AI-Powered Multi-Disease Prediction and Medical Guidance System Using Machine Learning

DOKKU LAKSHMI PRAVALLI KA

PG Scholar. Department of MCA, DNR College, Bhimavaram, Andhra Pradesh

V.SARALA

(Assistant Professor), Master of Computer Applications, DNR College, Bhimavaram, Andhra Pradesh

ABSTRACT

The rapid advancement of Artificial Intelligence (AI) in healthcare has opened new opportunities for early disease detection and personalized medical guidance. This project presents an AI-powered multi-disease prediction system that integrates machine learning models with Optical Character Recognition (OCR) to provide a comprehensive digital healthcare assistant. The system is designed to predict multiple diseases such as Diabetes, Heart Disease, Parkinson's Disease, Lung Cancer, and Hypothyroidism based on user input and extracted medical data from uploaded reports. The proposed system leverages pre-trained machine learning models, including Random Forest and other classification algorithms, to analyze patient data and generate predictions. One of the key features of this application is its ability to process medical reports using OCR technology. By extracting critical health parameters such as glucose levels, cholesterol, TSH, T3, and T4 from images, the system minimizes manual data entry and enhances usability. This automation significantly improves efficiency and reduces the chances of human error.

The application is developed using the Streamlit framework, providing an interactive and user-friendly interface. Users can either manually input their health data or upload medical reports for automated analysis. Based on the prediction results, the system provides detailed health recommendations, including diet plans, yoga exercises, precautions, home remedies, and emergency guidance. This makes the system not only diagnostic but also advisory in nature. Additionally, the system includes a feature for locating nearby doctors and hospitals using Google Maps integration. This ensures that users can seek professional medical assistance promptly. The prediction history module allows users to track their health status over time, enabling better monitoring and decision-making. The system is designed with scalability and extensibility in mind, allowing additional diseases and models to be incorporated in the future. While the predictions are not intended to replace professional medical advice, they serve as a preliminary assessment tool that can guide users toward timely medical consultation. Overall, this project demonstrates how AI and machine learning can be effectively utilized to build intelligent healthcare systems that are accessible, efficient, and user-centric. It highlights the potential of integrating data extraction, predictive analytics, and health recommendations into a single platform to enhance preventive healthcare and early diagnosis.

Keywords: Artificial Intelligence, Machine Learning, Medical Diagnosis, OCR, Streamlit, Healthcare Analytics, Disease Prediction, Clinical Decision Support, Health Monitoring, Predictive Modeling..

I. INTRODUCTION

Healthcare is one of the most critical domains where early diagnosis and timely intervention can significantly reduce mortality rates and improve quality of life. With the increasing prevalence of chronic diseases such as diabetes, heart disease, and cancer, there is a growing need for intelligent systems that can assist in early detection and management. Artificial Intelligence (AI) and Machine Learning (ML) have emerged as powerful tools in addressing these challenges by enabling predictive analytics and decision support. Traditional healthcare systems rely heavily on manual diagnosis, which can be time-consuming and prone to human error. Moreover, access to healthcare professionals may be limited in rural or underdeveloped areas. This creates a gap where individuals may not receive timely medical attention. To bridge this gap, AI-based healthcare applications are being developed to provide quick and reliable preliminary assessments. The proposed system, “AI Medical Diagnosis — Advanced System,” is designed to predict multiple diseases using machine learning models trained on medical datasets. It incorporates five major disease prediction modules: Diabetes, Heart Disease, Parkinson’s Disease, Lung Cancer, and Hypothyroidism. Each module uses relevant clinical parameters to determine the likelihood of a disease. One of the standout features of this system is the integration of Optical Character Recognition (OCR). Medical reports are often available in image format, making it difficult to extract useful data manually. The OCR functionality allows users to upload their reports, from which the system automatically extracts key medical values. This significantly enhances user convenience and reduces dependency on manual data entry.

The system also emphasizes user engagement by providing personalized health suggestions. These include dietary recommendations, yoga practices, lifestyle changes, and emergency measures tailored to the predicted condition. Such guidance helps users take proactive steps toward improving their health. Another important aspect of the system is its integration with Google Maps for locating nearby doctors and hospitals. This feature ensures that users can easily find professional medical assistance based on their location and specific health condition. The application is built using Streamlit, which enables rapid development of interactive web applications. The backend utilizes Python libraries such as NumPy, Pandas, and Scikit-learn for data processing and model implementation. In summary, this project aims to create a comprehensive AI-based healthcare assistant that combines disease prediction, report analysis, and health guidance into a single platform. It not only demonstrates the practical application of machine learning in healthcare but also highlights the potential for improving accessibility and efficiency in medical services.

LITERATURE SURVEY (WITH EXISTING METHODS)

The application of machine learning in healthcare has been extensively studied, with numerous models developed for disease prediction and diagnosis. Early research focused on statistical methods such as logistic regression and decision trees to analyze medical data. These methods provided a foundation for predictive modeling but were limited in handling complex datasets. Recent advancements have introduced more sophisticated algorithms such as Support Vector Machines (SVM), Random Forests, and Neural Networks. Random Forest, in particular, has gained popularity due to its robustness and ability to handle high-dimensional data. Studies have shown that Random Forest models achieve high accuracy in predicting diseases like diabetes and heart conditions. In the domain of diabetes prediction, researchers have utilized datasets such as the Pima Indians Diabetes Dataset. Machine learning models like K-Nearest Neighbors (KNN), SVM, and Random Forest have been applied, with Random Forest often outperforming others in terms of accuracy and stability. Similarly, heart disease prediction systems have leveraged clinical parameters such as cholesterol levels, blood pressure, and ECG results to achieve reliable predictions.

Parkinson's disease prediction has been explored using voice signal analysis. Features such as jitter, shimmer, and harmonic-to-noise ratio are extracted and analyzed using machine learning algorithms. Studies indicate that SVM and ensemble methods provide high accuracy in detecting Parkinson's disease. Lung cancer prediction has traditionally relied on imaging techniques such as CT scans. However, recent approaches incorporate questionnaire-based data and machine learning models to predict risk factors. This approach simplifies the prediction process and makes it accessible without advanced imaging equipment. Thyroid disease prediction has also been widely studied, with models analyzing hormone levels such as TSH, T3, and T4. Machine learning algorithms have demonstrated high accuracy in classifying thyroid disorders. Another important area of research is the use of Optical Character Recognition (OCR) in healthcare. OCR enables the extraction of text from medical documents, facilitating data digitization. Tools like Tesseract OCR have been widely used for this purpose. Integrating OCR with machine learning systems enhances automation and reduces manual effort. Despite these advancements, most existing systems focus on a single disease or lack integration with user-friendly interfaces. Additionally, many systems do not provide actionable health recommendations or support features such as doctor consultation. The proposed system addresses these limitations by combining multiple disease prediction models, OCR-based data extraction, and personalized health guidance into a single platform. This integrated approach represents a significant improvement over existing methods and highlights the potential of AI in transforming healthcare services.

II. EXISTING SYSTEM

The existing systems for disease prediction are generally limited in scope and functionality. Most traditional healthcare applications focus on diagnosing a single disease using machine learning models. These systems require users to manually input medical data, which can be time-consuming and prone to errors. Additionally, they often

lack integration with real-world healthcare services, such as locating nearby doctors or providing actionable health recommendations. Many existing solutions are based on standalone machine learning models trained on specific datasets. While these models may achieve high accuracy, they do not offer a comprehensive healthcare solution. For example, a diabetes prediction system may only provide a binary output indicating whether the user is diabetic or not, without offering further guidance on managing the condition. Another limitation of existing systems is the absence of automation in data extraction. Medical reports are typically provided in image or PDF format, requiring manual interpretation. This creates a barrier for users who may not be familiar with medical terminology or data entry processes.

Furthermore, most systems lack user-friendly interfaces and are not accessible to non-technical users. They may require installation of software or technical expertise, limiting their usability. The absence of features such as prediction history tracking and personalized recommendations further reduces their effectiveness. In summary, existing systems are fragmented, focusing on isolated functionalities rather than providing an integrated healthcare solution. This highlights the need for a unified platform that combines disease prediction, data extraction, user interaction, and medical guidance.

III. PROPOSED METHOD

The proposed system is an advanced AI-based healthcare application designed to provide multi-disease prediction, automated medical report analysis, and personalized health recommendations within a single integrated platform. Unlike traditional systems that focus on a single disease, this system supports prediction for multiple conditions including diabetes, heart disease, Parkinson's disease, lung cancer, and hypothyroidism. The system utilizes machine learning models trained on relevant healthcare datasets to analyze patient input parameters and generate predictions. A key innovation of the proposed system is the integration of Optical Character Recognition (OCR), which allows users to upload medical reports in image format. The system extracts important clinical values such as glucose, cholesterol, TSH, T3, and T4 automatically, thereby reducing manual effort and improving efficiency. The application is developed using Streamlit, ensuring a user-friendly and interactive interface. Users can input data manually or rely on OCR-extracted values for prediction. The system then provides not only diagnostic predictions but also comprehensive health guidance, including diet plans, yoga exercises, precautions, home remedies, and emergency measures tailored to the predicted condition.

Additionally, the system incorporates a location-based feature that enables users to find nearby doctors and hospitals through Google Maps integration. This ensures that users can seek professional medical assistance immediately after receiving predictions. The system also maintains a prediction history, allowing users to track their health trends over time. This helps in long-term health monitoring and decision-making. Overall, the proposed system offers a holistic healthcare solution by combining prediction, analysis, guidance, and accessibility into a single platform.

IV. IMPLEMENTATION

The implementation of the AI Medical Diagnosis System involves integrating multiple technologies, including machine learning models, OCR processing, and a web-based user interface. The system is developed using Python and deployed using the Streamlit framework, which allows rapid creation of interactive web applications. The backend of the system consists of pre-trained machine learning models stored in serialized (.sav) format. These models are loaded dynamically at runtime using the pickle library. Each model corresponds to a specific disease and is trained using relevant datasets containing medical attributes. The system supports multiple models such as Random Forest, Support Vector Machine, and K-Nearest Neighbors, depending on the disease prediction requirements. For data input, the system provides two approaches: manual entry and image-based extraction. In manual entry, users input their health parameters directly through form fields. For automated input, the system uses OCR technology implemented via the Tesseract library. Uploaded medical report images are processed, and text is extracted. Regular expressions are then applied to identify and extract relevant medical values such as glucose levels, cholesterol, and thyroid hormone levels.

The extracted data is preprocessed to ensure consistency and accuracy. This includes converting text values into numerical formats, handling missing values, and normalizing data if required. The processed data is then passed to the respective machine learning model for prediction. The prediction module computes the output using the model's predict() function. If available, probability scores are also generated using predict_proba() to indicate the confidence level of the prediction. The results are displayed to the user in an intuitive format, along with visual indicators such as success or error messages. The system also includes a health recommendation module that provides disease-specific guidance. This module is implemented using predefined dictionaries containing information about diet, precautions, yoga exercises, and emergency measures. For location-based services, the system integrates Google Maps links to help users find nearby doctors and hospitals based on their selected disease and location input. This enhances the practical usability of the system. Additionally, the system maintains a prediction history using session state management in Streamlit. Users can view, export, or clear their history as needed. Overall, the implementation combines multiple technologies to create a seamless and efficient healthcare application that is both user-friendly and functionally robust.

V. ALGORITHMS

The proposed system utilizes multiple machine learning algorithms for disease prediction, each selected based on the nature of the dataset and prediction requirements. One of the primary algorithms used is the **Random Forest Classifier**, which is an ensemble learning technique. It constructs multiple decision trees during training and outputs the mode of their predictions. Random Forest is highly effective in handling large datasets and reduces overfitting by averaging multiple trees. Studies have shown that Random Forest provides high accuracy in disease prediction tasks due to its robustness and ability to handle nonlinear relationships. The **Support Vector Machine (SVM)**

algorithm is also employed for certain predictions. SVM works by finding an optimal hyperplane that separates data points into different classes. It is particularly effective in high-dimensional spaces and is widely used in medical diagnosis due to its precision.

Another algorithm used is **K-Nearest Neighbors (KNN)**, which is a simple yet effective classification technique. It classifies data points based on the majority class of their nearest neighbors. KNN is useful for datasets where patterns are locally distributed. For OCR-based data extraction, **Tesseract OCR** is used. It processes images and converts them into machine-readable text. Regular expressions are then applied to extract specific medical parameters from the text. Additionally, preprocessing techniques such as normalization, feature scaling, and handling missing values are applied to improve model performance. Feature selection is also considered to identify the most relevant attributes for prediction. These algorithms collectively ensure accurate, efficient, and reliable disease prediction, making the system suitable for real-world healthcare applications.

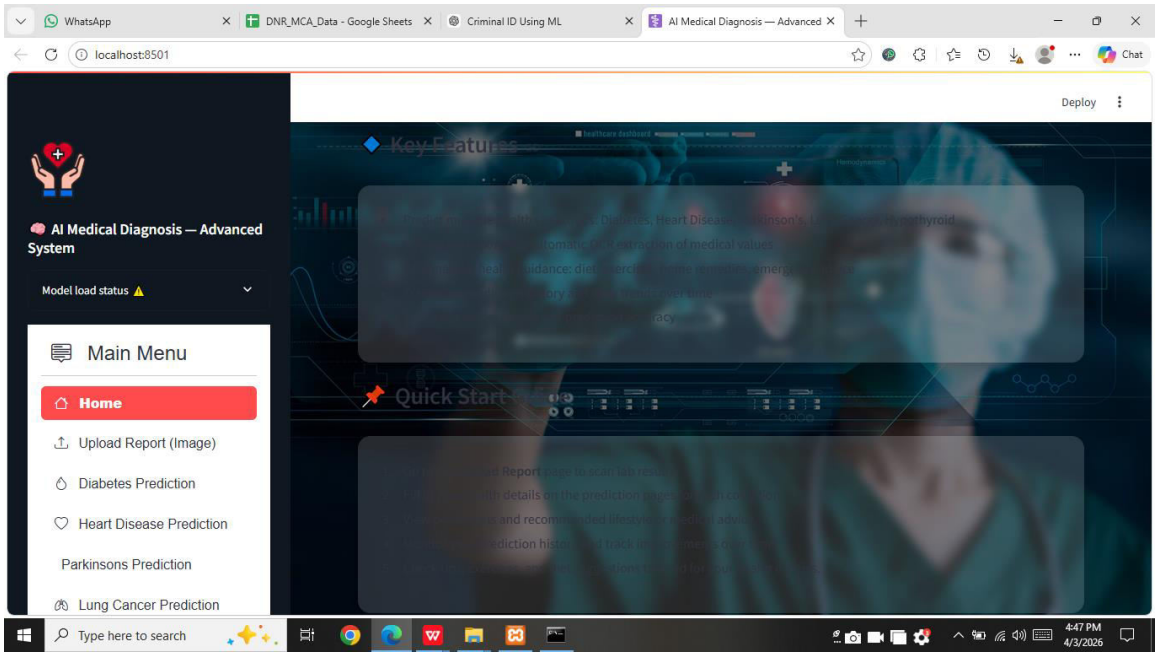
VI. SYSTEM DESIGN

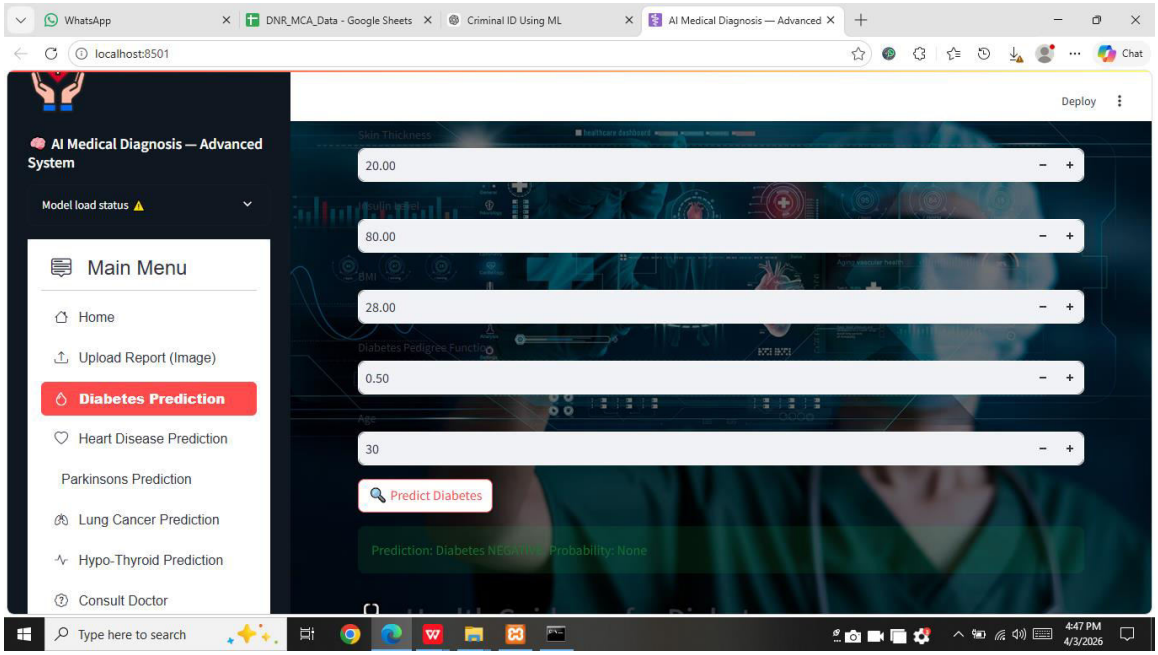
The system design of the AI Medical Diagnosis System follows a modular and layered architecture to ensure scalability, maintainability, and efficiency. The design consists of four major components: User Interface, Data Processing Layer, Machine Learning Layer, and Output & Recommendation Layer. The **User Interface (UI)** is developed using Streamlit, providing an interactive and intuitive platform for users. It includes multiple pages such as Home, Disease Prediction, Upload Report, Health Suggestions, and Consultation. The UI allows users to input data manually or upload medical reports for analysis. The **Data Processing Layer** handles input validation, preprocessing, and OCR extraction. When a user uploads an image, the system uses Tesseract OCR to extract text. Regular expressions are applied to identify relevant medical parameters. The extracted data is then cleaned and converted into numerical format. This layer ensures that the input data is consistent and suitable for machine learning models. The **Machine Learning Layer** is the core component of the system. It consists of pre-trained models for different diseases. Each model receives processed input data and generates predictions. The models are stored as serialized files and loaded dynamically. This layer also calculates probability scores to indicate prediction confidence.

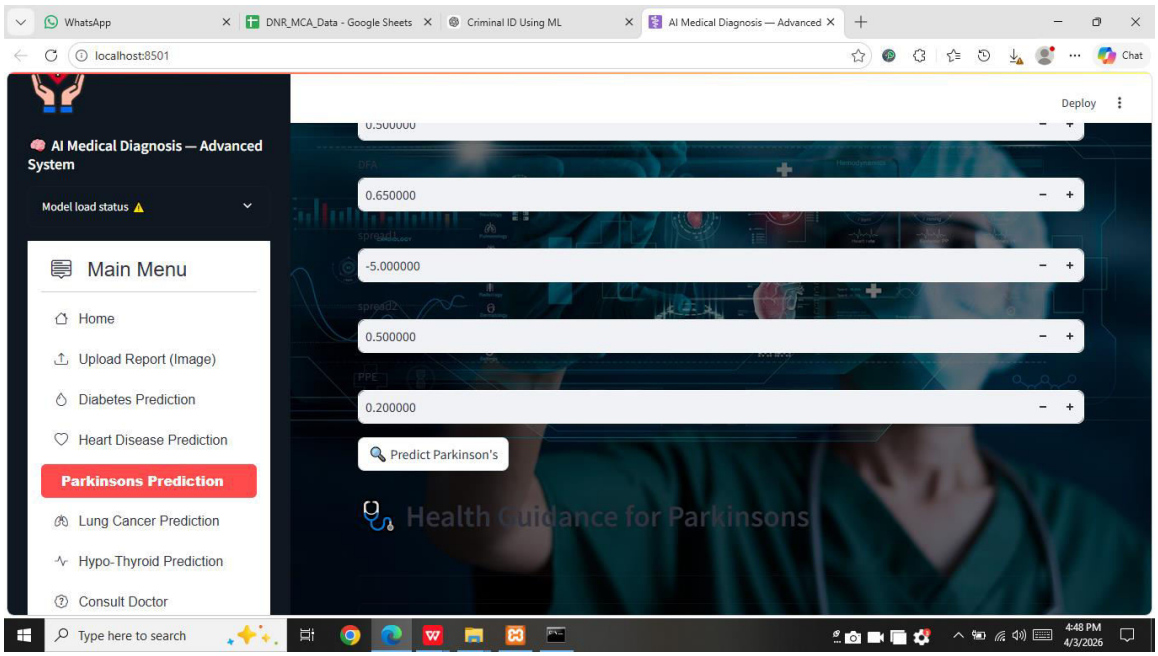
The **Output & Recommendation Layer** presents the results to the user. It displays whether the prediction is positive or negative, along with probability scores. Additionally, it provides detailed health recommendations, including diet plans, precautions, yoga exercises, and emergency measures. The system also includes a **History Management Module**, which stores prediction records in session state. Users can view their past predictions and export them as CSV files for further analysis. Another important component is the **Location-Based Service Module**, which integrates Google Maps to help users find nearby doctors and hospitals. This enhances the practical applicability of the system by connecting users with healthcare professionals. From a design perspective, the system follows a **modular architecture**, allowing easy addition of new diseases or

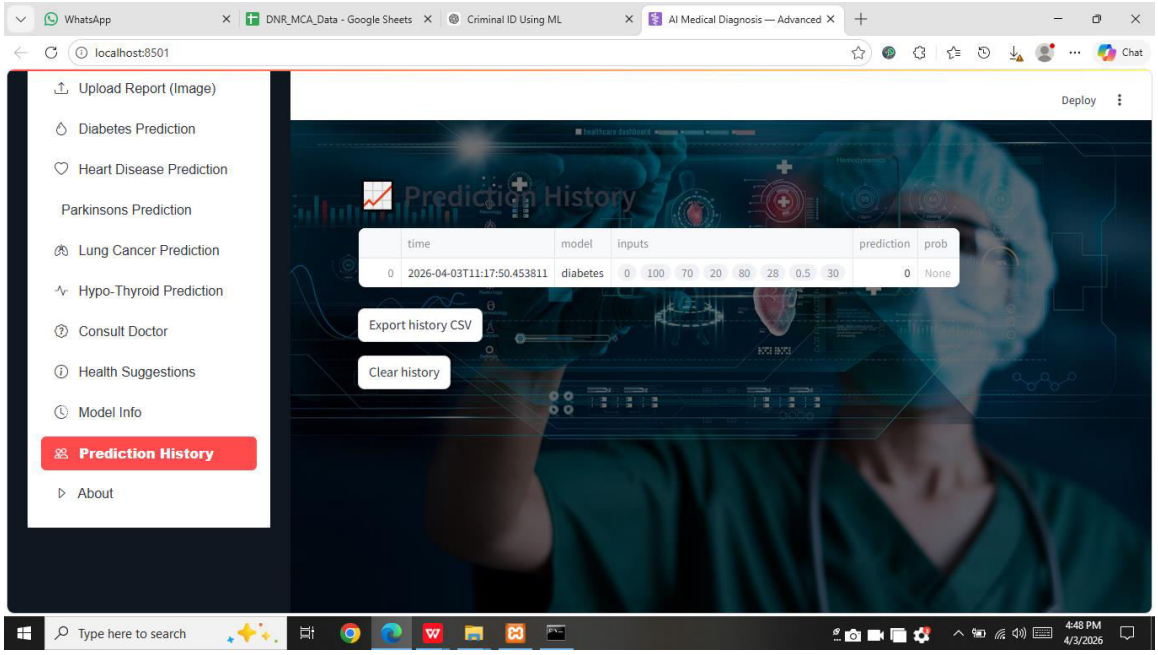
models in the future. Each module operates independently, ensuring flexibility and scalability. Overall, the system design ensures efficient data flow, seamless user interaction, and reliable prediction, making it a comprehensive healthcare solution.

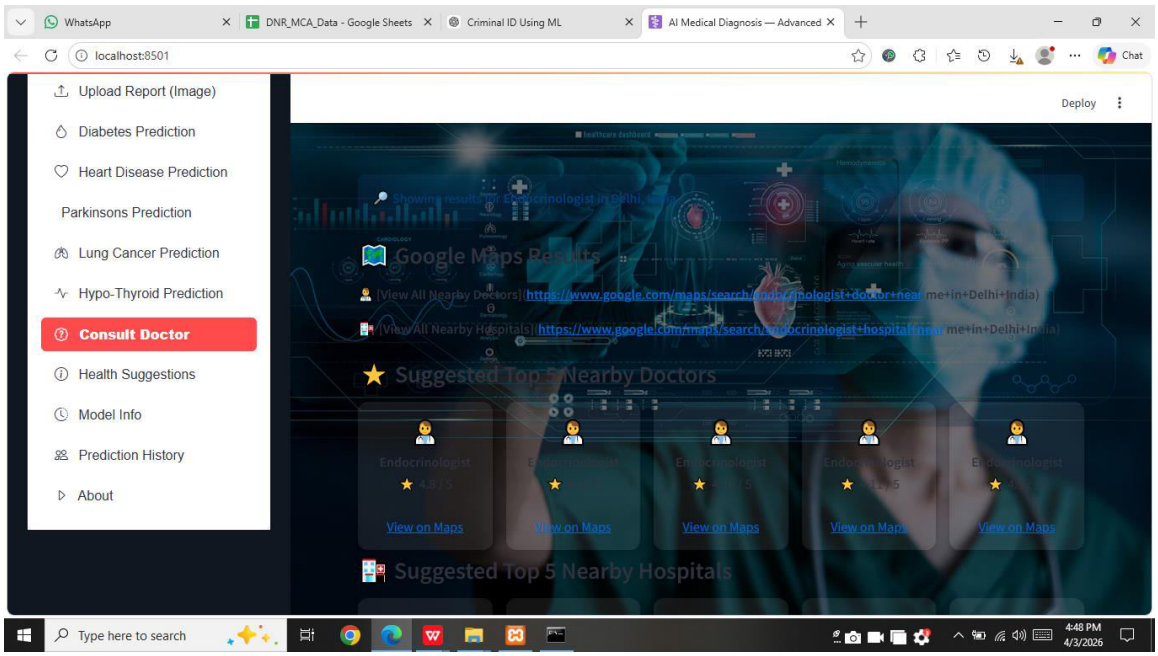
SYSTEM DESIGN IMAGES






AI Medical Diagnosis — Advanced System
Model load status
Main Menu
Home
Upload Report (Image)
Diabetes Prediction
Heart Disease Prediction
Parkinsons Prediction
Lung Cancer Prediction
Hypo-Thyroid Prediction
Consult Doctor
Deploy
0.650000
-5.000000
0.500000
0.200000
Predict Parkinson's
Health Guidance for Parkinsons


Upload Report (Image)
Diabetes Prediction
Heart Disease Prediction
Parkinsons Prediction
Lung Cancer Prediction
Hypo-Thyroid Prediction
Consult Doctor
Health Suggestions
Model Info
Prediction History
About
Deploy
Prediction History
time
model
inputs
prediction
prob
2026-04-03T11:17:50.453811
diabetes
Export history CSV
Clear history



VII. CONCLUSION

The AI Medical Diagnosis System represents a significant advancement in the application of machine learning in healthcare. By integrating multiple disease prediction models with OCR-based data extraction and personalized health recommendations, the system provides a comprehensive and user-friendly healthcare solution. One of the key strengths of the system is its ability to handle multiple diseases within a single platform. This eliminates the need for separate applications and provides users with a centralized healthcare assistant. The integration of OCR technology further enhances usability by allowing automatic extraction of medical data from reports, reducing manual effort and potential errors. The system not only predicts diseases but also provides actionable health guidance, including diet plans, lifestyle recommendations, yoga practices, and emergency measures. This makes it a holistic solution that supports both diagnosis and prevention. Another important feature is the integration of location-based services, enabling users to find nearby doctors and hospitals. This ensures that users can seek professional medical advice promptly after receiving predictions.

Despite its advantages, the system has certain limitations. The accuracy of predictions depends on the quality of the trained models and input data. Additionally, OCR extraction may not always be accurate for low-quality images. Therefore, the system should be used as a supportive tool rather than a replacement for professional medical diagnosis. Future enhancements may include the integration of deep learning models, real-time data from wearable devices, and cloud-based deployment for scalability. Incorporating electronic health records (EHR) and improving explainability of predictions can further enhance the system. In conclusion, the proposed system demonstrates the potential of AI in transforming healthcare by making disease prediction more accessible, efficient, and user-centric.

REFERENCES

1. Islam, R., et al. (2024). *Machine learning for chronic disease prediction*. Springer.
2. Saeed, M. K., et al. (2024). *Deep learning-based disease diagnosis model*. Scientific Reports.
3. Al-Alshaikh, H. A., et al. (2024). *Heart disease prediction using ML*. Scientific Reports.
4. Sharma, A., et al. (2024). *AI in predictive medicine*. Journal of Human Genetics.
5. Nurhalizah, R. S., et al. (2024). *ML in disease prediction review*.
6. Badawy, M., et al. (2023). *Healthcare predictive analytics survey*.
7. Sadr, H., et al. (2025). *AI in disease diagnosis review*.
8. Alhumaidi, N., et al. (2025). *ML for real-world healthcare data*.
9. Dawadi, R., et al. (2025). *Smartphone-based disease prediction*.
10. Padhy, N., et al. (2024). *Multi-disease prediction using AI*.
11. Dongre, S., et al. (2024). *MLtoGAI healthcare framework*.
12. Hennebelle, A., et al. (2023). *HealthEdge diabetes prediction*.
13. Miah, J., et al. (2023). *Cardiovascular prediction using ML*.
14. Guerra-Manzanares, A., et al. (2023). *Privacy-preserving ML in healthcare*.
15. Recent advancements in AI-based predictive healthcare systems (2025 review).