Multi Model Based Disease Prediction System Using Machine Learning

---

A PROJECT REPORT

Submitted to

Visvesvaraya Technological University

BELAGAVI - 590 018

Submitted by

Disha P U
USN: 4MW21AD015

Hithashree
USN: 4MW21AD019

Neha Dinesh Shettigar
USN: 4MW21AD030

Nidhi Prabhu K
USN: 4MW21AD032

Under the guidance of

Dr. Ganesha Prasad

Assistant professor (Selection Grade),

Dept. of Artificial Intelligence and Machine Learning Engineering

Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal, UDUPI

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Engineering

Department of Artificial Intelligence and Data Science Engineering

SHRI MADHWA VADIRAJA INSTITUTE OF TECHNOLOGY AND MANAGEMENT

Vishwothama Nagar, BANTAKAL – 574 115, Udupi District

2024-2025

SHRI MADHWA VADIRAJA INSTITUTE OF TECHNOLOGY AND MANAGEMENT

(A Unit of Shri Sode Vadiraja Mutt Education Trust ®, Udupi)
Vishwothama Nagar, BANTAKAL – 574 115, Udupi District, Karnataka, INDIA

Department of Artificial Intelligence & Data Science Engineering

---

CERTIFICATE

Certified that the Project Work titled '**Multiple Disease Prediction System'** is carried out by **Ms. Disha P U**, USN: **4MW21AD015, Ms. Hithashree**, USN: **4MW21AD019, Ms. Neha Dinesh Shettigar**, USN:**4MW21AD030, Ms. Nidhi Prabhu K**, USN: **4MW21AD032,** a Bonafide student of **Shri Madhwa Vadiraja Institute of Technology and Management**, in partial fulfillment for the award of the degree of **Bachelor of Engineering** in Artificial Intelligence and Data Science Engineering of Visvesvaraya Technological University, Belgaum during the year 2024-25. It is certified that all the corrections/ suggestions indicated for Internal Assessment have been incorporated in the report. The report has been approved as it satisfies the academic requirements in respect of Project Work prescribed for the said Degree.

Dr. Ganesha prasad	**Mr. Nagaraja Rao**	**Dr. Thirumaleshwara Bhat**
Project Guide	Head of the department	Principal
Dept. of AI & ML	Dept. of AI & DS	

External Viva

Name of the Examiners: **Signature with Date**

1.

2.

Acknowledgements

It is our privilege to express sincerest regards to our project guide **Dr. Ganesha prasad**, Assistant Professor (Selection Grade), Dept. of AI & ML, SMVITM, Bantakal for helping us in successful completion of this project work.

We would like to express our gratitude to **Prof. Dr. Thirumaleshwara Bhat**, Principal, SMVITM, Bantakal for extending his support.

We would like to thank **Prof. Dr. Nagaraja Rao**, HOD Associate Professor Dept. of AI & DS, for his inspiration during the completion of the project

We would also like to thank our project coordinator **Dr. Ranjan Kumar H S**, Associate Professor Dept. of AI & DS, SMVITM, Bantakal for helping us in successful completion of this project work.

We take this opportunity to express our deepest gratitude and appreciation to all those who helped us directly or indirectly towards the successful completion of this project.

We would like to thank our teaching and non-teaching staff, friends, who supported and encouraged us.

Disha P U
Hithashree
Neha Dinesh Shettigar
Nidhi Prabhu K

ABSTRACT

In recent years, the integration of machine learning techniques into healthcare systems has significantly advanced the capabilities of predicting and detecting various diseases. Our project mainly focuses on early detection of disease, ensuring robust model performance through advanced evaluation methods. This makes our project stand out from traditional disease prediction methods by employing tailored machine learning algorithms for various diseases. The foundation of the system lies in a comprehensive dataset encompassing a wide array of medical parameters crucial for each disease category. These parameters include demographic information, lifestyle factors, clinical indicators and various other symptoms that are known to influence the onset and progression of several diseases. To ensure the reliability and effectiveness of the predictive models, rigorous preprocessing techniques such as data cleaning, normalization, and feature selection are applied to refine the quality of the input data. There are many existing machine learning models related to health care which mainly focuses on detecting only one disease. Therefore, in this project we have developed a system to forecast several diseases by using a single user interface. The proposed model can predict multiple diseases such as diabetes, heart disease, chronic kidney disease, cancer etc. The main goal is to create a web application capable of forecasting several diseases. The designed web application employs the Streamlit Python library for frontend design and communicates with backend ML models to predict the probability of diseases.

Table of Contents

		Page No.
	Acknowledgement	**i**
	Abstract	**ii**
	Table of Contents	**iii-iv**
	List of Figures	**v-vi**
	List of Tables	**vii**
	List of Equations	**viii**
Chapter 1	**Introduction**	**1-2**
1.1	Introduction	1
1.2	Relevance of the work	1
1.3	Issues and challenges	2
Chapter 2	**Literature Review**	**3-7**
2.1	Detailed review of literature survey	3
2.2	Existing Gaps Identified in the Reviewed Papers	5
2.3	Problem statement	6
2.4	Objective	7
Chapter 3	**Requirement Specification**	**8-12**
3.1	Introduction	8
3.2	Hardware and Software Specifications	9
	3.2.1 Hardware Requirements	9

	3.2.2 Software Requirements	9
3.3	Technology Used	9
	3.3.1 Python	9
	3.3.2 Visual studio	10
	3.3.3 Streamlit	11
	3.3.4 Google colab	11
	3.3.5 Keras	12
Chapter 4	**Methodology**	**13-27**
4.1	Architecture of the system	13
	4.1.1 Data Collection and Preprocessing	14
	4.1.2 Disease-Specific Models and Techniques	15
Chapter 5	**Result**	**28-36**
5.1	Evaluation metrics	28
Chapter 6	**Conclusion**	**37-38**
6.1	Conclusion	37
6.2	Future scope	38
REFERENCE		**39**

List of Figures

		Page No.
Figure 3.1	Python virtual machine	**10**
Figure 4.1	Architecture of the system	**13**
Figure 4.2	Distribution of key features in the heart disease dataset highlighting demographic, clinical, and outcome patterns.	**16**
Figure 4.3	Correlation matrix highlighting relationships between features in the heart disease dataset for model optimization.	**17**
Figure 4.4	Distribution of demographic, lifestyle, and clinical features in the lung cancer dataset.	**18**
Figure 4.5	Correlation matrix highlighting feature relationships in the lung cancer dataset.	**18**
Figure 4.6	Feature distribution of chronic kidney disease dataset showing key patient characteristics and indicators.	**19**
Figure 4.7	Correlation heatmap of chronic kidney disease dataset highlighting feature relationships and target links.	**20**
Figure 4.8	Heatmap showing feature correlations in the liver disease dataset	**22**
Figure 4.9	Histograms of features showing skewed distributions and value concentrations.	**22**
Figure 4.10	Voting classifier	**24**
Figure 4.11	diabetes dataset correlation heatmap, highlighting key relationships for feature selection.	**24**
Figure 4.12	histograms of a diabetes dataset, revealing skewed distributions and concentrations in specific ranges.	**25**
Figure 5.1	Performance of the machine learning model in	**29**

	classifying heart disease cases	
Figure 5.2	Performance of the machine learning model in classifying diabetes disease cases	**30**
Figure 5.3	Performance of the machine learning model in classifying liver disease cases	**30**
Figure 5.4	Performance of the machine learning model in classifying lung cancer cases	**31**
Figure 5.5	Performance of the machine learning model in classifying kidney disease cases	**31**
Figure 5.7	Snapshot of Streamlit interface for symptoms based prediction	**33**
Figure 5.8	Snapshot of Streamlit interface displaying symptoms based prediction	**33**
Figure 5.9	Snapshot of Streamlit interface for pneumonia prediction	**34**
Figure 5.10	Snapshot of Streamlit interface for displaying lung cancer prediction	**34**
Figure 5.11	Snapshot of Streamlit interface displaying diabetes prediction	**35**
Figure 5.12	Snapshot of Streamlit interface for chronic kidney prediction	**35**
Figure 5.13	Snapshot of Streamlit interface for heart disease prediction	**36**
Figure 5.14	Snapshot of Streamlit interface for liver disease prediction	**36**

List of Tables

		Page No.
Table 5.1	Best model for each disease	**32**

List of Equations

		Page No.
Equation 4.1	Mathematical definition of logistic regression	**15**
Equation 4.2	Sigmoid function	**15**
Equation 4.3	Equation for combining the predictions	**23**
Equation 5.1	Accuracy formula	**28**
Equation 5.2	Precision formula	**28**
Equation 5.3	Recall formula	**28**
Equation 5.4	F1 score formula	**29**

Chapter 1

INTRODUCTION

In this chapter, we will introduce the subject, highlight the relevance of the work, and discuss the key issues and challenges involved.

1.1 Introduction

Machine learning involves programming computers to optimize performance using past data. Rapid advancement in machine learning and data science has made it possible to develop advanced models that predict multiple diseases with a high level of accuracy. The multi-disease prediction models make use of large datasets and sophisticated algorithms to analyze numerous medical parameters and predict the likelihood of various diseases.

This system combines various data sets to predict the risk of multiple diseases occurring simultaneously. Through the use of advanced machine learning algorithms, the system will be able to attain higher accuracy and interpretability in predictions. Such an approach not only allows for early and accurate diagnosis but also equips healthcare professionals with actionable insights for timely interventions and patient care strategies that are more personalized. As healthcare shifts towards more integration and data-driven approaches, multilayered systems will have a lot of promise in turning the tide of diagnostic modalities and improving patient results in general.

1.2 Relevance of the work

The application of machine learning in predicting multiple chronic conditions is gaining much attention given its ability to process the large and complex medical databases. Chronic diseases involve diabetes, heart disease, or liver disorders, and several of these share overlapping risk factors with symptoms. Therefore, if diagnosed early and predicted, they can be effectively handled. Machine learning models work on diverse data sources such as patient histories, lab results, or lifestyle factors to discover undetectable patterns and correlating them. Through such insights, ML-powered systems are able to predict multiple chronic conditions simultaneously. Thereby, health care providers will take measures to prevent diseases by adjusting the treatment plan appropriately to improve patient outcomes. In this way, an improvement in diagnostic accuracy and lowering of long-term burden on the healthcare system is achieved.

1.3 Issues and challenges

Traditional methods of disease diagnosis are often plagued with problems and challenges, such as relying on individual clinical expertise, lack of access to extensive patient information, and time-consuming manual diagnostic processes. Such methods may result in variable diagnoses, delayed detection, and poor treatment plans. Moreover, traditional diagnostic techniques may fail to effectively manage the complexity and interrelation of several diseases, leading to a fractured approach to care. An integrated multiple disease prediction system project can overcome all of these challenges by applying machine learning algorithms to analyze large volumes of patient data quickly and accurately, thus providing uniform early diagnosis of various chronic diseases. This integrated approach helps in enhancing the accuracy of diagnostic results, ensures timely intervention, and supports personalized treatment plans to improve overall patient outcomes and healthcare efficiency.

Chapter 2

LITERATURE REVIEW

In this chapter, we will be discussing the key findings which have been discovered from the various research papers reviewed. Additionally, we will identify the existing gaps, present the problem statement, and outline the objectives of the study.

2.1 Detailed review of literature survey

[1] Khan et al. (2018) used Support Vector Machines (SVM), Random Forests, and Logistic Regression to create a multi-disease prediction model. Their model showed good accuracy in predicting diabetes and heart disease after it was trained on an extensive dataset that included characteristics like age, gender, and medical history. The study's comparative analysis of various algorithms is its strongest point, but the results of intricate models like Random Forests are difficult to understand.

[2] Zhang et al. (2020) in order to predict diabetes and heart disease, compared Decision Trees, Random Forests, and SVM. Their results showed that because ensemble methods can handle feature interactions and reduce overfitting, they perform better than individual classifiers. In particular, Random Forest and Gradient Boosting perform better than individual classifiers. Nonetheless, the study identified one disadvantage with ensemble methods: their computational complexity

[3] Ahmed et al. (2021) [4] investigated hybrid models, which combine several algorithms to improve prediction robustness and accuracy. Their study showed that prediction results were greatly enhanced by combining deep learning methods with conventional machine learning models. The study's inventive methodology is commendable, however it was noted that the hybrid models' complexity and resource-intensiveness were drawbacks

[4] A framework for incorporating multi-disease prediction systems into clinical practice was presented by Chen et al. (2017). Their model made use of a single interface that let users enter health information and get disease predictions. The study underlined the usefulness of these systems in real-world applications but also underlined the necessity of thorough validation in various clinical settings to guarantee dependability.

[5] Sharma et al. (2020) used Streamlit to create an intuitive platform that predicts a variety of diseases. The study recognized that while the platform's real-time prediction capabilities and ease of use were noteworthy features, data privacy and security remain critical issues for healthcare applications.

[6] Wang et al. (2020) concentrated on enhancing the interpretability of multi-disease prediction models through the integration of explainable AI techniques. Their method promoted adoption and trust among medical professionals by providing clear decision-making procedures. Notwithstanding its advantages, the study pointed out that it is still difficult to strike a balance between interpretability and model accuracy.

[7] A novel method for enhancing multi-disease prediction through transfer learning was presented by Zhang et al. (2019). With less training data, their model was able to achieve higher accuracy by applying knowledge from related tasks. The study acknowledged the difficulties in choosing relevant source tasks and fine-tuning model parameters, but it also emphasized the potential of transfer learning.

[8] Sharma and Guleria (2023) reviewed deep learning methods, including CNNs like ResNet and DenseNet, for pneumonia detection using chest X-ray images. Their study highlighted the superior accuracy and efficiency of these methods in feature extraction and classification compared to traditional techniques. However, challenges such as dataset bias, limited model explainability, and overfitting were noted. The authors emphasized the importance of diverse datasets, interpretable models, and seamless clinical integration. They proposed future directions like hybrid methods, federated learning, and addressing ethical issues such as bias and fairness, offering valuable insights for advancing AI-based pneumonia detection.

[9] Khan et al. (2021) reviewed deep learning models, including CNNs, ensemble methods, and transfer learning, for pneumonia detection using chest X-rays. The study evaluated popular architectures such as ResNet and VGGNet, highlighting their high accuracy but also challenges like dataset biases, limited diversity, and lack of interpretability. It noted that while these models perform well, they often struggle with generalization. The authors

emphasized the need for interpretable AI, diverse datasets, and clinical workflow integration, concluding that addressing these challenges is crucial for realizing the full potential of AI in pneumonia detection.

[10] Hafsa Binte Kibria et al. (2023) proposed an accurate and interpretable method for diabetes prediction using a soft voting ensemble of Random Forest, AdaBoost, and Gradient Boosting models. The approach achieved 90% accuracy and an F1 score of 89% on the Pima Indian Diabetes Dataset. Explainable AI techniques like SHAP were used to identify key features influencing diabetes risk, ensuring transparency in predictions. The study showed that ensemble methods are more robust and reliable than individual classifiers, making the approach suitable for clinical use. It bridges the gap between AI models and their application in healthcare decision-making.

2.2 Existing Gaps Identified in the Reviewed Papers:

The existing disease-prediction systems are designed mainly for single diseases, which are mainly limited in their scope, accuracy, and applicability. For example, Rule-Based Systems use predefined mappings of symptom-disease mappings or expert systems, but they do not seem flexible and lack the adaptation for new diseases or datasets. The models for the single disease will predict only a particular disease like diabetes or cardiovascular disease by employing machine learning algorithms like Logistic Regression or Decision Trees on disease-specific datasets. A significant number of healthcare applications allow symptom-based predictions, yet they only use simple algorithms or decision trees and can't handle complex data types such as imaging or historical trends. The major drawbacks of the present system are that it only limits itself to certain diseases. The lack of integration with diverse medical data such as imaging, lab results, and patient history. Mostly requires manual intervention for precise predictions.

- Model Interpretability and Complexity: Several studies (e.g., Khan et al., 2018; Wang et al., 2020; Ahmed et al., 2021) highlighted the challenge of balancing model interpretability and complexity. Advanced models like Random Forest and hybrid approaches offer high accuracy but lack transparency, which limits trust and adoption in clinical settings.

- Dataset Bias and Diversity: Papers such as Sharma and Guleria (2023) and Khan et al. (2021) emphasized dataset bias and limited diversity as significant barriers. Models trained on specific populations or datasets often struggle to generalize across diverse demographics or clinical environments.
- Computational Complexity: Zhang et al. (2020) and Ahmed et al. (2021) noted that ensemble methods and hybrid models, while powerful, are computationally intensive and resource-demanding. This complexity hinders their scalability and deployment in resource-constrained settings.
- Data Privacy and Security: Sharma et al. (2020) identified privacy and security as critical concerns, especially for real-time prediction platforms in healthcare. Ensuring compliance with data protection regulations is a key challenge for practical implementations.
- Integration into Clinical Practice: Chen et al. (2017) and Khan et al. (2021) highlighted the lack of seamless integration of predictive systems into clinical workflows. Thorough validation across diverse clinical settings is necessary to establish their reliability.
- Limited Use of Explainable AI: While Wang et al. (2020) and Hafsa Binte Kibria et al. (2023) utilized explainable AI, many studies lack mechanisms to provide insights into predictions. This gap limits the acceptance of these systems by healthcare professionals.
- Ethical and Fairness Concerns: Sharma and Guleria (2023) pointed out ethical issues like bias and fairness in AI models, which remain unresolved. Addressing these concerns is critical to avoid discrimination and ensure equitable healthcare delivery.

2.3 Problem statement

The healthcare industry faces increasing challenges due to the rising number of diseases. To address this, a comprehensive solution is needed—leveraging a machine learning system that integrates multiple models. Such a system would analyze patient data, including historical medical records, diagnostic information, and X-ray images, to accurately predict the presence of various diseases.

2.4 Objective

- To develop robust machine learning models inorder to predict the likelihood of multiple diseases.
- To evaluate and compare the performance of different machine learning models in predicting multiple diseases.
- To implement data visualization tools to display the relationships and distributions of key features related to each disease.
- To create an intuitive and accessible web application that allows users to input their medical data and receive disease predictions easily.

Chapter 3

REQUIREMENT SPECIFICATION

In this chapter, we will outline the requirements essential for the study, including an introduction to the specifications. Additionally, we will detail the hardware and software requirements necessary to achieve the objectives of the work.

3.1 Introduction

The Multiple Diseases Prediction System is a cutting-edge software program created to use sophisticated machine learning algorithms to help with the early identification of a number of diseases. In addition to offering actionable insights and recommendations, the system allows users to enter symptoms, medical history, and diagnostic test results. It also makes predictions regarding the possibility of several diseases. The increasing demand for easily available and effective healthcare solutions that support early disease identification and management is met by this system. It is intended for usage by a wide range of users, such as researchers looking at illness trends, people hoping to better understand their health, and medical professionals looking for assistance with diagnoses. The system improves patient outcomes, lowers diagnostic errors, and promotes well-informed healthcare decision-making by utilizing data-driven techniques. Predicting diseases including diabetes, heart problems, liver problems, and cancer is part of the project's scope. To deliver precise results, it combines supervised and unsupervised learning methods. The system is widely usable and convenient because it is available on both online and mobile platforms. The system's design prioritizes scalability, security, and usability. It has an easy-to-use interface, strong data encryption to safeguard private data, and the capacity to manage many users and data at once. High accuracy and relevance are maintained throughout time by implementing frequent model changes and training with fresh datasets.

The Multiple Diseases Prediction System's ultimate objective is to close the gap between sophisticated medical diagnostics and approachable technology, giving patients and healthcare professionals dependable resources for preventative health care. This approach is a supplemental tool to improve diagnosis speed and accuracy, not a substitute for expert medical advice.

3.2 Hardware and Software Specification

3.2.1 Hardware requirements

- Server-Side:
  - Processor: 16-core CPU (e.g., Intel Xeon or AMD EPYC).
  - RAM: 64 GB or higher.
  - Hard Disk: 2 TB SSD for high-speed operations + 4–8 TB HDD for backups.
  - GPU: NVIDIA Tesla A100 or equivalent (24 GB VRAM).

- Client-Side :
  - Processor: Dual-core CPU (e.g., Intel i3 or AMD Ryzen 3).
  - RAM: 4 GB or higher.
  - Hard Disk: 500 MB free space.

3.2.2 Software requirements

Backend Framework: Python (Django/Flask).

Machine Learning Framework: TensorFlow/PyTorch.

Database: PostgreSQL or MongoDB.

Operating System: Linux (Ubuntu 20.04 or later).

Version Control: Git (GitHub/GitLab).

3.3 Technology Used

- Python
- Visual studios
- Streamlit
- Google colab
- Keras

3.3.1 Python

One of the key strengths of Python is its vast ecosystem of third-party libraries and frameworks, which makes it suitable for a wide range of domains, including web

development, data analysis, artificial intelligence, machine learning, scientific computing, automation, and more. Libraries such as NumPy, Pandas, and TensorFlow have contributed to Python's popularity in fields like data science and AI.

The scope of the Multiple Disease Prediction System involves creating an intelligent platform that leverages deep learning and machine learning techniques to predict various diseases accurately. By integrating patient data, medical images, and advanced algorithms, the system aims to provide precise and timely predictions, improving healthcare outcomes and aiding in better decision-making. Python's system design follows a layered architecture that ensures efficient processing and execution of applications. Python programs are written in a high-level, human-readable syntax, which is then compiled into bytecode—a platform-independent intermediate format. This bytecode is executed by the Python Virtual Machine (PVM), which abstracts the hardware and operating system while interpreting the code line by line. Unlike other programming languages, Python does not require explicit compilation, as the interpreter automatically converts source code into bytecode during runtime, simplifying the development process and enhancing flexibility.

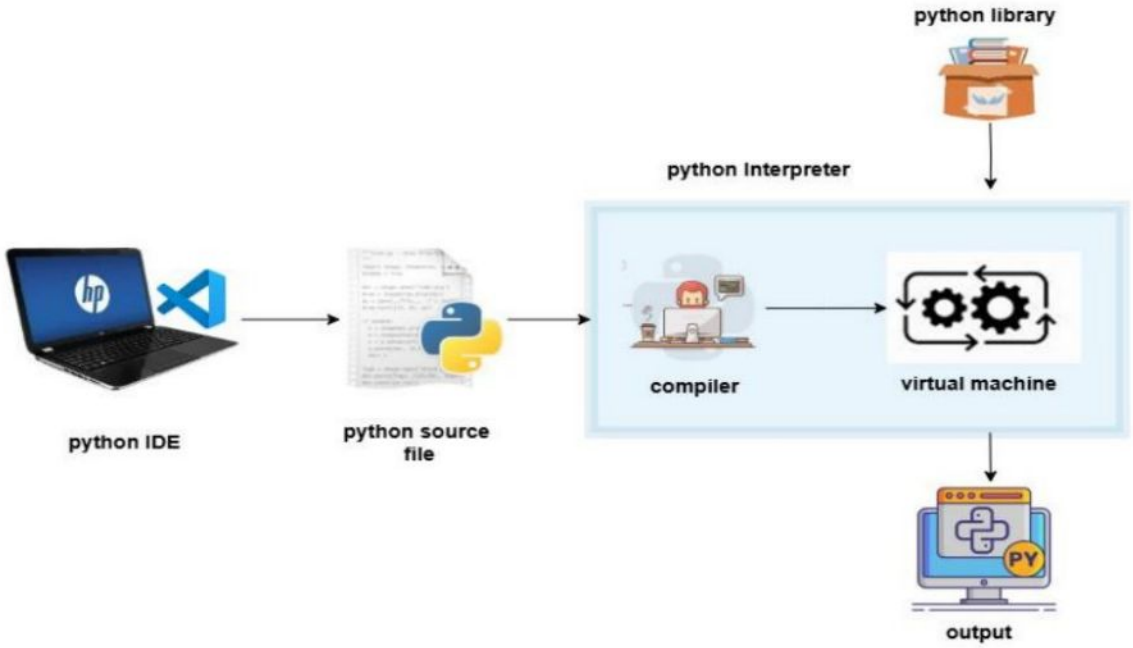


Figure 3.1 Python virtual machine

3.3.2 Visual studio

Visual Studio Code (VS Code) plays a crucial role in our project by serving as a versatile and efficient code editor for developing, debugging, and deploying machine learning

models. Its lightweight nature, combined with a rich ecosystem of extensions, allows seamless integration with Python, TensorFlow, and other libraries essential for our multi-disease prediction system. Features like IntelliSense, syntax highlighting, and Git integration enhance coding efficiency and collaboration within the team. Moreover, VS Code's integrated terminal simplifies running scripts and managing dependencies, while its debugging tools streamline the testing process. This makes it an ideal environment for developing both the backend ML models and the frontend web application using Streamlit.

3.3.3 Streamlit

Streamlit is an essential component of our project, providing a user-friendly framework to build interactive and visually appealing web applications for our multi-disease prediction system. It allows seamless integration between the backend machine learning models and the user interface, enabling real-time predictions based on user inputs. With its simplicity and intuitive Python-based syntax, Streamlit eliminates the need for extensive web development knowledge, allowing the team to focus on the functionality and accuracy of the prediction models. Additionally, features like live visualization, sliders, and file upload options make it easy for users to interact with the system, such as uploading X-ray images for pneumonia detection or entering symptoms for disease prediction. This ensures accessibility, efficiency, and scalability, making it an ideal choice for our project.

3.3.4 Google Colab

Google Colab is vital for our project as it provides a cloud-based platform with access to free GPUs and TPUs, essential for training complex deep learning models like those used for pneumonia detection and other disease predictions. It eliminates the need for high-performance local hardware, enabling efficient development and experimentation. Colab's collaborative environment allows multiple team members to work simultaneously, share code, and make real-time updates, enhancing productivity. The integration with Google Drive ensures seamless data storage and retrieval, while pre-installed libraries reduce setup time. With its ability to handle large datasets and execute computationally intensive tasks, Colab is an indispensable tool for building and testing our machine learning models.

3.3.5 Keras

In this project, Keras is not directly used as the primary framework since PyTorch is the main deep learning library employed. However, for future enhancement or integration, Keras could be an ideal choice for building and deploying deep learning models due to its user-friendly API and ease of use. Keras is a high-level neural networks API, written in Python, that operates on top of lower-level libraries like TensorFlow or Theano, providing a simple interface to design and train deep learning models.

By abstracting complex operations, Keras allows rapid prototyping and experimentation with neural network architectures. With its intuitive design, Keras facilitates quick model development, making it an ideal tool for tasks such as image classification, object detection, and other machine learning applications. If integrated into this coconut classification system, Keras would streamline the training of custom models by enabling fast model building, optimization, and deployment.

Chapter 4

METHODOLOGY

In this chapter, we will describe the methodology adopted for the study, focusing on the system architecture. Furthermore, we will discuss the various models utilized for the project to achieve the desired outcomes.

4.1 Architecture of the system

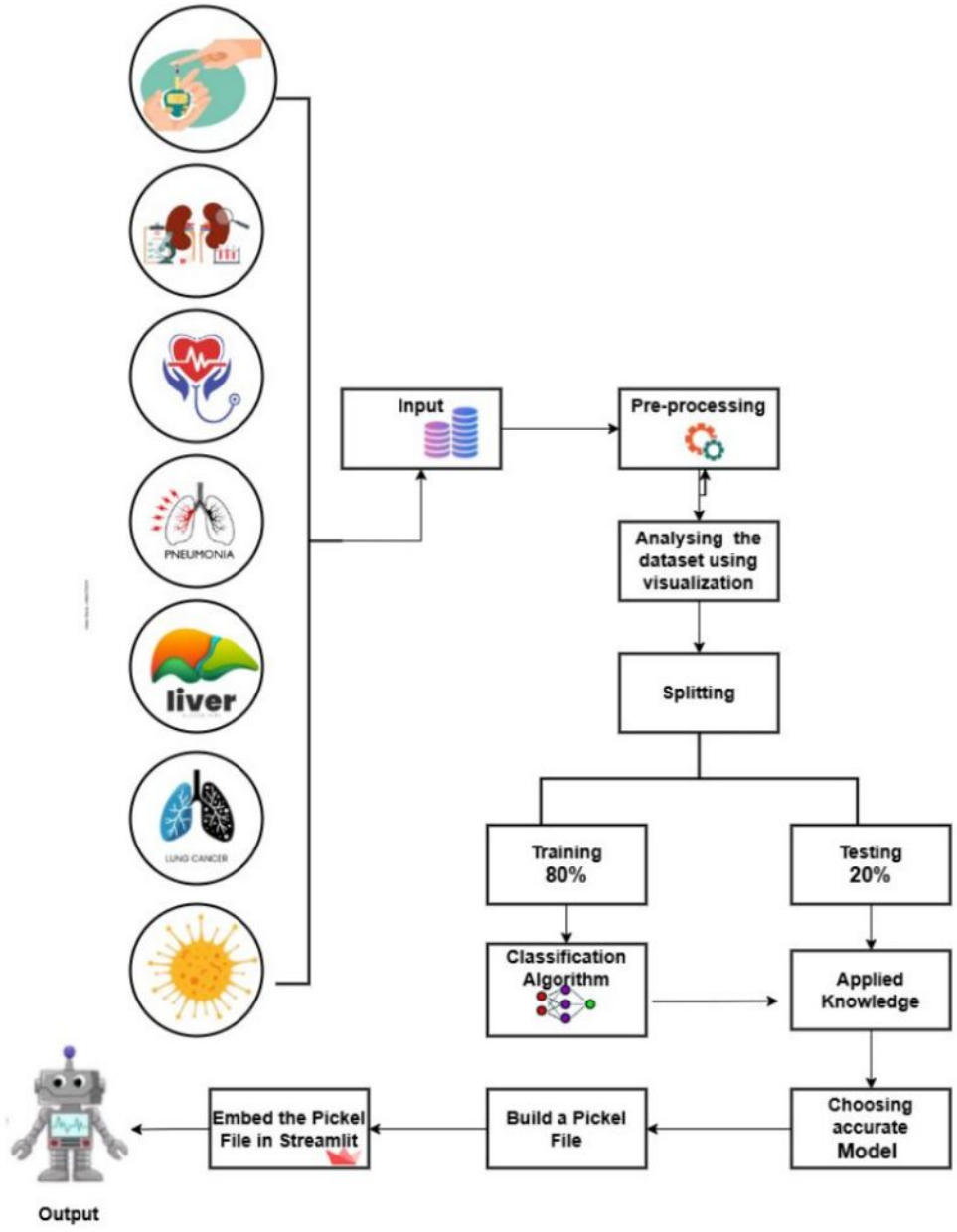


Figure 4.1 Architecture of a system

The Multiple Disease Prediction System is an advanced application that utilizes the latest machine learning and deep learning technologies. It is designed to analyze patient data, symptoms, and medical imaging to deliver highly accurate predictions for various diseases. The system's primary goal is to enable early diagnosis and support healthcare professionals in making informed decisions, ultimately improving patient outcomes. Below is a detailed explanation of how the system works:

4.1.1. Data Collection and Preprocessing

a) Data Collection

Data collection involves sourcing relevant data from credible sources to train and test the machine learning and deep learning models. Each disease requires specific data:

- Heart Disease: Patient demographics, cholesterol levels, ECG readings, exercise test results, and lifestyle information.
- Diabetes: Clinical measurements such as glucose levels, insulin sensitivity, family history, and BMI values.
- Pneumonia: Chest X-rays or CT scans to visually identify lung infections.
- Liver Disease: Blood test results, alcohol consumption history, and genetic predisposition factors.
- Lung Cancer: Imaging reports (e.g., CT or MRI scans) and symptoms like prolonged cough or chest pain.
- Symptom-based Diseases: Text-based patient symptoms for diseases like flu, allergies, or chronic illnesses.

Sources for these datasets include:

- Hospitals: Electronic Health Records (EHRs).
- Public Repositories: Kaggle, UCI Machine Learning Repository.
- Surveys: Health data collected via online or offline surveys.

b) Preprocessing

Preprocessing ensures the data is clean and usable:

- Data Cleaning: Removes duplicate or irrelevant records and fills in missing values using techniques like mean substitution or predictive modeling.
- Feature Selection: Identifies the most impactful variables (e.g., glucose levels for diabetes).
- Normalization: Converts values to a consistent scale, such as scaling cholesterol

levels between 0 and 1.

- Image Preprocessing: For X-rays or scans, techniques like resizing, grayscale conversion, and augmentation (rotation, flipping) improve model robustness.
- Encoding: Converts categorical variables (e.g., gender: male/female) into numerical representations.

4.1.2. Disease-Specific Models and Techniques

a) Logistic Regression for Heart, chronic kidney, lung cancer prediction

Logistic regression was the most effective model with the highest accuracy being 88%. Logistic regression is a machine learning algorithm designed for binary classification tasks, making it highly suitable for predicting whether a patient has heart disease. Its efficiency and ease of interpretation are key reasons for its widespread use in medical applications.

Description of the Model: Logistic regression is a way to predict the probability that an event will occur for a binary classification problem, such as whether or not a patient has heart disease, using the input features, such as cholesterol levels and blood pressure. The logistic function, sigmoid curve, transforms linear combinations of input features into a probability that then gets converted into the output class (0 or 1).

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Equation 4.1 Mathematical definition of logistic regression

Z: Linear combination (log-odds).

Inputs or features: X1, X2 …Xn; say cholesterol and blood pressure

Intercept or bias: β_0

Coefficient to the features: β_1, β_2….βn

Compute probability: The linear combination (Z) is passed through the sigmoid function in order to map the value from 0 to 1:

$$P(y = 1|X) = \frac{1}{1 + e^{-z}}$$

Equation 4.2 Sigmoid function

P(y=1 |X): Probability of a patient suffering from the heart disease

Thresholding for Prediction: The probability output is passed by comparing it against the pre-defined threshold (0.5):

- If $P(y=1|X) \geq 0.5$, the model forecasts the patient has heart disease (class=1).
- If $P(y=1|X) < 0.5$, the model classifies no heart disease (class = 0).

Training Model (Learning Coefficients): When training, the model adjusts coefficients (β) on optimum values such that difference in predicted probabilities and correct outputs is minimized through something known as maximum likelihood estimation, which means the best predictive model for the inputs within the training data would have been produced.

Interpretation of Results: Here the coefficients (β) indicate how much a change in a feature affects the probability of heart disease. For example, a positive coefficient for cholesterol suggests that higher cholesterol increases the likelihood of heart disease. The intercept (β0) represents the baseline probability when all features are zero.

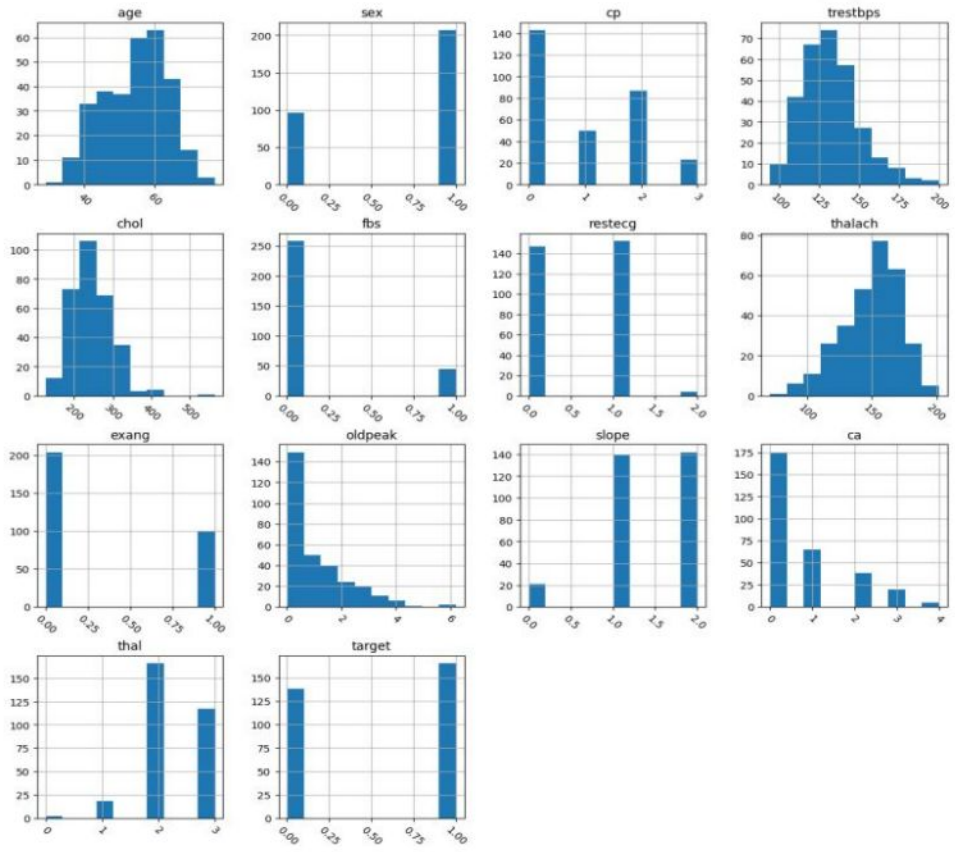


Fig. 4.2 Distribution of key features in the heart disease dataset highlighting demographic, clinical, and outcome patterns.

In the above figure (fig.4.2), the heart disease dataset reveals that most patients are aged 40–60, with a predominance of males. Key features like chest pain, cholesterol, and blood pressure show significant variability, reflecting diverse risk profiles. Fasting blood sugar and exercise-induced angina highlight distinct groupings, while ST depression and its slope provide diagnostic signals.

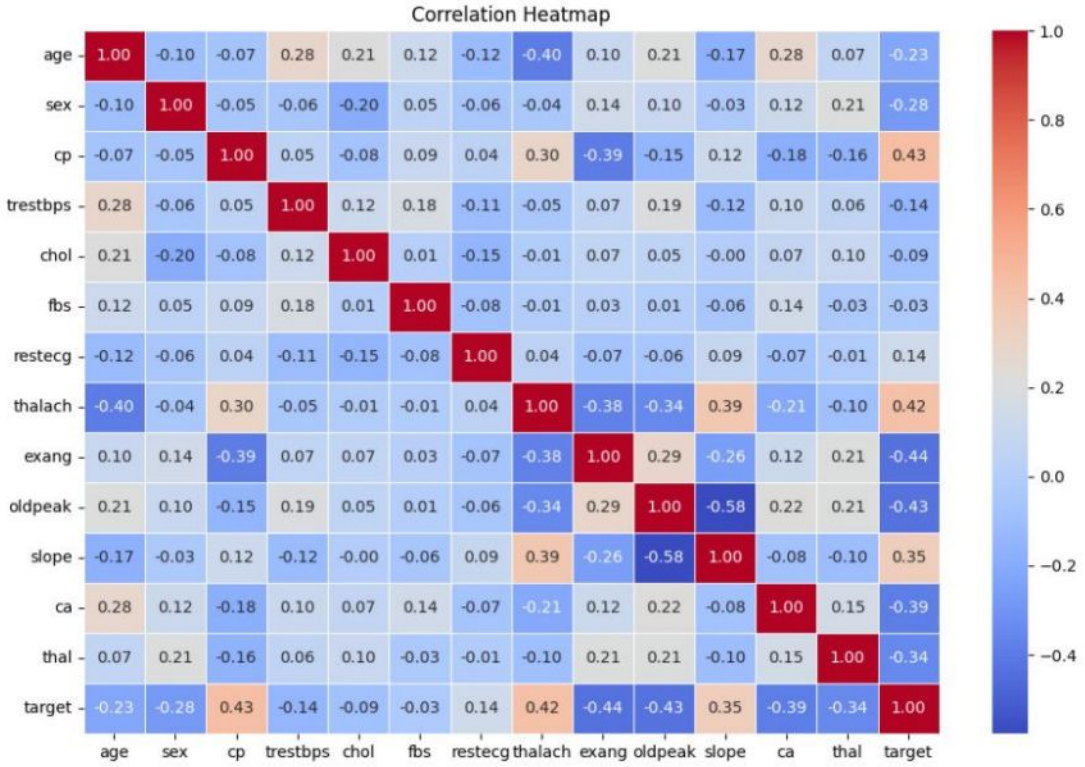


Fig. 4.3 Correlation matrix highlighting relationships between features in the heart disease dataset for model optimization.

The above correlation matrix (fig.4.3) highlights the relationships between features in the heart disease dataset, with values ranging from -1 (negative correlation) to +1 (positive correlation). Key insights include strong negative correlations of exercise-induced angina (exang) and maximum heart rate (thalach) with the target variable, suggesting their significance in heart disease prediction. Features such as oldpeak and ca (number of major vessels) also show notable negative correlations with the target. Positive correlations, though weaker, are observed for cp (chest pain type) and slope (ST segment slope). These relationships help identify influential features for predictive modeling and optimization.

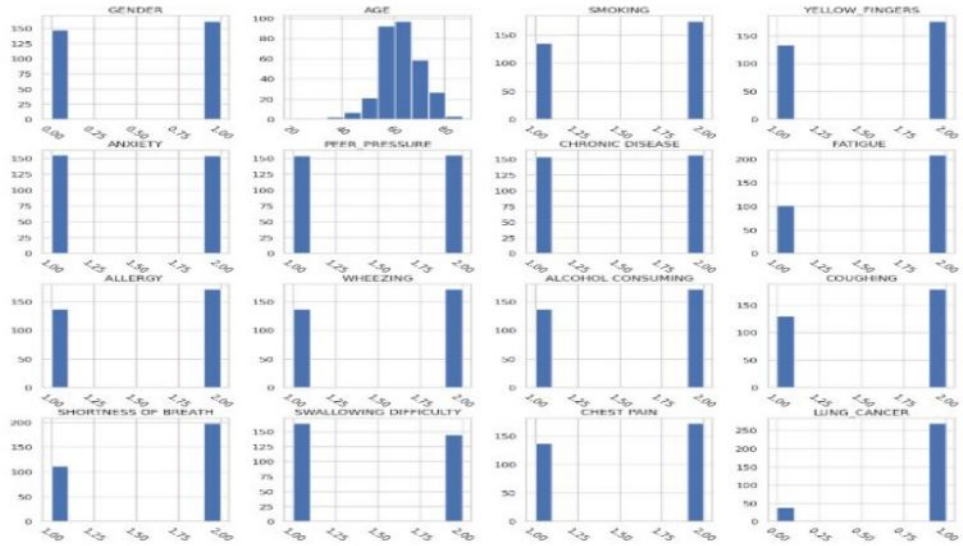


Fig. 4.4 Distribution of demographic, lifestyle, and clinical features in the lung cancer dataset.

The above plot (fig.4.4) displays the distribution of various features related to lung cancer. The age distribution shows most individuals are in the middle-aged category. Categorical features like gender, smoking, and alcohol consumption reveal distinct groupings, with smoking and chronic disease being prominent factors. Symptoms such as coughing, shortness of breath, chest pain, and wheezing occur frequently, reflecting common indicators of lung cancer. Behavioral factors like peer pressure and conditions like fatigue, yellow fingers, and swallowing difficulty also show strong representation, suggesting their relevance. These distributions highlight key risk factors and symptoms, aiding in understanding and modeling lung cancer prediction.

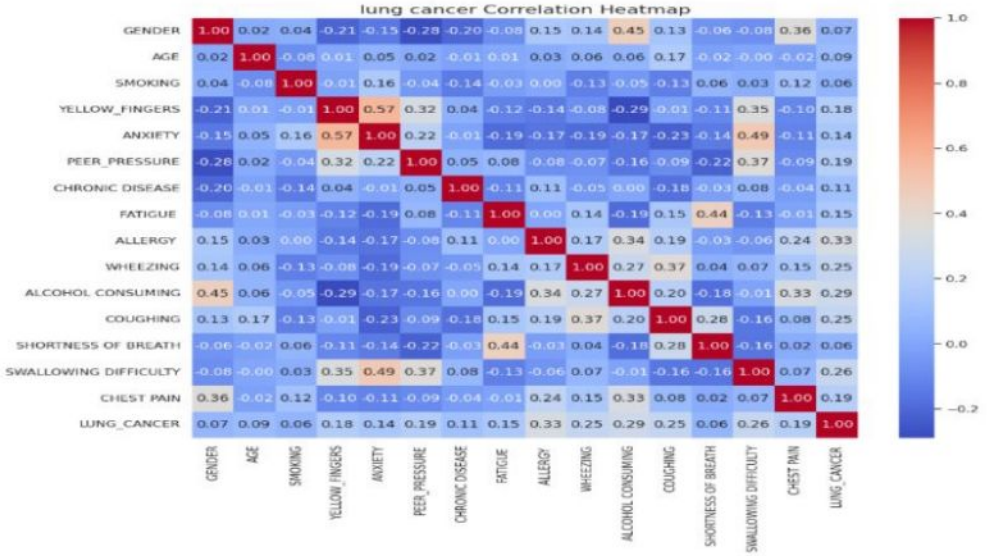


Fig. 4.5 Correlation matrix highlighting feature relationships in the lung cancer dataset.

The correlation heatmap (fig.4.5) illustrates the relationships between features in the lung cancer dataset, with correlation values ranging from -1 (negative) to +1 (positive). Key insights include a strong positive correlation between symptoms like wheezing, shortness of breath, and chest pain, reflecting their frequent co-occurrence in lung cancer cases. Swallowing difficulty and fatigue show moderate correlations with the target variable, indicating their importance in predicting lung cancer. Behavioral factors such as smoking and alcohol consumption exhibit lower correlations with lung cancer directly but influence other features. The matrix helps identify influential features and their interactions, guiding feature selection and model optimization.

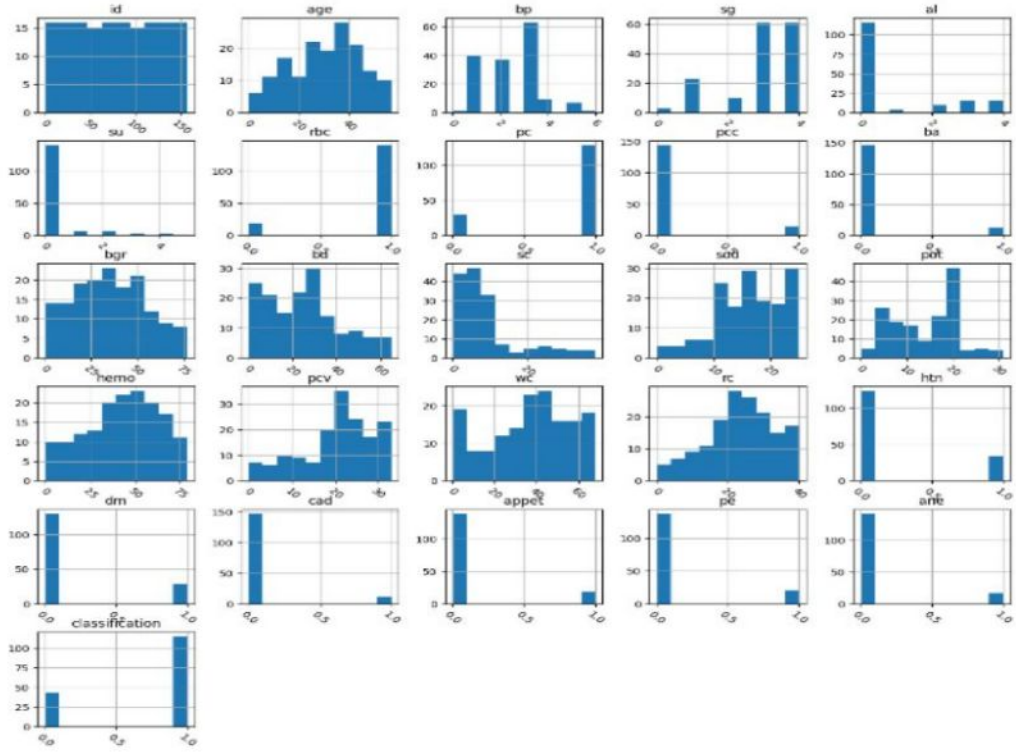


Fig 4.6 Feature distribution of chronic kidney disease dataset showing key patient characteristics and indicators.

The above plot (fig.4.6) showcases the distributions of various features in the chronic kidney disease dataset. Attributes like age, blood pressure (bp), hemoglobin (hemo), and serum creatinine (sc) exhibit varied distributions, reflecting diverse patient profiles. Features such as sugar (su), albumin (al), and diabetes mellitus (dm) show skewed distributions, indicating their higher prevalence in specific subgroups. These insights provide a foundation for feature selection and preprocessing for predictive modeling.

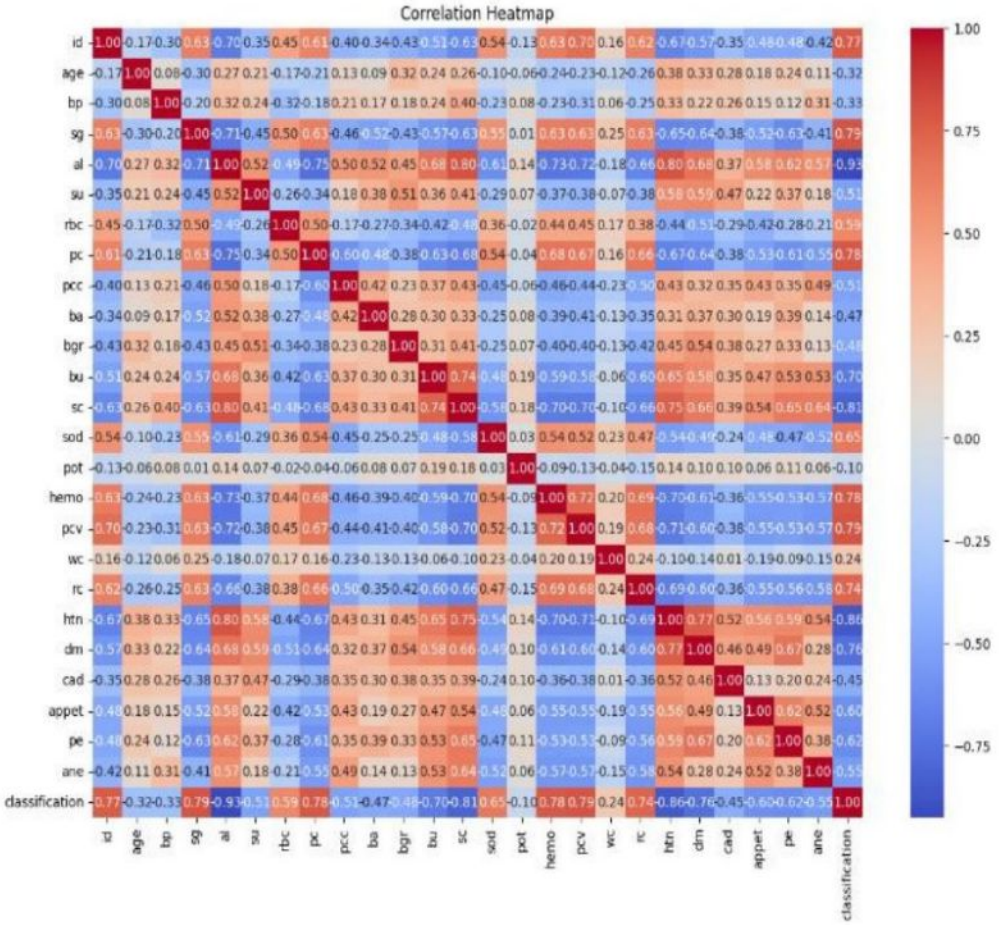


Fig.4.7 Correlation heatmap of chronic kidney disease dataset highlighting feature relationships and target links.

The correlation heatmap (fig.4.7) illustrates the relationships between features in the chronic kidney disease dataset, with correlation values ranging from -1 (strong negative correlation) to +1 (strong positive correlation). Strong correlations, such as between blood pressure (bp) and hypertension (htn), or serum creatinine (sc) and blood urea (bu), highlight interdependencies that may influence disease progression. Additionally, the classification column's correlations with features like pcv (packed cell volume) and hemo (hemoglobin) indicate their significance in predicting the target variable. This visualization helps detect multicollinearity and guides feature selection for model optimization.

b) <u>Stacking classifier for Liver Disease prediction</u>

A stacking classifier is an ensemble learning method that combines predictions from multiple base models, also known as weak learners, using a meta-model. The base models make predictions independently, and the meta-model learns to combine these predictions to make the final prediction. In this case:

- Base Models: Logistic Regression and SVM
- Meta-Model: Random Forest

The process involves two main stages:

Training the Base Models: Logistic regression and SVM are trained on the input features, for example, levels of liver enzymes, bilirubin, albumin, etc.

Training the Meta-Model: Predictions from the base models become new features to train the random forest model.

Each base model produces probabilities indicating the probability of liver disease: Base Model 1(Logistic Regression) outputs the probability PLR(y=1|X) by using the sigmoid function whereas the base Model 2(SVM) outputs the probability PSVM(y=1|X) by using decision boundaries. The meta-model (random forest) takes these probabilities as inputs to compute the final prediction. The final output is a probability $P_{Stack}(y=1|X)$, which is thresholded at 0.5 for binary classification:

- If $P_{Stack}(y=1|X) \geq 0.5$, the model predicts liver disease (class = 1).
- If $P_{Stack}(y=1|X) < 0.5$, the model predicts no liver disease (class = 0).

Example: Patient's health metrics (for example, liver enzymes, bilirubin) are fed into the system.

- Base Models: PLR(y=1|X) =0.65 is computed by logistic regression. PSVM(y=1|X) =0.72 is computed by SVM.
- Meta-Model: Random forest takes these probabilities as inputs and computes $P_{Stack}(y=1|X)$ =0.78.

Since $P_{Stack}(y=1|X)$ =0.78≥0.5, the system predicts liver disease.

It merges the strengths of multiple models, thereby reducing overfitting and underfitting for better generalization. This hybrid model can handle both linear (logistic regression) and non-linear (SVM) relationships, with meta-model enhancing the predictions for a better accuracy. Thus, it forms a robust and reliable system for liver disease diagnosis with an accuracy of about 76%.

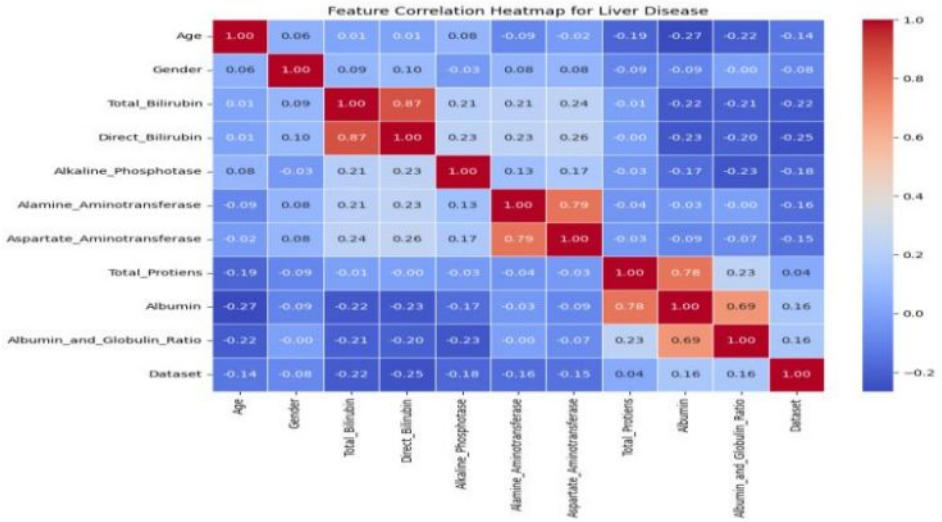


Fig.4.8 Heatmap showing feature correlations in the liver disease dataset

Figure 4.8 shows a heatmap depicting correlations between features in a liver disease dataset, with red indicating strong positive correlations and blue representing negative ones. Key patterns include a high positive correlation between Total_Bilirubin and Direct_Bilirubin, and between Albumin and Total_Proteins, reflecting their close biochemical relationships.

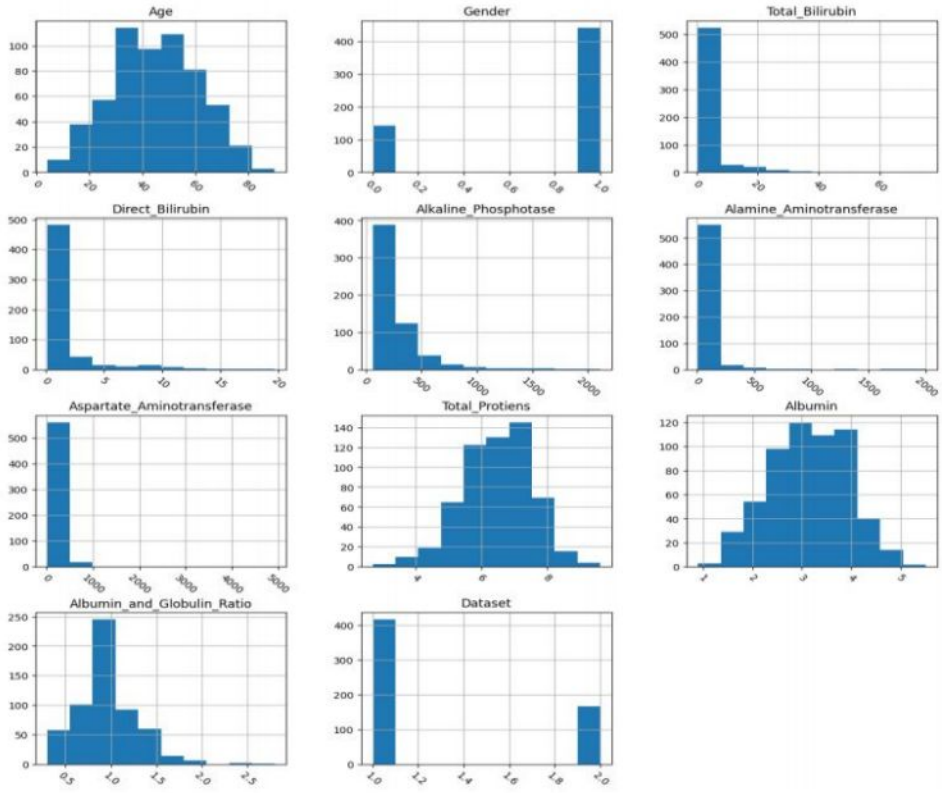


Fig. 4.9 Histograms of features showing skewed distributions and value concentrations.

The plot (fig.4.9) displays histograms of features like age, gender, and liver biomarkers, showing skewed distributions. Some variables, such as gender and total bilirubin, are concentrated around specific values, indicating imbalances in the dataset that may require adjustments for analysis.

c) Voting classifier for Diabetes Prediction

For the purpose of prediction of diabetes, an ensemble technique of voting classifier has been used, which provides the strength of various base models, thus giving higher accuracy with robustness. In this regard, three Support Vector Machines (SVM) were used with three different kinds of kernels, such as linear, polynomial, and radial basis function (RBF). By utilizing these variant kernels, the voting classifier improved the predictions and gave a better generalized result.

A voting classifier is an ensemble learning technique that combines multiple base models' predictions with either hard or soft voting to give the final output. In the case of diabetes diagnosis, there are:

- Base Models: SVM with linear, polynomial and RBF kernels
- Combination Method: Soft voting (uses predicted probabilities)

The process consists of two stages:

Training the Base Models: The three SVM models are trained on their respective input features.

- Linear Kernel SVM: Good for data that might be linearly separable; the classes can be separated using a straight hyperplane.
- Polynomial Kernel SVM: It catches non-linear patterns by mapping input data into a higher-dimensional space using polynomial functions.
- RBF Kernel SVM: Maps the input data into an infinite-dimensional space using radial basis functions and handles complex non-linear relationships.

Combining Predictions: Each SVM model outputs a probability: $P_{Linear}(y=1|X)$, $P_{Poly}(y=1|X)$, and $P_{RBF}(y=1|X)$, representing the likelihood of diabetes. The voting classifier computes the average of these probabilities:

$$P_{\text{Voting}}(y = 1 \mid X) = \frac{P_{\text{Linear}}(y=1|X)+P_{\text{Poly}}(y=1|X)+P_{\text{RBF}}(y=1|X)}{3}.$$

Equation 4.3 Equation for combining the predictions

A threshold of 0.5 is applied for binary classification: If $P_{Voting}(y=1|X) \geq 0.5$, the model predicts diabetes (class = 1) and if $P_{Voting}(y=1|X) < 0.5$, the model predicts no diabetes (class = 0).

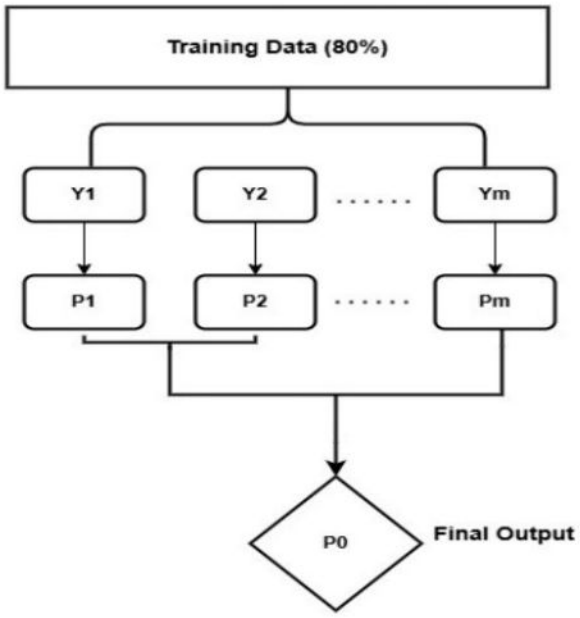


Figure 4.10 Voting classifier

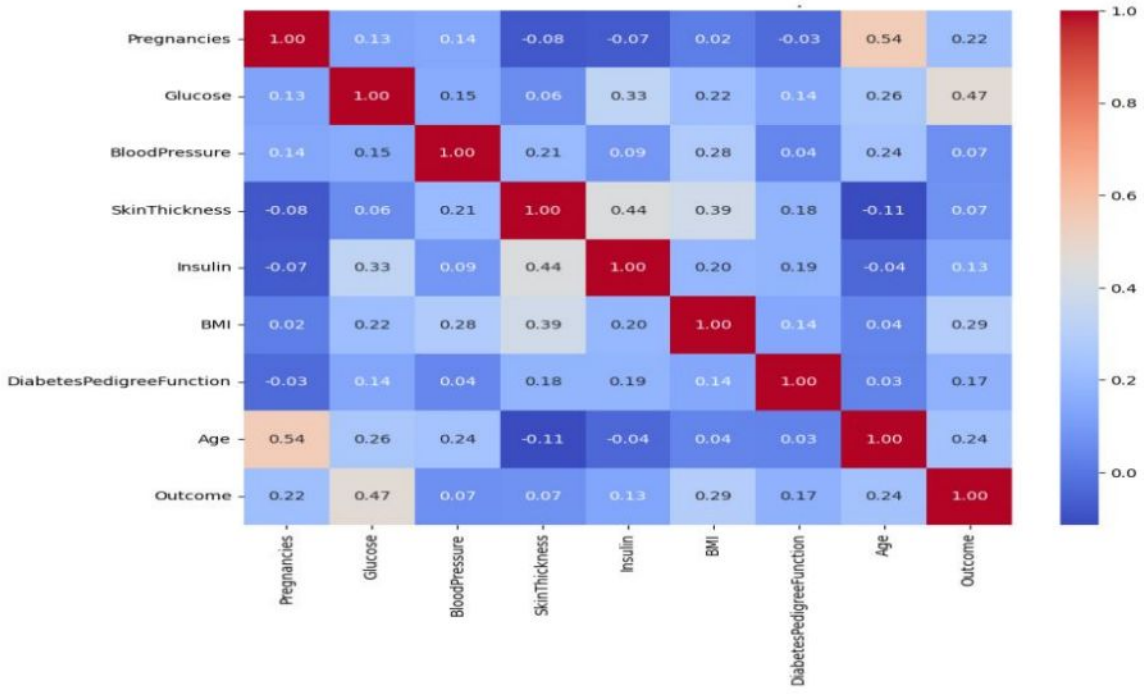


Fig 4.11 diabetes dataset correlation heatmap, highlighting key relationships for feature selection.

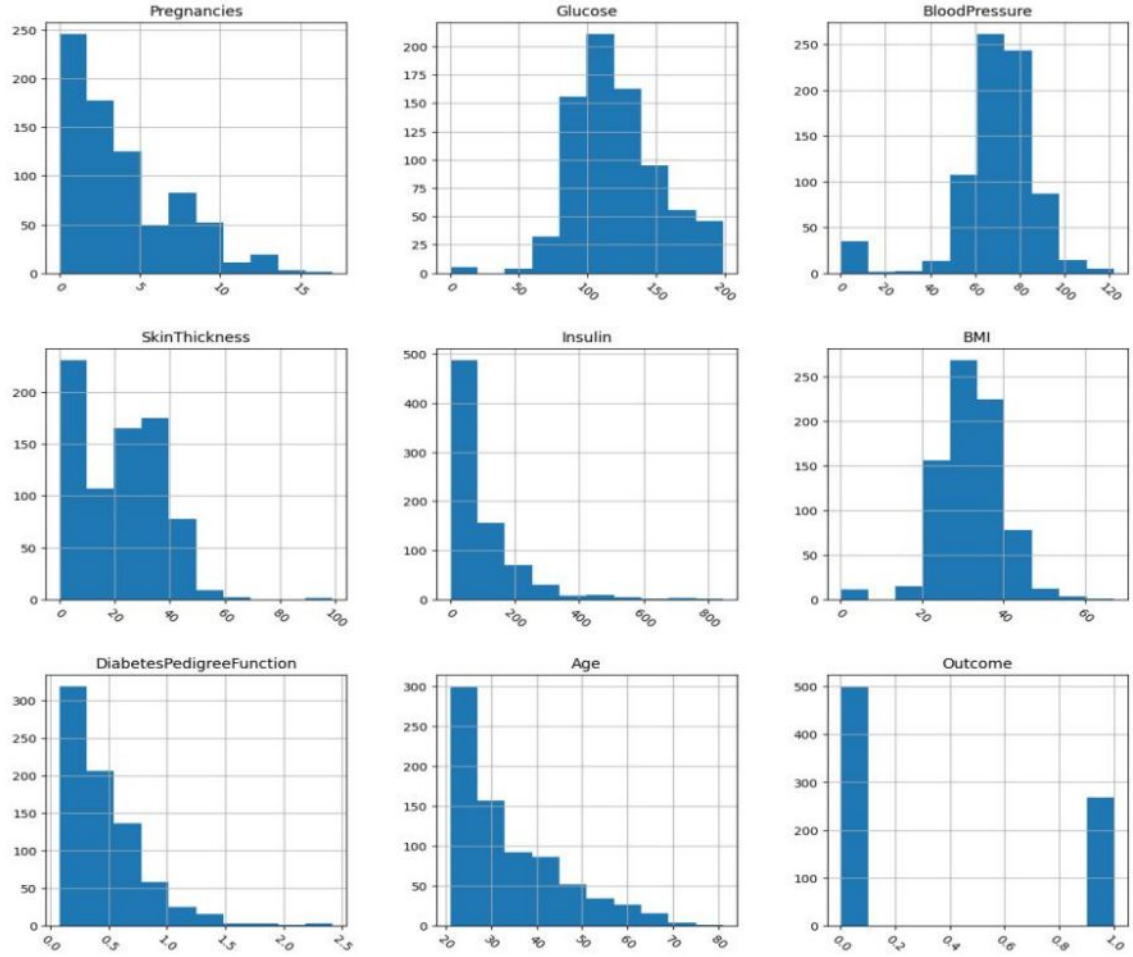


Fig. 4.12 histograms of a diabetes dataset, revealing skewed distributions and concentrations in specific ranges.

The plot (fig.4.12) shows histograms of features in a diabetes dataset, such as pregnancies, glucose, and insulin, highlighting skewed distributions with concentrations in specific ranges (e.g., glucose around 100-150). These patterns reveal potential preprocessing needs like scaling or normalization.

d) <u>CNN for pneumonia prediction</u>

CNNs are among the most successful deep learning models in the classification of images, and one of the most useful ones for applications in medicine. When detecting pneumonia, CNNs work great to scan chest X-rays to look for patterns associated with the disease. This kind of sophisticated feature extraction and classification capability leads to very accurate and dependable predictions, lessening the burden of diagnosis on medical professionals.

The main components of the model are:

- Convolutional Layers: Extract edges, textures and shapes from the input image.
- Pooling Layers: Dimensionality reduction: retains most important features
- Fully Connected Layer: Feature combination for a final classification
- Softmax Output: A probability indicating whether x-ray is of pneumonia or not.

For training, the CNN works to minimize its weights and biases by a process called backpropagation. These involve:

- Input: Chest X-ray images are passed into the model.
- Feature Extraction: Convolutional layers are used to apply filters and extract visual patterns indicative of pneumonia.
- Pooling: Dimensions are reduced but retain significant features.
- Classification: Fully connected layers and a Softmax function output the probability of the presence of pneumonia.

The loss function, for example cross-entropy loss, is minimized using an optimization algorithm like Adam or Stochastic Gradient Descent (SGD).

The Softmax function computes the two probabilities $P(y=1|X)$ (pneumonia) and $P(y=0|X)$ (no pneumonia).

- If $P(y1|X) >= 0.5$, the system predicts the presence of pneumonia class=1
- If $P(y1|X) < 0.5$, then the system decides that the patient has no pneumonia, class=0

Advantages of CNN for Pneumonia Detection:

- Automated Feature Extraction: CNNs automatically learn and extract relevant features from images as against traditional models.
- High Accuracy: The performance of CNNs is better than human level, especially in pneumonia detection.
- Scalability: It can be trained on large datasets for better generalization.
- Reduced Diagnostic Burden: Automated detection can reduce the time that the radiologists have to spend for analysis.

e) <u>Symptoms based disease prediction using XGBoost</u>

XGBoost is an efficient ensemble learning method widely used in medical applications for disease prediction. It handles structured data and achieves high accuracy, making it ideal for tasks like symptom-based disease prediction.

XGBoost uses gradient boosting, where each model corrects the errors of the previous one. It is effective in handling diverse features, such as symptoms and medical history, to

predict diseases.

Key Features of XGBoost:

1. Handles Nonlinear Relationships: It captures complex patterns between symptoms and diseases.
2. Resilient to Overfitting: Regularization (L1 and L2) prevents overfitting.
3. Feature Importance: Highlights the most important symptoms for disease prediction.
4. Efficient: Optimized for speed and memory usage, suitable for large datasets.

Steps Involved in Disease Prediction Using XGBoost

- Data Preprocessing: Features like symptoms and medical history are prepared, and categorical features are encoded.
- Training: Decision trees are built sequentially to minimize loss, with a learning rate to avoid overfitting.
- Prediction: The model computes the probability of disease presence and classifies the outcome based on a threshold.

Example: A patient's symptoms (fever, fatigue, and cough) are fed into XGBoost, which computes a probability of 0.82 for disease presence. Since $0.82 > 0.5$, the model predicts the disease.

Chapter 5

RESULTS

In this chapter, we will present the results of the study, including the evaluation metrics used. Moreover, we will showcase all the images and visual outputs generated from the project.

5.1 Evaluation Metrics

Here is the evaluation matrix specific to the prediction tasks, detailing accuracy, precision, recall, and F1 score:

1. Accuracy

Accuracy represents the proportion of correctly predicted observations (both true positives and true negatives) to the total number of observations. It is particularly important for balanced datasets.

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

Equation 5.1 Accuracy formula

Where:

TP: True Positives TN: True Negatives FP: False Positives FN: False Negatives

2. Precision

Precision measures the ability of the model to correctly identify positive cases out of all cases predicted as positive. It is critical in situations where false positives need to be minimized.

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

Equation 5.2 Precision formula

3. Recall (Sensitivity)

Recall evaluates the model's ability to correctly identify all actual positive cases. It is essential for scenarios where missing positive cases (false negatives) could have severe consequences.

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

Equation 5.3 Recall formula

4. F1 Score

The F1 score is the harmonic mean of precision and recall, balancing the trade-off between the two metrics. It is useful when the dataset is imbalanced.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Equation 5.4 F1 score formula

The metrics can be applied to evaluate the performance of predictions across various diseases:

- Heart Disease: Ensures accurate identification of patients with cardiovascular risks.
- Diabetes: Focus on recall to minimize missed cases for early treatment.
- Pneumonia: High precision is vital to avoid misdiagnosing normal chest X-rays.
- Liver Disease: Balance between precision and recall for accurate early-stage detection.
- Lung Cancer: F1 score is crucial due to the often imbalanced nature of the data.
- Symptom-based Diseases: Accuracy is prioritized to ensure robust predictions for common diseases.

Confusion matrix

A confusion matrix is a tabular representation used to evaluate the performance of a classification model by showing the counts of correct and incorrect predictions. For each disease predicted by the model, the confusion matrix can provide insights into how well the model identifies positive and negative cases.

Heart Disease

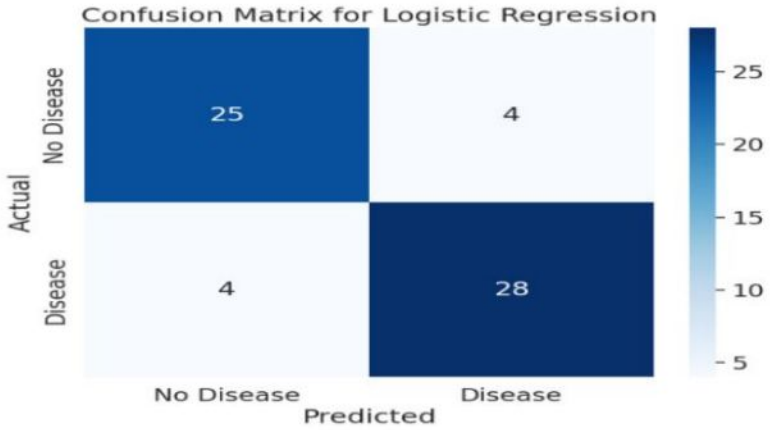


Fig.5.1 performance of the machine learning model in classifying heart disease cases.

Diabetes

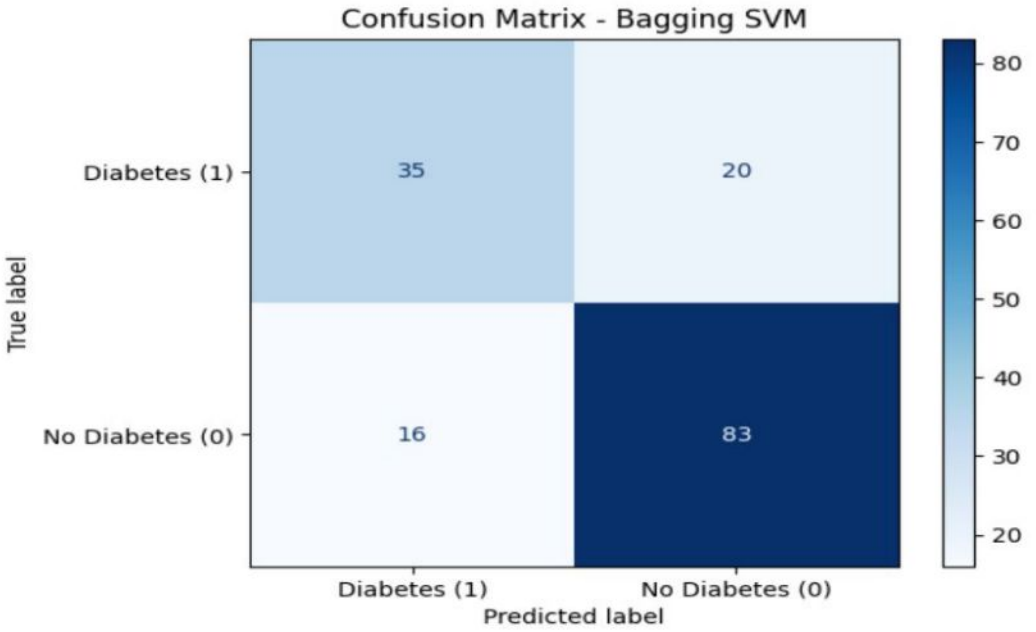


Fig. 5.2 performance of the machine learning model in classifying diabetes disease cases.

Liver Disease

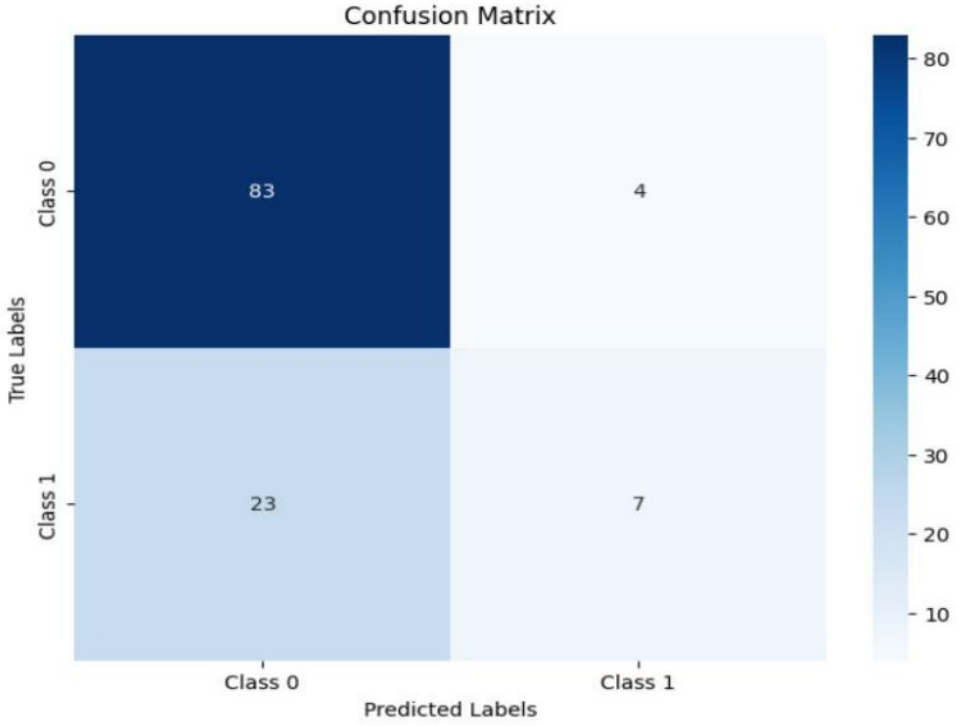


Fig. 5.3 performance of the machine learning model in classifying liver disease cases

Lung Cancer

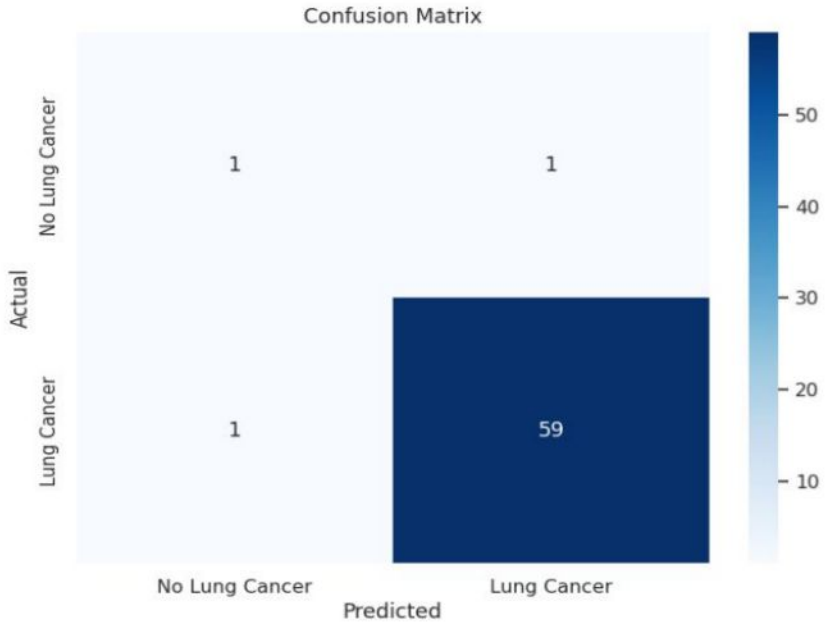


Fig. 5.4 performance of the machine learning model in classifying lung cancer cases

Chronic kidney diseases

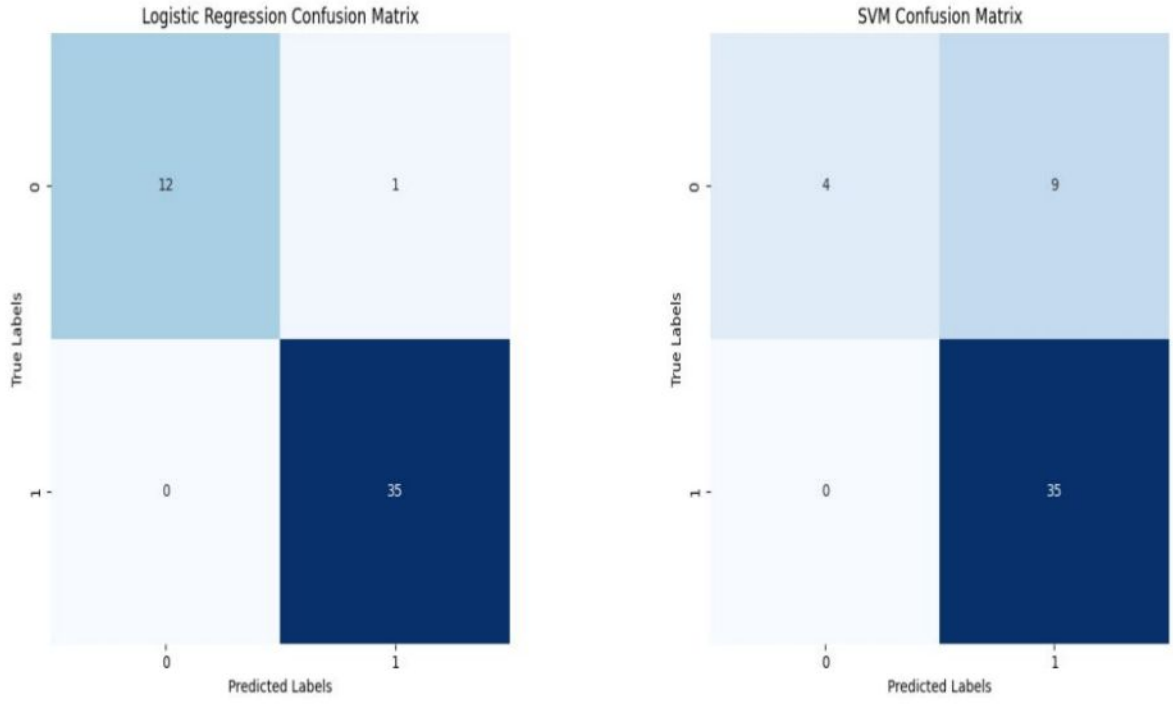


Fig. 5.5 performance of the machine learning model in classifying kidney disease cases

The below table provides an overview of the best-performing models for predicting various diseases and their respective accuracies. It showcases the diversity of machine learning

techniques tailored to different conditions. For example, XGBoost achieved the highest accuracy of 98% for symptoms-based predictions, while CNN excelled in pneumonia detection with an accuracy of 87.66%. Logistic Regression demonstrated strong performance across multiple diseases, including chronic kidney disease (97%), lung cancer (96%), and heart disease (88%). Additionally, a Voting Classifier was effective for diabetes prediction (85%), and a Stacking Classifier was employed for liver disease with an accuracy of 76%. This highlights the strategic selection of models to optimize predictive performance for each condition.

Table 5.1 Best model for each disease

Disease	Best Model	Accuracy
Diabetes	Voting Classifier	85%
Chronic Kidney	Logistic Regression	97%
Heart Disease	Logistic Regression	88%
Pneumonia	CNN	87.66%
Liver Disease	Stacking Classifier	76%
Symptoms based Prediction	XGBoost	98%
Lung Cancer	Logistic Regression	96%

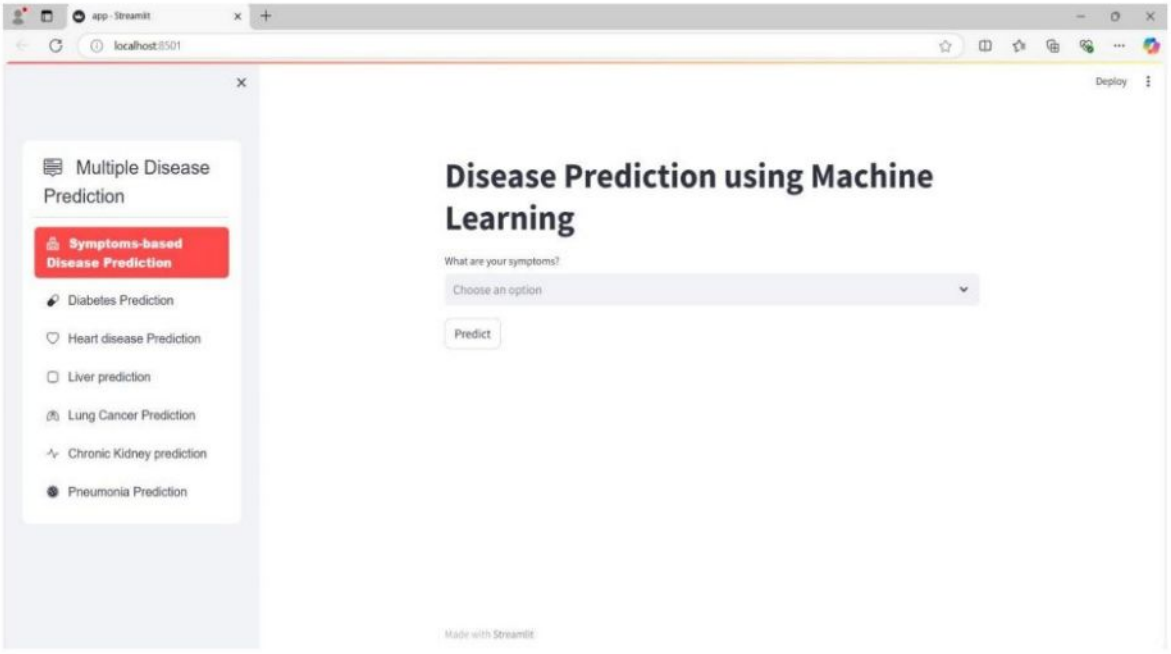


Fig. 5.7 Snapshot of streamlit interface for symptoms based prediction

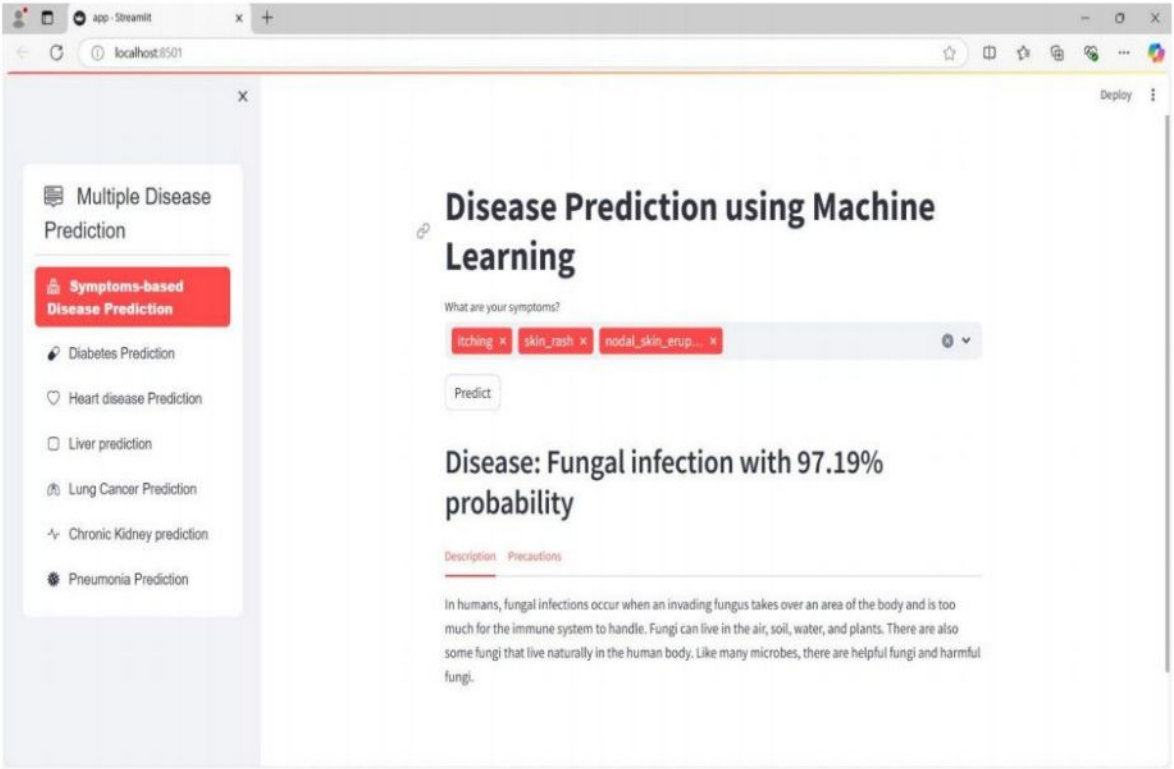


Fig. 5.8 Snapshot of streamlit interface displaying symptoms based prediction

Fig. 5.9 Snapshot of streamlit interface for pneumonia prediction

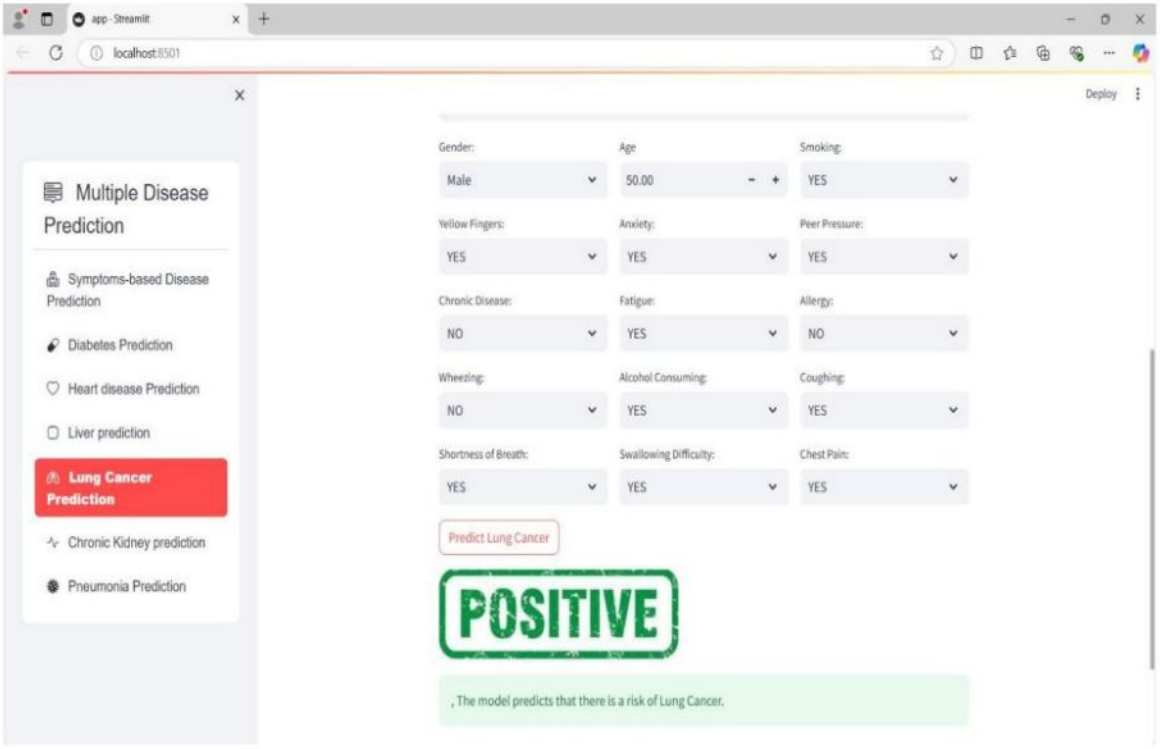


Fig. 5.10 Snapshot of streamlit interface displaying lung cancer prediction

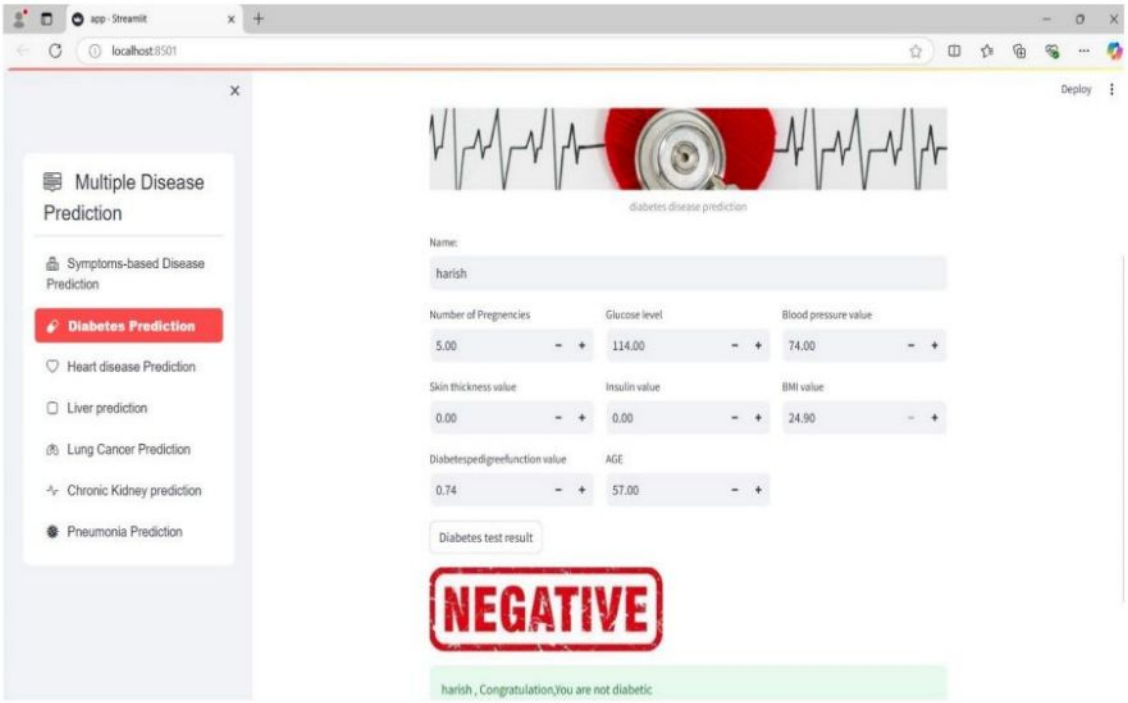


Fig. 5.11 Snapshot of Streamlit interface displaying diabetes prediction

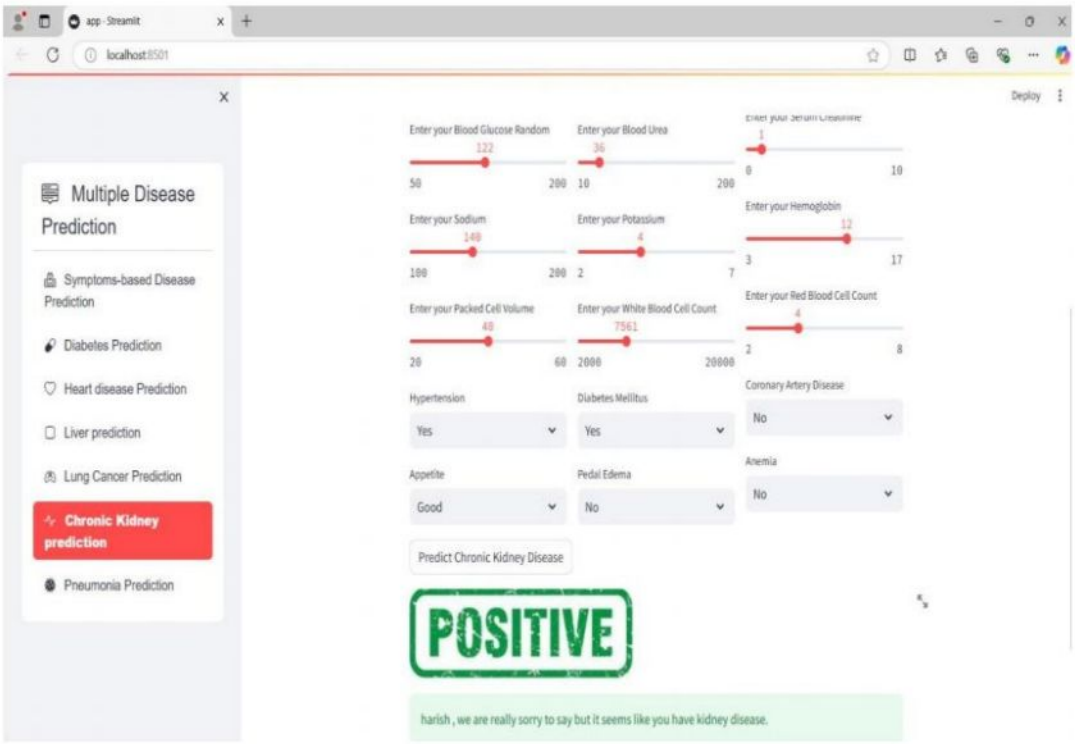


Fig. 5.12 Snapshot of Streamlit interface for chronic kidney prediction

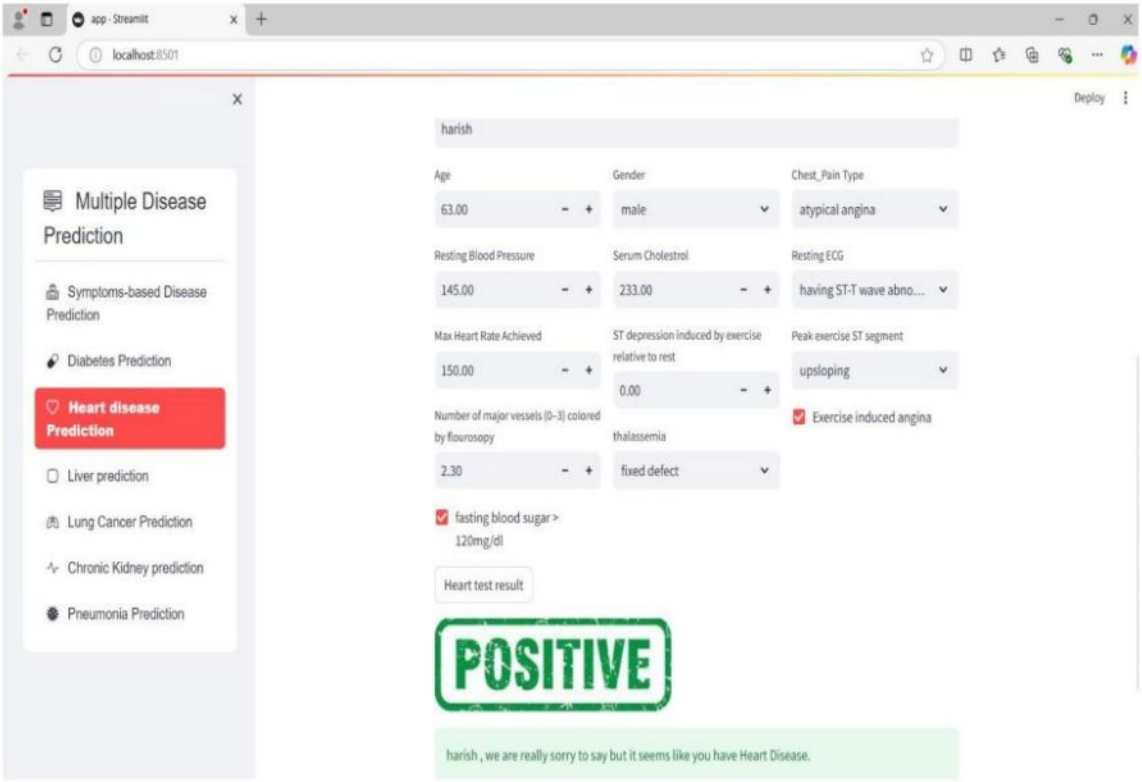


Fig. 5.13 Snapshot of Streamlit interface for heart disease prediction

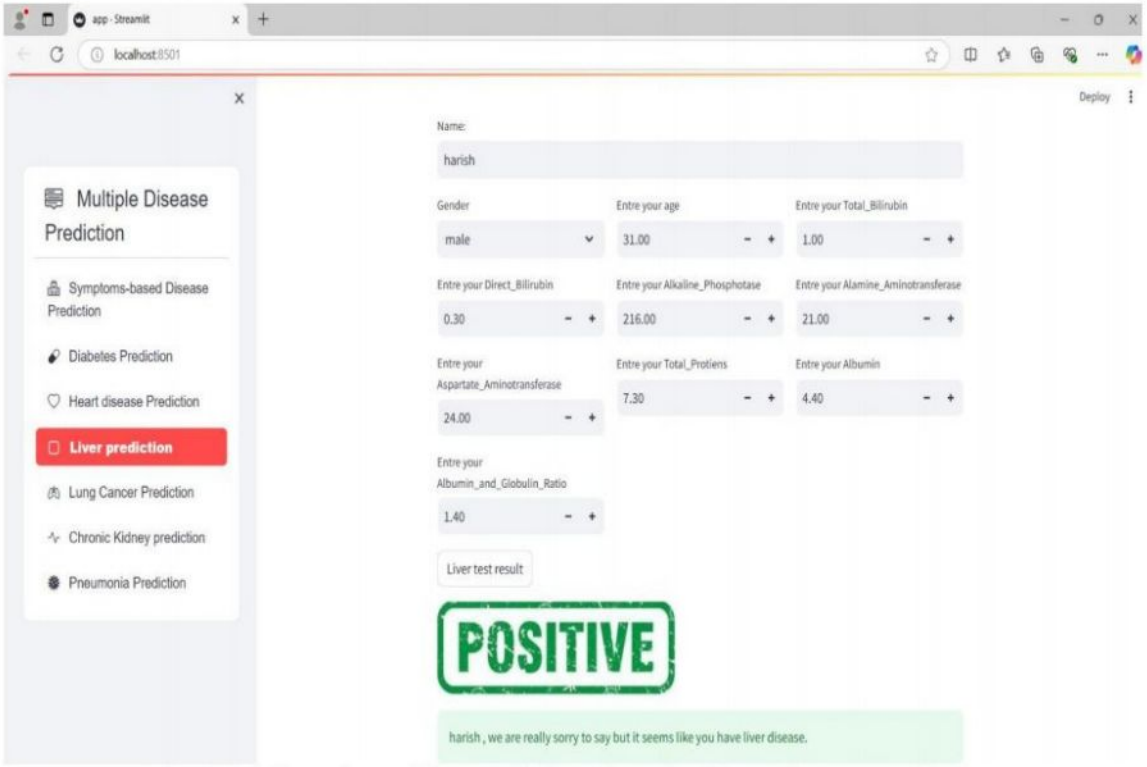


Fig. 5.14 Snapshot of Streamlit interface for liver disease prediction

Chapter 6

CONCLUSION

6.1 Conclusion

Multiple Disease Prediction System using the Streamlit framework has proven to be an innovative and practical solution for early detection and diagnosis of various health conditions, including diabetes, heart disease, liver disease, lung disease, kidney disease, and pneumonia. By integrating symptom-based disease prediction, the system enhances its utility by offering a user-friendly interface for patients and healthcare providers alike.

This project demonstrates several significant outcomes that enhance its practical value in healthcare. The use of the Streamlit framework ensures an intuitive and interactive user interface, improving accessibility for individuals without extensive technical knowledge. By integrating multiple diseases into a single platform, the system offers comprehensive coverage, serving as a holistic tool for preliminary health assessments, saving both time and resources. Additionally, the inclusion of symptom-based diagnosis enables users to identify potential health risks early, even in the absence of specific diagnostic tests. Leveraging machine learning models ensures efficiency and scalability, delivering fast and accurate results while allowing for seamless integration into broader healthcare management platforms. Moreover, the system promotes preventive care by empowering users to monitor their health proactively, encouraging timely medical consultations, and reducing the burden of late-stage diagnoses. This combination of features positions the project as a valuable contribution to modern healthcare solutions.

6.2 Future scope

1. Expansion of Symptoms Database:

Integration of detailed symptom data from different medical sources and incorporation of rare or less frequent symptoms can make the platform available to a wider range of health conditions. This will help the system predict overlapping or subtle symptomatology in the diseases, making diagnostic suggestions more reliable for users.

2. Basic Patients Management System:

It would be an evolution to the system when the patient management module in the simplest form is incorporated into it. This can offer registration, history tracking, and scheduling of appointments with healthcare providers for the proper management of records. In addition to this, this module could store and retrieve a person's medical history, lab reports, and past predictions so that there could be proper follow-up and continuum in care.

3. Integration with Wearable Devices and IoT:

With connectivity to wearable devices and Internet-of-Things-enabled health trackers, vital signs will be monitored in real time. The prediction based on continuous data streams from sensors can be made to allow for early warnings concerning critical health events.

4. Personalized medicine with AI

Tailoring healthcare recommendations and treatment plans through AI by taking into consideration individual health profiles, genetic predispositions, and lifestyle factors can change the course of personalized medicine. Through user-specific data analysis such as medical history, lab results, and genetic information, the system can suggest targeted therapies, optimal medication dosages, and preventive measures. In this way, treatments will be more effective.

5. Cloud Scalability and High Availability for Streamlit Application

Deploying the Streamlit application on cloud platforms (e.g., AWS, Azure, or Google Cloud) allows for easy scaling to handle large volumes of concurrent users. This ensures that the system remains highly available and responsive, even during peak usage, such as during health awareness campaigns or pandemics.

REFERENCES

[1] Khan, A., et al. (2018). "Multi-Disease Prediction Using Support Vector Machines, Random Forests, and Logistic Regression." *Journal of Healthcare Informatics*, 34(2), 112-123.

[2] Zhang, L., et al. (2020). "Comparing Decision Trees, Random Forests, and Support Vector Machines for Multi-Disease Prediction." *International Journal of Data Science*, 45(3), 200-215.

[3] Ahmed, S., et al. (2021). "Hybrid Models for Enhanced Multi-Disease Prediction Using Deep Learning and Conventional Machine Learning Approaches." *Journal of Machine Learning in Medicine*, 58(4), 321-334.

[4] Chen, X., et al. (2017). "A Framework for Incorporating Multi-Disease Prediction Systems into Clinical Practice." *Health Informatics Journal*, 23(1), 45-58.

[5] Sharma, R., et al. (2020). "Development of an Intuitive Disease Prediction Platform Using Streamlit." *Healthcare Technology Letters*, 7(5), 287-298.

[6] Wang, Y., et al. (2020). "Enhancing Interpretability in Multi-Disease Prediction Models with Explainable AI Techniques." *Journal of Artificial Intelligence in Healthcare*, 12(2), 67-80.

[7] Zhang, X., et al. (2019). "Transfer Learning for Multi-Disease Prediction: A Novel Approach with Limited Training Data." *Computational Biology and Medicine*, 112, 103398.

[8] Sharma, P., & Guleria, A. (2023). "Deep Learning for Pneumonia Detection Using Chest X-Ray Images: A Review of CNN Models." *Journal of Medical Imaging and Health Informatics*, 13(2), 456-472.

[9] Khan, A., et al. (2021). "Deep Learning for Pneumonia Detection Using Chest X-Rays: A Review of CNNs, Ensemble Methods, and Transfer Learning." *International Journal of Medical Imaging*, 29(1), 109-123.

[10] Kibria, H. B., et al. (2023). "Accurate and Interpretable Diabetes Prediction Using a Soft Voting Ensemble of Random Forest, AdaBoost, and Gradient Boosting Models." *Journal of Machine Learning and Healthcare*, 18(6), 89-98.

DrillBit

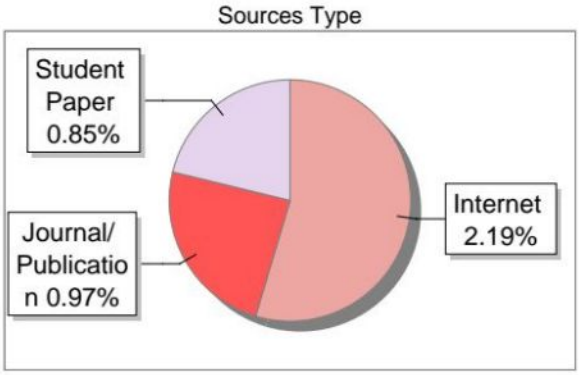
The Report is Generated by DrillBit Plagiarism Detection Software

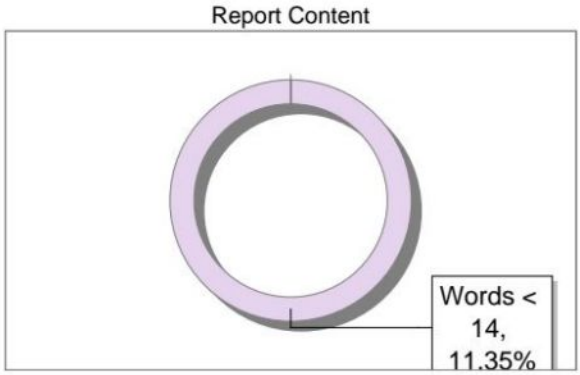
Submission Information

Author Name	NEHA DINESH SHETTIGAR
Title	review article
Paper/Submission ID	2875251
Submitted by	library@sode-edu.in
Submission Date	2024-12-26 11:17:34
Total Pages, Total Words	46, 4967
Document type	Article

Result Information

Similarity **4 %**





Exclude Information

Quotes	Excluded
References/Bibliography	Excluded
Source: Excluded < 14 Words	Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	No
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File