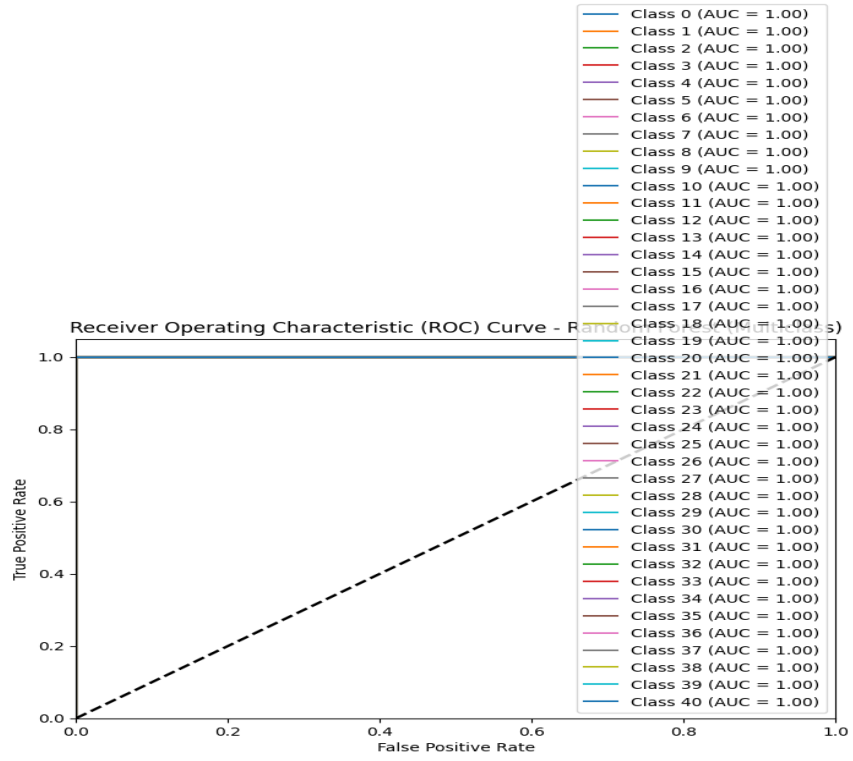


Figure: ROC Curve of Random Forest Model

## Appendix C

### Development of User Interface in Colab

I made an external interface for checking whether our system can be predicted well or not. This interface is designed to take user input for 17 symptoms, convert them to numeric format using a pre-defined mapping, and predict a disease using a pre-trained machine learning model. The predicted disease triggers the retrieval of comprehensive information, including a disease description, the specialist associated with the disease, a recommended set of precautions, and emergency contact details for a corresponding doctor. This external interface ensures a user-friendly experience by allowing an exit option during symptom input and handles incomplete user inputs gracefully. It serves as a comprehensive tool for predicting diseases and offering relevant suggestions based on user symptoms. Code for making a simple interface in Colab:

```
# Load the trained model
loaded_model = joblib.load('/content/multiple_model.joblib')

# Assuming you have a dataset that contains numeric values for symptoms
# Replace '/content/Symptom-severity.csv' with the actual path to your dataset
numeric_dataset_path = '/content/Symptom-severity.csv'
symptom_mapping_df = pd.read_csv(numeric_dataset_path)

# Create a dictionary to map symptoms to numeric values
symptom_mapping = dict(zip(symptom_mapping_df['Symptom'], symptom_mapping_df['weight']))

# User-defined input as object-type data
user_input_symptoms = []

# Take input for 17 symptoms using a loop
for i in range(1, 18):
    symptom = input(f"Enter value for symptom{i}: ")

    # Check if user wants to exit
    if symptom.lower() == 'exit':
        # If 'exit' is entered, set all remaining symptoms to 0 and break from the loop
        user_input_symptoms += ['0'] * (17 - i)
        break

    user_input_symptoms.append(symptom)

# Ensure user input contains values for all 17 symptoms
user_input_symptoms += ['0'] * (17 - len(user_input_symptoms))

user_input_numeric = [symptom_mapping.get(symptom, 0) for symptom in user_input_symptoms]
user_input_numeric = np.array(user_input_numeric).reshape(1, -1)

# Make a prediction
prediction = loaded_model.predict(user_input_numeric)
# Specialist_disease = df_specialist[df_specialist['Disease'] == prediction][['Specialist']]
# Print user input data and the predicted disease
description= df_description[df_description['Disease']==prediction[0]]
description = description.values[0][1]

specialist= df_specialist[df_specialist['Disease']==prediction[0]]
specialist = specialist.values[0][1]

doctor= df_doctor[df_doctor['Disease']==prediction[0]]
doctor = doctor.values[0][1]

precaution= df_precaution[df_precaution['Disease']==prediction[0]]
pre=np.where(df_precaution['Disease']==prediction[0])[0][0]
precaution_list=[]
for i in range(1,len(df_precaution.iloc[pre])):
    precaution_list.append(df_precaution.iloc[pre,i])

print("\nUser Input Data:")
for i, symptom_value in enumerate(user_input_symptoms):
    print(f"Symptom{i + 1}: {symptom_value}")

print("\nPredicted Disease:", prediction[0])
print("\nThe Disease Specialist: ",specialist)
print("\nThe Disease Discription: ",description)
print("\nRecommended Things to do at home: ")
for i in precaution_list:
    print(i)
print("\nIn case of an emergency, you may contact:\n",doctor)
```

This code snippet appears to be a part of a larger program designed to make predictions about a disease based on user input regarding symptoms. The provided symptoms and user Colab user interface look like:

```
Enter value for symptom1: vomiting
Enter value for symptom2: breathlessness
Enter value for symptom3: diarrhoea
Enter value for symptom4: chest_pain
Enter value for symptom5: exit
```

```
User Input Data:
Symptom1: vomiting
Symptom2: breathlessness
Symptom3: diarrhoea
Symptom4: chest_pain
Symptom5: 0
Symptom6: 0
Symptom7: 0
Symptom8: 0
Symptom9: 0
Symptom10: 0
Symptom11: 0
Symptom12: 0
Symptom13: 0
Symptom14: 0
Symptom15: 0
Symptom16: 0
Symptom17: 0
```

Predicted Disease: Heart attack

The Disease Specialist: Cardiologist

The Disease Discription: The death of heart muscle due to the loss of blood supply. The loss of blood supply is usually caused by a complete blockage of a coronary artery, one of the arteries that supplies blood to the heart muscle.

Recommended Things to do at home:  
call ambulance  
chew or swallow aspirin  
keep calm  
nan

In case of an emergency, you may contact:  
Prof. Dr. M. A. Baqui/  
Ibn Sina Specialized Hospital | Dhanmondi  
House : 68, Road: 15/A, Dhanmondi R/A, Dhaka, 1209, Bangladesh

## Appendix D

### PLAGIARISM

201-15-14188

#### ORIGINALITY REPORT

<b>18%</b>	<b>14%</b>	<b>8%</b>	<b>9%</b>
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

#### PRIMARY SOURCES

<b>1</b>	<b>dspace.daffodilvarsity.edu.bd:8080</b> Internet Source	<b>4%</b>
<b>2</b>	<b>docplayer.net</b> Internet Source	<b>1%</b>
<b>3</b>	<b>Submitted to Manipal University</b> Student Paper	<b>1%</b>
<b>4</b>	<b>Submitted to CSU Northridge</b> Student Paper	<b>1%</b>
<b>5</b>	<b>www.techscience.com</b> Internet Source	<b>&lt;1%</b>
<b>6</b>	<b>www.ije.ir</b> Internet Source	<b>&lt;1%</b>
<b>7</b>	<b>Submitted to University of Hertfordshire</b> Student Paper	<b>&lt;1%</b>
<b>8</b>	<b>www.ijert.org</b> Internet Source	<b>&lt;1%</b>
<b>9</b>	<b>Suwendu Kumar Nayak, Mamata Garanayak, Sangram Keshari Swain, Sandeep Kumar Panda, Deepthi Godavarthi. "An Intelligent</b>	<b>&lt;1%</b>