Executive Summary: SehatAI

Introduction

In rural Pakistan, limited access to diagnostic tools, digitized prescriptions, and predictive health monitoring intensifies healthcare disparities, with only 0.7 doctors per 1,000 people and 42% basic health coverage. SehatAI, developed for the URAAN AI Pakistan Techathon 1.0, is an AI-powered web application built with Python and Streamlit to deliver early detection of diabetes and pneumonia, alongside digitized prescription processing and medicine validation. By using machine learning, SehatAI enables health workers to screen chronic and infectious diseases and maintain digital records, enhancing care continuity in underserved rural areas of Pakistan.

Objective

SehatAI aims to bridge rural healthcare gaps by:

- Developing a convolutional neural network (CNN) for pneumonia detection from chest X-ray images.
- Creating a risk-scoring model for early diabetes detection using numerical inputs.
- Digitizing prescriptions with bilingual (Urdu/English) explanations to improve accessibility for low-literacy populations (56% rural literacy rate).

Purpose of Report

This report evaluates SehatAI's development, performance, and potential impact for rural Pakistan's healthcare system. It details the system's architecture, implementation across three phases (Planning/Design, Development/Integration, Testing/Deployment), and thir respective results. It identifies challenges and proposes future enhancements to support scalability and integration with national health systems like NADRA e-health.

Project Scope

SehatAI processes text-based inputs (e.g., glucose, BMI) for diabetes prediction, X-ray images for pneumonia detection, and prescription images for digitization, delivering advisory outputs for rural clinics. The system supports:

- Diabetes Prediction: Binary risk scoring (diabetes/non-diabetes).
- Pneumonia Detection: Binary classification (Normal/Pneumonia), with potential for multiclass (Normal/Bacterial/Viral).
- Prescription Digitization: OCR extraction and bilingual medicine information via Mistral AI API.

Methodology

SehatAI employs a client-server architecture with a Streamlit frontend and Python backend, organized into six modules:

1. Data Loading: Ingests tabular data (`diabetes.csv`) and X-ray images (`dataset.csv`).
2. Preprocessing: Scales diabetes features (StandardScaler) and resizes/augments X-rays (150x150 grayscale, ImageDataGenerator).

3. Diabetes Prediction: XGBoost Classifier (74.7% accuracy, AUC 80.8%) on Pima dataset (768 samples).
4. Pneumonia Prediction: Custom CNN (88% accuracy, F1: 0.87 macro) on Kaggle dataset (5,856 images).
5. Text-Generation (Prescription): Mistral AI API for OCR (`mistral-ocr-latest`) and bilingual medicine information (`mistral-small-latest`).

Implementation Phases

- Phase 1 (Planning/Design): Designed data pipelines and preprocessing (e.g., imputations, image resizing).
- Phase 2 (Development/Integration): Built XGBoost, CNN, and OCR modules, integrated via Streamlit.
- Phase 3 (Testing/Deployment): Evaluated models (74.7% diabetes, 88% pneumonia) and ensured model persistence.

Technologies

- **Languages and Framework**: Python 3.10, Streamlit.
- **Libraries**: pandas, NumPy, Scikit-learn, TensorFlow/Keras, XGBoost, OpenCV, Mistral AI.
- **API**: Mistral AI

Future Direction

To enhance SehatAI's impact:

- Model Optimization: Tune XGBoost and CNN (e.g., GridSearch, ResNet) to reach 80-99% diabetes and 90-96% pneumonia accuracy.
- Multiclass Pneumonia: Extend CNN to classify Normal/Bacterial/Viral Pneumonia.
- Prescription Enhancements: Improve OCR for low-quality images and add prescription validation (e.g., dosage checks).
- Mobile Support: Develop a mobile app for offline predictions, leveraging Pakistan's 70% mobile penetration.
- NADRA Integration: Link to "One Patient One ID" for secure record-keeping.
- Validation: Collaborate with WHO or local authorities for clinical validation using Pakistan-specific datasets.

Conclusion

SehatAI offers a scalable solution for rural Pakistan's healthcare challenges, achieving 74.7% diabetes accuracy, 88% pneumonia accuracy, and reliable prescription digitization with bilingual outputs. Despite challenges like class imbalance and data biases, the system's modular design and Streamlit interface enable effective screening and documentation, supporting health workers where only 50% of facilities have imaging equipment. Future enhancements will strengthen its role in reducing healthcare disparities, aligning with Pakistan's e-health initiatives.