FTL\_eth\_prep\_eng\_model\_names

AI-Powered Multimodal Diagnostic Tool for Endemic Infectious Diseases in Liberia

Group Members:

- Snoyonoh T. Barcon

- Peter Mulbah

- Jerry

# Data Preparation / Feature Engineering

## 1. Overview

This project focuses on developing an AI-powered multimodal diagnostic tool for endemic infectious diseases (malaria, TB, HIV/AIDS, typhoid fever) in Liberia. Data preparation and feature engineering are critical for ensuring model accuracy, fairness, and adaptability in low-resource contexts.

## 2. Data Collection

- Sources:  
 - Open-source image datasets (NIH Chest X-rays, Malaria blood smears).  
 - Clinical records (symptom-based data, anonymized).  
 - Public health datasets (WHO, Liberia MoH).  
 - Environmental/climatic datasets (NASA, World Bank).  
  
- Preprocessing during collection:  
 - De-identification of patient data.  
 - Conversion of raw microscopy/X-ray images into standardized formats.  
 - Extraction of structured clinical variables (age, sex, symptoms, test results).

## 3. Data Cleaning

Steps:  
- Handle missing values with imputation (mean/median for numeric, mode for categorical).  
- Remove duplicate patient entries.  
- Outlier detection using z-scores and IQR.  
- Noise reduction in image datasets via augmentation pipelines (denoising filters, contrast adjustments).

## 4. Exploratory Data Analysis (EDA)

- Univariate analysis: frequency of diseases, patient demographics.  
- Bivariate analysis: correlations between symptoms and disease outcomes.  
- Imaging analysis: pixel intensity distributions, example image grids.  
- Visualizations: histograms, heatmaps of correlations, class balance plots.  
(Code and example plots will be included in the repo.)

## 5. Feature Engineering

- Image Features: extracted via CNN convolutional layers (latent features).  
- Structured Data Features: categorical encodings (one-hot, binary indicators).  
- Temporal/Environmental Features: lagged climate variables for malaria/typhoid trends.  
- Interaction Features: symptom clusters (fever + cough + chest pain for TB prediction).

## 6. Data Transformation

- Scaling: MinMaxScaler for continuous features.  
- Encoding: One-hot encoding for categorical features.  
- Normalization: Z-score normalization before ANN/XGBoost models.  
- Train-Test Split: 80/20 with stratification on disease type.

# Model Exploration

## 1. Model Selection

CNN for images (TB/malaria detection), Ensemble ML (Random Forest, XGBoost) for structured clinical data, ANN/XGBoost for epidemiological forecasting. The multimodal approach captures imaging, structured, and environmental data simultaneously.

## 2. Model Training

- CNN: trained with augmentation (rotation, flipping, zoom).  
- Ensemble ML: cross-validation (k=5).  
- Forecasting: time-series split for malaria/typhoid case trends.  
- Hyperparameters tuned using GridSearchCV and Bayesian optimization.

## 3. Model Evaluation

- Metrics: Accuracy, F1-score, Precision, Recall.  
- Visualizations: ROC curves, confusion matrices, precision-recall curves.  
- Forecast models: MAPE and RMSE.  
- Target: ≥90% validation accuracy for CNN, balanced recall across diseases.

## 4. Code Implementation

Code in Python using TensorFlow/Keras, PyTorch, Scikit-learn, XGBoost, Pandas, Seaborn.  
  
Example snippet:  
```python  
from sklearn.preprocessing import MinMaxScaler  
scaler = MinMaxScaler()  
X\_scaled = scaler.fit\_transform(X)  
  
from tensorflow.keras import models, layers  
cnn = models.Sequential([  
 layers.Conv2D(32, (3,3), activation='relu', input\_shape=(128,128,3)),  
 layers.MaxPooling2D(2,2),  
 layers.Flatten(),  
 layers.Dense(128, activation='relu'),  
 layers.Dense(1, activation='sigmoid')  
])  
cnn.compile(optimizer='adam', loss='binary\_crossentropy', metrics=['accuracy'])  
```

# Sample EDA and Model Visualizations

Figure 1. Histogram showing distribution of disease cases (sample).

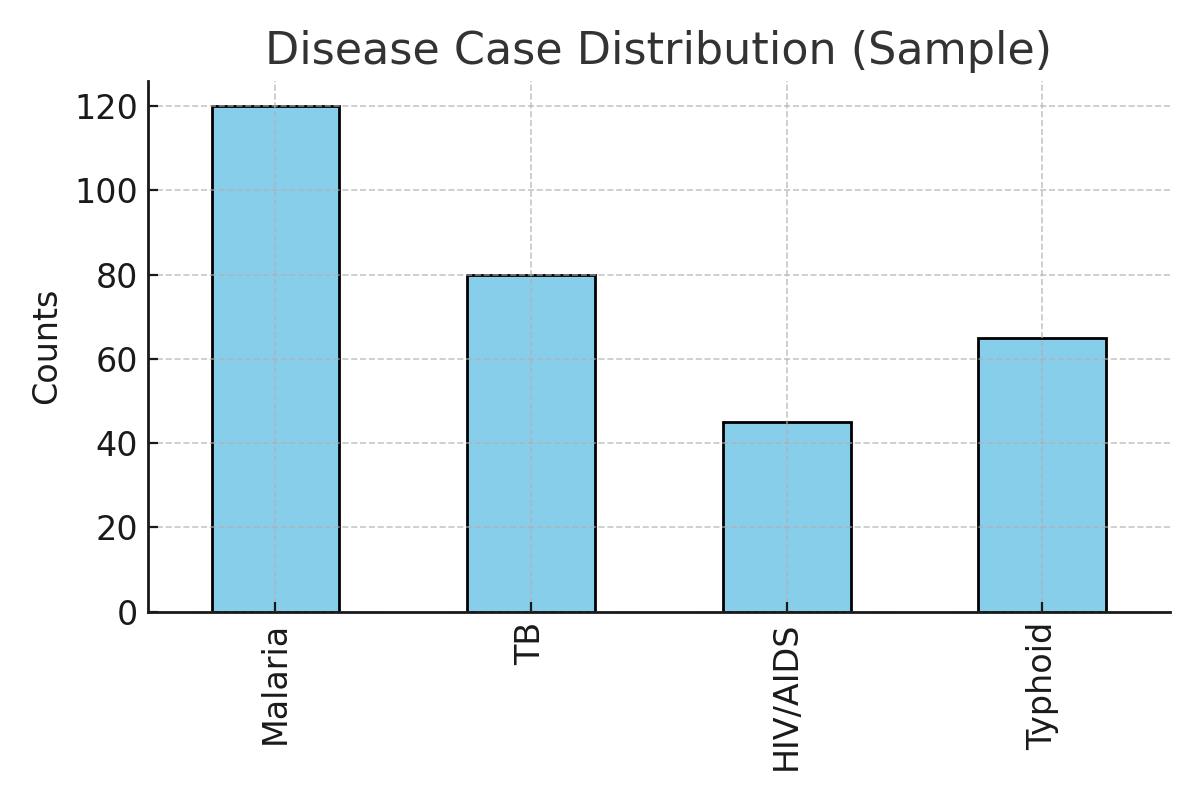


Figure 2. Heatmap of feature correlations in clinical data (sample).

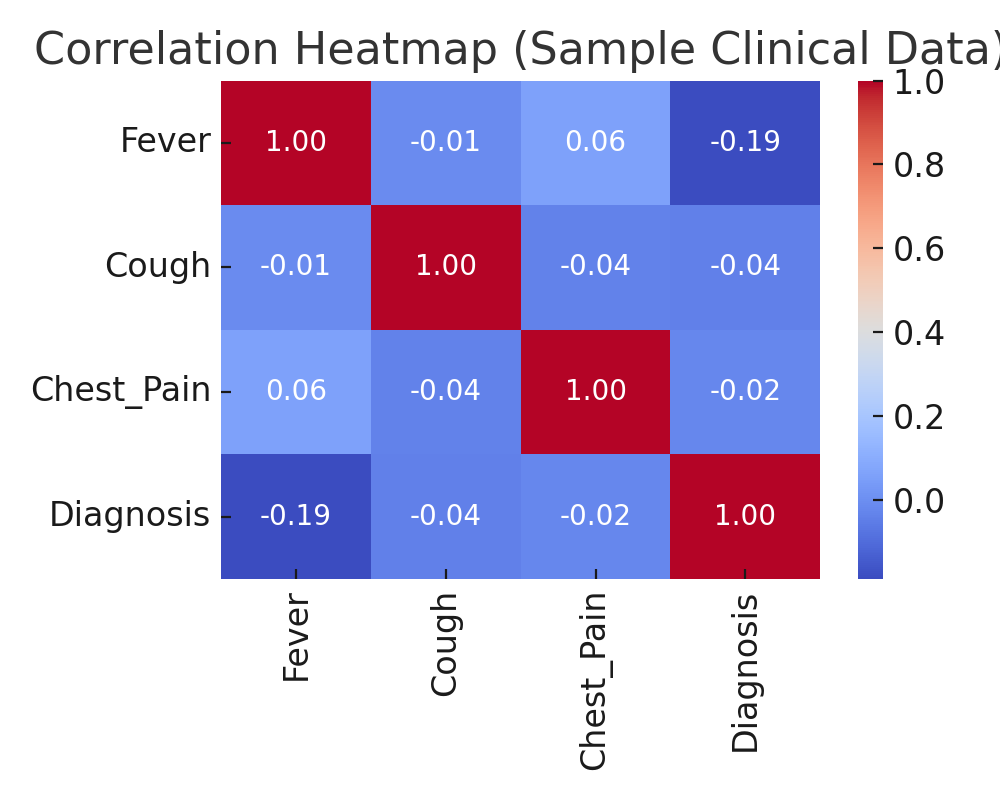


Figure 3. Confusion matrix from sample model evaluation.

