Emiliano Mixali Tofas Rodríguez – 175413

Bridging Accuracy and Trust: A Comparative Analysis of Explainable Deep Learning vs. Classical Computer Vision in Low-Resource Malaria Diagnostics

Artificial Intelligence - DSI5022

December 8, 2025

Autumn 2025

1. Problem Definition

1.1 Context and Motivation

Malaria remains one of the most significant global health burdens, particularly in low-resource settings. The World Health Organization (WHO) estimates hundreds of millions of cases annually, with the gold standard for diagnosis being the manual examination of thin blood smears via light microscopy. While effective, this process is highly dependent on the expertise of the microscopist, time-consuming, and prone to fatigue-induced errors, leading to misdiagnosis rates that can reach 50% in rural field conditions.

1.2 The Research Problem

The core problem this project addresses is not merely the automation of malaria detection, but the "Black Box" paradox in medical Artificial Intelligence (AI). While Deep Learning (DL) models—specifically Convolutional Neural Networks (CNNs)—have demonstrated state-of-the-art accuracy in medical imaging, they often lack interpretability. In a clinical setting, a high-accuracy prediction without a visual justification is often untrustworthy and ethically risky.

Furthermore, deep learning models are computationally expensive, raising questions about their deployability on edge devices (such as smartphones connected to microscopes) compared to "lighter," classical machine learning approaches.

1.3 Task and Dataset

This project seeks to develop an end-to-end AI system to perform binary classification of single-cell thin blood smear images into two categories: Parasitized and Uninfected.

The study will utilize the NIH Malaria Cell Images Dataset (Rajaraman et al., 2018), which contains 27,558 segmented cell images. This dataset offers a balanced class distribution, making it an ideal candidate for rigorous methodological comparison without the confounding variables of class imbalance.

2. Theoretical Justification and Relevance

2.1 Theoretical Framework: Representation Learning vs. Feature Engineering

This project is grounded in the theoretical debate between Symbolic/Classical AI (Feature Engineering) and Connectionist/Modern AI (Representation Learning).

- Classical Approach (Baseline): Relies on the hypothesis that domain-specific features (texture, color histograms, gradients) are sufficient to define the decision boundary. We will employ Histogram of Oriented Gradients (HOG) descriptors fed into a Random Forest classifier. This approach offers high interpretability and low computational cost but suffers from the "curse of dimensionality" and limits of human-designed features.

- Modern Approach (Transfer Learning): Relies on the Manifold Hypothesis, where high-dimensional data (images) lie on a lower-dimensional manifold. We will utilize Transfer Learning with a MobileNetV2 architecture pre-trained on ImageNet. The theoretical advantage here is the reuse of learned spatial hierarchies (edges -> shapes -> textures) to overcome data scarcity, a critical technique in medical AI (Pan & Yang, 2010).

2.2 Relevance to Advanced AI Concepts

This project fulfills the course requirements by addressing multiple advanced domains:

1. Dataset-Driven AI: Implementing a complete pipeline from raw pixel data to classification.
2. Advanced Deployment: Utilizing Transfer Learning to solve a domain-specific problem with limited compute resources.
3. Robustness and Ethics: A key contribution of this work will be the evaluation of robustness (testing the model against synthetic noise/blur to simulate poor microscope focus) and Explainability (XAI). We will implement Grad-CAM (Gradient-weighted Class Activation Mapping) to visualize the model's focus, ensuring it detects parasites rather than artifacts (e.g., staining errors).

3. Mini Literature Review

The application of AI to medical imaging is a mature field, yet the trade-off between performance and interpretability remains an active area of research.

Rajaraman et al. (2018) established the baseline for the NIH Malaria dataset, demonstrating that pre-trained CNNs (like ResNet50) could achieve accuracies exceeding 95%, significantly outperforming manual microscopy. However, their work focused primarily on accuracy metrics, leaving the "explainability" aspect largely unexplored.

Poostchi et al. (2018) provide a comprehensive survey of image analysis in malaria, highlighting that while Deep Learning yields superior performance, traditional methods based on color intensity and texture features remain relevant for low-power hardware deployments. This supports the necessity of our comparative analysis: does the marginal gain in CNN accuracy justify the computational cost?

Pan and Yang (2010) offer the foundational theory for Transfer Learning, arguing that knowledge transfer is most effective when the source and target domains share low-level feature characteristics (e.g., edges and curves). This justifies our choice of using ImageNet-trained weights, as the fundamental visual features of "objects" are transferable to "cells" in early network layers.

Selvaraju et al. (2017) introduced Grad-CAM, a technique to produce visual explanations for decisions from a large class of CNN-based models. This literature is critical to our methodology,

as it provides the mathematical basis for auditing our deep learning model, ensuring that the "Neural Network logic" aligns with human medical knowledge.

Finally, Lipton (2018) argues in "The Mythos of Model Interpretability" that "interpretability" is often ill-defined. He suggests that for high-stakes decision-making (like medical diagnosis), post-hoc explanations (like Grad-CAM) are necessary to verify that models are not relying on spurious correlations (biases).

4. Proposed Methodology and Experiments

4.1 Implementation Strategy

The project will be implemented in Python using PyTorch for deep learning and Scikit-Image/Scikit-Learn for the classical baseline. The pipeline will consist of:

1. Data Preprocessing: Standardization of image dimensions (224×224 for CNN, flatten for RF) and min-max normalization.
2. Model A (Classical Baseline): Feature extraction using HOG (to capture the gradient changes caused by the parasite's presence) followed by a Random Forest classifier.
3. Model B (Deep Transfer Learning): A MobileNetV2 backbone (chosen for its efficiency in mobile deployment) with a custom fully connected head. We will freeze early layers and fine-tune the final blocks.
4. Explainability Module: Integration of a Grad-CAM hook to generate heatmaps for False Positives and True Positives.

4.2 Experimental Design

The evaluation will go beyond simple train/test splits:

- Stratified K-Fold Cross-Validation: (k=5) to ensure statistical reliability of results.
- Robustness Testing: We will generate a "noisy test set" (adding Gaussian noise and Gaussian blur) to measure the degradation of performance for both models. This simulates real-world conditions where microscope lenses may be dirty or out of focus.

4.3 Evaluation Metrics

The models will be compared using:

- Performance: Accuracy, Recall (Sensitivity), Specificity, and Area Under the Curve (AUC-ROC).
- Efficiency: Inference time (milliseconds per image) and model size (MB).
- Interpretability: Qualitative analysis of Grad-CAM heatmaps.

5. References

1. Rajaraman, S., et al. (2018). "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images." *PeerJ*, 6, e4568.
2. Poostchi, M., et al. (2018). "Image analysis and machine learning for detecting malaria." *Translational Research*, 194, 36-55.
3. Pan, S. J., & Yang, Q. (2010). "A Survey on Transfer Learning." *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359.
4. Selvaraju, R. R., et al. (2017). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization." *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626.
5. Lipton, Z. C. (2018). "The Mythos of Model Interpretability." *Queue*, 16(3), 31–57.