Proceedings of IAM, V.14, N.2, 2025, pp.175-194
https://doi.org/10.30546/2225-0530.14.2.2025.2001

AN EXPLAINABLE AI-DRIVEN APPROACH FOR AUTOMATED MALARIA DETECTION FROM BLOOD SMEAR IMAGES

E. Basaran[1] , M.V. Aliev[1] , S.M. Nobari[2] , A. O. Akdemir [1]

[1] Department of Computer Engineering, Agrı ˙Ibrahim Cecen University, Turkiye
[2] Western Caspian University, Baku , Azerbaijan
erdalbasaran085@gmail.com , aocakakdemir@gmail.com, sabinamirzaei@gmail.com

Abstract Malaria remains one of the world's most critical parasitic diseases, posing a serious threat to global public health, particularly in developing regions. Traditional microscopic diagnostic methods, considered the gold standard, are labor-intensive and time-consuming, as well as heavily dependent on human expertise. This study aims to improve the accuracy and automation level of malaria detection by using advanced deep learning models such as convolutional neural networks (ResNet50, EfficientNetB0, InceptionResNetV2, and Xception) and Vision Transformer-based architectures (ViT-Base, Swin Transformer, DeiT, and PVT). The models were trained and tested on a microscopic blood image dataset containing four different types of malaria using the 5-fold cross-validation method. Among the CNN models, Xception achieved the highest accuracy rate of 98.10%, while Swin Transformer showed similar success among the Transformer-based models. Furthermore, the Gradient-Weighted Class Activation Mapping (Grad-CAM) technique was applied to visually highlight the most influential regions in the model's decision-making process and to increase model interpretability. The findings show that combining transformer-based architectures with explainable artificial intelligence significantly improves both classification performance and model transparency. This enhances the clinical reliability of automated malaria diagnosis systems, strengthening their potential for use in real-world applications

Keywords: Malaria, Deep learning, Vision transformer, XAI, Automated diagnostics.

AMS Subject Classification: 26D15, 26A51.

1. Introduction

Malaria the most important parasitic from diseases one was , global public health for big threat is considered . Especially development in doing which in countries millions every year human this to the disease is infected. Worldwide Healthcare According to the World Health Organization (WHO, 2024), according to, in 2023 263 million worldwide very malaria condition and approximately 597 thousand deaths event registered The disease was main reason various *Plasmodium* parasite are the types (the most very *P. falciparum* and *P. vivax)* these are infected *Anopheles* of the type which female mosquitoes bite with to the person passes [37].

Malaria exactly and on time diagnosis of the disease effective treatment and of infection in front of you purchase for decisive importance carries. Traditional microscopic diagnostics method of the 20th century from the beginning since " gold " standard " account " is done because this method *Plasmodium* infection as a result in erythrocytes (red blood cells (RBC) giving morphological changes obvious to do opportunity gives . In this method painted blood smears microscope under being checked and in erythrocytes which parasites number considered high to accuracy although this method very time leading , labor capacious and operator dependent one is a process . Resources limited which in clinics human mistakes diagnostic to the results serious impact to show can [37, 38].

Molecular methods for example , polymerase chain reaction (PCR) and fast diagnostic tests (RDT) also development But this methods high level laboratory equipment demand does and it is expensive, therefore also large-scale and either field at the level application for suitable They are not . That's why also , fast result giving , automated , accurate and profitable malaria detection to systems need increases [11,28].

In recent years Computer Vision (CV) and Deep In the fields of learning (DL) head giving fast developments microscopic pictures based on automated malaria detection systems to the creation opportunity In this direction Convolution Nerve Networks (CNN) and advanced architectures such as Faster R-CNN, YOLO, and Mask R-CNN are complex visual examples in recognition high results [5, 6]. These models picture features automatically removes and infected cells high precisely certain does. Many CNN-based research systems malaria detection 95% to 99 % accuracy where reported [21,13]. However, CNN models pixel at the level global addictions and in the pictures contextual-spatial their relationships understanding to do ability These restrictions are limited. removed to lift for natural language in processing use Transformer-based from architectures inspired new generation models offer. These models your picture various parts between long-distance addictions and global attention their relationships obvious can knows [1]. Vision Transformer (ViT), Data-efficient Image Transformer (DeiT), Pyramid Vision Transformer (PVT) and Like a Swin Transformer architectures , including malaria parasites in the detection , medical of images in the analysis promising results [2,30].

Accuracy from increasing additional, medical artificial in intelligence (AI) other The main problem is the lack of understanding. Clinical specialists only exactly result giving not , but also decision-making process visualized explanation who to systems need For this purpose, Gradient-weighted Class Activation Mapping (Grad-CAM) is used methods was prepared, which is also the model decided the most very impact who picture areas emphasizes. This is both doctors' AI systems trust increases , both also mistake analysis and algorithms improvement facilitates [15].

1.1. Research Gap

Last decade carried out numerous CNN-based research automated malaria to the discovery if directed also , this of works majority only infected and uninfected of cells binary attention to binary classification has reached and models explanation to be done knowledge interpretability and more deep analytical aspects very time review kidnapped .
CNN models perform very well in situations where global dependencies and ink texture are fully exposed to learning difficulties. This leads to decreased accuracy, especially when images are noisy or have lighting differences. This factor is also limited in number yet applied by researchers. The development of Explainable Artificial Intelligence (Explainable AI, XAI) approaches increases the trust of medical professionals in these systems by making the decision-making processes of artificial intelligence models more transparent and understandable, and makes a significant contribution to the effective use of artificial intelligence in clinical practices.

Therefore, the primary objective of this research is to provide a comprehensive review and analysis of CNN (ResNet50, Efficientb0, InceptionresnetV2, Xception) and Image Transformer-based models (Vit_base, Swim, Deit, PVT) for automatic malaria detection from microscopic images. The article describes classical CNN architectures, advanced Transformer-based approaches, and techniques such as Grad-CAM, with the aim of identifying future development directions for important malaria diagnostics through artificial intelligence exploration. Such a review will drive research into technological innovations and applications, leading to results that will increase the effectiveness of malaria control and contribute to global health services.

1.2. Background

In recent years Computer Computer Vision (CV) and Deep In the fields of Deep Learning (DL) obtained done important improvements microscopic blood from the descriptions use by doing malaria automatic in diagnostics revolutionary to changes reason It was. Both open source information databases (e.g. , NIH Malaria thin smear dataset), both also local from the given use who numerous research showed that Convolution Nerve Networks (CNN) and hybrid CNN- based models *Plasmodium* with infected red blood cells exactly certain can knows [8].
Initial research mainly to classic CNN architectures and transfer learning methods For example, Transfer learning from ResNet50 model with use done and infected and healthy cells more than 95% in separation high accuracy obtained [24]. The same in time [22], EfficientNet k-fold cross-validation method from the model with use making 97.57% detection achieved accuracy. More next in cases, CNN models ensemble ensemble learning methods with combining laboratory than 98% in datasets high performance obtained have done [16,25].

To these successes Although CNN models yet also one row with restrictions They are facing often due to staining differences , image of devices to its characteristics and information against data imbalance sensitive This is also true for real clinical models . in the data application when done performance cause of decrease (domain shift) These problems removed to lift for recent years two main research direction formed:

1. Resources limited which healthcare in their environments use for light, quantized and mobile to devices suitable models preparation.
2. Picture pixels between global addictions to model Transformer - based architectures application.

Medical visualization Vision Transformer (ViT) and his/her derivatives Data-efficient Image Transformer (DeiT), Pyramid Vision Transformer (PVT) and Swin Transformer fast attention to the center These models peculiar attention using the mechanism (self-attention) does and convolution limitations removed lifting from pictures multi -scale features and ink connections more efficient in the way to remove opportunity gives [29].
Comprehensive studies have shown that Vision Transformer models have achieved very successful results in tasks such as classification, segmentation and cell identification in the field of medical imaging [4,27]. From this addition, attention mechanism with CNN combining infection early in stages parasite structures more exactly obvious have succeeded in doing [21]. The same in time, multi-stage malaria detection for hybrid CNN–ViT model prepared and 99% accuracy obtained This result shows that CNN's local feature removal ability with ViT 's global attention mechanism to unite performance important to the extent increases [29].

Accuracy from being promoted other, medicine in the field artificial in the application of intelligence (AI) main from problems one explainability and clinical is trustworthy. Doctors model decided which one picture regions impact what you do finished For this purpose , Gradient-weighted Class Activation Mapping (Grad-CAM) and his/her improved versions Grad-CAM++ and Dropkey Grad-CAM developed . These methods model to the forecast the most very impact who regions Recent studies show that the Grad-CAM technique is superior to CNN and ViT architectures with together usage models explanation ability and diagnostic transparency important to the extent increases , medical to specialists while model behavior more deeply to understand opportunity gives [3,12].

2. Material and Methods

2.1. Dataset

In this study, a dataset available openly in the literature was used to identify malaria species [20]. Peripheral blood images were taken using a microscope. The dataset consists of four different malariavirus types: Falciparum, Malariae, Ovale,

and Vivax, and a total of 210 images. The images are 2598x1944 px and have a tiff extension. The images in the dataset are given in Figure 1.

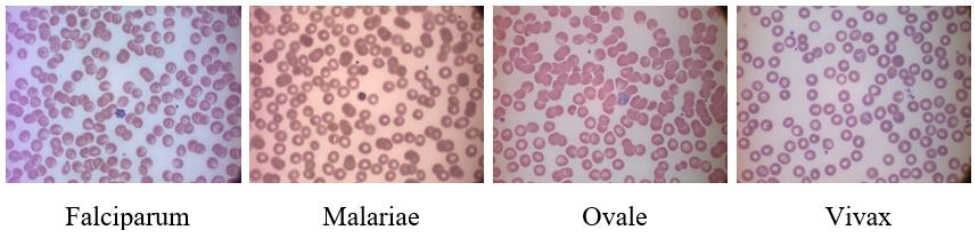

Figure 1. sample images in the dataset

2.2. Convolutional Neural Network

This study utilized the transfer learning strategy. Transfer learning utilizes four prominent convolutional neural network architectures to improve model generalization and classification accuracy on limited medical datasets: ResNet50, EfficientNetB0, InceptionResNetV2, and Xception. Transfer learning enables pre-trained models on large-scale datasets like ImageNet to effectively adapt to domain-specific tasks by reusing learned feature representations, thereby reducing training time and overfitting. ResNet50 provides redundant shortcut connections to overcome the vanishing gradient problems in deep networks, enabling the stable training of very deep architectures [17]. EfficientNetB0 employs a compound scaling method that uniformly balances network depth, width, and resolution to achieve high performance with optimal computational efficiency [33]. InceptionResNetV2 combines Inception modules with residual connections, offering both multi-scale feature extraction and enhanced convergence stability [23]. Meanwhile, Xception extends the Inception concept by employing depthwise separable convolutions, which improve computational efficiency and feature disentanglement across channels [34]. These CNN architectures, when fine-tuned through transfer learning, enable robust and interpretable feature extraction in medical image analysis, contributing to reliable diagnostic model development.

2.3. Vision Transformer

First of all natural language processing for created from the Transformer architecture inspired VIT, computer vision has revolutionized the field of. Vision Transformer (ViT) convolution from the layers use does not ; its instead entrance description small divides it into parts (patches) , each part feature to the vector converts and this vectors self-attention mechanism with global and local addictions via transformer encoder modeling processing does [19]. ViT and between CNN main difference characteristics removal CNNs place characteristics in a hierarchical manner to learn for local from filters use does, ViT and through self-attention far away regions between connections models and global structure place

limitation without understands. This is ViTs picture regions between far away relationships study demand who in tasks more superior does. With this so , some research shows that thin local details seizure CNNs in terms of yet also From ViTs more good result gives [10]. Attention within the Vit architecture is calculated as in equation 1

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

In the equation 1; Q, K and V relevant as **Query** , **Key** and **Value** matrices expression Here K**,** key vector size, The softmax function sets a weight for each part of the image, and this weight indicates how important one part is compared to another. d_k ; Key (K) represents the dimension (vector length) of vectors.

ViT model, picture various parts between global addictions to learn ability thanks to malaria with infected of cells microscopic in their descriptions ink patterns determination extremely for suitable one approach. The general diagram of the vision transformer model is given in Figure 2.

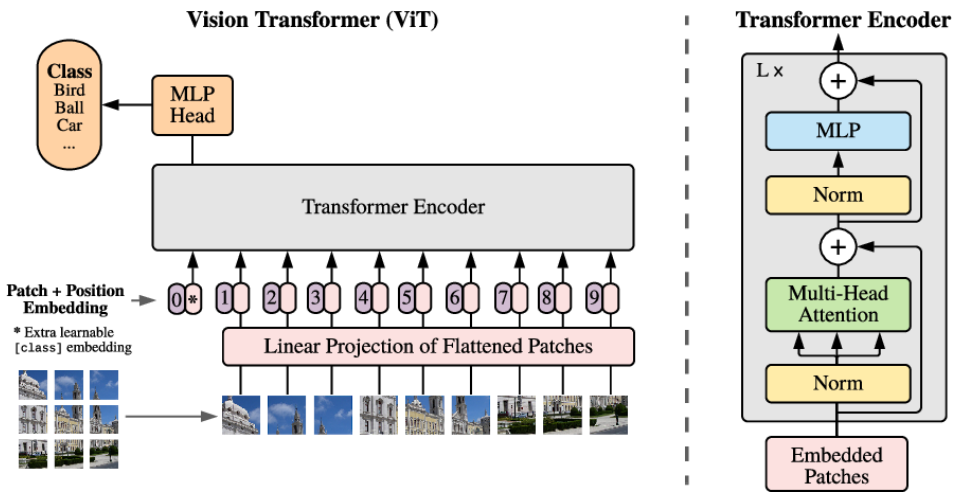

Figure 2. The Base Vision Transformer Model [9]

2.3. Base Vision Transformer (BaseViT)

ViT architecture standard one of the implementations and big voluminous from datasets picture learning image representations for intended This model is encoder Structure 12 Transformer from the floor consists of, each one 12 attentions per layer attention head is located and The embedding size is 768. This configuration to the model from pictures ink features hierarchical in the way to remove opportunity gives.

BaseViT architecture general structure so description to be done can: login picture first small patch-to divided , each one patch one feature to the feature vector turns and this vectors to the sequence position positional embeddings additional since it was done then they Transformer to encoder transmitted [14]. Every one encoder solid many-headed self-care block (multi-head self-attention block) and MLP (Multilayered Perceptron) from the block consists of ; these additional as layer normalization and residual residual connections also available [9]. These components combination to the model your picture various regions between both local, both also global connections to learn opportunity Malaria in their descriptions infected in cells color and texture in terms of thin differences observation that it was done according to BaseViT model this cells healthy from cells high precisely distinguish can knows.

2.4. Data-efficient Image Transformer (DeiT)

DeiT model, Touvron and others by presentation done and limited with data vision transformer models teaching for innovative one approach offer [32]. DeiT 's main distinctive from its characteristics one distillation token (distillation token) called special one token additional This token is patched and class tokens with together to the model included is done and teacher and student models between bridge role plays. Thus, the student model relatively small with datasets if taught yes, teacher model their knowledge more deep at the level to learn knows. DeiT, knowledge knowledge distillation frame application Here teacher model student to the model leadership does and ultimately student model less with data high accuracy obtained can knows [31]. Wide and targeted information increase**:** Teaching in the process DeiT various data augmentation from the methods use which is also the model available from the given more very information to remove conditions creates.

2.5. Pyramid Vision Transformer (PVT)

PVT dense prediction tasks (e.g., object detection and semantic segmentation) for multidimensional characteristics to be removed directed is a hierarchical transformer architecture. This model is a how many stage along entrance description size gradually reducing various on scales rich and high good quality features collector pyramid type hierarchical feature map forms. This structure multi-level Convolution Nerve Networks (CNNs) with similarity if it shows also, convolution operations without It was built [39]. Its instead Transformer's focus attention mechanism various description between resolutions place their addictions to model for use is being done. PVT model object detection and picture segmentation like tight prediction dense prediction tasks for both efficient, both also strong backbone in the state brings [35]. The general diagram of the PVT model is given in figure 3.

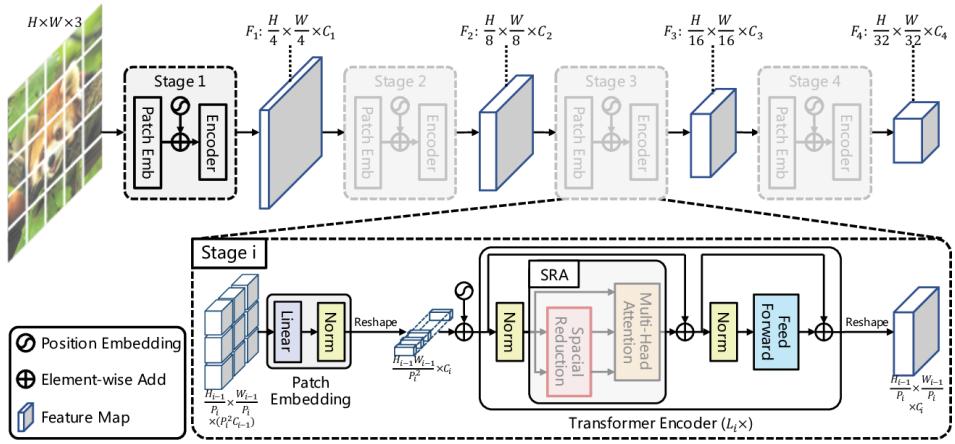

Figure 3. The general diagram of the PVT model [35].

2.6. Swin Transformer

Swin Transformer model slidden windows (shifted windows) the concept presentation by doing, calculating efficiency with model complexity between balance provide who one vision transformer architecture offer does [18]. In this architecture self- attention only small and each other with on top of each other unfallen windows inside It is calculated that this is also a calculation complexity important to the extent reduces. Neighbor windows between information exchange increase for, window positions consecutive in layers window size half until This mechanism **is** one window border tokens neighbor on the windows with tokens mutual in touch to be conditions creates and ultimately cross-window communication provide is done [36].

2.7. Gradient-weighted Class Activation Mapping (Grad-CAM)

Deep learning models main from problems one their decision-making in the process transparency lack of it. Grad-CAM (Gradient-weighted Class Activation Mapping) method model their decisions comment to do for strong one is a tool. This method convolution and either attention layers with related gradients calculating the model to the speech the most very impact who picture their regions emphasizes.

Grad-CAM also model debugging and reliability assessment (trust calibration) for effective in the way use be This can help researchers and to doctors model predictions behind logical more good finished to fall and to check opportunity creates. The Grad-CAM heatmap is a weighted combination of feature maps, followed by a ReLU. Its mathematical formula is given in Equation 2 [26].

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

Here, A_{ij}^{k} The activation value at spatial location $(i,)$ of the k-th feature map in the convolutional layer. Weight Calculation of coefficients (α_k) formula following is like :

$$\alpha_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^{k}} \quad (3)$$

Here, Z; The total number of spatial elements (pixels) in the feature map, used as a normalization factor when averaging the gradients.

3. Results

3.1. Performance metrics

Metrics derived from the confusion matrix were used to determine the performance of deep learning models in detecting malaria-infected cells. These are Accuracy **,** Sensitivity / Recall **,** Specificity**,** Precision indicator (Precision) and F1 score (F1-Score) These metrics include model predicted four possible to the conclusion mainly calculated : True Positive (TP) **,** True Negative (TN) **,** False Positive (FP) and Lie Negative (FN) [7].

The formulas (equations 4-8) and explanations of the metrics obtained from the confusion matrix are given below.

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

- Here:
  TP (True Positive):Correct certain done positive examples number (correct) obvious done infected cells)
- **TN (True Negative):** Correct certain done negative examples number (correct) obvious done healthy cells)
- **FP (False Positive):** Negative examples by mistake positive like classified made cases
- **FN (False Negative)**: Positive examples by mistake negative like classified made cases

Sensitivity (Recall): model positive samples (infected cells) correctly certain to do ability measures:

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

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

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

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

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

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

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

The 5-Fold Cross Validation method was used to conduct 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 results

In this study, eight different deep learning models were used to detect malaria from peripheral blood images. The performance results of four different CNN and four Vit models were examined. The CNN and Vit models were tested using the 5-fold cross-validation method. First, the dataset was analyzed using CNN models. The ResNet50, EfficientNetb0, InceptionResNetV2, and Xception architectures were used from the CNN models. To obtain effective results from the models, the epoch value of the entire dataset trained at once was set to 16, the mini batch-size value trained in each iteration was set to 16, and the learning rate was set to 1×10^{-4}. These settings were chosen to optimize the overall performance of the models. As a result of the experimental study, the best accuracy rate of 98.10% was achieved with the Xception CNN model. The performance values obtained with CNN models are given in Table 1. the confusion matrices are given in Figure 4 and train-val training and loss graphs of the models are given in Figure 5.

Table 1. Performance results of CNN models

CNN Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
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

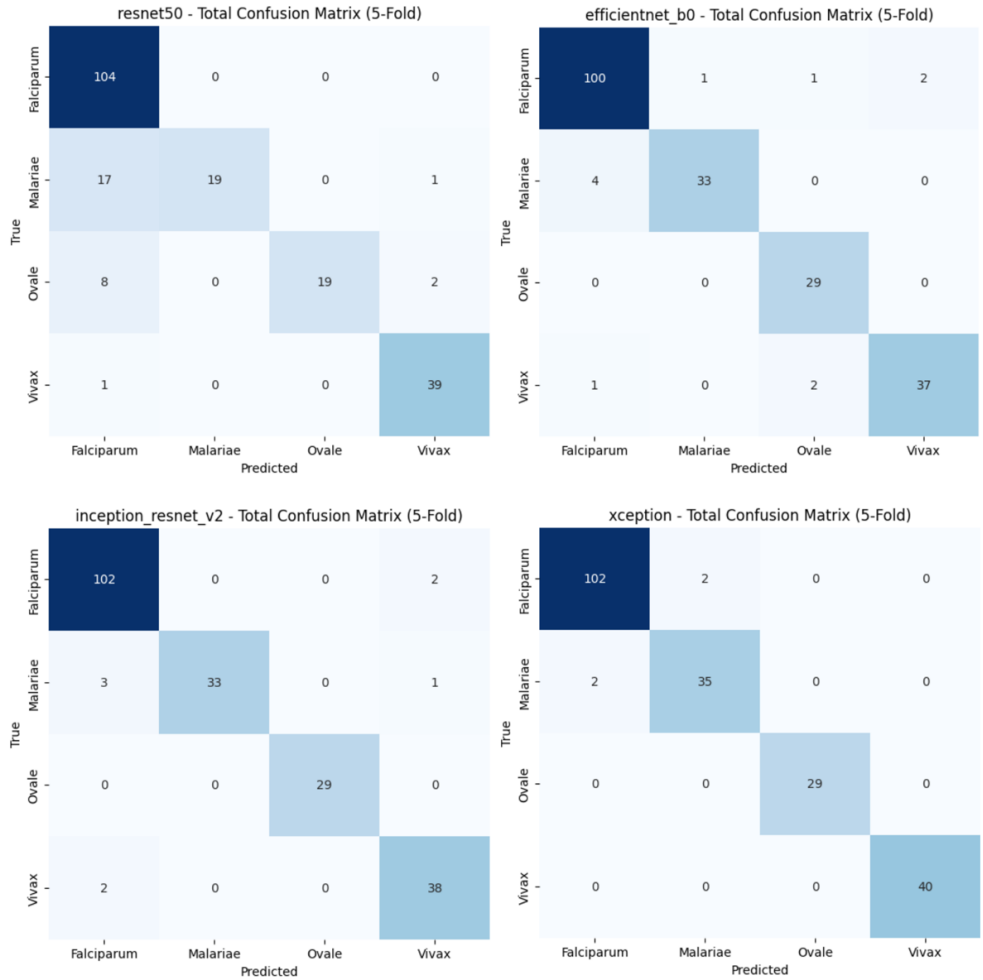

Figure 4. Confusion matrix of CNN algorithms

When examining the confusion matrices of the CNN models, it is observed that all models generally provide high classification accuracy, but the most balanced and error-free performance is achieved in the Xception model.

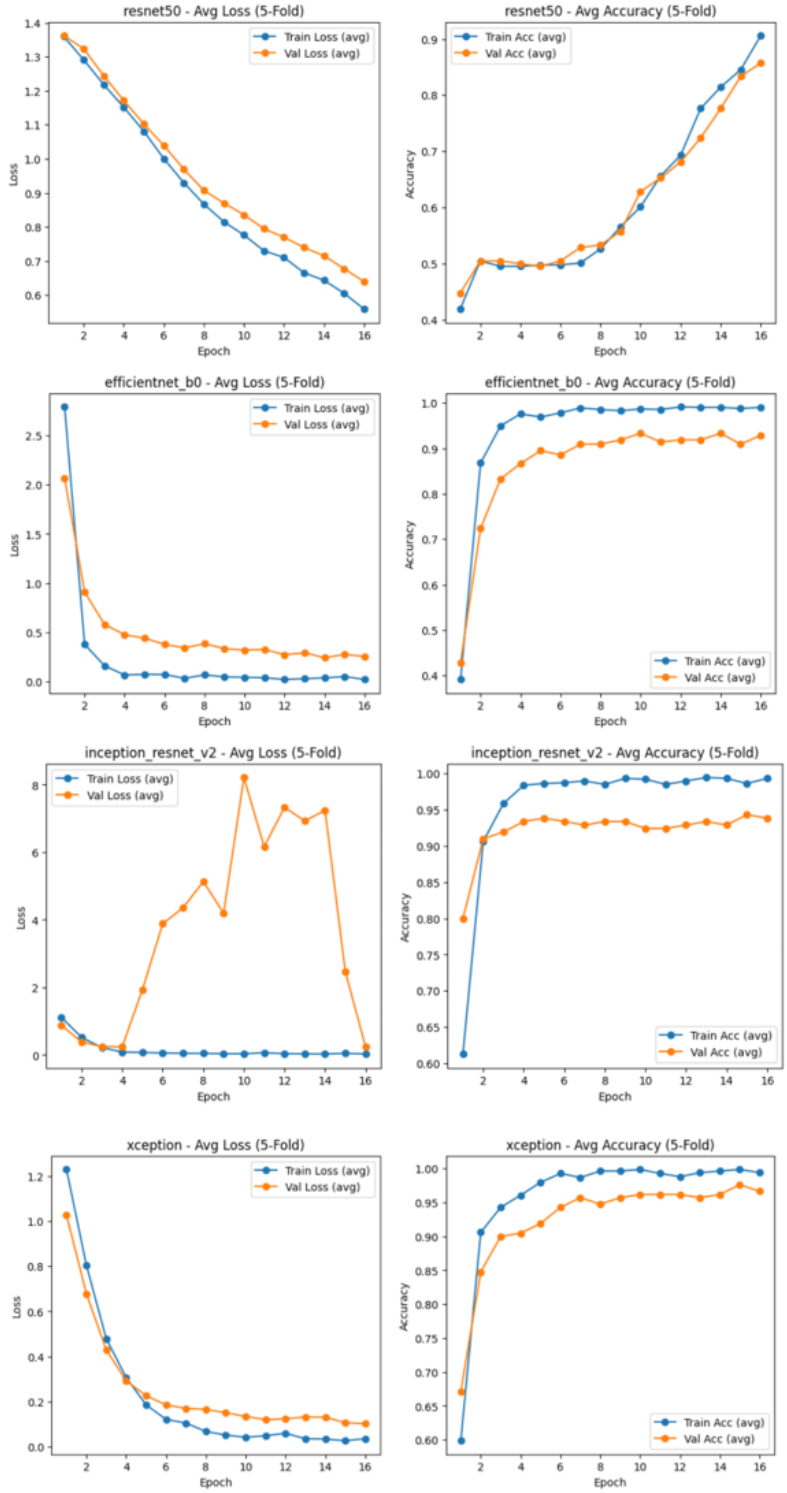

Figure 5. Trainin-validation training and loss graph

When examining the training and validation curves, it is observed that all models learn in a generally stable manner; however, despite fluctuations in validation loss, the Inception-ResNetV2 model generalizes more consistently and with higher accuracy compared to the Xception and EfficientNet-B0 models.

In the second phase of the experimental study, the data set was tested using the Vit_base model, Swim transformer, Deit, and Pyramid vision transformer algorithms. The 5-fold cross-validation method was chosen as the training and testing partitioning method for the study 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 the hyperparameters. The experimental study resulted in the best performance values with the Swim transformer algorithm, achieving an accuracy rate of 98.10%. The performance results obtained with the vit algorithms are given in Table 2. The confusion matrices are given in Figure 6, and the train-val training and loss graphs of the algorithms are given in Figure 7.

Table 2. Performance results of Vit algorithms

Vit Models	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**Precision (%)**	**F1-Score (%)**
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

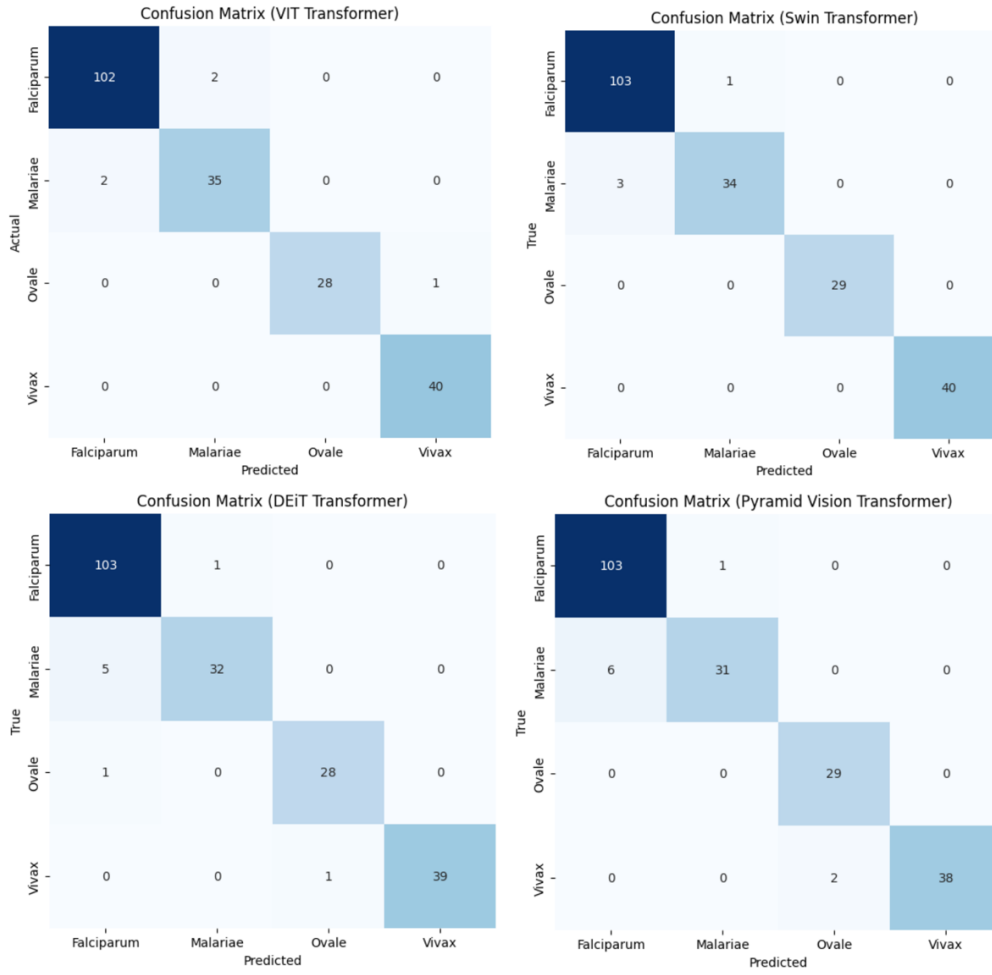

Figure 6. Confusion matrix of Vit algorithms

When examining the confusion matrices of Vision Transformer-based models, it is observed that all approaches achieve high classification performance, with Swin Transformer being the most balanced model with the lowest error rate.

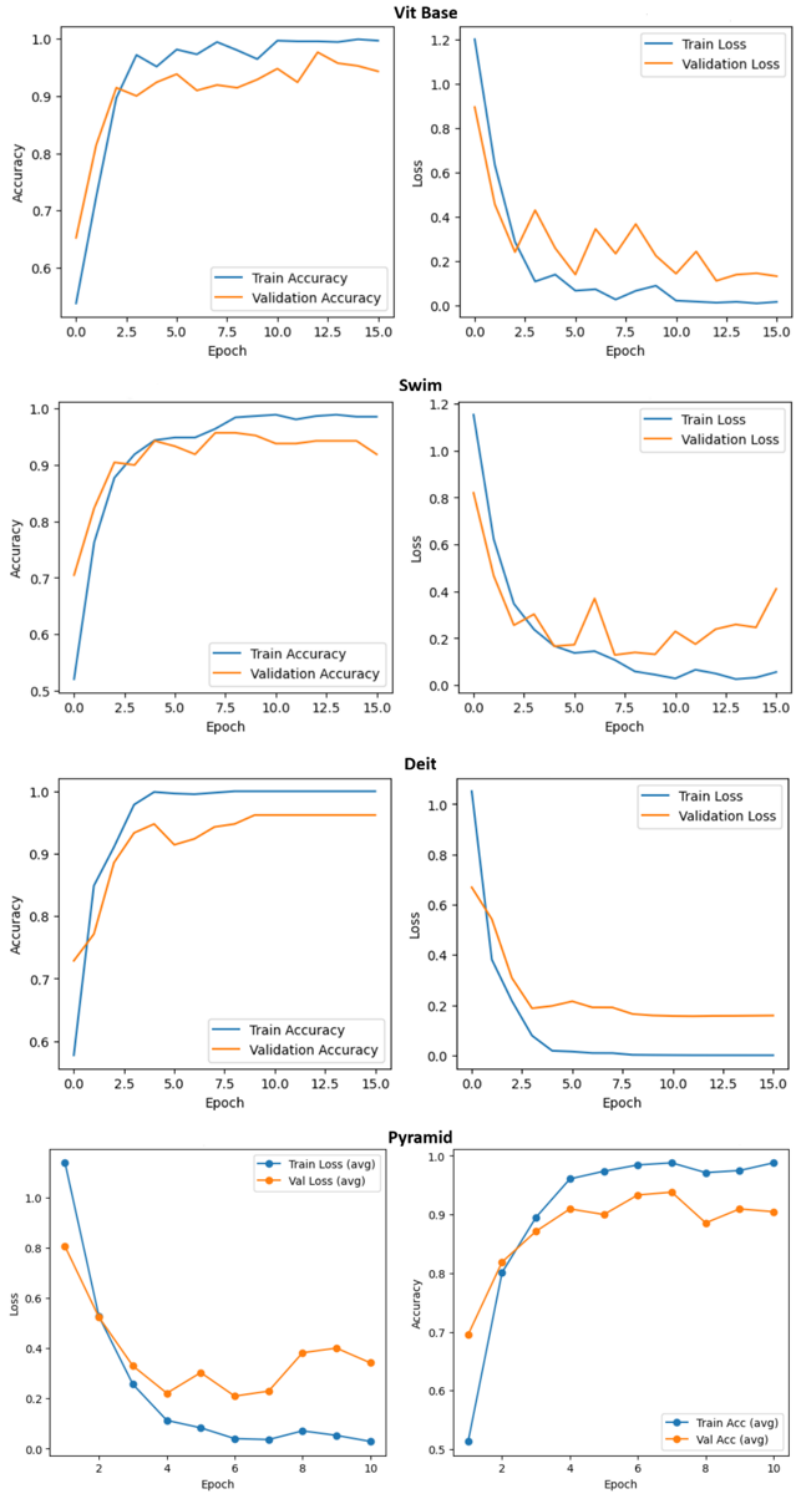

Figure 7. Accuracy and loss curves of Vit algorithms

Figure 7 shows that the Pyramid model exhibited a more stable generalization success compared to other models, demonstrating a more balanced performance in training and validation.

In the final stage of the experimental study, images were generated using the GRAD-CAM algorithm, an explainable artificial intelligence method, which yielded the best performance results with the Swim transformer algorithm. Models are made interpretable with GRAD-CAM images, enabling clinicians to interpret the images more easily. Random images generated with the GRAD-CAM algorithm are shown in Figure 8. When examining the images, the lesion areas colored red represent the region that is particularly effective in the models' decision-making. Here, it can be seen that the cells are colored red.

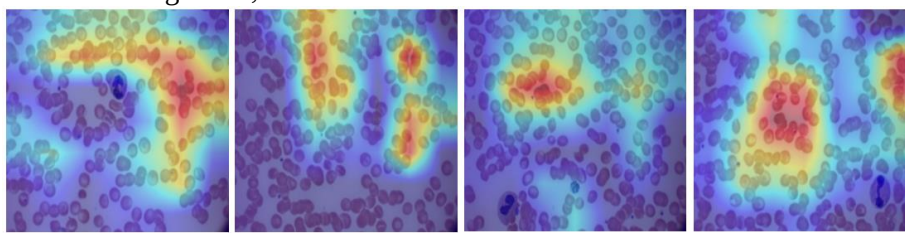

Figure 8. GRAD-CAM images produced with the swim transformer

4. Conclusion

Malaria remains a serious health problem, particularly for communities living in developing countries, and is one of the most significant parasitic infections threatening public health globally. Diagnostic methods using microscopes, which are still considered the gold standard today, have certain limitations due to the need for specialized personnel, the time required, and the intensive labor involved. This study aimed to increase the accuracy of malaria detection and automate the process. To this end, current 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. The limitation encountered while conducting the experiments was the low size of the data set.

The models used in the study were trained on a dataset consisting of microscopic blood images from four different types of malaria using the 5-fold cross-validation method, and their performance was evaluated. Among the convolutional neural network models, Xception yielded the most successful results with an accuracy rate of 98.10%; among the Transformer-based models, Swin Transformer demonstrated a similarly high level of performance. However, the Gradient-Weighted Class Activation Mapping (Grad-CAM) technique was applied to make the decision mechanisms of the models understandable and to increase their reliability in clinical applications. Thanks to this technique, it was possible to visually reveal which image regions the models considered in their diagnostic decisions.

The research results reveal that the use of Transformer-based deep learning architectures in conjunction with explainable artificial intelligence methods

significantly improves both classification success and model transparency. These developments enable the more reliable use of automated malaria diagnosis systems in clinical settings and their integration into real-life applications. Future studies plan to utilize large datasets and hybrid deep learning models.

References

1. Ahamed M.F. et al. Improving Malaria diagnosis through interpretable customized CNNs architectures, Sci. Rep., vol. 15, no. 1, p. 6484, 2025.9
2. Ahishakiye E., Kanobe F., Taremwa D., Nantongo B.A., Nkalubo L., and Ahimbisibwe S. Enhancing malaria detection and classification using convolutional neural networks-vision transformer architecture, Discov. Appl. Sci., vol. 7, no. 6, p. 612, 2025. 10
3. Awe O.O., Mwangi P.N., Goudoungou S.K., Esho R.V., and Oyejide O.S. Explainable AI for enhanced accuracy in malaria diagnosis using ensemble machine learning models, BMC Med. Inform. Decis. Mak., vol. 25, no. 1, p. 162, 2025. 21
4. Azad R. et al. Advances in medical image analysis with vision transformers: a comprehensive review, Med. Image Anal., vol. 91, p. 103000, 2024. 19
5. Başaran E., Cömert Z., and Celik Y. Convolutional neural network approach for automatic tympanic membrane detection and classification, Biomed. Signal Process. Control, vol. 56, p. 101734, Feb. 2020, doi: 10.1016/j.bspc.2019.101734. 5
6. Başaran E.,. Cömert Z, Çelik Y., Velappan S., and 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 Derg., vol. 22, no. 66, pp. 919–928, 2020, doi: 10.21205/deufmd.2020226625. 6
7. Başaran E., Çelik G. and. Toğaçar M. Regionally focused neural-coder model designed for the diagnosis of acute lymphoblastic leukemia disease, Measurement, p. 118176, 2025. 39
8. Çalışkan A. Diagnosis of malaria disease by integrating chi-square feature selection algorithm with convolutional neural networks and autoencoder network, Trans. Inst. Meas. Control, vol. 45, no. 5, pp. 975–985, 2023. 13
9. Dosovitskiy A., Beyer L., and Kolesnikov A. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. June 3, arXiv Prepr. arXiv2010.11929, 2021. 30
10. Dosovitskiy A.et al., An image is worth 16x16 words: {Transformers} for image recognition at scale, arXiv Prepr. arXiv2010.11929, 2020, Accessed: Mar. 07, 2025. [Online]. Available: https://arxiv.org/pdf/2010.11929/100029
11. Fong Amaris WM D.S., Martinez C., Cortés-Cortés L.J. Image features for quality analysis of thick blood smears employed in malaria diagnosis, Malar. J., vol. 21, no. 74, 2022. 3
12. Gao Y., Liu J., Li W., Hou M., Li Y., and Zhao H. Augmented grad-cam++:

super-resolution saliency maps for visual interpretation of deep neural network, Electronics, vol. 12, no. 23, p. 4846, 2023. 22
13. Goceri E. Deep learning based classification of facial dermatological disorders, Comput. Biol. Med., vol. 128, p. 104118, 2021. 8
14. Halder A., Gharami S., Sadhu P., Singh P.K., Woźniak M. and Ijaz M.F., "Implementing vision transformer for classifying {2D} biomedical images," Sci. Rep., vol. 14, no. 1, p. 12567, May 2024, doi: 10.1038/s41598-024-63094-9. 31
15. Islam M.R. et al. Explainable transformer-based deep learning model for the detection of malaria parasites from blood cell images, Sensors, vol. 22, no. 12, p. 4358, 2022. 12
16. Kaboré K.K., Guel D. and Somda F.H. Advancements in Deep Learning for Malaria Detection: A Comprehensive Overview., Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 6, 2025. 16
17. Karadeniz A.T., Çelik Y.and Başaran E. Classification of walnut varieties obtained from walnut leaf images by the recommended residual block based CNN model, Eur. Food Res. Technol., vol. 249, no. 3, pp. 727–738, 2023, doi: 10.1007/s00217-022-04168-8. 24
18. Liu Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv 2021, arXiv Prepr. arXiv2103.14030, vol. 10, 2021. 36
19. Liu Z., Mao H., Wu C.-Y., Feichtenhofer C., Darrell T., and Xie S. A convnet for the 2020s, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11976–11986. 28
20. Loddo A., Di Ruberto C., Kocher M., and Prod'Hom G. MP-IDB: The Malaria Parasite Image Database for Image Processing and Analysis BT - Processing and Analysis of Biomedical Information, N. Lepore, J. Brieva, E. Romero, D. Racoceanu, and L. Joskowicz, Eds., Cham: Springer International Publishing, 2019, pp. 57–65. 23
21. Maturana C.R. et al., Advances and challenges in automated malaria diagnosis using digital microscopy imaging with artificial intelligence tools: A review, Front. Microbiol., vol. 13, p. 1006659, 2022. 7
22. Mujahid M. et al. Efficient deep learning-based approach for malaria detection using red blood cell smears, Sci. Rep., vol. 14, no. 1, p. 13249, 2024. 15
23. Peng C., Liu Y., Yuan X., and Chen Q. Research of image recognition method based on enhanced inception-ResNet-V2, Multimed. Tools Appl., vol. 81, no. 24, pp. 34345–34365, 2022, doi: 10.1007/s11042-022-12387-0. 26
24. Rajaraman S. et al. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images, PeerJ, vol. 6, p. e4568, 2018. 14
25. Ramos-Briceño D.A., Flammia-D'Aleo A., Fernández-López G., Carrión-Nessi F.S. and Forero-Peña D.A. Deep learning-based malaria parasite detection: convolutional neural networks model for accurate species identification of Plasmodium falciparum and Plasmodium vivax, Sci. Rep.,

vol. 15, no. 1, p. 3746, 2025. 17
26. Selvaraju R.R., Das A., Vedantam R., Cogswell M., Parikh D., and Batra D. Grad-CAM: Why did you say that?, arXiv Prepr. arXiv1611.07450, 2016. 38
27. Shamshad F. et al. Transformers in medical imaging: A survey, Med. Image Anal., vol. 88, p. 102802, 2023. 20
28. Sukumarran D. et al. An optimised YOLOv4 deep learning model for efficient malarial cell detection in thin blood smear images, Parasit. Vectors, vol. 17, no. 1, p. 188, 2024. 4
29. Surendar P., Basha C.H., Nabi M. S.M., Kavitha L., Vijayakumar P. and Surender S. Multi-Modal Deep Learning for Malaria Diagnosis: Integrating CNNs and Vision Transformers for Enhanced Parasite Detection, in 2025 Third International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), IEEE, 2025, pp. 570–578. 18
30. Tan D. and Liang X. Multiclass malaria parasite recognition based on transformer models and a generative adversarial network, Sci. Rep., vol. 13, no. 1, p. 17136, 2023. 11
31. Toğaçar M. "BioTransNet: Detection of Plant Stress Through the Conversion of Biosensor Data into RGB Channels and Combination with Transformer Networks," J. Crop Heal., vol. 77, no. 5, p. 167, 2025, doi: 10.1007/s10343-025-01239-0. 33
32. Touvron H., Cord M., Douze M., Massa F., Sablayrolles A. and. Jégou H. "Training data-efficient image transformers & distillation through attention," in International conference on machine learning, PMLR, 2021, pp. 10347–10357. Accessed: Mar. 07, 2025. [Online]. Available: https://proceedings.mlr.press/v139/touvron21a 32
33. TR M., Gupta M., TA A., Kumar V., Geman O., and Kumar D. An XAI-enhanced efficientNetB0 framework for precision brain tumor detection in MRI imaging, J. Neurosci. Methods, vol. 410, p. 110227, 2024, doi: https://doi.org/10.1016/j.jneumeth.2024.110227. 25
34. Uzun M.Z., Başaran E., and Celık Y. Xception Derin Öğrenme Modeli ve Gabor Filtreleri ile ÇDÖÖE-DVM Algoritması Kullanılarak Mikro İfadelerin Tanınması TT - Recognition of Microexpressions Using Xception Deep Learning Model and Gabor Filters with RFECV-SVM Algorithm, J. Inst. Sci. Technol., vol. 13, no. 4, pp. 2339–2352, 2023, doi: 10.21597/jist.1252556. 27
35. Wang W. et al. Pyramid {Vision} {Transformer}: {A} {Versatile} {Backbone} for {Dense} {Prediction} without {Convolutions}, in 2021 {IEEE}/{CVF} {International} {Conference} on {Computer} {Vision} ({ICCV}), 2021, pp. 548–558. doi: 10.1109/ICCV48922.2021.00061. 35
36. Wei C., Ren S., Guo K., Hu H., and Liang J. High-Resolution Swin Transformer for Automatic Medical Image Segmentation, 2023. doi: 10.3390/s23073420. 37
37. World Health Organization. World malaria report 2024., Geneva: World Health Organization; 2024. Accessed: Jul. 08, 2025. Available:

https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2024 1

38. Yang F. et al. Deep learning for smartphone-based malaria parasite detection in thick blood smears, IEEE J. Biomed. Heal. informatics, vol. 24, no. 5, pp. 1427–1438, 2019. 2
39. Zhou Y., Jiang X., Xu G., Yang X., Liu X., and Li Z. PVT-SAR: An Arbitrarily Oriented SAR Ship Detector With Pyramid Vision Transformer, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 16, pp. 291–305, 2023, doi: 10.1109/JSTARS.2022.3221784. 34