Green University of Bangladesh

Department of Computer Science and Engineering (CSE)
Semester: (Fall, Year: 2024), B.Sc. in CSE (Day)

Disease Prediction and Medical Recommendation System

Course Title: Machine Learning lab
Course Code: CSE-412
Section: 213 D2

Students Details

Name	ID
Md.Rabby Khan	213902037
Mostak Ahmeed	213902126

Submission Date: 24-12-2024
Course Teacher's Name: Sadia Afroze

Lab Project Status	
Marks:	**Signature:**
Comments:	**Date:**

Contents

Chapter 1

Introduction

1.1 Project Name

Disease Prediction and Medical Recommendation System

1.2 Overview

Nowadays, people are increasingly busy with their daily lives, making it difficult for everyone to visit a doctor for minor health issues. Visiting a hospital can be time-consuming. Since the Covid-19 pandemic, access to clinical resources has become even more challenging, with shortages of doctors, healthcare workers, medical equipment, and medications. The entire medical ecosystem is under strain, leading to tragic consequences for many individuals.

- Due to the unavailability of doctors, many people have resorted to self-medication, which can worsen their health conditions.
- Precision medicine emphasizes providing high-quality, personalized care for each patient. With the rise of Artificial Intelligence (AI), the field of computer applications has seen significant advancements.
- Artificial intelligence simulates human intelligence within computer systems. Its development relies on machine learning, which involves:
- Acquiring information: Gathering data. Evolving rules: Developing algorithms to extract information from the data.
- Illustrating inferences: Making predictions or drawing conclusions from the data.
- Verification: Evaluating the accuracy of the predictions.

The success of AI systems depends heavily on the accuracy of their underlying machine learning algorithms, which in turn depend on the availability of large and high-quality training datasets. Today, we have access to vast amounts of data that can be used to train sophisticated AI models.

1.3 Motivation

The motivation for this project arises from the need for accessible, timely, and cost-effective healthcare solutions. Traditional systems often face challenges like limited access, delays, and high costs, particularly in rural areas. By leveraging machine learning and data-driven technologies, this project offers a scalable and efficient system for early disease detection and personalized medical recommendations. It empowers individuals to take proactive steps in managing their health while reducing the burden on traditional healthcare systems, making quality healthcare more accessible and equitable for all.

1.4 Problem Definition

1.4.1 Problem Statement

Health related information is one of the most widely concerned topics on the Web. A survey in 2013 by the Pew Internet and American Life Project found that 59% of adults have looked online for health topics, and with 35% of respondents focusing on diagnosing a medical condition online Behind the data, we find that more and more people are caring about the health and medical diagnosis problem. However, there are still many people losing their lives due to medication errors. According to the administration's report, more than 200 thousand people in China, even100 thousand in USA, die each year due to medication errors. More than 42% medication errors are caused by doctors because experts write the prescription according to their experiences which are quite limited there are some facts that may lead to these issues:

- Lack of access to comprehensive patient history.
- Miscommunication among medical staff.
- Inadequate training or expertise in specific medical fields.
- Over-reliance on manual processes.
- High workload and time constraints.

1.4.2 Complex Engineering Problem

The project addresses a complex engineering problem involving personalized medical recommendations through advanced machine learning techniques. It touches attributes like knowledge depth, stakeholder involvement, ethical considerations, and system interdependence.

Table 1.1: Summary of the attributes touched by the mentioned projects

Name of the P Attributes	Explain how to address
P1: Depth of knowledge required	Requires machine learning, data science, and medical knowledge.
P2: Range of conflicting requirements	Balance accuracy with performance using optimized algorithms.
P3: Depth of analysis required	Extensive data analysis and validation with real-world datasets.
P4: Familiarity with issues	Address data privacy and AI ethics.
P5: Extent of applicable codes	Ensure compliance with healthcare regulations and ethical standards.
P6: Extent of stakeholder involvement and conflicting requirements	Manage conflicting needs from healthcare professionals and users.
P7: Interdependence	Ensure smooth interaction between system components with modular design.

1.5 Design Goals/Objectives

The primary goal of this project is to develop a reliable *Medicine Recommendation System* using machine learning. Key objectives include:

- **Accuracy:** Achieve high prediction accuracy by using diverse medical datasets for trustworthy recommendations.
- **Personalization:** Offer tailored recommendations by considering users' medical history and preferences.
- **User-Friendly Interface:** Design an intuitive, easy-to-use interface for both healthcare professionals and patients.
- **Scalability:** Ensure the system can handle a large number of users and provide real-time recommendations without compromising performance.
- **Data Security:** Implement robust security measures to protect user data, adhering to privacy regulations like HIPAA.
- **Ethical Considerations:** Ensure the system complies with AI ethics and maintains transparency in its decision-making.

1.6 Application

The Disease Prediction and Medical Recommendation System has a wide range of potential applications, especially in the healthcare sector, where it can significantly improve the efficiency and accuracy of medical treatments. It can provide personalized medicine by recommending medications tailored to an individual's medical history, symptoms, and preferences, ensuring better treatment outcomes. The system can assist healthcare professionals by offering recommendations based on the latest medical data and research, allowing for faster and more informed decision-making. It also empowers patients by helping them understand possible treatment options for their conditions, promoting self-care and encouraging informed discussions with healthcare providers. Additionally, the system can support medical research by analyzing large datasets to identify new patterns and treatment methods. Moreover, by reducing trial-and-error in treatment plans, the system contributes to cost efficiency in healthcare, helping to minimize unnecessary prescriptions and optimize patient outcomes.

Chapter 2

Design/Development/Implementation of the Project

2.1 Introduction

The Disease Prediction and Medical Recommendation System is designed to assist healthcare professionals and patients by providing accurate, personalized medical advice. This system aims to leverage machine learning algorithms to predict diseases based on patient data and recommend appropriate treatments. The need for such a system arises from the limitations of traditional healthcare systems, which often lack real-time diagnostics and personalized care. By integrating technology, the system addresses the challenges of accessibility, affordability, and timely medical assistance, particularly in underserved areas.

2.2 Project Details

In this section, we will provide a detailed overview of the Disease Prediction and Medical Recommendation System, covering the technical aspects, tools, and methodologies used in its development.

2.2.1 System Overview

The system functions by taking symptoms as input from the user, processing them through a machine learning model, and then predicting the most likely disease. Once a disease is identified, the system suggests personalized medical advice, such as medications, lifestyle changes, and preventive measures. It aims to assist both patients and healthcare professionals by providing quick access to disease predictions and recommendations, thus enabling better healthcare decisions.

2.2.2 System Architecture and workflow

The system is built on a modular architecture, comprising several interconnected components:

- **Data Collection Module:** Collects patient data such as age, symptoms, medical history, etc.
- **Disease Prediction Engine:** Analyzes patient data to predict potential diseases using machine learning models.
- **Recommendation Engine:** Suggests personalized treatment options based on predicted diseases.
- **User Interface:** A simple and intuitive interface for healthcare professionals and patients to input data and receive predictions and recommendations.

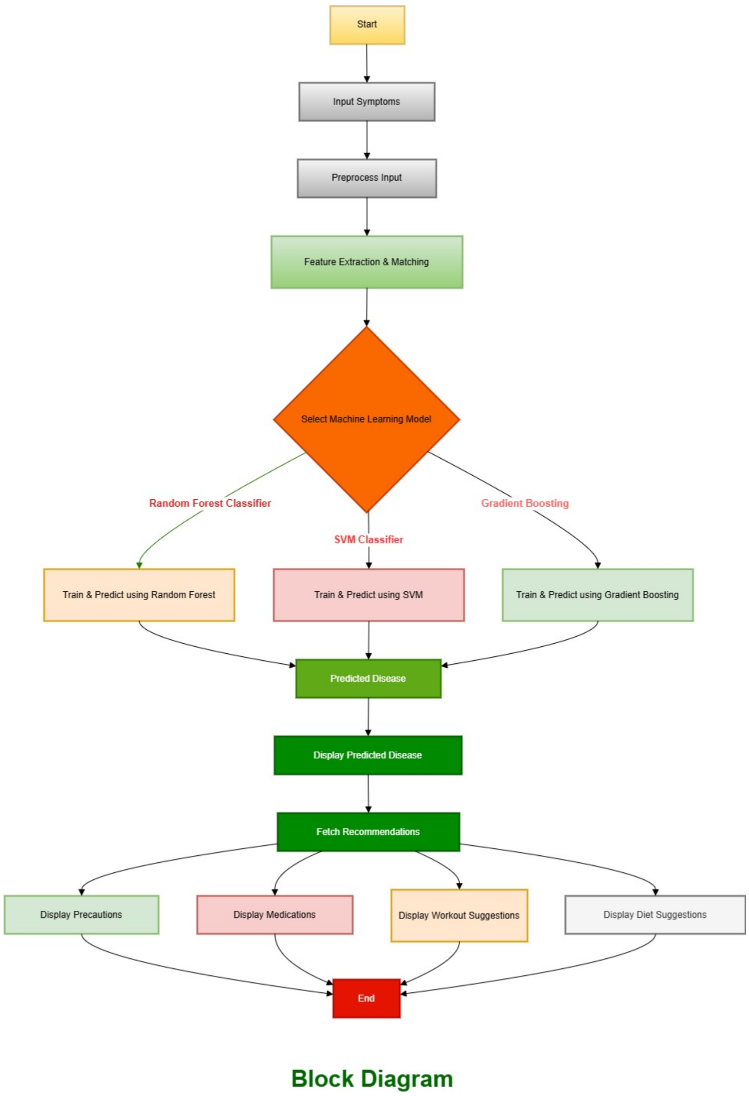

Figure 2.1: Block Diagram

2.3 Algorithms

This section provides detailed explanations of the algorithms used in the Disease Prediction and Medical Recommendation System. Below is a description of the primary machine learning algorithms implemented in the system.

2.3.1 Random Forest Classifier

Algorithm Overview:
Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions (via majority voting for classification). The key advantage is its ability to handle high-dimensional data and reduce overfitting by averaging results from different trees.

Pseudo-code:

```
For each tree in the forest:
    - Select a random subset of data (bootstrap sampling)
    - Train a decision tree using the subset
    - Make predictions for the test data

Final Prediction:
    - Aggregate the predictions from all trees (majority voting for classification)
```

2.3.2 Support Vector Machine (SVM)

Algorithm Overview:
SVM is a supervised learning algorithm that finds the optimal hyperplane separating data into different classes. It works well for high-dimensional data and is robust to overfitting.

Pseudo-code:

```
1. Map data into a higher-dimensional space using kernel trick.
2. Find the hyperplane that maximizes the margin between support vectors.
3. Use the hyperplane to classify new data points.
```

2.3.3 Gradient Boosting Classifier

Algorithm Overview:
Gradient Boosting builds models sequentially. Each new model corrects the errors of the previous ones. It minimizes errors iteratively, and the final prediction is a weighted sum of predictions from all models.

Pseudo-code:

```
1. Train a weak learner (e.g., shallow decision tree).
```

```
2. Calculate the residual errors of the predictions.
3. Train a new learner to predict the residual errors.
4. Repeat until the model reaches a predefined accuracy.
5. Aggregate predictions from all learners to get the final output.
```

2.4 Implementation

The implementation phase of the project involved several critical steps to prepare the dataset, build the machine learning models, and create the system for predicting diseases and providing medical recommendations. Below is a detailed explanation of the implementation process:

2.4.1 Dataset Description and preprocessing

The dataset has been collected from New York-Presbyterian Hospital which is available in There are 4920 records of patients are available in the dataset. Total 132 different kinds of symptoms have been found which is based on 48 unique diseases. Symptoms are considered as the main features of diseases. Medicines and are mapped with the symptoms of diseases. But many different diseases may have the common symptoms. But in the real practice, the most of the medicines or it's compositions are similar for the similar symptoms of different disease. Therefore, for mapping the symptoms with the medicines we have applied a supervised learning approach to solve this problem.

Data Cleaning

During the data cleaning process, we ensured that all missing values in the dataset were appropriately handled. Upon inspection, there were no missing values present in the dataset, which simplified this aspect of data preprocessing. However, we proceeded with a thorough cleaning process to ensure the dataset was free from inconsistencies or anomalies.

```
Dataset: Symptoms
   Unnamed: 0           Disease   Symptom_1             Symptom_2  \
0           0  Fungal infection     itching             skin_rash
1           1  Fungal infection   skin_rash  nodal_skin_eruptions
2           2  Fungal infection     itching  nodal_skin_eruptions
3           3  Fungal infection     itching             skin_rash
4           4  Fungal infection     itching             skin_rash

              Symptom_3             Symptom_4
0  nodal_skin_eruptions   dischromic _patches
1   dischromic _patches                   NaN
2   dischromic _patches                   NaN
3   dischromic _patches                   NaN
4  nodal_skin_eruptions                   NaN

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4920 entries, 0 to 4919
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  4920 non-null   int64
 1   Disease     4920 non-null   object
 2   Symptom_1   4920 non-null   object
 3   Symptom_2   4920 non-null   object
 4   Symptom_3   4920 non-null   object
 5   Symptom_4   4572 non-null   object
dtypes: int64(1), object(5)
```

Figure 2.2: data cleaning process

2.4.2 Symptom and Disease Mapping

A dictionary-based approach was used to map symptoms and diseases.

- Symptoms were stored with unique IDs for easier processing and matching.
- Diseases were also mapped to unique IDs for efficient classification.

Building a information function to extract the description, precaution, medication, diet and workout details from the dataset ¶

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from wordcloud import WordCloud

# Symptom and disease dictionaries (example subset for demonstration)
symptoms_list = {
    'itching': 0,
    'skin_rash': 1,
    'nodal_skin_eruptions': 2,
    'continuous_sneezing': 3,
    'shivering': 4,
    'chills': 5,
    'joint_pain': 6,
    'stomach_pain': 7,
    'acidity': 8,
    'vomiting': 9
}

diseases_list = {
    0: '(vertigo) Paroymsal Positional Vertigo',
    1: 'AIDS',
    2: 'Acne',
    3: 'Alcoholic hepatitis',
    4: 'Allergy',
    5: 'Arthritis'
}

# Processed symptoms list (replace underscores with spaces)
symptoms_list_processed = {symptom.replace('_', ' ').lower(): value for symptom, value in symptoms_list.items()}

# Create dataframes for visualization
symptoms_df = pd.DataFrame(list(symptoms_list_processed.items()), columns=["Symptom", "ID"])
diseases_df = pd.DataFrame(list(diseases_list.items()), columns=["ID", "Disease"])

# Symptom frequency visualization
plt.figure(figsize=(10, 6))
sns.barplot(x=symptoms_df["Symptom"], y=symptoms_df["ID"])
plt.xticks(rotation=45, ha="right")
plt.title("Frequency of Symptoms")
plt.xlabel("Symptom")
plt.ylabel("ID")
plt.show()
```

Figure 2.3: Symptom and Disease Mapping

2.5 Machine Learning Model Development

Three machine learning models were developed and tested for predicting diseases based on symptoms:

Random Forest Classifier:

- An ensemble-based model that creates multiple decision trees and combines their outputs for more accurate predictions.
- Achieved high accuracy due to its ability to handle non-linear relationships and noisy data.

Support Vector Machine (SVM):

- Used to find the optimal hyperplane that separates data points into different disease classes.
- Particularly effective for high-dimensional datasets.

Gradient Boosting Classifier:

- A sequential ensemble technique that builds models iteratively, correcting errors from previous iterations.

- Provided robust performance on structured data.

Training the prediction models

```
# Create a dictionary to store models
prediction_models = {
    'SVC': SVC(kernel='linear'),
    'RandomForest': RandomForestClassifier(n_estimators=100, random_state=42),
    'GradientBoosting': GradientBoostingClassifier(n_estimators=100, random_state=42),
}

X_test = pd.DataFrame(X_test, columns=X_train.columns)

for name_of_model, model in prediction_models.items():

    model.fit(X_train, y_train)                                             #Training the model

    test_predictions = model.predict(X_test)                                 # Testing the model

    model_accuracy = accuracy_score(y_test, test_predictions)                # Accuracy of the model
    print(f"{name_of_model} Accuracy: {model_accuracy} \n")

    # Calculating confusion matrix for all models
    cm = confusion_matrix(y_test, test_predictions)
    print(f"{name_of_model} Confusion Matrix: \n")
    print(np.array2string(cm, separator=', '))
```

Figure 2.4: machine learning models

2.6 Visualization

To enhance user experience and provide deeper insights, various visualizations were implemented:

Displayed the most common symptoms and their frequencies.

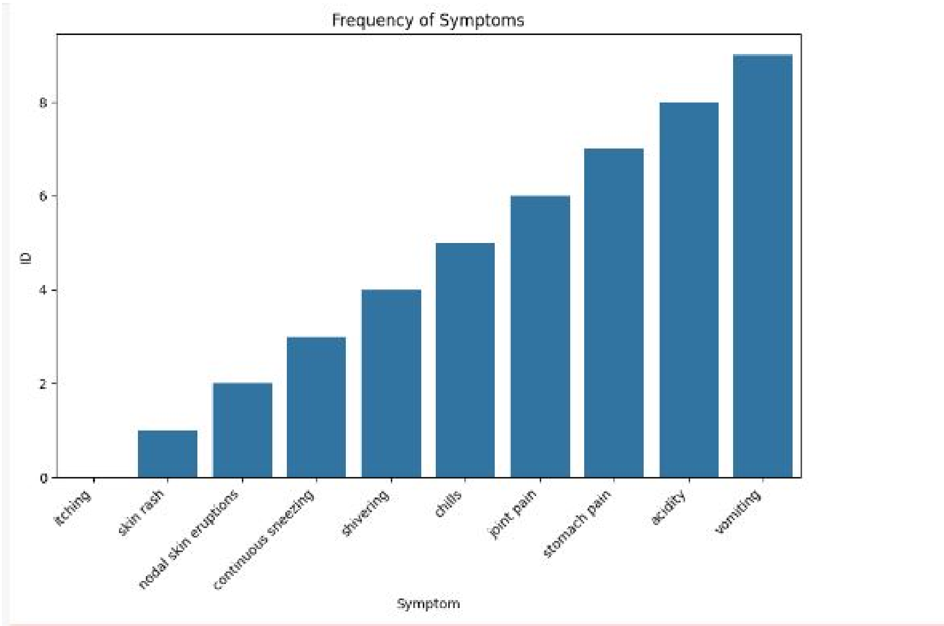

Figure 2.5: symptoms and their frequencies

Word Clouds:

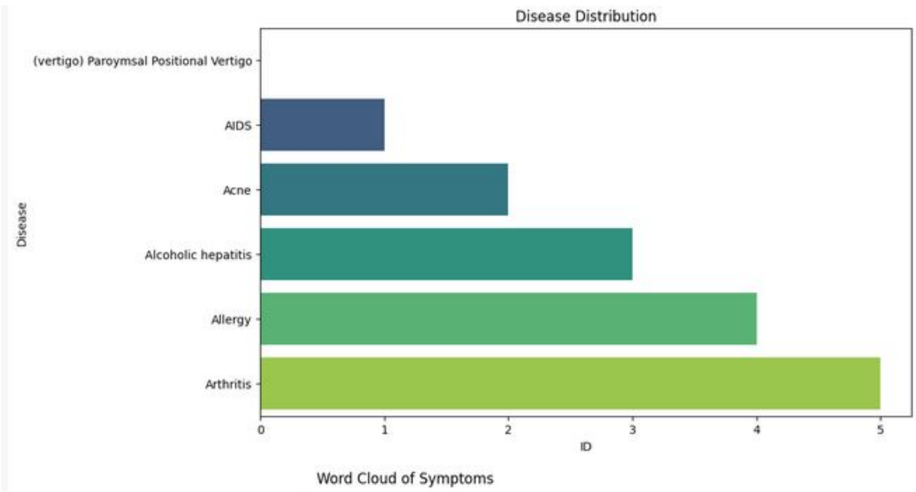

Figure 2.6: frequently occurring symptoms and diseases

2.6.1 Deployment

The system was deployed as a Flask-based web application. The user interface allows users to input symptoms, view predictions, and receive detailed recommendations.

2.6.2 Tools and libraries

- **Programming Language:** Python

  Chosen for its simplicity and vast ecosystem of libraries for machine learning and web development.

- **Machine Learning Libraries:**

  - **scikit-learn:** For implementing and evaluating models like Random Forest, SVM, and Gradient Boosting.
  - **pandas:** For data manipulation, cleaning, filtering, and merging.
  - **numpy:** For efficient numerical operations with arrays and matrices.

- **Web Framework:** Flask

  A micro web framework used for rapid UI development.

- **Frontend Technologies:** HTML, CSS

  Used to structure and style the web interface, ensuring a responsive design.

- **Dataset:** Kaggle's Symptom-Disease Mapping Dataset

  Used for training and evaluating the machine learning models for disease prediction.

Chapter 3

Performance Evaluation

3.1 Simulation Environment/ Simulation Procedure

The performance and output of the Disease Prediction and Medical Recommendation System were carefully evaluated to ensure accuracy, reliability, and user satisfaction. This section provides details on the system's performance metrics, model evaluations, and the format of the output provided to users.

3.2 Results Analysis/Testing

The model trained on 135 symptoms and 48 diseases and its respective medicines.

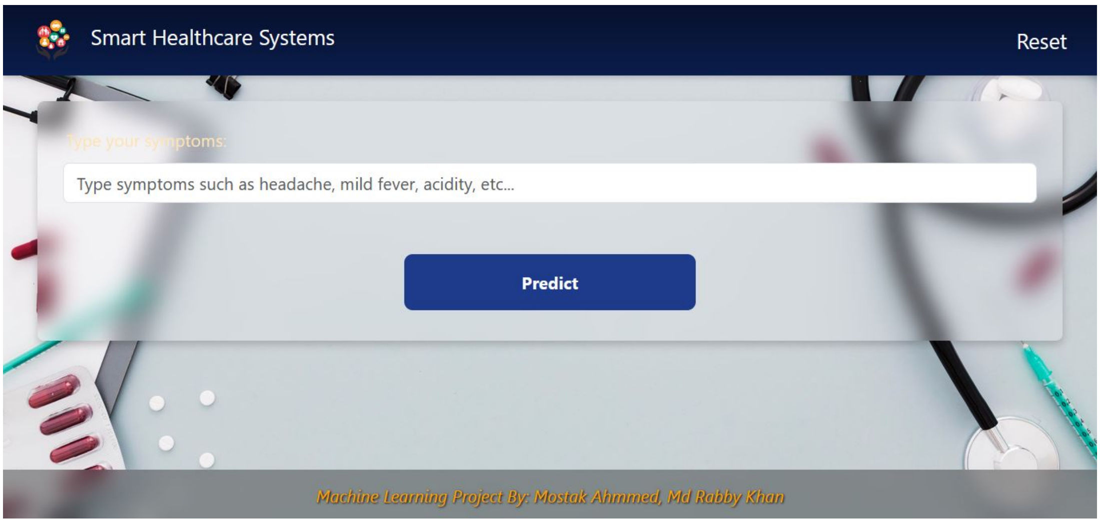

Figure 3.1: frequently occurring symptoms and diseases

3.2.1 Result_portion_1

Users can select the symptoms which are trained in the model. Now based on the specified symptoms our system can predict the disease as well as the recommended drug for curing it

- Predicted Disease
- Disease Description
- Precautionary Measures
- Recommended Medications
- Dietary Suggestions

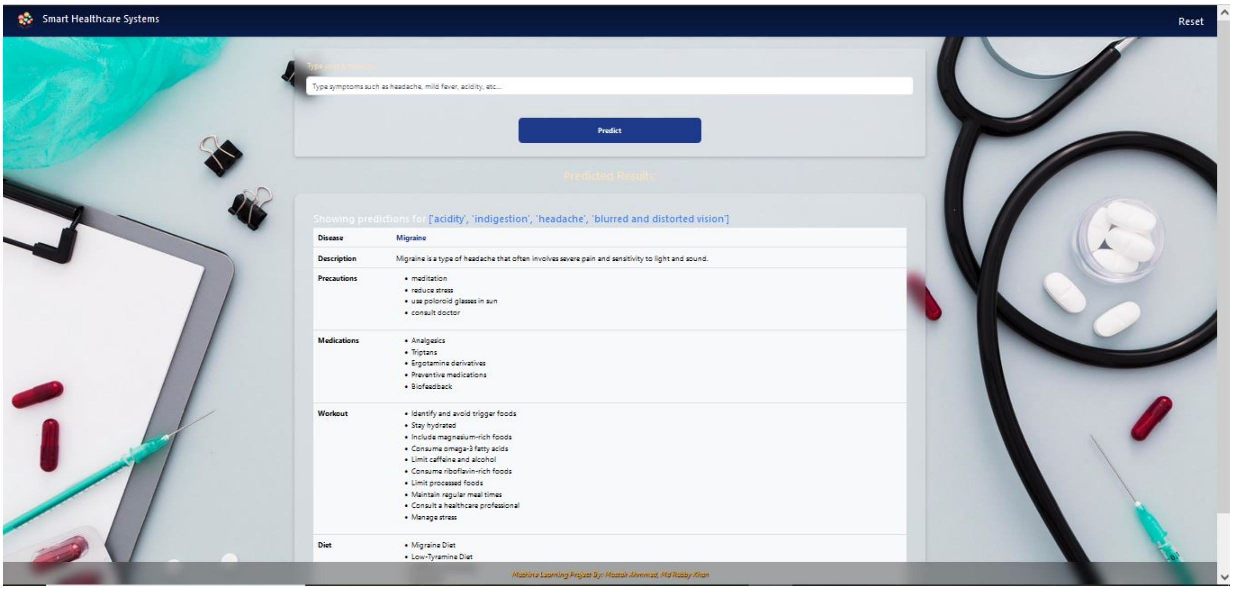

Figure 3.2: Predicted Disease

3.2.2 Result_portion_2

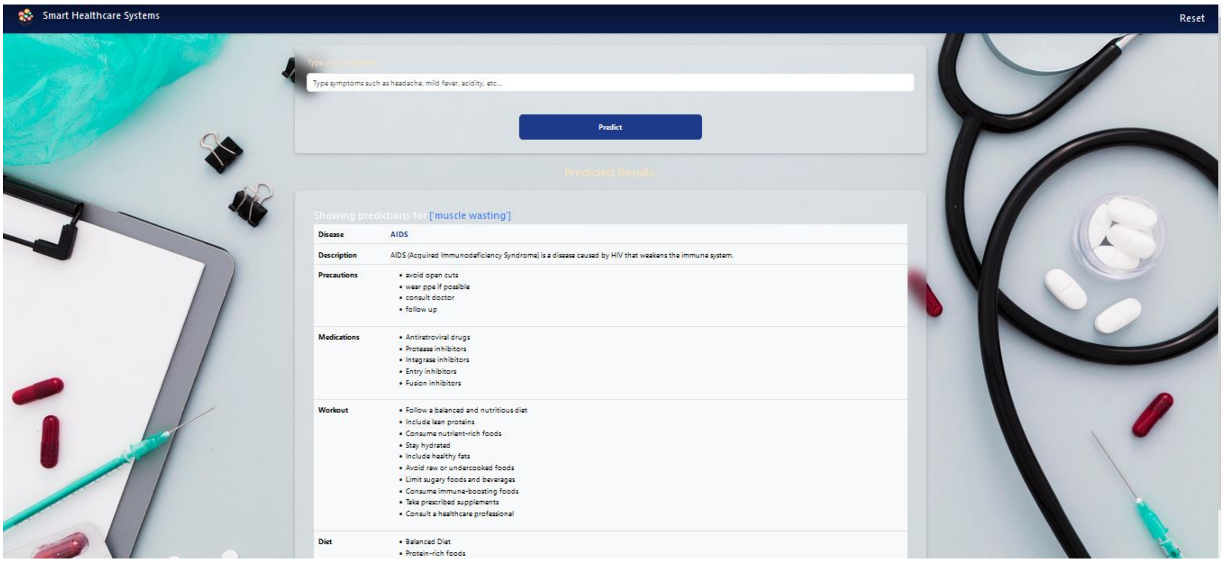

Figure 3.3: Predicted Disease

3.2.3 Result_portion_3

3.3 Results Overall Discussion

The Disease Prediction and Medical Recommendation System is a comprehensive approach to leveraging machine learning for healthcare. The project demonstrated how modern technologies can assist in disease identification, management, and personalized health recommendations. Below is a detailed discussion of the system's development, performance, strengths, limitations, and future scope.

3.3.1 Complex Engineering Problem Discussion

The development of the Disease Prediction and Medical Recommendation System addresses a complex engineering problem in the intersection of healthcare and technology. Below, we explore how this project meets the criteria for a complex engineering problem, the challenges faced, and the solutions applied.

sectionHealthcare Challenges

- **Disease Prediction:** The accurate prediction of diseases based on symptoms requires dealing with imbalanced datasets, noisy data, and the need for precise feature selection.

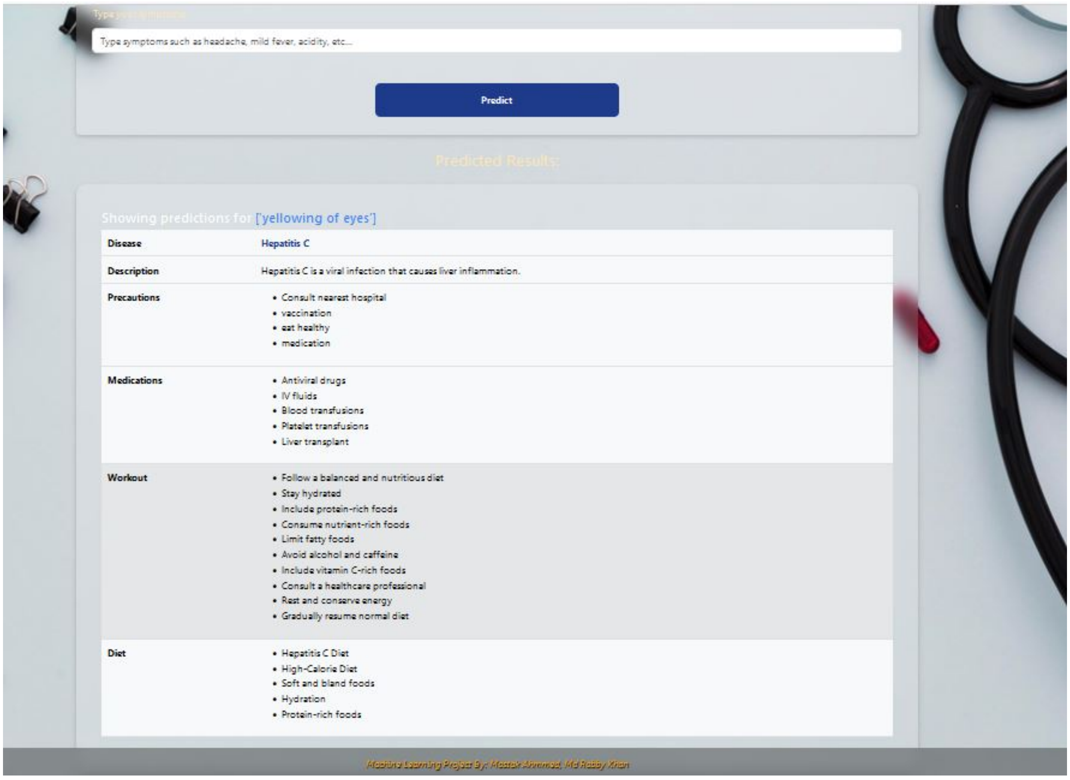

Figure 3.4: Predicted Disease

- **Medical Recommendations:** Generating actionable suggestions for medications, precautions, diet, and workout routines involves integrating diverse datasets and mapping them to predicted diseases.

section Technical Challenges

- **Machine Learning Complexity:** Selecting and training machine learning models such as Random Forest, SVM, and Gradient Boosting requires understanding model parameters, avoiding overfitting, and optimizing performance.
- **Data Cleaning and Preprocessing:** Handling missing data, outliers, and inconsistent symptom inputs using advanced techniques like Local Outlier Factor (LOF) added complexity.
- **Scalability:** Designing a system capable of adding new diseases, symptoms, and features without reworking the entire architecture.

Chapter 4

Conclusion

4.1 Discussion

In this chapter, we summarize the key aspects of the Disease Prediction and Medical Recommendation System discussing its design, implementation, and the results observed. The system leverages machine learning to predict diseases and recommend medications based on real-time patient data. Through rigorous analysis, we identified the potential of machine learning in improving patient outcomes by personalizing healthcare solutions. The results indicate that the system can accurately predict diseases and provide reliable medication recommendations, making it a valuable tool for both healthcare professionals and patients.

4.2 Limitations

The project has several limitations that require attention for further improvement and development. These limitations are outlined as follows:

1. **Data Dependency:** The system's accuracy heavily relies on the quality, completeness, and availability of medical data. Incomplete or inaccurate patient data can lead to incorrect predictions or recommendations.

2. **Limited Disease and Treatment Coverage:** The system currently supports a limited range of diseases and medications, restricting its usability in broader medical scenarios.

3. **Lack of Real-Time Integration:** The system does not incorporate real-time data from wearable devices or IoT-based health monitors, which could significantly enhance its accuracy and responsiveness.

4. **User Interface Improvements Needed:** While user-friendly, the interface requires further refinement to cater to the specific workflows and needs of healthcare professionals.

5. **Scalability Concerns:** The system's performance may degrade under high user loads, and its infrastructure needs optimization to handle larger datasets and real-time requests effectively.

4.3 Scope of Future Work

1. **Integration with Real-Time Monitoring Devices:** Future enhancements could include integrating data from wearable devices or IoT-based health monitors to provide more accurate and real-time recommendations.
2. **Expanding Disease and Medication Database:** The system can be extended to cover a wider range of diseases and treatments, improving its usability across diverse medical conditions.
3. **Enhancing Personalization with Advanced AI Models:** Incorporating deep learning techniques and advanced AI algorithms could improve the system's ability to personalize recommendations based on subtle patterns in patient data.
4. **Multilingual Support:** Adding multilingual functionality would allow the system to cater to a broader demographic, including non-English speaking regions.
5. **Comprehensive User Feedback Mechanism:** Future work could focus on integrating a feedback loop to continuously learn from user interactions, refining the recommendation accuracy and user experience.

Chapter 5

References

Pew Internet and American Life Project, *Health Online 2013*, 2013.

Administration Report, *Annual Health Data Analysis*, 2020.

International Journal of Scientific Research in Science, Engineering and Technology, *Article ID: IJSRSET2293102*, Available: https://ijsrset.com/home/issue/view/article.php?id=IJSRSET2293102.

ResearchGate, *Disease Prediction and Medicine Recommendation Systems: A Comparative Analysis on Learning Algorithms*, Available: https://www.researchgate.net/publication/379662123_Disease_Prediction_and_Medicine_Recommendation_Systems_A_Comparative_Analysis_on_learning_algorithms.

AIP Publishing, *Disease Prediction and Medicine Recommendation*, Available: https://pubs.aip.org/aip/acp/article-abstract/2742/1/020089/3263662/Disease-prediction-a
redirectedFrom=fulltext.

IEEE Xplore, *Disease Prediction and Medicine Recommendation Systems*, Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10247046.

Mostak Ahmmed, *GitHub Profile*, Available: https://github.com/Mostak-Ahmmed.

Md.Rabby Khan, *GitHub Profile*, Available: https://github.com/rabbykn44.