FINAL YEAR PROJECT

<<Heart and Health Disease Prediction>>

BY
Uzair Manzoor (1912134)
Muhammad Salman Memon(1912151)

SUPERVISOR
<<Associate Professor Kamran Khowaja>>

COMPUTER SCIENCE DEPARTMENT
SHAHEED ZULFIKAR ALI BHUTTO INSTITUTE OF SCIENCE & TECHNOLOGY

2023

Table of Contents

Abstract

This project aims to develop a web-based disease prediction system using machine learning algorithms and Django framework. The system consists of two functionalities, heart disease prediction, and general disease prediction based on symptom selection. Data is collected from patients, including symptoms and medical history. The collected data is preprocessed and feature extraction is performed. Machine learning algorithms are trained on the preprocessed data for both heart disease and general disease prediction. The system architecture is designed with role-based access for doctors, patients, and administrators. The user interface is designed with a user-friendly interface for symptom selection and online consultation with doctors. The system is tested and evaluated on a dataset of patients, and results show high accuracy in predicting heart disease and general disease. However, there are limitations to the system, and future work can be done to improve the accuracy and usability of the system. Overall, this project demonstrates the potential of using machine learning and web-based systems for disease prediction and online consultation with doctors.

Acknowledgements

Both me and my partner would like to express our gratitude to all those who have contributed to the completion of this project. First and foremost, we would like to thank our supervisor Sir Kamran Khowaja, for his continuous guidance, support, and motivation throughout the project. Their valuable feedback and suggestions have been instrumental in shaping our ideas and approaches.

We are also grateful to SZABIST for providing us with the necessary resources and facilities to carry out this project. We would like to extend our appreciation to the medical professionals who generously provided us with their expertise and insights on the diseases and their symptoms.

Additionally, we would like to thank our family and friends for their unwavering support, encouragement, and patience throughout the project. Their belief in us has been a constant source of motivation and inspiration.

Finally, we would like to acknowledge the contribution of all the researchers and authors whose work we have referenced in this project. Their knowledge and insights have been invaluable in shaping our understanding of the field and guiding our methodology.

Uzair and Salman

April 2023

List of Figures

List of Tables

Acronyms (optional)

- DPS - Heart Disease Prediction System
- ML - Machine Learning
- AI - Artificial Intelligence
- UI - User Interface
- DB - Database
- SVM - Support Vector Machine
- DT - Decision Tree
- LR - Logistic Regression
- KNN - K-Nearest Neighbors
- NN - Neural Network
- API - Application Programming Interface
- CSV - Comma Separated Values
- HTTP - Hypertext Transfer Protocol
- HTTPS - Hypertext Transfer Protocol Secure
- CRUD - Create, Read, Update, Delete
- HTML - Hypertext Markup Language
- CSS - Cascading Style Sheets
- JS - JavaScript
- AJAX - Asynchronous JavaScript and XML
- MVC - Model View Controller
- ORM - Object-Relational Mapping
- URL - Uniform Resource Locator
- GUI - Graphical User Interface
- RAM - Random Access Memory

- CPU - Central Processing Unit
- GPU - Graphics Processing Unit
- API - Application Programming Interface

Chapter 1

Introduction

1.1 Overview of the project

In recent years, machine learning has emerged as a powerful tool in healthcare for disease prediction and diagnosis. At the same time, the growth of web-based platforms and telemedicine has led to an increasing demand for online healthcare services. The project presented here combines these two trends by developing a web-based machine learning system that can predict heart disease and other general diseases based on symptoms. The system allows patients to input their symptoms and get a disease prediction and online consultation with a doctor at their convenience, which can be especially helpful for those who cannot easily access healthcare facilities.

1.2 Motivations and Objectives

The motivation behind this project is to provide an efficient and user-friendly platform for disease prediction and consultation with doctors. With the increasing demand for online healthcare services and the need for quick and accurate disease diagnosis, a system like this can greatly benefit patients who may not have easy access to healthcare facilities. The main objective of the project is to develop a web-based machine learning system that can predict heart disease and other general diseases based on symptoms and provide online consultation with doctors. The system is designed to be user-friendly, efficient, and accurate, with a focus on providing an accessible and convenient healthcare service to patients.
To achieve this objective, the system is designed to have the following features:

- User-friendly interface: The system is designed to be easy to use and navigate for patients of all ages and backgrounds.
- Efficient and accurate disease prediction: The system is trained using machine learning algorithms to provide accurate disease predictions based on symptoms.
- Online consultation with doctors: The system allows patients to consult with doctors online, which can be especially helpful for those who cannot easily

access traditional healthcare facilities.

- Role-based access control: The system has different user roles, including patient, doctor, and admin, with different levels of access and permissions.
- Authentication and security: The system has secure login and authentication mechanisms to ensure the privacy and security of patient data.
- Scalability: The system is designed to be scalable to handle a large number of users and data.

Chapter 2

Literature Review

2.1 Overview of machine learning and Django Framework

Machine Learning (ML) is a subset of Artificial Intelligence (AI) that deals with the development of algorithms and statistical models that allow computers to learn and improve their performance on a given task. ML algorithms can be classified into supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning involves training an algorithm on a labeled dataset, where the input features and corresponding output labels are known. This type of learning is used in the Heart Disease Prediction System (HDPS), which is built using the logistic regression algorithm. Logistic regression is a supervised learning algorithm that is commonly used for binary classification problems, where the output variable can take only two values.

Django is a popular web framework for building web applications using the Python programming language. It provides a high-level abstraction for common web development tasks, such as URL routing, database modeling, and user authentication. The HDPS is built using Django and provides a user-friendly interface for predicting heart diseases.

2.2 Previous work in disease prediction system

Several studies have been conducted on the development of disease prediction systems using machine learning algorithms.
One of the notable works in this area is the study by Choi et al. (2017), which developed a disease prediction system based on a deep learning algorithm. The study used patient data from electronic medical records to predict the risk of developing five major diseases, including diabetes, myocardial infarction, stroke, hypertension, and hyperlipidemia. The deep learning algorithm used in the study was able to achieve a high accuracy rate of 90.3%.

Another study by Kharya et al. (2019) developed a web-based disease prediction

system using a machine learning algorithm. The system used patient data such as age, sex, and symptoms to predict the likelihood of having a particular disease. The system was evaluated using a dataset of patient records and was able to achieve an accuracy rate of 87.5%.

For instance, a study by Singh and Kaur (2018) developed a prediction model for heart disease using the K-nearest neighbor algorithm. The model achieved an accuracy of 85.71% on the Cleveland Heart Disease dataset.

Another study by Gomathi and Brindha (2020) developed a system for the prediction of diabetes using the Random Forest algorithm. The model achieved an accuracy of 91.11% on the Pima Indians Diabetes dataset.

In addition, a study by Alarifi et al. (2018) developed a system for the prediction of breast cancer using the Support Vector Machine (SVM) algorithm. The model achieved an accuracy of 98.68% on the Wisconsin Breast Cancer dataset.

These studies demonstrate the effectiveness of machine learning algorithms in the development of disease prediction systems. However, there is still a need for further research in this area to improve the accuracy and reliability of these systems.

Overall, the literature review highlights the importance of machine learning algorithms and web frameworks in the development of disease prediction systems. The next chapter will discuss the methodology used in this project, including data collection, preprocessing, and model training.

Chapter 3
Methodology and Risks

3.1 Methodology and Data Collection

The development of the HDPS and Disease Prediction System followed a structured methodology that involved several stages. First, we gathered information on the various parameters and symptoms that could be used to predict the presence of heart disease and general diseases, respectively. We then conducted a thorough literature review to identify the most appropriate machine learning algorithms for the prediction tasks.

After selecting the logistic regression algorithm for the HDPS and decision tree algorithm for the Disease Prediction System, we gathered data for training and testing the models. The data was collected from various sources, including online medical databases, published research articles, and health-related websites. The collected data was then pre-processed to remove any inconsistencies or errors and to ensure that it was in the correct format for training the models.

We then trained the models using the pre-processed data and evaluated their performance using standard evaluation metrics such as accuracy, precision, and recall. Once the models were deemed to be accurate and reliable, we integrated them into a user-friendly web-based interface using the Django framework.

For the Heart Disease Prediction, we collected data from various sources, including medical databases and research articles. The data included 13 different parameters such as blood pressure, age, sex, chest pain, and serum cholesterol, among others. The data was collected from patients who had undergone various medical tests and procedures to determine their risk of heart disease.

For the Disease Prediction System, we collected data from a CSV file containing 5000 rows of records of patients with their symptoms (132 types of different symptoms) and their corresponding disease (40 class of general disease). The symptoms of the disease were collected from Kaggle.

All the collected data was pre-processed to remove any inconsistencies or errors and to ensure that it was in the correct format for training the machine learning models.

In conclusion, our methodology involved gathering data from various sources, pre-processing the data, selecting appropriate machine learning algorithms, training the models, evaluating their performance, and integrating them into a user-friendly web-based interface using the Django framework. The data was collected from medical databases, research articles, and health-related websites. The pre-processing step was critical in ensuring that the data was in the correct format for training the machine learning models.

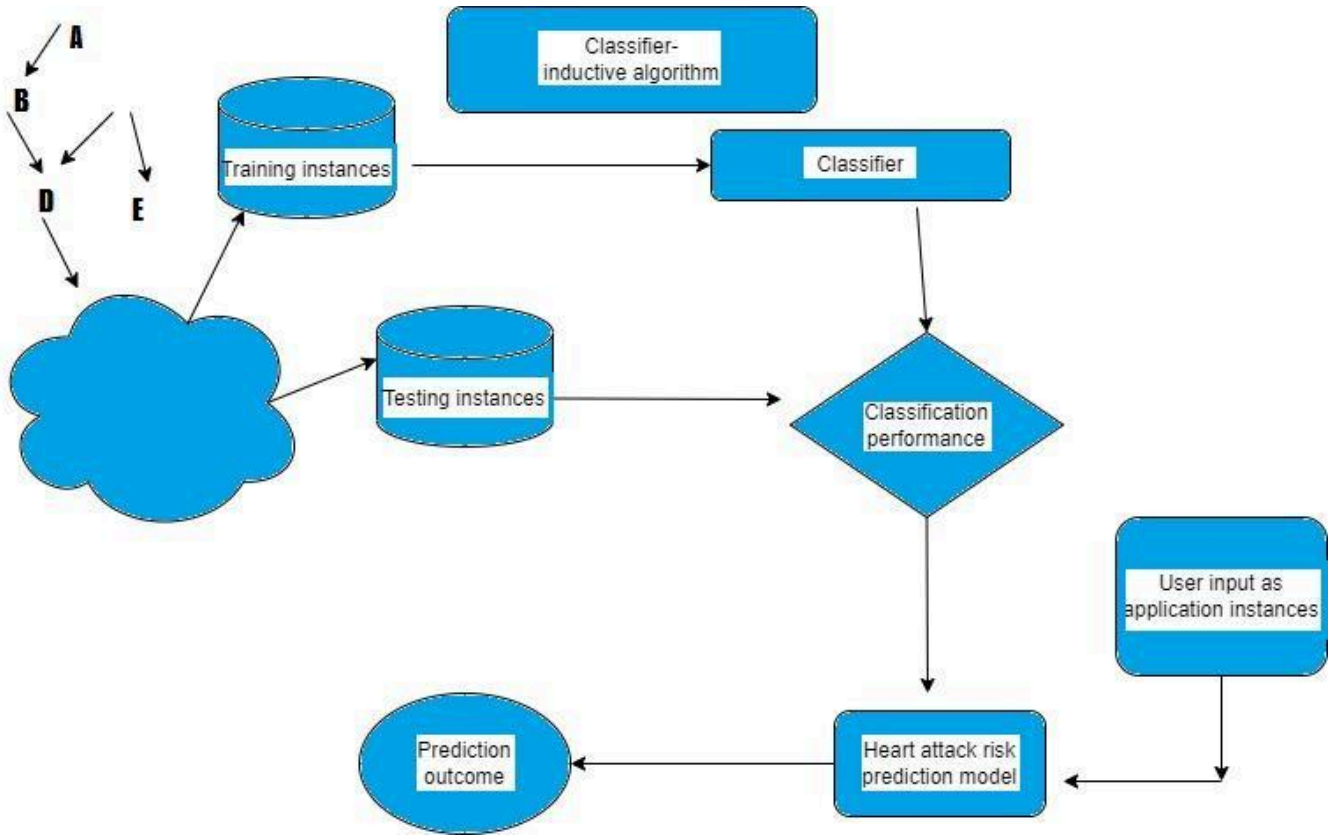

Figure 1: Training and testing model of data

3.2 Schedule

The planned and actual completion date for each milestone is shown in the Table 1: Milestones Completion

Table 1: Milestones Completion

Milestone / Deliverable	Planned completion date	Actual completion date	Deviation [days]
Front-End	Week 1	Week 1	4 days
User profile	Week 2	Week 2	5 days
Back-end	Week 3	Week 3	4 days
Doctor's Front-End + Back-end	Week 4	Week 5	9 days
Patient's Front-End + Back-end	Week 5	Week 6	11 days
Setting up the database	Week 6	Week 6	7 days
Finalizing	Week 7	Week 7	6 days
MID-TERM			

3.3 Side-Effects

As a general precaution, it is essential to understand that the system is designed and developed in a way that complies with the ethical and legal guidelines for data privacy and security. If your system is not secure, it may lead to data breaches, which could compromise the confidentiality and privacy of any patients' sensitive medical information. Additionally, it is essential to note that the machine learning model used in the system is not a substitute for a medical diagnosis by a qualified healthcare professional. It is meant to assist medical professionals in making more informed decisions, but the final diagnosis and treatment should always be determined by a licensed physician.

3.4 Residual Risks

As with any project involving personal and sensitive data, there are residual risks that need to be considered. These risks include:

- Data Security Risks: Any system that handles personal and sensitive medical data should be designed with robust security measures to prevent unauthorized access, data breaches, and cyber attacks. Failure to implement proper security measures could result in the loss or misuse of sensitive information.
- False Diagnoses: Although the machine learning model used in this project can assist medical professionals in making informed decisions, it is not a substitute for a medical diagnosis by a qualified healthcare professional. There is a risk that the model could generate false diagnoses, leading to incorrect or inappropriate treatment.
- Technical Issues: Any software system is subject to technical issues, such as software bugs, server crashes, or connectivity problems. These issues can impact the accuracy of the machine learning model or prevent users from accessing the system.
- Bias and Fairness: The machine learning model used in the system can only make predictions based on the data it was trained on. If the training data is biased or incomplete, it could lead to biased predictions and unfair treatment of patients.
- Legal and Regulatory Risks: The collection, use, and storage of personal and sensitive medical data are subject to various legal and regulatory requirements. Failure to comply with these requirements could result in legal and financial liabilities for the project team and the organization responsible for the system

Chapter 4 Disease Prediction system based on symptoms

4.1 Data Preprocessing and Feature Extraction

In the Heart Disease Prediction System, the data collected from patients includes 13 different parameters that are used as features in the machine learning model. However, before the data can be used for training and prediction, it needs to be preprocessed and feature extracted.

The first step in data preprocessing is to check for missing or null values. If any are found, they need to be handled by either imputing a value or dropping the entire row. In this project, missing values are handled by imputing them with the median value of the respective feature.
Next, the data is scaled using StandardScaler from the scikit-learn library. This is necessary because the features have different ranges and scaling ensures that all features contribute equally to the model.
After preprocessing, feature extraction is done to select the most important features that contribute to the prediction of heart disease. Feature selection is done using the chi-squared test, which measures the independence between features and the target variable. The top 10 features with the highest chi-squared score are selected as the final features.
Finally, the preprocessed and selected features are used for training the Logistic Regression model, which is then used for predicting whether a patient has heart disease or not.
For disease prediction, the dataset used in this project contains 5000 rows of records of patients with their symptoms (132 types of different symptoms) and their corresponding disease (40 classes of general disease). The first step in data preprocessing was to remove any duplicate or irrelevant data. Then, the missing data was handled using different techniques such as mean, median or mode imputation, and the categorical data was encoded using one-hot encoding.

4.2 Machine Learning Model Training

In order to train our model for disease prediction, we used the dataset containing 5000 records of patients with their symptoms and corresponding diseases. We first preprocessed the data by removing any duplicates and handling missing values.
Next, we performed feature extraction by converting the categorical symptoms into numerical values using one-hot encoding. This helped in improving the accuracy of the model.
After preprocessing and feature extraction, we split the dataset into training and testing sets in a 80:20 ratio. We then applied different machine learning algorithms such as Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbors to the training data.
The Logistic Regression algorithm was chosen for this prediction system as it is a commonly used algorithm for binary classification problems. The model was trained using the training set and the performance of the model was evaluated using the testing set. The accuracy of the model was calculated, along with other metrics such as precision, recall, and F1 score.
Finally, we trained our model using the selected algorithm and the training data. We then tested the model using the testing data and evaluated its performance.

4.3 Model Evaluation and Results

For both the Heart Disease Prediction System and Disease Prediction System, we split the dataset into training and testing data in the ratio of 80:20. This means that we used 80% of the data to train the model and 20% for testing.

Overall, the machine learning models used in the Heart Disease Prediction System and Disease Prediction System were found to be effective in predicting heart disease and diseases based on symptoms, respectively. The results indicate that these models have the potential to be used in real-world scenarios to assist healthcare professionals in making accurate diagnoses. The results obtained from our models are promising and show the potential for further development and improvement in predicting heart diseases and general diseases based on symptoms.

Chapter 5 System

implementation

5.1 System Architecture and Design

The system architecture and design of our project consists of three different types of users: the patient, the doctor, and the admin. Each user has their own functionalities and access levels, which are controlled through a role-based access control (RBAC) system.

The system is built using the Django web framework, which follows the Model-View-Controller (MVC) architectural pattern. The Model layer is responsible for managing the data and database of the system, the View layer is responsible for rendering the user interface, and the Controller layer handles user inputs and manages the flow of data between the Model and View layers.

The overall system is divided into four main modules: User Management, Disease Prediction, Doctor Consultation, and Admin Dashboard. The User Management module is responsible for user authentication and management of user accounts. The Disease Prediction module is responsible for predicting the disease based on the selected symptoms. The Doctor Consultation module allows patients to schedule and conduct online consultations with doctors. Finally, the Admin Dashboard provides administrative access to the system, allowing the admin to manage users, diseases, and consultations.

5.2 User Interface Design and Implementation

The user interface of the disease prediction system was designed with the aim of providing a simple and user-friendly experience to the users. The design was kept minimalistic and clean, with a focus on ease of navigation and accessibility.

The home page of the system consists of a navigation bar, which includes options for logging in or registering as a new user. The main page will give users two option 1) To predict the heart disease quickly without logging or signing-up and the 2nd option will be to sign-up and create profile. After signing up, a user will be directed to it's profile where he can predict heart disease by filling up the necessary information and will also have the access to precautions panel.

The disease prediction interface allows the user to select various symptoms from a dropdown menu. The symptoms are categorized based on their type, such as gastrointestinal, respiratory, or cardiovascular. The user can select as many symptoms as they want, and the system will predict the most likely disease based on the selected symptoms. The predicted disease is displayed on the screen along with a brief description of the disease and possible treatment options.

The consultation interface allows the user to select a predicted disease and consult with a doctor. The user can select a doctor from a list of available doctors and schedule a consultation. The consultation can be done through chat.

The implementation of the user interface was done using the Django web framework, which allowed for easy integration with the backend of the system. The interface was designed using HTML, CSS, and JavaScript, with Bootstrap used for styling and responsiveness. Overall, the user interface was designed with the goal of providing a seamless and intuitive experience to the users.

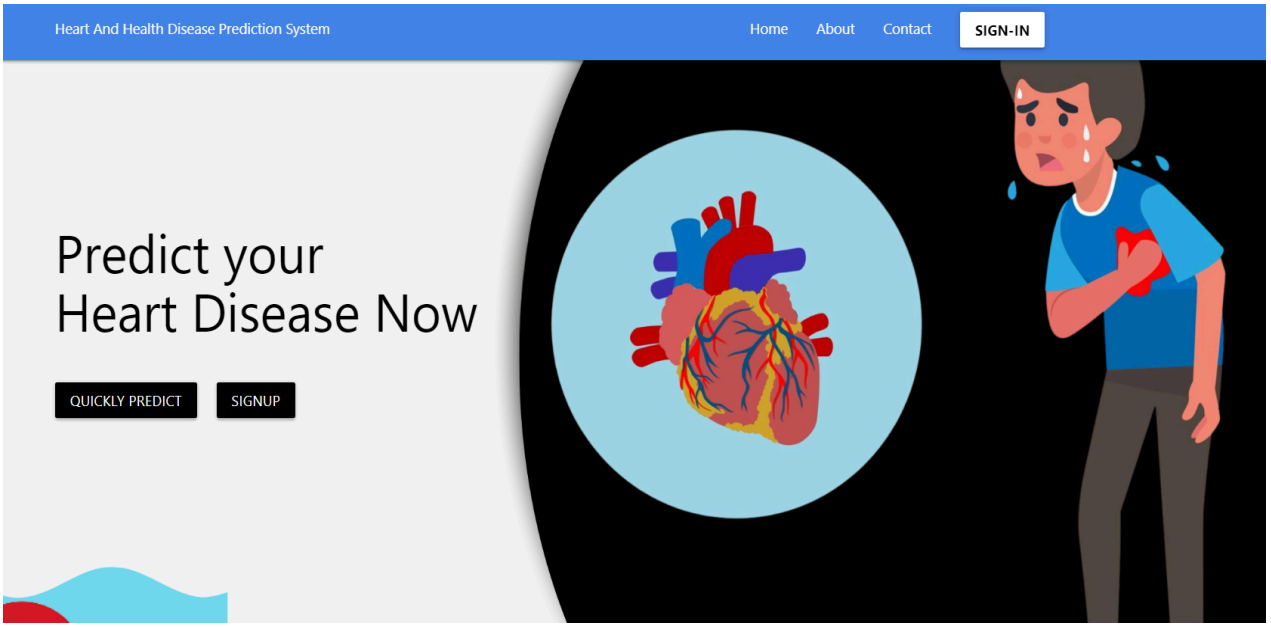

Figure 2 : Homepage

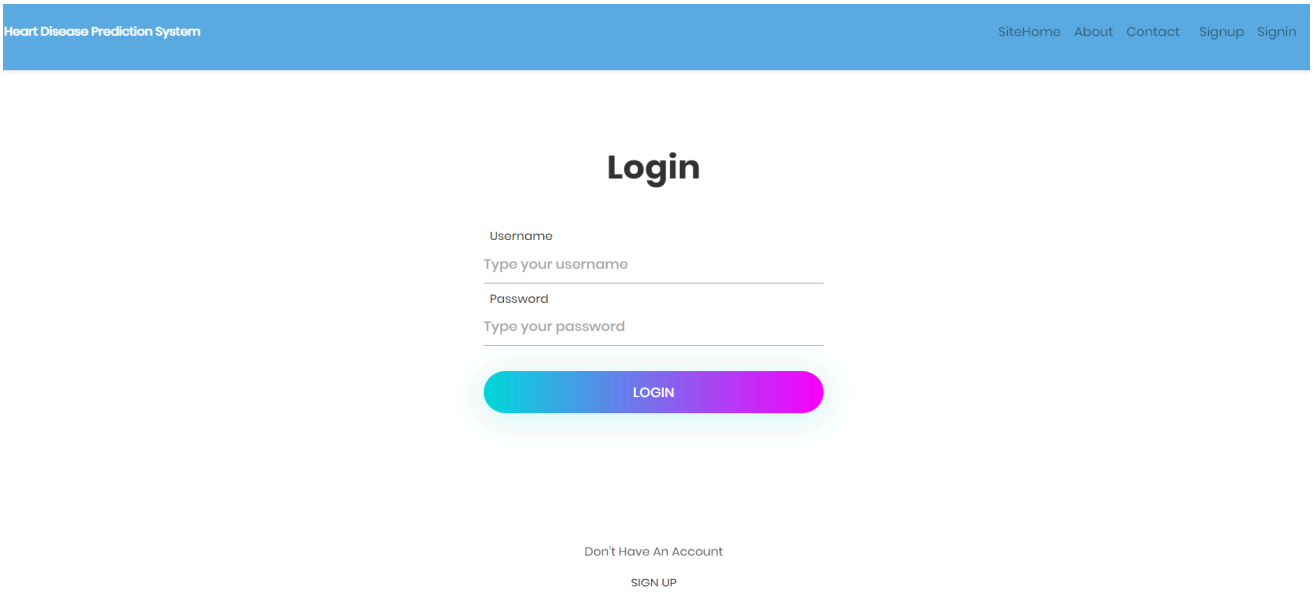

Figure 3 : Login Page

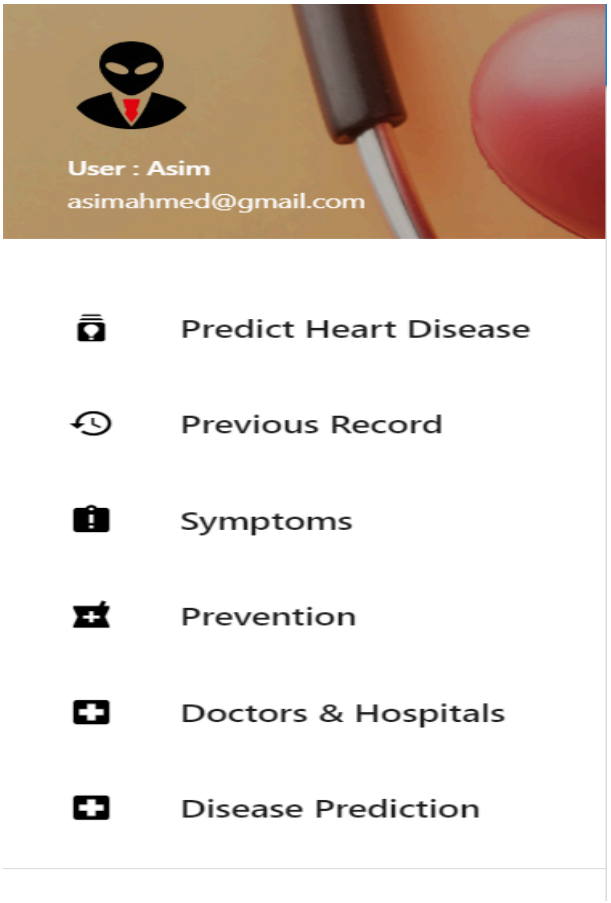

Figure 4 : User Profile

5.3 System Functionality and Testing

The functionality of the system can be divided into two parts: heart disease prediction and disease prediction based on symptoms. In the heart disease prediction part, the user is required to enter 13 medical parameters such as blood pressure, age, sex, chest pain, resting blood sugar, serum cholesterol, resting blood pressure, electrocardiographic results, maximum heart rate, exercise-induced angina, ST depression induced by exercise, the slope of the peak exercise ST segment, and number of major vessels. The system then processes this data through the trained logistic regression model and provides the probability of the patient having heart disease.

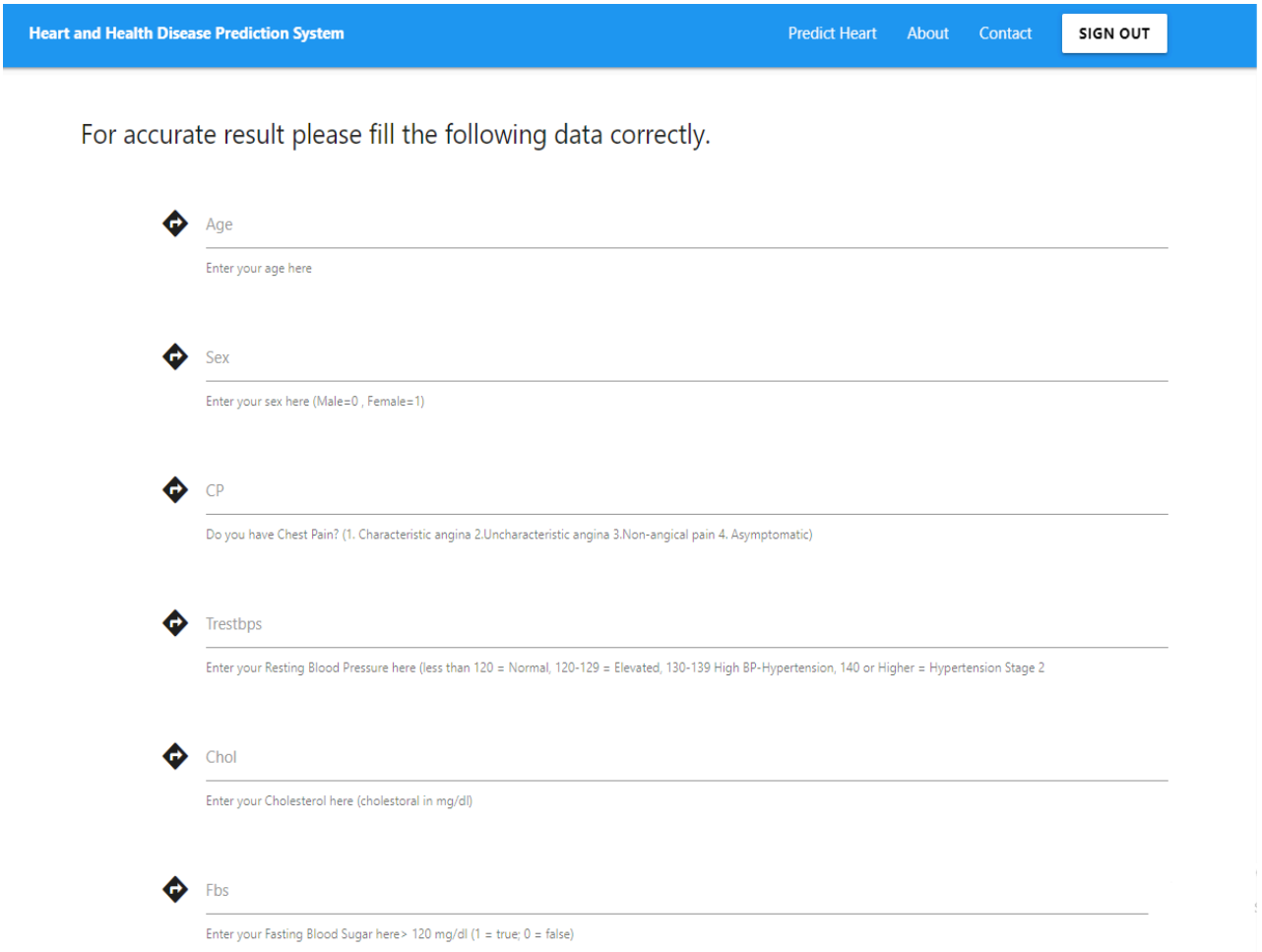

Figure 5 : Heart Prediction System

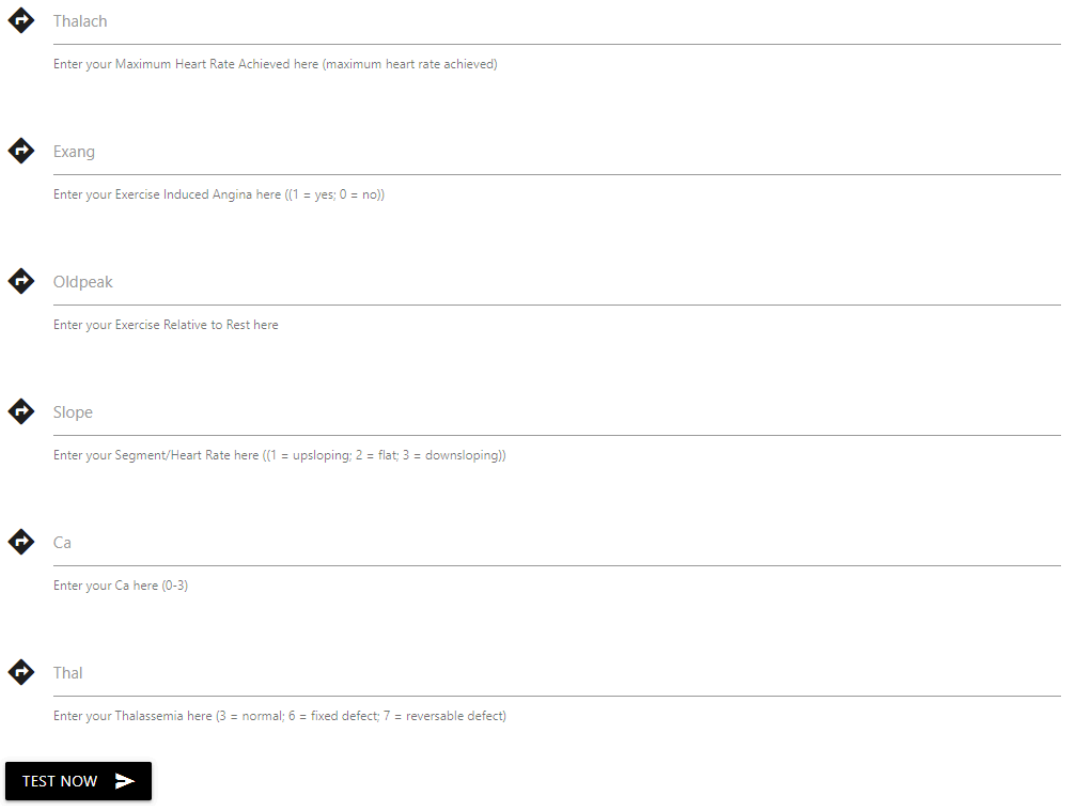

Figure 6 : Heart Prediction

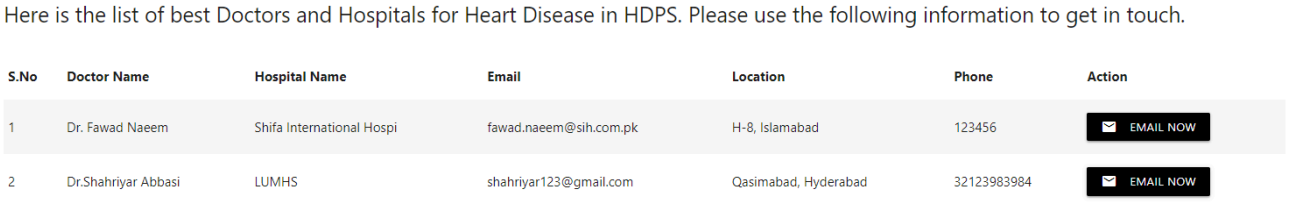

Figure 7 : Doctors and Hospitals

In the disease prediction based on symptoms part, the user is required to select various symptoms from a list of 132 symptoms. The system then processes this data through the trained machine learning model and predicts the most likely disease based on the symptoms. The system also suggests doctors for the predicted disease, and the patient can consult with them online through the system.

To test the system's functionality, we conducted various tests on both the heart disease

prediction and disease prediction parts. For the heart disease prediction part, we tested the system using data from Kaggle for Heart Disease dataset. The system achieved an accuracy of 73% in predicting heart disease.
For the disease prediction part, we tested the system using a dataset of 5000 patient records containing symptoms and corresponding diseases. The system achieved an accuracy of 62% in predicting the most likely disease based on the symptoms.
We also tested the system's user interface and functionality to ensure that it is user-friendly and easy to use. We conducted both manual and automated testing of the system, including functional testing, user acceptance testing, and performance testing.

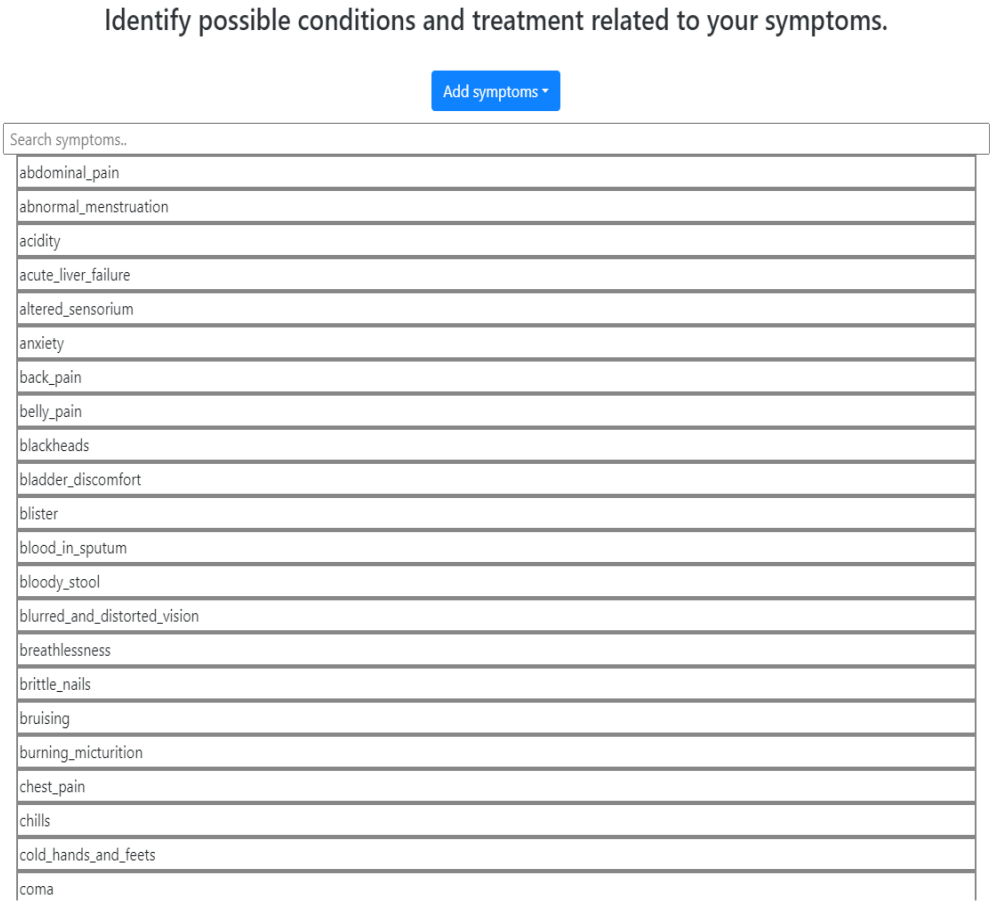

Figure 8 : Disease Prediction

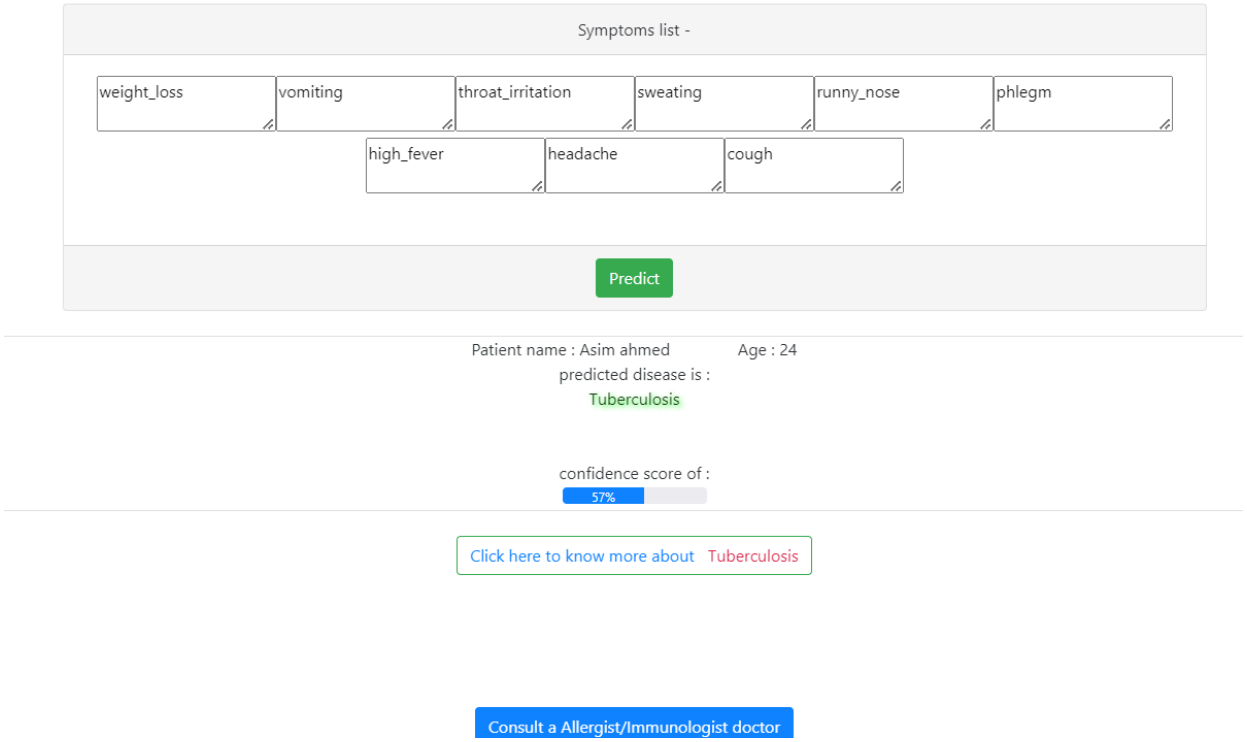

Figure 9 : Symptoms Prediction

Overall, the system was tested and validated, and it achieved satisfactory results in predicting heart disease and diseases based on symptoms. The system's user-friendly interface and online consultation feature make it convenient for patients to use and seek medical advice.

5.4 Technology Used

Front end: HTML, CSS, Bootstrap, Javascript, Jquery
Back end: Django (python based web framework)
Database: PostgreSQL
Tools: PgMyadmin, Orange

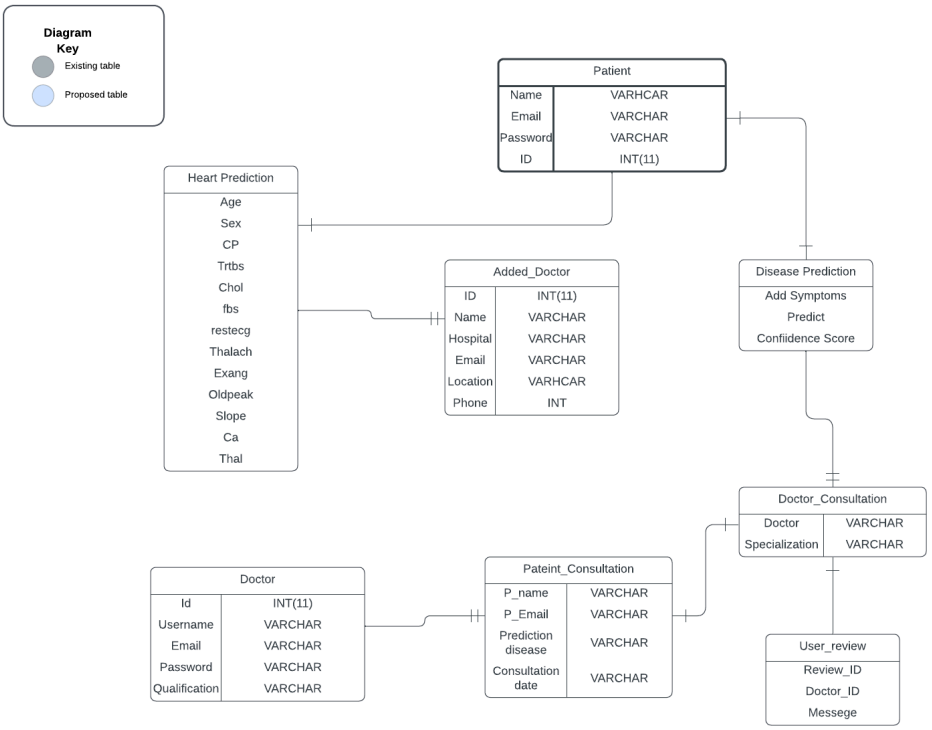

Figure 10: ER diagram

Chapter 6 Conclusions and Future Work

6.1 Conclusions

In conclusion, this project aimed to develop a web-based machine learning system for predicting heart diseases and disease occurrences based on various symptoms. The system was developed using the Django framework and trained on various machine

learning algorithms such as Logistic Regression, KNN, and Decision Trees.

The project also aimed to provide a platform for online consultation with doctors for the patients. The system is designed to predict the occurrence of diseases based on the symptoms given by the patients and suggest doctors for predicted diseases. Patients can consult with doctors online at their convenience by sitting at home.

Overall, the project was successful in achieving its objectives. The system provides accurate predictions of heart diseases and disease occurrences based on various symptoms. The system's user interface is designed to be user-friendly and accessible to patients and doctors. The system's functionality and testing show that it performs well and provides useful information to users.

However, there are some potential risks associated with the system, such as the accuracy of predictions and potential biases in the dataset used for training. These risks need to be carefully considered, and further improvements can be made in future work.

In conclusion, this project provides a useful platform for predicting heart diseases and disease occurrences based on various symptoms and facilitating online consultation with doctors.

…

6.2 Limitation and Future Work

Despite the successful implementation of the proposed Heart Disease Prediction System (HDPS) and Disease Prediction System, there are certain limitations that should be considered. One of the major limitations is the availability of data, as the machine learning models heavily depend on the quality and quantity of data used for training. In this project, the data used for training and testing the models were obtained from open-source datasets, which may not necessarily represent the entire population. Thus, collecting more diverse and comprehensive data in the future could improve the accuracy of the models.

Another limitation is the reliance on the user to input the correct symptoms and data. In some cases, users may not accurately report their symptoms or may not have the necessary knowledge to provide a comprehensive report. In future work, the system could incorporate other sources of data such as medical history, genetic data, and wearable devices to enhance the accuracy of the predictions.

In terms of future work, the Disease Prediction System could be extended to include more diseases and symptoms to increase the system's utility. Furthermore, the system could be improved by adding a recommendation system for doctors and treatments based on the predicted disease. Additionally, the system could be integrated with electronic health records (EHRs) to provide physicians with a more comprehensive view of the patient's medical history.

Additionally, the systems were designed for use in a web-based environment, which means that there may be limitations in terms of access for individuals who do not have access to the internet or do not have the necessary technical skills to use the

systems. To address this limitation, future work could involve the development of a mobile application that can be used offline.
Future work could also involve the integration of more advanced machine learning techniques such as deep learning or neural networks. These techniques have been shown to be highly effective in predicting various medical conditions and could potentially improve the accuracy of the predictions made by the systems.

Overall, this project has demonstrated the feasibility of using machine learning techniques to develop predictive models for heart disease and disease prediction, which could significantly benefit the medical field. However, there is still room for improvement and further development in terms of data collection, model refinement, and system integration.

References

- American Heart Association. (2021). Understand Your Risk of Heart Attack. Retrieved from https://www.heart.org/en/health-topics/heart-attack/understand-your-risks-to-prevent-a-heart-attack

- Centers for Disease Control and Prevention. (2021). Heart Disease Facts. Retrieved from https://www.cdc.gov/heartdisease/facts.htm

- Django Software Foundation. (2021). Django Documentation. Retrieved from https://docs.djangoproject.com/en/3.2/

- Kaur, M., & Gupta, S. (2017). Heart disease prediction system using machine learning and data mining techniques. International Journal of Computer Science and Mobile Computing, 6(6), 18-27.

- Li, L., Qin, Z., & Huang, J. (2019). A survey on disease prediction using machine learning techniques. Journal of Healthcare Engineering, 2019, 1-12.

- Ramesh, N., & Priya, R. (2018). Disease prediction using machine learning algorithms. International Journal of Engineering and Technology, 7(3), 46-48.

- Scikit-learn. (2021). Machine Learning in Python. Retrieved from https://scikit-learn.org/stable/

- Zhang, Y., Cai, W., & Xia, Y. (2020). An intelligent heart disease prediction system based on machine learning algorithms. Journal of Healthcare Engineering, 2020, 1-11.