INFOSYS SPRING BOARD VIRTUAL INTERNSHIP – 6.0

Project Title

AI/ML-Based Personalized Diet Plan Generator from Medical Report

Kalakata Shanmukha Priya

ABSTRACT

The increasing volume and complexity of medical reports make it difficult for patients to understand their health conditions and follow appropriate dietary guidelines. This project presents an AI/ML-based personalized diet plan generator that analyzes medical reports and provides customized dietary recommendations. The system extracts numerical health parameters and textual medical information using data processing and natural language processing techniques. A machine learning model predicts the patient's health status as normal or abnormal. Advanced NLP models accessed through OpenAPI platforms such as Hugging Face are used to understand doctor prescriptions, identify medical conditions, and extract dietary constraints. Based on the predicted health status and extracted medical intent, the system generates a personalized diet plan. A Streamlit-based web interface enables users to upload reports and receive clear, actionable dietary guidance, improving accessibility and decision-making in healthcare.

INTRODUCTION

Medical reports are an essential part of healthcare, containing vital information such as laboratory test results, clinical observations, and doctor prescriptions. However, these reports are often difficult for patients to understand due to medical terminology and unstructured presentation. As a result, patients may fail to follow appropriate dietary and lifestyle recommendations, which can negatively impact their health.

With the advancement of Artificial Intelligence, Machine Learning, and Natural Language Processing, it is now possible to automate the analysis of medical data and convert it into meaningful insights. Personalized diet planning based on medical reports can help individuals manage conditions such as diabetes, hypertension, obesity, and nutritional deficiencies more effectively.

This project proposes an AI/ML-based personalized diet plan generator that analyzes medical reports to predict health status and generate customized dietary recommendations. By combining machine learning models with NLP techniques and OpenAPI-based models such as those available through Hugging Face, the system transforms raw medical reports into actionable diet guidance through a user-friendly web interface.

Milestone 1: Data Collection and Structuring of Medical Reports

1. Objective

The objective of this milestone is to extract structured medical information from medical reports available in PDF, image, and text formats.

2. Problem Statement

Medical reports are often available in unstructured formats, making automated analysis difficult. Key challenges include:

- Variability in report layouts
- Presence of scanned documents
- Inconsistent representation of medical parameters
- Difficulty in extracting structured values from free text

This milestone addresses these challenges by designing an automated data collection pipeline capable of handling real-world medical data.

3. Data Sources

The system supports the following input formats:

- PDF medical reports
- Scanned medical report images (.jpg, .png)
- Plain text medical reports (.txt)

4. Methodology

The data collection and extraction pipeline consists of the following steps:

4.1 PDF and Image Text Extraction

- Digital PDFs are processed using pdfplumber.
- Scanned PDFs and images are converted into text using Tesseract OCR (pytesseract).

4.2 Text File Processing

Medical reports available in .txt format are directly read with minimal preprocessing, ensuring compatibility with legacy systems.

4.3 Medical Information Extraction

Key medical information is extracted from text using Regular Expressions (Regex).

- The extracted parameters include:
- Patient demographics (age, gender)
- Vital signs (blood pressure, BMI)
- Laboratory values (blood sugar, cholesterol, hemoglobin, WBC, RBC)

This ensures consistent and structured medical data across all input sources.

5. Results

5.1 Numeric Data Extraction

The system successfully extracts the following numerical attributes:

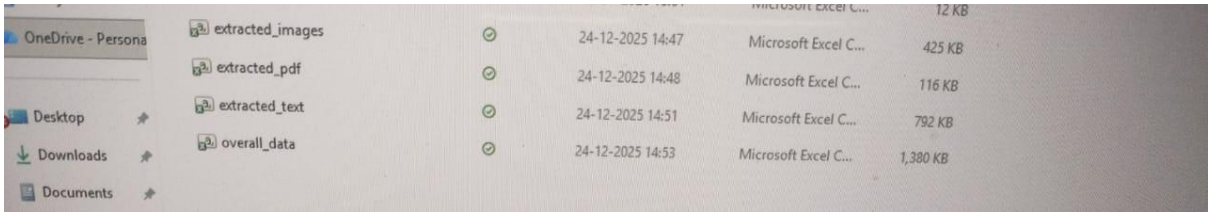
- Age
- Blood sugar
- Cholesterol
- Blood pressure (systolic/diastolic)
- BMI
- Hemoglobin
- WBC and RBC counts

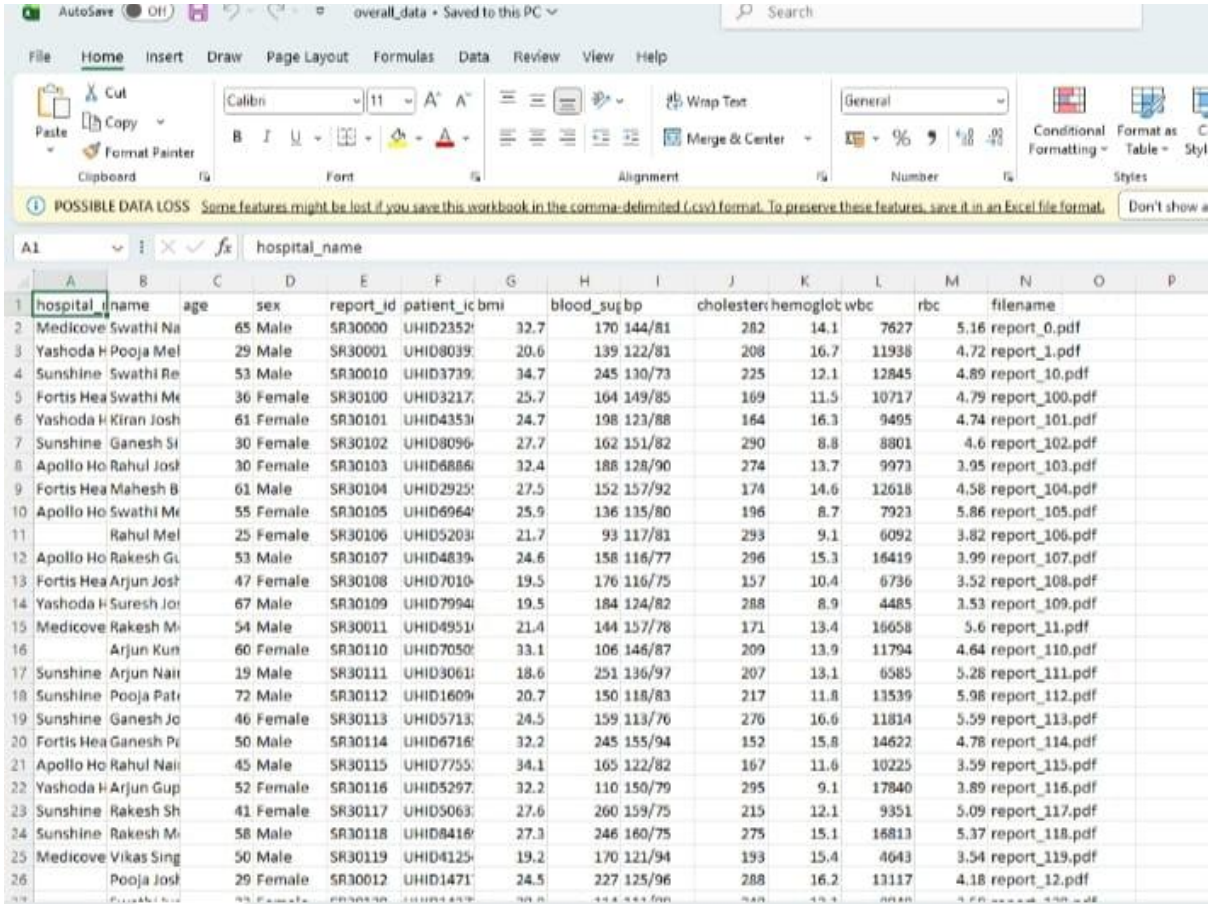
These features are used as inputs for further analysis and predictive modeling.

5.2 Doctor's Notes Extraction

Natural Language Processing (NLP) using SpaCy is applied to extract important information from clinical notes, including: Diagnosis, Prescriptions, Recommendations, Dietary advice.

6.Output





	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	hospital_n	name	age	sex	report_id	patient_ic	bmi	blood_sug	bp	cholesterc	hemoglot	wbc	rbc	filename		
2	Medicove	Swathi Na	65	Male	SR30000	UHID2352	32.7	170	144/81	282	14.1	7627	5.16	report_0.pdf		
3	Yashoda H	Pooja Mel	29	Male	SR30001	UHID8039	20.6	139	122/81	208	16.7	11938	4.72	report_1.pdf		
4	Sunshine	Swathi Re	53	Male	SR30010	UHID3739	34.7	245	130/73	225	12.1	12845	4.89	report_10.pdf		
5	Fortis Hea	Swathi Me	36	Female	SR30100	UHID3217	25.7	164	149/85	169	11.5	10717	4.79	report_100.pdf		
6	Yashoda H	Kiran Josh	61	Female	SR30101	UHID4353	24.7	198	123/88	164	16.3	9495	4.74	report_101.pdf		
7	Sunshine	Ganesh Si	30	Female	SR30102	UHID8096	27.7	162	151/82	290	8.8	8801	4.6	report_102.pdf		
8	Apollo Ho	Rahul Josh	30	Female	SR30103	UHID6886	32.4	188	128/90	274	13.7	9973	3.95	report_103.pdf		
9	Fortis Hea	Mahesh B	61	Male	SR30104	UHID2925	27.5	152	157/92	174	14.6	12618	4.58	report_104.pdf		
10	Apollo Ho	Swathi Me	55	Female	SR30105	UHID6964	25.9	136	135/80	196	8.7	7923	5.86	report_105.pdf		
11		Rahul Mel	25	Female	SR30106	UHID5203	21.7	93	117/81	293	9.1	6092	3.82	report_106.pdf		
12	Apollo Ho	Rakesh Gu	53	Male	SR30107	UHID4839	24.6	158	116/77	296	15.3	16419	3.99	report_107.pdf		
13	Fortis Hea	Arjun Josh	47	Female	SR30108	UHID7010	19.5	176	116/75	157	10.4	6736	3.52	report_108.pdf		
14	Yashoda H	Suresh Jos	67	Male	SR30109	UHID7994	19.5	184	124/82	288	8.9	4485	3.53	report_109.pdf		
15	Medicove	Rakesh M	54	Male	SR30011	UHID4951	21.4	144	157/78	171	13.4	16658	5.6	report_11.pdf		
16		Arjun Kun	60	Female	SR30110	UHID7050	33.1	106	146/87	209	13.9	11794	4.64	report_110.pdf		
17	Sunshine	Arjun Nair	19	Male	SR30111	UHID3061	18.6	251	136/97	207	13.1	6585	5.28	report_111.pdf		
18	Sunshine	Pooja Pate	72	Male	SR30112	UHID1609	20.7	150	118/83	217	11.8	13539	5.98	report_112.pdf		
19	Sunshine	Ganesh Jo	46	Female	SR30113	UHID5713	24.5	159	113/76	276	16.6	11814	5.59	report_113.pdf		
20	Fortis Hea	Ganesh Pa	50	Male	SR30114	UHID6716	32.2	245	155/94	152	15.8	14622	4.78	report_114.pdf		
21	Apollo Ho	Rahul Nair	45	Male	SR30115	UHID7755	34.1	165	122/82	167	11.6	10225	3.59	report_115.pdf		
22	Yashoda H	Arjun Gup	52	Female	SR30116	UHID5297	32.2	110	150/79	295	9.1	17840	3.89	report_116.pdf		
23	Sunshine	Rakesh Sh	41	Female	SR30117	UHID5063	27.6	260	159/75	215	12.1	9351	5.09	report_117.pdf		
24	Sunshine	Rakesh M	58	Male	SR30118	UHID8416	27.3	246	160/75	275	15.1	16813	5.37	report_118.pdf		
25	Medicove	Vikas Sing	50	Male	SR30119	UHID4125	19.2	170	121/94	193	15.4	4643	3.54	report_119.pdf		
26		Pooja Josh	29	Female	SR30012	UHID1471	24.5	227	125/96	288	16.2	13117	4.18	report_12.pdf		

6. Conclusion

Milestone 1 successfully delivers a robust and automated data collection framework for medical reports. By handling multiple input formats and standardizing extracted medical information.

Milestone 2: Data Processing, Model Building, and Evaluation

1. Objective

The objective of Milestone 2 is to preprocess the extracted medical data and develop machine learning models to predict patient health-related outcomes. This milestone focuses on cleaning and transforming raw medical data, training multiple classification models, evaluating their performance using standard metrics, and selecting the best-performing model.

2. Data Preprocessing

The following preprocessing steps were applied to improve data quality:

Preprocessing Steps Performed:

- Handling missing values to ensure data completeness
- Removal of noise and inconsistencies introduced during OCR and text extraction
- Feature normalization and scaling for uniform value ranges
- Encoding categorical features such as gender into numerical format
- Outlier handling to reduce the impact of extreme values

These preprocessing steps ensure improved data quality and enhance model performance.

3. Model Building

Multiple machine learning models were implemented and trained to identify the most accurate and reliable model.

Models Used

- Random Forest
- Light Gradient Boosting Machine (LightGBM)
- Extreme Classifier Gradient Boosting (XGBoost)

Each model was trained using the preprocessed dataset with appropriate hyperparameters to ensure fair performance comparison.

4. Model Evaluation

- The dataset was split into training and testing sets.

Models were evaluated using standard performance metrics such as:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion matrix

5.Model Selection

Based on evaluation results:

- Random Forest showed stable performance.
- XGBoost achieved improved predictive accuracy.
- LightGBM demonstrated the best overall performance in terms of accuracy and computational efficiency.

Hence, LightGBM was selected as the final model for health status prediction.

6.Outcome

- Successfully preprocessed medical datasets.
- Trained and evaluated multiple machine learning models.
- Selected an optimal prediction model for integration with NLP-based diet recommendation modules.

7.Conclusion

This milestone successfully completed:

- Data cleaning and preprocessing
- Implementation of multiple machine learning models
- Model evaluation and comparison
- Hyperparameter tuning for performance improvement

The optimized LightGBM model provides a strong foundation for the next milestone, where predictions will be integrated into the AI-based personalized diet plan generation system.

Milestone 3: NLP-Based Medical Text Analysis and Diet Plan Generation

1.Objective

The objective of this milestone is to analyze unstructured doctor prescriptions and medical notes, extract relevant medical intent, and generate personalized diet plans using Natural Language Processing (NLP) and Generative AI techniques. The system aims to convert complex medical text into structured, patient-friendly dietary recommendations**.**

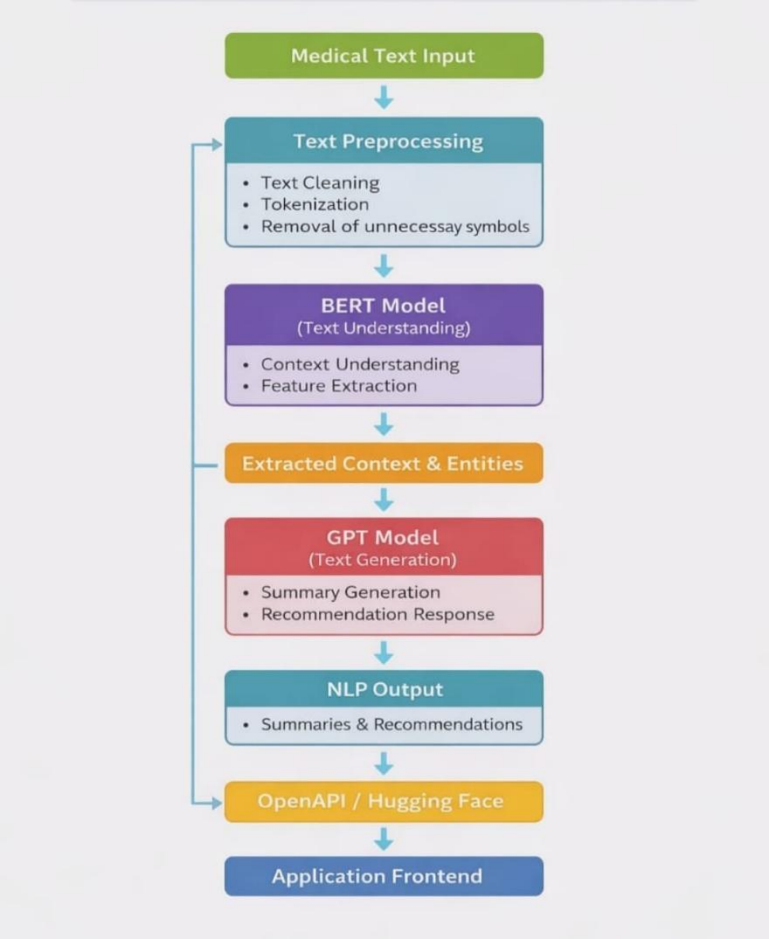
2.Problem Statement

Doctor prescriptions and medical notes are unstructured and vary in format and terminology, making automated processing difficult. Therefore, an NLP-based system is required to:

- Extract medical entities
- Identify dietary constraints
- Determine patient intent
- Generate personalized diet recommendations

This milestone addresses the conversion of unstructured medical text into structured data and diet plans using NLP and generative AI.

3.Methodology:



4. Medical Entity Recognition and Intent Extraction

4.1 BERT-Based Named Entity Recognition (NER)

A BERT-based NER model is used to extract key medical entities from prescription text, including:

- Diseases and medical conditions
- Nutritional components
- Dietary restrictions
- Medications and supplements

4.2 Medical Intent Classification

BERT is used to classify the intent of the prescription text, such as:

- Dietary restriction
- Nutritional recommendation
- Lifestyle advice

This step converts unstructured text into structured medical intent.

5. Structured Medical Intent Representation

The extracted information is organized into a structured format containing:

- Medical condition(s)
- Dietary restrictions
- Recommended nutrients

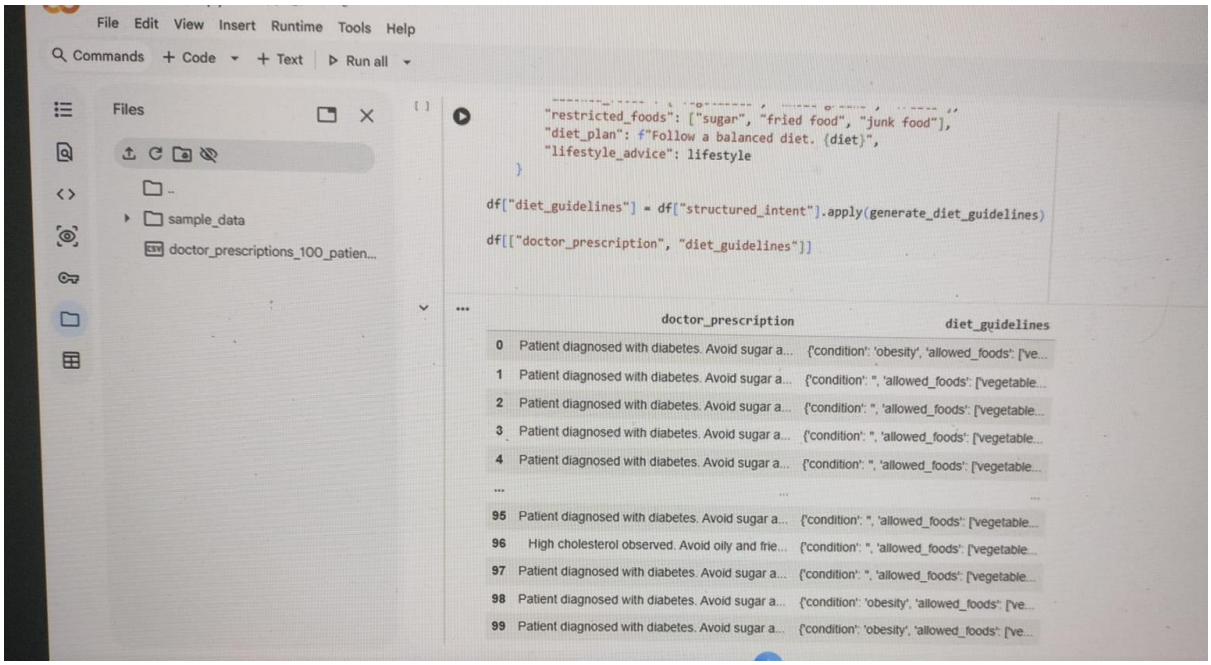
6. GPT-Based Diet Plan Generation

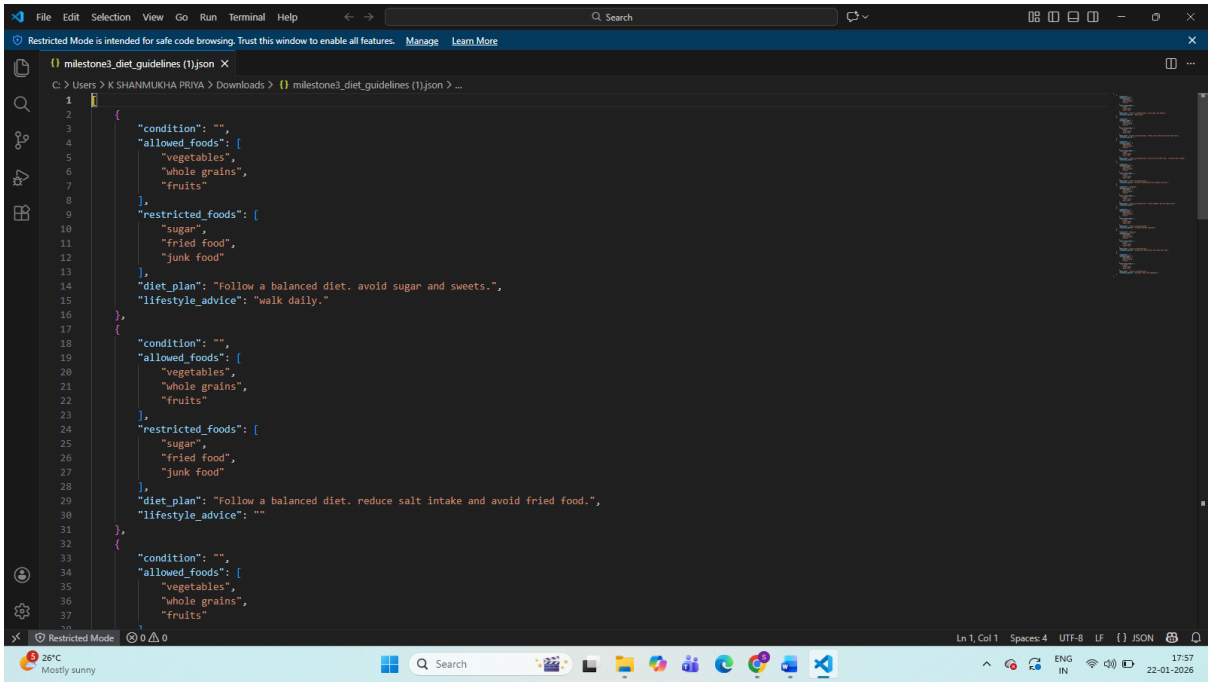
The structured medical intent is provided as input to a GPT-based generative model to produce:

- Personalized meal suggestions
- Foods to include and avoid
- Portion control guidance
- Basic lifestyle tips

The generated diet plan is aligned with medical conditions and extracted prescription details.

7.Output





8.Conclusion

This milestone successfully applies NLP techniques to extract medical entities and intent from unstructured prescriptions. By integrating BERT-based models with GPT-based generation, the system produces structured and personalized diet plans. The approach improves automation, accuracy, and usability of medical diet recommendations.

Milestone 4: Deployment and System Integration

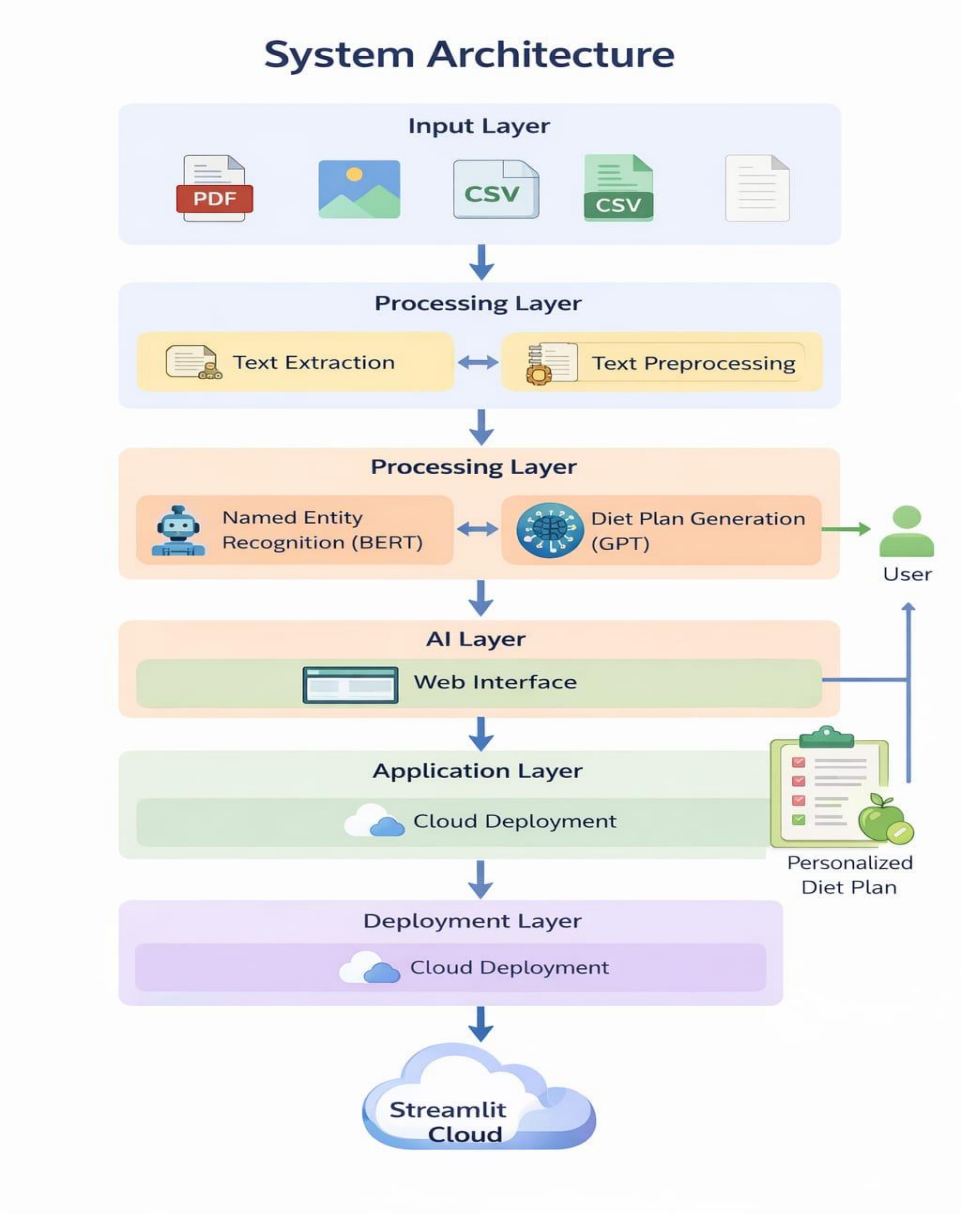
1.Objective

The objective of this milestone is to integrate all project modules into a single system and deploy it as a user-friendly web application.

2.Problem Statement

Earlier milestones developed individual ML and NLP components. These components must be combined into a unified system that works end-to-end and is accessible to non-technical users.

3.System Architecture:



4.System Integration

- All data extraction, prediction, and NLP modules are combined into a single end-to-end workflow.
- Ensures smooth data flow between preprocessing, analysis, and output generation.

5.Deployment Platform

- The application is deployed using Streamlit Cloud.
- Supports cloud-based execution with GitHub integration for easy updates.

6.User Interface

- Allows users to upload files such as PDFs, images, CSVs, and text.
- Displays extracted information and generated diet plans.
- Provides options to download personalized outputs.

7.Testing and Validation

- Tested using multiple input formats including PDFs, images, CSV files, and text.
- All system components performed as expected.

8.Challenges and Solutions

- Deployment issues were handled by optimizing dependencies.
- Lightweight NLP pipelines were used to improve performance.
- New models or features can be integrated without major changes.
- System handles user data securely during processing.

9.Final Outcome

A fully functional, cloud-deployed AI-powered personalized diet plan generator.

Conclusion

This milestone successfully transforms the project into a real-world, deployable AI healthcare solution.