

Enhancing Malaria Diagnosis: A Hybrid Deep Learning Approach Integrating CNN-Swin Transformer, and Grad-CAM

Syed Noor Hussain Shah, Muhammad Shahan Ibad *, Sumaira Johar, Omar Bin Samin, Afsheen Khalid, Sami Ur Rahman

Center of Excellence in IT, Institute of Management Sciences (IMSciences), Peshawar, Peshawar, Pakistan

* Corresponding author: shaniims2022@gmail.com

Received: 09 October 2025, Revised: 04 March 2026, Accepted: 31 March 2026, Published: 01 April 2026

KEY WORDS	ABSTRACT
Malaria Detection Hybrid Deep Learning Swin Transformer Explainable AI Red Blood Cells (RBCs)	Malaria remains one of the most pressing global health challenges, demanding diagnostic solutions that are accurate, interpretable, and deployable in clinical practice. Most existing studies rely on traditional machine learning, which depends on handcrafted features and struggles to capture the heterogeneous morphology of malaria parasites. Deep learning approaches, particularly Convolutional Neural Networks (CNNs), have improved feature extraction but are limited by their local receptive fields, weak generalization across diverse blood smear images, and the absence of interpretability mechanisms. Moreover, only limited efforts have been made to utilize red blood cell (RBC) images for robust detection, and prior works seldom explored hybrid architectures or explainable AI (XAI), both of which are crucial for healthcare adoption. To address these gaps, this study proposes a hybrid deep learning framework that integrates CNNs for local feature representation with the Swin Transformer to capture global contextual dependencies. Additionally, XAI techniques, including Grad-CAM, are employed to provide transparent visual explanations by highlighting image regions most relevant to predictions. Experiments conducted on a benchmark malaria cell image dataset demonstrate state-of-the-art performance, achieving 96.8% accuracy, 96.7% precision, 96.9% recall, and a 96.8% F1-score. Confusion matrix analyses confirm balanced sensitivity and specificity, while interpretability visualizations validate the model's reliability in reducing each false negative and false positive. The results establish the proposed CNN-Swin Transformer with XAI integration as the first effective RBC-based hybrid framework for malaria detection, offering a robust, accurate, and interpretable diagnostic solution well-suited for both advanced clinical workflows and resource-constrained healthcare environments.

1. Introduction

Malaria has been one of the earliest diseases associated with humans. From the year 2700 B.C. in China, texts have described symptoms of fever that are close to those of people who have experienced the disease. Evidence that ancient civilizations around the world (India, Greece, and Mesopotamia) suffered from similar fevers has also been found in their historical records. The name malaria comes from Italy and means "bad air". This name came about because,

in medieval Europe, it was believed that the disease was transmitted through the "bad air" produced by marshes and swamps. In modern terms, we gained a greater understanding of malaria when Charles Laveran, a French army doctor, recognised that Plasmodium parasites existed in the blood of an affected individual in 1880 and received a Nobel Prize in Physiology or Medicine in 1907 for this work [1]. Soon after, Ronald Ross confirmed in 1897 that Anopheles mosquitoes transmit malaria, establishing the vector-borne nature of the disease [2]. During the

early 20th century, malaria caused over 100 million cases annually worldwide, especially in tropical and subtropical regions [3], making it one of humanity's most devastating infectious diseases.

Malaria remains one of the world's most serious public health challenges. According to World Health Organization (WHO) in 2023, estimation is around 263 million patients of Malaria globally, resulting of 597,000 departures. [4] The disease continues to disproportionately affect the WHO African Region, which accounted for about 94% of cases and 95% of deaths globally in 2023, with children under five bearing the heaviest burden. [5] Over recent years, progress in reducing incidence and mortality has not only slowed but in some locations reversed due to factors such as climate change expanding vector habitats, insecticide and drug resistance, and disruptions from humanitarian crises and the COVID-19 pandemic [4]. The number of malaria cases reported throughout the country of Pakistan has greatly increased over the past several years. After several areas experienced major flooding in 2022, malaria cases increased from 2020 when more than 2 million cases were recorded in the country during 2023 according to data provided by PMU common management unit. [6] In total, there were an estimated 2,726,491 confirmed malaria infections, 64.2% of which were caused by *Plasmodium vivax* (PV) while 32.1% were due to *Plasmodium falciparum* (PF), and an estimated 3.7% of cases were diagnosed as Mixed infections. [7] The provinces of Sindh and Balochistan had the highest numbers of reported malaria cases; therefore, people living in rural areas, especially in flood-affected areas will find it even more difficult to access proper diagnosis and treatment for malaria [8].

In recent years, technology has become increasingly important in combating malaria. Machine learning has been used to identify malaria from images of blood smeared on glass slides. However, this has not been an easy process. Most conventional models perform poorly at detecting infected RBCs because the parasite is located within intricate patterns that resemble the normal RBC structure or staining artifacts found on the slides. Nearly all previous studies approached this problem by treating it as a simple binary decision: was the RBC infected? The subtle differences in red blood cell (RBC) shapes and properties can make it difficult to differentiate between them using traditional methods. In our previous studies, we identified RBC morphological descriptors and have seen many different types of shapes. Here we present a novel, cross-disciplinary approach to the analysis of RBC shapes by incorporating an advanced architecture, the Swin

Transformer. The Swin Transformer can be thought of as an advanced CNN model. In addition to capturing local features within a relatively small distance (9 pixels in our case), it has the ability to identify long-range spatial relationships and similarities across larger distances. This allows us to design a model capable of accurate and reliable classification of RBC shapes — irrespective of the effects of staining or how the RBCs were captured in the image. By applying Explainable Artificial Intelligence (XAI) methods to our RBC classification model, we can provide - in addition to accurate results — the reasoning for each of our classifications, thus providing trust in the results to the healthcare provider using our system. Throughout the course of this project, our goal was not only to create a system capable of providing accurate results, but also to develop the necessary trust in our systems and how they are ultimately applied in practice.

The structure of the research article is made up of proceeds. In Section 2, we review related literature about malaria detection, including the weaknesses and problems of both traditional machine learning and traditional deep learning. In Section 3, we present our novel hybrid framework that combines Convolutional Neural Networks (CNNs) with the Swin Transformer and the use of Explainable Artificial Intelligence (XAI) methods. In Section 4, we provide details about the dataset, the preprocessing techniques we used to prepare it, and the experimental design and conditions. In Section 5, we report the results of our study, including how it compares to existing methods. In Section 6, we consider the broader impact of our approach and the applicability of the results, as well as the limitations of this work. Finally, in Section 7, we provide a conclusion to the research and suggest areas for future studies.

2. Literature Review

Researchers conducted a comparative study about malaria detection using deep learning technologies (in particular, CNNs) in which they used the NIH Malaria Dataset and a total of 27,558 images comprising blood smear images with definitive labeling of the sample image. They partitioned this dataset into three groups, training set, validation set and test set. The researchers implemented pre-processing steps on all data prior to using with CNN by collecting images from the dataset, resizing images to 300 pixels square, normalizing pixel values to between 0 and 1, and finally producing the appropriate number of labels per image. The researchers assessed various Models of CNN, including: CNN Baseline; CNN with a slimmed down architecture (Model 1); CNN with batch normalization (Model 2); CNN using data augmentation techniques

(Model 3); and CNN using VGG16 and transfer learning (Model 4). All models were implemented in TensorFlow/Keras using 10-fold cross validation, Binary crossentropy loss functions and Adam as the optimization. The different accuracy levels achieved by the respective models all have their own strengths and weaknesses; however model 1 achieved the highest level of accuracy (99%). In addition to an overall perfect recall, precision, and f1 score - model 1 produced the lowest resource requirements therefore making it a highly valid candidate for detecting malaria. Model 2 has a number of advantages including faster training, having balanced performance across the five different metrics, while model 4 offers an excellent level of feature extraction, however does incur a higher cost to compute. In contrast, Model 3 produced poor performance results due to excessive use of augmentation. The authors concluded that lightweight CNN models like Model 1 provide a scalable and accurate solution for malaria detection and recommended continued work using diverse datasets with advanced CNN architectures such as ResNet and EfficientNet, and incorporating multimodal data for marijuana detection. Ethical considerations were raised concerning the clinical use of these models [9].

The other study referred to was also about using machine learning/deep learning for advanced malaria detection through blood cell images. In this example the data set used for training comprised of 27,558 balanced blood cell images and was provided by the NIH/Kaggle Repository[10]. This project used Logistic Regression (LR), Support Vector Machine (SVM) and Inception-V3 as algorithms on the NIH/Kaggle repository dataset consisting of 27,558 balance blood cell images. Preprocessing steps were carried out on the blood cell images; these included rescaling, normalising and augmenting the dataset. The HOG features were extracted for SVM and transfer learning was employed with Inception-V3. The results for the performance of these algorithms for detecting malaria varied greatly; LR obtained a 65.38% accuracy, SVM raised this level to 84% and Inception-V3 outperformed both of the other algorithms with an accuracy of 94.52% after only five epochs of training. This illustrates the capability of deep learning techniques to classify/recognise images in the medical field, such as the malaria case. Moreover, the authors suggested that using more sophisticated models such as Inception-V3 can be used to perform better diagnosis in situations where there is an absence of enough resources and as such recommended future explorations on the implementation of ensemble techniques, optimising for real-time and clinical scalable deployments.

The optimized deep convolutional neural network (CNN) model developed in the new study, which is based on earlier work, included rotation-based features and was trained using a RotNet [11] self-supervised learning approach. There were 27,558 images (balanced) in the NIH Malaria Dataset with an 80/20 ratio split on training and test sets. In this dataset, both the infected and non-infected cells had equal amounts of images; therefore, each set of the dataset included all of the images for both types of cells. The authors trained the CNN to predict the rotation angle of an image (0 degrees, 45 degrees, 90 degrees, and 135 degrees) as a pretext task; after completing this task, the CNN learned features were transferred to various downstream classifiers, which included Modified ResNet50, DenseNet201, EfficientNet, Vision Transformer (ViT) and a custom CNN. The Modified ResNet50 performed the best among the classifiers, achieving 98.4% accuracy and AUC of 99.9% when using the entire dataset and 96.6% accuracy when using only 10% of the labelled data, outperforming traditional supervised approaches. Overall, the study reported a robust AUC of 99.8%, and also reported that RotNet significantly reduced the reliance on manual annotations; although limitations in computational complexity and generalizability to noisy clinical settings were also reported, the authors concluded that rotation-based self-supervised learning represents an approach that is both scalable and exceeding expectations for performance in medical image analysis and could be used in applications beyond malaria, such as segmentation and greater diagnostics.

A hybrid neural network was developed for the identification of malaria from blood smear photographs. This model combined Capsule Neural Networks (CapsNets) with Convolutional Neural Networks (CNN's) and provided improved performance on the dataset collected from the National Institutes of Health with 27,558 unique specimens. Images were trained using a number of sophisticated techniques that included generating more samples by creating 'augmented' images with the original specimen's 'flipping', 'rotating', 'normalising', 'balancing' (creating similar quantities of images for each class of specimen) and applying algorithmic parameters. The final version of the high performing hybrid model achieved a test set accuracy of 99% with a test set precision of 97.6% and a test set recall of 97.3%. The F1-score for the test set was 1.98 and the AUC for the test set was 97.4%. This hybrid model has produced a significant improvement over the performance of the Deep Selective Convolutional Network (DSCN-Net) used in previous studies (e.g. reduced False Alarm Rate (FAR) of 0.97%) and found

that CapsNets can provide more effective diagnostic capabilities for detecting malaria than conventional CNN's can provide[12]. Future work was recommended on optimizing CapsNet for deployment in resource-limited environments, expanding dataset diversity, and extending its applications to other domains such as cancer detection and mobile health integration.

In addition, Biosensor technologies have also been suggested to detect malaria besides software-based solutions. An example is that a study introduced a VLSI-based CMOS-compatible terahertz metasurface biosensor to detect cancer and malaria at two modes [13]. Elements that make up the biosensor included black phosphorus, graphene, silver, and gold on a silicon dioxide substrate with both experimental and theoretical measurements of absorption being analyzed with random forest regression. The biosensor operated in the frequency range of 0.189-0.423 THz with an excellent sensitivity of 1000 GHz/RIU to cancer and 143 GHz/RIU to malaria and quality factors of between 6.750-10.846 with a detection limit of 0.065 RIU and a high signal stability. The frequency changes of 30 GHz (cancer) and 20 GHz (malaria) provided frequency responses and robust resonance responses, and the R^2 scores were 94 percent predicting refractive index and 89 percent predicting incident angle changes. The authors had mentioned that this dual-mode biosensor was accurate and sensitive to diagnostics and offered concerns on the next research on miniaturisation, low cost and use of clinical samples to validate its findings. On the same note, a study on a hybrid graphene-copper terahertz metasurface biosensor machine learning-integrated device enhanced detection of malaria [14]. The authors used COMSOL Multiphysics to simulate the transmittance responses of the biosensor varying against the graphene's potential, the angle of incidence, the dimensions of the resonator and the concentration of the pathogen. This biosensor achieved impressive performance with a maximum sensitivity of 429 GHz/RIU, a figure of merit at 3.968 RIU⁻¹, a quality factor at 10.352, and a frequency shift of 70 GHz. The results indicated the biosensor's ability to adapt to various environmental conditions while demonstrating a strong correlation between frequency shift and the refractive index (RI). In addition, the use of linear regression to provide the predictive model yielded an $R^2 = 0.99806$, while Support Vector Regression yielded a perfect $R^2 = 100\%$. Thus, as indicated previously, these results support the potential of this biosensor in non-invasive and real-time malaria diagnostics, especially in low-resource settings. It was suggested the future work will include

extending this research to multiple testing systems and validation in clinical settings.

This study used previous biosensor work to establish the first terahertz, machine learning-assisted biosensor based on an MXene and graphene composite, which provides high sensitivity for detecting malaria [15]. The authors performed intermediate angle absorption measurements with angles between 0° and 40°, in addition to using COMSOL Multiphysics to conduct finite element modelling. They also used the Kubo equation to calculate the conductivity of graphene, as well as the Lorentz-Drude model to estimate the permittivity of MXenes. To predict the absorption coefficient, a (1D-CNN) was created with mean squared error (MSE) loss, L2 regularization (L2), dropout, and fivefold cross validation. This model showed an increase in the absorption coefficient from 0.193 to 1.094, a sensitivity of 500 GHz/RIU, a figure of merit (FOM) of 22.727 RIU⁻¹, and consistency in absorption spectra angles up to 40° (greater than or equal to 0°). Additionally, the 1D-CNN was determined to have an R^2 value of 1.00 for both training angles and greater than 98% accuracy on previously unseen intermediate angles, demonstrating its ability to provide data independent of the angle tested. The lowest detectable limit (LDL) was established as 0.042 with the biosensor being specifically designed for high specificity, label-free operation (i.e., with no marker needed to identify malaria) and fast point-of-care (POC) diagnostic testing. The authors of this study concluded that the 1D-CNN is well positioned as a candidate for the next generation of malaria detection because of its Internet of Things (IoT) compatibility and adaptability, and recommended future work to include research into multiplexed biomarker detection, microfluidic integration, and clinical validation. Another study investigated the development of an automated multi-model framework for malaria detection through combining deep learning (DL) algorithms and feature fusion [16]. The framework's implementation used the Kaggle "Cell Images for Detecting Malaria" dataset, consisting of 27,558 balanced, microscopic images, which were analyzed using three different pre-trained DL models (ResNet50, VGG16, and DenseNet201) to perform feature extraction. Fused feature vectors from all three models resulted in an 8064-dimensional feature vector. The features from this feature vector were reduced to 3135 using Principal Component Analysis (PCA). Hybrid classification was accomplished using both Support Vector Machines (SVM) and Long Short-Term Memory (LSTM) algorithms based on the results from the feature fusion.

Researchers in a separate study developed a phase classification system to enhance detection of malaria parasites using lensless inline digital holographic microscopy (LiDHM) [17]. To identify the malaria parasites, an established method of purification was performed - magnetic-activated cell sorting (MACS) was used to isolate infected red blood cells (RBCs) with *Plasmodium falciparum* (strain HB3); positively identified with Giemsa-stain microscopes. Holograms created by reconstructing the LiDHM holographic images, were analyzed via a phase-supported reconstruction algorithm written in Python to minimize artifacts, use adaptive thresholding, watershed segmentation, and morphological transformations. The authors extracted morphological and optical phase features from individual particles for parasitemia classification. A novel multi-feature thresholding approach improved parasitemia detection from 48% to 61%, while maintaining a high (98-100%) level of specificity. The authors found that there were very strong correlations ($r = 0.90$) between the optical phases, the refractive indices, and the total dry masses of particles, with phase being the primary metric to evaluate the diagnostic performance of each type of particle. The authors concluded that phase classification is a more sensitive and specific diagnostic indicator than morphology only, and that further studies should assess the ability of phase classification to classify early stage parasites, and implement a supervised classification process.

The next step in the development of deep learning approaches was the research of a CNN-based species-level malaria detection model with thick blood smear images [18]. The data was 5,941 images of Giemsa-stained microscopy (3,024 x 4,032 pixels each) of 350 patients in Bangladesh, labeled into 190,399 cell images. Preprocessing involved noise reduction, enhancement and edge localization, and imbalance in the classes was overcome by undersampling and StratifiedKFold cross-validation. The suggested CNN of seven channels, which included convolutional, batch normalization, Leaky ReLU, pooling, dropout, and fully connected layers, was trained using Adam optimization, and cross-entropy loss. The findings indicate that the accuracy of 99.51, precision and recall of 99.26 and specificity of 99.63 were impressive, better than three and six channel variants. The model proved to have better classification of *P. falciparum*, *P. vivax* and uninfected cells, eliminating the need to rely on human microscopy. In a complement to this, there was another study that examined interpretable customized CNN architectures to aid in the diagnosis of malaria using red blood cell smears [19]. The authors have then suggested Parallel CNN (PCNN), Soft Attention PCNN (SPCNN) and

Soft Attention after Functional Block PCNN (SFPCNN) using 26,161 re-annotated Giemsa-stained thin smear images of the Chittagong Medical College Hospital and other data sets to validate them. The preprocessing involved resizing, CLAHE, dilation, normalization, and standardization with training based on Adam optimization and dropout. Of these, it was found that SPCNN outperformed other custom models as well as benchmark transfer learning architectures, obtaining 99.37% accuracy, precision and recall with an AUC of 99.95, and with computational efficiency (2.207M parameters, 26 MB). It was also superior in the classification of species, detection of stages, and external validation, with the interpretability analyses (Grad-CAM, SHAP) demonstrating the diagnostic consistency.

An exploration of hybrid architectures included the use of an ensemble model that blended CNNs (ResNet18-based convolutional neural networks) with Vision Transformers (ViTs) for a total of 27,558 balanced microscopic red blood cell images from Mendeley Data. The CNN extracted local spatial features and the ViT captured global dependencies; these two types of features were concatenated and sent through the fully connected layers of the model using different forms of regularization, such as dropout, SMOTE, and early stopping. This model was trained in TensorFlow using the Adam optimization technique, while it obtained results of 99.64% accuracy, 99.51% F1-score, and an AUC of 1.00, which outperformed the respective models from both architectures independently. These results underline the effectiveness of hybrid CNN-ViT frameworks for scalable AI-based malaria detection. In addition, research published in Sensors developed an automated system for parasite and leukocyte detection using Romanowsky-stained blood smears [21]. Based on 1,000 thick smear images of *Plasmodium vivax* from Colombia's National Institute of Health, the pipeline included image quality assessment (via histogram and GLCM features), leukocyte detection (using OTSU thresholding and morphological operations), and parasite classification (using custom CNNs compared with VGG-19, MobileNetV2, and ResNet-50). According to the study conducted by a group of experts, the performance of the system was evaluated on different parameters and found to have a very high level of accuracy. These include 95% accuracy of quality assessment on images, 88.92% F1 for the identification of leukocytes, 98.65% accuracy with detecting parasites, and 86% accuracy for parasite stages classification. Besides, it was noted that this

system can be easily extended, allowing the implementation of such a system on mobile applications for the diagnosis of Malaria in resource-poor settings.

Furthermore, a similar study has shown that an integrated AI framework consisting of Vision Transformers and probabilistic models can be used for detecting and forecasting the treatment of Malaria. Using NIH Malaria Database images, supplemented with synthetic samples generated by GANs, the system incorporated ViTs for detection, self-supervised learning for feature extraction, Hidden Markov Models for life cycle modeling, Bayesian neural networks for uncertainty quantification, and graph neural networks for treatment prediction. With 99.1% accuracy for binary classification and 97% for stage-level classification, the framework represents a promising opportunity for improving malaria treatment through personalized applications. To address resource constraints, an ultra-compact SqueezeNet1.1 was developed, using the Kaggle/NIH dataset of 27,558 images of red blood cells. Three variants of the model were tested, which utilized a combined total of over 54 times fewer parameters than the original SqueezeNet1.1 model. Though significantly less parameter-heavy than the original model, the performance was very similar, with SqueezeNet1.1 achieving 97.12% accuracy and the fastest inference time for Variant 1, which achieved 96.55% accuracy with six times fewer parameters. All authors emphasized that, in addition to accuracy, reducing the number of parameters in a CNN model to achieve efficiency is of utmost importance in developing portable malaria diagnostics systems[23].

3. Methodology

Deep Learning has become more widely used in Medical Imaging Classification since its development; one key area of focus is Malaria Diagnosis through Microscopic Blood Smears. Traditional Convolutional Neural Networks (CNN) perform well with regards to Accuracy, however, CNNs do not provide as much information about how they generate their predictions (Interpretability) and they struggle to extend their features over long distances (Long Range Feature Extraction). To improve both Interpretability and Diagnostic Performance for this project, the Integration of CNN and Swin Transformer was used in conjunction with Explainable AI (XAI) Techniques.

The conceptual workflow of the proposed methodology is presented in Figure 1.

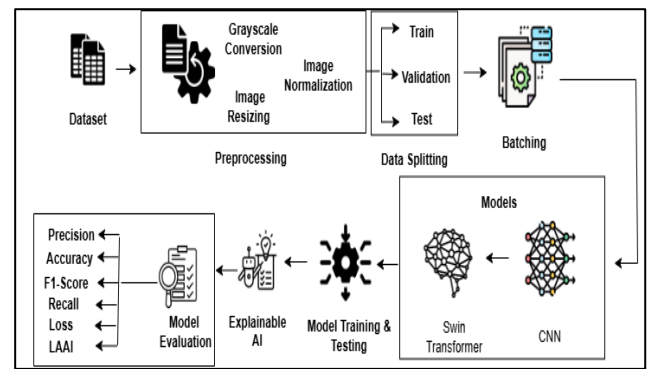


Fig 1. Indented methodology.

3.1 Dataset

The Malaria Cell Images dataset (27,558 cell images) on Kaggle is compiled from publicly available data. The dataset contains 27,558 images of cells, which are roughly evenly split between Parasitized (cells with malaria parasites) and Uninfected (cells without malaria parasites). The dataset contains images captured using a variety of staining techniques and lighting. Thus, it provides a good training dataset to develop models that can be easily adapted to a clinical setting.

Table 1

Dataset Details.

Feature Type	Count
Total Images	27,558
Classes	2 (Parasitized, Uninfected)
Image Type	Microscopic Blood Smear
Resolution	Variable
Missing values	None
Balanced Dataset	Yes

This dataset is categorized into two types of data, namely:

- Parasitized cells (i.e. red blood cell images that have been infected by malaria parasite Plasmodium).
- Uninfected cells (i.e. red blood cell images that are healthy and do not contain parasites). Images are presented in RGB color space and all images have different dimensions; therefore, image preprocessing steps need to be performed to resize, convert to greyscale, and normalize all of the images prior to training the model. The dataset is also well balanced for both classes so that models can be trained from an appropriately representative distribution of the two types of cells.

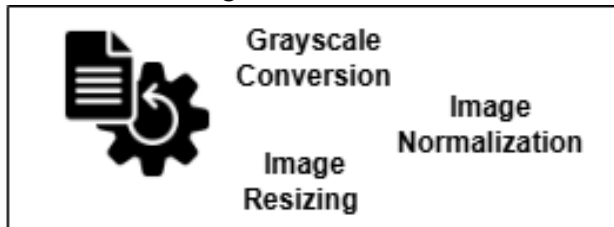
Table 2

Class Detail.

Class	Class Records	Total Records
Parasitized	13,779	27,558
Uninfected	13,779	

3.2 Data Preprocessing

The "Malaria Cell Images" dataset includes 27,558 microscopic blood smear images. Images in this collection were equally split between two classes, Parasitized and Uninfected, as indicated by Table 2. The unstructured nature of the raw data presented challenges to prepare it for machine learning model training because the images had different resolutions and color intensities. A significant amount of effort was invested in the pre-processing or cleaning of the images to enable an image with a standard resolution (to be determined later) and color intensity normalization so that all pixel values were scaled to a common range of [0,1] as shown in Fig 2. The pixel intensity normalization allowed for the standardization of pixel intensity variations and the acceleration of machine learning model convergence. To reduce the computational load on the machine learning model, the data for training was converted from RGB (Red, Green, Blue) format to grayscale while still retaining some of the critical features of discriminative texture associated with malaria parasite infection. Because of the nearly equal distribution of both Parasitized and Uninfected image types, there was no need to implement synthetic oversampling techniques for either class of cells; to improve model generalization and limit potential overfitting. Data augmentation techniques were applied to the raw data, such as random rotation, horizontal and vertical flips, zoom, and minor alterations in brightness, to enable the model to recognize invariant features of Parasitized and Uninfected cells regardless of orientation and scale.

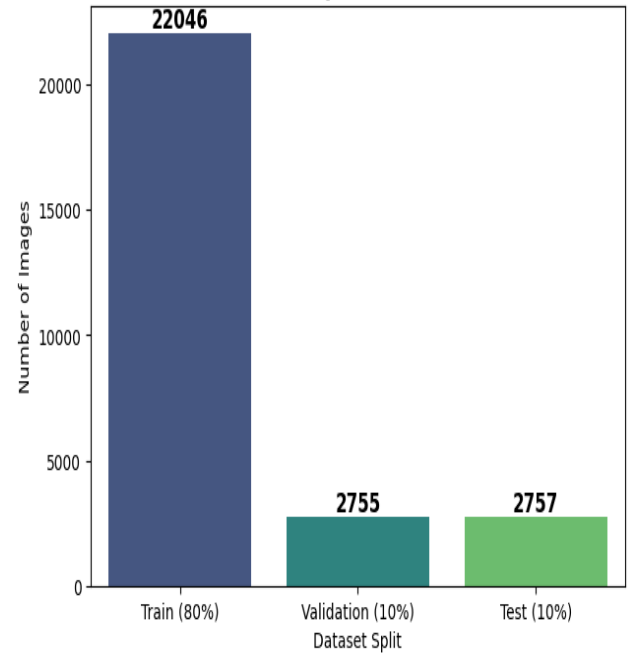


Preprocessing

Fig 2. Preprocessing.

The dataset was split into three different sections: 10% validation, 10% test, and 80% training (as shown in Figure 3). The training section included the extraction of another validation set to help with hyperparameter fine tuning. This design provided a way to evaluate the models on unseen sample data. To maintain the same level of quality throughout the preprocessing pipeline, © Mehran University of Engineering and Technology 2026

the training section was free from corrupted and/or unreadable images. The resulting dataset had been prepared in a clean, augmented, and balanced state for use in deep learning classification algorithms.

**Fig 3.** Dataset Split Distribution.

3.3 Deep Learning Models

For the accurate classification of malaria infected and Malaria free Red blood Cell images, this study makes use of newly developed Deep Learning architectures. These photos show significantly different physical characteristics between infected and uninfected. Therefore, multi-scale and hierarchical representation of these features are necessary, as well as the ability to acquire local pattern recognition from either a topographical or spatial standpoint. A Multi-scale hybrid Approach to classification was used, featuring the combination of the Convolutional Neural Network (CNN) with Swin Transformer, and XAI for maximising both interpretability and accuracy of classification.

3.3.1 Convolutional Neural Network (CNN)

The use of Convolutional Neural Networks (CNN) for image classification tasks is based on their ability to extract spatial features using the convolution and pooling operations. Therefore, in this research paper, it is accepted that the layers of CNNs will act as the initial stage of feature extraction which captures low level features that include the edges, textures, and cellular structures.

Mathematically the convolution operation can be defined as follows:

$$S(I, j) = (X * K)(i, j) = \sum_m \sum_n X(I + m, j + n) K(m, n)$$

Where:

- X represents the input image,
- For learnable filter use K,
- S (i, j) is the resulting feature map at position (i, j).

Following convolution, pooling layers reduce dimensionality while preserving salient features. ReLU introduces non-linearity, and fully connected layers map features to the binary output: parasitized vs. uninfected.

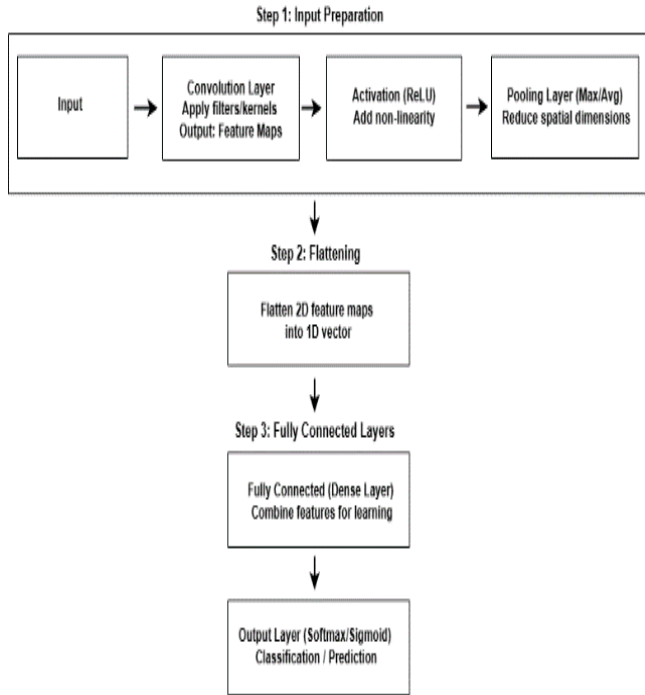


Fig 4. CNNs Workflow.

Flowchart of the CNN Architecture is depicted in figure 4

Step 1: Preparatory Input

- Feeding the model (input): Inputting raw/prepared data to the network. This may be an image, text representation, or signal.
- Convolutional Layer: The convolution layer extracts local features from the input through the convolution operation performed by filters/kernels that slide over the data. The result is a set of feature maps (e.g., edges, textures, patterns, etc.) that indicate where various features are located in the original input image.
- Activation Layer (ReLU): The activation layer adds a non-linear element to the output of each convolution layer to allow the CNN to learn complex patterns.
- Pooling Layer (Max or Average): The pooling layer reduces the spatial dimensions

of each feature map and retains only the most important features. Pooling reduces the number of computations required to perform the subsequent layers of the CNN.

Step 2: Flattening

- Flattening: The 2D outputs of all the pooling layers are flattened into a single 1D vector so that they can then be used as input for the fully connected layers. Flattening bridges the convolutional and pooling layers and the fully connected layers and is what prepares the information for classification.

Step 3: Fully Connected Layers

- The Flat Shapes/Dense Layer takes the one-dimensional vectors generated from the flattening process and understands the input as a whole by learning higher-level combinations of features that were detected previously by the convolutional and pooling layers.

Output Layer:

- Activation Function such as Sigmoid or Softmax is applied to determine what output is assigned to an element in the dense output layer.
- Produces the classification (final prediction) for the CNN.

3.3.2 Swin Transformer

Swin Transformer uses hierarchical feature maps and computes self-attention within non-overlapping windows, which are shifted across layers to enable cross-window connections. This design captures both local and global contextual relationships among cell structures. While CNNs excel at local feature extraction, they have limitations in modeling long-range dependencies. To overcome this, the Swin Transformer (Shifted Window Transformer) was employed.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (\text{Equation 2})$$

Where:

- Q, K, V are query, key, and value matrices,
- For key vector d_k is the dimensionality used.

By incorporating the process, the Swin Transformer captures fine-grained parasite details while maintaining computational efficiency through window-based attention.

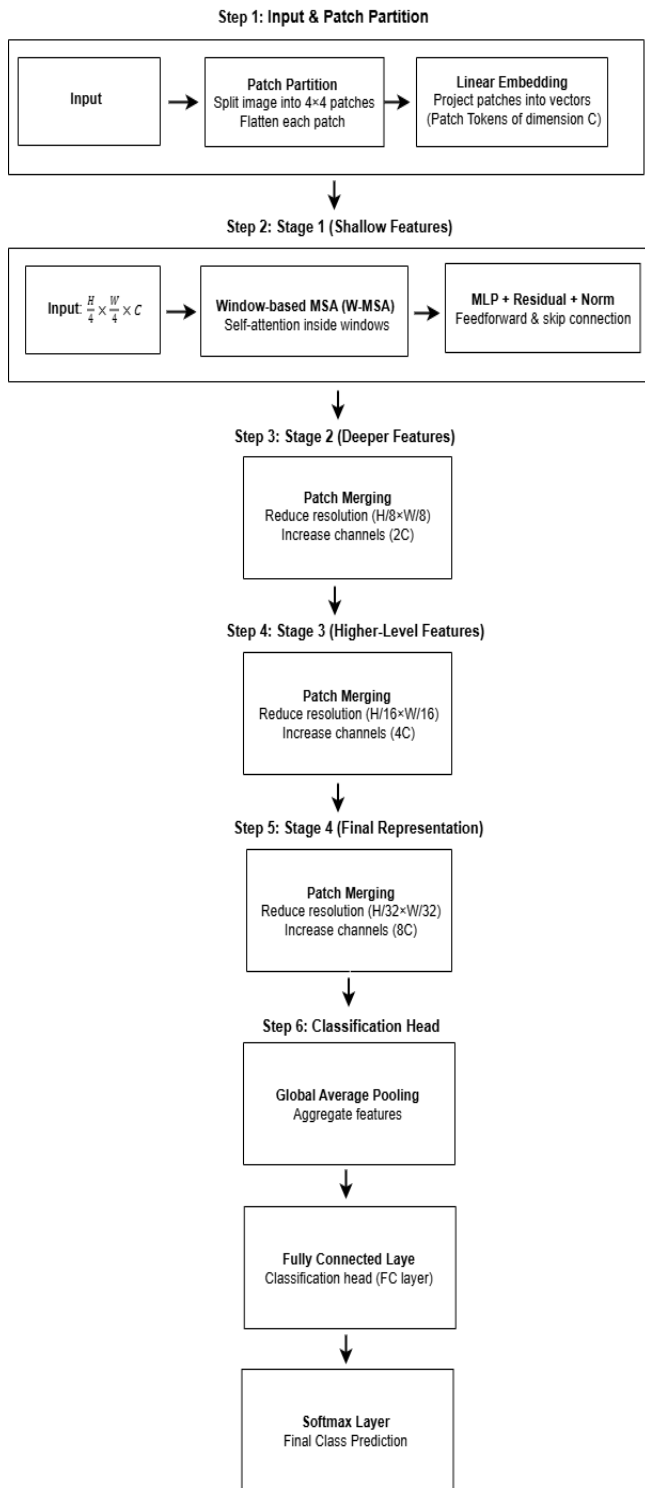


Fig 5. Swin Transformer Workflow.

Swin transformer architecture flow chart as shown in Figure 5:

Step 1: Input & Patch Partition

- The model starts with an input image (e.g., $224 \times 224 \times 3$).
- The image is split into small patches (e.g., 4×4).

- The initial patch tokens of dimension C are created by flattening each patch and projecting it into a vector using linear embedding.

This step converts the 2D image into a sequence of vectors that can be processed like words in NLP.

Step 2: Stage 1 (Shallow Features)

- Input blocks of first Swin Transformer are input tokens of size $H/4 \times W/4 \times C$.
- Window-based Multi-Head Self Attention (W-MSA): Self-attention on within-fixed windows, which makes it computationally efficient.
- Shifted Window MSA (SW-MSA) (implied here): Windows slide across blocks to permit cross-window connections, to have global context with time.
- MLP + Residual + Normalization: Adds non-linearity, stabilizes training, and retains skip connections. At this step, the network discovers the low-level features such as edges and corners, and even textures (as it is the case with the first CNN layers).

Step 3: Stage 2 (Deeper Features)

- Patch Merging helps to cut the resolution in half ($H/8W/8$) and channels ($2C$) in half. Currents of W-MSA and SW-MSA layers are used.
- This enables the model to absorb bigger shapes in the image including shapes or small objects.

Step 4: Stage 3 (Higher-Level Features)

- Another patch down sample combines further ($H/16XW/16$) and adds channels ($4C$).
- The self-attention layers proceed with bigger receptive fields. At this age, the network begins to internalize parts and shapes of objects, such as faces, wheels or leaves.

Stage 4 (Final Representation)

- is the last stage. The last patch combining is lower resolution ($H/32XW/32$) and more channels ($8C$).
- The lowest level, in which the model captures worldly, abstract characteristics of the entire picture. In this case, Swin Transformer is a holistic representation of the input.

Step 6: Classification Head

- Global Average Pooling The features of all tokens are combined into a single global representation. A Fully Connected Layer (FC) converts this representation to class logits.

Lastly, a Softmax Layer creates the probabilities of every class. This provides the ultimate forecast, e.g., the word cat or the word dog in image classification.

3.4 Explainable AI (XAI)

The Explainable AI (XAI) methods combined were used to provide transparency and trustworthiness in the model's predictions. The use of XAI methods (Gradient-weighted Class Activation Mapping (Grad-CAM)) showed which parts of the input images influenced the model's decision most directly, increasing the interpretability of the model and specifically highlighting parasite-infected areas within parasitized cells or clean area within uninfected cells. The localization map $Lc_Grad-CAM$ was created for the class ccc using the following formula:

$$Lc_{Grad-CAM} = ReLU(\sum_k \alpha_k^c A^k)$$

Where:

- A^k represents the k^{th} feature map,
- α_k^c is the weight obtained by global average pooling of gradients,
- ReLU ensures only positive influences are visualized.

The inclusion of XAI ensures that the diagnostic model is not a “black box” but provides human-interpretable insights, making it valuable for medical practitioners in decision support systems.

3.5 Evaluation Metrics

Validating effectiveness of proposed malaria detection framework, many performance metrics were adopted. Accuracy quantifies the overall proportion of correctly classified blood smear images. Precision measures how many of the predicted malaria-positive samples are actually parasitized, ensuring the reliability of positive predictions. Recall (Sensitivity) focuses on identification ability of model to actual malaria-infected cells, which is crucial in medical diagnosis to minimize false negatives. The F1-Score Averages Precision and Recall to produce an equally weighted metric into a single measure. Additionally, during training, the Loss function was also monitored to check model convergence and optimization. For interpretability purposes, Local Attribution-based

Explainable AI (LAAI) was also incorporated to check the degree through which the model places focus on notable visual regions responsible for making predictions and thus making the system more interpretable to the practicing physicians.

3.5.1 Accuracy: Calculate number of samples accurately grouped per total dataset, which is an indicator of overall correctness.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

3.5.2 Precision: Reports how many of the positive cases predicted are actually positive while emphasizing reduction of false positives.

$$Precision = \frac{TP}{TP + FP}$$

3.5.3 Recall (Sensitivity): Estimates the strength of the model to correctly classify actual positive cases while minimizing false negatives.

$$Recall = \frac{TP}{TP + FN}$$

3.5.4 F1-Score: The harmonic mean of precision and recall, treating both equally to get more accurate performance evaluation.

$$F1-Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

3.5.5 Loss: Computes the difference of predicted and actual values while training, guiding the model's optimization process.

$$Loss = -\frac{1}{N} \sum_{i=1}^N [y_i \log(y'_i) + (1 - y_i) \log(1 - y'_i)]$$

3.5.6 LAAI (Latency-Aware Accuracy Index):

Balances precision with computational delay, trading off both predictive capability and efficiency.

$$LAAI = \frac{Accuracy}{1 + Latency}$$

4. Results & discussion

In this study, we examined the malaria cell image dataset to assess the performance of a deep learning model. The model featured a hybrid structure composed of Convolutional Neural Networks (CNN) that extracted local features and a Swin Transformer that captured the global dependencies. Performance measures (precision, recall, F1 score, specificity, loss and accuracy) were used for the evaluation of model performance during testing. Learning curves and validation curves for the performance measures were obtained from eight cycles of training, and these

results provide a detailed description of the learning characteristics of the model as well as its ability to generalize based on the data.

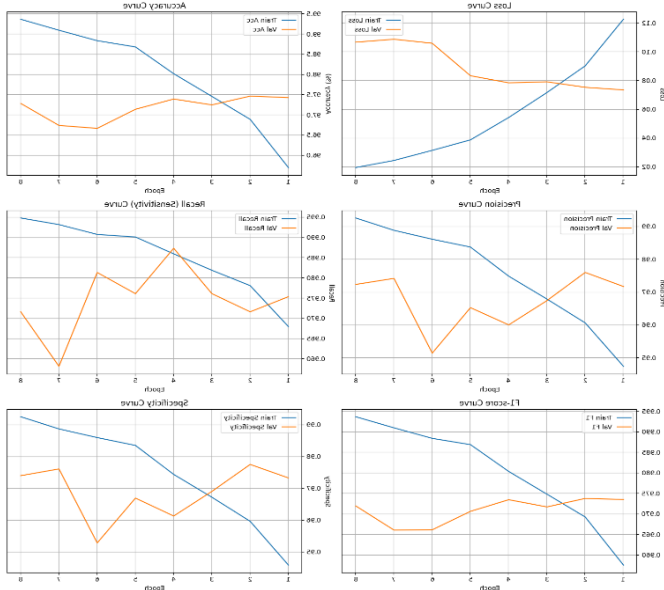


Fig 6. Training and Validation Metrics.

Figure 6 depicts six curves of learning and validation performances of the model. Loss curve indicates that loss gradually decreases during training and the loss curve achieves successful optimization whereas the validation loss level-offs in the fourth epoch indicating insignificant overfitting. Accuracy curve The accuracy steadily increases in training, more than 99 percent, whereas validation in the 96 per cent range and 97.5 per cent range all prove that there is high stability on unseen samples. The curves of precision and recall show that the model is very reliable in differentiating between malaria-infected cells and normal cells. The training precision is above 99 and the validation precision is approximately between 96 and 98. In line with this, recall is very sensitive with recall trainability of nearly 99.5% and the respective validation recall of nearly 96, guaranteeing minimum false negatives which is a very significant aspect of medical diagnosis. The F1-score curve strikes a balance between precision and recall with consistent values ranging between 97 percent and 98 percent in the validation, demonstrating the fact that the model is firm in the positive samples, and also in the negative samples. Lastly, the specificity curve demonstrates the capability of the model suitable classification of the uninfected cells with specificity in training over 99% and specificity in the validation over 96%. These results validate the integration of CNN with the Swin Transformer accelerates the capability of the model to meet high precision, high sensitivity, and high specificity all simultaneously. This makes the intended approach viable to the actual

real-world malaria detection problems, while keeping minimum false negatives coupled with reliability in classification has been an imperious dimension.

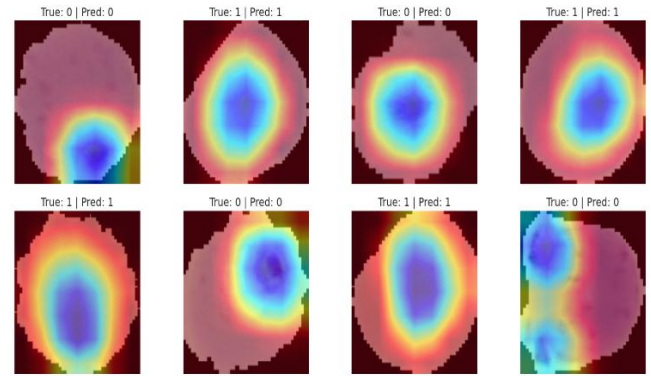


Fig 7. Grad-CAM Visualizations on Test Samples.

Figure 7 shows Grad-CAM visualizations of eight test images, each having the true label and predicted label from the model. The Grad-CAM visualizations give insight into the model's decision-making through an explanation of which parts of the input images were most responsible for the classification decision. The large majority of the predictions visualized prove correct, with six out of eight images showing agreement between true and predicted labels. In these correctly classified samples, the Grad-CAM heatmaps focus on the central regions of the cell images, particularly around the cellular structures where parasitic features are likely present. This suggests that the model successfully learned to attend to diagnostically relevant areas, supporting the reliability of its decisions.

Two inaccurate samples (True: 1, Pred: 0 and True: 1, Pred: 0) illustrate the model's omission to identify infected cells. Both exhibited Grad-CAM activations that weren't very focused, or shifted outward to regions outside of the cell body, suggesting that detection of subtle or ambiguous elements of infection may have been challenging for the model. Thus, these samples illustrate the issues that may arise when attempting to identify between borderline/visually general samples. Additionally, while detecting general characteristics for the entire sample set produces consistently large number of Grad-CAM heatmaps, it also suggests that the model utilized significant visual characteristic(s) present within the sample set. The importance of using significant visual characteristics in medical imaging cannot be underestimated, as having a solid understanding of what the model uses as basis for its predictions will support the development and/or implementation of clinical grade workflow. Additionally, the presence of focused Grad-CAM activations in samples classified correctly stated earlier (based on the performance metrics) confirms the model's interpretive properties in addition to the

performance results (i.e., accuracy, precision, etc.). Therefore, the results support the potential use of the model by malaria diagnosticians, as both trustworthiness and effectiveness are essential to have within a diagnostic laboratory.

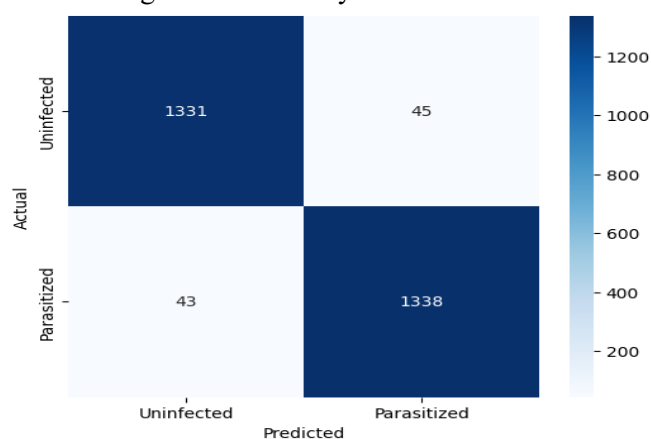


Fig 8. Confusion Matrix on Test Set.

In order for the predictions of the model to be verifiable and understandable, Explainable AI (XAI) Methods have also been incorporated. XAI Methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) were used to Illustrate the specific parts of the input images where the model's decision was most influenced by. The method has improved the interpretability of the model by highlighting the location of the parasite-infected regions of the parasitized cells or the location of the clean regions in the uninfected cells. In general, the method of Grad-CAM produces a localization map L_c Grad-CAM for class c such that: Additionally, very few misclassifications occurred: only 45 uninfected cells were incorrectly classified as being parasitized, while only 43 parasitized cells were incorrectly classified as uninfected. This data highlight many important characteristics of the model's overall performance. It showed high sensitivity with correct classification of 1338 out of 1381 parasitized samples, with 96.89% recall of infected cells. This highlights strong ability of the model to reduce false negatives, a factor which is very critical in medicine diagnostics because missing an infected case has serious consequences. Similarly, high specificity was also shown by the model with proper classification of 1331 out of 1376 non-infected samples, with 96.73% recall and with the confidence that healthy cases were scarcely predicted as infected cases. The balanced confusion matrix also indicates there is no bias of the model either to or away from either class with continued equal performance on parasitized and non-infected cases respectively. Additionally, the relatively small number of incorrect predictions 45 false positives and 43 false negatives demonstrates strong generalizability with low overfitting and emphasizes strong ability of the model to distinguish infected from healthy cells.

These results also affirm the earlier quantitative results from visualizations and from performance metrics, confirming the observation that the CNN-Swin Transformers-based model provides strong and accurate classification. High specificity, precision, and recall scores affirm its possible suitability for use in practical malaria detection software, wherein diagnostic reliability and accuracy are paramount.

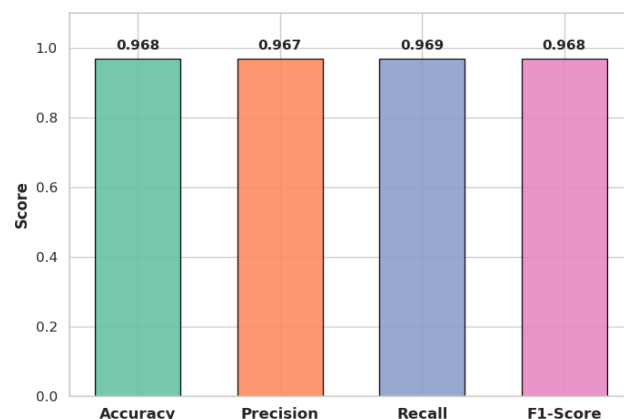


Fig 9. Test Set Performance Metrics.

Figure 9 shows the overall performance of the constructed CNN-Swin Transformer model on test set using four major performance metrics: accuracy, precision, recall, and F1-score. The model achieved an accuracy of 96.8%, presenting superior generalizability in differentiating infected and non-infected cells. With precision of 96.7%, it shows a low number of false positives, such that out of all the cells identified as malaria-infected, the overwhelming majority were parasitized, hence minimizing unwarranted treatments. The value of recall of 96.9% indicates the effectiveness of the model in parasitized cells detection, an area pivotal in keeping the number of false negatives to a minimum coupled with avoiding lost infections. Moreover, the perfect balance in terms of precision-recall is demonstrated by a very high F1-Score of 96.8%, which is reflective of how stable and reliable this model will be when applied in clinical situations. Furthermore, the “set” of performance indices, which fall within fluctuations of 96.7 to 96.9, confirm the model's stability and reliability when tested on new data. Together with results garnered from the confusion matrix and heatmaps of Grad-CAM, these results provide further evidence of the transferability of the developed methodology towards practical malaria detection tasks occurring outside of controlled laboratory environments. The integration of CNN and Swin Transformer settings has resulted in a model with superior precision but also interpretability and reliability both of immense value in medicine.

These performance measures, all clustered in the 96.7–96.9% range, confirm the model's stability and reliability on unseen data. Combined with data from

Table 3
CNN–Swin Transformer Model.

Model	Accuracy	Precision	Recall	F1-Score	Latency (s)
CNN+ Swin Transformer	96.8%	96.7%	96.9%	96.8%	0.21

the confusion matrix and Grad-CAM visualizations, this constitutes compelling evidence in favor of the suitability of the proposed method to realistic malaria detection problems. The combination of CNN and the Swin Transformer architectures has led to a model which, aside from operating with high precision, also preserves interpretability and reliability both necessary in the field of medical diagnostics.

5. Conclusion

This study suggested a hybrid deep learning model consisting of Convolutional Neural Networks (CNNs) and the Swin Transformer, augmented with Explainable AI (XAI) mechanisms, for automated malaria detection from images of cells. The model led to superior prediction performance on a number of criteria of evaluation, by accuracy of 96.8%, F1-score of 96.8%, precision of 96.7%, and recall of 96.9%. The results show the effectiveness of the combination of CNNs for localized extraction with the hierarchical self-attention of the Swin Transformer for extracting global contextual relationships an integration effective enough to capture the inherent heterogeneity and complexity of microscopic blood smear images. Importantly, aside from numerical performance, the incorporation of Grad-CAM visualizations enhanced the interpretability of the model through the regular identification of diagnostically significant regions of cells. This aspect augments clinical trustworthiness but also makes model implementation possible in clinical practice, where transparency in decision-making is as critical as diagnostic accuracy. Lastly, confusion matrices showed balanced classification of parasitized and uninfected cells with minimum rates of false negatives and false positives, stressing the robustness and reliability of the model. From a translational perspective, the proposed framework holds substantial potential in real-world clinical settings, particularly in resource-constrained environments where access to expert microscopists is limited. Its ability to deliver both accurate and interpretable results position it as a promising decision-support tool for early malaria detection, which can significantly improve treatment outcomes and reduce mortality. Nevertheless, this study also opens pathways for future exploration.

By broadening the application of our method for parasite developmental stage multi-class

classification, we would have the opportunity to obtain even more thorough diagnostic data. Additionally, validating our model using larger, heterogeneous, multi-institutional databases would improve both the generalizability and clinical reliability of our model. Using lightweight, portable, or point-of-care diagnostic tools has a significant role in allowing for the translation of our study into viable and scalable solutions for use in endemic areas. The proposed CNN-Swin Transformer model augments the current state of the art in AI-based malaria diagnostics while balancing accuracy, robustness, and interpretability. It provides a valuable contribution to the accumulation of evidence demonstrating the application of deep learning in medical image analysis. Additionally, it establishes a footing for the creation of clinically deployable, transparent, and equitable diagnostic technology to enhance global health.

Additional information and declarations

Author Contributions

- **Syed Noor Hussain Shah:** Conceptualized the research framework, designed and conducted experiments, performed literature review, formulated methodology, led data visualization, interpreted results, prepared figures/tables, drafted and revised the manuscript, and approved the final version.
- **Muhammad Shahan Ibad:** Assisted in experiments, data collection, and preprocessing; contributed to methodology refinement; prepared figures/tables; reviewed and revised the manuscript; and approved the final version.
- **Sumaira Johar:** Provided supervision, expert guidance, critical revisions, and final approval of the manuscript.
- **Omar Bin Samin:** Provided supervision, expert guidance, critical revisions, and final approval of the manuscript.
- **Afsheen Khalid:** Provided supervision, expert guidance, critical revisions, and final approval of the manuscript.
- **Sami Ur Rahman:** Assisted in experiments, data collection, and preprocessing; contributed to methodology refinement; prepared figures/tables; reviewed and revised the manuscript; and approved the final version.

Data Availability

The following information was supplied regarding data availability:

The “Malaria Cell Images” dataset is available at Kaggle: <https://www.kaggle.com/datasets/iarunava/cell-images-for-detecting-malaria>.

6. References

- [1] NobelPrize.org, “Alphonse Laveran – Facts,” 1907. [Online]. Available:

<https://www.nobelprize.org/prizes/medicine/1907/laveran/facts/>

- [2] Centers for Disease Control and Prevention, "The history of malaria," U.S. Department of Health & Human Services, n.d. [Online]. Available: <https://www.cdc.gov/malaria/history/index.html>
- [3] World Health Organization, "Malaria history," n.d. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/malaria>
- [4] The Lancet, "The Lancet," 2023.
- [5] World Health Organization, *World Malaria Report 2023*, 2023.
- [6] Common Management Unit Pakistan, "Malaria Technical Unit – Annual statistics 2023," Islamabad: Government of Pakistan, 2024. [Online]. Available: <https://www.cmu.gov.pk/dmc-directorate-of-malaria-control/>
- [7] Common Management Unit Pakistan, "Flood-related malaria report or situation update," Islamabad: Government of Pakistan, 2024. [Online]. Available: <https://www.cmu.gov.pk>
- [8] The Lancet Planetary Health, "Coverage of the 2022 Pakistan floods."
- [9] Y. Benachour, F. Flitti, and H. M. Khalid, "Enhancing Malaria Detection Through Deep Learning: A Comparative Study of Convolutional Neural Networks," *IEEE Access*, 2025.
- [10] M. Hamza *et al.*, "Leveraging Machine Learning and Deep Learning for Advanced Malaria Detection Through Blood Cell Images," 2025.
- [11] S. Kumar *et al.*, "Optimized deep CNN with rotation-driven features for malaria parasite detection," *Neural Computing and Applications*, pp. 1–20, 2025.
- [12] S. A. Kumar *et al.*, "Application of hybrid capsule network model for malaria parasite detection on microscopic blood smear images," *Multimedia Tools and Applications*, vol. 84, no. 7, pp. 3459–3485, 2025.
- [13] J. Wekalao and A. Rajakannu, "VLSI-integrated CMOS-compatible high-performance terahertz metasurface biosensor for dual-mode detection of cancer and malaria with machine learning optimization," *Plasmonics*, pp. 1–18, 2025.
- [14] P. Karuppasamy *et al.*, "Design and optimization of a hybrid graphene-copper terahertz metasurfaces biosensor for high-sensitivity malaria detection," *Plasmonics*, pp. 1–14, 2025.
- [15] S. S. Birunda *et al.*, "Machine Learning-Enhanced Terahertz Biosensor with Mxene-Graphene Conjugate for High-Sensitivity Malaria Detection," *Plasmonics*, pp. 1–14, 2025.
- [16] O. R. Shahin *et al.*, "Automated multi-model framework for malaria detection using deep learning and feature fusion," *Scientific Reports*, vol. 15, no. 1, p. 25672, 2025.
- [17] C. Kyeremah *et al.*, "Enhanced detection of malaria infected red blood cells through phase driven classification," *Scientific Reports*, vol. 15, no. 1, p. 30733, 2025.
- [18] D. A. Ramos-Briceño *et al.*, "Deep learning-based malaria parasite detection," *Scientific Reports*, vol. 15, no. 1, p. 3746, 2025.
- [19] M. F. Ahamed *et al.*, "Improving Malaria diagnosis through interpretable customized CNNs architectures," *Scientific Reports*, vol. 15, no. 1, p. 6484, 2025.
- [20] E. Ahishakiye *et al.*, "Enhancing malaria detection and classification using convolutional neural networks-vision transformer architecture," *Discover Applied Sciences*, vol. 7, no. 6, p. 612, 2025.
- [21] J. Sora-Cardenas *et al.*, "Image-Based Detection and Classification of Malaria Parasites and Leukocytes," *Sensors*, vol. 25, no. 2, p. 390, 2025.
- [22] M. Rejili *et al.*, "Malaria Parasite Detection Using Vision Transformer (ViT) and Deep Learning-Based Image Analysis," 2025.
- [23] S. B. Nettur *et al.*, "UltraLightSqueezeNet: A Deep Learning Architecture for Malaria Classification," *IEEE Access*, 2025.