ABSTRACT

The rapid advancement of artificial intelligence and machine learning has opened new avenues for improving healthcare delivery. Early and accurate detection of diseases remains one of the most significant challenges in modern medicine, particularly in regions with limited access to specialized diagnostic facilities. This project presents an AI-Powered Healthcare Analytics and Disease Prediction System — a web-based application developed using Python and Flask that integrates multiple machine learning models, optical character recognition, and natural language processing to provide a comprehensive healthcare assistance platform.

The system is capable of predicting the likelihood of several common diseases including diabetes, heart disease, and Parkinson's disease based on clinical inputs provided by the user. The underlying prediction models were trained using publicly available medical datasets in a Jupyter Notebook environment, utilizing the Scikit-learn library. The trained models are saved as serialized pickle files and loaded into the Flask application for real-time predictions.

In addition to structured data-based prediction, the system includes a medical report analysis feature that allows users to upload scanned or digital reports. Tesseract OCR is used to extract textual content from these documents, after which a spaCy-based natural language processing pipeline identifies medically relevant keywords, abnormal values, and potential risk indicators. Based on the extracted information, the system assigns a risk score — categorized as Low, Medium, or High — and generates personalized health recommendations.

The frontend of the application is built using plain HTML and CSS without any JavaScript frameworks, ensuring simplicity and accessibility. All prediction data and user history are stored in a MySQL database, enabling health trend tracking over multiple sessions. The system is designed purely as an educational and assistive tool and does not replace professional medical consultation.

Results demonstrate that the diabetes prediction model achieves an accuracy of approximately 85%, the heart disease model reaches 87%, and the OCR extraction accuracy stands at around 90% for clearly scanned documents. This project illustrates how accessible web technologies and machine learning can be combined to create meaningful healthcare support tools that are both practical and easy to use.

TABLE OF CONTENTS

Certificate of Recommendation	i
Declaration	ii
Acknowledgement	iii
Abstract	iv
Table of Contents	v
List of Figures	vi
List of Tables	vii
Chapter 1: Introduction	1
Chapter 2: Problem Statement	4
Chapter 3: Objectives of the Project	6
Chapter 4: Scope of the Project	8
Chapter 5: Literature Review	10
Chapter 6: Proposed System	14
Chapter 7: System Architecture	17
Chapter 8: Flowchart of the System	20
Chapter 9: Methodology	22
Chapter 10: Tools and Technologies Used	26
Chapter 11: System Modules	28
Chapter 12: Implementation Details	32
Chapter 13: Source Code	37
Chapter 14: Output Screenshots	41
Chapter 15: Results and Discussion	44
Chapter 16: Limitations	46
Chapter 17: Future Scope	48
Chapter 18: Conclusion	50
Chapter 19: References	52