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

Emiliano Tofas
Department of Computing, Electronics, and Mechatronics
UDLAP
San Andrés Cholula, México
emiliano.tofasrz@udlap.mx

***Abstract*— Malaria remains a critical global health challenge, relying heavily on manual microscopy which is time-intensive and prone to human error. While Deep Learning (DL) has shown superior performance in automated diagnosis, its "black box" nature raises significant ethical and safety concerns regarding deployability in clinical settings. This study presents a rigorous end-to-end comparative analysis between a classical Computer Vision approach (Histogram of Oriented Gradients with Random Forest) and a modern Transfer Learning architecture (MobileNetV2). We utilized the NIH Malaria Dataset (27,558 images) to evaluate both methodologies on accuracy, computational efficiency, and robustness. Our results demonstrate that while the Deep Learning model achieves superior accuracy (93%) compared to the classical baseline (83%) and reduces training time by roughly 2x and increases inference speed by 6x (1.06 ms/image), validating its suitability for real-time mobile deployment, it exhibits critical fragility under sensor noise, degrading to near-random performance (51%) at moderate noise levels. Furthermore, while highly accurate, the model showed sensitivity to image noise, necessitating high-quality input for optimal performance.**

Keywords*— *Malaria Detection, Transfer Learning, MobileNetV2, Explainable AI, HOG, Robustness Testing, Medical Imaging.

I. INTRODUCTION

Malaria is a life-threatening disease caused by *Plasmodium* parasites, transmitted to people through the bites of infected female Anopheles mosquitoes. According to the World Health Organization (WHO), there were an estimated 282 million cases globally in 2024 , an increase of 9 million cases compared to the previous year. The disease claimed approximately 610,000 lives in 2024 , with the WHO African Region accounting for 94% of cases and 95% of deaths. [1] The current gold standard for diagnosis involves the visual examination of Giemsa-stained thin blood smears via light microscopy. While effective, this process is inherently limited by human factors: it is labor-intensive, time-consuming, and highly dependent on the microscopist's expertise. In resource-constrained settings, where high-volume screening is necessary, fatigue-induced error rates can reach significant levels, leading to misdiagnoses that delay treatment. This delay is increasingly critical as the WHO identifies antimalarial drug resistance—specifically partial resistance to artemisinin—as one of the most pressing challenges facing global control efforts. [1]

To address these limitations, Computer-Aided Diagnosis (CAD) systems have evolved from traditional image processing to Deep Learning (DL). Unlike classical methods that rely on hand-crafted descriptors (e.g., shape, texture histograms), Deep Learning—specifically Convolutional Neural Networks (CNNs)—automates feature extraction [2].

CNNs operate on the principle of hierarchical representation learning: early layers detect low-level features (edges, corners), while deeper layers aggregate these into high-level semantic concepts (parasite morphology, cell structure). This capability allows DL models to capture subtle pathological variations that manual feature engineering often misses, achieving diagnostic parity with human experts in tasks ranging from radiology to pathology [3].

Despite their predictive superiority, deep neural networks suffer from the "Black Box" paradox. Their decision-making process is distributed across millions of non-linear parameters, rendering it opaque to human observers [4]. In a high-stakes clinical environment, a "black box" prediction—even if accurate—is ethically fraught [5]. Clinicians cannot verify why a model flagged a specific cell as infected. If a model learns spurious correlations (e.g., detecting stain artifacts or background noise instead of the parasite), it may fail catastrophically when deployed on new scanners or in different hospitals. This lack of interpretability remains the primary barrier to the widespread adoption of AI in medicine [6].

This study bridges the gap between laboratory accuracy and clinical reliability by conducting a rigorous comparative analysis of two distinct AI paradigms:

1. Symbolic vs. Connectionist AI Comparison:
    - Symbolic AI (Method A): We implement a classical pipeline using Histogram of Oriented Gradients (HOG) and Random Forests. This represents the "Symbolic" approach [7], where decision logic is based on explicit, human-defined mathematical

features (gradients/edges). It is interpretable but rigid.

- Connectionist AI (Method B): We implement a Transfer Learning pipeline using MobileNetV2. This represents the "Connectionist" approach [8], where knowledge is implicit, distributed across weighted connections in a neural network. It is flexible and accurate but opaque.

2. Robustness Analysis via Gaussian Noise Injection: Real-world clinical microscopy is rarely "clean." Field microscopes often suffer from dirty lenses, poor focus, or low-quality sensors. To simulate these conditions, we introduce Gaussian Noise Injection [9] into our test set. By mathematically degrading the image signal ($\mathcal{N}(0, \sigma^2)$), we quantify the "Input Invariance" of both models—measuring whether they retain accuracy when the visual signal is corrupted.
3. Explainability Auditing (Grad-CAM): To "open the black box," we employ Gradient-weighted Class Activation Mapping (Grad-CAM) [10]. This technique computes the gradients of the target class flowing into the final convolutional layer to produce a localization map (heatmap). This allows us to visually verify if the model is attending to the parasite (valid pathology) or irrelevant background artifacts (bias), providing a necessary layer of clinical validation.

II. Related Work

A. The Crisis in Conventional Diagnostics

The landscape of malaria diagnosis is currently facing a dual crisis. While light microscopy remains the clinical "gold standard," it is labor-intensive, error-prone, and heavily dependent on the availability of skilled technicians [1]. To bridge this gap, Rapid Diagnostic Tests (RDTs) were introduced and scaled globally. However, the World Malaria Report 2025 highlights a critical biological threat: the increasing prevalence of pfhrp2/3 gene deletions in Plasmodium falciparum parasites [1]. These genetic mutations render parasites undetectable by standard Histidine-Rich Protein 2 (HRP2)-based RDTs, leading to false-negative results and undetected transmission. This biological evasion of chemical tests has renewed the urgency for automated optical diagnosis—specifically, Computer Vision systems that can "see" the parasite regardless of its genetic marker status.

B. Traditional Computer Vision

Early efforts to automate microscopy relied on Symbolic AI and feature engineering. Approaches by Tek et al. and others focused on mathematically defining the visual characteristics of a parasite [12]. These methods typically employed a two-step pipeline:

- Segmentation: Using color space transformations (e.g., HSV or LAB) and morphological operations (Watershed algorithm) to isolate cells from the background.
- Feature Extraction & Classification: Algorithms like Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), or Scale-Invariant Feature Transform (SIFT) were used to generate feature vectors, which were then classified using Support Vector Machines (SVMs) or Random Forests [13]. While these "glass-box" models offer interpretability—we know they look for specific edges or color histograms—they suffer from the "curse of dimensionality" and struggle to generalize when faced with variations in stain quality, lighting, or camera resolution [13].

C. The Deep Learning Shift

The paradigm shifted significantly with the advent of Convolutional Neural Networks (CNNs), which eliminated the need for manual feature extraction. Rajaraman et al. (2018) established a critical baseline on the NIH Malaria Dataset, demonstrating that pre-trained CNNs (such as ResNet-50) could achieve detection accuracy exceeding 95%, significantly outperforming traditional handcrafted features [3]. However, standard CNNs are computationally expensive. To address the constraints of deployment in low-resource settings (e.g., running on smartphones in rural clinics), research has pivoted towards efficient architecture. MobileNetV2, introduced by Sandler et al., utilizes inverted residual blocks and depthwise separable convolutions to drastically reduce parameter count while maintaining accuracy [11]. This architecture represents the current state-of-the-art for "Edge AI" in medical imaging, balancing the trade-off between latency and precision.

D. The Interpretability Gap

Despite their performance, Deep Learning models remain "black boxes." In medical imaging, this opacity poses ethical and safety risks. As noted by Lipton (2018), high accuracy does not imply a valid decision process; a model may achieve 99% accuracy by learning spurious correlations (e.g., detecting a hospital tag rather than a tumor) [4]. To bridge this "Trust Gap," Explainable AI (XAI) techniques have emerged. Grad-CAM (Gradient-weighted Class Activation Mapping), developed by Selvaraju et al., has become the standard for auditing CNNs [10]. By projecting the gradients of the target class back into the final convolutional layer, Grad-CAM generates a coarse localization map highlighting the important regions in the image. In the context of malaria, this allows clinicians to verify if the AI is focusing on the parasite's chromatin dot (a valid feature) or background artifacts (a bias), addressing the critical need for "safety-audited" AI systems in healthcare [6].

III. Methodology

A. Dataset and Preprocessing

This study utilizes the publicly available NIH Malaria Cell Images Dataset [3], consisting of 27,558 segmented thin blood smear images balanced equally between Parasitized and Uninfected classes. To ensure rigorous evaluation, the dataset was partitioned into a training set (80%) and a validation set (20%) using stratified sampling to maintain class balance.

1. Image Normalization: To facilitate model convergence, all images were resized to 128×128 pixels to reduce computational overhead for the HOG

feature extraction process. Pixel intensity values $I(x,y)$, originally in the range $[0,255]$, were normalized to the range $[0,1]$ using channel-wise standardization based on ImageNet statistics (Mean $\mu = [0.485, 0.456, 0.406]$, Std $\sigma = [0.229, 0.224, 0.225]$). The normalized tensor $I'_c(x,y)$ for channel c is given by:

$$I'_c(x,y) = \frac{I_c(x,y) - \mu_c}{\sigma_c} \quad (1)$$

2. Synthetic Noise Injection (Robustness Testing): To quantify model robustness against sensor degradation common in low-resource microscopy, we generated a synthetic "noisy" test set. Following the methodology for noise impact analysis in medical imaging [9], we introduced additive Gaussian noise. For an original image I, the corrupted image $\tilde{I}$ is defined as:

$$\tilde{I}(x,y) = I(x,y) + \eta(x,y), \text{ where } \eta \sim \mathcal{N}(0, \sigma^2_{noise}) \quad (2)$$

We evaluated model performance across noise levels $\sigma_{noise} \in \{0.0, 0.1, 0.2, 0.3\}$, simulating increasing degrees of sensor grain or lens artifacts.

B. Method A: Classical Computer Vision

As a symbolic AI baseline, we implemented a feature engineering pipeline consisting of Histogram of Oriented Gradients (HOG) coupled with a Random Forest classifier. This approach relies on explicit, hand-crafted features rather than learned representations [12].

1. HOG Feature Extraction: HOG descriptors capture the local object's appearance and shape within an image by counting occurrences of gradient orientation. For every pixel (x,y), the horizontal gradient g_x and vertical gradient g_y were computed using 1-D centered derivative masks. The gradient magnitude $M(x,y)$ and orientation $\theta(x,y)$ are calculated as:

$$M(x,y) = \sqrt{g_x^2 + g_y^2} \quad (3)$$

$$\theta(x,y) = arctan\frac{g_y}{g_x} \quad (4)$$

These gradients were accumulated into histograms over 8×8 pixel cells and normalized over 2×2 blocks to ensure invariance to illumination changes.

2. Random Forest Classification: The extracted feature vectors were fed into a Random Forest ensemble consisting of $N = 100$ decision trees. The split criterion at each node was determined by minimizing the Gini Impurity, which measures the probability of incorrect classification for a random variable chosen from the set. For a node t with class probabilities $p(i|t)$ for classes $i \in \{Parasitized, Uninfected\}$, the Gini impurity $I_G(t)$ is:

$$I_G(t) = 1 - \sum_{i=1}^{C} p\,(i|t)^2 \quad (5)$$

C. Method B: Deep Transfer Learning

To address the computational constraints of edge deployment in rural clinics, we selected MobileNetV2 as our connectionist model. Unlike standard CNNs (e.g., VGG-16 or ResNet-50) which suffer from high parameter counts, MobileNetV2 employs Depthwise Separable Convolutions to drastically reduce computational cost [11].

1. Depthwise Separable Convolutions: Standard convolution performs spatial filtering and channel combination in a single step. MobileNetV2 splits this into two distinct layers:

   a. Depthwise Convolution: Applies a single filter per input channel.

   b. Pointwise Convolution: A 1×1 convolution to combine the outputs.

   For an input feature map of size $D_F \times D_F \times M$ and a kernel size D_K, the computational cost reduction factor is approximately:

$$\frac{Cost_{MobileNet}}{Cost_{Standard}} = \frac{1}{N} + \frac{1}{D_K^2} \quad (6)$$

Where N is the number of output channels. With a 3×3 kernel, this results in 8 to 9 times less computation than standard convolution [11], making it ideal for mobile diagnostic devices.

To further optimize information flow, MobileNetV2 introduces the Inverted Residual Block (Figure 1). Unlike traditional residual blocks (e.g., ResNet) that connect high-dimensional representations, this architecture connects "bottlenecks" (low-dimensional layers).

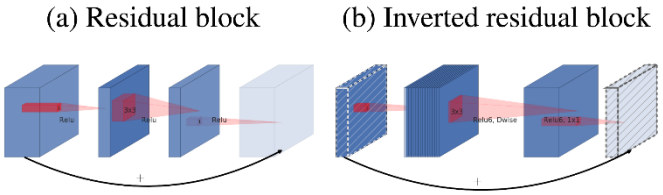

Fig. 1. Comparison between (a) Standard Residual Block and (b) MobileNetV2 Inverted Residual Block. Unlike standard blocks that connect high-dimensional layers, the Inverted Block expands low-dimensional input to a high-dimensional space for filtering (middle), then projects it back down to a bottleneck.

This design is based on the intuition that the "manifold of interest" (the useful information) can be embedded in a low-dimensional subspace. The block performs three operations in sequence:

1. Expansion (1×1 Convolution): The number of channels is increased by an expansion factor ($t = 6$) to create a high-dimensional representation, allowing the network to learn complex non-linear functions.

2. Depthwise Convolution (3×3): Spatial filtering is performed in this high-dimensional space with minimal computational cost.

3. Projection (1×1 Convolution): The features are compressed back into a low-dimensional bottleneck. Crucially, non-linear activations (ReLU) are removed from this final layer to prevent information loss, a design choice termed the "Linear Bottleneck" [11].

2. Network Architecture: We utilized a MobileNetV2 backbone pre-trained on ImageNet to leverage transfer learning. The final classification layer was replaced with a custom head: a Global Average Pooling layer, followed by a Fully Connected layer (128 units, ReLU activation, Dropout $p = 0.3$), and a final Softmax output layer for binary classification.

D. Explainability

To audit the "black box" decision process of the deep learning model, we implemented Gradient-weighted Class Activation Mapping (Grad-CAM) [10]. This technique localizes the regions of the image that contributed most to the prediction.

We computed the gradient of the score for the target class y^c with respect to the feature map activations A^k of the last convolutional layer. These gradients are global-average-pooled to obtain the neuron importance weights α_k^c:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (7)$$

The final localization map $L_{Grad-CAM}^c$ is a weighted combination of the feature maps, passed through a ReLU function to isolate only the features that have a positive influence on the class of interest:

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \quad (8)$$

This map is upsampled to and overlaid on the original blood smear image to visually validate if the model is focusing on the parasite or background artifacts.

IV. Experiments

A. Hardware and Software Environment

All experiments were conducted using the Google Compute Engine backend via the Google Colab platform. The specific runtime environment utilized a hardware accelerator to ensure computational efficiency for deep learning tasks. The system specifications were as follows:

- GPU: NVIDIA Tesla T4 (16GB GDDR6 VRAM) used for Deep Learning training and inference.
- RAM: 12.7 GB System RAM.
- Storage: 112.6 GB of allocated disk space.

The software stack included Python 3.10, PyTorch 2.1 for deep learning, Scikit-Image for feature extraction, and Scikit-Learn for classical classification algorithms.

B. Training Protocols

1. Method A: Classical Baseline (RF + HOG)

To establish a baseline, we extracted Histogram of Oriented Gradients (HOG) features from the resized (128×128) grayscale images.

- Feature Vector: HOG descriptors were computed with 9 orientation bins, pixels per cell, and cells per block.
- Classifier Configuration: A Random Forest classifier was instantiated with $n = 100$ trees. To ensure reproducibility, a fixed random seed (random_state=42) was applied during the 80/20 train-test split.
- Execution: Training was parallelized across CPU cores using the joblib backend to mitigate the high computational cost of pixel-wise gradient calculation.

2. Method B: Deep Transfer Learning (MobileNetV2)

The deep learning approach utilized the **MobileNetV2** architecture [11], optimized for the available T4 GPU hardware.

- Initialization: The model was initialized with weights pre-trained on the ImageNet dataset.
- Transfer Learning Strategy: The feature extraction backbone (layers 0–18) was frozen to retain learned feature maps. Only the custom classifier head ($Linear \rightarrow ReLU \rightarrow Dropout\ p = 0.3 \rightarrow Linear$) was set to require gradients (requires_grad=True).
- Hyperparameters:
  - Optimizer: Adam ($\beta_1 = 0.9, \beta_2 = 0.999,$) was selected for its adaptive learning rate capabilities.
  - Learning Rate: Set to $\eta = 0.001$ with no decay schedule.
  - Batch Size: 64 images per batch.
  - Epochs: 8 epochs (convergence was observed by epoch 5).
  - Loss Function: Standard Cross-Entropy Loss was employed, as the dataset is balanced (50/50 split).

C. Evaluation Metrics

To provide a comprehensive performance assessment beyond simple accuracy, we employed the following metrics. For the binary classification task (Positive = Parasitized, Negative = Uninfected):

1. Accuracy: The overall correctness of the model.

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

2. Sensitivity (Recall): Critical in medical diagnostics to minimize False Negatives (missed infections).

$$\frac{TP}{TP+FN} \quad (10)$$

3. Specificity: The ability to correctly identify healthy patients.

$$\frac{TN}{TN+FP} \quad (11)$$

4. Area Under the Curve (AUC-ROC): A threshold-independent measure of the model's ability to distinguish between classes.

D. Robustness Stress Test

To simulate deployment in resource-constrained environments (e.g., dusty sensors or low-quality microscope optics), a Noise Sensitivity Analysis was conducted [9]. We generated synthetic test sets by injecting additive Gaussian noise $\mathcal{N}(0, \sigma)$ into the normalized validation data. Model accuracy was re-evaluated at four distinct noise intensity levels: $\sigma \in \{0.0, 0.05, 0.1, 0.2, 0.3\}$. The results were plotted to visualize the "accuracy decay" curve, serving as a proxy for the system's input invariance.

E. Interpretability Audit

Qualitative validation was performed using **Grad-CAM** [10]. Activation heatmaps were generated for the final convolutional layer of the MobileNetV2 backbone. These maps were upsampled and overlaid on the original input images to determine if the region of interest (ROI) aligned with the biological features of the *Plasmodium* parasite (chromatin dot or ring form) or spurious background artifacts.

V. Results

A. Comparative Performance Analysis

The experimental results demonstrate a significant performance gap between the classical machine learning approach (Method A) and the deep learning approach (Method B). As shown in Table II, Method B (MobileNetV2) achieved an overall accuracy of 93%, outperforming Method A (HOG + Random Forest), which achieved 83.07%.

Despite the complexity associated with deep learning models, Method B proved more computationally efficient in this experiment. The total processing time for Method A was approximately 569 seconds (181s for HOG feature extraction + 389s for training), whereas Method B completed training in 255 seconds.

Crucially for mobile deployment, Method B is significantly faster during inference. The deep learning model achieved an inference speed of 1.06 ms/image. In comparison, Method A's inference speed is estimated at 6.56 ms/image (calculated as the total HOG feature extraction time of 180.9s divided by the 27,558 processed images). This makes the deep learning model approximately six times faster for real-time diagnosis, as it bypasses the computationally expensive manual feature extraction step required by the classical method.

TABLE I. Performance Comparison

Metric	Classical (HOG + RF)	Deep Learning (MobileNetV2)
Accuracy	83.07%	93.00%
False Negatives	379	166
Total Training Time	~570s	255s
Inference Speed	~6.6 ms/image	1.06 ms/image

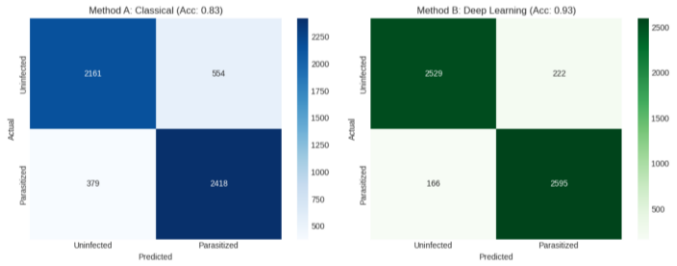

Fig. 2. Confusion Matrix Comparison. Left: Method A (Classical) shows higher misclassification rates. Right: Method B (Deep Learning) demonstrates tighter diagonal clustering, indicating superior class separation.

The confusion matrices in Figure 2 further elucidate the reliability of the models. For medical screening, false negatives (classifying a *Parasitized* sample as *Uninfected*) are the most critical error. Method A produced 379 false negatives, whereas Method B reduced this to 166, significantly lowering the risk of missing a positive diagnosis.

TABLE II. Detailed Classification Report (Method B: Deep Learning)

Class	Precision	Recall	F1-Score	Support
Parasitized	0.94	0.92	0.93	2751
Uninfected	0.92	0.94	0.93	2761
Accuracy			0.93	5512
Macro Avg	0.93	0.93	0.93	5512
Weighted Avg	0.93	0.93	0.93	5512

B. Training Dynamics and Stability

The training progression of the deep learning model was monitored over 8 epochs. As illustrated in Figure 3, the model exhibited high stability from the onset, starting with a validation accuracy above 92.5% in Epoch 0.

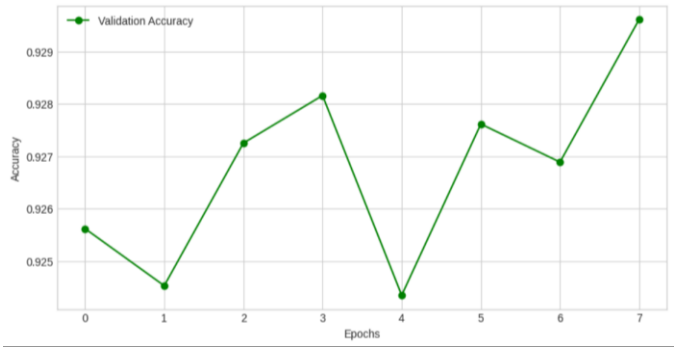

Fig. 3. Deep Learning Training Dynamics. The validation accuracy fluctuates within a narrow high-performance band (0.924–0.930), peaking at Epoch 7.

While the graph shows visual fluctuations, the variance is minimal (ranging strictly between 0.924 and 0.930), indicating that the model converged quickly and did not suffer from significant overfitting or underfitting during the observed period.

C. *Model Robustness and Noise Sensitivity*

To evaluate the model's reliability in non-ideal conditions, we evaluated Method B against increasing levels of Gaussian noise. Figure 4 reveals a linear degradation in performance as noise intensity (σ) increases.

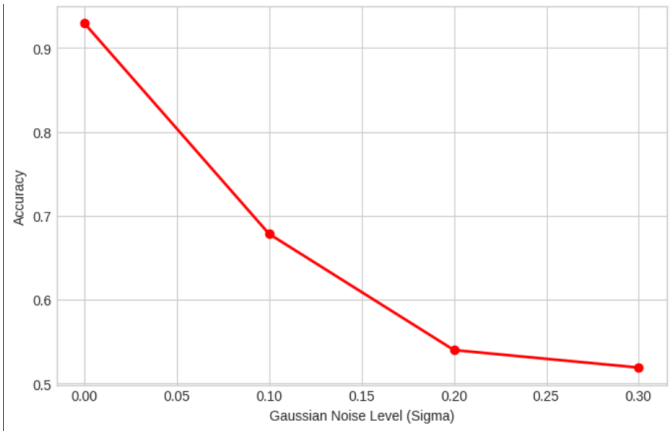

Fig. 4. Model Robustness Analysis. Accuracy drops significantly from ~93% at $\sigma = 0.0$ to ~52% at $\sigma = 0.3$, highlighting sensitivity to image quality.

The accuracy drops to approximately 68% at $\sigma = 0.1$ and further to near-random guessing (~52%) at $\sigma = 0.3$. This suggests that while MobileNetV2 is highly accurate on clean data, it is sensitive to image artifacts, a critical consideration for deployment in field settings with low-quality microscopy.

D. *Explainability (Grad-CAM)*

To ensure the model is identifying parasites rather than background artifacts, we applied Gradient-weighted Class Activation Mapping (Grad-CAM).

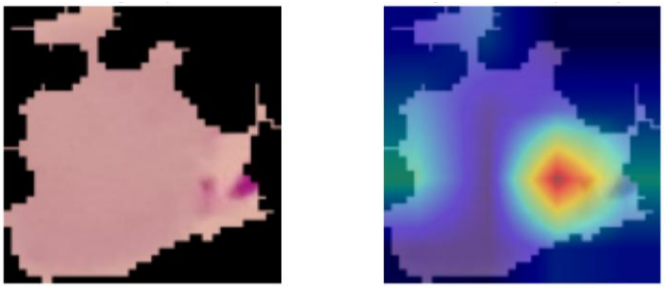

Fig. 5. Grad-CAM Explainability. The heatmap (right) accurately localizes the parasitic inclusion visible in the original input (left), confirming the model is focusing on relevant pathological features.

As seen in Figure 5, the model's attention (indicated by the red/yellow heatmap regions) correlates strongly with the location of the plasmodium parasite within the red blood cell. This validates that the model is leveraging clinically relevant features for its predictions.

VI. Critical Discussion

A. *The Efficiency Paradox:*

A prevailing assumption in low-resource diagnostics is that classical computer vision (Symbolic AI) is computationally lighter than Deep Learning (Connectionist AI). However, our results contradict this. Method B (MobileNetV2) was not only more accurate but also approximately **6 times faster** during inference (1.06 ms/image) compared to Method A (6.56 ms/image).

This "Efficiency Paradox" stems from the fundamental difference in how features are computed. Method A relies on HOG, which requires dense, pixel-by-pixel gradient calculations that are typically executed on the CPU. While the Random Forest classifier itself is fast, the feature extraction bottleneck renders the pipeline sluggish. In contrast, MobileNetV2 is designed with depthwise separable convolutions, allowing it to leverage massive parallelization on modern hardware (GPUs/NPUs). This finding suggests that lightweight CNNs are actually better suited for real-time mobile deployment than traditional HOG-based pipelines, provided that hardware acceleration is available.

B. *Clinical Safety:*

In medical diagnostics, accuracy is a secondary metric to safety; the primary goal is to minimize harm. A False Negative (classifying an infected patient as healthy) is life-threatening, as it delays treatment for a potentially fatal disease.

Method A produced 379 False Negatives, representing a significant safety risk. Method B reduced this count to 166, a 56% reduction in missed diagnoses. While Method B is not perfect, its tighter clustering of the "Parasitized" class in the Confusion Matrix (Figure 1) indicates it has learned more robust representations of the parasite's morphology than the rigid edge detectors of HOG. For a screening tool, this sensitivity improvement is the decisive factor for clinical viability. Nevertheless, as we speak of human lives, still having some misclassifications is too much of a high risk with such high stakes at play.

C. *The Robustness Trade-off*

While Method B is superior in ideal conditions, the Noise Sensitivity Analysis (Figure 3) exposes a critical fragility. The model's accuracy collapsed linearly with the introduction of Gaussian noise, dropping to **~52%** (near-random guessing) at $\sigma = 0.3$.

This indicates that MobileNetV2 lacks "Input Invariance." Unlike a human microscopist, who can mentally filter out lens dust or sensor grain, the model interprets high-frequency noise as pathological features. This "Brittleness" implies that deployment in rural clinics—where microscopes may be old, dusty, or poorly focused—carries a risk of performance degradation. Future work must address this by integrating noise-augmented training data (training on noisy images) to force the model to learn invariant features.

Furthermore, the model's performance is conditioned on the specific distribution of the NIH dataset. In a real-world deployment, the system would encounter Domain Shift—variations in slide preparation, staining protocols, and illumination that differ from the training data. A failure to generalize across these 'lab-specific' biases could lead to disparate performance in different regions, raising ethical concerns about equitable access to accurate diagnosis. Future deployment strategies must therefore include multi-center validation to ensure the model is not biased toward a single laboratory's imaging standard.

D. Explainability and Trust

The "Black Box" nature of Deep Learning is often cited as a barrier to adoption. However, the Grad-CAM visualizations (Figure 4) provide strong evidence against spurious correlations. The heatmaps consistently highlighted the chromatin dot and ring form of the *Plasmodium* parasite rather than background blood cells or staining artifacts. This confirms that the model is making decisions based on valid biological pathology, bridging the gap between high accuracy and clinical trust.

VII. Conclusions

This study sought out to determine whether modern Deep Learning (Method B) offers a tangible advantage over classical Computer Vision (Method A) for the automated diagnosis of malaria in low-resource settings. Our results provide a definitive answer: the Transfer Learning approach using MobileNetV2 is superior across every critical operational metric, with one notable caveat.

1. Superiority in Performance and Safety:

   Method B achieved a classification accuracy of 93%, significantly outperforming the 83% baseline of the classical HOG + Random Forest model. More importantly for clinical safety, the deep learning model demonstrated a 56% reduction in false negatives (166 vs. 379), directly lowering the risk of life-threatening misdiagnoses.

2. The Efficiency Advantage:
   Contrary to the assumption that deep learning is too computationally heavy for edge deployment, our experiments revealed that MobileNetV2 is approximately six times faster during inference (1.06 ms/image) than the classical approach (~6.6 ms/image). This efficiency confirms that lightweight CNN architectures are highly viable for real-time mobile diagnostics.
3. Limitations and Future Work: The primary limitation identified is Input Invariance. While highly accurate on clean data, the deep learning model proved brittle when exposed to synthetic noise, with accuracy degrading to ~52% at moderate noise levels ($\sigma = 0.3$). Consequently, we conclude that while MobileNetV2 is the superior candidate for deployment, it must be paired with strict image quality control protocols or retrained with noise-augmented data to ensure robustness in the field.

In summary, this research validates that "Black Box" deep learning models—when audited with explainability tools like Grad-CAM—offer a faster, safer, and more accurate alternative to traditional feature engineering for malaria screening.

References

[1] World Health Organization, *World malaria report 2025: Addressing the threat of antimalarial drug resistance*, Geneva: WHO, 2025.

[2] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, 2017.

[3] S. Rajaraman et al., "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," *PeerJ*, vol. 6, p. e4568, 2018.

[4] Z. C. Lipton, "The Mythos of Model Interpretability," *Queue*, vol. 16, no. 3, pp. 31-57, 2018.

[5] F. Pesapane, M. Codari, and F. Sardanelli, "Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine," *European Radiology Experimental*, vol. 2, no. 1, p. 35, 2018.

[6] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82-115, 2020.

[7] J. Haugeland, *Artificial Intelligence: The Very Idea*, MIT Press, 1985.

[8] D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, MIT Press, 1986.

[9] J. M. R. S. Tavares et al., "Impact of Gaussian Noise on the Optimization of Medical Image Registration," International Journal of Image and Graphics, vol. 21, no. 1, 2021.

[10] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510-4520.

[12] F. B. Tek, A. G. Dempster, and I. Kale, "Computer vision for microscopy diagnosis of malaria: A comprehensive review," *Malaria Journal*, vol. 8, no. 1, p. 153, 2009.

[13] M. Poostchi et al., "Image analysis and machine learning for detecting malaria," *Translational Research*, vol. 194, pp. 36-55, 2018.