Data Collection and Preprocessing Phase

Date	03 August 2025
Project Title	Anemia Sense – Machine Learning for Precise Anemia Recognition
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Report

Elevate your health data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed, clinically relevant outcomes in every analysis and patient management decision.

Data Collection Plan

Section	Description
Project Overview	The Anemia Sense project leverages machine learning to recognize and manage anemia using clinical and demographic patient data. Utilizing datasets containing features such as age, gender, hemoglobin, RBC count, and additional lab parameters, the objective is to build a predictive model that can accurately detect anemia, support early diagnosis, and enable remote monitoring for effective healthcare delivery.
Data Collection Plan	Search, review, and download publicly available hematology datasets focused on anemia detection. Prioritize datasets with comprehensive clinical parameters and representative demographic data. Ensure datasets are relevant for medical ML applications, including CSV files from reputable repositories (e.g., Kaggle). Conduct initial profiling to confirm suitability for predictive analytics and patient stratification.

Raw Data Sources Identified

The raw data sources for this project include clinical anemia datasets, primarily obtained from Kaggle, a leading platform for healthcare and data science competitions and repositories. The sample data includes diverse patient histories and laboratory results essential for advanced ML-based anemia detection.

Raw Data Sources Report

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Anemia Dataset	The dataset comprises patient records including demographics (age, gender), hemoglobin and RBC count, various red cell indices (MCV, MCH), and verified anemia status labels. Suitable for predictive modeling and clinical analytics.	https://www.kaggle.com/datasets/biswaranjanrao/anemia-dataset	CSV	50–200 kB (varies)	Public