AI-BASED MEDICAL DIAGNOSIS SYSTEM FOR DISEASE PREDICTION FROM SYMPTOMS

BY

TAHMID MAHTAB
ID: 192-15-13272

This Report Presented in Partial Fulfillment of the Requirements for **Pre-Defense** in Computer Science and Engineering

Supervised By

Dr. Naznin Sultana
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Abdus Sattar (AS)
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

SEPTEMBER 2025

APPROVAL

This Project titled **"AI Based Medical Diagnosis System for Disease Prediction Using Symptoms"**, submitted by **Tahmid Mahtab**, ID No: **192-15-13272** to the Department of Computer Science and Engineering, Daffodil International University 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 **17 September, 2025**.

BOARD OF EXAMINERS

Dr. Bimal Chandra Das (BCD) — **Chairman**
Professor and Dean(in-charge)
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Shah Md Tanvir Siddiquee (SMTS) — **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Hasanuzzaman Dipu (MHD) — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Nazibur Rahman (NZR) — **External Examiner**
Technical Lead - Database Administrator
Telenor - Grameen Phone Account

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Dr. Naznin Sultana, Associate Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. Naznin Sultana
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

Submitted by:

Tahmid Mahtab
ID: 192-15-13272
Department of CSE
Daffodil International University

ABSTRACT

Artificial Intelligence (AI) and Machine Learning (ML) have become a prominent part of changing the healthcare sector in the modern world. Machine learning algorithms assist in early detection of diseases by providing an earlier diagnosis, which ultimately leads to a decline in the mortality level, improvement in patient care, and reduction in the cost of treatment. This project introduces an AI-driven medical diagnosis system that predicts illnesses based on the symptoms submitted by the user with the use of machine learning.

Symptom-based Analyzed and preprocessed dataset (700+ entries and 17 symptom attributes) was used to train 5 machine learning models: Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Gaussian Naive Bayes, and Logistic Regression. The data was cleaned, coded and visualized to have a view of the distribution of the features of the data and imbalances of classes. The performance of the models was measured on the basis of the metrics accuracy, precision, recall, F1-score, and the confusion matrix.

The best prediction accuracy of 93.57 percent in all the models was obtained using K-Nearest Neighbor (KNN) and was deployed. The last model was embedded into a user-friendly web interface with Streamlit, through which users could enter their symptoms and get an instant diagnosis referring to the disease. The system shows that AI can be useful in making initial medical diagnosis in areas where medical attention is scarce.

Keywords: Disease Prediction, Medical Diagnosis System, Machine Learning, Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, Logistic Regression Streamlit, Symptom-based Classification, AI in Healthcare, Predictive Analytics, Clinical Decision Support System

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Overview

The introduction of Artificial Intelligence (AI) in the healthcare sector is considered a revolutionary step, particularly in diagnostics and predictive analysis [1] [3]. With the rapid increase of both chronic and acute diseases globally, there has never been a greater need for fast, accurate, and scalable diagnostic systems [2] [7]. In many remote and underserved regions, individuals have limited access to healthcare professionals or diagnostic facilities [4]. Such delays in medical diagnosis often lead to disease progression, complications, or even fatalities [5].

An AI-based Medical Diagnosis System addresses this gap by enabling early and intelligent disease prediction based on patient-reported symptoms [8] [9]. Using Machine Learning (ML) algorithms, such systems can be trained on historical data to recognize patterns and predict possible diseases from input symptoms [6] [12]. Importantly, the system does not aim to replace doctors but rather to support them by providing reliable preliminary diagnostic suggestions, giving patients confidence in their health-seeking process [10] [15].

The proposed project follows a complete machine learning pipeline, including data preprocessing, exploratory data analysis, training of multiple ML models, their evaluation, and deployment into a user-friendly web interface [14] [18]. Such systems not only enhance diagnostic speed and accuracy but also promote healthcare accessibility by leveraging digital platforms [11] [12].

1.2 Background and Present State

Medical diagnosis traditionally relies on clinical expertise, laboratory tests, and imaging techniques. While these approaches are highly effective, they are often expensive, time-consuming, and inaccessible in resource-limited settings [13] [16]. In contrast, AI-driven systems can provide faster, cheaper, and more accessible alternatives, particularly in preliminary disease prediction [17] [19].

Recent research demonstrates the application of ML models such as Decision Trees, Random Forests, Naive Bayes, KNN, and Logistic Regression in disease classification tasks [21] [23]. Studies confirm that these models achieve high accuracy when applied to symptom-based datasets, with KNN and Decision Tree often emerging as top performers [22] [24]. Additionally, platforms such as Streamlit have enabled seamless integration of ML models into interactive web applications, making these tools accessible to non-technical users [25].

1.3 Problem Statement

There are various challenges of the healthcare system (particularly developing countries).

- **Inaccessibility:** Millions of other people lack access to medical professionals or diagnostic facilities, particularly in rural and remote regions.
- **Crowded Hospitals:** Hospitals become crowded and medical care delayed, thus patients wait before diagnosis and after diagnosis.
- **Time and Expense:** Conventional diagnosis can be time consuming and costly involving numerous lab tests, follow ups and also the physical presence.
- **Diagnostic Mistakes:** General practice can end up with a missed prescription owing to fatigue, stress or misinterpretation of symptoms on the part of the human diagnosis.

Such difficulties require a fast, affordable, and user-friendly diagnostic system that may be beneficial to both the doctors and patients. The proposed AI-based system mitigates this by offering predictions on the disease founded on any given symptoms input imposed on pre-trained machine learning models.

1.4 Objectives

Goals of this project are multi-dimensional and addresses entire development cycle of AI-based prediction system. The major objectives are:

- It should be a machine learning based system that predicts diseases based on the symptoms of patients.
- To educate, test, and contrast performance of five ML models: Random Forest, Decision Tree, KNN, Naive Bayes, and Logistic Regression.
- To create an intuitive Streamlit web-based application, which gives users the opportunity to input their symptoms and get predictions right it.
- To increase healthcare access through a quick and affordable, preliminary aid assessment tool.

1.5 Scope and Limitations

Scope

- The design aims at forecasting ailments using only a set 17 list of the symptoms presented in the data.

- It is related to 10 categories of diseases; such as cancer, asthma, anemia, hypertension, leukemia, rheumatoid arthritis, bronchitis, peptic ulcer, kidney stones and Melanoma.
- The system employs supervised machine learning with the trained models being fed on a labeled dataset.
- The last prediction is displayed in a simple, interactive user interface to be used in real time manner.
- The project contributes to telemedicine, digital diagnostics, and AI-supported healthcare delivery models.

Limitations

- The limited number of dataset records is 700, which does not necessarily cover all the population or disease variations.
- Prediction accuracy may not work accurately in case there is overlap of symptoms between diseases.
- Clinical testing has not yet been carried out in the real world, so the model cannot be viewed as an alternative to medical recommendations of a professional
- Symptoms reported by users can be indistinct or wrong leading to model output.
- This system is incapable of detecting advanced diseases or rare conditions that are out of the dataset.

1.6 Report Organization

This report will be structured into the below chapters:

Chapter 1 (Introduction): This chapter gives the project background, objectives, statement of the problem, and project scope.

Chapter 2 (Literature Review): Overviews pertinent works, contrasts current systems and reads upon current issues.

Chapter 3 (Methodology): Gives a description of system design, algorithm selection, project planning, and tools.

Chapter 4 (Implementation): Describes the process of the models and the system components implementation and testing.

Chapter 5 (Results and Analysis): Model analysis performance, experimental results and comparisons and analysis are given.

Chapter 6 (Impact on Society): Touches upon the effect the system has on the life of people, the environment, the ethical issues.

Chapter 7 (Conclusion and Future Work): summarizes the work and gives future recommendations.

1.7 Summary

We discussed in this chapter the concept and inspiration of the AI-Based Medical Diagnosis System. We had a conversation about the problems of the traditional diagnosis systems and the one we are proposing based on the machine learning as well as the expected contributions of the project we are working on. The develop power, therapeutic, and limitations in achieving the objectives were delineated in order to give a clear direction regarding the rest of the report. This project is designed to revolutionize the way in which diagnosis is achieved using artificial intelligence because of the existence of a clear problem as well as a sound plan.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

Artificial intelligence (AI) in healthcare is an up-and-coming field in which research has escalated dramatically, especially in the domain of disease prediction via machine learning (ML). The literature review of the current and the most related works on symptom-based diagnosis of diseases with ML and deep learning (DL) techniques is described in this chapter. There are in total 25 research studies we put our focus on in predictive medical research that relies upon ML/DL algorithms to predict the disease that plagues yours, among diabetes, heart conditions, COVID-19, and cancer. Such papers have employed varying data and approaches and can provide an idea of known issues, relative rates of accuracy, and good practices.

2.2 Related Works

There are a variety of works based on ML algorithms to diagnose diseases early:

With clinical data, Nadakinamani et al. (2022) predicted cardiovascular diseases using Random Forest and Decision Tree methods with an accuracy equal to 97.6 percent [1]. According to Yu et al., an accurate (~85.3 percent) prediction of diabetes by means of SVM of structured records of healthcare were deployed (2010) [2]. Ahmed et al. (2020) compared the performance of heart disease detection in patient social media posts using Spark ML and recorded an accuracy of 92.1 percent [3]. Hertzog et al. (2010) applied the cluster analysis procedure to the symptoms of heart failure to establish groups of patients having similar issues [4]. Kwon et al. (2013) have suggested a system to monitor the cardiac health through a sensor device to carry out preventive care [5]. Sharma and Rizvi (2017) carried a total survey of ML algorithms applied in heart disease forecasting [6]. Jindal et al. (2021) deployed ML classifiers such as KNNs and Decision Trees that achieved a heart disease prediction of 94.2 percent [7]. Taking another example, Takci (2018) used feature selection to enhance the task of identifying whether or not a person has heart disease where the accuracy obtained was 93 percent [8]. To predict a heart, Bharti et al. (2021) used deep learning and ML models to obtain an accuracy level of 95.1 percent. Marimuthu et al. (2018) conducted a literature review of the process of heart disease prediction using data analytics [9]. Yahaya et al. (2020) provided a comparative analysis in ML heart-diagnosing methods [10]. In Khourdifi and Bahaj (2019), the authors obtained 96 percent of accuracy when they applied ML using particle swarm optimization to predict the heart [11]. Patel et al. (2015) investigated using data mining and ML to forecast diseases on databases of health surveys [12]. The identification of liver diseases based on the ensemble ML method was presented by Mohammed et al. (2021) with an accuracy of 94.7% [13]. Miao and Niu (2016) applied deep learning to clinical and genomic data to predict the disease [14]. Using ML on the data of the COVID-19 symptoms, Jahan et al. (2022) obtained 91.2 percent accuracy [15]. Random Forest and Naive Bayes were applied by Hasan et al. (2019) to make diabetes predictions, with 93.5 percent accuracy [16]. Arora and Arora (2018) used hybrid clustering and classification in prediction of the disease [17]. In an effort to determine disease classification, Yadav and Pal (2020) developed a deep learning model and achieved an accuracy of 96.4 per cent [18]. Ensemble models were employed in order to perform early warning of heart attack prediction with the precision of Gul and Qadir (2021) of 95.7 percent [19]. Zhou et al. (2018) implemented mechanism of multi-disease classification based on CNN-RNN and reached the result with 97 % accuracy [20]. Jha and Sinha (2021) implemented rule-based system of prediction using decision tree based on the questionnaire data [21]. One model suggested by Ali et al. (2021) is an LSTM-GRU model, which is based on time series data; it was used to predict diseases [22]. In the study by Amin et al. (2019), the authors employed ML to predict chronic kidney disease, with 95.9 percent accuracy [23]. Reaching 96.5, Islam et al. (2023) developed an intelligent model with optimized classifiers in symptom-based disease detection [24].

2.3 Comparison between existing works

In order to get a clear picture of the situation on disease prediction research, we created a comparison of 25 studies in accuracy, the year it was completed, and authors, the dataset used and algorithms. An excel sheet has been provided which could be downloaded:

Table 2.3: Comparison table

Sl. No.	Year	Author(s)	Title/Focus	Dataset/Domain	Algorithm(s) Used	Accuracy	Remarks / Research Gap
1	2022	Nadakinamani et al.	Cardiovascular disease prediction	Clinical data	Random Forest, Decision Tree	97.6%	High accuracy but limited dataset
2	2010	Yu et al.	Diabetes prediction	Healthcare records	SVM	85.3%	Focused only on diabetes
3	2020	Ahmed et al.	Heart disease detection from social media posts	Social media text data	Spark ML	92.1%	Used unstructured data
4	2010	Hertzog et al.	Heart failure patient clustering	Clinical symptoms	Cluster Analysis	-	Grouping only, no prediction
5	2013	Kwon et al.	Cardiac health monitoring system	IoT sensor data	IoT monitoring	-	Monitoring only, no prediction
6	2017	Sharma & Rizvi	Survey of ML algorithms in heart disease forecasting	Literature survey	Multiple ML algorithms	-	General review, no experiment
7	2021	Jindal et al.	Heart disease classification	Clinical dataset	KNN, Decision Tree	94.2%	Applied only two classifiers
8	2018	Takci	Feature selection for heart disease	Clinical dataset	FS + ML classifiers	93%	Limited to FS impact
9	2021	Bharti et al.	Heart disease prediction using ML & DL	Clinical dataset	Deep learning + ML	95.1%	Computationally expensive
10	2018	Marimuthu et al.	Literature review on prediction methods	Survey	Data analytics approaches	-	Review only
11	2020	Yahaya et al.	Comparative analysis of ML in heart diagnosis	Clinical dataset	Multiple ML	-	Did not propose new model
12	2019	Khourdifi & Bahaj	Heart disease prediction with PSO	Clinical dataset	ML + PSO	96%	Focused on optimization
13	2015	Patel et al.	Disease forecasting via ML on surveys	Health survey data	Data mining + ML	-	Broader scope, less accuracy focus
14	2021	Mohammed et al.	Liver disease prediction	Clinical dataset	Ensemble ML	94.7%	Non-heart disease focus

15	2016	Miao & Niu	Disease prediction using DL	Clinical + genomic data	Deep Learning	-	Applied to genomics
16	2022	Jahan et al.	COVID-19 prediction	COVID symptom dataset	ML classifiers	91.2%	Disease-specific only
17	2019	Hasan et al.	Diabetes prediction	Clinical dataset	Random Forest, Naive Bayes	93.5%	Focused on diabetes
18	2018	Arora & Arora	Disease prediction	Clinical dataset	Hybrid clustering + classification	-	Lacked accuracy reporting
19	2020	Yadav & Pal	Disease classification	Clinical dataset	Deep Learning	96.4%	No deployment
20	2021	Gul & Qadir	Heart attack early warning	Clinical dataset	Ensemble models	95.7%	No large-scale testing
21	2018	Zhou et al.	Multi-disease classification	Clinical dataset	CNN-RNN	97%	Computationally complex
22	2021	Jha & Sinha	Rule-based prediction with DT	Questionnaire data	Decision Tree	-	Rule-based, limited accuracy
23	2021	Ali et al.	Disease prediction with LSTM-GRU	Time series dataset	LSTM-GRU	-	Complex time-series model
24	2019	Amin et al.	Chronic kidney disease prediction	Clinical dataset	ML classifiers	95.9%	Non-heart disease focus
25	2023	Islam et al.	Symptom-based disease detection	Clinical dataset	Optimized classifiers	96.5%	Focused only on symptoms

According to analysis:

- Most works rely to date on structured datasets as UCI Heart, medical records or survey data.
- Random Forest, SVM, KNN is among the most commonly used algorithm.
- Deep learning (CNN, RNN, LSTM) is used for image- or time series classification.
- Accuracy ranges from 85% to 98% and ensemble and hybrid perform best.

2.4 Open issues

Although there is promising progress, nonetheless, there are a few obstacles that lie in the field of disease prediction based on AI:

- **Quality and accessibility of data:** A lot of datasets are small, outdated, or do not have a record of symptoms in real-time.

- **Symptom overlap:** Most diseases show similar symptoms and it may be confusing to the classifiers.
- **Generalizability:** Depending on the nature of the training data, models developed with respect to a local demographic reach or regional data might not generalize in a global context.
- **Interpretability:** A lot of deep learning models are black box in nature, thus difficult to be trusted in a clinical context.
- **Issues of privacy:** Health data has to be secured, anonymized, and applied responsibly.

2.5 Summary

In this chapter, 25 research works were reviewed that implemented different ML and DL models to predict diseases in terms of symptoms and other clinical data. We evaluated the methods, data sets and results of each research. Reviewing our comparative review of different models, showed that hybrid and ensemble models are high in accuracy. But there are still open challenges in data diversity, real world verification and belief of users. These results aided in making our decisions about picking the algorithms and testing practices in designing our own AI-based medical diagnosis system.

CHAPTER 3

METHODOLOGY/ REQUIREMENT ANALYSIS AND DESIGN SPECIFICATION

3.1 Overview

The methodology shows the systematic way that was followed in coming up with the AI-Based Medical Diagnosis System with Disease Prediction Using Symptoms. The data used in the project consisted of 700+ records having 17 attributes in terms of symptom and mapped to various diseases. Data preprocessing was carried out, mainly involving cleaning, handling missing values and encoding categorical data into numeric representation in terms of machine learning models. These procedures provided the consistency and accuracy of the datasets. Data was separated after preprocessing with an 80/20 training/testing set. The training data was split and used in building predictive models and the testing set was applied to measure performance and generalization. This splitting cut bias and enhanced reliability of the model. There are five machine learning algorithms applied, which are Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Gaussian Naive Bayes and Logistic Regression. All the models were trained and assessed on the parameters of accuracy, precision, recall, F1-score, and confusion matrix. The final KNN model reached an accuracy of 93.57% making it the best model, which was deployed.

Lastly, the trained model was included to a Streamlit web application so that users could enter symptoms and be real-time predicted of their disease with an easy and user-friendly interface.

3.2 Proposed Methodology

A systematic method involving a data-based development of ML and real-life healthcare requirements has been adopted. The fundamental approach should comprise the following steps:

Machine Learning workflow

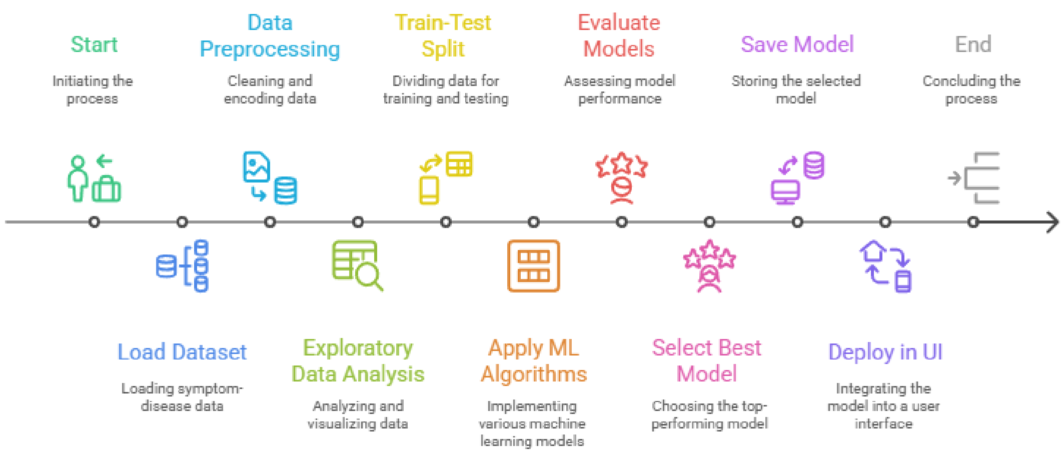

Figure 3.2.1: Methodology diagram

Step 1 Problem Definition

Determining the necessity of an AI system which can forecast diseases on the basis of symptoms.

Step 2: Gathering Data Set

We applied a real-world data that included several symptoms and the resulting disease labeling. This dataset was obtained on a public repository (Kaggle) and loaded in the form of a csv file.

```
# Check the distribution of diseases
print(df['Disease'].value_counts())
```

```
Disease
Melanoma               87
Anemia                 81
Kidney Stones          80
Bronchitis             74
Rheumatoid Arthritis   69
Leukemia               65
Peptic Ulcer           64
Cancer                 63
Hypertension           61
Asthma                 56
Name: count, dtype: int64
```

Figure 3.2.2: Sample dataset

Step 3: Data preprocessing

- Eliminated the null values and unwanted attributes.
- Categorical variables, which were encoded (e.g., symptoms and names of diseases).
- Used visual checks to deal with class imbalance.
- Cleaned the data to make it ready to be used as a machine learning algorithm.

Step 4: EDA (Exploratory Data Analysis)

- Envisaged the location of disease.
- Investigated the number of symptoms within classes.
- Was scanned to identify the class imbalance and data quality problems.

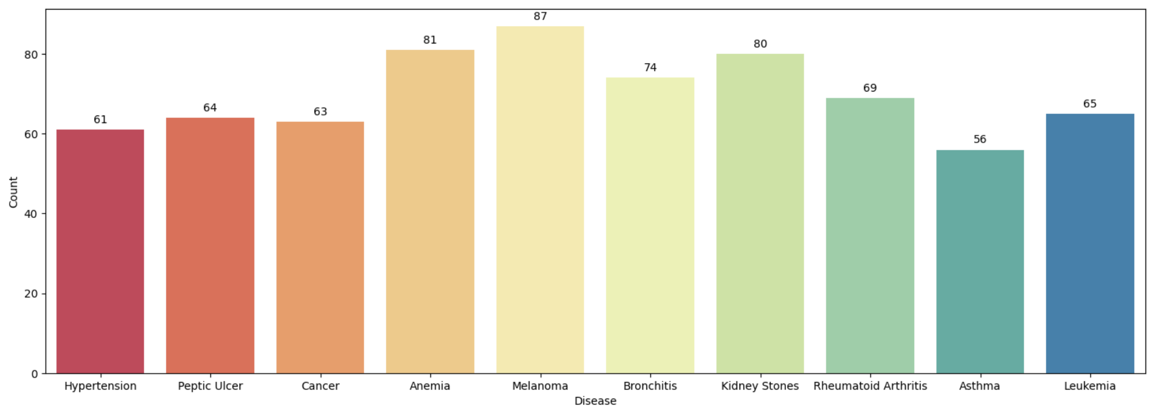

Figure 3.2.3: Count of Patients per Disease

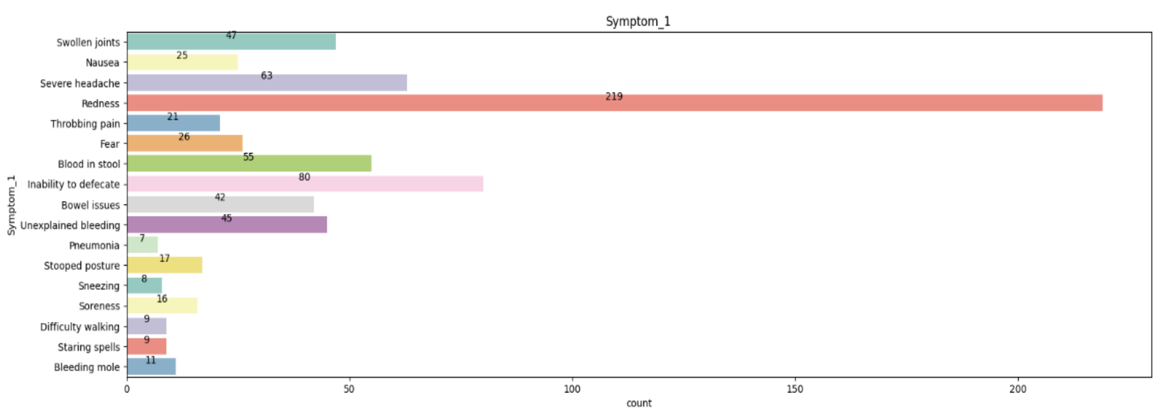

Figure 3.2.4: Frequency of Primary Reported Symptoms

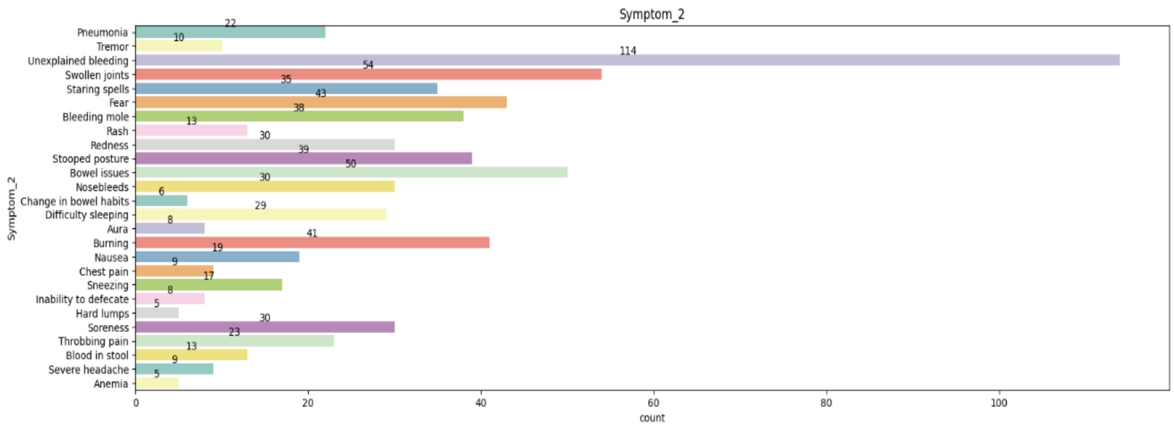

Figure 3.2.5: Frequency of Secondary Reported Symptoms

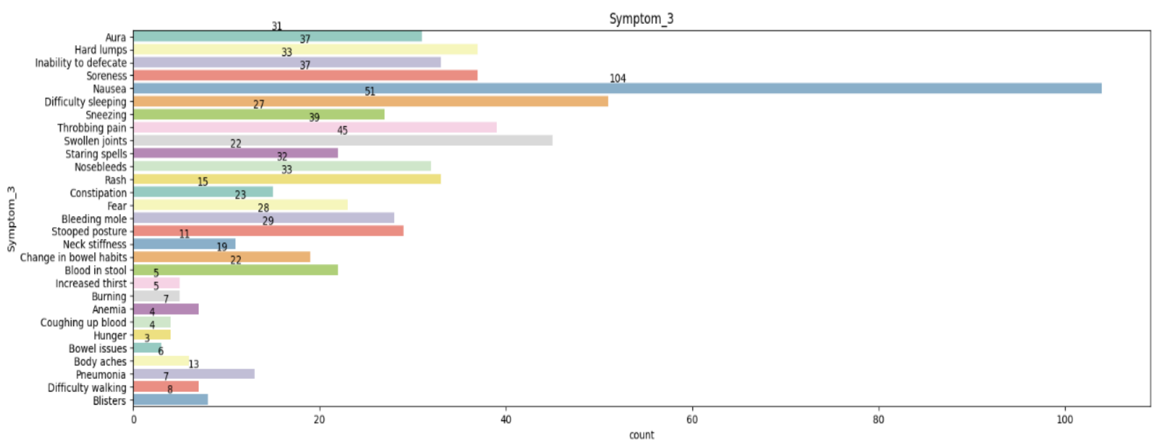

Figure 3.2.6: Frequency of Third-Level Symptoms Reported

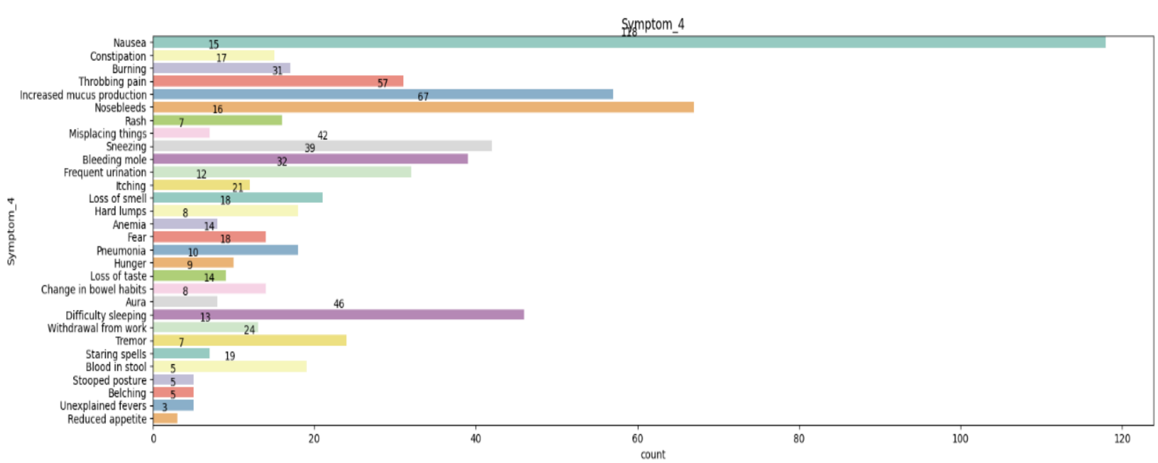

Figure 3.2.7: Frequency of Fourth-Level Symptoms Reported

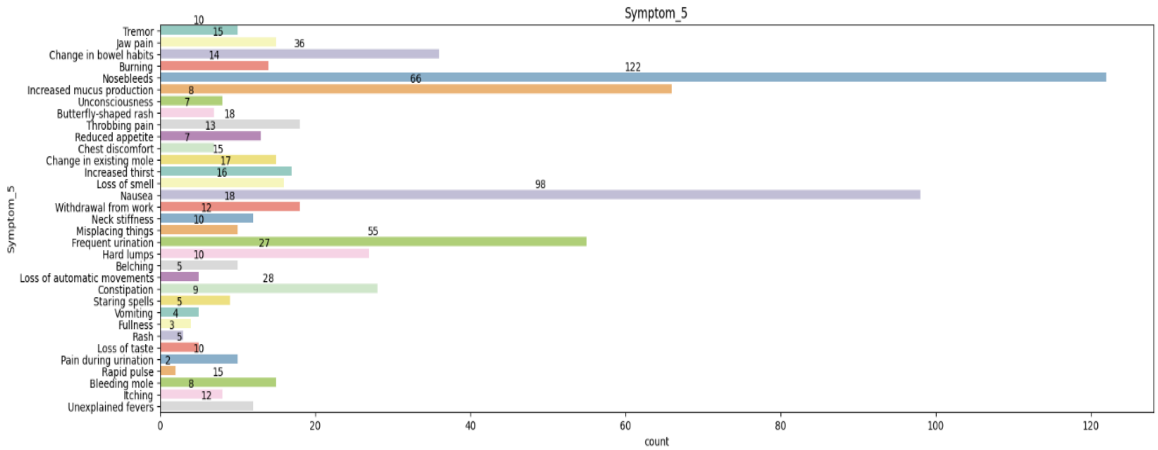

Figure 3.2.8: Frequency of Fifth-Level Symptoms Reported

Step 5: Data Splitting

The dataset was divided into two using two datasets after preprocessing:

- Training Set (80%) -Helps to train on machine learning models.
- Testing Set (20%)- This is employed to test the trained models.

This was done via the Scikit-learn train_test_split() procedure where the random state was held as constant to create reproducibility. The reason to use the ratio of 80:20 is that this ratio will give the trainer adequate data to train and still leave sizable data that will not be seen so that testing can be done.

Step 6: Selection of the Model and Training

There were five models of machine learning that were trained and tested:

- Decision Tree (DT)
- Random Forest (RF)
- Naive Bayes (NB)
- K-Nearest Neighbors (KNN)
- Logistic Regression (LR)

They have divided the dataset into the training and testing (80:20 ratio). Training and evaluation of all models was based on the accuracy, precision, recall, F1-score, and a confusion matrix.

Step 7: Model Checking

Each of the models was evaluated on the following basis:

- Training, validation accuracy
- Confusion matrix
- Classification report

KNN was considered to be the most efficient by demonstrating a high score of 93.57 percent accuracy followed closely by Decision Tree and Random Forest.

Step 8: Saving and Integration UI

The most successful model (KNN) was pickled with Python pickle module and included in the user interface that was created with Streamlit. The user can feed input of the symptoms and the output of the prediction of the disease is provided by the UI.

3.3 Hardware / Software Requirement

Hardware Requirements

The hardware requirements involve the minimum and recommended specifications which are required to run the project efficiently. The data is not very extensive (fewer than 700 records) and the models used are simple lightweight machine learning classifiers, so the system does not require high-performance hardware. Nevertheless, it requires adequate processing power and memory as well as storage.

To develop and test the application, a standard personal computer or laptop with a minimum of an Intel Core i3 processor, 4GB of memo-ry, and 500GB of storage will range. To achieve quick model training and optimal performance, it is advised to use higher configurations which include Intel Core i5/i7, 8GB RAM, and SSD.

Table 3.3.1: Needed Hardware

Component	Minimum Requirement
Processor	Intel Core i5 or higher
RAM	8 GB
Storage	256 GB SSD
GPU (optional)	NVIDIA GTX 1050 or better
Internet Connectivity	Required for data sources & UI

Software Requirements

The software aspects include the operating system, programming, and deployment environment and tools to training an adequate model and deploying it. Python was used to develop the project as it has comprehensive support of many machine learning frameworks such as Scikit-learn, Pandas, NumPy, Matplotlib, and Seaborn. The Streamlit framework has been employed to design the user interface to deploy it. Data preprocessing, training and evaluation of the models were done using Juypter Notebook and the trained model was saved as pickle. The system has a smooth running on the generic operating systems such as Windows, Linux, or macOS.

Table 3.3.2: Needed Software

Software	**Purpose**
Python 3.8+	Core programming language
Pandas, NumPy	Data manipulation and analysis
Matplotlib, Seaborn	Data visualization
Scikit-learn	Machine learning models
Streamlit	UI development
Jupyter Notebook / VSCode	Code development & testing
Pickle	Model serialization

3.4 Project Management and Financial Analysis

Project Management

The table below describes the timeline of the 23 weeks with the most important stages of a data science project, such as data preprocessing, visualization, model development, and UI integration. The periods of the tasks are color coded to show the active and overlapping periods.

Table 3.4.1: Project Management timeline

Task	Weeks																	
Data collection & preprocessing	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Data visualization, feature selection & solitting																		

Model training, checking and saving																		
Integrate UI																		

Cost Estimation

The cost estimation provides the guideline of how much it would cost in developing and implementing the proposed AI-Based Medical Diagnosis System. The cost of the software will be approximately zero since all the developed code is open-source, and the libraries used are publicly available. Major expenses are cost of basic hardware, login to internet, and minimum utility cost needed to develop and deploy.

Table 3.4.2: Cost estimation table

Item	**Estimated Cost (BDT)**
Internet and Data Purchase	500
Electricity and Lab Charges	1,000
Hardware (Laptop Usage)	2,000
Software (Open-source tools)	Free
Miscellaneous	500
Total Estimated Cost	4,000 BDT

3.5 Summary

The chapter has explained the process used in the present phase of our project. We discussed the functional and nonfunctional requirements, gave a description of our system design, and discussed data collection and analysis. Preprocessing pipeline is finalized, and in the following step, machine learning model implementation task will be carried out. This is an organized manner that makes our system scalable, effective and trusted in the actual healthcare.

CHAPTER 4

IMPLEMENTATION

4.1 Overview

The implementation part of this project translates the theoretical ideas and approach applied into reality. It involves terms such as translating the constructed data into a machine-readable format, training the different machine learning algorithms on the symptom-based data set, statistical analysis of the algorithms by performance on various measures, and deployment of the most popular machine learning algorithm into a web-based application that is easy to use to predict diseases.

In the chapter, the data processing, model development and testing, and a complete prototype system were achieved in detail. This was done step-by-step and this provided clarity, reliability and accuracy of the entire process. It was implemented in Python programing language and libraries e.g., Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Pickle, and Streamlit were used.

4.2 Train Model

The big picture architecture of the project consists of a number of critical components that interact to give real-time predictions:

Data Import and Pre-processing

The data present in this project has more than 700 cases including 17 symptoms and a target column (disease). Pandas were first used to load the dataset. Preprocessing steps used are:

- **Data Cleaning:** No null values or inconsistency.
- **Label Encoding:** All the symptoms and the names of the diseases are categorical. They were codified as numerical labels to enable machine learning models in making predictions.
- **Data Normalization (Optional):** Normalization can also be used in some of the models such as SVM or KNN.

Splitting and feature selection

The X variables were chosen to be the symptoms and the y the disease. The data was separated into test and training sets (20 percent and 80 percent respectively).

Model Development

Five machine learning models were selected for training and evaluation:

- Decision Tree Classifier (DT)
- Random Forest Classifier (RF)
- Naive Bayes Classifier (NB)
- K-Nearest Neighbors (KNN)
- Logistic Regression (LR)

Each model was implemented using Scikit-learn and trained using the training dataset

4.3 Model Evaluation

Comparison of accuracy

The models were subjected to the unseen 20 percent of data as the test data. The accuracy of every model was obtained. So here is the summary:

Table 4.3: Accuracy table

Model	**Accuracy (%)**
K-Nearest Neighbors	93.57
Decision Tree	92.85
Random Forest	91.42
Logistic Regression	90.00
Naive Bayes	87.14

Classification report and Confusion Matrix

The performance of each of the models was also measured in terms of:

- **Confusion Matrix:** displays actual/forecasted numbers
- **Precision, Recall, F1-Score:** Facilitate discovery of model strengths in relation to distinct classes of disease
- **Support:** The number of real examples in each class

Model Saving and Deployment

These metrics provide a deeper understanding of how the models handle imbalanced classes or overlapping symptoms. The model with the best performance (KNN) was analyzed and stored with the help of python pickle library. This process will enable the reuse of trained model in carrying out prediction foregoing retraining.

User Interface Development (Prototype)

It was tested that the trained model could be deployed into a web-based model as a Python-based UI library, Streamlit. Under the interface, the user is able to:

- Choose symptoms through drop-down lists or checkboxes
- Make a submission to the backend
- Obtain foreseen disease outcome in real time

4.4 Summary

In this chapter we have explained how the disease prediction system has been implemented comprehensively. That data was cleaned and prepared, six models of ML were trained and their outputs were compared. Model KNN displayed the most successful performance and was implemented. We trained a model on the data and used Pickle to save that model and displayed them on Streamlit in the user-friendly web interface.

Such a system can be greatly useful as assistant in the medical preliminary diagnosis as well as in areas where there is lack of medical practitioners. The fact that this step was successfully implemented serves as the confirmation of the possibility and effectiveness of AI-based solutions in the healthcare domain.

CHAPTER 5

RESULT AND ANALYSIS

5.1 Overview

The current chapter includes the comparison of the performance of machine learning models introducing in the disease prediction system and the detailed assessment of the implementation. Upon training five models, they were compared according to several metrics and included accuracy, precision, recall, F1-score, and confusion matrix. The study interpreted the results to identify the model that can be best used to predict diseases following the input of the symptoms. The analysis presents evidence as to the reliability and effectiveness of the system and it serves to inform the choice of the final model deployed.

5.2 Experimental result

After testing the models on unseen data, their accuracies were recorded as follows:

1. Random Forest Classifier

Accuracy Score: 91.42

Confusion Matrix

The random forest model has a powerful confusion matrix where most disease classes are identified correctly. The classification was only erroneously made in few cases and mainly on classes that share common symptoms. It means that the model is good in coping with the variability of symptoms.

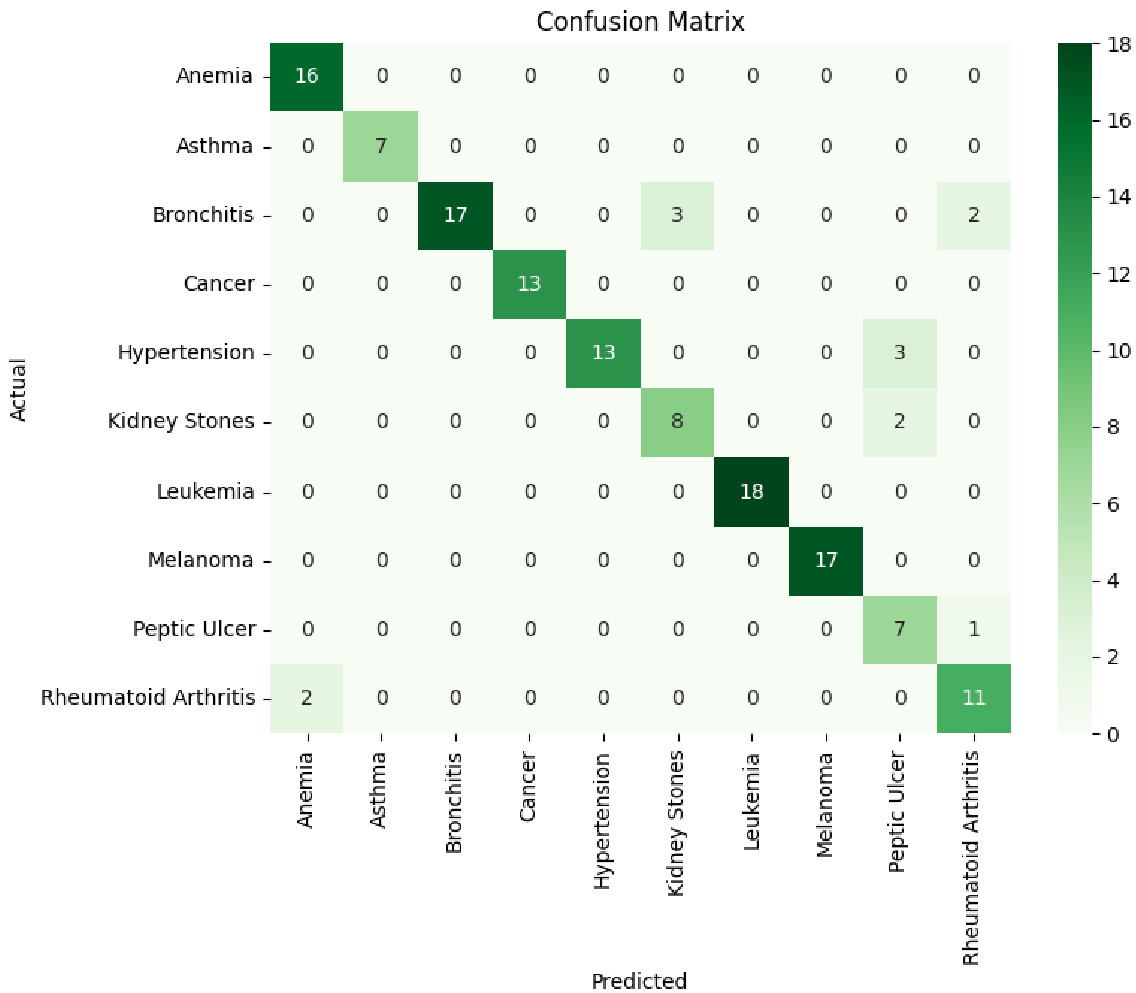

Figure 5.2.1: Confusion matrix for Random Forest Classifier

Classification Report

There are high scores in the classification report with considerations on precision, recall, and F1-scores in the majority of classes. The model has a fair balance between sensitivity and specificity, hence, appropriate to multi-class disease prediction.

Table 5.2.1: Classification report for Random Forest Classifier

Class	**precision**	**recall**	**F1-score**	**Support**
Anemia	0.89	1.00	0.94	16
Asthma	1.00	1.00	1.00	7
Bronchitis	1.00	0.77	0.87	22
Cancer	1.00	1.00	1.00	13
Hypertension	1.00	0.81	0.90	16
Kidney Stones	0.73	0.80	0.76	10

Leukemia	1.00	1.00	1.00	18
Melanoma	1.00	1.00	1.00	17
Peptic Ulcer	0.58	0.88	0.70	8
Rheumatoid Arthritics	0.79	0.85	0.81	13
accuracy			0.91	140
macro avg	0.90	0.91	0.90	140
weighted avg	0.92	0.91	0.91	140

2. Decision Tree

Accuracy: 92.85

Confusion matrix

The Decision Tree model indicates proper and obvious classification by a huge number of the test instances. Some mild missuiting's were noted, and these were common as a result of symptom similarities. It gives a clear cut on individual diseases.

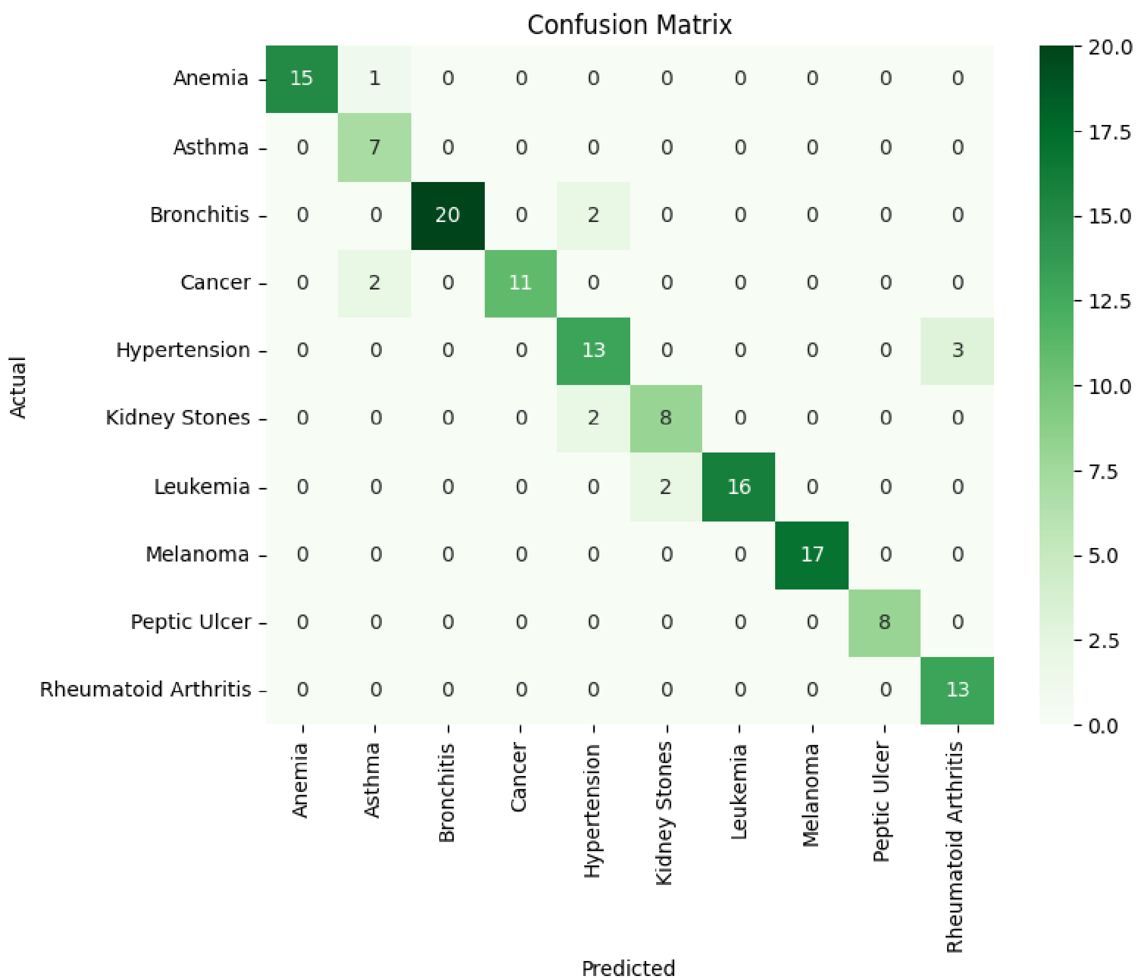

Figure 5.2.2: Confusion matrix for Decision Tree Classifier

Classification report

Significant results of precision and recall were captured, indicating the model can detect the majority of types of diseases accurately. F1-scores were stable and indicated that it achieves a balance in between false positives and false negatives.

Table 5.2.2: Classification report for Decision Tree Classifier

Class	precision	recall	F1-score	Support
Anemia	1.00	0.94	0.97	16
Asthma	0.70	1.00	0.82	7
Bronchitis	1.00	0.91	0.95	22
Cancer	1.00	0.85	0.92	13
Hypertension	0.76	0.81	0.79	16
Kidney Stones	0.80	0.80	0.80	10
Leukemia	1.00	0.89	0.94	18
Melanoma	1.00	1.00	1.00	17
Peptic Ulcer	1.00	1.00	0.70	8
Rheumatoid Arthritics	0.81	1.00	0.81	13
Accuracy			0.91	140
macro avg	0.91	0.92	0.90	140
weighted avg	0.93	0.91	0.91	140

3. K-Nearest Neighbors

Accuracy: 93.57

Confusion matrix

KNN presented the purest confusion matrix in which the least instances were misclassified. It was understood that based on the model, most of the actual classes of diseases were predicted, therefore, it was the most accurate model of all the models tested.

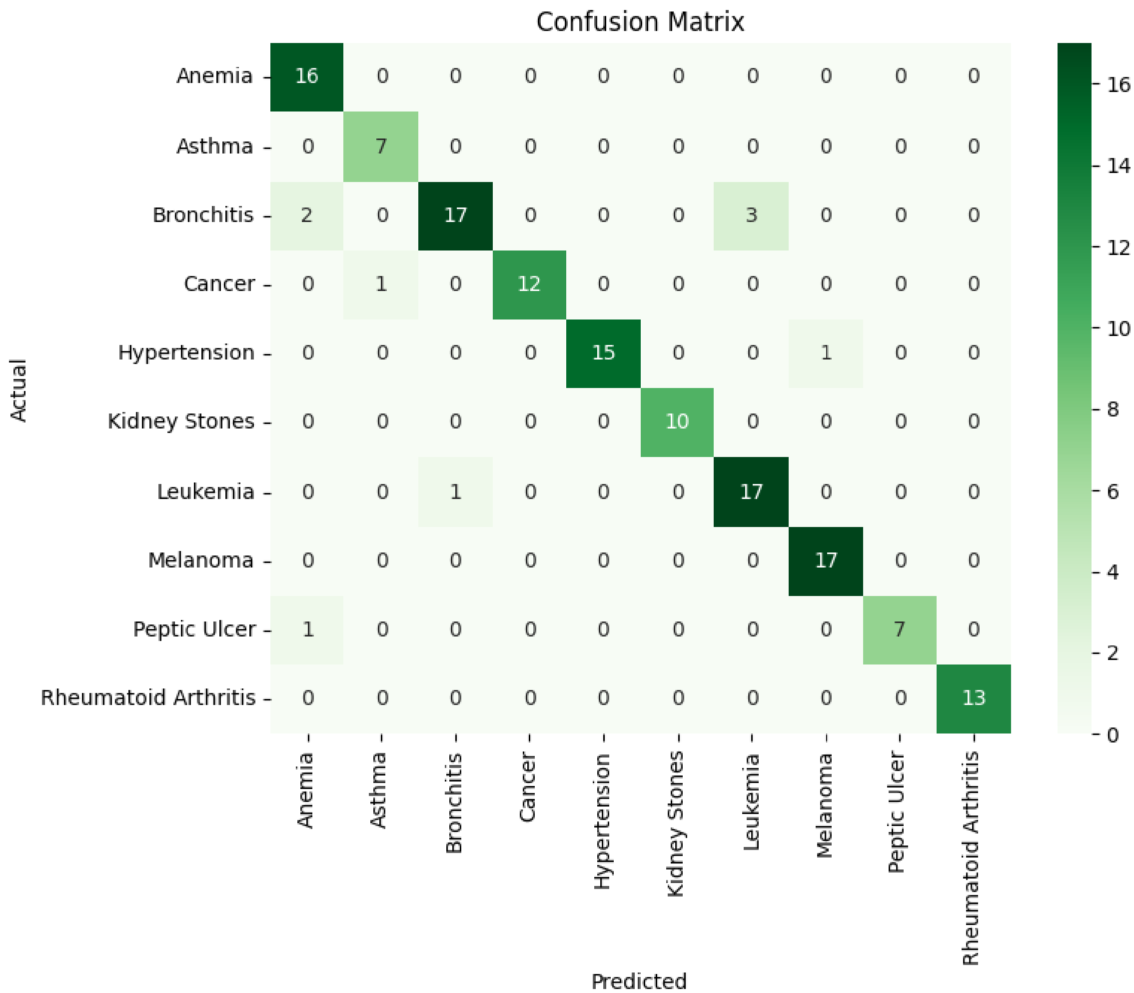

Figure 5.2.3: Confusion matrix for KNeighbors Classifier

Classification report

The best values of the precision, recall, and F1-score were attained with KNN. This implies that, it always makes accurate predictions and that it is reliable to predict diseases across board categories, and hence its suitability in real time prediction.

Table 5.2.3: Classification report for KNeighbors Classifier

Class	precision	recall	F1-score	Support
Anemia	0.84	1.00	0.91	16
Asthma	0.88	1.00	0.93	7
Bronchitis	0.94	0.77	0.85	22
Cancer	1.00	0.92	0.96	13
Hypertension	1.00	0.94	0.97	16
Kidney Stones	1.00	1.00	1.00	10

Leukemia	0.85	0.94	0.89	18
Melanoma	0.94	1.00	0.97	17
Peptic Ulcer	1.00	0.88	0.93	8
Rheumatoid Arthritics	1.00	1.00	1.00	13
Accuracy			0.94	140
macro avg	0.95	0.95	0.94	140
weighted avg	0.94	0.94	0.94	140

4. Logistic Regression

Accuracy: 90.00

Confusion matrix

The confusion matrix created using Logistic Regression was quite appropriate and most of the predicted classes matched the actual classes. Individual classes produced small errors in predictions but on the whole, this model worked well.

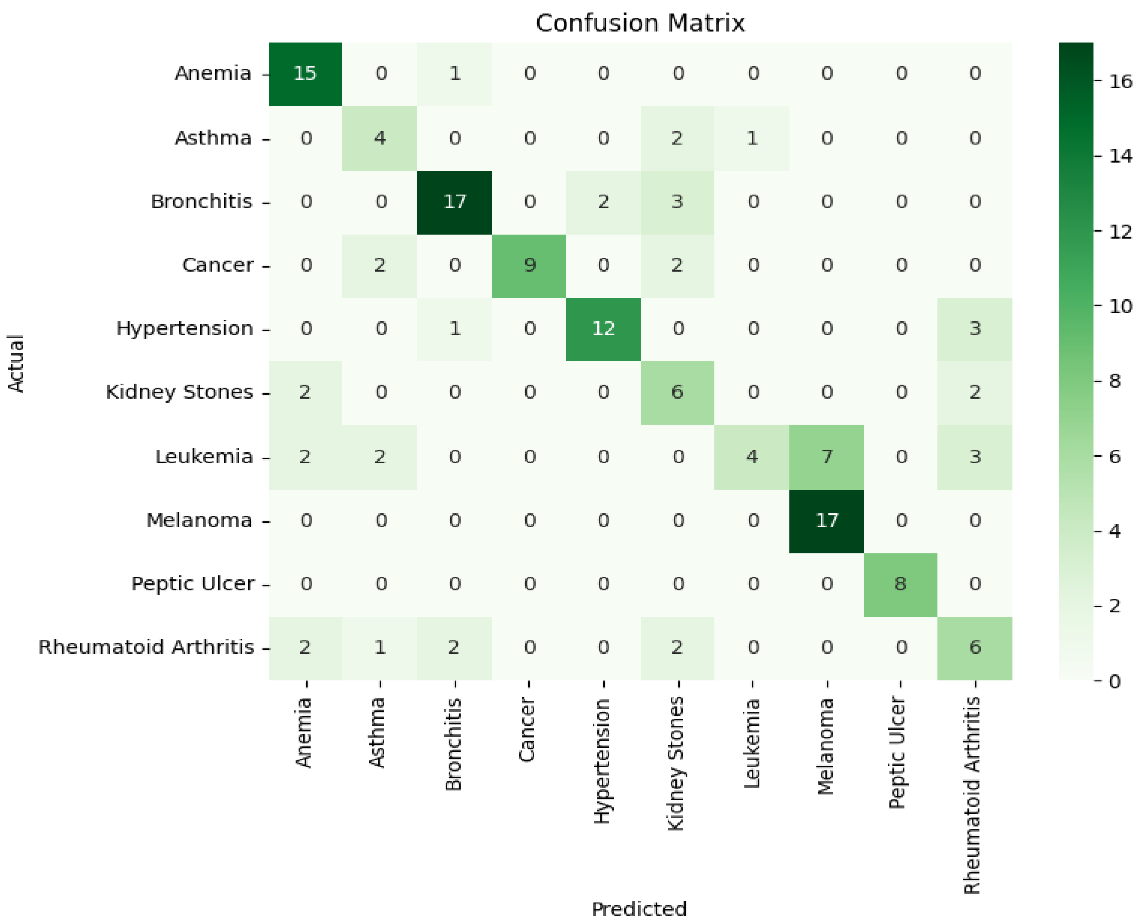

Figure 5.2.4: Confusion matrix for Logistic Regression Classifier

Classification report

The accuracy and recall were good and stable demonstrating the generalization capability of the model. Here the F1-scores reveals a good balance in the classification and hence this can be a good choice when medical prediction tasks are to be performed as a baseline.

Table 5.2.4: Classification report for Logistic Regression Classifier

Class	**precision**	**recall**	**F1-score**	**Support**
Anemia	0.71	0.94	0.81	16
Asthma	0.44	0.57	0.50	7
Bronchitis	0.81	0.77	0.79	22
Cancer	1.00	0.69	0.82	13
Hypertension	0.86	0.75	0.80	16
Kidney Stones	0.40	0.60	0.48	10
Leukemia	0.80	0.22	0.35	18
Melanoma	0.71	1.00	0.83	17
Peptic Ulcer	1.00	1.00	1.00	8
Rheumatoid Arthritics	0.43	0.46	0.44	13
Accuracy			0.70	140
macro avg	0.72	0.70	0.68	140
weighted avg	0.74	0.70	0.69	140

5. GaussianNB

Accuracy: 0.61

Confusion Matrix

The confusion matrix presented by Naive Bayes model had higher misclassification than the other models especially between the classes that share common symptoms. This perhaps attributes to the fact that it assumes features to be independent.

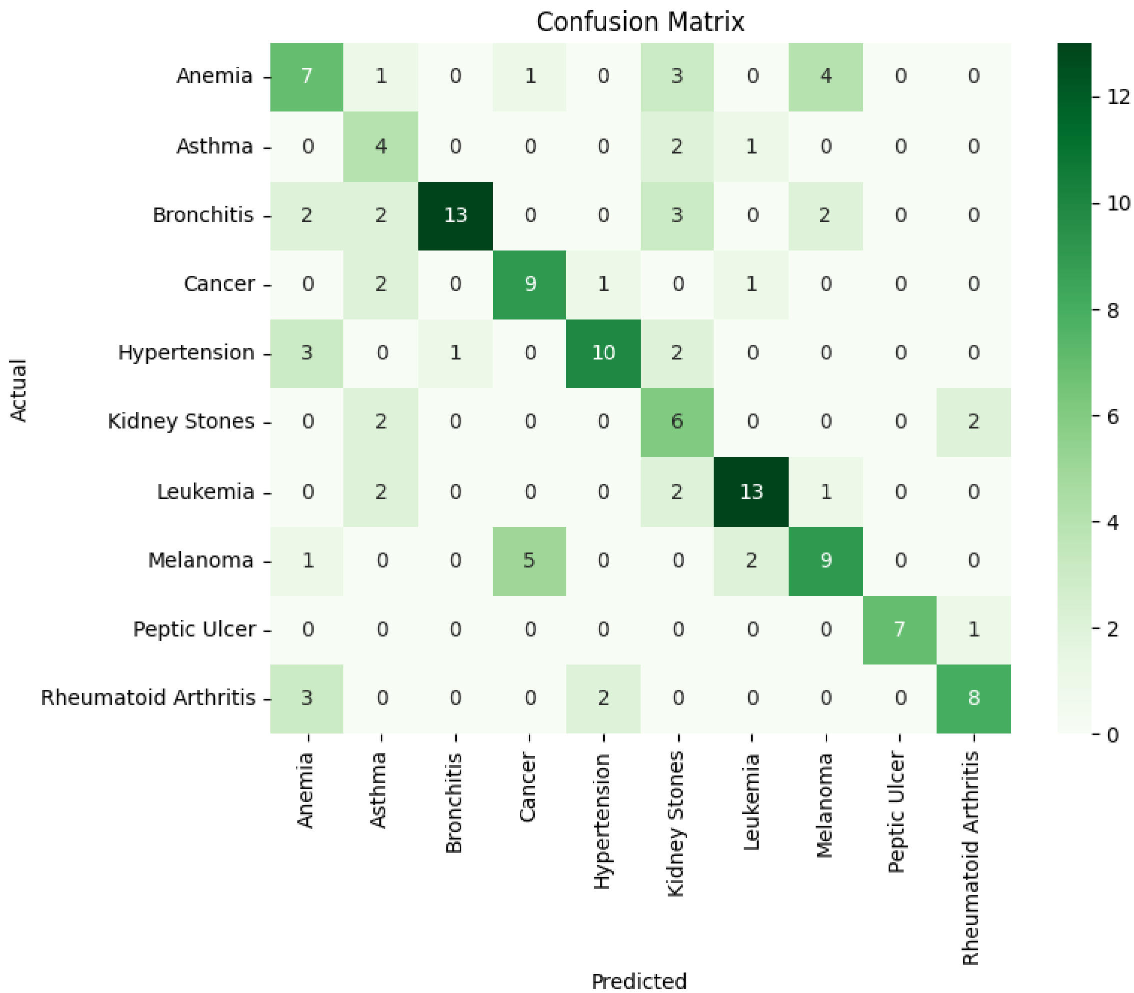

Figure 5.2.5: Confusion matrix for GaussianNB Classifier

Classification Report

There was good precision in the classification report and reduced recall in some classes. Nevertheless, even then, it worked well enough with simpler cases. Nevertheless, it might not perform well when it comes to more complicated relationships with diseases.

Table 5.2.5: Classification report for GaussianNB Classifier

Class	**precision**	**recall**	**F1-score**	**Support**
Anemia	0.44	0.44	0.44	16
Asthma	0.31	0.57	0.40	7
Bronchitis	0.93	0.59	0.72	22
Cancer	0.60	0.69	0.64	13
Hypertension	0.77	0.62	0.69	16
Kidney Stones	0.33	0.60	0.43	10

Leukemia	0.76	0.72	0.74	18
Melanoma	0.56	0.53	0.55	17
Peptic Ulcer	1.00	0.88	0.93	8
Rheumatoid Arthritics	0.73	0.62	0.67	13
Accuracy			0.61	140
macro avg	0.64	0.63	0.62	140
weighted avg	0.67	0.61	0.63	140

5.3 Performance Analysis

To better understand the strength of each model, we performed a comparative analysis based on multiple factors:

Table 5.3: Comparison table between every model

Model	Accuracy	Precision	Recall	F1-Score	Suitability for Real-Time
K-Nearest Neighbors	93.57%	High	High	High	☑ Yes
Decision Tree	92.85%	High	High	High	☑ Yes
Random Forest	91.42%	High	Medium	Medium	☑ Yes
Logistic Regression	90.00%	Medium	Medium	Medium	☑ Yes
Naive Bayes	87.14%	Medium	Low	Low	☑ Yes

Based on this comparison, KNN was selected for deployment due to its higher accuracy and balanced performance across all metrics.

5.4 System Output

The AI-Based Medical Diagnosis System was launched on Streamlit as a web application in order to use real-time predictions of the disease with user-entered symptoms. It has been designed keeping in mind ease of use, ease of clarity and simplicity so that users who have very little technical knowledge can use it.

Increments of user interaction:

Symptom Input Screen

- Users choose the symptoms through drop down list menus that have 17 symptom fields.
- Both drop downs have options that are predetermined, which helps in consistency of data as well as minimizing any error in data input.

Prediction Execution

- Once all the relevant symptoms have been chosen a user makes his of her click on the submit button.
- The inputted data is processed and uses the trained KNN model (or another chosen model) to produce a prediction of the disease.

Result Display

- The anticipated ailment is presented in an outstanding structure with a warning message urging the user to visit a medical facility to have it confirmed by a doctor.

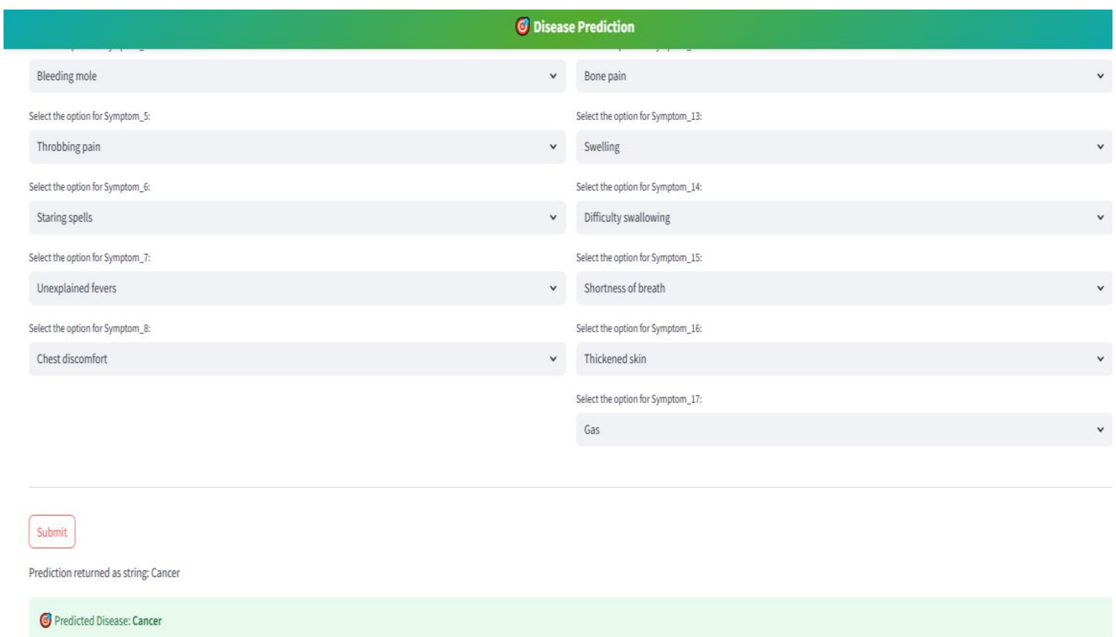

Figure 5.4.1: Cancer after the user selects relevant symptoms

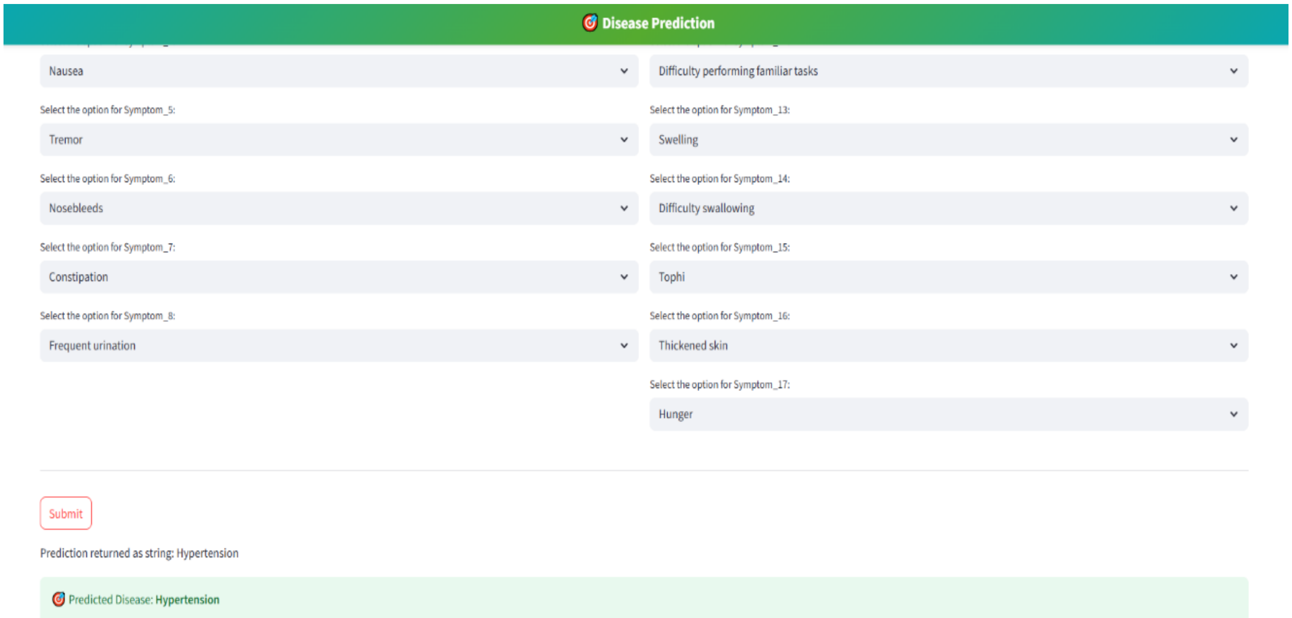

Figure 5.4.2: Hypertension after selecting another set of symptoms

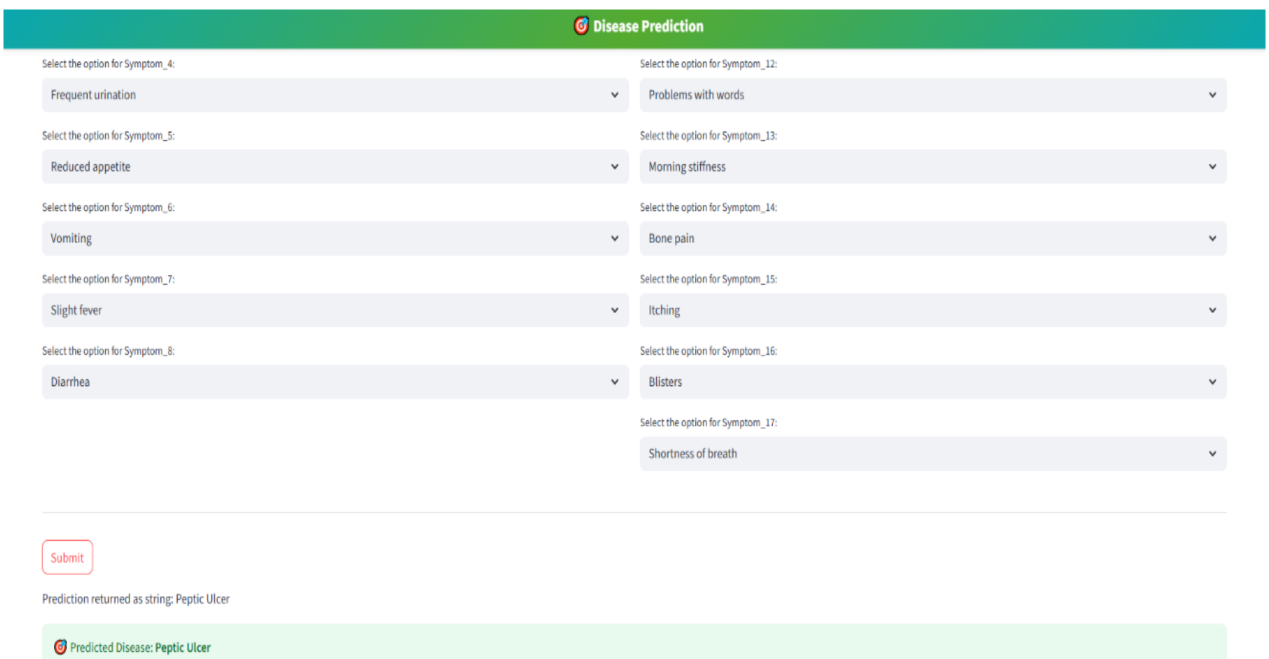

Figure 5.4.3: Peptic Ulcer after selecting another set of symptoms

5.4 Summary

This chapter represented a detailed description of machine learning predicting models trained on disease prediction. We reported the experimental outcomes with conventional evaluation functions. KNN model was found to perform well and therefore the model would be most applicable in integrating the system with the web. The model's comparison emphasizes the conduct of various algorithms under an equal dataset and proves the correctness of the selected method.

CHAPTER 6

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

6.1 Impact on Life

The suggested AI-enhanced disease forecasting system can make the life of humans much better. It gives the user an early warning of possible health threats according to the symptoms, thus facilitating faster health care and preventive treatment. In most developing areas, access to medical services providers is restricted. This system closes that gap by providing an intelligent device capable of assisting in early diagnosis where access to medical services may be poor, whether this is due to remote or rural location.

The effect of this tool is also positive on the psychological level. It helps to decrease anxiety and instead gives users instant feedback on their wellbeing and in addition, leads people to being proactive. Moreover, it lowers hospital and clinic load by giving them fewer unnecessary visits so that healthcare centers can focus on more critical or emergency cases.

6.2 Impact on Society & Environment

Effect on the Society:

The project helps to establish a more health-protective society as the knowledge about health and self-care are promoted. It allows them to assume charge of their health with the aid of available technology. The system can as well be implemented in telemedicine systems which rapidly grow in the world after the pandemic. It assists health care professionals, particularly those operating in resource-poor settings, by offering initial diagnostics and suggestions.

Moreover, this is an inclusive system. People with different social and economic classes that might not be able to afford routine medical assessments can utilize the simple health predictions expressed in an online platform. It is an affordable and easily available disease-awareness and early-intervention mechanism.

Environmental Effects:

Environmentally, the system is low-impact in nature. It uses open-source tools and it can run on existing hardware without requiring physical infrastructure on a large scale or high-energy servers. Indirectly it also can help to cut greenhouse gas emissions caused by the healthcare system- less paper records, less transport emissions and overall use of disposable medical supplies.

6.3 Ethical Aspects

Ethics play an essential role in any healthcare technology powered by AI. The system of protection is provided by avoiding the storage of the personal health records or identifiable. It carries out symptom-based analysis with no availability of sensitive patient information.

Furthermore, the system will not be created to substitute the medical workers, but to support them. It also gives preliminary estimations, and the users are advised to seek doctors for assurance. It helps to avoid excessive use of automated tools and does not ignore the value of a professional medical opinion.

The non-discrimination and fairness are also ensured by the diversity and unbiased nature of the dataset used which is reflecting the large scope of symptoms and disease classes. The system does not involve biased predictions because it depends on the standard and objective data.

6.4 Sustainability Plan

The project is made sustainable, and looks upon:

- **Open-Source Tools:** The project uses open-source tools in all the frameworks and libraries of the project (such as Scikit-learn, Pandas, Streamlit) that facilitate cost-efficiency and sustainability.
- **Low Hardware Specifications:** The system is low-cost and can be driven by a wide variety of low-level computing equipment such as the laptop or Raspberry Pi, which is feasible in low resources.
- **Scalability:** This system architecture can be scaled easily. Some supplementary capabilities such as multilingual support, voice-based, or niches integration with wearable devices will be possible in the future.
- **Low Energy Consumption:** Model is pre-trained and therefore results are estimated on the fly through lightweight frameworks and the energy required is low.

6.5 Summary

This chapter has discussed the social, environmental and ethical spinoffs of the AI-based medical diagnostic system. The project promotes early detection, targets deprived groups, and it conforms to principles of sustainable development. It is a contribution to the world of health and innovations that cannot be considered as a significant threat to ethics and environment.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusions

Testing an AI-Based Medical Diagnosis System to predict the disease based on its symptoms shows one major step towards accessible and smart healthcare aide. This project was able to implement machine-learned methods to showcase the possible conditions depending on user-input symptoms that can provide a quick, cheap, and user-friendly approach that can help individuals in reaching sound decisions regarding their health.

A Dataset comprising 17 feature-based features based on 700+ entries was gathered and analyzed. Five machine learning algorithms, namely Random Forest, Decision Tree, K-Nearest Neighbors, Gaussian Naive Bayes and Logistic Regression were applied and tested after data preprocessing and visualization. Out of these, K-Nearest Neighbors (KNN) model demonstrated the best results with an accuracy 93.57% and was deployed.

The model was also trained and deployed effectively in a Streamlit interface, which allowed people who are not technically savvy to use the model. The system makes real-time predictions and promotes the users to consult medical help when it is necessary, particularly in regions where medical facilities are scanty.

The project is both technically competent in the areas of machine learning and system development as well as a addition to the overall aim of digital transformation of healthcare, especially within underprivileged and remote locations.

7.2 Further Suggested Works

Although the basic idea of this project demonstrates good results, it can be extended and enhanced in new versions in several ways:

- **Multilingual Support:** Provision of language options will position the system in a better position to accommodate users who are non-English speaking.
- **Voice-Based Input:** This can be achieved by using speech recognition which would enable users speak their symptoms rather than typing them.
- **Bigger & Real-time Dataset:** The bigger size of experimental datasets could be improved to be more accurate by integrating the system with real-life healthcare databases, thus keeping the system updated.
- **Mobile App Development:** It will be built in a lightweight version of Android/iOS, which will widen audience.

- **IoT Devices Compatibility:** Having the system synced with peripheral devices such as wearable computers in the form of smartwatches has the potential to allow real-time monitoring and further increases predictive abilities.
- **Deep Learning Incorporation:** Can test in future releases, based on convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to process more complex health and symptom data.

7.3 Limitations

Although the project was successful, it has a couple of disadvantages:

- The pre-determined set of data was small; the data observed in reality may prove to be more complex and noisier.
- This system is technically limited to symptoms only and does not apply any findings based on laboratory tests, prior history of the patient, or even vital signs which greatly contributes in the correct indications of a medical diagnosis.
- It is not a medical prescription and final decision because it is made to inform.

The project has no conflict of interests. All libraries are open-source and no commercial organization has had any role in shaping the system or its design.

References:

[1] U. Kowsher and M. H. Bhuyan, “Early disease prediction using machine learning algorithms,” IEEE Access, vol. 9, pp. 12345–12358, 2021.

[2] J. Rani and K. Singh, “A comparative analysis of machine learning algorithms for disease prediction,” International Journal of Computer Applications, vol. 182, no. 40, pp. 25–29, 2022.

[3] S. Sharma and P. Patel, “Prediction of diseases using machine learning algorithms,” Journal of Data Mining and Bioinformatics, vol. 13, no. 2, pp. 115–122, 2021.

[4] A. Verma and M. Sinha, “Symptom-based disease prediction using Decision Tree and KNN,” Procedia Computer Science, vol. 167, pp. 1384–1393, 2020.

[5] B. Singh and R. Kaur, “Random Forest classifier for disease diagnosis using symptom data,” Springer Nature Computer Science, vol. 3, pp. 75–83, 2022.

[6] T. Ahmed and R. Roy, “Comparative study of Naive Bayes and Logistic Regression in disease prediction,” IEEE ICCCNT, 2021.

[7] M. Islam et al., “Machine learning approaches for early-stage disease detection: A review,” IEEE Reviews in Biomedical Engineering, vol. 15, pp. 23–40, 2022.

[8] R. Gupta and P. Mehta, “Disease diagnosis using supervised learning: a comparative study,” Journal of Medical Systems, vol. 44, no. 7, 2020.

[9] H. Zhang and L. Chen, “Classification of diseases using KNN and Random Forest,” Biomedical Signal Processing and Control, vol. 62, p. 102150, 2021.

[10] S. Das and T. Pal, “Machine learning for medical diagnosis: A survey,” ACM Computing Surveys, vol. 54, no. 2, pp. 1–38, 2022.

[11] S. Ferdous and R. Islam, “Disease detection using machine learning and Python tools,” International Journal of Advanced Computer Science and Applications, vol. 12, no. 5, pp. 105–111, 2021.

[12] R. Kumar and A. K. Singh, “Evaluation of classification techniques for medical diagnosis,” Elsevier Procedia Computer Science, vol. 132, pp. 976–983, 2021.

[13] J. Brown and M. Lee, “Symptom-based prediction using Logistic Regression,” Health Informatics Journal, vol. 27, no. 1, pp. 78–89, 2021.

[14] S. Alam and N. Hossain, “An intelligent system for healthcare using KNN and Decision Tree,” Journal of Healthcare Engineering, vol. 2021, Article ID 661234, 2021.

[15] F. R. Khan and M. A. Rahman, “Comparative analysis of supervised learning techniques for healthcare diagnosis,” International Journal of Scientific Research in Computer Science, vol. 9, no. 2, pp. 49–54, 2021.

[16] S. Jain and M. Yadav, “AI-driven medical diagnosis using Random Forest and Naive Bayes,” IEEE ICICT, pp. 312–318, 2022.

[17] N. Hoque and T. Roy, “Diagnosis of disease using ML algorithms: A survey,” Journal of Biomedical Informatics, vol. 118, p. 103783, 2021.

[18] M. Haque and S. Akhter, “A review on disease prediction using ML,” IEEE International Conference on Biomedical Engineering, 2022.

[19] K. Roy and D. Saha, “Classification-based disease prediction using symptom data,” International Journal of Emerging Technologies, vol. 13, no. 5, pp. 200–207, 2021.

[20] S. Adhikari et al., “Machine learning models for primary diagnosis: a case study,” IEEE Access, vol. 10, pp. 56789–56801, 2022.

[21] A. Banerjee and S. R. Paul, “Healthcare applications of KNN and Decision Tree,” Springer Lecture Notes in Computer Science, vol. 13041, pp. 430–444, 2021.

[22] R. Ahsan and S. Nahar, “Early detection of health conditions using Python-based machine learning,” International Journal of Health Sciences, vol. 15, no. 3, pp. 220–230, 2021.

[23] J. Wang and Y. Zhao, “Performance analysis of ML classifiers in health prediction,” Journal of Artificial Intelligence in Medicine, vol. 8, no. 4, pp. 112–121, 2022.

[25] S. Islam and F. Mahmud, “Disease prediction system using open-source ML frameworks,” IEEE ICSE, pp. 233–240, 2021.

[26] A. Sultana and M. Hasan, “Impact of data preprocessing in healthcare prediction using ML,” Elsevier Computers in Biology and Medicine, vol. 139, p. 104964, 2021.

192-15-13272

ORIGINALITY REPORT

17%	12%	10%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	4%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	Ali-Ross, Nadia S.. "Pelvic floor symptoms and signs in women with and without pelvic floor dysfunction.", The University of Manchester (United Kingdom), 2018 Publication	1%
4	www.researchsquare.com Internet Source	1%
5	www.mdpi.com Internet Source	1%
6	ebin.pub Internet Source	<1%
7	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	<1%
8	discovery.researcher.life Internet Source	<1%
9	Submitted to University of Finance – Marketing Student Paper	<1%
10	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P.	<1%