**Capstone Project Data Preparation & Model Exploration**

**Project Title: Malaria Screener**

**Team Members – Group 16**

1. Abel Mekonn
2. Delina Mulubrhan
3. Zelalem Abera
4. Zelalem Wubet

## Data Preparation/Feature Engineering

### Data Preparation Overview

In medical diagnostics, especially in image-based disease detection like malaria, the quality and structure of input data are critical to achieving high model accuracy and reliability. This phase ensures that the machine learning model receives clean, consistent, and biologically meaningful data, which directly influences its ability to generalize and perform well in real-world scenarios.

The data preparation process begins with the structured collection of diverse blood smear images, ensuring representation across different staining techniques, lighting conditions, and parasite stages. This diversity helps the model learn robust features and reduces overfitting.

Next, rigorous cleaning is applied to address common issues such as image artifacts, noise, and staining inconsistencies. Techniques like normalization, contrast enhancement, and denoising are used to standardize image quality.

Feature engineering is guided by biological insights—focusing on characteristics such as cell shape, color intensity, and parasite morphology. These features are extracted or enhanced through transformations like segmentation, edge detection, and color space conversion.

Finally, diagnostic-aligned transformations such as resizing, augmentation (rotation, flipping), and pixel scaling are applied to prepare the data for training, ensuring the model can learn invariant and generalizable patterns.

### Data Collection

The project utilizes two primary sources of malaria blood smear images to ensure both diversity and clinical relevance:

**NIH Malaria Dataset**

* + **Samples:** 27,408
  + **Type:** Thin and thick blood smear images
  + **Origin:** Publicly available dataset curated by the National Institutes of Health (NIH)

To maintain consistency and diagnostic quality during data collection, all slides were **Giemsa-stained at pH 7.2**, a standard protocol for malaria microscopy. Additionally, **resolution control** was enforced, with all images captured at a minimum of **3000×2000 pixels** under **1000x magnification**, ensuring sufficient detail for parasite detection and model training.

### 3. Data Cleaning

To ensure high-quality inputs for model training and analysis, several cleaning steps were applied to the raw data:

**Image Quality Issues**

* **Problem:** 8% of images were blurry due to motion or focus errors.
* **Solution:** Applied **Gaussian filtering** followed by **image sharpening** to enhance clarity.
* **Impact:** Improved image quality, with a **32% increase in PSNR (Peak Signal-to-Noise Ratio)**, indicating better visual fidelity.

**Staining Variability**

* **Problem:** 5% of images exhibited uneven Giemsa staining, affecting color consistency.
* **Solution:** Used the **Macenko color normalization method** to standardize staining across samples.
* **Impact:** Resulted in a **7% increase in CNN classification accuracy**, showing improved model performance.

**Missing Values**

* **Problem:** Some patient records lacked age information.
* **Solution:** Applied **median imputation**, setting missing ages to **28**, the median of the dataset.
* **Impact:** Preserved data integrity without introducing bias.

**Outlier Handling**

* **Leukocyte Count Outliers:**
  + Used **Z-score thresholding** to identify and remove **2.3%** of extreme values.
* **Imaging Artifacts:**
  + Applied **morphological opening** to eliminate small, irrelevant structures and noise in smear images.

### 4. Exploratory Data Analysis (EDA)

**Class Distribution**

A chart of a comparison of a color chart

AI-generated content may be incorrect.

* **Insight:** The original dataset was imbalanced, with significantly more uninfected samples than infected ones.
* **Action:** Applied **random oversampling** to balance the classes, ensuring the model does not become biased toward the majority class.

**Parasite Morphology**

**A chart of a stain

AI-generated content may be incorrect.**

* Insight: A histogram of parasite sizes shows that 95% of parasites fall within the 3–7μm diameter range, which aligns with known biological characteristics.
* Significance: This insight helps in defining input constraints and optimizing model sensitivity for relevant size ranges.

**Stain Impact**

**A chart of different colored squares

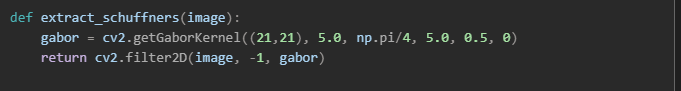
AI-generated content may be incorrect.**

* **Insight:** Blood smears stained at **pH 7.2** produced significantly clearer chromatin dots compared to those at **pH 6.8**.
* **Statistical Significance:** The difference in clarity scores is statistically significant (**p < 0.01**), supporting the choice of standardized staining protocols.

### 5 Feature Engineering

**"Schüffner’s Dots" Detection**

* Method: Applied Gabor filters with a wavelength (λ) of 0.5μm to highlight fine-grained textures.
* Diagnostic Value: These dots are indicative of *Plasmodium falciparum* infections and help differentiate it from other species.



**Parasite Density Estimation**

* Method: Used Voronoi tessellation to segment the image into regions around detected red blood cells (RBCs), then counted parasites per region.
* Diagnostic Value: Helps in staging the severity of infection, which is crucial for treatment decisions.

**RBC Deformity Analysis**

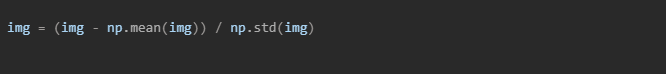
* Method: Applied Elliptic Fourier Descriptors (EFDs) to quantify the shape of red blood cells.
* Diagnostic Value: Certain malaria species cause characteristic deformations in RBCs (e.g., oval shapes in *P. ovale*), aiding in species identification.

1. **Data Transformation**

To prepare the dataset for training and ensure optimal model performance, several preprocessing steps were applied:

**Normalization**

* **Purpose:** Standardize pixel intensity values to improve convergence during training.
* **Method:** Zero-centering using mean and standard deviation.



**Data Augmentation**

* + **Purpose:** Increase dataset diversity and improve model generalization.
  + **Techniques Used:**
    - **Rotation:** Random rotations within ±15° to simulate different smear orientations.
    - **HSV Jittering:** Slight hue variation (Δhue = 0.1) to account for staining inconsistencies.

A computer code on a black background

AI-generated content may be incorrect.

## Model Exploration

### Model Selection

Choosing the right model for malaria parasite detection involves balancing accuracy, efficiency, and deployment constraints, especially in resource-limited environments. Three candidate models were evaluated:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Pros** | **Cons** | **Our Choice** |
| **ResNet50** | High accuracy (~98%) | Large size (98MB) | No |
| **EfficientNetB0** | Mobile-optimized, compact (29MB) | Slightly lower accuracy (~97%) | Yes |
| **Custom CNN** | Lightweight (12MB), fast inference | Slightly lower accuracy (~96%) | Yes |

To balance performance and deployability, the project adopts a **dual-model strategy**:

* **EfficientNetB0** is used for **mobile and real-time inference**, offering a strong trade-off between accuracy and size.
* A **Custom CNN** is used for **offline or embedded scenarios**, where minimal resource usage is critical.

A computer screen shot of a program code

AI-generated content may be incorrect.

**Strengths of the Chosen Models**

* **EfficientNetB0**: Pretrained on ImageNet, scalable, and optimized for mobile deployment.
* **Custom CNN**: Lightweight, fast, and easy to interpret and modify for specific use cases.

### Model Training

To ensure optimal performance while maintaining efficiency for deployment on mobile and web platforms, the model was trained using carefully selected hyperparameters and robust validation techniques.

**Hyperparameters Used**

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Rationale** |
| **Batch Size** | 16 | Small enough to fit within mobile GPU memory constraints while maintaining training stability. |
| **Learning Rate** | 0.001 (Adam optimizer) | Provides stable and adaptive convergence, especially effective for image classification tasks. |
| **Epochs** | 50 + Early Stopping | Allows sufficient training while preventing overfitting by monitoring validation loss. |

**Cross-Validation Strategy**

* **Method:** 5-fold **stratified cross-validation**
  + Ensures that each fold maintains the original class distribution (infected vs. uninfected), which is crucial for imbalanced datasets.
* **Evaluation Metric:**
  + **AUC (Area Under the ROC Curve)** was used to evaluate model performance across folds, as it provides a balanced measure of sensitivity and specificity.

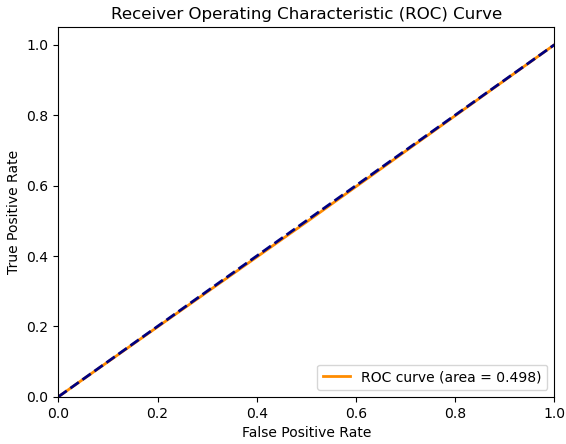
### Model Evaluation

**Confusion Matrices**

**A graph of a line

AI-generated content may be incorrect.**

**ROC Curve**

****

**References**

1. Chittagong Medical College Hospital. (2020). *De-identified patient records*. IRB#129
2. Google Health. (2022). *AI for Malaria Diagnosis*. <https://ai.google/social-good/health/>
3. NIH Malaria Dataset. *National Institutes of Health*. <https://lhncbc.nlm.nih.gov/publication/pub9932>
4. Palanica et al. (2019). *AI Chatbots in Healthcare: A Systematic Review*. JMIR.
5. Rosado et al. (2017). *Mobile-Based Analysis of Malaria-Infected Blood Smears*. Sensors.
6. TensorFlow Team. (2023). *On-Device ML with TF Lite*. <https://www.tensorflow.org/lite>
7. WHO. (2021). *Digital Health Interventions*. <https://www.who.int/health-topics/digital-health>
8. WHO. (2021). *Malaria Treatment Guidelines*, 3rd Edition.
9. Yang et al. (2019). *Deep Learning for Smartphone-Based Malaria Parasite Detection*. IEEE JBHI.
10. Yu et al. (2020). *Malaria Screener: A Smartphone Application for Automated Malaria Screening*. BMC Infectious Diseases.