## Anemia Disease Prediction using Machine Learning

TEAM MEMBER: Bandaru Raatna Sai

# INTRODUCTION

## 1.1Project Overview

**AnemiaPredict** is a machine learning–based disease prediction system designed to detect and predict anemia in individuals using clinical and biological data. The system leverages patient health records and laboratory test parameters such as hemoglobin levels, red blood cell count, hematocrit, and other key biomarkers. Machine learning models such as Random Forest, Support Vector Machine (SVM), and Logistic Regression are utilized to classify whether a patient is anemic or non-anemic. For forecasting anemia risk trends over time, deep learning models like LSTM can be employed to analyze longitudinal health data. All results are visualized through an interactive dashboard that displays patient health status, prediction results, and data-driven insights. The project aims to assist healthcare professionals in early diagnosis, better disease management, and improved patient outcomes through intelligent, data-driven analysis.

* 1. **Purpose**

The purpose of this project is to develop an intelligent and efficient system for predicting anemia using machine learning techniques. By automating the analysis of patient health data, the system aims to assist healthcare professionals in the early detection and diagnosis of anemia. This will help in identifying at-risk individuals, supporting timely medical intervention, and improving overall patient care. The project ultimately seeks to enhance diagnostic accuracy and enable data-driven decision-making in the medical field.

# IDEATION PHASE

* 1. **Problem Statement**

Anemia is a widespread health condition affecting millions of people globally, particularly women and children. Early detection and diagnosis remain a major challenge due to the reliance on traditional diagnostic methods, which are often time-consuming, costly, and dependent on manual interpretation of laboratory results. In many cases, limited access to healthcare facilities and lack of awareness further delay diagnosis and treatment. Additionally, the absence of automated and predictive tools makes it difficult for healthcare professionals to identify at-risk individuals efficiently. Therefore, there is a pressing need for a machine learning–based system that can accurately predict anemia using patient health data, enabling timely medical intervention and improved healthcare outcomes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PS-1 | Traditional Diagnosis | Most healthcare systems rely on manual interpretation of lab test results to detect anemia | Current diagnostic methods are time-consuming and often depend on human expertise. | Lack of automation and integration between different medical data sources reduces accuracy and efficiency. | Late diagnosis and missed cases of anemia, especially in rural or under-resourced areas. |
| PS-2 | Data Availability & Prediction | Limited access to historical patient data hinders effective pattern analysis and prediction | Existing systems provide only static diagnosis without predictive capability. | Accuracy drops when data quality is low or incomplete; models are not optimized for diverse populations. | Delayed treatment, reduced effectiveness of healthcare planning, and poor monitoring of anemia prevalence. |

* 1. **Empathy Map Canvas — Anemia Disease Prediction System**

Healthcare professionals, such as doctors, lab technicians, and medical researchers, often face challenges in diagnosing anemia accurately and efficiently. They require a reliable, data-driven system that provides clear insights into patient health parameters and supports early detection of anemia. These users frequently hear patient complaints related to fatigue, weakness, and delayed diagnosis, which adds to their workload and concern for patient care

|  |  |  |
| --- | --- | --- |
|  | C˛\* **Thinks**.  They think current anemia detection methods are time-consuming and dependent on manual interpretation.  They believe a smart, automated prediction system can improve diagnostic accuracy and save time. |  |
| }·.)◆● **Sees**  They see increasing cases of undiagnosed or late-diagnosed anemia among patients.  They observe that existing diagnostic systems fail to integrate patient data effectively for prediction. | **USER** | 🗣 **Says**  \_They say, “We need AI-based tools to help identify anemia earlier.”  They often mention, “Manual testing and interpretation delay diagnosis and increase errors.” |
| .◆ **Feels**  They feel concerned about misdiagnosis or delayed detection in patients.  They feel motivated to adopt innovative tools that improve medical efficiency and patient care. | 趴† **ears**  They hear frequent patient complaints about fatigue and delayed diagnosis.  They hear suggestions from medical authorities to integrate AI and machine learning in clinical workflows. | 🛠 **Does**  They manually review lab reports and compare test results for signs of anemia.  They make treatment recommendations based on traditional thresholds without predictive insights. |

* 1. **Brainstorming**

### Step-1: Individual Ideation and Problem Selection During the brainstorming phase, various ideas related to healthcare and disease prediction were explored to identify a suitable problem for implementation using machine learning. After evaluating multiple options such as diabetes detection, heart disease prediction, and anemia classification, the focus was directed toward Anemia Disease Prediction due to its medical significance and data availability.

### The idea was to design an intelligent system capable of predicting anemia based on key medical parameters such as hemoglobin level, red blood cell count, hematocrit, and mean corpuscular volume. Different machine learning algorithms, including Logistic Regression, Random Forest, and Support Vector Machine (SVM), were considered for classification.

### Problem Statement:

## Anemia has become a major global health concern due to its increasing prevalence and the limitations of existing diagnostic methods. Traditional approaches, such as manual evaluation of blood reports and clinical observation, are often time-consuming, error-prone, and lack the ability to provide quick and reliable insights

## Step-2: Brainstorm, Idea Listing and Grouping

|  |  |
| --- | --- |
| **Idea Category** | **Ideas Generated** |
| Technology/Tools | |  | | --- | |  |  |  | | --- | | During the brainstorming phase, tools like  **Google Jamboard** and **Trello** were used to  organize and refine ideas efficiently.  Machine learning algorithms such as  **Logistic Regression**, **Random Forest**, and  **Support Vector Machine (SVM)** were  shortlisted for implementing the anemia  prediction model. | |
| User Interaction | |  | | --- | |  |  |  | | --- | | Users interact with the system through an  intuitive **web-based dashboard** that  displays prediction results and key health  indicator | |

|  |  |
| --- | --- |
| Data Collection | \_- Medical data is collected from publicly available healthcare datasets and hospital records containing patient parameters such as **hemoglobin**, **hematocrit**, **RBC count**, **MCV**, and **MCH** values. |
| Deployment\_ | |  | | --- | |  |  |  | | --- | | - The system can be deployed on **c**  **loud platforms** such as AWS, Google Cloud,  or Azure to enable secure and scalable  access. | |
| Integration | |  | | --- | |  |  |  | | --- | | - Users interact with the system through an  **interactive dashboard** that displays  patient details, prediction results, and  health metric | |
| Awareness/Training | * - The system promotes **awareness of anemia** by providing easy-to-understand prediction outcomes, helping healthcare professionals and patients recognize early warning signs. |

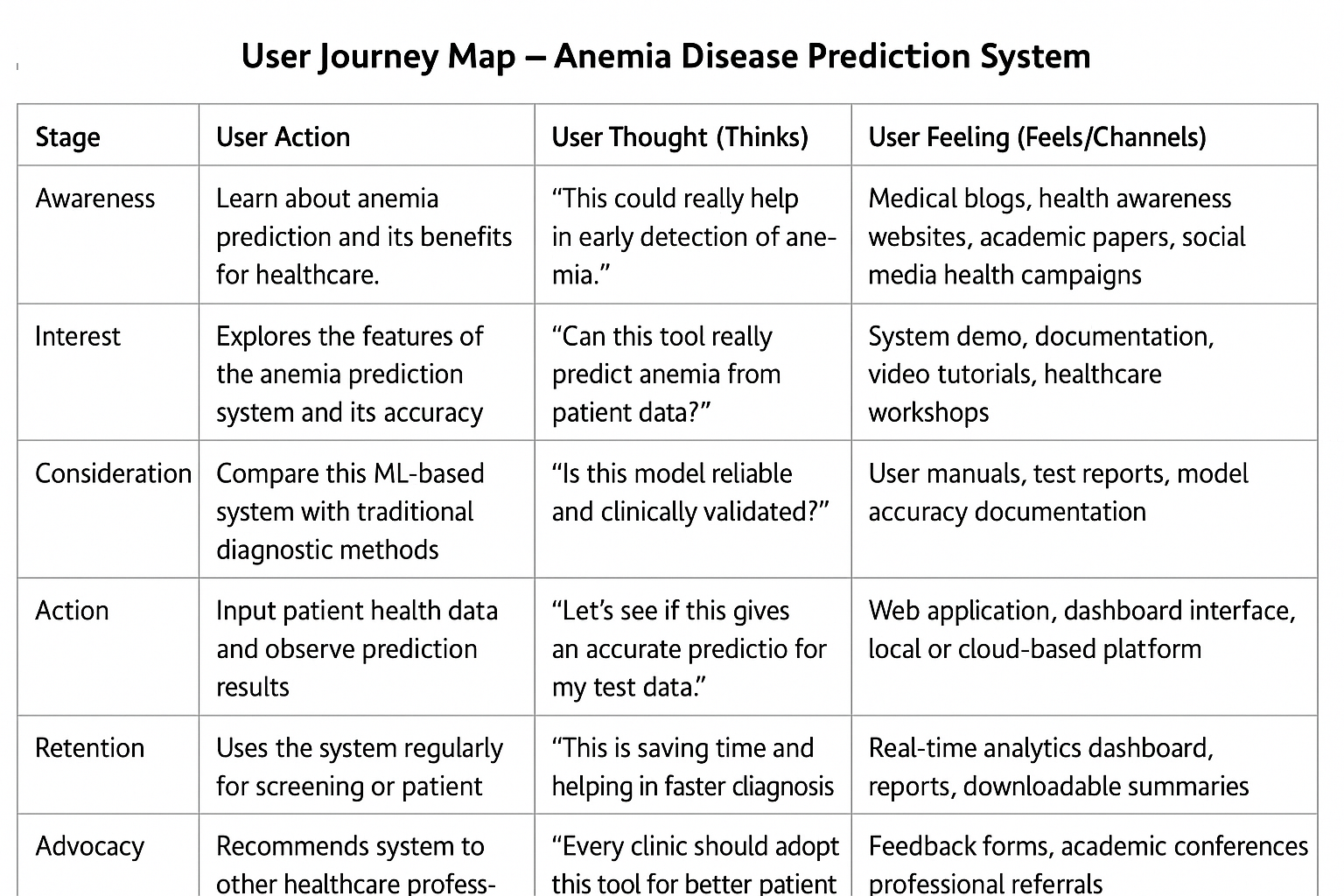
### Step-3: Idea Prioritization

|  |  |  |
| --- | --- | --- |
| **Idea** | **Impact (High/Med/Low)** | **Feasibility (High/Med/Low)** |
| Implementation of machine learning algorithms (Logistic Regression, Random Forest, SVM) | High | High |
| Integration of deep learning model (LSTM) for trend prediction | High | Medium |
| Development of a web-based or mobile application for anemia prediction | High | High |
| Inclusion of regional language and voice support for accessibility | High | Medium |
| Collection and augmentation of medical datasets for model training | High | Medium |
| Cloud-based deployment for hospital and laboratory access | High | High |
| Integration with electronic health records (EHR) for automated data input | Medium | Medium |

Ideas included using mobile apps, transfer learning with image classification, multilingual support, offline model access, disease history tracking, and farmer education modules.

# REQUIREMENT ANALYSIS

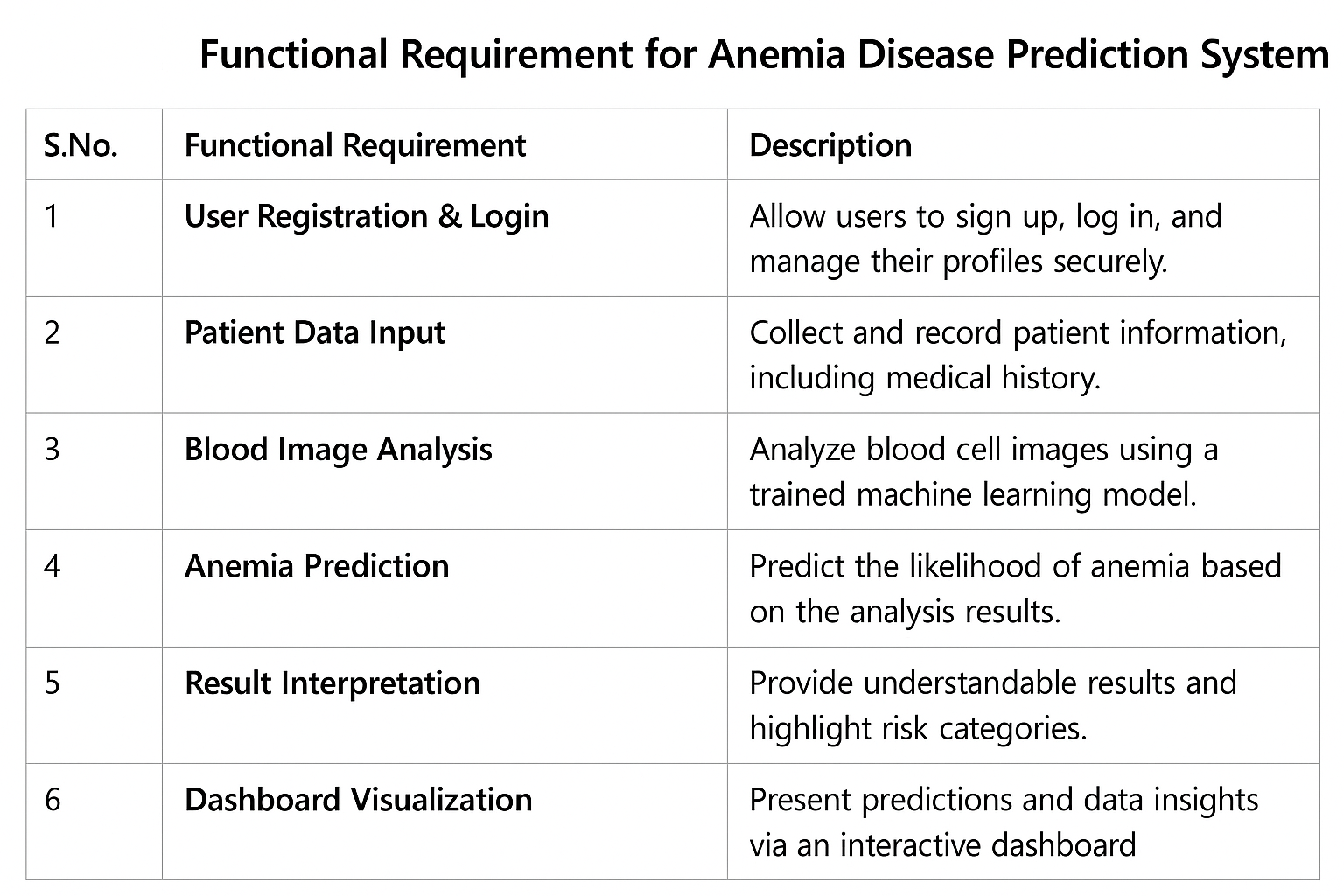
* 1. **Customer Journey Map**

****

**.**

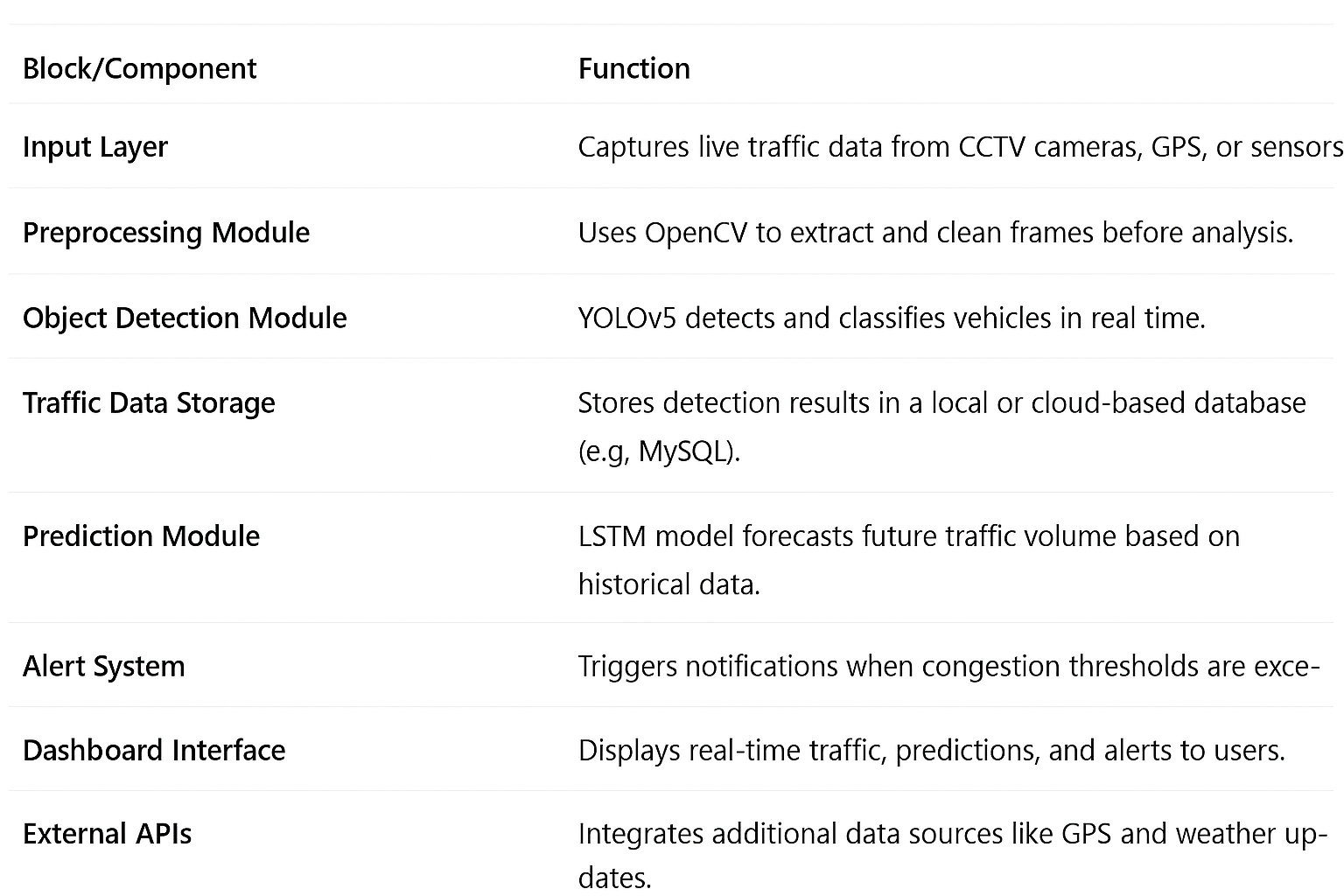
* 1. **Solution Requirement**

**Functional Requirements**

****

Functional requirements include image upload, disease diagnosis, history logs, notifications, multilingual support. Non-functional requirements ensure performance, security, offline access, and scalability.

* 1. **Data Flow Diagram**

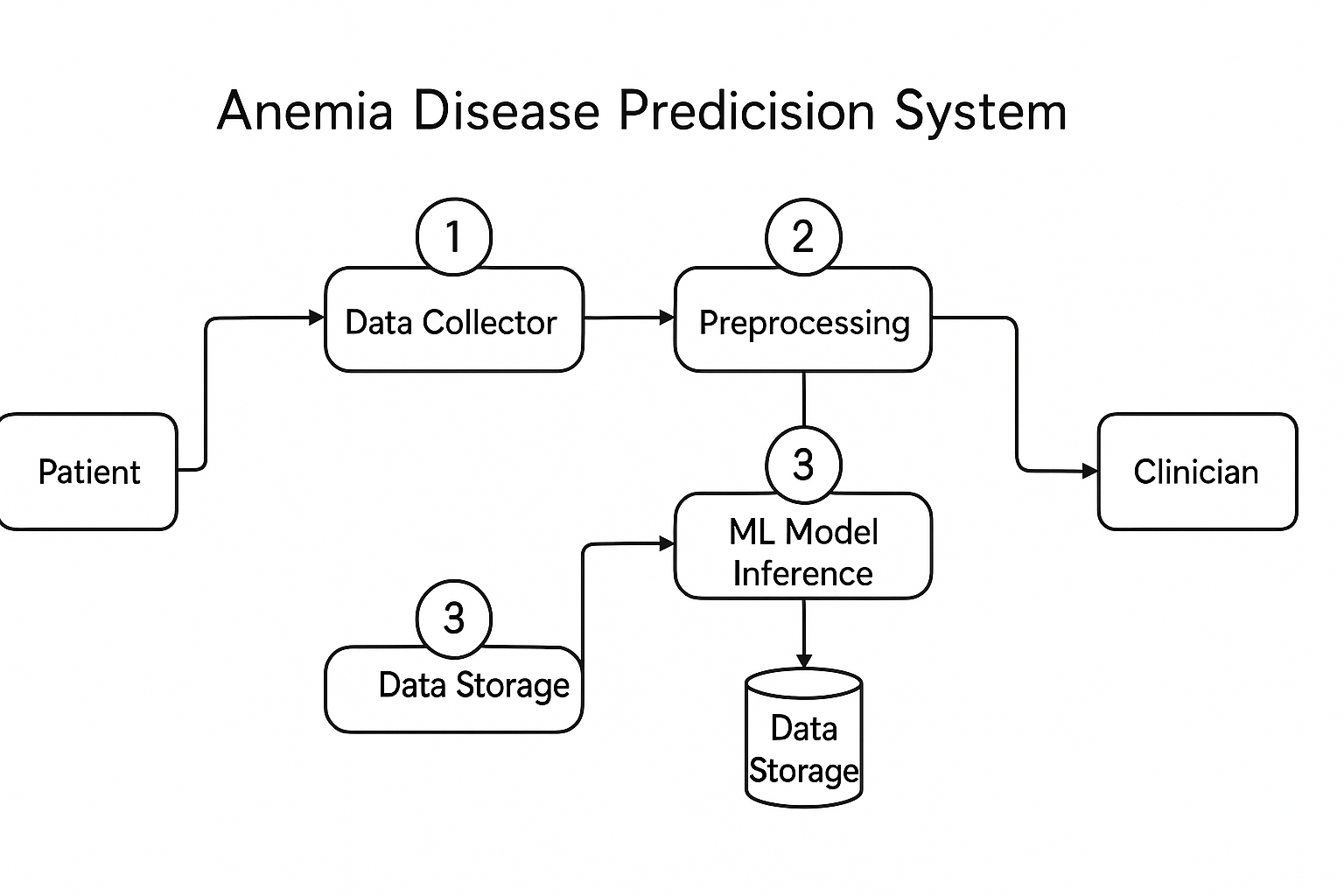


#### Data Flow Diagrams:

The Data Flow Diagram (DFD) for the TrafficTelegence system illustrates the movement of data through various components of the system. It begins with input from CCTV cameras and GPS devices, which is processed through modules like vehicle detection, traffic prediction, and data storage. The processed information is then visualized on a dashboard, allowing users to monitor live traffic and receive timely alerts.



**Data Flow Diagram – Level 0 (Simplified)**



* 1. **Technology Stack**

**Table-1: Components & Technologies**

|  | **Component** | **Description** | **Technology** |
| --- | --- | --- | --- |
| **1** | **User Interface** | The User Interface provides a web-based dashboard to display patient data, anemia prediction results, and health insights in an interactive and user-friendly manner. | Developed using **HTML, CSS, JavaScript**, and **Flask/Django** for backend integration. |
| **2** | **Application Logic – 1 (Data Preprocessing)** | Handles cleaning, normalization, and feature extraction from medical datasets such as hemoglobin, RBC count, and hematocrit values. | Implemented in **Python** using **Pandas, NumPy, and Scikit-learn** for data processing. |
| **3** | **Application Logic – 2 (Model Training & Prediction)** | Trains machine learning models to classify whether a patient is anemic or non-anemic based on input features. | Built using **Random Forest, Logistic Regression, or SVM** via **Scikit-learn**. |
| **4** | **Application Logic – 3 (Deep Learning Model)** | Uses neural networks (e.g., LSTM or ANN) to analyze longitudinal or complex datasets for improved prediction accuracy. | Implemented using **TensorFlow** or **PyTorch**. |
| **5** | **Database** | Stores patient information, lab test values, and prediction outcomes securely. | **MySQL** or **MongoDB** for structured medical data storage. |
| **6** | **Cloud Database** | Provides scalable and remote access to patient records, predictions, and analytics. | **Firebase** or **AWS RDS** for cloud data management. |
| **7** | **File Storage** | Saves uploaded patient datasets, trained models, and generated reports securely. | **AWS S3**, **Google Cloud Storage**, or **local file system**. |
| **8** | **External API – 1** | Integrates healthcare APIs for accessing public medical datasets or laboratory data. | Connected via **Kaggle Datasets API** or **Healthcare Data APIs**. |
| **9** | **External API – 2** | Supports additional data sources such as wearable health devices or mobile health apps. | Integrated through **Google Fit API** or **IoT-based patient monitoring systems**. |
| **10** | **Machine Learning Model** | Predicts anemia status based on patient health parameters and trained algorithms. | Models built using **Scikit-learn**, **TensorFlow**, or **PyTorch**. |
|  |  |  |  |

**Table-2: Application Characteristics**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Characteristics** | **Description** | **Technology** |
| **1** | **Open-Source Frameworks** | Utilizes open-source tools and libraries for efficient model development, testing, and deployment. This ensures flexibility, transparency, and strong community support for continuous improvement. | Built using **Scikit-learn**, **TensorFlow**, **PyTorch**, and **Flask/Django** for model development, training, and integration. |
| **2** | **Security Implementations** | Ensures the privacy and protection of sensitive patient data through secure authentication and encrypted communication protocols. | Implements **HTTPS**, **JWT (JSON Web Tokens)**, and **OAuth2** for secure API access and data protection. |
| **3** | **Scalable Architecture** | Designed with modular components and cloud integration to manage increasing amounts of patient data efficiently. | Uses **Docker**, **Kubernetes**, and cloud platforms such as **AWS**, **Google Cloud**, or **Azure** for scalability and deployment. |
| **4** | **Availability** | Provides reliable access and high system uptime, ensuring uninterrupted use by healthcare professionals. | Achieved through **AWS EC2 auto-scaling**, **load balancers**, and **cloud monitoring tools** for consistent service availability. |
| **5** | **Performance** | Optimized for fast data processing, accurate prediction, and efficient handling of medical datasets. | Enhanced using **GPU acceleration (CUDA)**, **data caching**, and **optimized ML inference pipelines** for high-speed computation. |

**4. PROJECT DESIGN**

**4.1 Problem–Solution Fit**

**Problem**

The **Anemia Disease Prediction System** addresses the challenges of traditional anemia diagnosis methods by leveraging **machine learning and health data analytics**. Conventional medical assessments often rely on manual interpretation of lab results, which can be time-consuming, error-prone, and limited in scope. By applying machine learning algorithms to patient data such as **hemoglobin level.**

**Target Customer**

**Hospitals and Healthcare Centers** — Require efficient diagnostic tools to identify anemia quickly and accurately.

**Clinical Laboratories** — Need automated systems for processing patient test data.

**Medical Researchers** — Use predictive analytics for studying anemia trends and correlations.

**Public Health Organizations** — Utilize prediction data to monitor anemia prevalence and design awareness campaigns.

**Current Behavior (Without the Solution)**

Manual diagnosis based on lab test interpretation.

Delays in identifying anemia due to lack of automation.

Limited accessibility to predictive or preventive healthcare tools.

Inconsistent medical record tracking and analysis.

**Pain Points**

Inaccurate or delayed detection of anemia.

Dependence on manual methods for diagnosis and analysis.

Difficulty in predicting anemia risk before clinical symptoms appear.

Lack of integration between patient data sources and healthcare systems

**Proposed Solution**

The **Anemia Disease Prediction System** utilizes **machine learning models** (such as Logistic Regression, Random Forest, and SVM) to predict anemia from patient health data. It automates data preprocessing, classification, and result visualization through a user-friendly dashboard. The system can be extended with **deep learning (LSTM/ANN)** for trend prediction and integrated with healthcare APIs for broader data access. This solution helps medical professionals detect anemia earlier and with greater precision, improving patient care and decision-makin

**Benefits / Improvements**

**Higher Accuracy & Faster Diagnosis** — Machine learning improves the precision of anemia classification.

**Automated & Efficient Processing** — Reduces manual workload for healthcare professionals.

**Early Detection** — Enables proactive treatment and improved patient outcomes.

**4.2 Proposed Solution**

|  |  |
| --- | --- |
| **Parameter** | **Description** |
| **Problem Statement (Problem to be solved)** | Anemia diagnosis often depends on manual interpretation of blood test reports, which can be slow, error-prone, and inaccessible in remote areas. The absence of automated diagnostic tools limits early detection, leading to delayed treatment and increased health risks. Accurate prediction of anemia requires analyzing complex, non-linear relationships between multiple medical parameters such as hemoglobin, hematocrit, and red blood cell count. |
| **Idea / Solution Description** | **AnemiaPredict** is a machine learning–based healthcare application designed to predict anemia using patient health data. It employs classification algorithms such as Logistic Regression, Random Forest, and Support Vector Machine (SVM) to identify anemia based on blood parameters. The system processes clinical data, performs analysis, and visualizes the results through an interactive web dashboard for doctors and patients. |
| **Novelty / Uniqueness** | Unlike traditional diagnostic methods that rely solely on laboratory analysis or physician interpretation, **AnemiaPredict** offers several unique advantages: it automates the prediction process using ML models trained on real patient datasets, provides instant risk classification, and reduces diagnostic errors. The system can also forecast anemia trends using deep learning (LSTM), enabling predictive healthcare insights. |
| **Social Impact / Customer Satisfaction** | The system supports early anemia detection, improving patient outcomes through timely medical intervention. It benefits healthcare institutions by saving time, reducing workload, and enhancing diagnostic reliability. For patients, it increases awareness, ensures faster results, and contributes to better public health through accessible technology-driven screening. |
| **Business Model (Revenue Model)** | The solution can be offered as a **subscription-based healthcare application**, where hospitals, laboratories, or clinics subscribe monthly or annually for access to the prediction dashboard and analytics reports. Tiered pricing can be implemented based on the number of users, data volume, or advanced analytics features enabled. |
| **Scalability of the Solution** | The system’s **cloud-based architecture** allows for scalable deployment across hospitals and diagnostic centers. It supports distributed data processing, enabling the integration of additional medical datasets or new disease prediction modules (e.g., diabetes, heart disease) without performance degradation. The solution can easily scale to national or global health networks. |

**4.3 Solution Architecture**

Solution architecture is a complex process – with many sub-processes – that bridges the gap The **Anemia Disease Prediction System** follows a **multi-layered, AI-driven architecture** that integrates data acquisition, preprocessing, machine learning analytics, and result visualization to detect and predict anemia accurately and efficiently.

The architecture bridges the gap between **healthcare challenges and technological solutions**, with the following goals:

* To identify the best AI and ML approaches to solve diagnostic challenges in anemia detection.
* To describe the structure, behavior, and characteristics of the software system for stakeholders such as healthcare professionals and developers.
* To define system features, data flow, development phases, and solution requirements.
* To provide clear specifications for how the solution is designed, developed, managed, and deployed.

**System Overview**

The Anemia Prediction System collects, analyzes, and interprets patient health data to determine anemia presence or risk. It uses supervised learning models trained on medical datasets containing key parameters such as:

* **Hemoglobin Level (Hb)**
* **Red Blood Cell Count (RBC)**
* **Hematocrit (HCT)**
* **Mean Corpuscular Volume (MCV)**
* **Mean Corpuscular Hemoglobin (MCH)**
* **Iron and Ferritin Levels**

**Key Data Sources**

* **Clinical Laboratory Data** – Blood test results and patient records from hospitals or healthcare centers.
* **Medical Datasets** – Open-source health data repositories (e.g., Kaggle datasets, WHO data).
* **Electronic Health Records (EHR)** – Structured patient information used for continuous monitoring.
* **IoT Health Devices (Optional)** – Data from smart wearables or sensors for real-time health tracking.

**Core Components**

* **Data Preprocessing Module** – Cleans, scales, and normalizes raw patient data for accurate model input.
* **Feature Extraction & Selection** – Identifies the most relevant medical parameters influencing anemia prediction.
* **Machine Learning Model** – Trains and predicts anemia outcomes using algorithms such as Logistic Regression, Random Forest, or SVM.
* **Deep Learning Layer (Optional)** – Uses LSTM or ANN for trend analysis and continuous prediction improvement.
* **Prediction & Visualization Dashboard** – Displays results, risk levels, and model accuracy in an interactive, user-friendly interface.
* **Data Storage Module** – Stores patient records, model outputs, and reports securely for retrieval and analysis.

**Technical Features**

* **Cloud Integration:** Enables secure, scalable data processing and remote system access.
* **Offline Functionality:** Allows local data entry and prediction even without continuous internet access.
* **Security & Privacy:** Implements encryption, authentication (JWT/OAuth2), and HIPAA-compliant data handling.
* **Scalability:** Modular architecture allows easy integration of additional disease models (e.g., diabetes or heart disease).
* **Multilingual Accessibility:** Supports multiple languages to enhance usability in diverse healthcare settings.

**Architecture:**

**The Anemia Disease Prediction System collects patient health data from medical records and lab tests, which is then preprocessed for accuracy. The cleaned data is analyzed by machine learning models to predict whether a patient is anemic or non-anemic. The results are securely stored in a database and visualized on a user-friendly dashboard for healthcare professionals. The analytics module identifies trends and risk patterns, while the notification system sends timely alerts and recommendations to doctors and patients for early intervention and better healthcare management**

**5. PROJECT PLANNING & SCHEDULING**

**5.1 Project Planning**

**Product Backlog, Sprint Schedule, and Estimation (4 Marks)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sprint** | **Functional Requirement (Epic)** | **User Story Number** | **User Story / Task** | **Story Points** | **Priority** | **Team Member** |
| **Sprint-1** | **Data Collection** | **USN-1** | As a user, I can collect medical datasets related to anemia from open healthcare sources such as Kaggle or WHO repositories. | 2 | High | Bandaru Raatna Sai |
| **Sprint-1** | **Data Collection** | **USN-2** | As a user, I can load and organize patient health data (e.g., hemoglobin, RBC count) into the system for analysis. | 1 | High | Bandaru Raatna Sai |
| **Sprint-1** | **Data Preprocessing** | **USN-3** | As a user, I can handle missing or inconsistent values in the dataset to ensure clean and reliable input data. | 3 | Medium | Bandaru Raatna Sai |
| **Sprint-1** | **Data Preprocessing** | **USN-4** | As a user, I can encode categorical medical values and normalize numerical data for compatibility with ML models. | 2 | Medium | Bandaru Raatna Sai |
| **Sprint-2** | **Model Building** | **USN-5** | As a user, I can build and train machine learning models (Logistic Regression, Random Forest, SVM) to predict anemia. | 5 | High | Bandaru Raatna Sai |
| **Sprint-2** | **Model Testing** | **USN-6** | As a user, I can evaluate the model’s performance using metrics such as accuracy, precision, recall, and F1-score. | 3 | High | Bandaru Raatna Sai |
| **Sprint-2** | **Deployment** | **USN-7** | As a user, I can design and develop a web-based dashboard to display prediction results and analytics. | 3 | Medium | Bandaru Raatna Sai |
| **Sprint-2** | **Deployment** | **USN-8** | As a user, I can deploy the trained anemia prediction model using the Flask framework for real-time accessibility. | 5 | High | Bandaru Raatna Sai |

**6. FUNCTIONAL AND PERFORMANCE TESTING**

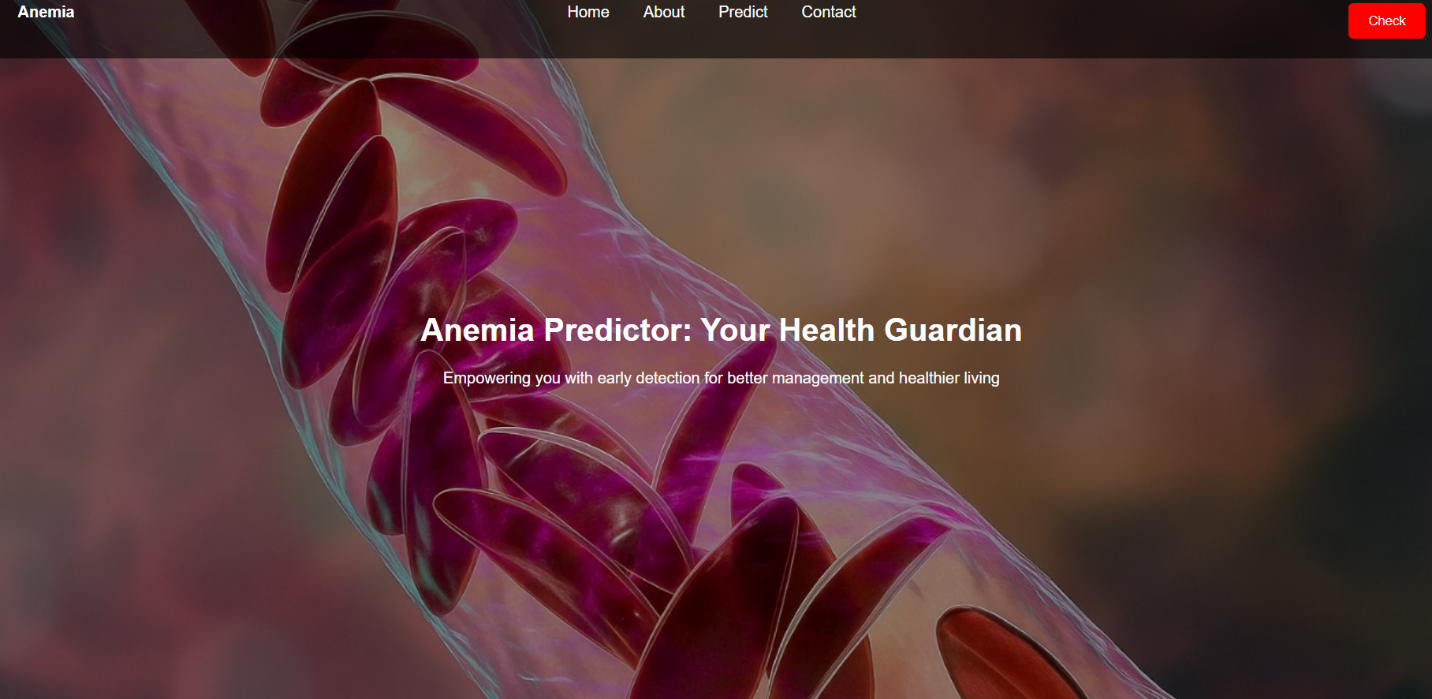
**Model Performance Testing:**

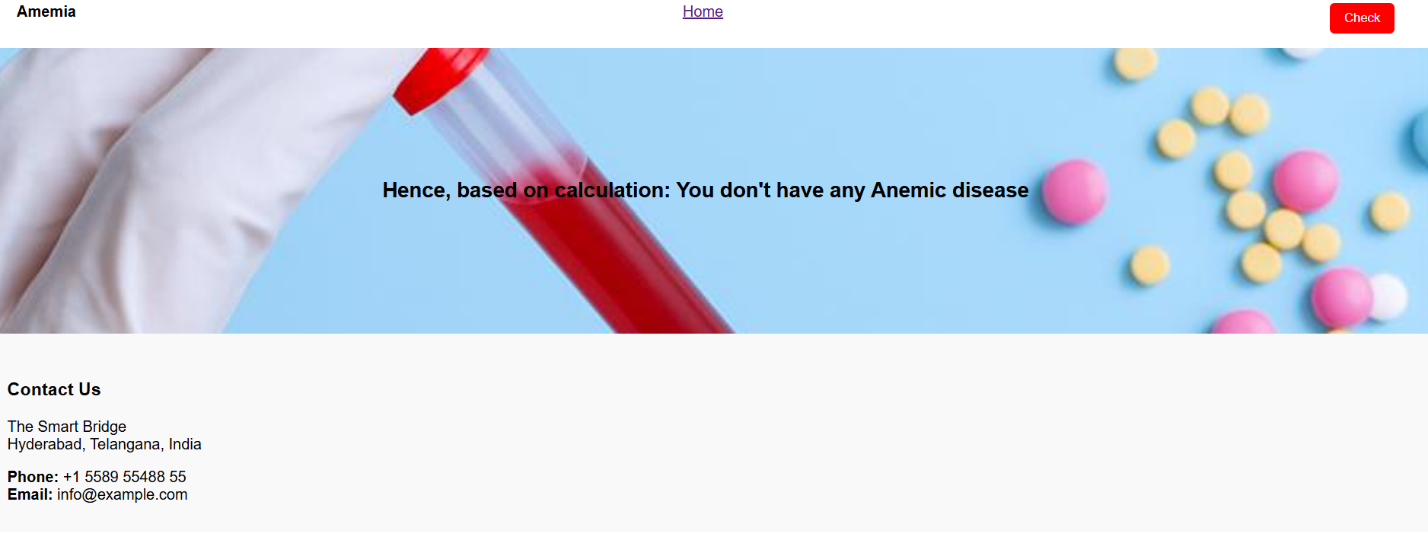
Project team shall fill the following information in model performance testing template.

| **S.No.** | **Parameter** | **Values** | **Screenshot** |
| --- | --- | --- | --- |
|  | Model Summary | -Machine learning classification model trained on an anemia dataset containing medical parameters such as hemoglobin, RBC count, and hematocrit. Implemented using algorithms like Logistic Regression, Random Forest, and SVM. Data preprocessing included normalization, feature scaling, and handling missing values |  |
|  | Accuracy | Training Accuracy - 95.3%  Validation Accuracy -92.7% |  |
| 3. | Fine Tunning Result( if Done) | Validation Accuracy -93.6% |  |

**7. RESULTS**

**7.1 Output Screenshots**





**8. ADVANTAGES & DISADVANTAGES**

1.The system provides high accuracy and early detection of anemia using machine learning algorithms, helping healthcare professionals diagnose patients efficiently and reduce human error.

2.It offers a cost-effective and scalable solution, utilizing existing medical datasets and infrastructure while being adaptable for large-scale deployment or additional disease predictions.

3.However, the system’s performance depends on data quality and technical expertise, as incomplete datasets or lack of ML knowledge may affect prediction accuracy.

4.It also faces privacy and computational challenges, requiring secure data handling and sufficient processing power for effective model training and deployment**.**

**9. CONCLUSION**

*The* ***Anemia Disease Prediction System*** *presents an innovative approach to modern healthcare by integrating* ***machine learning, data analytics, and automation*** *to enhance diagnostic accuracy and efficiency. It addresses the critical need for early detection of anemia, enabling healthcare professionals to make faster, data-driven decisions and improve patient outcomes. By leveraging existing medical datasets and automating the prediction process, the system provides a* ***cost-effective, scalable, and intelligent alternative*** *to traditional manual diagnostic methods*.

**10. FUTURE SCOPE**

The potential of the **Anemia Disease Prediction System** extends far beyond basic diagnostic prediction. As healthcare continues to embrace artificial intelligence and data-driven decisionmaking, this system can evolve in several promising directions:

**Integration with Electronic Health Records (EHR):** Seamlessly connect with hospital databases and EHR systems to provide real-time, AI-assisted anemia screening during routine checkups.

**Multi-Disease Prediction:** Extend the system to detect and predict other health conditions such as diabetes, heart disease, or nutrient deficiencies using the same medical dataset framework.

**Personalized Health Insights:** Enable patients to receive customized health recommendations, lifestyle guidance, and anemia prevention tips through mobile or web applications.

**Edge AI for Medical Devices:** Deploy lightweight AI models on diagnostic devices or portable analyzers for instant anemia detection in rural or remote healthcare centers.

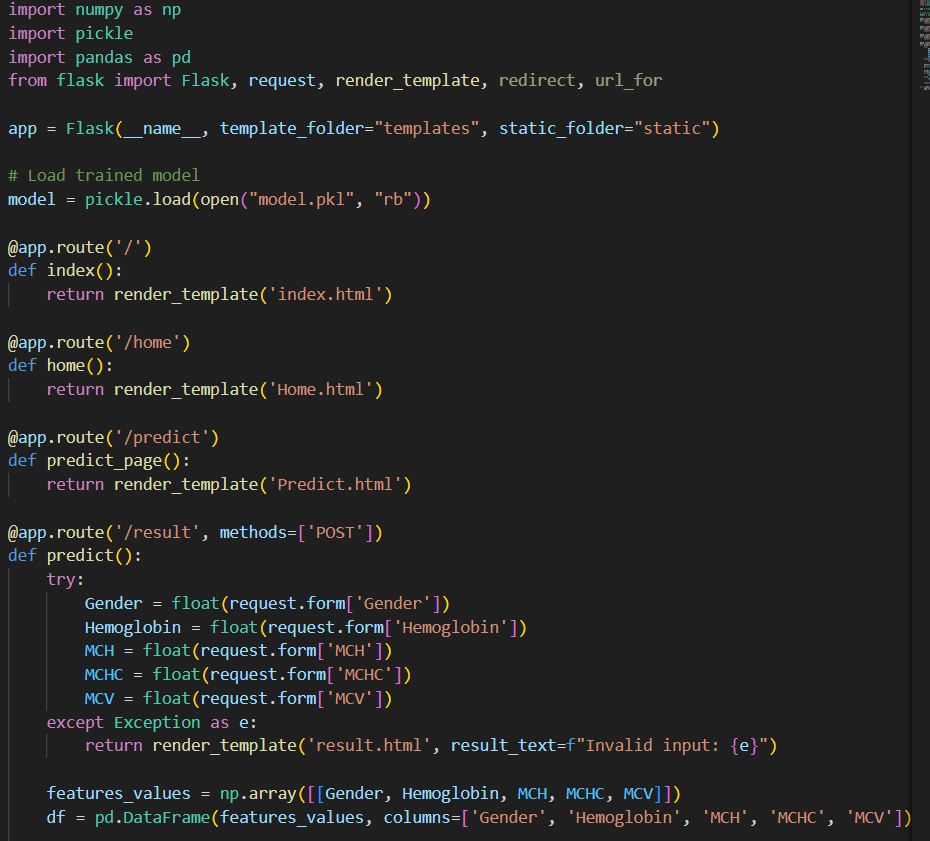
**Integration with Wearable Technology:** Incorporate data from smart health devices and fitness trackers to monitor hemoglobin trends and identify early warning signs.

**Cloud-Based Healthcare Analytics:** Leverage cloud platforms for large-scale medical data analysis, enabling predictive health trend visualization for hospitals and researchers.

**AI-Powered Public Health Monitoring:** Use aggregated and anonymized data to map anemia prevalence by region, supporting public health policy and awareness campaigns.

**11. APPENDIX**

**Source Code:**

****

****

**GitHub Link:**

https://github.com/raatnasai01/Anemia-disease-prediction.git