Disease Prediction using Machine Learning

Submitted as Third Year Mini Project 2A by

Group Number: TE16	
Group Member Names	**Roll No.**
Kartik Babu	**01**
Harsh Chaurasiya	**08**
Harikrishna Gurrapu	**10**
Soham Handore	**11**

Supervisor: **Mrs. Amrita Jhaveri**

Department of Electronics and Computer Science

V.E.S. Institute of Technology

(An Autonomous Institute affiliated to University of Mumbai, approved by AICTE & Recognized by Govt .of Maharashtra)

A.Y. 2024 - 25

CERTIFICATE

This is to certify that the project entitled **"Disease Prediction using Machine Learning"** is a bonafide work of **Kartik Babu(01), Harsh Chaurasiya(08), Harikrishna Gurrapu(10) , Soham Handore(11)** submitted to the V.E.S. Institute of Technology as a Third Year Mini Project 2A during the academic 2024-25.

(Name and sign)
Supervisor/Guide

(Name and sign)
Head of Department

(Name and sign)
Principal

PROJECT REPORT APPROVAL

This project report entitled **"Disease Prediction using Machine Learning"** by **Kartik Babu (01), Harsh Chaurasiya (08), Harikrishna Gurrapu (10), Soham Handore (11)** is approved as **Third Year Mini project 2A** during Academic year 2024-25.

Examiners

1.----------

2.----------

3.----------

DECLARATION

We declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

---------- ----------

(Name of students and Roll No.) (Name of students and Roll No.)

---------- ----------

(Name of students and Roll No.) (Name of students and Roll No.)

Date:

INDEX

CHAPTER 1
INTRODUCTION

CHAPTER 1
INTRODUCTION

In recent years, the integration of technology into healthcare has transformed how diseases are diagnosed and managed. One of the most impactful advancements is the use of **Machine Learning (ML)** to predict disease risk based on various medical data parameters. This shift allows healthcare professionals to utilize vast amounts of patient data, such as health indicators, lab results, and symptoms, to identify early signs of critical conditions like heart disease, kidney disease, and liver disease.

Key points to highlight:

- **Complexity of Disease Diagnosis:** Traditional diagnostic methods often struggle with the intricate interplay of biological and environmental factors that affect the human body. Machine learning tools, such as decision trees, neural networks, and ensemble methods, provide an advanced approach to analyzing these complexities by recognizing patterns in the data that may not be obvious to human experts.
- **Data-Driven Health Insights:** These algorithms are trained on extensive datasets, enabling them to learn from numerous cases and continuously improve their predictive capabilities. This allows for more reliable predictions and the discovery of subtle trends or correlations that may go unnoticed through conventional methods.
- **Proactive Healthcare:** By providing actionable insights, machine learning models assist doctors in making better-informed decisions. They enable earlier interventions and facilitate personalized treatment plans that are specifically tailored to each patient's needs, helping to reduce the burden of chronic diseases and improve patient outcomes**.**
- **Aim and Scope**: The purpose of this report is to explore how machine learning models can be applied to predict the risk of developing heart, kidney, and liver diseases. By highlighting the advantages of early detection and personalized care, this investigation aims to showcase how ML can help foster a proactive approach in healthcare.
- **Significance**: The potential of these models extends beyond diagnostics, offering healthcare systems a revolutionary tool for enhancing the accuracy of predictions and improving patient car

CHAPTER 2
LITERATURE REVIEW

CHAPTER 2
LITERATURE REVIEW

- Recent advancements in disease prediction using machine learning have garnered significant attention, showcasing the transformative potential of these technologies in healthcare. A variety of studies demonstrate how machine learning (ML) models can enhance the accuracy and reliability of disease predictions across different medical domains.

- For instance, Zhou et al. [2023] in the *Journal of Medical Systems* revealed compelling evidence that deep learning models have a distinct advantage over traditional predictive methods when it comes to forecasting cardiovascular events. Their research involved analyzing extensive datasets, which allowed them to uncover complex patterns that simpler models often overlook. This study emphasizes the ability of deep learning to process large volumes of data and deliver nuanced insights, ultimately leading to improved patient outcomes in cardiology.

- In the realm of oncology, Sharma et al. [2022] reviewed the efficacy of support vector machines (SVM) and random forests in predicting cancer recurrence. Published in *IEEE Access*, their analysis highlights how these algorithms can effectively integrate genomic and clinical data, paving the way for more personalized treatment approaches. By utilizing these advanced predictive techniques, healthcare providers can better assess patient risk and tailor follow-up care, thus potentially improving survival rates.

- Patel et al. [2024], in their work featured in *Nature Medicine*, underscored the critical role of longitudinal data in enhancing the accuracy of disease predictions. Their findings indicate that tracking patients over time allows for a deeper understanding of health trajectories, which in turn leads to more precise risk assessments. This study advocates for the integration of long-term data collection into predictive modeling, reinforcing the notion that context and temporal factors significantly influence patient health outcomes.

- Moreover, the landscape of infectious diseases has also been transformed through machine learning applications. Kim et al. [2023], in a study published in *Frontiers in Public Health*, examined the effectiveness of Long Short-Term Memory(LSTM) networks and XGBoost algorithms in forecasting outbreaks such as COVID-19.

CHAPTER 3
METHODOLOGY

CHAPTER 3

METHODOLOGY

This project follows a systematic approach to developing a disease prediction model using machine learning. The primary goal is to predict the risk of heart, kidney, and liver diseases based on patient data. The methodology is divided into several key stages, ensuring a structured flow from initial research to the final implementation.

1. **In-depth Study of the Topic**: Before beginning the technical aspects, a thorough investigation into the medical conditions—heart, kidney, and liver diseases—was conducted. This study included understanding the biological markers, risk factors, and symptoms commonly associated with these diseases. Reviewing current medical literature and examining how machine learning models have been used for disease prediction helped shape the framework of this project.

2. **Data Collection:** High-quality and relevant data is essential for developing an effective machine learning model. We collected patient datasets, including demographic information, medical history, lab reports, and symptom logs. The data was curated from various sources, including medical databases and research publications. Preprocessing steps were applied to clean the data, address missing values, and ensure it was ready for feeding into the machine learning models.

3. **Choosing Frontend and Database for the Application**: For the user interface, **React.js** was chosen for its flexibility and efficiency in building interactive web applications. The platform allows users to input patient data and view predictions seamlessly. It offers a modern, responsive design, ensuring the application is both user-friendly and accessible.

4. On the Backend, **Python** was utilized for its extensive libraries in machine learning, while **Express.js** was integrated for handling API calls and managing server operations efficiently.

MongoDB was selected as the database due to its scalability and ability to handle large amounts of unstructured patient data. It stores information such as patient details and prediction results in an efficient, document-based format.

5. **Developing and Training the Machine Learning Model**: Using Python's powerful machine learning libraries like Scikit-learn and TensorFlow, several models were developed and tested. Decision trees, neural networks, and ensemble methods were explored to determine which algorithms provided the most accurate predictions. The models were trained on the collected patient data, with the performance of each model evaluated through metrics such as accuracy, precision, and recall.

6. **Integration and Testing**: The machine learning models were integrated into the backend of the application using Python and Express.js. Data flows between the frontend (React.js) and backend are handled through secure API calls. Extensive testing was conducted to ensure that the system performs well under real-world conditions, handling various inputs, and producing reliable prediction.

CHAPTER 4 BLOCK DIAGRAM & WORKING

CHAPTER 4
BLOCK DIAGRAM & WORKING

4.1. Block Diagram:

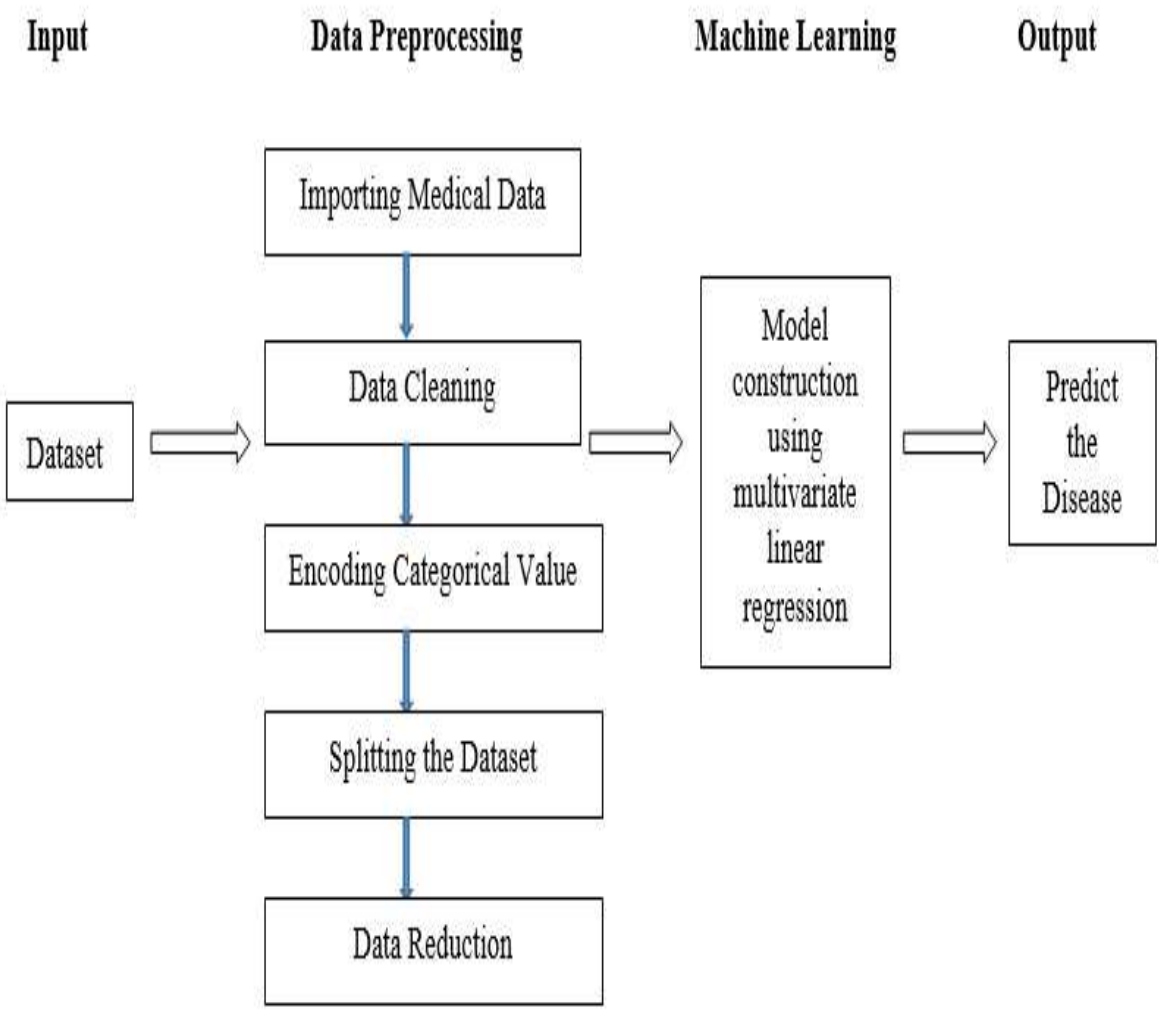

Fig.1. Block Diagram of our project

WORKING OF THE PROJECT

This project aims to create efficient and user-friendly disease prediction software using machine learning, designed to assist both healthcare professionals and patients. With separate portals for doctors and patients, the software enhances communication, data management, and predictive healthcare, focusing on risks related to kidney, liver, and heart diseases. Below is a more detailed look at how the system operates.

1. **User Portals: Doctor and Patient**
    - **Doctor Portal**: Aimed at helping healthcare professionals manage patient records, analyze reports, and generate insights. Doctors can access individual patient profiles, review lab reports, and receive machine-learning-based predictions.
    - **Patient Portal**: Provides an easy-to-use interface for patients to upload reports, input health parameters, and view their risk assessments.

2. **Data Input and Submission**
    - **Patient Submissions**: Patients upload their blood reports and enter basic health parameters such as blood pressure, age, gender, and BMI. These inputs form the foundation for disease prediction.
    - **Automated Report Analysis**: The Machine learning models analyze the submitted data and generate risk assessments for heart, kidney, and liver diseases.

3. **Backend Processing and Machine Learning Integration**
    - **Frontend and Backend Setup**: The software uses React.js for the frontend, providing a seamless experience for both patients and doctors. The backend is powered by Python, with Express.js managing server requests and MongoDB storing patient data.
    - **Machine Learning Models**: Trained machine learning algorithms process patient data to predict disease risk. These models have been developed using Python libraries like Scikit-learn and TensorFlow, ensuring high accuracy in detecting potential health

risks.

4. **Predictive Insights and Reports**
    - **For Doctors**: Doctors receive detailed insights, including risk levels for specific diseases based on their patients' health data. These insights help in crafting personalized treatment plans and advising on early interventions.
    - **For Patients**: After submission, patients are provided with a summarized health report, including predictions on disease risks. This enables them to take proactive measures or consult their healthcare provider for further analysis.

5. **Communication and Care Coordination**
    - **Patient-Doctor Interaction**: The doctor portal includes communication tools to facilitate discussions regarding patient reports and predictions. This feature enhances coordination between patients and their doctors, ensuring timely interventions and adjustments in care plans.

By employing this dual-portal system and machine learning capabilities, the software streamlines the disease prediction process, empowering both patients and doctors. It fosters a proactive healthcare.

4.1 Advantages and Applications

1. **Early Detection:** Machine learning models can analyze patient data to detect early signs of heart, liver, or kidney diseases, enabling timely intervention.
2. **Improved Accuracy:** The project uses advanced algorithms to provide more accurate predictions compared to traditional diagnostic methods.
3. **Personalized Healthcare:** By evaluating individual health data, the system can offer customized recommendations based on personal risk factors.
4. **Cost-Effective:** Automated predictions reduce the need for extensive medical tests, making healthcare more affordable and accessible.
5. **Scalability:** The system can handle large volumes of patient data, making it suitable for deployment in hospitals and clinics for widespread disease screening.
6. **Real-Time Predictions**: The system provides immediate feedback on whether a patient is at risk for heart, liver, or kidney diseases, supporting faster decision-making.

Some Applications are as follows:

- **Early Disease Detection:** Identifies risks for heart, kidney, and liver diseases before symptoms appear.
- **Personalized Treatment Plans:** Enables customized treatment based on individual risk factors.
- **Proactive Health Monitoring:** Allows patients to regularly monitor their health and risk levels.
- **Enhanced Doctor Decision-Making:** Provides predictive insights to assist doctors in diagnoses and treatment.
- **Remote Patient Management:** Facilitates continuous care and monitoring without frequent in-person visits.

CHAPTER 5
HARDWARE & SOFTWARE OVERVIEW

CHAPTER 5
SOFTWARE

In this disease prediction project, each of the tools used plays a specific role in building an efficient and responsive system for predicting heart, kidney, and liver disease risks. Here's how each programming tool contributes and why it's important:

1. **Python**:

   Python is the backbone of this project's machine learning algorithms. It's used to analyze patient data, build, and train predictive models. Libraries like **scikit-learn**, **TensorFlow**, and **Pandas** are key here. For example, **scikit-learn** helps with building algorithms like decision trees, while **TensorFlow** is used for deep learning models. **Pandas** make handling and organizing large medical datasets easier. Python's flexibility and strong library ecosystem make it ideal for tasks like model training, data analysis, and validation.

2. **MongoDB**:

   MongoDB is used as the database because it's designed to handle unstructured or semi-structured data, which is common in medical records. Instead of organizing data in strict rows and columns like traditional databases, MongoDB stores patient data, test results, and predictions in a more flexible format. This allows the system to efficiently store and retrieve vast amounts of medical data, making it scalable and capable of handling large datasets as more patients are added to the system.

3. **React**:

   React is the tool used to create the front-end of the project, which is what users (both doctors and patients) interact with. It makes the user interface dynamic and responsive. For example, when a doctor inputs patient data or a patient uploads their reports, React ensures the application runs smoothly and updates in real-time without needing to reload the page. React's component-based structure allows developers to reuse code and keep the interface fast and organized, which enhances the overall user experience.

4. **Node.js**:

   Node.js is used to handle the server-side operations of the project. It bridges the gap between the front end (React) and the backend (MongoDB). When a patient uploads their medical data, Node.js takes that information, processes it, and stores it in the database. Similarly, it retrieves the data and passes it to the machine learning model for prediction. This ensures real-time communication between the different parts of the system so that the data flows smoothly, and users can quickly see their predictions.

5. **Express**:

   Express is a framework built on top of Node.js, making it easier to set up and manage the web server. It helps in building the APIs that allow the frontend (React) to communicate with the backend (Python, MongoDB). Express handles tasks like routing HTTP requests and ensuring secure data transmission between the server and the client (whether it's a doctor or a patient). This is essential for making sure the data is processed efficiently and securely between the user interface and the database.

CHAPTER 6
RESULT AND DISCUSSION

CHAPTER 6
RESULT & DISCUSSION

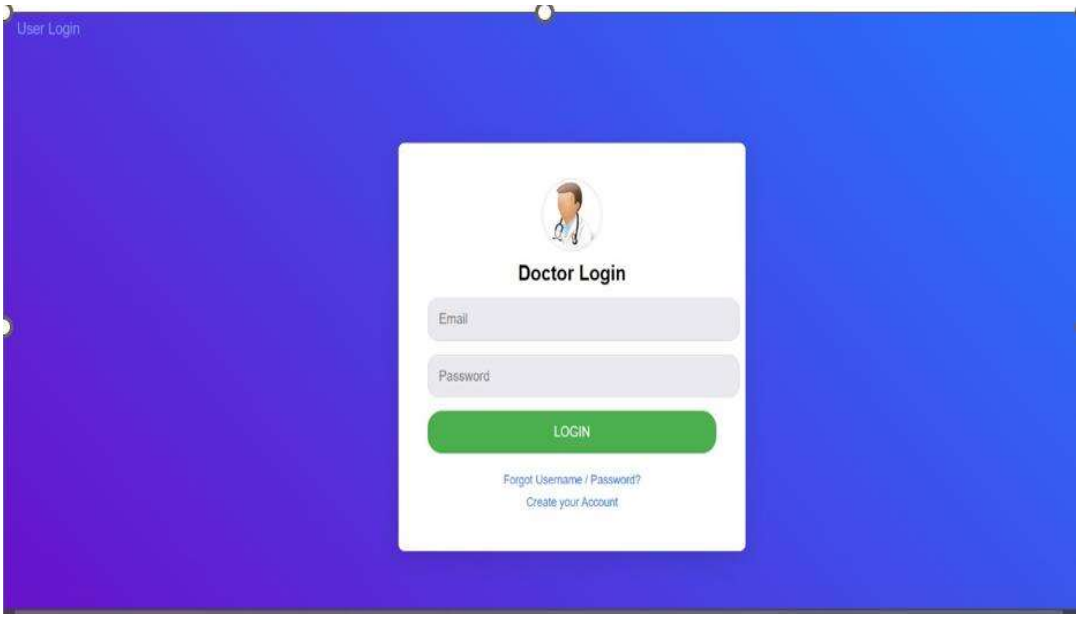

Figure 2 – Login page for doctor

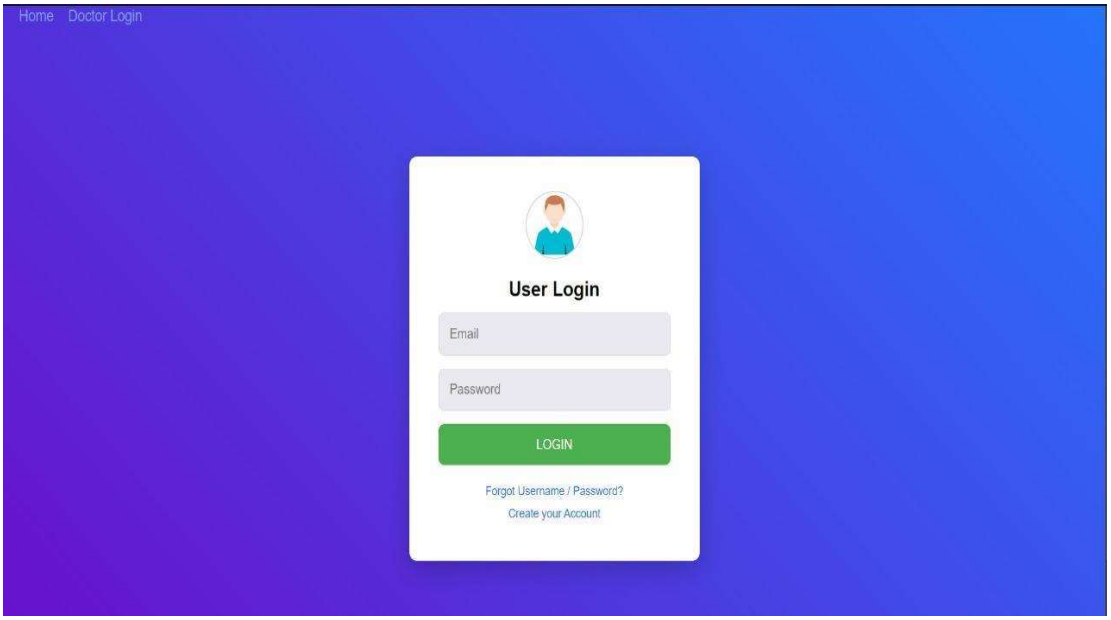

Figure 3 – Login page for User

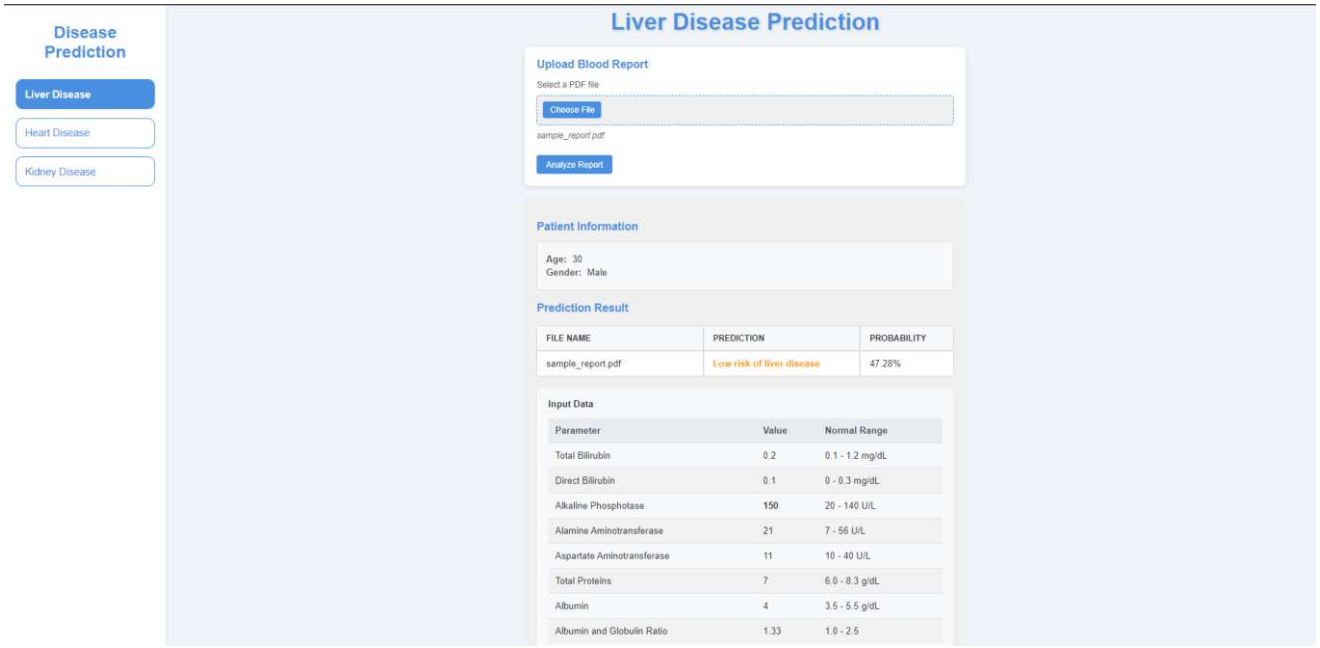

Figure 4 - User uploads blood report and gets prediction

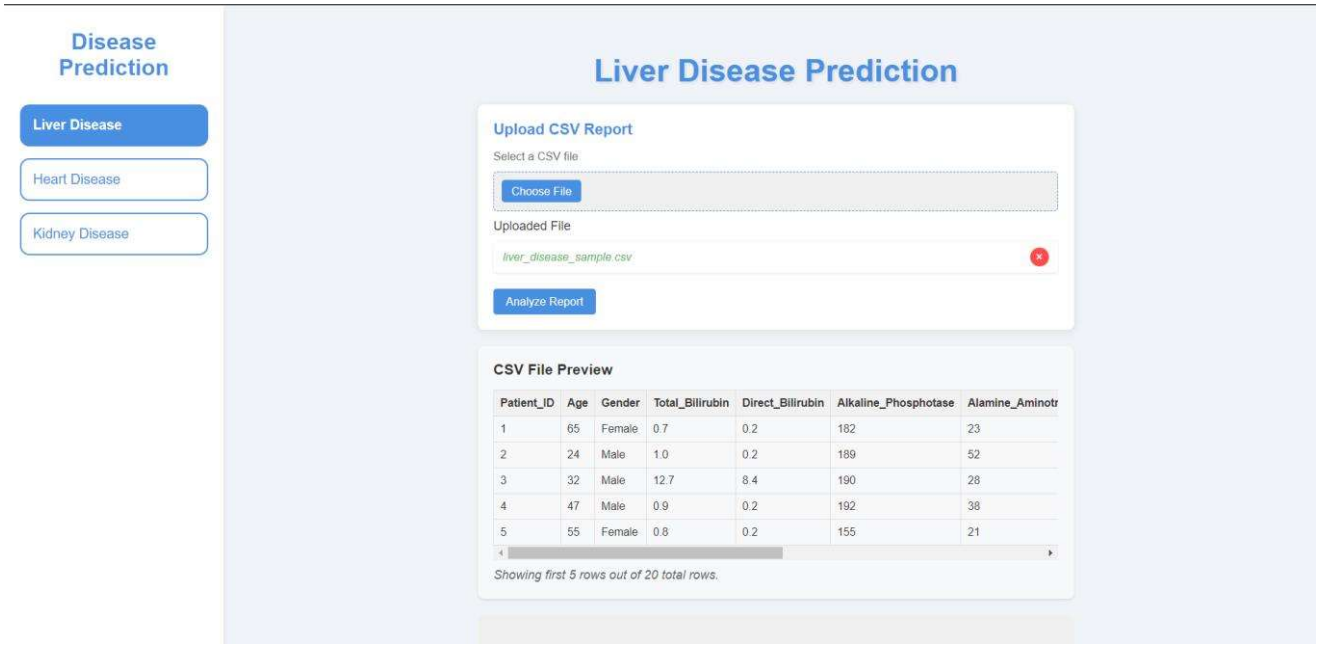

Figure 5 – doctor uploads multiple patient data reports

Prediction Results

PATIENT ID	PREDICTION	PROBABILITY
11	Likely to have liver disease	99.65%
8	Likely to have liver disease	99.23%
7	Likely to have liver disease	98.81%
15	Likely to have liver disease	98.56%
14	Likely to have liver disease	97.86%
9	Likely to have liver disease	96.29%
16	Likely to have liver disease	94.67%
3	Likely to have liver disease	93.33%
6	Likely to have liver disease	91.00%
10	Likely to have liver disease	86.84%
20	Likely to have liver disease	81.06%
18	Likely to have liver disease	76.88%
17	Likely to have liver disease	65.88%
5	Likely to have liver disease	61.71%
1	Likely to have liver disease	59.88%
4	Likely to have liver disease	57.77%
12	Likely to have liver disease	56.49%
2	Likely not to have liver disease	49.02%
13	Likely not to have liver disease	48.75%
19	Likely not to have liver disease	46.41%

Figure 6 – Predictions obtained (sorted on the basis of criticality)

Discussion and Results

The following were the results made through the analysis of the project.

- **Accurate Predictions**: The machine learning models integrated into the system have successfully analyzed patient data and provided accurate predictions for potential disease risks. By using advanced algorithms like decision trees and neural networks, the system was able to detect patterns in health parameters, lab reports, and patient history that may indicate the early onset of diseases.

- **Improved Patient Care**: Doctors using the platform have gained access to deeper insights into their patients' health. The predictive insights generated by the machine learning models allow healthcare professionals to create more personalized and data-driven treatment plans. This leads to better management of diseases and improved patient outcomes through timely interventions.

- **Enhanced Patient Engagement**: The patient portal has empowered individuals to be more proactive in managing their health. By uploading their reports and health data, patients can track their disease risk and take preventive measures. This has contributed to a more informed and health-conscious patient base.

- **Efficiency and Scalability**: The use of MongoDB for the backend has proven efficient in managing large volumes of patient data. The system is scalable, meaning it can handle increasing numbers of users and medical records without compromising performance. Additionally, React.js has enabled a seamless and user-friendly experience on the frontend, ensuring fast interactions and real-time updates.

- **Real-World Application**: This project has strong potential for real-world use, particularly in telemedicine and remote patient monitoring. The dual-portal system facilitates easy communication between doctors and patients, enhancing overall healthcare delivery. By

integrating predictive models into routine care, both patients and doctors benefit from a more data-driven approach to disease prevention.

- **Challenges and Limitations**: Despite the success, some challenges remain. The accuracy of predictions can depend heavily on the quality of the data provided. For example, incomplete or inaccurate medical records can impact the model's effectiveness. Additionally, further validation with larger datasets and real-world trials will be needed to ensure the model performs consistently.

TESTING MODEL

Testing Model Overview:

- **Data Preparation**
- **Data Splitting**: Divide the dataset into training, validation, and testing subsets to ensure unbiased evaluation.
- **Preprocessing**: Clean and preprocess the data, including normalization, handling missing values, and encoding categorical variables.
- **Model Training**
- Algorithm Selection: Choose appropriate machine learning algorithms (e.g., decision trees, neural networks) for training.
- Training Phase: Train the model on the training dataset to learn patterns and relationships within the data.
- **Validation**
- Hyperparameter Tuning: Optimize model parameters using the validation set to improve accuracy and prevent overfitting.
- Cross-Validation: Employ techniques like k-fold cross-validation to ensure model robustness.
- **Testing**
- Evaluation Metrics: Assess model performance using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC on the testing set.
- Confusion Matrix: Analyze the confusion matrix to understand true positives, false positives, true negatives, and false negatives.
- **Model Deployment**
- Integration Testing: Test the integrated system (frontend, backend, and database) to ensure seamless interaction.
- User Acceptance Testing (UAT): Conduct testing with real users (doctors and patients) to gather feedback and ensure the model meets their needs.

CHAPTER 7
CONCLUSION & FUTURE SCOPE

CHAPTER 7
CONCLUSION & FUTURE SCOPE

Conclusion:

This machine learning-based disease prediction project marks a major step forward in healthcare by helping detect heart, liver, and kidney disease risks early. Its personalized insights based on individual health data enable better prevention and treatment plans, while real-time feedback allows doctors to act quickly. The system is scalable, making it suitable for large hospitals, and it continuously improves by learning from new data. Overall, this tool has the potential to make healthcare more proactive, affordable, and precise, improving patient outcomes.

Future Scope:

The project can be expanded in several exciting ways:

1. More Diseases: Predict risks for diabetes, cancer, and other conditions.
2. Wearable Integration: Use data from smart devices for real-time health tracking.
3. Global Data: Improve accuracy by incorporating diverse medical datasets.
4. Advanced AI: Use deeper AI to suggest personalized treatments.
5. Telemedicine: Enhance virtual care with predictive insights.
6. Patient Tools: Add features like health reminders and goals.
7. Data Security: Strengthen privacy with blockchain.
8. Population Analytics: Help hospitals manage resources by predicting public health trends.
9. Clinical Trials: Identify candidates for research studies.
10. Personalized Medicine: Integrate genetic data for highly customized care.

In short, this project has laid the groundwork for making healthcare smarter and more focused on prevention, with many exciting possibilities for future growth

CHAPTER 8
REFERENCES

CHAPTER 8
REFERENCES

1. M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, "Machine Learning Techniques for Heart Disease Prediction," *Procedia Computer Science*, vol. 85, pp. 108-115, [2016]. doi: 10.1016/j.procs.2016.08.256.

2. M. Madhavi Latha and R. Jeevan, "Prediction of Liver Disease Using Classification Algorithms," *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 1-5, [2018]. doi: 10.14419/ijet.v7i4.40.27659.

3. P. Anuradha, A. Rashmi, and R. Sridevi, "Kidney Disease Prediction Using Machine Learning Algorithms," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, [2020], pp. 1-5. doi: 10.1109/ICIMIA48430.2020.9074988.

4. X. Li, J. Liu, and J. Liu, "An Integrated Machine Learning Model for Early Prediction of Chronic Kidney Disease," *Informatics in Medicine Unlocked*, vol. 18, no. 100293, [2020]. doi: 10.1016/j.imu.2020.100293.

5. S. Rawat, P. Gupta, and R. Mehra, "Hybrid Machine Learning Models for Predicting Liver Disease," *Computer Methods and Programs in Biomedicine*, vol. 196, no. 105670, [2020]. doi: 10.1016/j.cmpb.2020.10567.

6. A. Gupta, R. Ranjan, and S. S. Raj, "Diabetes Prediction Using Machine Learning Algorithms," in Proceedings of the 2021 International Conference on Smart Technologies for Sustainable Development (ICSTSD), [2021], pp. 1-5. doi: 10.1109/ICSTSD50832.2021.9420540.

7. R. Singh, M. Kaur, and P. S. Gill, "Heart Disease Prediction Using Machine Learning Techniques: A Review*," Journal of King Saud University - Computer and Information Sciences*, [2022]. doi: 10.1016/j.jksuci.2022.05.002.

8. N. Shinde, A. Patil, and S. Sharma, "Breast Cancer Prediction Using Machine Learning Algorithms," *International Journal of Computer Applications*, vol. 975, no. 8887, [2020]. doi: 10.5120/ijca2020920780.

9. K. B. V. Reddy, S. R. M. Krishna, and A. R. Rao, "A Comprehensive Review on Machine Learning Techniques for Diabetes Prediction," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 4675-4686, [2021]. doi: 10.1007/s12652-020-02520-6.

10. T. S. R. K. Prasad, P. K. Das, and D. R. Rao, "Application of Machine Learning Techniques for Predicting Chronic Diseases," *Journal of Biomedical Informatics*, vol. 113, no. 103635, [2021]. doi: 10.1016/j.jbi.2021.103635.