DISEASE PREDICTION SYSTEM USING MACHINE LEARNING ALGORITHMS

A PROJECT REPORT

Submitted by

JEFFREY HAMLIN V (2116210701094)
HARINI V (2116210701071)

in partial fulfillment for the award of the

degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI-602105

BONAFIDE CERTIFICATE

Certified that this Thesis titled **“DISEASE PREDICTION SYSTEM USING MACHINE LEARNING ALGORITHMS**” is the bonafide work of “**JEFFREY HAMLIN V (2116210701094), HARINI V (2116210701071)”** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr . K.Anand M.E.,Ph.D.,

PROJECT COORDINATOR

Professor

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Chennai - 602 105

Submitted to Project Viva-Voce Examination held on_________________

Internal Examiner **External Examiner**

ABSTRACT

To create a reliable predictive model for early disease identification, "Disease Prediction Using Machine Learning" was brought out. This research aims to reliably forecast the possibility of a disease by analyzing medical data, such as patient demographics, symptoms, and diagnostic test results, using machine learning algorithms. Collecting data, preprocessing, choosing features, training the model, and evaluating it are important tasks. For real-time prediction, the top-performing model will be included into an intuitive user interface. The model will be updated and improved continuously to guarantee its responsiveness to changing healthcare requirements. In the end, this project seeks to transform the diagnosis and treatment of diseases by providing a proactive strategy for tailored healthcare and better patient outcomes. Improvements in patient care and healthcare analytics will result from collaboration between data scientists and medical experts.

ACKNOWLEDGMENT

First, we thank the almighty god for the successful completion of the project. Our sincere thanks to our chairman **Mr. S. Meganathan B.E., F.I.E.,** for his sincere endeavor in educating us in his premier institution. We would like to express our deep gratitude to our beloved Chairperson **Dr. Thangam Meganathan Ph.D.,** for her enthusiastic motivation which inspired us a lot in completing this project and Vice Chairman **Mr. Abhay Shankar Meganathan B.E., M.S.,** for providing us with the requisite infrastructure.

We also express our sincere gratitude to our college Principal,

Dr. S. N. Murugesan M.E., PhD., and **Dr. P. KUMAR M.E., PhD, Director computing and information science , and Head Of Department of Computer Science and Engineering** and our project coordinator **Dr. K.Ananthajothi M.E.,Ph.D.,** for her encouragement and guiding us throughout the project towards successful completion of this project and to our parents, friends, all faculty members and supporting staffs for their direct and indirect involvement in successful completion of the project for their encouragement and support.

JEFFREY HAMLIN V

HARINI V

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The emergence of machine learning (ML) has resulted in significant progress in numerous sectors, including healthcare. A ground-breaking project called "Disease Prediction Using Machine Learning" has the potential to completely transform the medical diagnostics industry. With healthcare data becoming more and more accessible, creative ways to make the most of this abundance of knowledge are desperately needed. The main objective of this research is to forecast the possibility of various diseases by analyzing a variety of datasets that include patient records, clinical signs, and diagnostic results using machine learning (ML) algorithms.

By facilitating prompt intervention and proactive management techniques, early illness identification has enormous promise for enhancing patient outcomes and lowering healthcare expenses. Conventional diagnostic techniques sometimes depend on subjective evaluations and may overlook minute patterns or clues that might indicate the beginning of a disease. On the other hand, machine learning techniques provide a data-driven method that may reveal hidden correlations and patterns in intricate medical data, improving the precision and effectiveness of illness prediction.

This project's scope includes a number of crucial elements, such as feature selection, data collecting, preprocessing, model construction, and assessment. The project's goal is to produce strong prediction models that can precisely detect illness risks by extracting pertinent insights through rigorous data preprocessing and feature engineering. In order to identify the best method for illness prediction, a variety of machine learning methods will be investigated and assessed, including decision trees, random forests, support vector machines, and neural networks.

Moreover, the incorporation of the established models into intuitive interfaces would enable smooth communication with medical specialists, allowing for the prediction of diseases in real time and the making of well-informed decisions. Predictive models will be updated and refined continuously to guarantee their flexibility in response to changing patient demographics and healthcare trends, enhancing their usefulness in clinical practice.

1.1 PROBLEM STATEMENT

Accurate illness prediction using machine learning is the project's goal. Managing complicated healthcare data, picking pertinent characteristics, guaranteeing model generalization, and producing forecasts that are understandable are among the difficulties. The objective is to use ML-driven predictive analytics in clinical practice to enhance patient outcomes and healthcare quality.

1.2 SCOPE OF THE WORK

The project includes gathering data, preparing it, and choosing features. Next, machine learning models for illness prediction are developed and assessed. It entails investigating different algorithms, integrating models into intuitive user interfaces, and iteratively improving them to guarantee flexibility in response to changing healthcare requirements.

1.3AIM AND OBJECTIVES OF THE PROJECT

With the use of machine learning algorithms, the "Disease Prediction Using Machine Learning" project seeks to create precise prediction models that estimate the likelihood of diseases based on extensive medical data. Obtaining a variety of datasets, choosing pertinent features, and training decision trees, random forests, and neural networks are among the goals. Model performance will be validated by evaluation criteria including accuracy and precision, guaranteeing generalization and resilience.

Healthcare practitioners will be able to access real-time forecasts with ease thanks to integration into an intuitive interface, and model flexibility will be maintained through ongoing improvement based on input and advances in machine learning techniques. We will examine ethical issues related to model interpretability and data privacy, guaranteeing responsible implementation in clinical contexts. The project's ultimate goal is to enhance patient outcomes by advancing healthcare analytics and enabling early illness identification and individualized treatment plans.

1.4 RESOURCES

This project has been developed through widespread secondary research of accredited manuscripts, standard papers, business journals, white papers, analysts' information, and conference reviews. Significant resources are required to achieve an efficacious completion of this project.

The following prospectus details a list of resources that will play a primary role in the successful execution of our project:

- A properly functioning workstation (PC, laptop, net-books etc.) to carry out desired research and collect relevant content.
- Unlimited internet access.
- Unrestricted access to the university lab in order to gather a variety of literature including academic resources (for e.g. Prolog tutorials, online programming examples, bulletins, publications, e-books, journals etc.), technical manuscripts, etc.

1.5 MOTIVATION

In the current healthcare environment, prompt diagnosis is essential to both bettering patient outcomes and cutting costs. This initiative intends to transform illness diagnosis and management by using machine learning algorithms to predict disease likelihood based on extensive medical data. Early identification allows medical practitioners to act quickly and provide individualized treatment programs based on the requirements of each patient. Predictive modeling also provides physicians with insightful information that helps them make better decisions and use resources more effectively. The project's ultimate goal is to close significant gaps in the state of healthcare today, opening the door to a day when predictive analytics will be essential to the delivery of precision medicine and the improvement of patient care.

CHAPTER 2

LITERATURE SURVEY

[1] explores the utilization of vast clinical data in medical decision support, focusing on cardiac disease prediction, particularly coronary heart disease. It reviews various machine learning algorithms employed for disease classification and risk prediction, emphasizing their characteristics and disparities. While machine learning techniques demonstrate broad applicability in cardiac diseases, limitations arise due to the non-uniformity of medical data, constraining each method to specific contexts. The article concludes with a summary of findings regarding heart disease prediction, highlighting the challenges and future directions in leveraging machine learning for improved medical decision-making.

[2] addresses the critical challenge of cardiovascular disease prediction using machine learning (ML) techniques. With heart disease ranking among the leading causes of global mortality, accurate prediction is imperative. ML has shown promise in analyzing vast clinical datasets, and recent developments in IoT have expanded its applications. However, existing studies provide only limited insights into heart disease prediction. The paper proposes a novel approach to identifying significant features using ML, enhancing prediction accuracy. Through the hybrid random forest with a linear model (HRFLM), the proposed model achieves an impressive accuracy level of 88.7%, underscoring its potential in improving cardiovascular disease prediction.

[3] introduces an advanced methodology for early prediction of chronic diseases, such as heart attack, diabetes, breast cancer, and kidney disease, by combining cutting-edge techniques. Beginning with Feature Engineering using Recursive Feature Elimination (RFE) and Support Vector Machine (SVM), irrelevant features are eliminated to simplify data complexity. The refined dataset is then fed into the robust eXtreme Gradient Boosting (XGBoost) classifier, known for its efficiency in predicting complex relationships. Hyperparameter tuning using Bayesian optimization further optimizes model performance. The proposed approach significantly enhances early prediction of chronic diseases, demonstrating the efficacy of the methodology in improving predictive accuracy.

[4] addresses cardiovascular concerns by proposing the TLV (Two-Layer Voting) model for disease prediction. Utilizing two datasets, including Kaggle's heart disease dataset and UCI's heart disease dataset, the study introduces new features like Pulse Pressure (PP), Body Mass Index (BMI), and Mean Arterial Pressure (MAP) to enhance results. The TLV model employs ensemble methods of hard and soft voting, combining statistical methods in layer 1 for feature selection and classification algorithms in layer 2. Hyperparameter tuning using GridSearchCV further optimizes performance. Results show that the TLV methodology with soft voting achieves 99.03% accuracy with UCI's dataset and 88.09% with Kaggle's CVD dataset, surpassing existing CAD disease prediction studies in accuracy and performance.

[5] proposes DWRF, a novel metabolite-disease association prediction algorithm, to streamline the identification of disease pathogenesis. DWRF integrates semantic and information entropy similarities of diseases and metabolites, employing DeepWalk for metabolite feature extraction based on metabolite-gene association networks. Subsequently, a random forest algorithm infers metabolite-disease associations. Experimental results demonstrate DWRF's efficacy through superior performance metrics such as area under the curve, leave-one-out cross-validation, and five-fold cross-validation. Case studies further affirm DWRF's reliability in predicting metabolite-disease associations, offering a promising approach to expedite disease diagnosis and treatment.

CHAPTER 3

SYSTEM DESIGN

3.1 GENERAL

In this section, we would like to show how the general outline of how all the components end up working when organized and arranged together. It is further represented in the form of a flow chart below.

3.2 SYSTEM ARCHITECTURE DIAGRAM

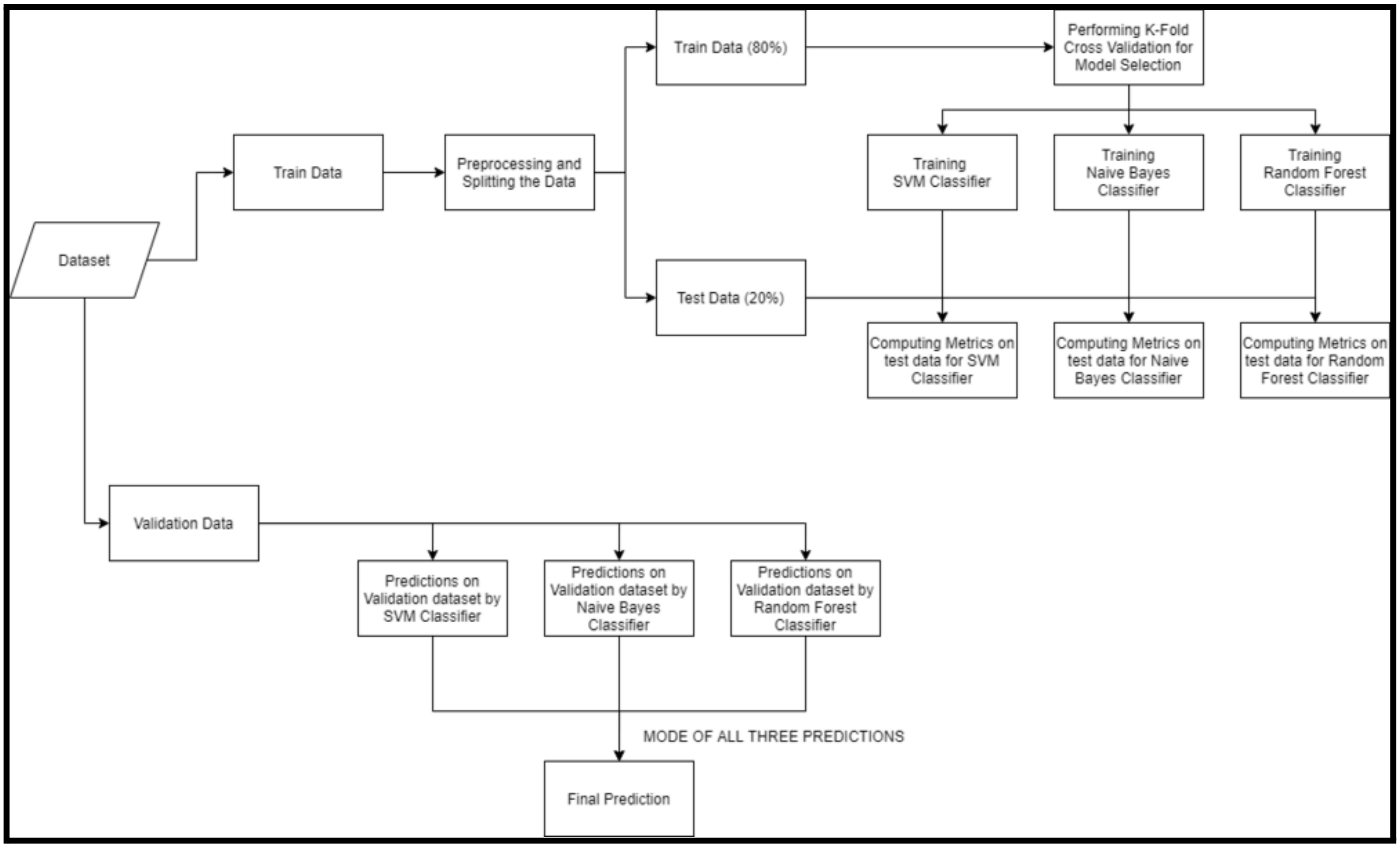

Dataset
Train Data
Preprocessing and Splitting the Data
Train Data (80%)
Test Data (20%)
Performing K-Fold Cross Validation for Model Selection
Training SVM Classifier
Training Naive Bayes Classifier
Training Random Forest Classifier
Computing Metrics on test data for SVM Classifier
Computing Metrics on test data for Naive Bayes Classifier
Computing Metrics on test data for Random Forest Classifier
Validation Data
Predictions on Validation dataset by SVM Classifier
Predictions on Validation dataset by Naive Bayes Classifier
Predictions on Validation dataset by Random Forest Classifier
MODE OF ALL THREE PREDICTIONS
Final Prediction

Fig 3.1: System Architecture

3.3 DEVELOPMENTAL ENVIRONMENT

3.3.1 HARDWARE REQUIREMENTS

The hardware requirements may serve as the basis for a contract for the system's implementation. It should therefore be a complete and consistent specification of the entire system. It is generally used by software engineers as the starting point for the system design.

Table 3.1 Hardware Requirements

COMPONENTS	SPECIFICATION
PROCESSOR	Intel Core i5
RAM	8 GB RAM
GPU	NVIDIA GeForce GTX 1650
MONITOR	15" COLOR
HARD DISK	512 GB
PROCESSOR SPEED	MINIMUM 1.1 GHz

3.3.2 SOFTWARE REQUIREMENTS

The software requirements document is the specifications of the system. It should include both a definition and a specification of requirements. It is a set of what the system should rather be doing than focus on how it should be done. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating the cost, planning team activities, performing tasks, tracking the team, and tracking the team's progress throughout the development activity.

Python IDLE, and **Chrome** would all be required.

3.4 SEQUENCE DIAGRAM

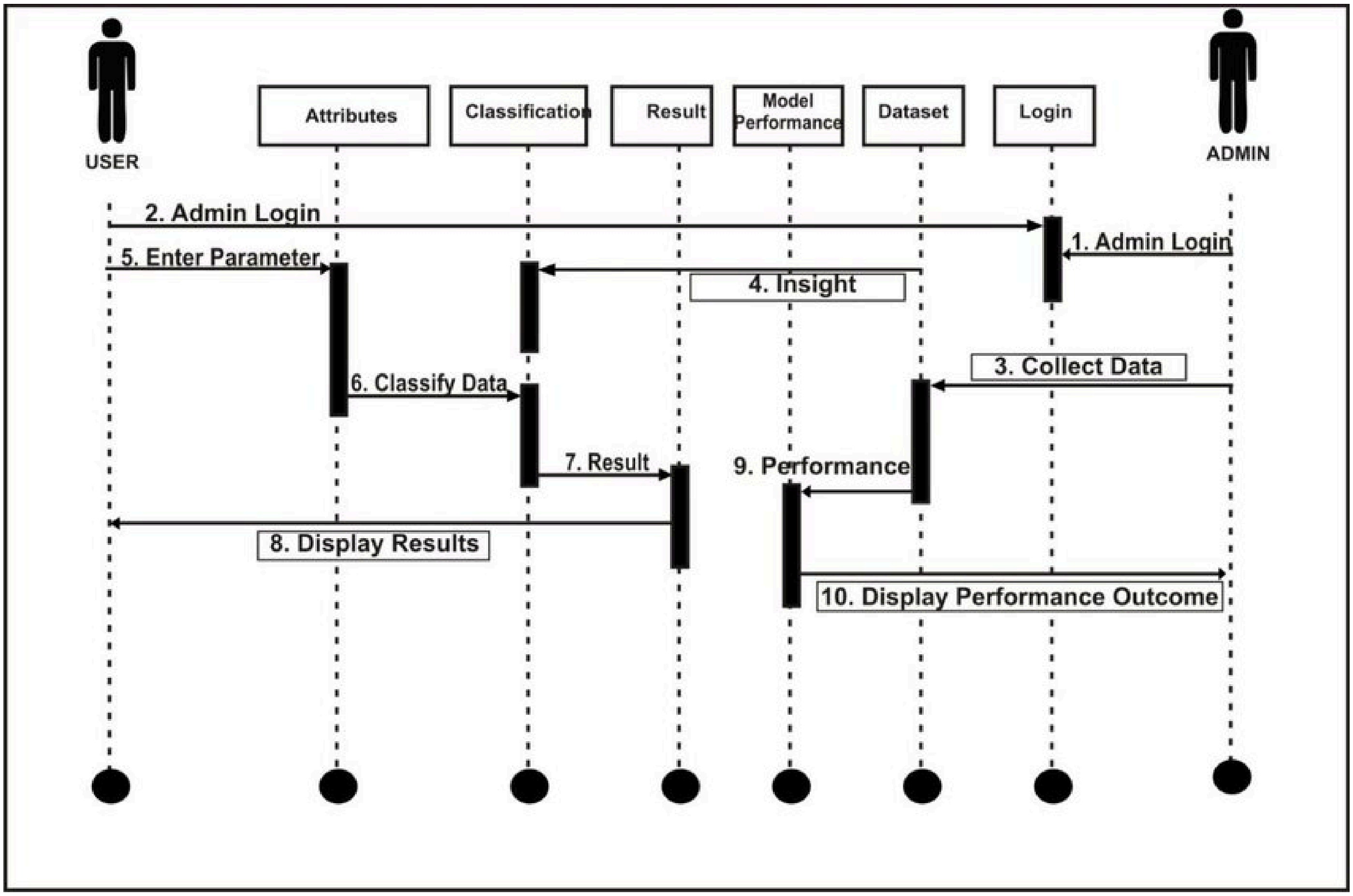


Fig 3.4: Sequence Diagram

CHAPTER 4

PROJECT DESCRIPTION

4.1 METHODOLOGY

We use a systematic approach in our machine learning (ML) methodology for illness prediction in order to efficiently use ML algorithms for precise prediction. In order to manage missing values and outliers, extensive medical datasets that include patient records, clinical symptoms, and diagnostic test results are first gathered and preprocessed. Then, important traits that greatly aid in illness prediction are found using feature selection approaches. Using the preprocessed data, a variety of machine learning techniques, including decision trees, random forests, support vector machines, and neural networks, are trained and assessed. Metrics including accuracy, precision, recall, and F1-score are used to evaluate the performance of the model.

The performance of the chosen models is optimized by hyperparameter adjustment. To increase prediction accuracy even further, ensemble approaches may be investigated. Ultimately, the top-performing model is included into an intuitive interface for real-time illness forecasts, guaranteeing healthcare practitioners' use and accessibility. The objective of this technique is to offer a strong and effective framework for illness prediction, enabling early diagnosis and tailored therapies to enhance patient outcomes.

4.2 MODULE DESCRIPTION

Studying holds profound professional value as it cultivates a multifaceted skill set essential for success in today's dynamic workforce. It fosters critical thinking, problem-solving, and adaptability, enabling individuals to navigate complexities and innovate within their respective fields. Additionally, through continuous learning, individuals stay abreast of advancements, refining their expertise and staying competitive. Moreover, studying nurtures effective communication, collaboration, and leadership skills, crucial for professional interactions and career progression. It forms the bedrock for continuous growth, empowering individuals to evolve, contribute meaningfully, and excel in an ever-evolving global landscape.

4.2.1 DISEASE MODULE

The disease module contains the predicted diseases. After the prediction process is conducted by the machine learning algorithm, it determines the disease based on the provided symptoms and presents the result.

4.2.2 DESCRIPTION MODULE

The description module contains the predicted diseases. After the prediction process is conducted by the machine learning algorithm, it determines the disease based on the provided symptoms and presents the description of that particular disease.

4.2.3 PRECAUTION MODULE

The precaution module contains the predicted diseases. After the prediction process is conducted by the machine learning algorithm, it determines the disease based on the provided symptoms and presents the precautions of that particular disease.

4.2.4 WORKOUT MODULE

The workout module contains the predicted diseases. After the prediction process is conducted by the machine learning algorithm, it determines the disease based on the provided symptoms and presents the workouts to prevent that particular disease.

4.2.5 DIET MODULE

The diet module contains the predicted diseases. After the prediction process is conducted by the machine learning algorithm, it determines the disease based on the provided symptoms and presents the diet that is required to be followed to avoid that particular disease.

CHAPTER 5
RESULTS AND DISCUSSIONS

5.1 OUTPUT

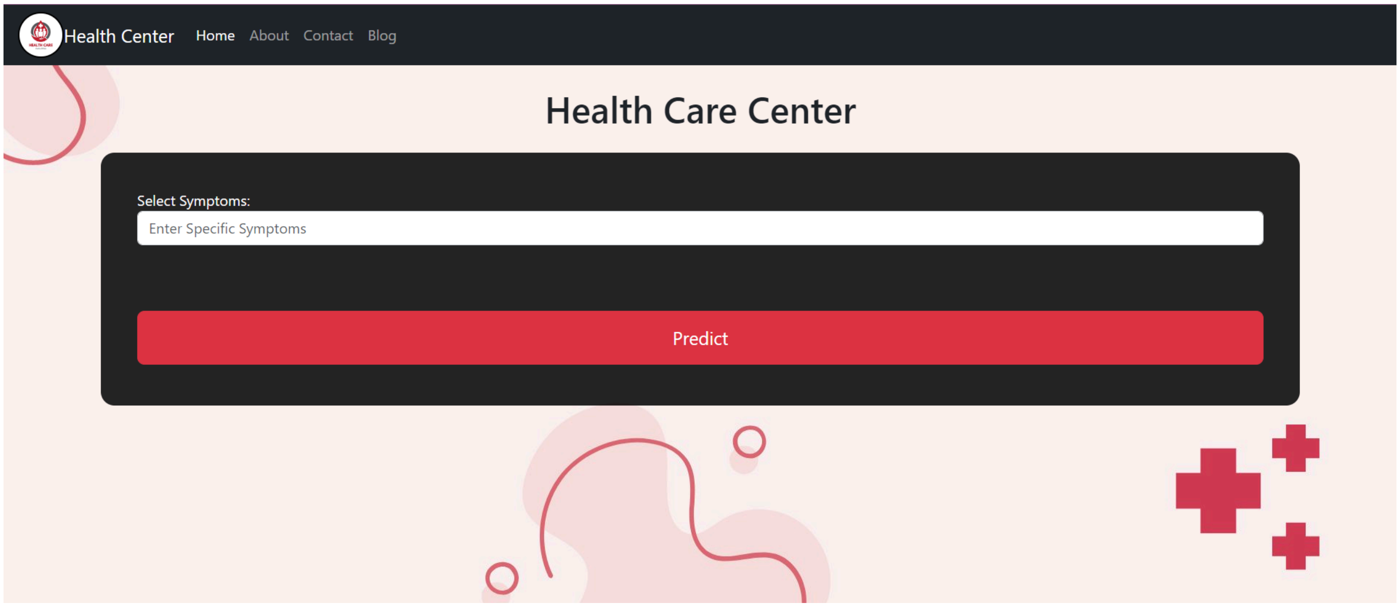


Fig 5.1 Home

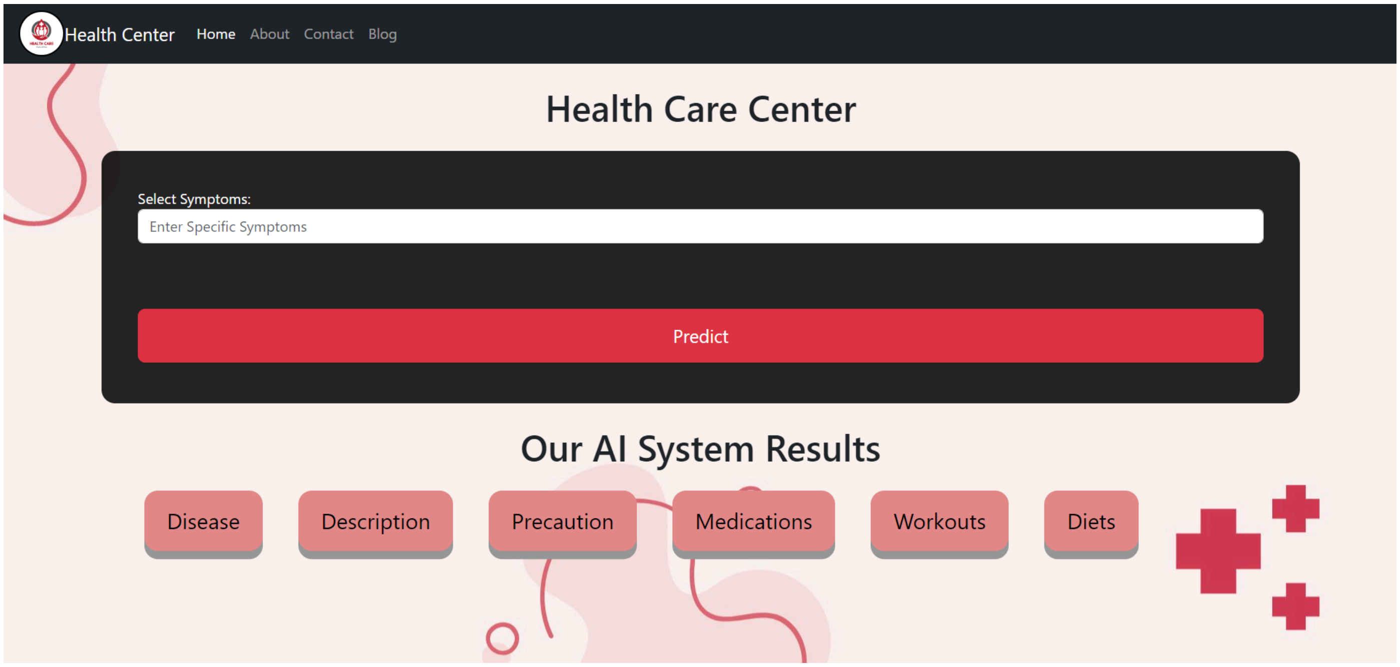
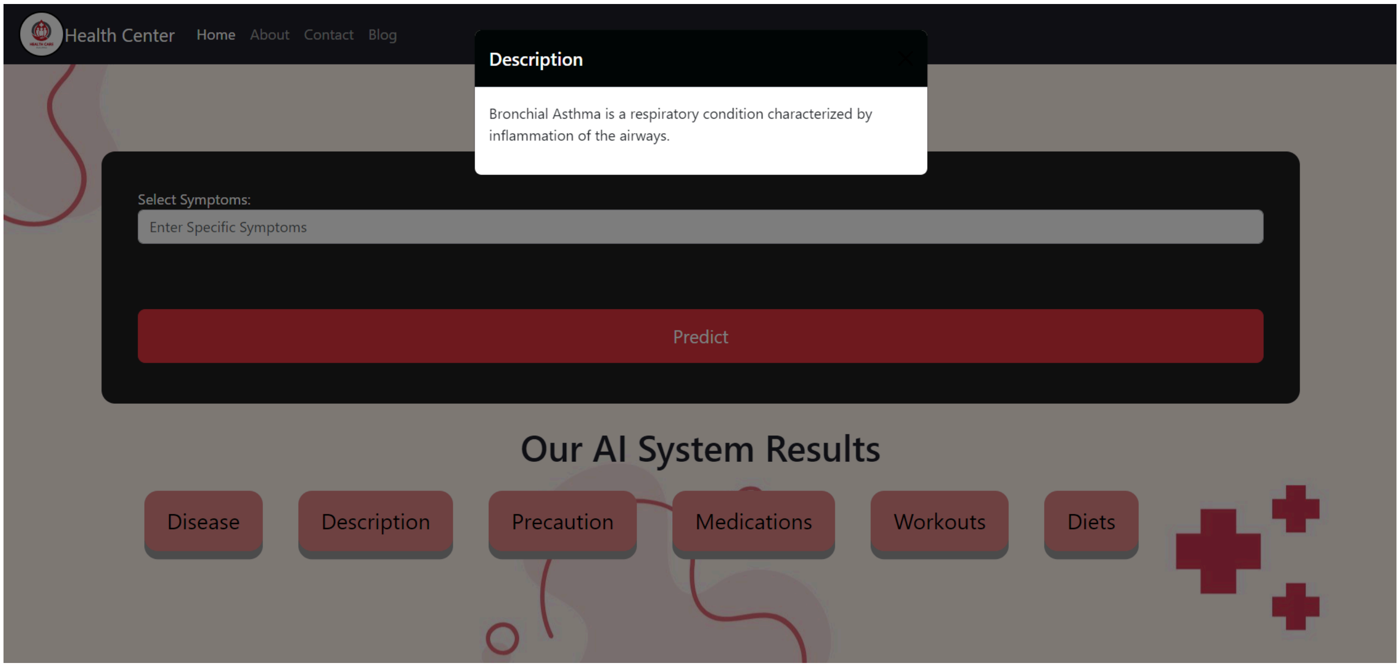


Fig 5.2 After Prediction



Description

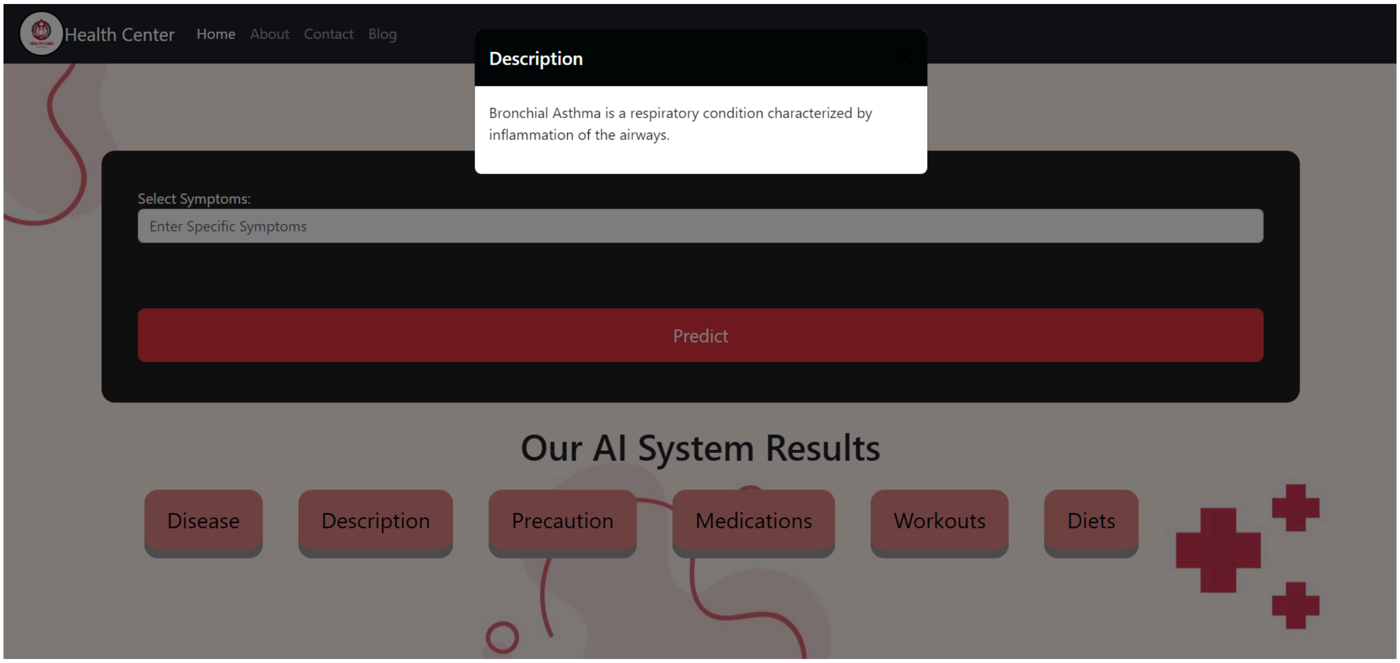
Bronchial Asthma is a respiratory condition characterized by inflammation of the airways.

Select Symptoms:

Enter Specific Symptoms

Our AI System Results

Fig 5.3 Disease Module



Description

Bronchial Asthma is a respiratory condition characterized by inflammation of the airways.

Select Symptoms:

Enter Specific Symptoms

Our AI System Results

Fig 5.4 Description Module

5.2 RESULT

The disease prediction project encompasses five distinct modules: disease, description, workout, precautions, and diet. Each of these modules plays a pivotal role in the predictive process, collectively aimed at providing comprehensive insights into potential health conditions. Leveraging machine learning algorithms, the system effectively predicts the disease corresponding to the input symptoms, offering users a clear understanding of their health status. Furthermore, it goes beyond mere diagnosis, furnishing detailed descriptions of the identified diseases, tailored workout regimens, necessary precautions to mitigate risks, and personalized dietary recommendations. Through seamless integration and analysis of symptom data, the project empowers users with actionable insights, fostering informed decisions regarding their well-being and promoting proactive healthcare management.

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENT

6.1 CONCLUSION

In conclusion, the development and implementation of a disease prediction system utilizing machine learning algorithms mark a significant advancement in the realm of healthcare technology. Through the amalgamation of advanced computational techniques and medical knowledge, this system holds immense potential in revolutionizing disease diagnosis and prognosis. By accurately analyzing symptoms and patterns, it enables early detection of various health conditions, thereby facilitating timely intervention and treatment. Furthermore, the system's ability to generate personalized recommendations, including workout routines, precautions, and dietary guidelines, enhances the user's understanding of their health status and empowers them to make informed lifestyle choices. The integration of multiple modules, encompassing disease prediction, description, workout, precautions, and diet, ensures a holistic approach to health management. Moreover, its user-friendly interface and accessibility make it a valuable tool for both healthcare professionals and individuals seeking to monitor and improve their well-being. As technology continues to evolve, the disease prediction system serves as a testament to the transformative potential of machine learning in reshaping the landscape of healthcare delivery, ultimately contributing to better health outcomes and improved quality of life for individuals worldwide.

6.2 FUTURE ENHANCEMENT

Looking forward, enhancing the disease prediction system using machine learning algorithms involves several key strategies. Integrating real-time health data from wearables or electronic health records can improve prediction accuracy by adapting to changing health statuses promptly. Personalization efforts, including factors like genetics and lifestyle, can increase user engagement and adherence to recommendations. Expanding the database of diseases and symptoms ensures the system predicts a broader spectrum of conditions accurately. Advanced analytical techniques like deep learning refine predictions from complex datasets, while integration with telemedicine platforms enhances access to healthcare services, particularly in remote areas. Emphasizing preventive healthcare by identifying high-risk individuals and promoting lifestyle modifications fosters proactive health management. Through these enhancements, the system evolves into a sophisticated tool, aiding in early detection, risk mitigation, and personalized health interventions for users worldwide.

APPENDIX

SOURCE CODE:

```
from flask import Flask, request, render_template, jsonify  # Import jsonify
import numpy as np
import pandas as pd
import pickle

# flask app
app = Flask(__name__)

# load databasedataset======================================
sym_des = pd.read_csv("datasets/symtoms_df.csv")
precautions = pd.read_csv("datasets/precautions_df.csv")
workout = pd.read_csv("datasets/workout_df.csv")
description = pd.read_csv("datasets/description.csv")
medications = pd.read_csv('datasets/medications.csv')
diets = pd.read_csv("datasets/diets.csv")
# load model===============================================
svc = pickle.load(open('models/svc.pkl','rb'))


#=================================================================
# custome and helping functions
#============================helper funtions=================
def helper(dis):
   desc = description[description['Disease'] == dis]['Description']
   desc = " ".join([w for w in desc])

   pre = precautions[precautions['Disease'] == dis][['Precaution_1', 'Precaution_2',
'Precaution_3', 'Precaution_4']]
   pre = [col for col in pre.values]

   med = medications[medications['Disease'] == dis]['Medication']
   med = [med for med in med.values]

   die = diets[diets['Disease'] == dis]['Diet']
   die = [die for die in die.values]

   wrkout = workout[workout['disease'] == dis] ['workout']
```

```
    return desc,pre,med,die,wrkout

# Model Prediction function
def get_predicted_value(patient_symptoms):
    input_vector = np.zeros(len(symptoms_dict))
    for item in patient_symptoms:
        input_vector[symptoms_dict[item]] = 1
    return diseases_list[svc.predict([input_vector])[0]]

# creating routes========================================

@app.route("/")
def index():
    return render_template("index.html")

# Define a route for the home page
@app.route('/predict', methods=['GET', 'POST'])
def home():
    if request.method == 'POST':
        symptoms = request.form.get('symptoms')
        # mysysms = request.form.get('mysysms')
        # print(mysysms)
        print(symptoms)
        if symptoms =="Symptoms":
            message = "Please either write symptoms or you have written misspelled
symptoms"
            return render_template('index.html', message=message)
        else:

            # Split the user's input into a list of symptoms (assuming they are
comma-separated)
            user_symptoms = [s.strip() for s in symptoms.split(',')]
            # Remove any extra characters, if any
            user_symptoms = [symptom.strip("[]' ") for symptom in user_symptoms]
            predicted_disease = get_predicted_value(user_symptoms)
            dis_des, precautions, medications, rec_diet, workout =
helper(predicted_disease)

            my_precautions = []
            for i in precautions[0]:
                my_precautions.append(i)
            return render_template('index.html', predicted_disease=predicted_disease,
dis_des=dis_des,
                                  my_precautions=my_precautions, medications=medications,
my_diet=rec_diet,
                                  workout=workout)

    return render_template('index.html')

# about view funtion and path
@app.route('/about')
def about():
    return render_template("about.html")
# contact view funtion and path
@app.route('/contact')
def contact():
    return render_template("contact.html")

# about view funtion and path
@app.route('/blog')
def blog():
    return render_template("blog.html")

if __name__ == '__main__':

    app.run(debug=True)
```

REFERENCES

[1]“HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System,” IEEE Journals & Magazine | IEEE Xplore, 2020. https://ieeexplore.ieee.org/document/9144587

[2]“Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,” IEEE Journals & Magazine | IEEE Xplore, 2019. https://ieeexplore.ieee.org/document/8740989

[3]“Synergistic Feature Engineering and Ensemble Learning for Early Chronic Disease Prediction,” IEEE Journals & Magazine | IEEE Xplore, 2024. https://ieeexplore.ieee.org/document/10511078

[4]“An Integrated Two-Layered Voting (TLV) Framework for Coronary Artery Disease Prediction Using Machine Learning Classifiers,” IEEE Journals & Magazine | IEEE Xplore, 2024. https://ieeexplore.ieee.org/document/10500702

[5]“Metabolite-disease association prediction algorithm combining DeepWalk and random forest,” TUP Journals & Magazine | IEEE Xplore, Feb. 01, 2022. https://ieeexplore.ieee.org/document/9515700