**1. System Design Document (SDD)**

**Project Name:** SehatAI – AI-powered Healthcare Support System  
**Version:** 1.0  
**Date:** September 2025

**1. Introduction**

This document explains the **system architecture, components, and workflows** for SehatAI. It covers the AI models, backend services, frontend application, and the data pipeline used to process health-related information.

**2. High-Level Architecture Diagram**

+-------------------+

| Doctor/Staff |

| (End User) |

+---------+---------+

|

v

+--------------------+

| Frontend Dashboard|

| (Streamlit / Flask)|

+---------+----------+

|

------------------------------------------------

| | |

v v v

+-------------------+ +-------------------+ +-------------------+

| X-ray Analyzer | | Prescription OCR | | Risk Scoring Model|

| (CNN, PyTorch) | | (Tesseract OCR) | | (Scikit-learn ML) |

+---------+---------+ +---------+---------+ +---------+---------+

| | |

v v v

Diagnosis JSON Digital Text JSON Risk Score JSON

| | |

-------------------------------------------------

|

v

+--------------------+

| Results Database |

| (SQLite/Postgres) |

+--------------------+

|

v

+--------------------+

| Reports Dashboard |

| Export (PDF/CSV) |

+--------------------+

**3. System Components**

1. **Frontend**
   * Streamlit web dashboard.
   * Upload interface for X-rays/prescriptions.
   * Patient form input for demographics.
   * Report display & download.
2. **Backend**
   * REST API (Flask).
   * Routes: /xray, /ocr, /risk.
   * Handles inference requests from frontend.
3. **AI Models**
   * CNN (ResNet/DenseNet) for chest X-ray classification.
   * Tesseract OCR (Urdu + English) for prescriptions.
   * Logistic Regression/Random Forest for risk scoring.
4. **Data Pipeline**
   * Input data preprocessing (resize X-rays, binarize text, normalize tabular data).
   * Model inference.
   * Output formatting.
5. **Database**
   * SQLite/PostgreSQL for Phase-1.
   * Stores patient data, results, and digitized prescriptions.

**4. Workflow**

1. User uploads data (X-ray image / prescription / patient info).
2. Data is preprocessed.
3. AI models generate predictions.
4. Results stored in database.
5. Dashboard displays results in real time.
6. Reports exported if required.

**📄 2. Data Documentation**

**Project Name:** SehatAI – AI-powered Healthcare Support System  
**Version:** 1.0  
**Date:** September 2025

**1. Data Sources**

1. **Chest X-rays**
   * **NIH Chest X-ray Dataset** (112,000 images).
   * **Kaggle Tuberculosis Dataset** (~700 images).
2. **Prescriptions (OCR)**
   * **UCOM Urdu Handwritten Dataset**.
   * Synthetic handwritten Urdu/English prescription samples.
3. **Risk Scoring**
   * **Pakistan Demographic and Health Survey (PDHS)**.
   * **WHO Global Health Survey** datasets.

**2. Data Schema**

**X-rays:**

* image\_id (string)
* file\_path (string)
* label (TB / Pneumonia / Normal)

**Prescriptions:**

* prescription\_id (string)
* file\_path (string)
* text\_extracted (string)

**Risk Scoring:**

* patient\_id (string)
* age (int)
* gender (string)
* weight (float)
* blood\_pressure (float)
* diabetes\_risk (binary: 0/1)

**3. Preprocessing Steps**

* **X-rays:**
  + Resize to 224x224.
  + Normalize pixel values.
  + Augmentation (flip, rotate).
* **Prescriptions:**
  + Convert to grayscale.
  + Binarize (thresholding).
  + Noise removal.
* **Risk Data:**
  + Remove missing values.
  + Encode categorical features.
  + Normalize numerical fields.

**4. Data Cleaning Rules**

* Drop images with missing or corrupted labels.
* Remove incomplete patient records.
* Standardize all prescriptions in UTF-8 format.

**5. Data Versioning Plan**

* **Tool:** DVC (Data Version Control).
* **Storage:** GitHub with Git-LFS for large files.
* Each dataset release tagged (v1.0, v1.1).
* Training/validation/test splits logged.

**📄 3. Model Card (AI/ML Models)**

**Project Name:** SehatAI – AI-powered Healthcare Support System  
**Version:** 1.0  
**Date:** September 2025

**1. Model Description**

1. **X-ray Analyzer**
   * Type: Convolutional Neural Network (CNN).
   * Architecture: ResNet-50 (PyTorch).
   * Task: Classify X-rays into TB, Pneumonia, Normal.
2. **Prescription OCR**
   * Type: OCR Engine.
   * Library: Tesseract OCR with Urdu + English language pack.
   * Task: Extract text from handwritten prescriptions.
3. **Risk Scoring Model**
   * Type: Logistic Regression / Random Forest (Scikit-learn).
   * Task: Predict diabetes/heart disease risk.

**2. Model Architecture & Hyperparameters**

**X-ray Analyzer (ResNet-50):**

* Learning Rate: 0.001
* Batch Size: 32
* Epochs: 10–15
* Loss Function: CrossEntropyLoss
* Optimizer: Adam

**Prescription OCR (Tesseract):**

* Pretrained weights (LSTM-based OCR).
* Language Pack: Urdu + English.

**Risk Scoring (Logistic Regression):**

* Solver: liblinear
* Penalty: L2 Regularization
* Features: Age, Gender, BP, BMI, Lifestyle factors.

**3. Training Data**

* **X-rays:** NIH Chest X-ray Dataset, Kaggle TB.
* **OCR:** UCOM Urdu Handwriting Dataset + synthetic data.
* **Risk Scoring:** PDHS & WHO survey data.

**4. Metrics**

* **X-ray Analyzer:** Accuracy ≥ 85%, F1-score.
* **OCR:** Accuracy ≥ 80%, Word Error Rate (WER).
* **Risk Scoring:** AUC ≥ 0.8, Precision-Recall.

**5. Limitations**

* Limited to TB & Pneumonia (Phase-1).
* OCR may fail on very messy handwriting.
* Risk model depends on available health survey data.

**6. Ethical Considerations**

* Models should **support doctors**, not replace them.
* Misdiagnosis risk must be mitigated with human validation.
* Data must remain anonymized to protect patient privacy.
* Avoid bias in training (balanced male/female and regional samples).

✅ These three documents (SDD, Data Documentation, Model Card) ensure your project is **transparent, structured, and ready for Techathon Phase-1 submission**.