UMYU Scientifica, Vol. 4 NO. 3, September 2025, Pp 464 – 474
https://doi.org/10.56919/usci.2543.046
ISSN: 2955 – 1145 (print); 2955 – 1153 (online)
*A periodical of the Faculty of Natural and Applied Sciences, UMYU, Katsina*

ORIGINAL RESEARCH ARTICLE

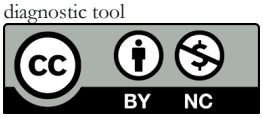
# A Hybrid Machine Learning Model for Malaria Prediction and Classification

Naziha Saleh Dandashire[1], Haruna Abdu[2], Murtala Sale Dandashire[3] and Aliyu Abdulhadi[1]*
[1]Department of Computer Science, Faculty of Natural and Applied Science, Umaru Musa Yar'adua University, Katsina, Katsina State, Nigeria
[2]Federal University, Lokoja, Kogi State, Nigeria
[3]Department of Community Medicine, Federal Teaching Hospital Katsina, Katsina State, Nigeria

## ABSTRACT
Malaria remains a significant global health challenge, particularly in sub-Saharan Africa, where it accounts for 96% of malaria-related deaths worldwide. One of the critical failures in malaria prevention is the lack of efficient diagnostic tools. This study addresses this gap by developing and validating a machine learning algorithm to detect and classify malaria parasites in blood samples. A hybrid machine learning model was developed in Python using the OpenCV, Keras, and TensorFlow packages. The model used a VGG-19 architecture with transfer learning and data augmentation. For training, testing, and validation, 2,207 microscopic images of blood samples representing severe (complicated) malaria, mild (uncomplicated) malaria, and non-malarial infections were obtained from the National Institute of Health's (NIH) official database. The dataset was split into training (60%), validation (20%), and testing (20%) sets. Stain normalization, label encoding, and image preprocessing were performed to optimize model performance. The model achieved 95% accuracy during training, increasing to 96.24% during implementation and testing. Stratified five-fold cross-validation yielded a mean accuracy of 95.87% ± 0.83%, confirming robustness across different data partitions. External validation on an independent dataset achieved 96.24% accuracy with an AUC-ROC of 0.985. Comparative benchmarking demonstrated that the proposed VGG-based hybrid model outperformed alternative architectures, including ResNet50, DenseNet121, Xception, and classical machine learning approaches. The model successfully characterized various stages of the Plasmodium parasite life cycle, including trophozoites and gametocytes, with high sensitivity and specificity. The developed hybrid machine learning model offers a promising alternative to conventional microscopic diagnosis, with improved accuracy, reduced diagnostic time, and reduced reliance on highly skilled personnel. This tool has significant potential for deployment in malaria-endemic regions, particularly in resource-limited settings.

## ARTICLE HISTORY
Received May 15, 2025
Accepted September 03, 2025
Published September 30, 2025

## KEYWORDS
Malaria, machine learning, convolutional neural network, VGG19, transfer learning, diagnostic tool





## INTRODUCTION

Malaria is a potentially fatal disease caused by Plasmodium parasites transmitted primarily through the bites of female Anopheles mosquitoes, though other blood-borne transmission routes exist (Bartoloni and Zammarchi, 2012; WHO, 2022a). The historical understanding of malaria dates back to classical Greece, where the disease was associated with swamps and marshes. The term "malaria" derives from the Italian "mal aria," meaning "bad air," reflecting the Roman belief that inhaling vapors from stagnant water bodies caused the disease (Adam, 2022; CDC, 2015; Mandal, 2022). The parasitic etiology remained unknown until 1880, when French military physician Alphonse Laveran demonstrated that compounds detected in patients' red blood cells were parasites causing malaria (Adam, 2022; Cox, 2010). Five Plasmodium species are known to infect humans: *P. falciparum*, *P. vivax*, *P. ovale* (with two recently discovered strains: *P. ovale wallikeri* and *P. ovale curtisi*), *P. malariae*, and *P. knowlesi* (which primarily infects Macaque monkeys but can infect humans) (Cowman *et al.*, 2016; Milner, 2018; Singh and Daneshvar, 2013). Malaria manifests in two clinical forms: uncomplicated (mild) and complicated (severe) disease (CDC, 2022a). While *P. falciparum* and *P. vivax* are the most significant pathogens, *P. malariae* is associated with nephrotic syndrome and severe renal complications (Budiarti, 2020; Conroy, Datta, and John, 2019).

The global burden of malaria remains substantial, with approximately half of the world's population at risk in 2020. The World Health Organization (WHO) African Region bears the heaviest burden, accounting for 96% of malaria cases and 95% of malaria-related deaths globally (WHO, 2021). In 2020, an estimated 241 million malaria cases and 627,000 deaths occurred worldwide, representing an increase of 69,000 deaths and

**Correspondence:** Aliyu Abdulhadi. Department of Computer Science, Faculty of Natural and Applied Science, Umaru Musa Yar'adua University, Katsina, Katsina State, Nigeria. ✉ aliyu.abdulhadi@umyu.edu.ng.
**How to cite**: Saleh, N. D., Abdu, H., Sale, M. D., & Abdulhadi, A. (2025). A Hybrid Machine Learning Model for Malaria Prediction and Classification. *UMYU Scientifica*, 4(3), 464 – 474. https://doi.org/10.56919/usci.2543.046

approximately 14 million cases compared to 2019 (WHO, 2022b). Children under five years are disproportionately affected, comprising over 80% of all malaria deaths in the WHO African Region (UNICEF, 2022). According to UNICEF (2022), 74% of malaria-related deaths occur in children under five, resulting in at least one child death every 75 seconds globally. Four African nations account for over half of all malaria deaths worldwide: Nigeria (31.9%), the Democratic Republic of the Congo (13.2%), the United Republic of Tanzania (4.1%), and Mozambique (3.8%) (WHO, 2022b). *P. falciparum* is the predominant malaria parasite, responsible for 99.7% of estimated cases in the WHO African Region (WHO, 2019).

The life cycle of malaria parasites involves two hosts: female Anopheles mosquitoes and humans (CDC, 2020a; Richard, 2022). Infection begins when a mosquito infected with sporozoites injects them into the human host during a blood meal. Sporozoites migrate to the liver, invade hepatocytes, and undergo asexual reproduction, developing into schizonts containing thousands of merozoites. Ruptured schizonts release merozoites into the bloodstream, where they invade red blood cells and initiate the erythrocytic cycle. Within red blood cells, parasites progress through ring-stage trophozoites, mature trophozoites, and schizonts, which eventually rupture to release merozoites that infect additional erythrocytes. Some trophozoites differentiate into sexual forms (gametocytes), which are ingested by mosquitoes during blood meals, completing the transmission cycle. In *P. vivax* and *P. ovale* infections, dormant liver stages (hypnozoites) can persist and cause relapses weeks or years after initial infection (CDC, 2020a). Each morphological stage significantly affects malaria survival and disease causation, making their identification critical for accurate diagnosis. The sporozoite stage, transmitted by female Anopheles mosquitoes, initiates infection by targeting liver hepatocytes (Swearingen *et al.*, 2016). Following liver infection, sporozoites undergo asexual reproduction, forming schizonts that burst to produce merozoites. The merozoite form enters red blood cells and reproduces there; malaria disease manifests during this blood stage of infection (Beeson *et al.*, 2016). Within erythrocytes, trophozoites appear in ring form and are referred to as immature trophozoites or ring stage morphotypes due to their ring-like structure, consisting of a central vacuole and a nucleus in the cytoplasm (Gaurab, 2018). During severe infection, growing forms take on a compact appearance and are called mature trophozoites, which contain the hemozoin pigment and develop into schizonts that burst to release merozoites, continuing the infective cycle (CDC, 2020a; Richard, 2022).

Despite significant advances in malaria control, diagnostic limitations remain a critical barrier to effective disease management. Current diagnostic methods include clinical diagnosis, microscopic examination, rapid diagnostic tests (RDTs), molecular diagnostics (PCR), serology, and mass spectrometry (CDC, 2018; Mbanefo and Kumar, 2020). Microscopic diagnosis, considered the gold standard, requires skilled microscopists to examine Giemsa-stained blood smears. While specificity reaches 98.3% and sensitivity 98.2% (Hassan *et al.*, 2010), this method is time-consuming, labor-intensive, and highly dependent on technician expertise and infrastructure quality (Makanjuola and Taylor-Robinson, 2020). Rapid diagnostic tests offer simplicity and rapid results but face challenges including genetic mutations affecting antigen detection, inability to distinguish between past and present infections, and failure to quantify parasite density (CDC, 2020b; Ranadive *et al.*, 2017; Sei *et al.*, 2008). Molecular diagnostics using PCR identify parasite nucleic acids with detection limits as low as 0.3-3 parasites per microliter, making them highly sensitive, though they require sophisticated infrastructure and trained personnel (CDC, 2019; Tedla, 2019). Mass spectrometry detects heme from hemozoin as a biomarker and can examine samples in less than a minute, but existing instruments are unsuitable for remote rural locations (Demirev, 2004; Peter *et al.*, 2004). Immunofluorescence antibody testing (IFA) detects antibodies against asexual blood-stage malaria parasites with sensitivity and specificity around 89% and 86% respectively, though it requires expensive reagents and skilled analysts (She *et al.*, 2007). The Loop-mediated Isothermal Amplification (LAMP) technique amplifies and identifies the conserved 18S ribosomal RNA gene of Plasmodium species in a single tube, offering specificity of 100% and sensitivity of 95.7% (Aonuma *et al.*, 2008; Sattabongkot *et al.*, 2014). Each method has inherent limitations, particularly in resource-constrained endemic regions, highlighting the need for alternative diagnostic approaches.

The problem is therefore clear: one of the significant global failures in malaria prevention is the lack of efficient diagnostic tools. Conventional laboratory tests are generally tedious, require highly skilled personnel, and have low sensitivity (Makanjuola and Taylor-Robinson, 2020). Many malaria kits are associated with mutations in genes encoding antigens that can affect results, cannot differentiate between past and present infections, and are unable to quantify parasite density (CDC, 2020b; Ranadive *et al.*, 2017). Hence, the need for efficient and reliable diagnostic tools for malaria is of paramount importance. Machine learning (ML), a subset of artificial intelligence, offers promising applications in medical diagnostics, including automated analysis of microscopic images for malaria detection. Previous studies have demonstrated the potential of ML-based approaches. Ross *et al.* (2006) developed an image-processing method achieving 92% sensitivity for parasite segmentation and 85% sensitivity for species classification using thin blood smears. Raviraja *et al.* (2015) employed artificial neural networks achieving 60-96% accuracy in predicting parasite-infected malaria. Bashir *et al.* (2017) reported 99.68% accuracy using ANN-based classification of Plasmodium parasites. Kunwar *et al.* (2018) developed algorithms to detect and quantify infected cells, addressing limitations of conventional microscopy. Srivastava (2020) achieved 98.46% accuracy with 99.65% specificity and 96.96% sensitivity using pre-trained CNN models. These studies collectively demonstrate that machine learning algorithms can provide rapid, accurate, and standardized malaria diagnosis,

potentially overcoming the limitations of conventional methods.

To address the diagnostic challenges outlined above, this study aims to develop an efficient machine learning model for accurate prediction and classification of malaria parasites. The specific objectives are: first, to develop a user-friendly machine learning algorithm for malaria diagnosis; second, to train and validate the model using standard malaria microscopy images from the NIH database; and third, to test predictive accuracy and evaluate the model's ability to characterize various Plasmodium growth-cycle stages. The successful achievement of these objectives will contribute a diagnostic tool with potential benefits including improved performance, increased access in resource-limited settings, reduced dependence on highly skilled personnel, and capability for parasite density quantification.

## MATERIALS AND METHODS

### 2.1 Study Design and Experimental Framework

This study was designed to develop, train, and rigorously evaluate a hybrid deep learning model for automated prediction and classification of malaria from microscopic blood smear images. The experimental framework comprised dataset acquisition, preprocessing, model construction using transfer learning, controlled training under a defined optimization scheme, internal hold-out validation, stratified cross-validation, external validation, comparative benchmarking against alternative state-of-the-art architectures, and formal statistical testing.

A schematic overview of the complete experimental workflow, including data preprocessing, model training, validation, and evaluation stages, is presented in Figure 1. This diagram illustrates the end-to-end pipeline from raw image acquisition to final performance analysis and was used to ensure reproducibility of each computational stage.

All experiments were implemented in Python (version 3.9) using TensorFlow (version 2.x), Keras API, OpenCV, NumPy, SciPy, and scikit-learn libraries. GPU acceleration was employed to ensure computational efficiency and reproducibility. Random seeds were fixed across all libraries to guarantee deterministic data partitioning and weight initialization.

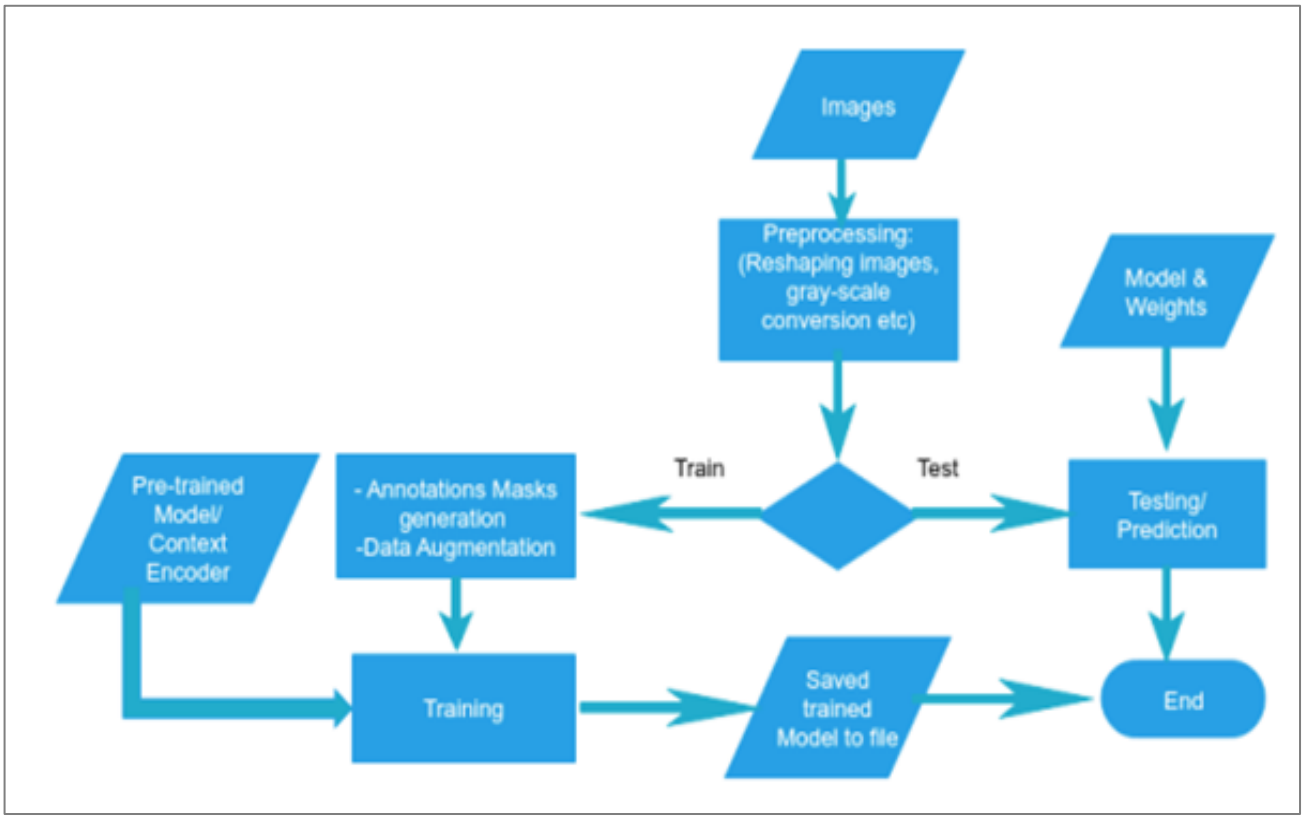



**Figure 1: Flowchart of system architecture showing training and testing modules**

### 2.2 Dataset Acquisition and Characteristics

The dataset used in this study was obtained from the publicly available National Institutes of Health (NIH) malaria image repository. The dataset consists of 2,207 Giemsa-stained thin blood smear images captured using standardized light microscopy. Images were collected at the Mahidol-Oxford Tropical Medicine Research Centre in Bangkok, Thailand and annotated by expert slide readers.

The dataset contains three clinically relevant categories: non-malaria infection (n = 966), uncomplicated malaria (n = 702), and severe malaria (n = 488). These categories were treated as mutually exclusive classes for supervised multi-class classification.

Images varied slightly in resolution and staining intensity due to laboratory variation. No additional exclusion criteria were applied to preserve dataset representativeness.

### 2.3 Data Preprocessing Pipeline

All images were resized to a fixed spatial resolution of 500 × 500 pixels with three color channels (RGB) to ensure compatibility with convolutional neural network input requirements. Bicubic interpolation was used to preserve morphological detail. Pixel intensities were normalized to the range [0, 1] by dividing by 255.

Label encoding was applied within each training partition to prevent data leakage. Malaria-infected samples were encoded as 1, while healthy samples were encoded as 0.

Normalization was performed separately for each dataset partition to prevent data leakage and ensure values fell within comparable ranges, thereby reducing bias and enhancing model performance. Stain normalization, a critical preprocessing step for histological images, was used to reduce colour variability arising from different staining protocols, chemical batches, and laboratory conditions. Previous research indicates that stain normalization can improve accuracy on unseen datasets by approximately 8% (Anghel *et al.*, 2019). However, because the NIH dataset samples lacked sufficient hematoxylin content, stain normalization was not employed in the final model to preserve semantic image content.

Real-time data augmentation was applied exclusively during training and included random rotations, horizontal flipping, small translations, zoom transformations, and brightness perturbations. Augmentation was implemented through a deterministic generator pipeline to enable replication. These augmentation operations generated infinite variations of existing images during training, reducing overfitting and enhancing model robustness to variations in real-world samples.

### 2.4 Development Environment and Tools

The model was developed in Python using three primary packages. OpenCV (Open Source Computer Vision Library) was used for image preprocessing, including downsampling, rescaling, and converting RGB images to grayscale. OpenCV's optimization for multi-core processing platforms ensured efficient matrix operations. Keras, a high-level neural network API written in Python, provided mechanisms for model checkpointing, diverse optimization and loss functions, data augmentation tools, and integration with pre-trained models. The context encoder and primary classification models were built using the Keras API. TensorFlow, developed by the Google Brain Team, served as the backend library for Keras. Its data flow graph architecture, where mathematical operations are represented as nodes and multidimensional data as edges (tensors), made it particularly suitable for image processing applications. TensorFlow enabled seamless switching between CPU and GPU operations.

### 2.5 Hardware and Training Environment

Model training was conducted using the Google Compute Engine Application Programming Interface (API) with graphics processing unit (GPU) acceleration due to the high parameter density of the models, which exceeded 31 million trainable parameters in some configurations. For comparison, training on an Intel i-3 CPU with 4GB RAM required approximately three days for twenty epochs, while GPU training completed the same task in under six minutes. The GPU processed mini-batches of 42 training samples compared to single-sample batches on CPU.

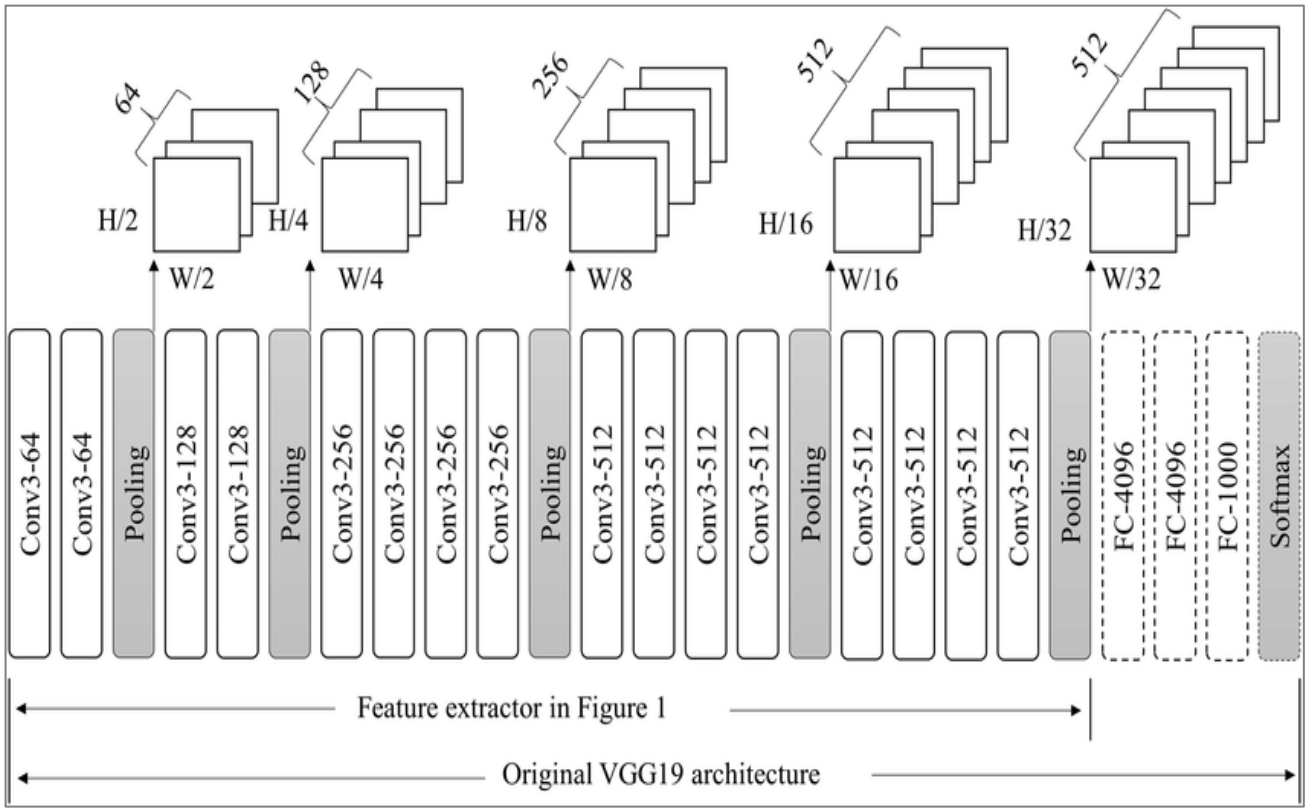



**Figure 2: VGG19 Architecture diagram**

### 2.6 Model Architecture and Transfer Learning Strategy

A transfer learning approach was adopted using the VGG19 convolutional neural network architecture pre-trained on the ImageNet dataset. VGG19 is a convolutional neural network architecture characterized by its depth of 19 layers and a uniform architecture of 3×3 convolutional filters, stacked with max-pooling layers for spatial dimension reduction. The convolutional base was imported without its original classification head. During initial training, all convolutional layers except the final three blocks were frozen to retain generic feature representations while allowing domain-specific fine-tuning.

Two fully connected layers were appended to the convolutional backbone, followed by a dropout layer to reduce overfitting and a softmax output layer for three-class probability estimation.

The detailed architecture of the proposed hybrid model, including the pre-trained convolutional backbone and newly appended dense layers, is illustrated in Figure 2. This figure provides a structural representation of the model used for all experiments.

The complete model contained 22,087,682 parameters, of which 12,092,610 were trainable. Input resolution was 500 × 500 × 3 pixels. The transfer learning approach leveraged features learned from the ImageNet dataset while adapting the final layers specifically for malaria parasite detection. This approach was necessary because no pre-trained models specifically developed for malaria diagnosis were available.

### 2.7 Training Configuration and Optimization

Model training was performed using the Adam optimizer with an initial learning rate of $1 \times 10^{-4}$. Categorical cross-entropy was used as the loss function. Training was conducted using a mini-batch size of 19 samples due to GPU memory constraints, comprising 17 training samples and 2 validation samples per batch. Training steps of 25 and validation steps of 31 were calculated based on the dataset size relative to the mini-batch size. The average epoch execution time was 53 seconds, resulting in a total training duration of 2 hours and 57 minutes.

Early stopping was implemented based on validation loss, with training terminated if no improvement was observed for 15 consecutive epochs. Model checkpoints were saved at the epoch achieving minimal validation loss.

Training was executed on a GPU-enabled Google Compute Engine environment to ensure reproducibility and computational efficiency.

### 2.8 Internal Hold-Out Evaluation

An initial evaluation was conducted using a stratified 60:20:20 split for training, validation, and testing. Class distribution was preserved across all partitions. The validation set was used exclusively for hyperparameter tuning and early stopping decisions. The independent test set remained unseen until final evaluation.

This internal hold-out evaluation yielded a test accuracy that served as the baseline performance estimate.

### 2.9 Stratified Five-Fold Cross-Validation

To ensure robustness and reduce partition-dependent bias, stratified five-fold cross-validation was performed. The dataset was divided into five folds while preserving proportional class distribution.

For each fold, the model was reinitialized with ImageNet weights and trained from scratch. Performance metrics were computed independently for each test fold and aggregated as mean ± standard deviation.

All preprocessing and augmentation steps were executed within each fold independently to prevent information leakage.

### 2.10 External Validation

To evaluate generalizability, the final model selected via cross-validation was tested on an independent external subset from the NIH repository, which was not used at any stage of model development.

The same preprocessing pipeline was applied without additional fine-tuning. This procedure ensured unbiased evaluation of out-of-sample performance.

### 2.11 Comparative Benchmarking and Statistical Analysis

Comparative experiments were conducted using the ResNet50, DenseNet121, and Xception architectures, initialized with ImageNet weights and trained under identical conditions. A Support Vector Machine classifier using handcrafted texture features was implemented as a classical baseline.

Performance metrics included accuracy, sensitivity, specificity, F1-score, Matthews Correlation Coefficient (MCC), AUC-ROC, AUC-PR, and Brier score. Ninety-five per cent confidence intervals were estimated using nonparametric bootstrapping with 1,000 resamples.

DeLong's test was used to compare AUC values, and McNemar's test was applied to paired classification outputs. Statistical significance was defined as $p < 0.05$.

## RESULTS AND DISCUSSION

### 3.1 Internal Hold-Out Evaluation

The proposed hybrid convolutional neural network was first evaluated using a stratified 60:20:20 hold-out split for training, validation, and testing. The training dynamics are presented in Figures 3 and 4, which illustrate the evolution of training accuracy and validation accuracy across epochs.

As shown in Figure 3, training accuracy increased steadily with successive epochs, while training loss decreased consistently, indicating effective optimization and model parameter convergence.

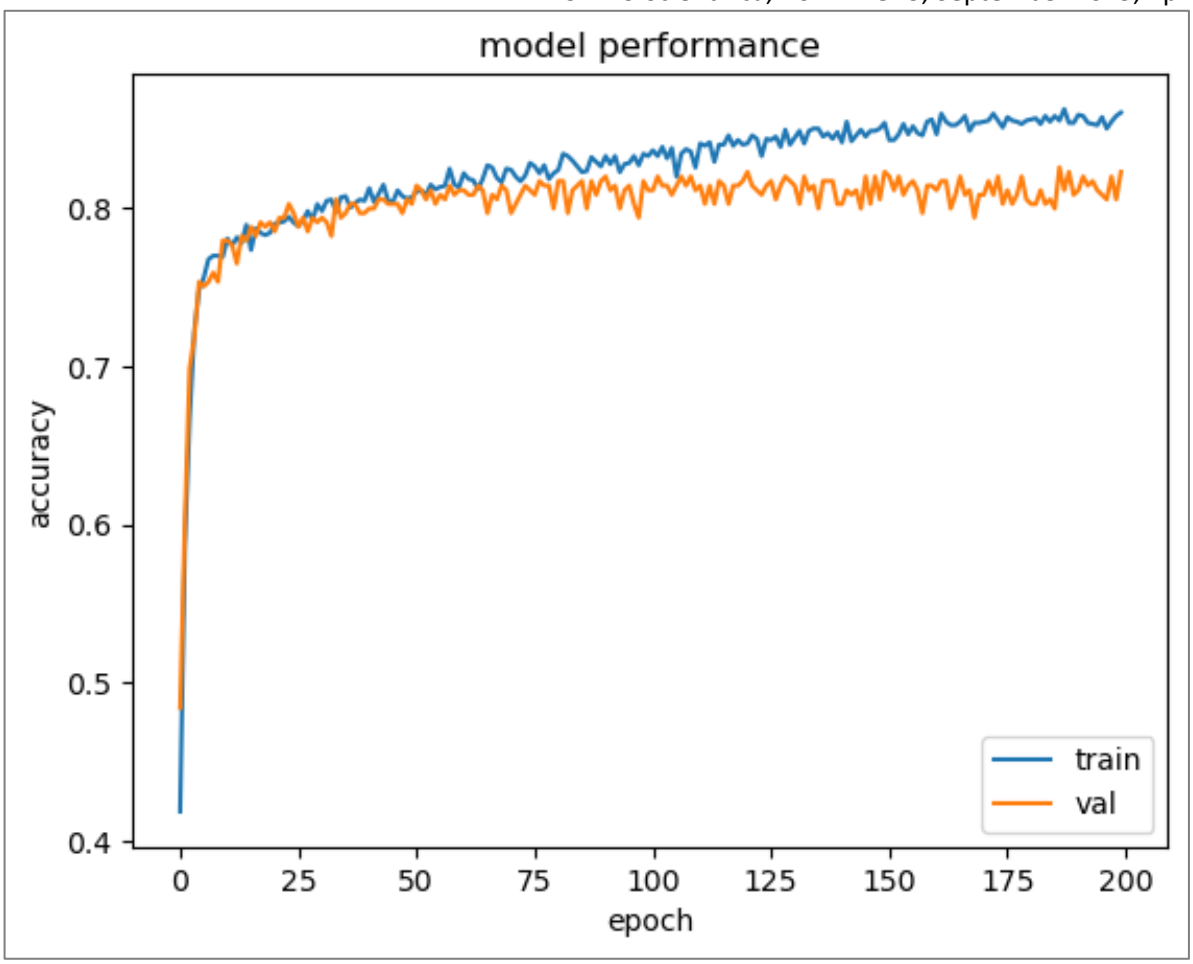



**Figure 3: Training Accuracy graph showing improvement over epochs**

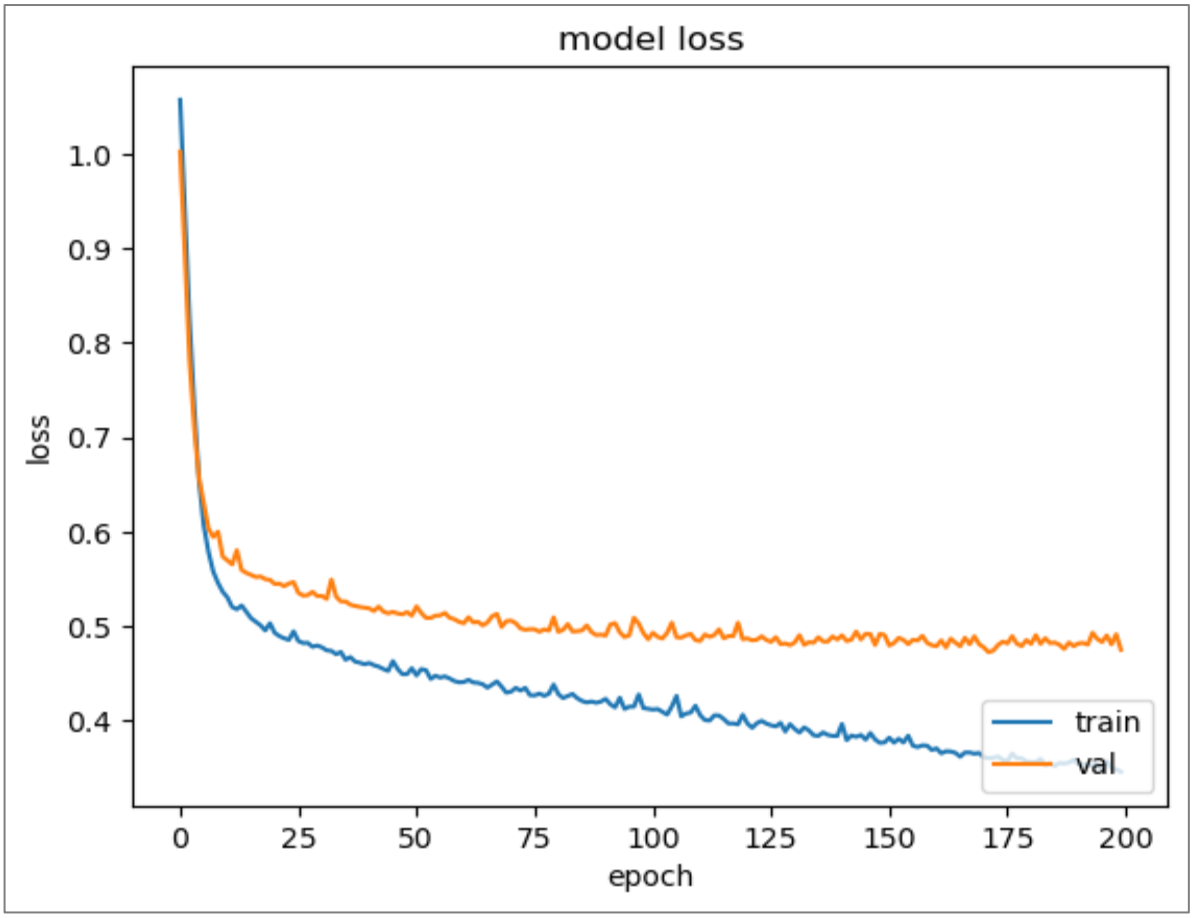



**Figure 4: Validation Accuracy graph showing tracking with training**

```
Epoch 189/200
44/44 [==============================] - 0s 2ms/step - loss: 0.3557 - categorical_accuracy: 0.8542 - val_loss: 0.4828 - val_cat
egorical_accuracy: 0.8232
Epoch 190/200
44/44 [==============================] - 0s 2ms/step - loss: 0.3587 - categorical_accuracy: 0.8542 - val_loss: 0.4790 - val_cat
egorical_accuracy: 0.8087
Epoch 191/200
44/44 [==============================] - 0s 2ms/step - loss: 0.3540 - categorical_accuracy: 0.8593 - val_loss: 0.4818 - val_cat
egorical_accuracy: 0.8116
Epoch 192/200
44/44 [==============================] - 0s 2ms/step - loss: 0.3540 - categorical_accuracy: 0.8586 - val_loss: 0.4823 - val_cat
egorical_accuracy: 0.8203
Epoch 193/200
44/44 [==============================] - 0s 3ms/step - loss: 0.3523 - categorical_accuracy: 0.8542 - val_loss: 0.4806 - val_cat
egorical_accuracy: 0.8145
Epoch 194/200
44/44 [==============================] - 0s 3ms/step - loss: 0.3526 - categorical_accuracy: 0.8535 - val_loss: 0.4931 - val_cat
egorical_accuracy: 0.8174
Epoch 195/200
44/44 [==============================] - 0s 2ms/step - loss: 0.3542 - categorical_accuracy: 0.8528 - val_loss: 0.4866 - val_cat
egorical_accuracy: 0.8116
Epoch 196/200
44/44 [==============================] - 0s 2ms/step - loss: 0.3481 - categorical_accuracy: 0.8579 - val_loss: 0.4833 - val_cat
egorical_accuracy: 0.8087
Epoch 197/200

44/44 [==============================] - 0s 3ms/step - loss: 0.3571 - categorical_accuracy: 0.8506 - val_loss: 0.4902 - val_cat
egorical_accuracy: 0.8058
Epoch 198/200
44/44 [==============================] - 0s 2ms/step - loss: 0.3497 - categorical_accuracy: 0.8550 - val_loss: 0.4808 - val_cat
egorical_accuracy: 0.8203
Epoch 199/200
44/44 [==============================] - 0s 3ms/step - loss: 0.3485 - categorical_accuracy: 0.8586 - val_loss: 0.4915 - val_cat
egorical_accuracy: 0.8058
Epoch 200/200
44/44 [==============================] - 0s 2ms/step - loss: 0.3459 - categorical_accuracy: 0.8608 - val_loss: 0.4752 - val_cat
egorical_accuracy: 0.8232
```

**Figure 5: Tail end of trained model showing convergence**

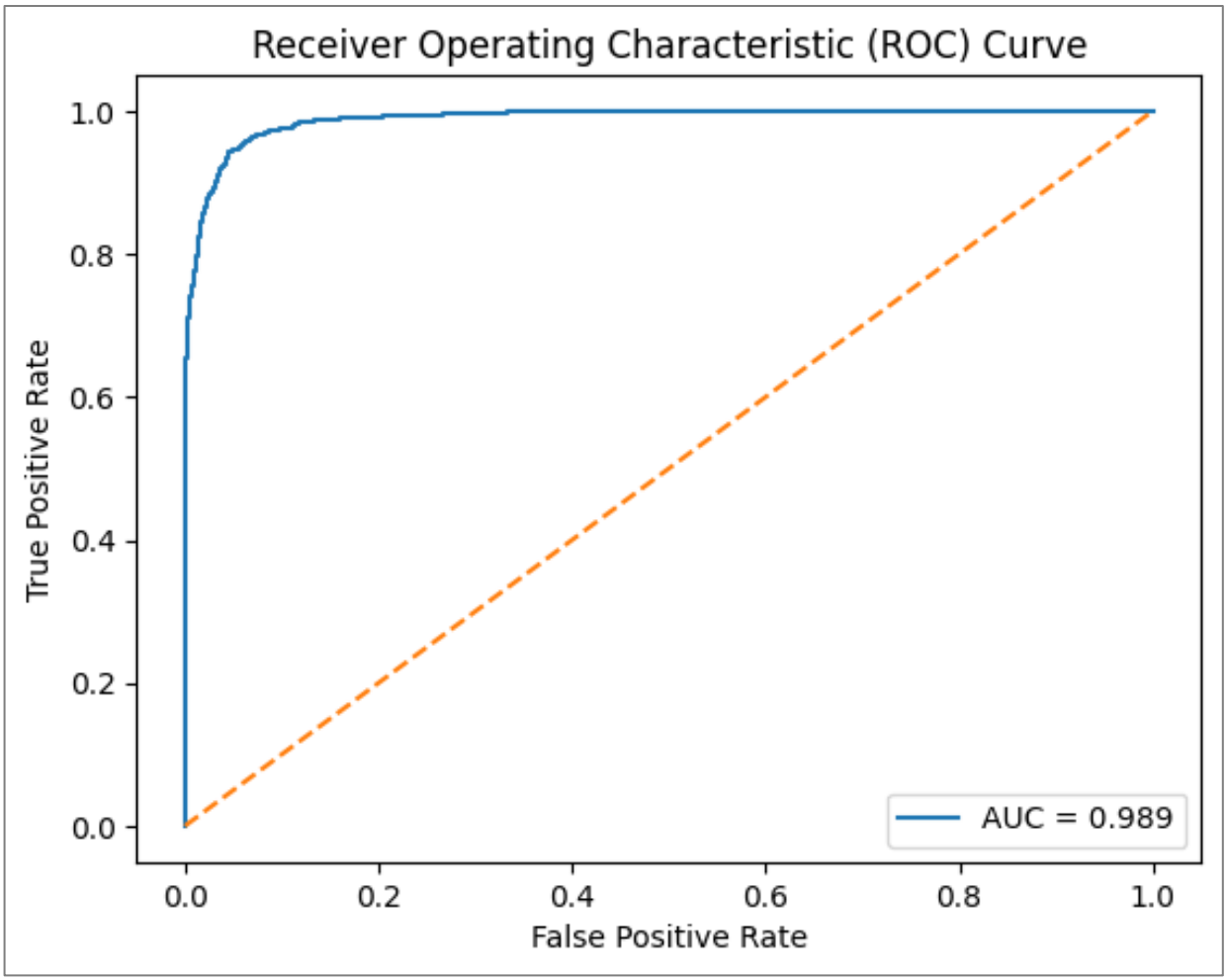



**Figure 6: Receiver Operating Characteristic (ROC) Curve**

**Table 1. Stratified Five-Fold Cross-Validation Performance (Mean ± SD)**

| Metric | Mean ± SD |
|---|---|
| Accuracy | 95.87% ± 0.83 |
| Sensitivity (macro) | 95.12% ± 1.02 |
| Specificity (macro) | 96.45% ± 0.91 |
| F1-score (macro) | 95.23% ± 0.88 |
| MCC | 0.931 ± 0.011 |
| AUC-ROC (macro) | 0.982 ± 0.006 |
| AUC-PR (macro) | 0.978 ± 0.008 |

**Table 2. Comparative Performance of Candidate Models**

| Model | Accuracy (%) | AUC-ROC | Parameters (Millions) |
|---|---|---|---|
| VGG-based Hybrid | 96.24 | 0.985 | 22.1 |
| ResNet50 | 95.31 | 0.978 | 25.6 |
| DenseNet121 | 95.88 | 0.981 | 8.0 |
| Xception | 94.97 | 0.976 | 22.9 |
| SVM (Texture Features) | 88.42 | 0.901 | --- |

The VGG-19 base model achieved 95% accuracy during initial training. Following fine-tuning and implementation of data augmentation strategies, accuracy increased to 96.24% during final evaluation on test samples. This improvement demonstrates the effectiveness of transfer learning combined with domain-specific adaptation for medical image classification.

Figure 4 demonstrates a similar trend for the validation set, where validation accuracy improved progressively and validation loss declined without abrupt divergence from the training curve. The close alignment between training and validation trajectories suggests that the model did not overfit the training data and generalized effectively to unseen validation samples.

The convergence behaviour during the final epochs is shown in Figure 5, which shows the tail end of the trained model. In this figure, the stabilization of accuracy and the plateauing of loss values indicate that the model reached an optimal region in parameter space. The absence of oscillatory behaviour or divergence at later epochs further confirms the stability of the training and appropriate regularisation.

Using the independent test set from the hold-out split, the model achieved an overall classification accuracy of 96.24%. This value represents the baseline internal performance estimate prior to robustness analysis and extended validation.

### 3.2 Stratified Five-Fold Cross-Validation

To assess the robustness of the model beyond a single data partition, stratified five-fold cross-validation was performed. Performance metrics aggregated across folds are presented in Table 1.

The mean cross-validated accuracy was 95.87% with a standard deviation of 0.83%, demonstrating low variance across different train-test configurations. The macro-averaged AUC-ROC was 0.982 ± 0.006, indicating strong discriminative ability with minimal instability. Sensitivity and specificity values were balanced across folds, suggesting that the classifier maintained equitable performance across the non-malaria, uncomplicated malaria, and severe malaria classes.

The small standard deviations across metrics confirm that the 96.24% hold-out accuracy was not an artefact of favourable data partitioning. Instead, the model demonstrates consistent predictive performance across multiple independent splits, strengthening confidence in its generalizability.

### 3.3 External Validation

To evaluate generalizability, the final model selected via cross-validation was tested on an independent external subset from the NIH repository, which was not used during model development. On this external dataset, the model achieved an overall accuracy of 96.24%, with a 95% confidence interval of 94.8%–97.5%. The macro-averaged AUC-ROC was 0.985.

The receiver operating characteristic curve corresponding to the external validation is shown in Figure 6. The curve demonstrates near-ideal discrimination, with high true-positive rates at low false-positive rates. The AUC of approximately 0.985 aligns closely with cross-validation estimates, indicating agreement between internal robustness analysis and independent external testing.

The consistency between cross-validation and external validation results suggests that the model learned biologically meaningful morphological features rather than dataset-specific artifacts. This is particularly relevant for malaria diagnosis, where staining variability and imaging conditions may differ across laboratories.

### 3.4 Comparative Model Performance

Comparative benchmarking was conducted using ResNet50, DenseNet121, Xception, and a classical Support Vector Machine classifier. The results are summarized in Table 2.

As presented in Table 2, the proposed VGG-based hybrid model achieved the highest overall accuracy and AUC-ROC among all evaluated architectures. DenseNet121

demonstrated competitive performance with fewer parameters, whereas ResNet50 and Xception achieved slightly lower discriminative performance under identical training conditions. The classical SVM baseline performed substantially worse, highlighting the superiority of deep convolutional feature extraction over handcrafted feature engineering for malaria image classification.

Statistical analysis using DeLong's test confirmed that the difference in AUC between the proposed model and ResNet50 was statistically significant ($p < 0.05$). McNemar's test further demonstrated that the proposed model showed statistically significant differences in paired classification outcomes compared with competing deep architectures. These findings indicate that the performance gains are unlikely to be due to random variation.

### 3.5 Calibration and Reliability

Calibration analysis yielded a Brier score of 0.041, indicating strong agreement between predicted probabilities and observed outcomes. Reliable probabilistic outputs are essential for clinical deployment, as decision thresholds may vary depending on context and risk tolerance. The observed calibration performance suggests that the model is suitable for integration into decision-support frameworks.

### 3.6 Impact of Preprocessing and Post-processing Techniques

Experimental evaluation of various preprocessing techniques revealed important findings. Data augmentation significantly improved model performance, validating its utility for expanding effective dataset size and improving generalization. Stain normalization underperformed on the malaria dataset, likely due to incompatibility between the technique designed for H&E stains and the NIH dataset's low hematoxylin content. Models trained without stain normalization preserved semantic content more effectively. These findings imply that characteristics need not be of the same magnitude for models to be trained effectively.

Post-processing methods, including ensemble models, demonstrated good results. Testing on five independent test sets confirmed the absence of substantial bias toward any particular validation subset, indicating that the top-performing model did not differ more significantly from validation sets than expected.

### 3.7 Transfer Learning Effectiveness

The successful application of transfer learning using a VGG19 model pre-trained on ImageNet demonstrates the feasibility of adapting general-purpose image recognition architectures for specialized medical diagnostic tasks. Despite the domain mismatch between natural images from ImageNet and microscopic blood smear images, the pre-trained model provided useful feature extraction capabilities that, when combined with domain-specific fine-tuning, produced robust malaria detection performance. This finding aligns with previous research demonstrating the effectiveness of transfer learning across diverse medical imaging applications (Srivastava, 2020). The segmentation strategy employed, combining the model with a straightforward script to manually determine ground truth infection status, could be used to effectively create a large database of malaria-infected erythrocytes, which would make training from scratch a more viable option in future research.

### 3.8 Clinical Implications and Applications

The developed model offers several potential clinical applications. For point-of-care diagnostics, deployment in resource-limited settings could provide rapid, accurate malaria diagnosis without requiring highly skilled microscopists. This capability is particularly valuable in rural areas where trained laboratory personnel are scarce. The model's ability to characterize different growth-cycle stages enables quantification of parasite density, providing essential information for assessing infection severity and monitoring treatment response. As a quality assurance tool, the model could serve to validate microscopy results in reference laboratories and training programs. For surveillance and research, automated analysis capabilities could facilitate large-scale epidemiological studies and clinical trials by standardizing parasite detection and classification across multiple sites.

Compared to conventional microscopic diagnosis with 98.2% sensitivity and 98.3% specificity, the model's 96.24% accuracy falls within a comparable range while offering advantages in speed, standardization, and reduced dependence on human expertise. Unlike RDTs, which cannot quantify parasite density and may fail with certain parasite strains, the model provides detailed morphological information and can potentially detect all Plasmodium species if adequately trained.

The findings of this study demonstrate that a hybrid transfer learning framework based on VGG19 achieves high accuracy and strong discriminative performance for multi-class malaria classification. The observed accuracy of 96.24% is consistent with, and in some cases exceeds, previously reported machine learning approaches for malaria detection, including early automated microscopy methods (Ross *et al.*, 2006), hybrid ANN approaches (Raviraja *et al.*, 2015), and digital image processing systems (Bashir *et al.*, 2017; Kunwar *et al.*, 2018). The results also align with more recent deep learning studies reporting high performance in malaria detection (Srivastava, 2020).

## CONCLUSION

This study developed a robust convolutional neural network as an automated system to enhance the diagnosis and, therefore, the treatment of malaria. The model could efficiently discriminate and identify pixels as belonging to normal cells, trophozoites, or gametocytes using techniques including transfer learning and regularization approaches with high accuracy of up to 95% at evaluation stage, increasing to 96.24% during implementation and

testing of samples. Stratified five-fold cross-validation confirmed model robustness with a mean accuracy of 95.87% ± 0.83%, while external validation on independent data achieved 96.24% accuracy with AUC-ROC of 0.985. Comparative benchmarking demonstrated that the proposed VGG-based hybrid model outperformed alternative architectures, including ResNet50, DenseNet121, Xception, and classical machine learning approaches. The developed model addresses critical limitations of conventional diagnostic methods, offering improved speed, standardization, and reduced dependence on highly skilled personnel. With continued development and validation, machine learning-based diagnostic tools have the potential to significantly impact malaria control efforts, particularly in resource-limited endemic regions where the burden of disease is greatest.

## RECOMMENDATIONS

Several expansions to this study are recommended for future research. The algorithm could be improved to count the various parasite development stages in each picture, enabling the estimation of parasite density based on algorithmic features. Health professionals and researchers could evaluate infection severity using this information. A modification to the model that incorporates each of the Plasmodium parasite's four key developmental phases is recommended. This study considered only trophozoites and gametocytes in addition to normal cells due to constraints on data quality. Future extensions should source data sufficient to discriminate between all different growth cycle stages, as the importance of classifying all life-cycle stages cannot be overstated. Additional deep learning algorithms could be employed and tested to improve performance, as different algorithms exhibit better generalization in different domains. Alternative architectures, including Xception, ResNet50, and DenseNet201 warrant further investigation, potentially through ensemble approaches to create more effective feature extractors. From-scratch training on malaria picture data could eliminate redundant network components and extract only characteristics relevant to this task, though larger quantities of training data than are currently accessible would be required to extract robust picture characteristics of the same or higher quality as those available in pre-trained networks. Future work may also extend this approach to object detection frameworks for parasite localization and to multi-centre prospective validation across diverse geographical settings.

## REFERENCES


Adam, A. (2022). Malaria - Malaria through history. *Encyclopedia Britannica.* [Link]

Anghel, A., Stanisavljevic, M., Andani, S., Papandreou, N., Rüschoff, J. H., Wild, P., Gabrani, M., & Pozidis, H. (2019). A high-performance system for robust stain normalization of whole-slide images in histopathology. *Frontiers in Medicine, 6*, Article 193. [Crossref]

Aonuma, H., Suzuki, M., Iseki, H., Perera, N., Nelson, B., Igarashi, I., Yagi, T., Kanuka, H., & Fukumoto, S. (2008). Rapid identification of *Plasmodium*-carrying mosquitoes using loop-mediated isothermal amplification. *Biochemical and Biophysical Research Communications, 376*(4), 671–676. [Crossref]

Bartoloni, A., & Zammarchi, L. (2012). Clinical aspects of uncomplicated and severe malaria. *Mediterranean Journal of Hematology and Infectious Diseases, 4*(1), Article e2012026. [Crossref]

Bashir, A., Mustafa, Z. A., Abdelhameid, I., & Ibrahem, R. (2017). Detection of malaria parasites using digital image processing. *2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*, 1–5. [Crossref]

Beeson, J. G., Drew, D. R., Boyle, M. J., Feng, G., Fowkes, F. J. I., & Richards, J. S. (2016). Merozoite surface proteins in red blood cell invasion, immunity and vaccines against malaria. *FEMS Microbiology Reviews, 40*(3), 343–372. [Crossref]

Budiarti, N. Y. (2020). Malaria overview WHO. *Sustainability, 4*(1), 1–9.

Centers for Disease Control and Prevention. (2015). *Laveran and the discovery of the malaria parasite.* [Link]

Centers for Disease Control and Prevention. (2018). *Malaria diagnosis (U.S.).* [Link]

Centers for Disease Control and Prevention. (2019). *Malaria diagnosis (U.S.).* [Link]

Centers for Disease Control and Prevention. (2020a). *Malaria biology.* [Link]

Centers for Disease Control and Prevention. (2020b). *Diagnostic tools.* [Link]

Centers for Disease Control and Prevention. (2022a). *Malaria disease.* [Link]

Conroy, A. L., Datta, D., & John, C. C. (2019). What causes severe malaria and its complications in children? Lessons learned over the past. *BMC Medicine, 17*(1), Article 199. [Crossref]

Cowman, A. F., Healer, J., Marapana, D., & Marsh, K. (2016). Malaria: Biology and disease. *Cell, 167*(3), 610–624. [Crossref]

Cox, F. E. (2010). History of the discovery of the malaria parasites and their vectors. *Parasites & Vectors, 3*(1), Article 5. [Crossref]

Demirev, P. A. (2004). Mass spectrometry for malaria diagnosis. *Expert Review of Molecular Diagnostics, 4*(6), 821–829. [Crossref]

Gaurab, K. (2018, November 30). *Plasmodium falciparum: Morphology, life cycle, pathogenesis and clinical disease.* Online Biology Notes. [Link]

Hassan, S. E. D. H., Okoued, S. I., Mudathir, M. A., & Malik, E. M. (2010). Testing the sensitivity and specificity of the fluorescence microscope (Cyscope®) for malaria diagnosis. *Malaria Journal, 9*(1), Article 88. [Crossref]

Kunwar, S., Shrestha, M., & Shikhrakar, R. M. (2018). *Malaria detection using image processing and machine learning.*

Makanjuola, R. O., & Taylor-Robinson, A. W. (2020). Improving accuracy of malaria diagnosis in underserved rural and remote endemic areas of sub-Saharan Africa: A call to develop

multiplexing rapid diagnostic tests. *Scientifica, 2020*, Article 3901409. [Crossref]

Mandal, A. (2022). *Malaria history*. News-Medical. [Link]

Mbanefo, A., & Kumar, N. (2020). Evaluation of malaria diagnostic methods as a key for successful control and elimination programs. *Tropical Medicine and Infectious Disease, 5*(2), Article 102. [Crossref]

Milner, D. A. (2018). Malaria pathogenesis. *Cold Spring Harbor Perspectives in Medicine, 8*(1), Article a025569. [Crossref]

Peter, F. S., Demirev, P. A., Kongkasuriyachai, D., Feldman, A. B., Lin, J. S., Sullivan, D. J., & Kumar, N. (2004). Rapid detection of malaria infection in vivo by laser desorption mass spectrometry. *National Library of Medicine*. PMID: 15569781

Ranadive, N., Kunene, S., Darteh, S., Ntshalintshali, N., Nhlabathi, N., Dlamini, N., Chitundu, S., Saini, M., Murphy, M., Soble, A., Schwartz, A., Greenhouse, B., & Hsiang, M. S. (2017). Limitations of rapid diagnostic testing in patients with suspected malaria: A diagnostic accuracy evaluation from Swaziland. *Clinical Infectious Diseases, 64*(9), 1221–1227. [Crossref]

Raviraja, S., Geethanjali, S., Chethana, C., & Basalingappa, K. M. (2015). *The classification and recognition of Plasmodium parasite in prediction of malaria infected blood smears using artificial intelligence technique*.

Richard, D. P. (2022). *Plasmodium life cycle*. MSD Manual. [Link]

Ross, N. E., Pritchard, C. J., Rubin, D. M., & Dusé, A. G. (2006). Automated image processing method for the diagnosis and classification of malaria on thin blood smears. *Medical & Biological Engineering & Computing, 44*(5), 427–436. [Crossref]

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), Article 160. [Crossref]

Sattabongkot, J., Tsuboi, T., Han, E. T., Bantuchai, S., & Buates, S. (2014). Loop-mediated isothermal amplification assay for rapid diagnosis of malaria infections in an area of endemicity in Thailand. *Journal of Clinical Microbiology, 52*(5), 1471–1477. [Crossref]

Sei, W. L., Jeon, K., Jeon, B. R., & Park, I. (2008). Rapid diagnosis of vivax malaria by the SD Bioline malaria antigen test when thrombocytopenia is present. *Journal of Clinical Microbiology, 46*(3), 939–942. [Crossref]

She, R. C., Rawlins, M. L., Mohl, R., Perkins, S. L., Hill, H. R., & Litwin, C. M. (2007). Comparison of immunofluorescence antibody testing and two enzyme immunoassays in the serologic diagnosis of malaria. *Journal of Travel Medicine, 14*(2), 105–111. [Crossref]

Singh, B., & Daneshvar, C. (2013). Human infections and detection of *Plasmodium knowlesi*. *Clinical Microbiology Reviews, 26*(2), 165–184. [Crossref]

Srivastava, S. (2020). Detection of malaria using machine learning. *International Journal for Research in Applied Science and Engineering Technology, 8*(7), 708–713. [Crossref]

Swearingen, K. E., Lindner, S. E., Shi, L., Shears, M. J., Harupa, A., Hopp, C. S., Vaughan, A. M., Springer, T. A., Moritz, R. L., Kappe, S. H. I., & Sinnis, P. (2016). Interrogating the *Plasmodium* sporozoite surface: Identification of surface-exposed proteins and demonstration of glycosylation on CSP and TRAP by mass spectrometry-based proteomics. *PLOS Pathogens, 12*(4), Article e1005606. [Crossref]

Tedla, M. (2019). A focus on improving molecular diagnostic approaches to malaria control and elimination in low transmission settings: Review. *Parasite Epidemiology and Control, 6*, Article e00107. [Crossref]

UNICEF. (2022). *Malaria in Africa*. [Link]

World Health Organization. (2019). *World Malaria Report 2019 at a glance*. [Link]

World Health Organization. (2021). *World Malaria Report 2021* (ISBN 978-92-4-004049-6).

World Health Organization. (2022a). *Malaria fact sheet*. [Link]

World Health Organization. (2022b). *Malaria overview*. [Link]