Armauer Hansen Research Institute
Department of Malaria and Neglected Tropical Disease (MNTD)

Project Title: AI-Based Data Anomaly Detection System.

Submitted to DSWB-AHRI Hackathon-2025 Team: ***requirement of to develop accurate, reusable and scalable solutions that can serve to solve real time problems the institute is engaged in.***

Submitted By: Mihret Ebabu

Armauer Hansen Research Institute

Addis Ababa, Ethiopia

May-13-2025

Contents

Introduction

Ensuring the integrity of health data is critical to accurate clinical decision-making and effective public health interventions, especially in low- and middle-income countries (LMICs) where data systems are often fragmented and manually managed.
This project addresses the pressing need for automated, scalable, and interpretable anomaly detection by developing an AI-driven framework capable of identifying missing, inconsistent, and contextually implausible data in real-world healthcare datasets. By leveraging a hybrid approach—combining deep learning models, time-series analysis, and rule-based logic—the proposed system aims to enhance data reliability, reduce error rates, and support the trustworthiness of digital health ecosystems under constrained computing environments.

Background

In many low- and middle-income countries (LMICs), health data is often collected manually, stored in fragmented systems, and prone to errors. Data anomalies—such as missing entries, outliers, and inconsistencies—pose serious risks to healthcare decision-making, clinical outcomes, and research accuracy. Common examples include illogical timestamps, medically implausible values (e.g., adult weight 5kg), and device-generated artifacts. Identifying these anomalies accurately and efficiently is essential to improve data quality, reliability, and trust in digital health ecosystems.

Objectives

The primary objective of this project is to design a **robust, AI-driven anomaly detection** system capable of automatically identifying missing data, point anomalies, and contextual or collective inconsistencies in real-world health datasets.
The system will be interpretable, scalable, and deployable in environments with limited computing infrastructure, and demonstrate high precision and recall for clinically relevant anomalies.

Proposed Solution

We propose to develop a **hybrid anomaly detection framework** using unsupervised deep learning (**Autoencoders**) for point anomalies and Long and Short-Term Memory (LSTM)-based time-series models to capture contextual and temporal inconsistencies.
We will also implement **rule-based** checks for obvious logical violations. The system will be designed to **generate interpretable alerts** and can be generalized across datasets with **minimal reconfiguration**.

Detailed Methodology

Data Preparation Pipeline

- Analyze schema and completeness of the dataset.
- Handle missing values using imputation or marking for detection.
- Normalize date/time formats and categorical values.
- Generate statistical summaries for outlier detection.

Proposed Models

- **Autoencoder (AE):** For unsupervised detection of point anomalies based on reconstruction error.
- **LSTM-based sequence model**: for detecting temporal anomalies in time-series data such as patient vitals.
- **Rule-Based Logic**: E.g., Discharge date before admission, male with pregnancy, zero pulse for extended duration.

Programming Language & Tools

- Python 3.9+: Libraries: `Pandas`, `NumPy`, `Scikit-learn`, `TensorFlow/Keras`, `Matplotlib`, `Seaborn`, `PyOD`.
- Development in JupyterLab environment.

Evaluation and Testing

- Use synthetic datasets and real health data samples with embedded anomalies.
- Evaluation Metrics: Precision, Recall, F1-score, Confusion Matrix, ROC-AUC (where applicable)
- Cross-validation withheld-out anomaly-labeled datasets

Data Quality Dashboards

- Visualize anomalies by site, time, or user.
- Track trends in anomalies to identify training or system issues.

Audit Trail & Logs

- Track who entered or edited which data and when.
- Helps trace the source of anomalies and prevent recurring errors.

Expected Outcomes

This project is expected to produce a **modular**, **interpretable** anomaly detection system capable of handling real-world healthcare datasets. Deliverables will include a Python-based toolset for **preprocessing**, **anomaly detection**, **visualization**, and **reporting**, along with a documented Jupyter Notebook and demo-ready output for evaluation.

Selected Milestone Functionality

The key milestone selected for submission is the implementation of an Autoencoder-based anomaly detection module.
Includes:

Building the AE model architecture

- **Designing the structure of the Autoencoder** (encoder-decoder layers, activation functions, etc.).

Data Handling & Preparation Functions

- **Training on normalized clinical data**
  Feeding preprocessed, scaled data into the Autoencoder model to learn normal patterns.

Anomaly Scoring & Thresholding Functions

- **Generating reconstruction loss threshold**
  Calculating reconstruction errors and determining the cutoff point to classify anomalies.

Anomaly Detection Functions

- **Identifying and flagging high-loss anomalies**
  Detecting data points with errors above the threshold and marking them as potential anomalies.

Interpretation & Visualization Functions

- **Visualizing results with interpretability support**
  Creating interpretable plots (e.g., loss distributions, anomaly scores, feature contributions) to explain why data was flagged.

Project Summary: AI-Based Data Anomaly Detection System

Submitted by: Mihret Ebabu
Institution: Armauer Hansen Research Institute (AHRI), Department of Malaria and Neglected Tropical Diseases
Event: DSWB-AHRI Hackathon 2025

This project proposes the development of an **AI-driven anomaly detection system** to improve the integrity and reliability of healthcare data, especially in low- and middle-income countries where data is often fragmented and manually collected. The system combines **Autoencoders**, **LSTM-based time-series models**, and **rule-based logic** to detect various anomalies, including missing values, outliers, and contextual inconsistencies.

The framework will be:

- **Scalable**, for use across different datasets and sites
- **Interpretable**, offering clear anomaly explanations
- **Lightweight**, suitable for low-resource settings

Key features include:

- A data preparation pipeline (cleaning, normalization, imputation)
- Autoencoder-based detection for point anomalies
- LSTM for sequence-based anomaly detection
- Rule checks for logical violations (e.g., male with pregnancy)
- Visual dashboards to track anomaly trends
- Audit trails to trace data issues back to their source

Tools & Environment: Python 3.9+, TensorFlow/Keras, PyOD, Pandas, NumPy, Scikit-learn, JupyterLab
Evaluation Metrics: Precision, Recall, F1-score, ROC-AUC, confusion matrix

Milestone Delivered: An Autoencoder module that preprocesses health data, learns normal patterns, flags anomalies based on reconstruction loss, and visualizes the results for interpretability.