Journal of Nanotechnology and Innovative Materials

Research Article

Explainable Vision Transformer-Based Automated Malaria Detection from Blood Smear Images: Toward Intelligent Nanotechnology-Enabled Diagnostic Systems

Sabina M. Nobari[1], Sabina R. Azizova[1], and Shokrollah A. Ghadyani[2]

[1]Department of Information Technologies, School of Advanced Technologies and Innovation Engineering, Western Caspian University, Ahmad Rajabli 17A, III Parallel, AZ1001 Baku, Azerbaijan
[2]Tekdata Co., No. 12, Marzbannameh Alley, Mofatteh Street, Tehran, Iran

Received: 23.06.2026 Accepted: 28.06.2026 Published: 31.07.2026

https://doi.org/10.54414/LCGJ5320

Abstract

Malaria remains one of the world's most significant parasitic diseases, requiring rapid and accurate diagnosis to reduce mortality. Recent advances in digital microscopy, artificial intelligence, and nanotechnology-enabled diagnostic platforms have created new opportunities for automated malaria detection. In particular, nano-enabled biosensors and lab-on-chip technologies increasingly rely on intelligent image analysis algorithms to improve diagnostic performance. In this study, convolutional neural network architectures (ResNet50, EfficientNetB0, InceptionResNetV2, and Xception) and Vision Transformer models (ViT-Base, Swin Transformer, DeiT, and Pyramid Vision Transformer) were evaluated for multiclass malaria detection using microscopic blood smear images. A five-fold cross-validation strategy was employed, and model interpretability was enhanced using Gradient-weighted Class Activation Mapping (Grad-CAM). Among the CNN models, Xception achieved an accuracy of 98.10%, while the Swin Transformer demonstrated comparable performance among transformer-based architectures. The results indicate that combining explainable artificial intelligence with transformer-based models improves both classification accuracy and model transparency, supporting their future integration into nano-enabled diagnostic systems, intelligent digital microscopy, and point-of-care biomedical devices.

Keywords: malaria diagnosis, deep learning, vision transformer, explainable artificial intelligence, nanotechnology, nanobiosensors, medical image analysis, digital microscopy

1. Introduction

Malaria remains one of the most significant parasitic infectious diseases worldwide and continues to pose a major public health challenge, particularly in tropical and subtropical regions. Millions of people get this disease every year. According to the World Health Organization (WHO, 2024), in 2023 there were 263 million malaria cases worldwide and about 597 thousand deaths worldwide. The main cause of the disease is various types of Plasmodium parasites (most commonly P. falciparum and P. vivax). Malaria is transmitted to humans through the bite of infected female *Anopheles* mosquitoes [1]. Effective treatment of malaria in full and on time and the diagnosis of the infection are of absolute importance. Since the beginning of the traditional microscopic diagnostic method of the 20th century, the "gold standard" has been established; this method is open to allowing morphological changes in Plasmodium infection as a result of Erythrocytes (red blood cells-RBC). In this method, stained blood is checked under a microscope; then, whichever parasite number in erythrocytes is considered high accuracy. This method is a process that is labor-intensive and operator-dependent process, although it requires a significant amount of time. Sources say that human errors in clinics can affect the results, and serious effects can be demonstrated [1], [2].

Molecular methods such as polymerase chain reaction (PCR) and rapid diagnostic tests (RDT) also improve, but these methods require high-level laboratory equipment and are expensive, so both areas need

large-scale and appropriate-level implementation. Therefore, increases in systems that require rapid results, automated, accurate, and cost-effective malaria detection [3], [4]. In recent years, Computer Vision (CV) and Deep Learning (DL) have significantly advanced the automated analysis of microscopic blood smear images. Various deep learning architectures, including Convolutional Neural Networks (CNNs), Faster R-CNN, YOLO, and Mask R-CNN, have been successfully applied to automated malaria detection [5]. These models automatically extract discriminative image features from microscopic blood smear images, enabling accurate classification of malaria parasites. Many CNN-based research systems have reported that malaria detection is between 95% and 99% accurate [6], [7]. However, CNN models have limitations in capturing long-range dependencies and understanding global contextual relationships between distant pixels and image regions. These limitations restrict their ability to effectively model complex spatial relationships within images.

To overcome these limitations, a new generation of models inspired by Transformer architectures has been introduced. Originally developed for natural language processing (NLP), these models have demonstrated remarkable performance in computer vision tasks. These models show an image of various parts between long-distance dependencies and global attention, the open relationships [8].

The promising results of the images in the analysis, including the architecture of Vision Transformer (ViT), Data Efficient Image Transformer (DeiT), Pyramid Vision Transformer (PVT), and Swin Transformer (malaria parasites in the testpit) [9], [10].

One of the major limitations of AI in medical diagnosis is the lack of interpretability (or explainability). Clinical specialists need not only a complete, but also a visualized explanation of the decision-making process. To this end, Gradient is mainly used in Class Activation Mapping (Grad-CAM) methods, but also helps decide on the most impactful areas highlighted by the model. This is both an increase in the trust of doctors in AI systems, as well as facilitating error analysis and improvement of algorithms [11].

Moreover, in recent years, nanotechnology has significantly contributed to the development of rapid malaria diagnostic systems through nanoparticle-based biosensors, plasmonic sensing platforms, microfluidic lab-on-chip devices, and portable nano-enabled imaging technologies [12]. These emerging diagnostic tools generate high-resolution microscopic data that require intelligent computational algorithms for accurate interpretation. Consequently, artificial intelligence and deep learning have become essential complementary technologies for improving the diagnostic performance of nanotechnology-based medical devices.

1.1. Research Gap

In the last decade, a large number of CNN-based research conducted for the discovery of automated malaria detection is also directed, the majority of these studies have focused only on binary classification of infected and uninfected cells, and the examination of multiple analytic time patterns that were revealed to be far deeper than analytic methods that could be construed.

CNN models perform very well in situations where global additions and ink tissue are fully exposed to learning challenges. This leads to a reduction in accuracy, especially when the images are noisy or there are lighting differences. This factor is also limited in the number yet applied by the researchers. The development of explainable Artificial Intelligence (Explainable AI, XAI) approaches, by making the decision-making processes of artificial intelligence models more transparent and understandable, increases the trust of medical professionals in these systems and makes a significant contribution to the effective use of artificial intelligence in clinical applications. The objective of this study is to develop and evaluate an explainable artificial intelligence framework for automated malaria detection by comparing CNN- and Vision Transformer-based architectures on microscopic blood smear images. The article describes classic CNN architectures, advanced Transformer-based approaches, and techniques such as Grad-CAM to identify future development aspects for significant malaria diagnostics through artificial intelligence research. Although recent studies have demonstrated the effectiveness of CNN and Vision Transformer architectures for malaria detection, comparative evaluations incorporating explainable artificial intelligence remain limited.

Furthermore, few studies have discussed how these AI models can complement emerging nanotechnology-based diagnostic platforms, including nano-enabled biosensors and lab-on-chip systems. Therefore, this study evaluates multiple deep learning architectures while emphasizing their potential integration into future intelligent nanodiagnostic devices.

1.2. Background

In recent years, significant improvements in Computer Vision (CV) and Deep Learning (DL) have changed the landscape by using microscopic blood descriptions to automatically diagnose malaria. Both open-source information databases (e.g., NIH Malaria thin smear dataset) and a large number of studies, both local to the given use, were able to pinpoint the CNN-based model Plasmodium hybrid with the Convolutional Neural Networks (CNN) and infected red blood cells.

For example, the first research for classic CNN architectures and transfer learning methods achieved high accuracy in separating 95% more than infected and healthy cells by more than 95% with the ResNet50 model [13]. At the same time [14], accuracy was achieved with 97.57% detection using the EfficientNet k-fold cross-validation method. Subsequently, CNN models combine learning methods, achieving more than 98% accuracy in the resulting high-performance datasets [15], [16], [17].

Despite these successes, CNN models are still restricted in a row. They often face differences in staining due to characteristics and image information; devices sensitive to data imbalance apply to these real clinical models. Performance decreases when models are applied to new datasets (domain/field shift), and researchers have proposed two main approaches to address this issue.

1. Resources are limited, which health services around them use for light, quantity, and mobility to suit model preparation devices.
2. Image pixels between global dependencies-based architectures application for Model Transformer.

Medical visualization Vision Transformer (ViT) and its/derivatives Data-efficient Image Transformer (DeiT), Pyramid Vision Transformer (PVT), and Swin Transformer, pay close attention to the center. These models employ self-attention mechanisms to capture long-range dependencies and global contextual information, enabling more effective image feature extraction than conventional convolutional neural networks. Extensive studies have shown that Vision Transformer models achieve strong results on tasks such as classification, segmentation, and cell identification in medical imaging [18], [19]. Recent studies have further demonstrated the effectiveness of Vision Transformer architectures for 2D biomedical image classification, confirming their applicability to medical imaging tasks [20]. From this supplement, the attention mechanism more prominently combines parasite structures in the early stages of infection with CNN [7]. At the same time, multi-stage malaria detection was prepared for the hybrid CNN–ViT model, and 99% accuracy was achieved. This result is due to combining the performance of CNN and ViT, which is important to the extent that CNNs and ViT's ability to remove local properties with a global attention mechanism [21]. In addition to the other benefits of accuracy, intelligence (AI) is the main application of artificial medicine, explaining and clinical reliability. Despite the high diagnostic accuracy of deep learning models, their decision-making process is often considered a "black box," limiting their acceptance in clinical practice. Healthcare professionals require not only accurate predictions but also interpretable explanations of how these predictions are generated. To address this challenge, Explainable Artificial Intelligence (XAI) techniques, particularly Gradient-weighted Class Activation Mapping (Grad-CAM), have been widely adopted to visualize the image regions that contribute most to a model's predictions. Improved variants, such as Grad-CAM++ and DropKey Grad-CAM, further enhance localization accuracy and visualization quality. Recent studies have demonstrated that integrating Grad-CAM with both Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) architectures significantly improves model interpretability and diagnostic transparency, enabling clinicians to better understand, validate, and trust AI-assisted medical diagnoses [22], [23].

1.3. Role of Nanotechnology in Malaria Diagnosis

Nanotechnology has emerged as a promising approach for improving malaria diagnosis by enabling rapid, sensitive, and portable diagnostic platforms [24]. Gold nanoparticles, magnetic nanoparticles, quantum dots, and nanostructured biosensors have been extensively investigated for detecting Plasmodium antigens, DNA, and hemozoin biomarkers. In addition, lab-on-chip and microfluidic systems integrated with nanoscale materials facilitate point-of-care diagnosis in resource-limited settings. Although these technologies significantly improve sample preparation and signal detection, automated interpretation of microscopic and biosensor-generated images remains challenging. Therefore, explainable deep learning models, such as Vision Transformers combined with Grad-CAM, provide valuable computational support for next-generation nanotechnology-enabled malaria diagnostic systems.

2. Materials and Methods

2.1. Dataset

In this study, a publicly available malaria dataset obtained from the literature [25] was used to classify different *Plasmodium* species. The dataset consists of microscopic images of peripheral blood smear samples acquired using a light microscope. The dataset consists of four different types of malaria parasites: P. falciparum, P. malariae, P. ovale, and P. vivax, and a total of 210 images. The images are 2598x1944 pixels and have a TIFF extension. Images in the dataset are shown in Figure 1.

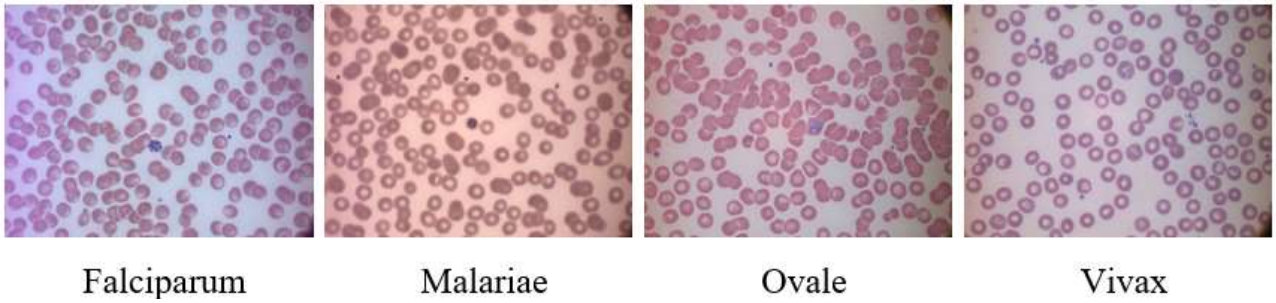

Figure 1. Sample images in the dataset.

2.2. Convolutional Neural Network

Transfer learning has been widely adopted in medical image analysis because it enables convolutional neural networks to achieve high classification accuracy even when only limited annotated datasets are available [26]. This strategy utilizes four distinct convolutional neural network architectures, ResNet50, EfficientNetB0, InceptionResNetV2, and Xception, to enhance the model's generalization capability and classification accuracy on limited medical datasets. By reusing learned feature representations, transfer learning enables models pre-trained on large-scale datasets like ImageNet to effectively adapt to domain-specific tasks, thereby reducing training time and the risk of overfitting.

ResNet50 provides redundant shortcut connections to overcome gradient problems lost in deep networks, and ensures stable training of very deep architectures [27]. EfficientNetB0 uses a compound scaling method that evenly balances network depth, width, and resolution to achieve high performance with optimal computing efficiency [28]. InceptionResNetV2 now combines Inception modules with links, offering both multi-scale feature extraction and improved convergence stability [29]. Meanwhile, Xception extends the Inception concept using deep detachable folds that improve computational efficiency and can disentangle channels [30]. Recent studies have also demonstrated that modern convolutional architectures, such as ConvNeXt, achieve competitive performance while narrowing the gap between CNNs and Vision Transformers in image classification tasks [31].

2.3. Vision Transformer

Natural language processing, created primarily from the Transformer architecture, inspired ViT, revolutionized computer vision. The evolution of Vision Transformer (ViT) from the use of layers; instead, the introduction breaks the description into small pieces (patches), each part converts to a vector, and these vectors make the global and local dependencies by means of transformer encoder modeling processing.

2.4. Self-Attention Mechanism

One of the fundamental differences between Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) is the use of the self-attention mechanism instead of convolutional filters for feature extraction. Self-attention enables ViTs to capture both local and long-range dependencies by modeling relationships among image patches across the entire image. Consequently, ViTs can effectively learn global contextual information, overcoming the limited receptive field of conventional CNNs. This capability has led to superior performance in many computer vision tasks, particularly when large training datasets are available. However, some studies have reported that CNNs may still outperform ViTs in capturing subtle local features when training data are limited or fine-grained image details are critical [32].

Attention within the ViT architecture is calculated as in Equation 1:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

In Equation 1, Q, K, and V are the Query, Key, and Value matrices. Here, K is the key vector size; the Softmax function sets a weight for each part of the image and shows how much this weight is compared to the other parts.

Vision Transformers are particularly effective for medical image analysis because they capture both local image features and long-range spatial relationships through self-attention mechanisms. This capability enables accurate identification of subtle morphological changes associated with malaria-infected erythrocytes. The general diagram of the vision transformer model is given in Figure 2.

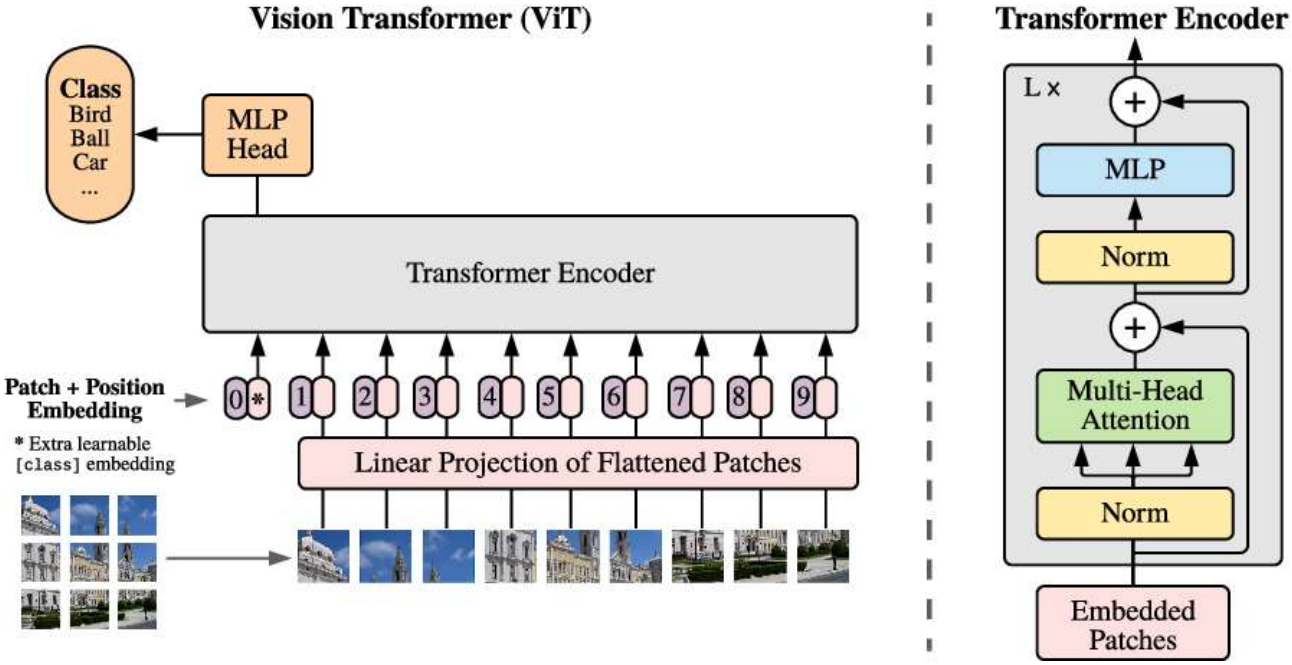

Figure 2. The base vision transformer model [32].

2.5. Base Vision Transformer (BaseViT)

ViT architecture is one of the standard applications for image representations intended to learn images from large-volume datasets. The encoder of the ViT-Base model consists of 12 Transformer layers, each containing 12 self-attention heads, with an embedding dimension of 768. Unlike convolutional neural networks (CNNs), ViT does not rely on convolutional operations for feature extraction; instead, it employs a Transformer-based architecture to capture both local and global dependencies among image regions. The overall structure of the ViT-Base architecture can be described as follows: the input image is first divided into a sequence of fixed-size patches. Each image patch is linearly projected into an embedding vector, and learnable positional embeddings are added to preserve spatial information. These patch embeddings are then fed into the Transformer encoder. Each encoder block consists of a multi-head self-attention (MSA) module and a multilayer perceptron (MLP) block, with layer normalization and residual connections applied to improve training stability and information flow [1], [32]. These components combine to model your image, both local and global connections to learn the presence of malaria, in their definition, infected cells, observing color and texture in terms of subtle differences. The observation made by the BaseViT model can distinguish these cells from healthy cells in a healthy, sensitive way.

2.6. Data-Efficient Image Transformer (DeiT)

The DeiT model, proposed by Touvron and others, is designed for presentation and limited to data vision transformer models taught for an innovative approach offered [33]. The main distinguishing feature of DeiT

is the introduction of a distillation token, which is a learnable token designed to facilitate knowledge transfer between a teacher model and a student model. This token is appended to the sequence of patch embeddings along with the class token and is used to transfer information from the teacher network to the student network during training. Unlike conventional Vision Transformers that require large-scale datasets, DeiT employs a knowledge distillation framework that enables the student model to effectively learn from a teacher model, reducing the dependency on extensive training data [34]. Through this approach, DeiT achieves competitive performance while requiring fewer computational resources and smaller datasets. In addition, DeiT incorporates advanced data augmentation strategies during training, which improve model generalization and enhance its ability to extract robust visual representations under different conditions.

2.7. Pyramid Vision Transformer (PVT)

PVT is a hierarchical transformer architecture for multi-dimensional properties to be driven (e.g., object detection and semantic segmentation) predictive tasks. This model has as many stages as the input description size, collecting a variety of rich and high-quality features collector pyramid type hierarchical feature map form. If this structure shows similarities to multi-level Convolutional Neural Networks (CNNs), convolution operations were built without it [35]. Instead, Transformer's focus attention mechanism and various explanations between resolutions place their dependencies on the model for use. The PVT model brings intensive forecasting tasks [36] for both efficient and powerful backbones in image segmentation, such as object detection and tight prediction. The overall diagram of the PVT model is given in Figure 3.

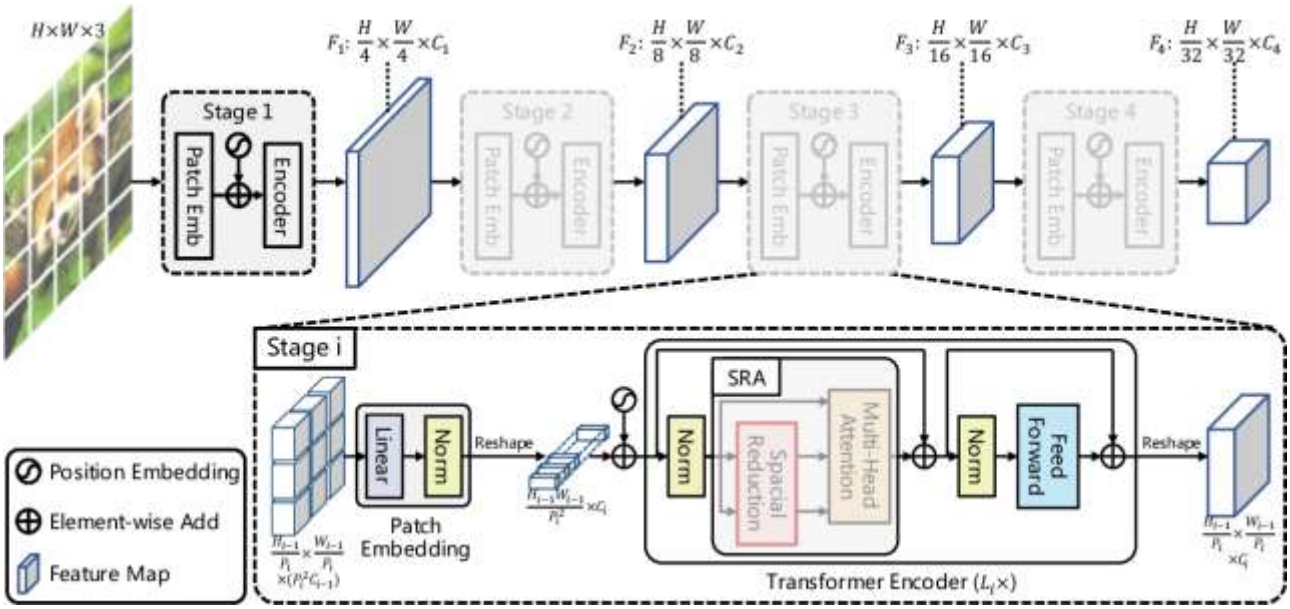

Figure 3. The general diagram of the PVT model [36].

2.8. Swin Transformer

The Swin Transformer model enables a vision transformer architecture to whom it offers, by making sliding windows (shifted windows), balancing efficiency with model complexity, the concept presentation [37]. In this architecture, self-care is not only small and includes other windows located on top of each other, this also reduces significantly, which is calculated to be a computational complexity. The number of neighbouring windows between the exchange of information increases. This mechanism involves neighboring window border markers on windows; they need to be mutually touched in order to create conditions and ultimately provide cross-window communication [38].

2.9. Gradient-Weight Class Activation Mapping (Grad-CAM)

Deep learning models are mainly affected by problems, including a lack of transparency in the process. The Grad-CAM (Gradient-weighted Class Activation Mapping) method is a powerful tool for understanding how to model their decisions. This method emphasizes evolution and layers of attention with related gradients, mapping the model to speech, highlighting the most impact depicting their region. Grad-CAM can also be used

for model debugging and reliability evaluation (trust calibration) to be effective in the way of use. This can help researchers and doctors model the predictions behind what is logically better finished. The Grad-CAM heat map is a weighted combination of feature maps followed by a ReLU. Its mathematical formula is given in Equation 2 [39].

$$L_{ij}^{Grad-CAM} = ReLU\left(\sum_{k} \alpha_k A_{ij}^{k}\right) \qquad (2)$$

Weight Calculation of coefficients (α_k) formula is like:

$$\alpha_k = \frac{1}{Z}\sum_{i}\sum_{j}\frac{\partial y^c}{\partial A_{ij}^{k}} \qquad (3)$$

3. Results and Discussion

3.1. Performance Metrics

Metrics derived from the confusion matrix were used to determine the performance of deep learning models in the detection of malaria-infected cells. These are Accuracy, Sensitivity/Remembering, Specificity, Precise indicator (Precision, and F1 score (F1-Score). These metrics include four possible predicted models to the essentially calculated result: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [40]. The formulas (balances 4-8) of the metrics derived from the Confusion matrix and their descriptions are given below.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \qquad (4)$$

Here:
TP (True Positive): Correct some positive samples made number (accurate, as open infected cells)
Data TN (True Negative): The number of correctly classified specific negative samples (true), made of clear healthy cells)
DBFP (False Positive): Negative examples with positive error, such as classified cases
DB FN (False Negative): Positive examples with negative error, such as classified cases
Sensitivity (Repeat): Model positive samples (infected cells) to make accurate capability measures precise:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \qquad (5)$$

Specificity: Model negative samples (healthy cells) correctly certain to do ability measuring is an indicator:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \qquad (6)$$

Precision: correct predicted positive examples general positive predictions to the number which ratio shows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (7)$$

F1 score (F1-Score) Precision and Sensitivity (Recall) harmonic middle like certain is being and this two indicator between balance creates:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \qquad (8)$$

Five layers of Cross Verification method was used to perform experimental studies with neural networks. It is a model validation method in which the dataset is divided into five equal parts, one part used for testing and the remaining four for training, and this process is repeated five times to calculate the average performance.

3.2. Experimental Test Results

This study employs eight different deep learning models to detect malaria using peripheral blood images. The performance of four CNN (Convolutional Neural Network) models and four ViT (Vision Transformer) models was evaluated using 5-fold cross-validation. First, the dataset was analyzed using CNN models; the architectures employed included ResNet50, EfficientNet-B0, InceptionResNetV2, and Xception. To achieve optimal results, the training parameters were set as follows: 16 epochs, a mini-batch size of 16, and a learning rate of 1×10^{-4}; these settings were chosen to maximize overall model performance. Experimental results indicate that Xception achieved the best performance among the CNN models, reaching an accuracy of 98.10%. The performance results (confusion matrices) for the CNN models are presented in Table 1 and Figure 4, while the training/validation processes and loss curves are illustrated in Figure 5.

Table 1. Performance results of CNN models.

CNN Models	Accuracy	Sensitivity	Specificity	Precision	F1-Scor
ResNet50	86.19	79.48	95.32	93.18	89.47
EfficientNetb0	94.76	94.79	98.04	94.77	94.48
InceptionResNetV2	96.19	95.22	98.39	97.26	96.07
Xception	98.10	98.02	99.23	98.11	98.01

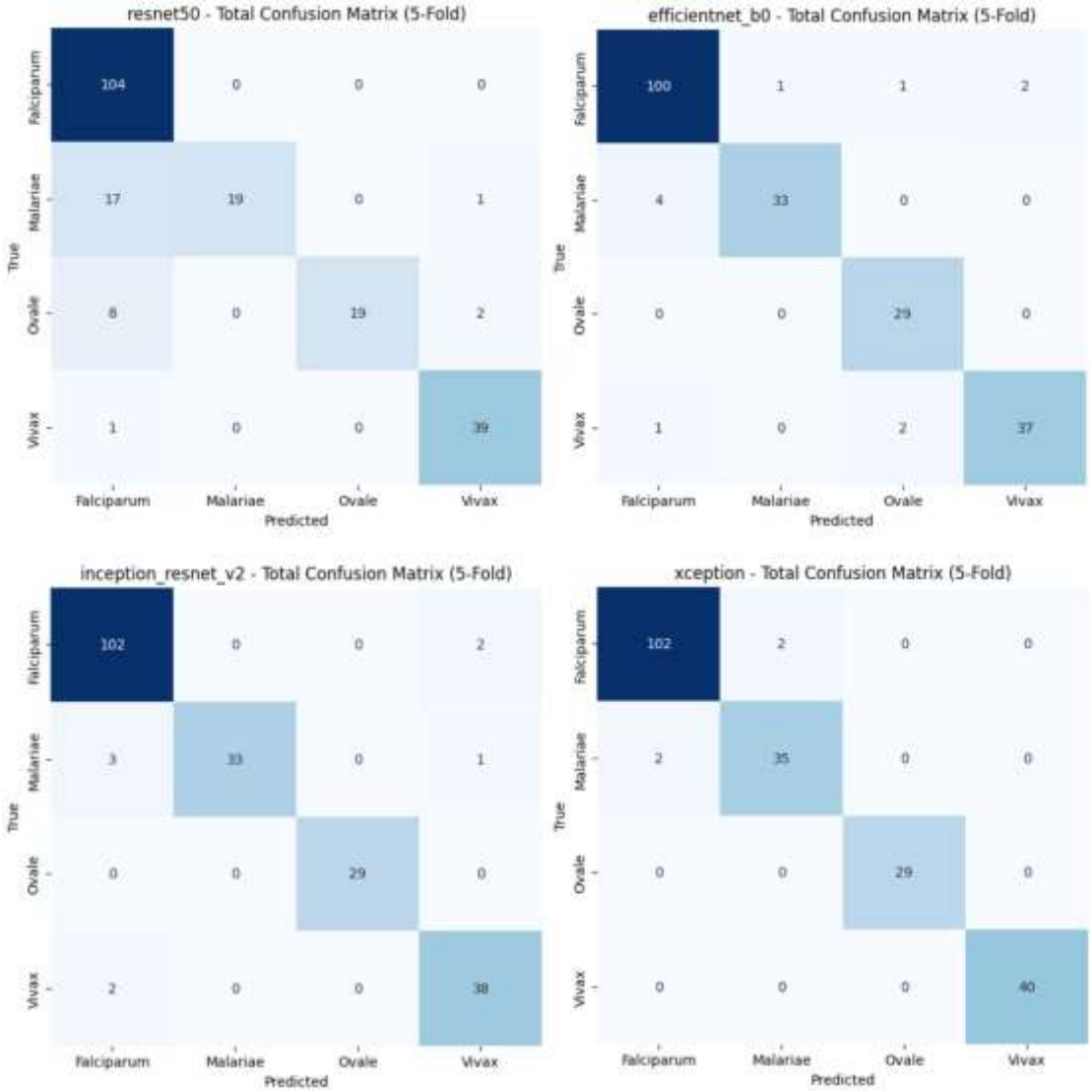

Figure 4. Confusion matrix of CNN algorithms.

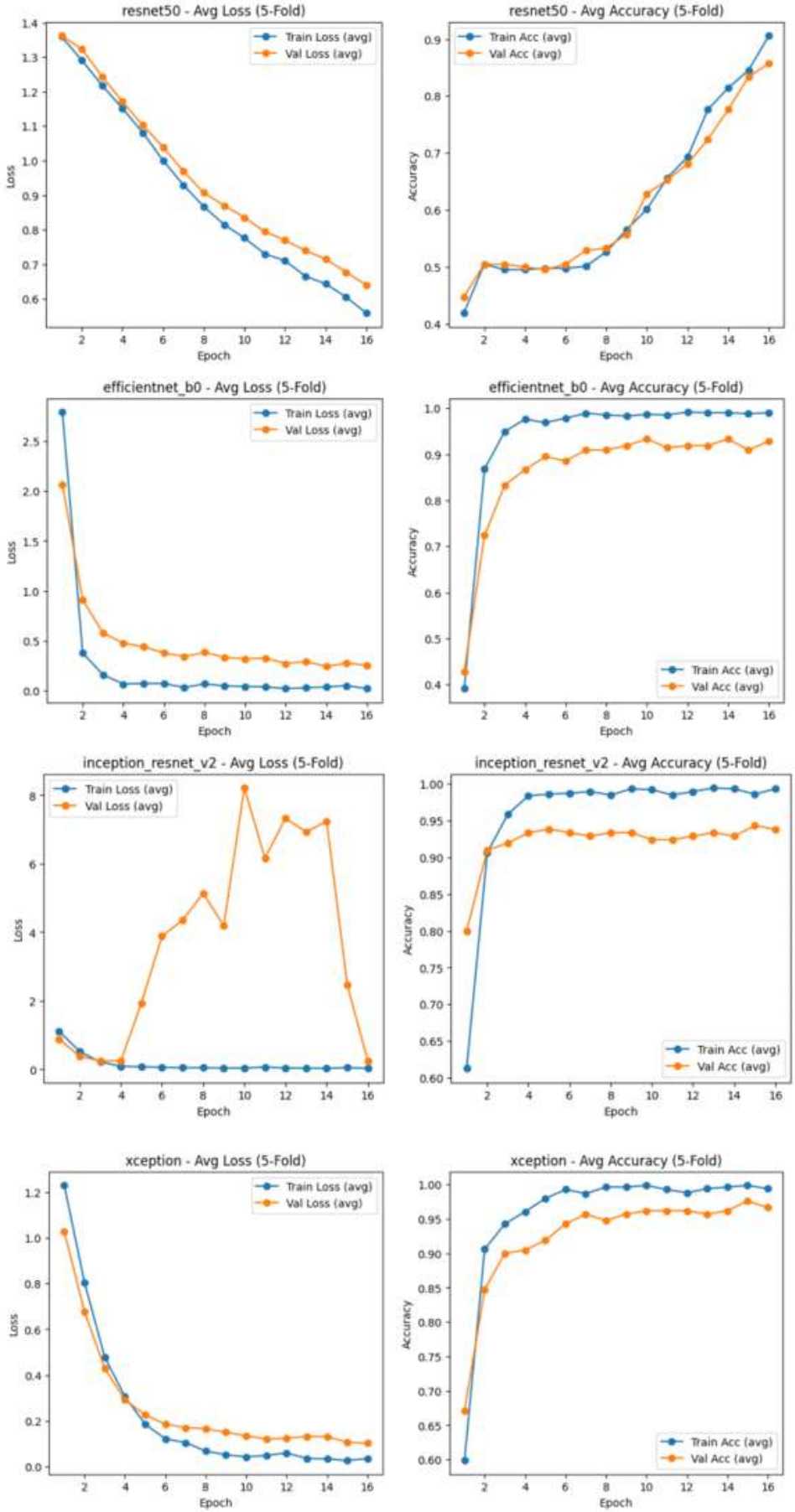

Figure 5. Training and validation accuracy and loss curves.

In the second phase of the experimental work, the dataset was tested using the ViT_base model, Swin Transformer, DeiT, and Pyramid Vision Transformer algorithms. The 5-layer cross-validation method was chosen as the training and test partitioning method for the working dataset. The epoch, mini-batch size, and learning rate for the four ViT models were set to 16, 16, and 1×10^{-4}, respectively, to match those of the CNN models. The optimizer was set to the Adam algorithm, and the loss function was set to cross-entropy loss for hyperparameters. The experimental work resulted in the best performance values with the Swim transformer algorithm and achieved an accuracy rate of 98.10%. Performance results obtained with ViT algorithms are given in Table 2. Confusion matrices are given in Figure 6, and the training of algorithms and the loss graphs are given in Figure 7.

Table 2. Performance results of ViT algorithms.

ViT Models	Accuracy	Sensitivity	Specificity	Precision	F1-Scor
ViT base	96.72	97.31	99.09	97.56	97.42
Swim	98.10	97.73	99.15	98.58	98.13
Deit	96.19	94.89	98.30	97.00	95.86
Pyramid Vit	95.71	93.98	98.18	96.42	94.85

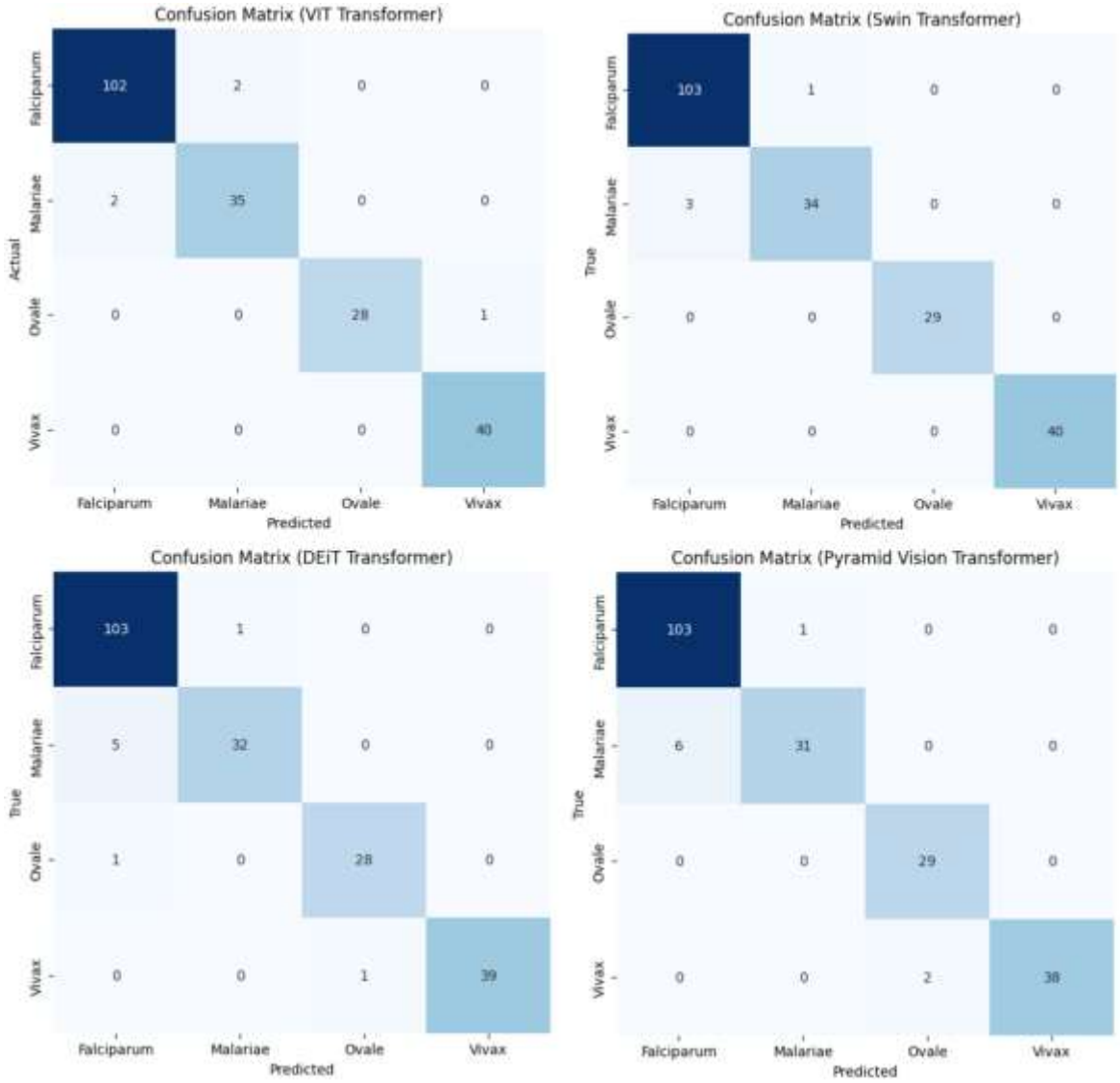

Figure 6. Confusion matrix of ViT algorithms.

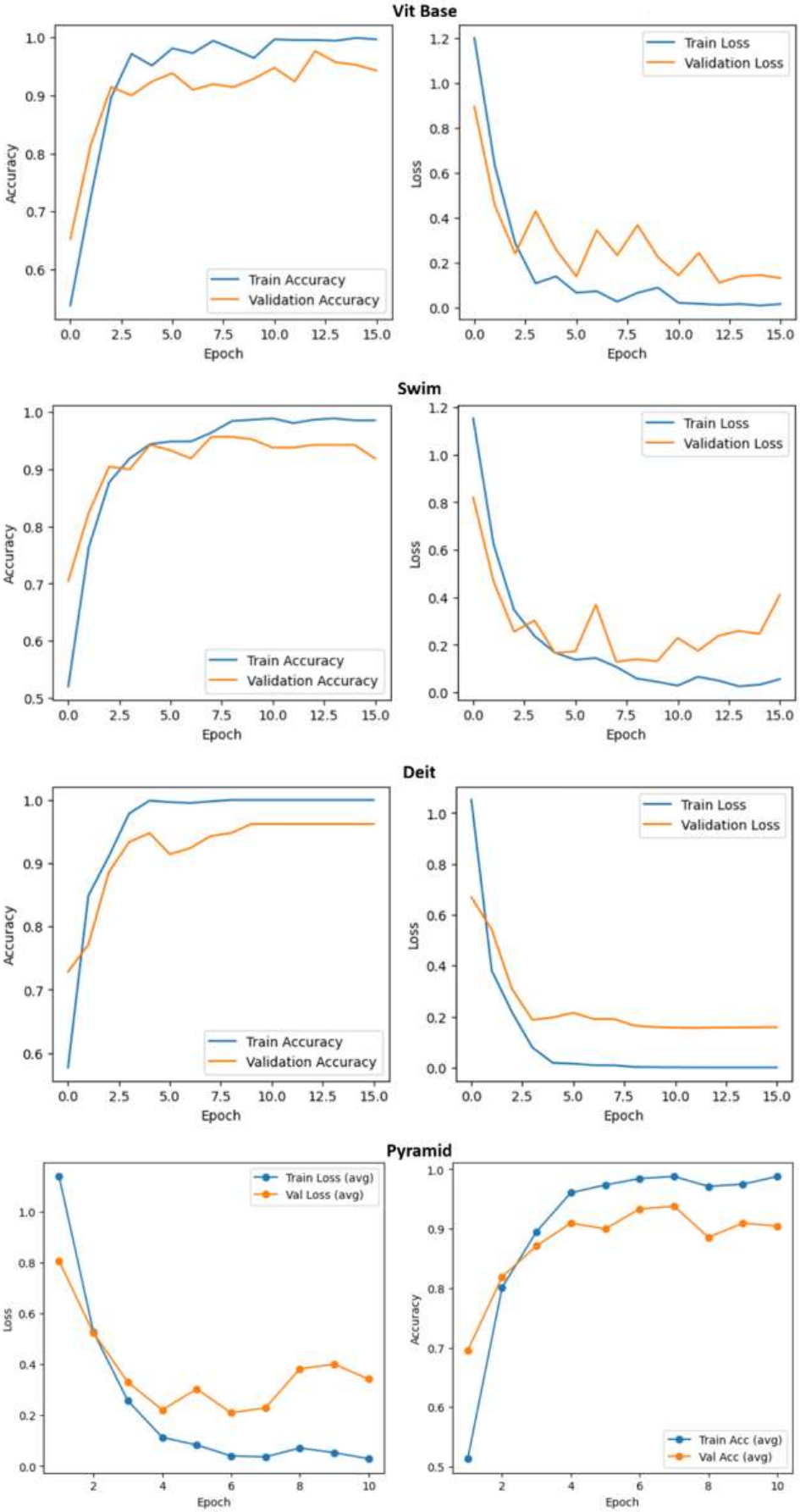

Figure 7. Training and validation loss graph of ViT algorithms.

At the last stage, images were produced using the GRAD-CAM algorithm, an explainable artificial intelligence method that gives the best performance results with the Swin Transformer algorithm. Models can be interpreted with GRAD-CAM images, allowing clinicians to interpret images more easily. Random images generated by the GRAD-CAM algorithm are shown in Figure 8. When examining the images, red lesion areas represent the region that is particularly effective in the decision-making processes of the models. Here, it can be seen that the cells are red in color.

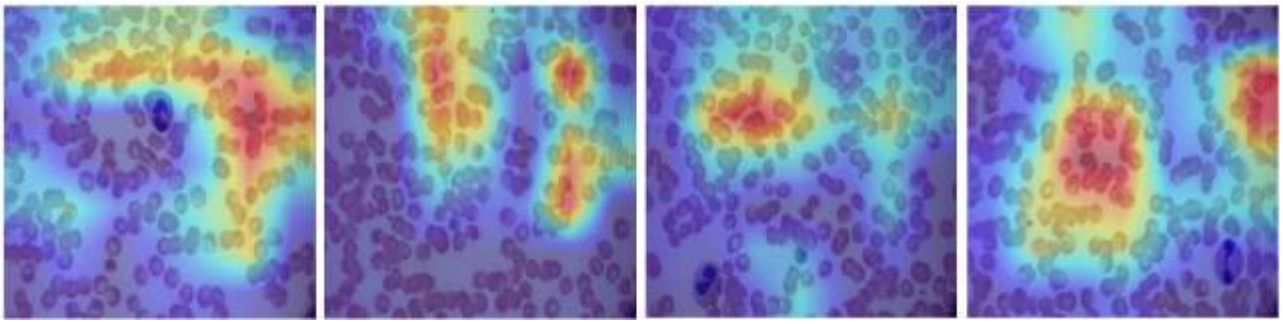

Figure 8. GRAD-CAM images produced with the Swin Transformer.

3.3. Implications for Nanotechnology

Although this study focused on microscopic blood smear image analysis, the proposed explainable AI framework has considerable potential for integration with emerging nanotechnology-enabled diagnostic platforms. Modern malaria diagnostic technologies increasingly employ nanoparticle-based biosensors, microfluidic devices, and portable lab-on-chip systems that generate digital imaging data requiring automated interpretation. Vision Transformer architectures combined with explainable artificial intelligence can provide reliable decision support for these systems by improving classification accuracy while simultaneously increasing clinical interpretability. Therefore, integrating advanced AI algorithms with nanotechnology-based diagnostic platforms represents a promising direction for future malaria diagnostics.

4. Conclusion

Malaria remains a serious health problem, especially for communities living in developing countries, and is one of the most important parasitic infections that threaten public health globally. Diagnostic methods using microscopes, still considered the gold standard nowadays, have certain limitations due to the need for specialist staff, the time required, and the associated labor intensity. This study is aimed at improving the accuracy of malaria detection and automating the process. For this purpose, up-to-date deep learning techniques such as convolutional neural network architectures (ResNet50, EfficientNetB0, InceptionResNetV2, and Xception) and Vision Transformer-based approaches (ViT-Base, Swin Transformer, DeiT, and PVT) were used. This study employed 5-fold cross-validation to train and evaluate models using a dataset of microscopic blood images covering four different types of malaria. Among the convolutional neural network models, Xception achieved the best performance with an accuracy of 98.10%, while the Swin Transformer demonstrated equally excellent performance among the Transformer-based models. Furthermore, the study utilized Gradient-weighted Class Activation Mapping (Grad-CAM) to elucidate the models' decision-making mechanisms, thereby enhancing their reliability for clinical application. This technique enabled researchers to visually demonstrate the specific image regions the models focused on when making diagnostic decisions. The research results reveal that using Transformer-based deep learning architectures in conjunction with explainable artificial intelligence methods significantly improves both classification success and model transparency. These developments enable automated malaria diagnostic systems to be used more reliably in clinical settings and integrated into real-life applications.

At the same time, nanotechnology represents an innovative approach with significant potential for improving the diagnosis and treatment of malaria. Nanomaterial-based delivery systems can enhance the targeted transport of antimalarial drugs to infected cells, improve drug bioavailability, reduce systemic side effects, and increase therapeutic efficacy. In addition, nanotechnology-based platforms contribute to the development of early diagnostic tools and next-generation vaccine strategies. Although many of these technologies remain at the preclinical or clinical research stages, they hold promise for future applications in

malaria prevention and treatment. In particular, targeted drug delivery systems based on liposomes, polymeric nanoparticles, and lipid-based nanocarriers enable more efficient delivery of antimalarial agents to infected hepatocytes and erythrocytes. By improving drug accumulation at infection sites while minimizing toxicity to healthy tissues, these nanocarrier systems represent a promising strategy for enhancing malaria therapy.

Future research should investigate the integration of explainable Vision Transformer architectures with nanotechnology-enabled diagnostic systems, including nano-biosensors, plasmonic sensing platforms, microfluidic lab-on-chip devices, and portable digital microscopy. Such multidisciplinary approaches have the potential to improve point-of-care malaria diagnosis by increasing diagnostic accuracy, interpretability, and clinical reliability while facilitating the deployment of intelligent nanodiagnostic technologies in resource-limited settings.

Author Contributions

All authors contributed to the conception and development of the study, participated in the preparation and revision of the manuscript, and approved the final version of the manuscript.

Conflict of Interest

The authors declare that they have no conflicts of interest related to this work.

Funding

This research received no external funding.

Acknowledgment

The authors would like to thank their institutions and colleagues for their support and assistance during the preparation of this study.

Abbreviations

Natural Language Processing (NLP), Gradient-weighted Class Activation Mapping (Grad-CAM), World Health Organization (WHO), Red Blood Cells (RBC), Polymerase Chain Reaction (PCR), Rapid Diagnostic Tests (RDT), Computer Vision (CV), Deep Learning (DL), Convolutional Neural Networks (CNNs), Vision Transformer (ViT), Data Efficient Image Transformer (DeiT), Pyramid Vision Transformer (PVT), Base Vision Transformer (BaseViT), Multi-Head Self-Attention (MSA), Multilayer Perceptron (MLP), True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

References

[1] World Health Organization. World malaria report 2024. Geneva: World Health Organization; 2024. [cited 2025 July 08]. Available from: https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2024

[2] Yang F, Poostchi M, Yu H, Zhou Z, Silamut K, Yu J, Maude RJ, Jaeger S, Antani S. Deep learning for smartphone-based malaria parasite detection in thick blood smears. IEEE journal of biomedical and health informatics. 2019 Sep 23;24(5):1427-38. https://doi.org/10.1109/jbhi.2019.2939121

[3] Fong Amaris WM, Martinez C, Cortés-Cortés LJ, Suárez DR. Image features for quality analysis of thick blood smears employed in malaria diagnosis. Malaria journal. 2022 Mar 5;21(1):74. https://doi.org/10.1186/s12936-022-04064-2

[4] Sukumarran D, Hasikin K, Khairuddin AS, Ngui R, Sulaiman WY, Vythilingam I, Divis PC. An optimised YOLOv4 deep learning model for efficient malarial cell detection in thin blood smear images. Parasites & Vectors. 2024 Apr 16;17(1):188. https://doi.org/10.1186/s13071-024-06215-7

[5] Başaran E, Cömert Z, Çelik Y, Velappan S, Toğaçar M. Determination of tympanic membrane region in the middle ear otoscope images with convolutional neural network based YOLO method. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi. 2020 Sep 22;22(66):919-28. https://doi.org/10.21205/deufmd.2020226625

[6] Maturana CR, De Oliveira AD, Nadal S, Bilalli B, Serrat FZ, Soley ME, Igual ES, Bosch M, Lluch AV, Abelló A, López-Codina D. Advances and challenges in automated malaria diagnosis using digital microscopy imaging with artificial intelligence tools: A review. Frontiers in Microbiology. 2022 Nov 15;13:1006659. https://doi.org/10.3389/fmicb.2022.1006659

[7] Goceri E. Deep learning-based classification of facial dermatological disorders. Computers in Biology and Medicine. 2021 Jan 1;128:104118. https://doi.org/10.1016/j.compbiomed.2020.104118

[8] Ahamed MF, Nahiduzzaman M, Mahmud G, Shafi FB, Ayari MA, Khandakar A, Abdullah-Al-Wadud M, Islam SR. Improving Malaria diagnosis through interpretable customized CNNs architectures. Scientific Reports. 2025 Feb 22;15(1):6484. https://doi.org/10.1038/s41598-025-90851-1

[9] Ahishakiye E, Kanobe F, Taremwa D, Nantongo BA, Nkalubo L, Ahimbisibwe S. Enhancing malaria detection and classification using convolutional neural networks-vision transformer architecture. Discover Applied Sciences. 2025 Jun 6;7(6):612. https://doi.org/10.1007/s42452-025-06704-z

[10] Tan D, Liang X. Multiclass malaria parasite recognition based on transformer models and a generative adversarial network. Scientific Reports. 2023 Oct 10;13(1):17136. https://doi.org/10.1038/s41598-023-44297-y

[11] Islam MR, Nahiduzzaman M, Goni MO, Sayeed A, Anower MS, Ahsan M, Haider J. Explainable transformer-based deep learning model for the detection of malaria parasites from blood cell images. Sensors. 2022 Jun 8;22(12):4358. https://doi.org/10.3390/s22124358

[12] Baptista V, Peng WK, Minas G, Veiga MI, Catarino SO. Review of microdevices for hemozoin-based malaria detection. Biosensors. 2022 Feb 11;12(2):110

[13] Rajaraman S, Antani SK, Poostchi M, Silamut K, Hossain MA, Maude RJ, Jaeger S, Thoma GR. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. PeerJ. 2018 Apr 16;6:e4568. https://doi.org/10.7717/peerj.4568

[14] Mujahid M, Rustam F, Shafique R, Montero EC, Alvarado ES, de la Torre Diez I, Ashraf I. Efficient deep learning-based approach for malaria detection using red blood cell smears. Scientific Reports. 2024 Jun 10;14(1):13249. https://doi.org/10.1038/s41598-024-63831-0

[15] Çalışkan A. Diagnosis of malaria disease by integrating chi-square feature selection algorithm with convolutional neural networks and autoencoder network. Transactions of the Institute of Measurement and Control. 2023 Mar;45(5):975-85. https://doi.org/10.1177/01423312221147335

[16] Kaboré KK, Guel D, Somda FH. Advancements in Deep Learning for Malaria Detection: A Comprehensive Overview. International Journal of Advanced Computer Science & Applications. 2025 Jun 1;16(6):57. https://doi.org/10.14569/ijacsa.2025.0160608

[17] Ramos-Briceño DA, Flammia-D'Aleo A, Fernández-López G, Carrión-Nessi FS, Forero-Peña DA. Deep learning-based malaria parasite detection: convolutional neural networks model for accurate species identification of Plasmodium falciparum and Plasmodium vivax. Scientific Reports. 2025 Jan 30;15(1):3746. https://doi.org/10.1038/s41598-025-87979-5

[18] Azad R, Kazerouni A, Heidari M, Aghdam EK, Molaei A, Jia Y, Jose A, Roy R, Merhof D. Advances in medical image analysis with vision transformers: a comprehensive review. Medical image analysis. 2024 Jan 1;91:103000. https://doi.org/10.1016/j.media.2023.103000

[19] Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H. Transformers in medical imaging: A survey. Medical Image Analysis. 2023 Aug 1;88:102802. https://doi.org/10.1016/j.media.2023.102802

[20] Halder A, Gharami S, Sadhu P, Singh PK, Woźniak M, Ijaz MF. Implementing vision transformer for classifying 2D biomedical images. Scientific Reports. 2024 May 31;14(1):12567. https://doi.org/10.1038/s41598-024-63094-9

[21] Surendar P, Basha CH, Nabi MS, Kavitha L, Vijayakumar P. Multi-modal deep learning for malaria diagnosis: integrating CNNs and vision transformers for enhanced parasite detection. In2025 Third International Conference on Augmented Intelligence and Sustainable Systems (ICAISS) 2025 May 21 (pp. 570-578). IEEE. https://doi.org/10.1109/icaiss61471.2025.11042087

[22] Awe OO, Mwangi PN, Goudoungou SK, Esho RV, Oyejide OS. Explainable AI for enhanced accuracy in malaria diagnosis using ensemble machine learning models. BMC Medical Informatics and Decision Making. 2025 Apr 11;25(1):162. https://doi.org/10.1186/s12911-025-02874-3
[23] Gao Y, Liu J, Li W, Hou M, Li Y, Zhao H. Augmented grad-cam++: Super-resolution saliency maps for visual interpretation of deep neural network. Electronics. 2023 Nov 30;12(23):4846. https://doi.org/10.3390/electronics12234846
[24] Obeagu EI, Okoroiwu GI, Ubosi NI, Obeagu GU, Onohuean H, Muhammad T, Adias TC. Revolution in malaria detection: unveiling current breakthroughs and tomorrow's possibilities in biomarker innovation. Annals of Medicine and Surgery. 2024 Jul 17;86(10):5859
[25] Loddo A, Di Ruberto C, Kocher M, Prod'Hom G. MP-IDB: the malaria parasite image database for image processing and analysis. InSipaim–Miccai Biomedical Workshop 2018 Sep 20 (pp. 57-65). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-13835-6_7
[26] Başaran E, Cömert Z, Çelik Y. Convolutional neural network approach for automatic tympanic membrane detection and classification. Biomedical Signal Processing and Control. 2020 Feb 1;56:101734. https://doi.org/10.1016/j.bspc.2019.101734
[27] Karadeniz AT, Çelik Y, Başaran E. Classification of walnut varieties obtained from walnut leaf images by the recommended residual block-based CNN model. European Food Research and Technology. 2023 Mar;249(3):727-38. https://doi.org/10.1007/s00217-022-04168-8
[28] Mahesh TR, Gupta M, Anupama TA, Geman O. An XAI-enhanced efficientNetB0 framework for precision brain tumor detection in MRI imaging. Journal of Neuroscience Methods. 2024 Oct 1;410:110227. https://doi.org/10.1016/j.jneumeth.2024.110227
[29] Peng C, Liu Y, Yuan X, Chen Q. Research of image recognition method based on enhanced inception-ResNet-V2. Multimedia Tools and Applications. 2022 Oct;81(24):34345-65. https://doi.org/10.1007/s11042-022-12387-0
[30] Uzun MZ, Başaran E, Çelik Y. Xception Derin Öğrenme Modeli ve Gabor Filtreleri ile ÇDÖÖE-DVM Algoritması Kullanılarak Mikro İfadelerin Tanınması. Journal of the Institute of Science and Technology. 2023 Jan 12;13(4):2339-52. https://doi.org/10.21597/jist.1252556
[31] Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. In2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR) 2022 Jun 18 (pp. 11966-11976). IEEE. https://doi.org/10.1109/cvpr52688.2022.01167
[32] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020 Oct 22. https://doi.org/10.48550/arXiv.2010.11929
[33] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning (ICML 2021). Proceedings of Machine Learning Research. Vol. 139. PMLR; 2021. p. 10347-57. [cited 2025 March 07]. Available from: https://proceedings.mlr.press/v139/touvron21a
[34] Toğaçar M. Biotransnet: Detection of plant stress through the conversion of biosensor data into RGB channels and combination with transformer networks. Journal of Crop Health. 2025 Oct;77(5):167. https://doi.org/10.1007/s10343-025-01239-0
[35] Zhou Y, Jiang X, Xu G, Yang X, Liu X, Li Z. PVT-SAR: An arbitrarily oriented SAR ship detector with pyramid vision transformer. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2022 Nov 14;16:291-305. https://doi.org/10.1109/jstars.2022.3221784
[36] Wang W, Xie E, Li X, Fan DP, Song K, Liang D, Lu T, Luo P, Shao L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In2021 IEEE/CVF International Conference on Computer Vision (ICCV) 2021 Oct 10 (pp. 548-558). IEEE. https://doi.org/10.1109/iccv48922.2021.00061
[37] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. In2021 IEEE/CVF International Conference on Computer Vision (ICCV) 2021 Oct 10 (pp. 9992-10002). Ieee. https://doi.org/10.1109/iccv48922.2021.00986
[38] Wei C, Ren S, Guo K, Hu H, Liang J. High-resolution Swin transformer for automatic medical image segmentation. Sensors. 2023 Mar 24;23(7):3420. https://doi.org/10.3390/s23073420

[39] Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: Why did you say that?. arxiv preprint arxiv:1611.07450. 2016 Nov 22. https://doi.org/10.48550/arXiv.1611.07450

[40] Başaran E, Çelik G, Toğaçar M. Regionally focused neural-coder model designed for the diagnosis of acute lymphoblastic leukemia disease. Measurement. 2025 Dec 1;256:118176. https://doi.org/10.1016/j.measurement.2025.118176