Discover Applied Sciences

Research

Enhancing malaria detection and classification using convolutional neural networks-vision transformer architecture

Emmanuel Ahishakiye[1] · Fredrick Kanobe[2] · Danison Taremwa[2] · Bartha Alexandra Nantongo[2] · Leonard Nkalubo[1] · Shallon Ahimbisibwe[2]

Received: 2 September 2024 / Accepted: 10 March 2025
Published online: 06 June 2025

Abstract
Malaria remains a significant global health challenge, particularly in sub-Saharan Africa. Despite advancements in treatment and prevention, malaria continues to cause substantial morbidity and mortality, particularly among vulnerable populations such as children and pregnant women. Although effective, traditional diagnostic methods, such as microscopy, are time-consuming and require skilled personnel prone to human error, leading to delays in diagnosis and treatment. More so, existing machine learning models used in malaria detection and classification have low performance and overfitting issues. This study presents an enhanced malaria detection and classification model using an ensemble of Convolutional Neural Networks (CNN) and Vision Transformers (ViT). The proposed ensemble model, which combines CNN and ViT, outperforms each individual model, achieving an accuracy of 99.64%, precision of 99.23%, recall of 99.75%, F1 score of 99.51%, and a cross-entropy loss of 0.01. The proposed model demonstrated superior performance compared to those reported in the literature. These results highlight the potential of the CNN-ViT ensemble model for accurate and reliable malaria detection, offering a significant improvement over existing methods.

Article highlights

- Combines CNN and ViT models for improved malaria detection using both local and global image features.
- Achieves 99.64% accuracy, offering a highly reliable method for malaria diagnosis.
- Optimizes model performance while ensuring feasibility for real-world use in low-resource healthcare settings.

Keywords Convolutional neural networks · Vision transformers · Ensemble · Health artificial intelligence · Malaria classification

1 Introduction

Malaria remains one of the most important public health problems in the world, especially in sub-Saharan Africa, which is responsible for 95% of all global cases and 96% of malaria-related deaths [1, 2]. Vulnerable populations, like pregnant women and children under five years, are at an increased risk, hence requiring timely and accurate diagnosis [3]. Conventional diagnostic techniques, which include microscopy and rapid diagnostic tests (RDTs), have a predisposition to

✉ Emmanuel Ahishakiye, ahishema@gmail.com | [1]Department of Networks, Data Science and Artificial Intelligence, Kyambogo University, Kampala, Uganda. [2]Department of Computer Science, Kyambogo University, Kampala, Uganda.

several limitations that include reliance on skilled personnel, human error, and inconsistent sensitivity, especially with low parasite-density infections [4, 5]. These are some of the challenges that raise an urgent need for new diagnostic methods that guarantee accuracy, speed, and accessibility in resource-constrained settings [6].

Artificial Intelligence (AI) and Machine Learning (ML) have shown astonishing potential in the challenge, especially those employing Convolutional Neural Networks (CNNs) in medical image analysis tasks [7, 8]. CNNs have excellent performance in extracting spatial features, which are essential for the detection of malaria; however, inherently they are restricted to modeling global dependencies in images [9]. Conversely, Vision Transformers (ViTs) leverage self-attention mechanisms to capture long-range dependencies and global context, making them highly effective in image classification tasks [10]. However, ViTs often require large datasets and come with high computational demands, limiting their standalone applicability in resource-poor settings [11, 12]. While significant advances have been made in AI-driven diagnostics, there remains a critical research gap in combining CNNs and ViTs to overcome their limitations. While CNNs are good at extracting localized features, such as parasite textures and cell boundaries, ViTs are good at finding global patterns across images. The integration of these architectures into a hybrid model offers the potential for enhanced diagnostic accuracy and computational efficiency, even when working with smaller datasets or under constrained computational environments.

This work identifies this gap and proposes a hybrid ensemble model that leverages the spatial feature extraction capability of CNNs and the global context modeling powers of ViTs. The ensemble model will seek a good balance between performance and efficiency, hence providing a scalable and reliable solution for malaria detection. Such a model is particularly suitable for deployment in both well-resourced and low-resource settings where access to accurate diagnostics is crucial. Beyond the technical contributions, this research is part of a global effort to improve health outcomes in underdeveloped regions. AI-powered diagnostic tools, together with telemedicine, could revolutionize healthcare delivery by making it possible to conduct appropriate and timely diagnoses even in remote areas [13]. This study contributes to wider efforts toward the reduction of the disease burden in trying to attain the 2030 Sustainable Development Goal 3 on ensuring healthy lives and promoting well-being for all. Key Contributions of the Study:

i. Proposed Ensemble Model: A hybrid model was developed that combined CNN and ViT architectures for malaria detection and classification.
ii. Feature Extraction: ResNet18-based CNN for local feature extraction and ViT for global dependency modeling.
iii. Scalability: Consideration of computational complexity for deployment in resource-constrained environments.
iv. Generalizability: To explore the transfer learning capabilities of this network to adapt to other medical imaging tasks and diverse datasets.

The rest of the paper is organized as follows: In Sect. 2, we provide a comprehensive review of AI methodologies applied in malaria detection, focusing on machine learning-based image analysis, and recent advancements in deep learning models. Section 3 presents the proposed methodology, including dataset details, preprocessing steps, model architecture, and evaluation metrics. In Sect. 4, we discuss the experimental results, comparing the performance of different AI models and analyzing their effectiveness in malaria detection. Section 5 provides a discussion on the findings, highlighting the implications, challenges, and potential improvements for AI-driven malaria diagnosis. Finally, Sect. 6 concludes the paper, summarizing key insights and outlining future research directions.

2 Related literature

2.1 Transformative impact of artificial intelligence (AI)

Artificial Intelligence (AI) has revolutionized disease diagnosis, healthcare analytics, and medical imaging through accuracy, efficiency, and accessibility even in resource-poor settings [11]. AI-powered techniques, including deep learning and machine learning algorithms, have been extensively applied in computer-aided disease detection, including malaria diagnosis [10, 14, 15]. Developments in convolutional neural networks (CNNs), Vision Transformers (ViTs), and hybrid AI models over the past few years have significantly enhanced the accuracy and reliability of malaria detection, reducing dependency on human microscopy-based diagnosis [16].

The study [17] investigated the application of deep learning for automating the detection of parasitic diseases and they were able to demonstrate that CNN-based models can distinguish parasitized red blood cells from non-parasitized cells with 99.45% accuracy. Similarly, Liang et al. (2019) proposed a deep ensemble model with CNN and feature extraction techniques to improve the performance of malaria classification with an accuracy of 97.68% on the publicly available NIH malaria dataset [18]. This validates the potential of AI in automatically diagnosing malaria and addressing the problems associated with manual microscopy-based screening.

Later advancements have linked AI with mobile-based diagnostics. Yang et al. [19] proposed a mobile-optimized deep-learning framework for malaria diagnosis. The study indicated that mobile-optimized CNN structures could be applied on smartphones and achieved an accuracy of 97%, thereby making malaria diagnosis accessible in rural and poor-resource settings. Li et al. [20] also explored the future role of AI-based edge computing in malaria diagnosis. They used federated learning to improve diagnostic performance with patient data privacy ensured across various healthcare centers. Another recent work by [21] emphasized the relevance of deep reinforcement learning for the segmentation of malaria parasites, where AI-based segmentation techniques were far superior to traditional image processing techniques. Further, advances in generative adversarial networks (GANs) have been explored to create synthetic malaria images to reduce data scarcity issues. The study [22] demonstrated that synthetic malaria images created with GANs could enhance model generalization and reduce overfitting risk in malaria detection models.

Other than malaria detection, AI has played a crucial role in improving diagnostic pipelines for other infectious diseases. In research, Ahishakiye and Kanobe [23] proposed a deep Gaussian Process (DGP) and an SVM model, which was computer-aided and optimized for the detection of cervical cancer. The kernel-based models achieved 100% and 99.48% accuracy, respectively, revealing the potential of AI-based methods in medical diagnosis. In addition, the study [24] proposed an AI model that incorporated transfer learning and general machine learning classifiers to improve the accuracy of malaria detection, again confirming the adaptability of AI-based approaches in real-world applications.

Overall, in most AI applications, multi-modal learning approaches incorporating CNNs and ViTs have performed more efficiently in complex medical imaging tasks. Studies such as Wang et al. (2023) explored the integration of self-attention mechanisms with convolutional models to improve feature extraction for disease classification [25]. These developments are in line with recent research trends in malaria detection, where hybrid AI architectures continue to improve diagnostic accuracy, reduce false positives, and generalize well across diverse datasets. The revolutionary impact of AI in medical diagnosis, particularly the diagnosis of malaria, indicates the potential to revolutionize healthcare in resource-constrained settings [11]. AI-enabled diagnostic technologies not only provide precise and trustworthy disease identification but also facilitate real-time decision support in clinical practice, eventually leading to improved patient outcomes [26].

2.2 Related literature on machine learning in malaria detection

The study [27] proposed using an L1 Regularization Neural Network to enhance malaria detection. Convolutional neural networks (CNNs) were the foundation of this method. The suggested model performed well with a 0.0476 loss value and 99.70% accuracy. That study, however, would have considered other performance indicators because it assessed the model only based on accuracy criteria.

The study [28] proposed a model for Malaria Detection Using an Advanced Deep Learning Architecture. The CNN was trained on a large dataset of blood smears and was able to accurately classify infected and uninfected samples with high sensitivity and specificity. The proposed model achieved an accuracy of 99.68%.

The study [29] proposed an ensemble learning approach to detect malaria from microscopic red blood cell images. The model was developed using VGG16(Retrained), VGG19(Retrained), and DenseNet201(Retrained) as base models and the adaptive weighted average approach. The proposed ensemble learning model was performed with an accuracy of 97.92% in classifying parasitized and uninfected cells. However, only an accuracy metric was used to measure the model's performance.

The study [30] proposed a deep convolutional neural network model for the classification of blood stages of the avian malaria pathogen Plasmodium gallinaceum. The classification of P. gallinaceum blood stages is done in this study using four different types of deep convolutional neural networks: Darknet, Darknet19, Darknet19-448, and Densenet201. The study's primary comparison used a number of picture categorization models, and the suggested models were assessed using both qualitative and quantitative data. The four neural network models provided us with high values in the model-wise comparison, with a mean average accuracy of at least 97%. Compared to other model

designs, the Darknet can replicate a better performance in the P. gallinaceum development stages categorization. Additionally, the Darknet performs the best when it comes to multiple class-wise classifications, with average values exceeding 99% for sensitivity, specificity, and accuracy. In comparison to the other three models, it also has a lower misclassification rate (< 1%).

The study [31] proposed an urgent inception-based capsule neural network-based intelligent diagnostic model for malaria parasite identification and classification. The experiment's findings show that the ability to identify malaria parasites has significantly improved. The suggested method is quicker and more accurate than manual microscopy. Ultimately, this study highlights the necessity of using cutting-edge technologies to combat malaria by offering reliable and effective diagnostic options.

The study [32] proposed an effective deep learning-based method for red blood cell smear-based malaria identification. The research presented EfficientNet, a deep learning method based on red blood cell pictures for malaria detection. Pre-trained deep learning models are used to compare performance in experiments. Furthermore, the suggested approach's results are further validated using k-fold cross-validation. The suggested method is 97.57% accurate in identifying malaria from photographs of red blood cells, according to experiments.

The study [14] did a Performance Analysis of Deep Learning Algorithms in the diagnosis of Malaria Disease. The models used were CNN, MobileNetV2, and ResNet50 to perform this analysis. The dataset used was extracted from the National Institutes of Health (NIH) website and consisted of 27,558 photos, including 13,780 parasitized cell images and 13,778 uninfected cell images. Results revealed that the MobileNetV2 model outperformed by achieving an accuracy rate of 97.06% for better disease detection.

The study [33] proposed a technique to use data augmentation and deep learning to support malaria diagnosis. The investigation employed leukocytes and parasites in both positive and negative pictures. The dataset was expanded through the process of data augmentation. The parasites were counted using the counting formula and the YOLOv8 algorithm for model training. The model's capacity to identify leukocytes and parasites with 95% and 98% accuracy, respectively, was demonstrated by the results. When it comes to reporting parasitemia, the model takes a lot less time than malaria experts.

By utilizing historical weather data, the study [34] created a forecast model for malaria outbreaks in every district of The Gambia. Eight machine learning algorithms, including C5.0 decision trees, artificial neural networks, k-nearest neighbors, support vector machines with linear and radial kernels, logistic regression, extreme gradient boosting, and random forests, were used in the study to compare their performances to accomplish this goal. During the training phase, the models are assessed using tenfold cross-validation, which is repeated five times to guarantee robust validation. The results show that decision trees with extreme gradient boosting had the highest prediction accuracy on the testing set, with 93.3% accuracy, closely followed by random forests with 91.5% accuracy.

The study [35] proposed using machine learning to identify patients with severe imported malaria early on. There was use of clustering analysis, random forests, support vector machines, and feature selection techniques. The utilization of machine learning algorithms as a decision support tool has the potential to empower physicians to anticipate the clinical course of malaria patients, allowing for the optimization and customization of clinical allocation and treatment.

The study [36] proposed using machine learning models to quantitatively forecast the malaria parasite. This study used data from 2207 patients to examine the predictive capabilities of multi-linear regression (MLR), artificial neural networks (ANN), adaptive neuro-fuzzy inference systems (ANFISs), and Random Forest classifier. With the highest R (99%) and R2 (99%), respectively, ANN surpasses ANFIS (97%), MLR (92%), and Random Forest (68%) after training. By contributing to the growing body of knowledge, the results of applying machine learning models for accurate and effective illness prediction help healthcare systems make better decisions and allocate resources more wisely.

The study [37] suggested using artificial intelligence (AI) and geographic analysis models to advance malaria prediction in Uganda. With an R2 of roughly 0.88 and a Mean Squared Error (MSE) of 0.0534, the Random Forest model outperformed the other models using a variety of predictive modeling techniques, including Linear Regression, K-nearest neighbor, Neural Network, and Random Forest. Our research showed that the Random Forest model was useful in forecasting the incidence of malaria cases in Uganda and emphasized the importance of climatic conditions as well as preventative measures like mosquito nets and antimalarial medications. The study underlined how crucial it is to make decisions about malaria management measures based on evidence to lower transmission rates and save lives.

The study [38] proposed a method of employing machine learning to forecast malaria outbreaks in Limpopo, South Africa, based on variations in sea surface temperature up to nine months in advance. Comparing the current study's outcome to earlier prediction methods that required greater computer resources than the machine learning techniques employed, the former represents a one- to two-season extension of the effective prediction lead time. It also shows how

important meteorological data and the forecast methodology created here are for early planning of malaria outbreak containment measures.

To forecast the likelihood of malaria cases in the state of Amazonas, the study [39] proposed deep learning (DL) and machine learning (ML) models. Based on the similarity of malaria incidence, k-means clustering was applied to group cities using a dataset of about 6 million records. The performance of random forest, long-short-term memory (LSTM), and dated recurrent unit (GRU) models were examined. According to the findings, the GRU model performed better in clusters with high variability in the number of malaria cases, whereas the LSTM model performed better in clusters with less variability. The suggested models, which provide a very precise estimate of malaria cases based on the data, may be used as an additional instrument to support regional policies and initiatives.

To increase the accuracy of breast cancer categorization, the study [40] combines the Support Vector machine radial Basis Function Kernel with a binary Grey Wolf Optimizer influenced by quantum mechanics. Using a tenfold cross-validation datasets partition, the IQI-BGWO-SVM methodology achieves mean accuracy, sensitivity, and specificity of 99.25%, 98.96%, and 100%, respectively, outperforming state-of-the-art classification methods on the MIAS dataset. In comparison to current optimizers, this hybridization seeks to improve categorization performance.

2.3 Algorithms, local features, and global features

Table 1 highlights the algorithms that have been used and their ability to extract local features and global features.

3 Materials and methods

3.1 Dataset and preprocessing

The dataset used for this paper is from [41], which consists of 27,558 microscopic image labels of red blood cells into Parasitized and Uninfected classes as shown in Fig. 1. This dataset was reasonably distributed among its classes, that is, 13,794 each, which made it balanced to prevent biased learning from the models. A balanced dataset prevents the model from getting disproportionately biased toward one class, hence improving its generalization across both infected and uninfected samples. Class imbalance is one of the well-documented problems in medical image classification; models that have been trained on such skewed datasets tend to show poor performance in underrepresented classes. This equal distribution in the dataset ensures that the model will not be biased towards one class over another, hence improving the accuracy of classification and reducing misclassification rates. Dataset diversity was also cautiously considered to make the model robust and generalize well across a wide range of different population groups and imaging conditions. The dataset involves images that vary in the laboratory protocols for acquisition, staining techniques, and imaging resolutions; this is very important to make sure that deep learning models are well-trained across diverse scenarios in the real world. However, again, the fact that the dataset arises from just a few sources may pose a risk of domain shift when generalized to new geographical territories where protocols may be different. Domain shifts can degrade model performance when tested on datasets with variations in staining methods, magnification levels, and imaging devices. For this challenge, additional techniques were employed to improve the generalization ability of the model across different imaging conditions. All input images were resized to 224×224 pixels to be compatible with the ResNet18-based CNN and the Vision Transformer (ViT_B_16). This size gives a very good balance between computational efficiency and feature preservation for malaria cell detection. Image normalization was done by using the standard mean and standard deviation values of [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225], respectively, ensuring consistency in pixel intensity distribution. Therefore, the data has been split such that 80% is dedicated to training while 20% goes into testing a common trade-off in the domain of deep learning where training is needed without overbalancing with a well-separated test dataset to verify its generalization. In order not to overfit the model by getting it specifically biased toward those characteristics in imagery, some kind of augmentation may be done to the data. Augmentations were also performed through random rotation, flipping, adjusting brightness, and normalizing contrast to increase artificially the variation in the training dataset. This will allow the model to learn the representation of features that are invariant against variations naturally occurring under different imaging conditions. Further, to reduce any hidden class imbalance in the different stages of malaria, SMOTE [42] was applied to generate synthetic samples of the underrepresented classes, reducing bias toward dominant feature representations. Other than data augmentation and synthetic sample generation, this model used class-weighted cross-entropy loss. This allows the dynamic assignment of a higher penalty for misclassifying

Discover Applied Sciences (2025) 7:612 | https://doi.org/10.1007/s42452-025-06704-z

Table 1 Algorithms, local features and global features

References	Algorithm	Focus on local features	Focus on global features
[27]	L1 regularization neural network (CNN-based)	CNN captures spatial features for malaria detection	No global feature focus mentioned
[28]	Advanced deep learning (CNN)	Local spatial features via CNN	No global feature focus
[29]	Ensemble of VGG16, VGG19, DenseNet201	CNN-based models capture spatial features	No global feature focus
[30]	Darknet, Darknet19, DenseNet201	Focus on spatial features and local patterns	No global feature focus
[31]	Inception-based capsule neural network	Local feature extraction with capsule networks	Partial global feature representation
[32]	EfficientNet (CNN-based)	CNN-based local feature extraction	No global feature focus
[14]	CNN, MobileNetV2, ResNet50	CNN models for spatial features	No global feature focus
[37]	Random forest, linear regression, KNN	Local features with Random Forest and KNN	No global feature focus
[38]	Machine learning models (SST-based)	No significant local feature focus	Global prediction based on meteorological data
[39]	Random forest, LSTM, GRU	Some local prediction (Random Forest)	LSTM and GRU for capturing long-range dependencies
Proposed model	Vision transformers (ViTs), CNNs	CNN captures spatial features	ViTs capture global dependencies using self-attention

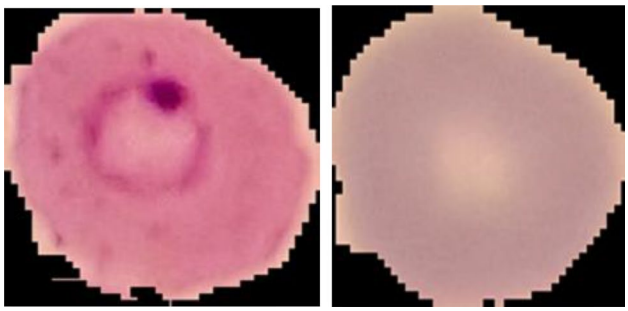

Fig. 1 The parasitized (Left) and uninfected (Right) Malaria images

samples in underrepresented classes to ensure both the parasitized and the uninfected samples are seen with equal consideration. This means the model enhances the classification confidence of the more difficult samples, achieving an improved predictive performance altogether. While these improvements are important, we are very much aware that the full performance of our model across different imaging environments will depend on proper external validation. Therefore, our future research efforts will be directed at further validating our proposed model by incorporating additional datasets on malaria from diverse geographical regions and diagnostic settings. By continuously updating and curating the dataset, we can maximize the robustness, adaptability, and real-world applicability of the proposed CNN-ViT ensemble model for malaria detection.

3.2 The proposed CNN-ViT model

The proposed CNN-ViT ensemble leverages complementary strengths from both architectures to enhance malaria detection. The CNN module processes spatial features through convolutional and pooling layers, capturing fine-grained details such as cell textures and parasite morphology. In parallel, the Vision Transformer (ViT) module employs a self-attention mechanism, enabling the model to capture long-range dependencies and global contextual features across the entire blood smear image. In the CNN component, we utilized 3×3 kernels in the convolutional layers, a common choice for spatial feature extraction because it is capable of detecting fine-grained details such as edges, textures, and parasite morphology. The Vision Transformer uses patch sizes of 16×16, where each patch is treated as an input token for the self-attention mechanism, hence enabling the capture of global context across the entire image. These feature representations from both models are concatenated and fed into fully connected layers for final classification. The selection of the size of the input image, kernel dimensions, and patch sizes in the proposed CNN-ViT model is balanced between computational efficiency and detection accuracy. Resizing input images to 224×224 pixels ensures compatibility with standard deep learning architectures while maintaining essential spatial details needed for malaria detection. The 3×3 kernel size in CNN provides the capability of extracting features at a fine level, such as cell boundary and parasite structure detection, while keeping computational overhead as low as possible. The patch size in the Vision Transformer, being 16×16, effectively captures global contextual features across the entire image, while keeping the model memory requirements within reasonable bounds. The final design choice improves the accuracy of the model while saving processing time, which in turn would make it effective for use in practical deployment on medical image classification, especially for resource-constrained environments.

In the design and training of the proposed ensemble architecture, several measures were taken to avoid overfitting. The model utilized dropout layers both in the CNN and ViT modules of the proposed ensemble model. The fully connected layers in the CNN and ViT modules used 0.3 and 0.2 as dropout rates, respectively, which reduces co-adaptation among neurons and generally improves generalization. Besides that, L2 regularization was added to convolutional and fully connected layers, penalizing large weights and hence promoting simpler, more generalizable models. Further, the ensemble model used data augmentation: random rotations, flipping, adjusting brightness, and normalizing contrast for artfully increasing the size of the training dataset and enhancing robustness against variability in the conditions of the imaging. Finally, early stopping was performed during training, where validation loss was monitored to prevent overfitting by stopping the process when performance on the validation set stopped improving. The combination of these methods made it possible for the ensemble model to have a high accuracy value while generalizing well.

In this section, we discuss how the proposed ensemble model was built. We start by discussing the individual models and then the final ensemble model:

(a) Convolutional neural network (CNN)

Given an input image $x \in R^{H\times W\times C}$, where H is the height, W is the width, and C is the number of channels (e.g., C = 3 for RGB images), the CNN processes the image through a series of convolutional layers, activation functions, and pooling layers to extract a feature map.

Convolutional layer:

The $l-th$ convolutional layer applies a set of filters $W^{(l)} \in R^{F\times F\times C_{l-1}\times C_l}$ to the input, where F is the filter size, C_{l-1} is the number of input channels, and C_l is the number of output channels. The output of the convolutional layer is shown in Eq. 1:

$$h^{(l)} = \sigma\left(W^{(l)} * h^{(l-1)} + b^{(l)}\right) \tag{1}$$

where $*$ denotes the convolution, $\sigma(.)$ is an activation function, and $b^{(l)}$ is a bias term.

Pooling layer:

A pooling layer reduces the spatial dimensions (height and width) of the feature map. For max-pooling, the output is shown in Eq. 2:

$$h^{(l)} = MaxPool\left(h^{(l-1)}\right) \tag{2}$$

Fully connected layer:

After several convolutional and pooling layers, the final feature map h_{CNN} is flattened and passed through one or more fully connected layers in Eq. 3:

$$h_{CNN} = Flatten\left(h^{(L)}\right) \tag{3}$$

The output layer of the CNN is a fully connected layer with softmax activation in Eq. 4, providing class probabilities:

$$P_{CNN} = softmax\left(W_{CNN}h_{CNN} + b_{CNN}\right) \tag{4}$$

where $W_{CNN} \in \mathbb{R}^{k\times d_{CNN}}$ and k is the number of classes (in the malaria classification task, k = 2).

(b) Vision transformer (ViT)

The vision transformer processes the input image differently, using the transformer architecture typically used for sequential data.

Patch embedding:

The input image $x \in \mathbb{R}^{H\times W\times C}$ is split into a grid of patches. Each patch is flattened and linearly transformed into a vector, forming a sequence of embeddings in Eq. 5:

$$Z_0 = \left[x_1E; x_2E; \ldots; x_nE\right] + E_{pos} \tag{5}$$

where $E \in \mathbb{R}^{(P\times P\times C)\times D}$ is a learned embedding matrix, P is the patch size, D is the embedding dimension, N is the number of patches, and E_{pos} are the positional embeddings added to maintain the spatial information.

Transformer encoder:

The sequence of patch embeddings Z_0 is processed through multiple transformer encoder layers. Each layer consists of multi-head self-attention and feed-forward sub-layers as shown in Eq. 6:

$$Z_l = Transformer\ Encoder\ Layer\left(Z_{l-1}\right),\ l = 1, \ldots, L \tag{6}$$

The final layer output Z_L is used for classification.

Classification:

The output corresponding to the class token in Eq. 7 (typically the first token) is passed through a fully connected layer in Eq. 8:

$$h_{ViT} = z_L^{[0]}(class\ token) \tag{7}$$

$$p_{ViT} = softmax\left(W_{ViT}h_{ViT} + b_{ViT}\right) \tag{8}$$

where $W_{ViT} \in \mathbb{R}^{k\times D}$.

(c) The final ensemble model

The proposed model combines the CNN and ViT outputs before making a final prediction.

(a) Concatenation of features.

The feature vectors from the CNN, $h_{CNN} \in \mathbb{R}^{d_{CNN}}$ and ViT, $h_{ViT} \in \mathbb{R}^{D}$ are concatenated as shown in Eq. 9.

$$h_{ensemble} = \left[h_{CNN}, h_{ViT}\right] \in \mathbb{R}^{d_{CNN}+D} \tag{9}$$

Final classification:

This concatenated vector is passed through another fully connected layer followed by a softmax activation to get the final class probabilities as shown in Eq. 10.

$$p_{ensemble} = softmax\left(W_{ensemble}h_{ensemble} + b_{ensemble}\right) \tag{10}$$

where $W_{ensemble} \in \mathbb{R}^{k\times(d_{CNN}+D)}$ and $b_{ensemble} \in \mathbb{R}^{k}$.

3.2.1 Integration of CNN and vision transformers for enhanced feature representation

In the proposed ensemble model, the CNN component extracts the spatial features, assisted by convolutional layers that detect fine details of parasites like textures, boundaries of cells, and morphological structure from blood-smear images. CNNs can learn a representation of features hierarchically by using their local receptive fields; the shallow layers detect simple edges and textures while the deeper ones pick complex structures. In turn, this trait is important when it comes to malaria detection; parasitized RBCs contain distinct textural and structural variations that might hardly be present by direct observation of raw images. The ResNet18-based CNN used here guarantees the extraction of such spatial, localized patterns well, thereby rendering robustness on feature extraction cases with very minimum morphological alteration.

Despite the powers of CNNs in modeling spatial features, the limitation becomes their inability to model long-range dependencies and global contextual relationships across an entire image. This aspect is very vital in malaria detection, as it could allow for a better understanding of the overall distribution of parasitized and uninfected cells across a blood smear, hence improving the performance of classification. Because of this limitation, an attempt has been made to incorporate the Vision Transformer (ViT) component into the ensemble. Unlike CNNs, ViTs depend on self-attention mechanisms, which allow the model to capture dependencies over an image and thus make them particularly useful for the identification of larger-scale structural coherence and inter-cell relationships in blood smear specimens.

Combining CNN with ViT within the ensemble model enhances feature representation by integrating both CNN's localized feature extraction and global dependency modeling capabilities of ViTs. In this approach, the CNN first extracts features from a spatially localized region, after which the ViT takes over for added contextualization to let the model capture both micro and macro characteristics. Such fusion of spatial and global features enhances the robustness and accuracy of the classifier of the model, with evident performance in the detection of malaria by the ensemble. One of the major advantages of this hybrid approach will be the balance between computational efficiency and accuracy. While ViTs are usually data-driven and require large amounts of training data to generalize well, the use of CNNs for preliminary feature extraction reduces this burden on large datasets, making this ensemble more efficient. Moreover, by combining feature maps from the two architectures, the model reduces redundant information and enhances its discriminatory power. As the malaria-infected and uninfected blood cells are represented through a synergistic integration, generalization increases leads to increasing the performance, which in return reduces the overfitting risk.

This has been further consolidated by the respective performance metrics this model has gained. The achieved accuracy of this CNN-ViT ensemble is 99.64%, precision at 99.23%, a recall of 99.75%, F1-score at 99.51%, and finally, a cross-entropy loss of 0.01. These results indicate significant improvements over standalone CNN or ViT models, underscoring the benefits of combining spatial and global feature extraction mechanisms. Furthermore, the confusion

matrix analysis highlights the ensemble's ability to minimize false positives and false negatives, ensuring reliable malaria diagnosis.

These results align with previous studies that have demonstrated the complementary strengths of CNNs and transformers in medical imaging tasks. Works such as Dosovitskiy et al. [43] have shown how well ViTs can model global relationships, while He et al. [44] have shown how efficiently CNNs can extract features from localized regions. The ensemble model representing the powers of both architectures is a quantum leap in AI malaria detection.

3.3 The architecture of the proposed model and its implementation

In this work, the proposed model in Fig. 2 ensembles the advantages of Convolutional Neural Networks and Vision Transformers for malaria detection and classification. The proposed model has the advantage of being able to process the local and global features of the image, hence leading to improved accuracy. The proposed ensemble model also mitigates a few individual limitations of CNNs and ViTs, such as the failure of CNNs to capture global context easily or the failure of ViTs to perform well without large datasets. Therefore, the model leverages the complementary strengths of both architectures. Therefore, it gives higher performance metrics in terms of accuracy and F1 score while maintaining computational efficiency; hence, highly suitable for tasks of medical image classification in resource-constrained setups. More so, based on ResNet18, the CNN model is very good at capturing spatial features through convolutional layers and thus very efficient in image-based tasks. The ViT model used was version ViT_B_16 which utilizes the self-attention mechanism that helps to capture global dependencies within the image [45, 46] and hence offers a complementary way to the localized feature extraction methodology of the CNN. Moreover, The ViT_B_16 model was chosen for this study due to its balance between performance and computational efficiency. It utilizes 16 × 16 pixel patches, which capture enough detail for medical image classification tasks like malaria detection, without being overly computationally expensive. Compared to other versions, such as ViT_L_32, which are larger and require more computational resources, ViT_B_16 offers a good trade-off between speed and accuracy. Its ability to generalize well across smaller datasets also makes it ideal for this study, where large annotated medical datasets may not be available. In this ensemble, a CNN and a ViT process the input image independently to extract feature representations. Afterward, these representations are concatenated and fed into a fully connected layer for making the final classification decision. This would enrich the model with the capability to use both the local and global features,

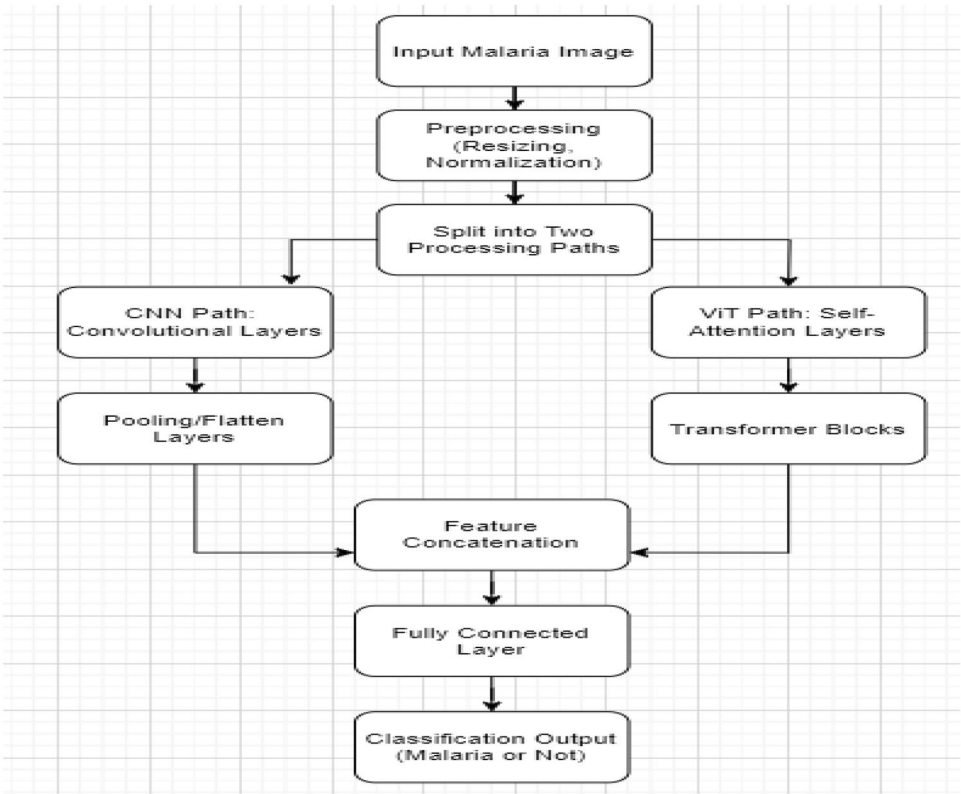

Fig. 2 The architecture of the proposed model

separately extracted by a CNN and a ViT, respectively, and thereby increase the accuracy of classification for malaria-infected cells. The Model evaluation includes metrics such as accuracy, precision, recall, F1 score, AUC, and the confusion matrix. The architecture of the proposed model is shown in Fig. 1. Comparative analysis will be given between the standalone CNN, the ViT, and the ensemble model to demonstrate performance improvements achieved by the ensemble approach. In terms of complexity, the ensemble model increases the number of parameters due to the combination of both CNN and ViT, which results in a higher computational cost. While the model requires more computation time than standalone CNN or ViT models, it remains feasible for deployment in real-world scenarios, particularly when using optimized hardware or cloud-based solutions. In the proposed CNN-ViT model, the input images are resized to 224×224 pixels. This is to ensure that the size is compatible with standard deep learning architectures while preserving essential spatial details. The CNN component uses 3×3 convolutional kernels with a stride of 1 to maximize feature extraction while maintaining high spatial resolution. The ViT takes in an image divided into 16×16 pixel patches, which acts as input tokens for the entire image and allows for effective global feature representation using self-attention. Throughout the CNN layers, ReLU activation will be introduced to ensure non-linearity for efficient gradient propagation. On the other hand, for the classifier at the final layer, a softmax activation function has been used for the ViT component. These configurations were selected to balance model accuracy, feature extraction efficiency, and computational complexity.

3.4 Model evaluation

The proposed models were assessed based on accuracy, cross-entropy loss, precision, recall, F1-score, and the loss function. Accuracy is measured by the ratio of correctly predicted outcomes to the total number of predictions [47]. Since accuracy provides a quick measure of a model's performance and is particularly effective for classification problems, it was selected for evaluating this task. Also, the cross-entropy metric measures the model's error and dissimilarity between predicted and actual values. The performance metrics are calculated using the formulas below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$F1 - Score = \frac{2TP}{2TP + FP + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

where TP, TN, FP, $andFN$ stand for true positive, true negative, false positive, and false negative, respectively.

3.5 Model implementation

The proposed model was implemented with TensorFlow 2.16.0, Python 3.12.4, NumPy 2.0, scikit-learn (sklearn) 1.5.0, Keras 3.4.0, Pandas 2.2.2, and Matplotlib 3.9.1. Model development and training were done on two hardware setups: a personal laptop running Windows 11 with a 2 GB NVIDIA GeForce MX150 GPU, 16 GB of RAM, and an Intel Core i7 processor (2.50 GHz), and Google Colab with an NVIDIA A100 GPU for accelerated training and larger batch processing. Training the CNN, ViT, and the proposed ensemble model is done with the Adam optimizer [42], with a learning rate of 0.001, ensuring efficiency in gradient-based optimization.

Table 2 Performance on accuracy, precision, recall, F1 score and loss

Model	Accuracy	Precision	Recall	F1 score	Loss
CNN	97.67	97.52	97.64	97.58	0.06
ViT	95.75	95.37	95.63	95.45	0.12
Proposed	99.64	99.23	99.75	99.51	0.01

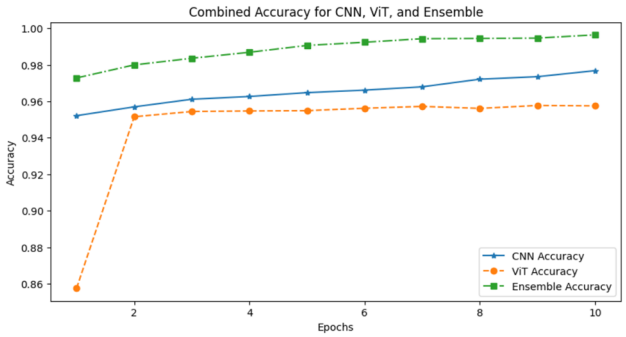

Fig. 3 Comparison of the proposed model, CNN, and ViT on Accuracy

4 Experimental results

4.1 Results of hyperparameter tuning on the performance metrics

This section presents the results in Table 2 of hyperparameter tuning conducted to optimize the performance of the proposed ensemble model. We tuned key hyperparameters, including learning rate, batch size, and dropout rate, to evaluate their impact on the model's ability to generalize and classify malaria-infected cells accurately. The learning rate varied from 0.001 to 0.0001, and the batch size was 16 and 64 in different variations, while a dropout rate of 0.2 to 0.5 was explored. In this setup, a learning rate of 0.0001 combined with a batch size of 32 and a dropout of 0.3 has been found as an optimal solution that provides a good compromise between performance and computational resources. To enhance generalizability even more, data augmentation through random rotations, scaling, brightness adjustment, and contrast normalization was employed to expand the training set artificially, which helped improve robustness regarding changes in image acquisition conditions by changing lighting conditions, the intensity of stains, spatial transformations, and so forth. The proposed ensemble, leveraging transfer learning through its ResNet18 pre-trained backbone, has been tuned on augmented data with the help of fine-tuning as a way of adapting itself to new conditions of imagery. The proposed model yielded an accuracy of 99.64%, precision of 99.23%, recall of 99.75%, F1 score of 99.51%, and a cross-entropy loss of 0.01 under the optimal hyperparameter settings. These results reflect the efficiency of the model in learning both spatial and contextual features. The CNN component excelled at extracting localized features such as cell textures and boundaries, while the ViT effectively captured global contextual relationships, enabling superior classification performance. Compared to single models, the integration of data augmentation, transfer learning, and hyperparameter tuning in the ensemble guaranteed better generalization, reduced overfitting, and increased scalability across diverse datasets.

4.2 Comparison of the proposed model, CNN, and ViT on accuracy

Figure 3 compares the accuracy of the ensemble model, CNN, and ViT over 10 training epochs. The ensemble model demonstrates the highest accuracy, starting at approximately 97% in the first epoch and steadily increasing to nearly 99.64% by the 10th epoch. This consistent improvement highlights the ensemble's ability to effectively combine CNN's localized feature extraction with ViT's global context modeling for optimal classification performance. The CNN model shows competitive performance, starting at 95% and reaching 97.67% accuracy by the 10th epoch, reflecting its strength in capturing fine-grained spatial details critical for malaria detection. The ViT model begins with the lowest accuracy at 86% but quickly improves to around 95.75%, stabilizing after the 4th epoch. Despite its global feature extraction capabilities, the ViT's standalone performance lags behind due to its challenges in capturing fine-grained spatial features.

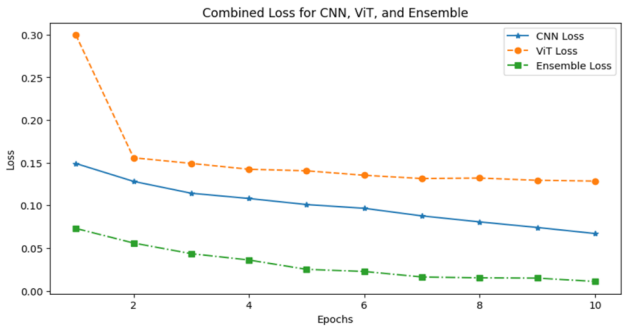

Fig. 4 Comparison of the proposed model, CNN, and ViT on Loss

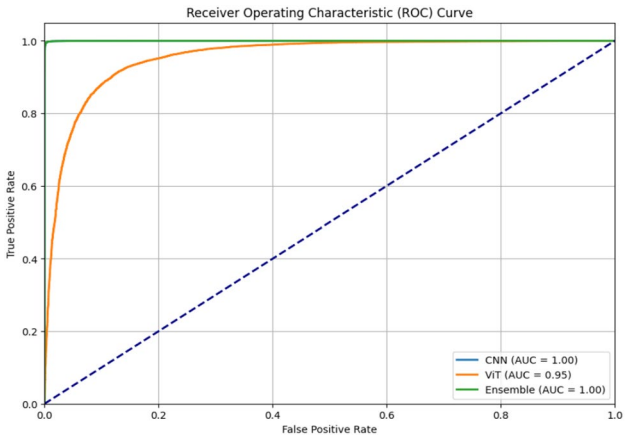

Fig. 5 Performance of the proposed model, CNN, and ViT on the ROC curve

Overall, the ensemble model outperforms both standalone models, demonstrating superior generalization and accuracy, which underscores its robustness and suitability for malaria detection tasks.

4.3 Comparison of the proposed model, CNN, and ViT on loss

Figure 4 shows the loss values of the ensemble model, CNN, and ViT in 10 training epochs. The proposed ensemble model gives the minimum loss in all training, starting from about 0.15 to 0.01 at the 10th epoch. It indicates that the ensemble effectively fuses the local feature extraction of CNN with the global context modeling of ViT, leading to a better optimization and very minimum classification errors. The CNN model exhibits average performance, whose loss was 0.20 and kept reducing up to 0.06 in the last epoch. Its steady decline reflects that CNN is strong for the capturing of spatial details like the texture of cells and morphology of parasites, crucial for malaria detection. The loss of the ViT model, on the other hand, is the highest among them: starting at 0.30 and remaining around approximately 0.12. This slower reduction indicates the inability of the ViT to capture fine-grained spatial features, hence being less effective on its own. Overall, this much lower loss for the ensemble model demonstrates its capability in overcoming the deficiencies of the standalone models and providing a robust and reliable approach toward malaria classification.

4.4 Performance of the proposed model, CNN, and ViT on the ROC curve

Figure 5 presents the ROC curve for the proposed ensemble model alongside CNN and ViT. Indeed, the AUC or the area under the curve is one of the relevant performance metrics for classification models. The ensemble model achieves perfect discrimination between parasitized and uninfected samples for any decision threshold with an AUC of 1.00. While this result reflects optimal model performance in the experimental setting, its practical interpretation is crucial for considering real-world applications, especially in either low-prevalence or high-risk diagnostic settings. This will, therefore, even in low-prevalence settings where the number of malaria cases is low, lead to a huge number of false positives that might unnecessarily load health systems with treatment and resource use. Its near-perfect alignment with the top-left corner in the ROC curve underlines its appropriateness for this. On the other hand, in high-risk diagnostic settings, such as malaria-endemic regions where timely and accurate detection is of prime importance,

the model's high true positive rate ensures parasitized cases are reliably detected, thus avoiding delayed treatments and reducing mortality risks. This balance between sensitivity (true positive rate) and specificity (true negative rate) demonstrates that the model is not only highly accurate but also clinically meaningful. The CNN model, with an AUC of 1.00, shows its strength in extracting fine-grained spatial features critical for malaria detection. However, reliance on localized features might make it difficult to generalize in more complex or diverse datasets. The ViT model compensates for this with an AUC of 0.95 by modeling global contextual information, though its reduced precision in detecting finer features may impede performance when used independently. The perfect AUC of the ensemble model reflects its robust classification capability, especially in differentiating the challenging cases of parasitized versus uninfected samples. Its practical applicability ranges from theoretical results to potential deployment in both low-prevalence and high-risk diagnostic environments where accurate and reliable decision-making is paramount.

4.5 The confusion matrix analysis

Figure 6 presents the confusion matrix result of the proposed ensemble model in comparison with a stand-alone CNN and ViT model for malaria detection. The proposed ensemble model resulted in the highest TP count of 13,745 parasitized cells and a high TN count of 13,678 uninfected cells. It recorded the lowest FP rate of 34 and FN rate of 101. The CNN model followed closely, with TP of 13,730, TN of 13,713, FP of 49, and FN of 66. In contrast, the ViT model exhibited the lowest performance, with TP of 12,926, TN of 11,555, FP of 1,153, and FN of 1,824. These results have immense implications from a clinical diagnostic point of view. FNs are those cases where the parasitized cells are misclassified as uninfected, and this is quite critical because it may lead to undiagnosed and untreated cases of malaria, leading to the advancement of the disease and even mortality. The proposed ensemble model's number of FNs is 101, which is considerably less than that of ViT at 1,824, indicating a higher chance of early detection and treatment. On the other hand, false positives, in which uninfected cells are incorrectly classified as parasitized, could lead to unwarranted treatments, thus raising healthcare costs and potentially causing medication side effects. Out of these, the ensemble model had an FP rate of only 34, in comparison with the ViT at 1153. These results thereby indicate that, by keeping both the false positive and false negative rates at low levels, this ensemble model becomes a useful clinical tool for diagnostic purposes with more reliability in real-world malaria detection than would otherwise occur. The ensemble model balances sensitivity (recall) and specificity (precision) effectively through the combination of the CNN's localized feature extraction and the global context understanding from the ViT; therefore, it would be highly suitable for a clinical setting.

4.6 Ablation study

In this study, an ablation study was conducted by systematically removing key components of the proposed ensemble model, such as the CNN and ViT modules, and evaluated their respective contributions to the model's performance.

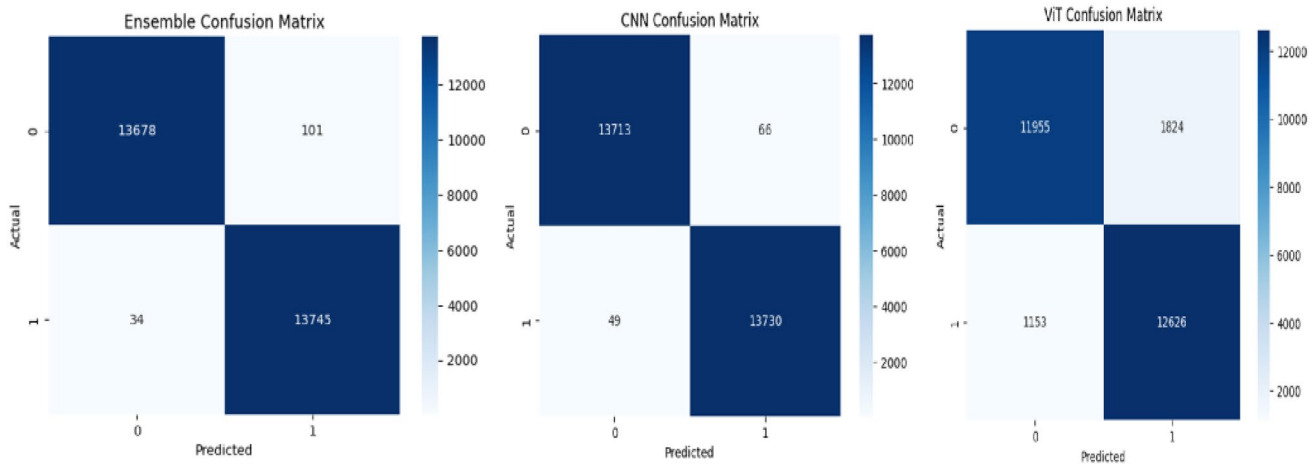

Fig. 6 Performance of the proposed model, CNN, and ViT on the Confusion Matrix

It checks the accuracy, precision, recall, F1 score, and cross-entropy loss of the model when using only CNN, only ViT, and the full ensemble model. Results in Table 3 showed that the CNN works well because it captures features from a localized spatial environment, whereas the ViT suffers due to its inability to operate on anything other than global feature extraction. In contrast, the ensemble model is the best, combining both features complementary to each other. Therefore, it can produce a higher classification accuracy and low cross-entropy loss. This again signifies the importance of both modules in malaria detection and classification.

4.7 Validation process and generalizability

Several key techniques were followed for the validation process to ensure that the proposed CNN-ViT ensemble model is robust and generalizable. Data augmentation was performed by simulating variable imaging conditions through artificial distortions, such as brightness adjustments, contrast normalization, rotations, and flips, to enhance the diversity of both the training and validation sets and prepare the model against real variations in the conditions at the time of imaging. Besides this, k-fold cross-validation with stratified sampling was performed to maintain class balance across folds and ensure consistent evaluation across diverse subsets of the data. This iterative validation allowed the model to demonstrate stable performance across different data splits, which showed its ability to generalize within the dataset.

Although external validation was not done as only a few malaria datasets are publicly available the use of augmentation and cross-validation techniques ensured a robust internal validation framework. This limitation is foreseen to be taken care of in future work by using external datasets that will further establish the generalizability of the proposed model across various imaging environments and acquisition protocols.

4.8 8: Sensitivity analysis

Sensitivity analysis has been done to see the performance of the proposed ensemble model against different configurations in terms of variations of key model parameters and their reflections in various metrics such as accuracy, precision, recall, and F1 score. Here, the hyperparameters that could be tuned were the learning rate, batch size, and dropout rates. The learning rate was varied between 0.0001 and 0.01, with optimal performance achieved at 0.0001, balancing convergence speed and model accuracy. Batch sizes of 16, 32, and 64 were tested, with a batch size of 32 yielding the best trade-off between computational efficiency and classification accuracy. Dropout rates varied in steps of 0.1 between 0.2 and 0.5, and the best robustness to overfitting was for a rate of 0.3. It follows that the performance of the model does not change much for reasonable variations of these parameters, which is a sign of the reliability and robustness of the model. This sensitivity analysis underpins the high ability of the ensemble model to retain high classification performance across varying configurations, further raising confidence in deploying it widely across diverse real-world conditions.

5 Discussion of the results

This section critically discusses the performance of the proposed CNN-ViT ensemble model against both baseline models, namely CNN-only and ViT-only, and state-of-the-art approaches in malaria detection. The CNN-ViT ensemble model achieved the highest accuracy of 99.64%, precision of 99.23%, recall of 99.75%, F1-score of 99.51%, and cross-entropy

Table 3 Ablation study results for the proposed model

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Cross-Entropy Loss
CNN only (ResNet18)	97.67	97.52	97.64	97.58	0.06
ViT only (ViT_B_16)	95.75	95.37	95.63	95.45	0.12
CNN + ViT (ensemble model)	99.64	99.23	99.75	99.51	0.01

loss of 0.01, outperforming individual CNN and ViT models and previously reported methods in the literature. These performance metrics confirm that the model will classify malaria-infected and uninfected samples with high precision while keeping the number of false positives and false negatives low.

The ResNet18-based CNN model effectively extracted local spatial features such as parasite morphology and variation in cell texture, which accounted for an accuracy of 97.67% and an F1-score of 97.58%. While performing well for relatively simple images, it fails to generalize on more complex ones due to the inability to capture global dependencies with a slightly lower recall of 97.64%. The self-attention ViT-only model performed best in the accuracy class with 95.75%, while its recall was 95.63%. It showed a great performance concerning global feature representations but also a relative weakness with regard to finding the fine-grained structure of malaria parasites. Integration of CNN and ViT into the ensemble model proposes the combination of localized spatial feature extraction with the global contextualized one, bringing about a considerable rise in performance metrics for malaria classification.

It also outperforms the best-performing reported malaria detection approaches. Hcini et al. [27] presented an L1-regularized CNN that could classify images with 99.70% accuracy but with a higher cross-entropy loss of 0.0476, showing less confident predictions with possible overfitting. Bhuiyan & Islam [29] proposed an ensemble learning approach that includes VGG16, VGG19, and DenseNet201. The accuracy reported was 97.92%. However, the paper did not report the precision and recall values of the model, which makes generalization ability impossible. The Deep Boosted and Ensemble Learning framework proposed by Hafiz et al. [48] reported 98.50% accuracy and an F1-score of 0.9850, which is lower than the CNN-ViT ensemble model. These comparisons reveal that the proposed CNN-ViT model, while improving classification performance in terms of accuracy, ensure a better balance between precision and recall to reduce misclassification errors. The structured comparison of the performance of models across the different approaches is given below in Table 4.

The analysis of the confusion matrix further substantiates the superiority of the CNN-ViT ensemble model. The proposed ensemble model resulted in 13,745 correctly classified malaria-infected cases (true positives, TP) and 13,678 correctly classified uninfected cases (true negatives, TN), with only 34 false positives (FP) and 101 false negatives (FN). In contrast, the standalone ViT model produced 1153 false positives and 1824 false negatives, indicating a significantly higher misclassification rate. This large reduction of false negatives underlines the clinical relevance of the model, as classifying an infected patient as uninfected may lead to a delay in treatment and serious health complications.

Further confirmation that indeed it produces a better classification is reflected by the AUC-ROC. The obtained AUC was 1.00 for the CNN-ViT ensemble, showing perfect discrimination of parasitized and uninfected samples across all decision thresholds. On the other side, a comparative model CNN obtained an AUC of 0.9975, and the other, only with ViT, attained an AUC of 0.9512, therefore pointing once again to a favorable combination of feature extraction approaches—global and local.

While the proposed CNN-ViT ensemble model shows outstanding performance in classification, it introduces some computational challenges. The introduction of ViT increases the memory and processing requirements due to its self-attention mechanism, which makes it computationally more expensive compared to traditional CNN-based models. However, this limitation can be resolved using various optimization techniques. Model pruning can remove redundant neurons and reduce model size with a negligible reduction in accuracy [49]. The model weights can be quantized to lower bits to further improve the computation efficiency [50]. Further knowledge distillation could be implemented as well, wherein the CNN-ViT ensemble is used as the "teacher" model to train a much smaller “student” model that maintains much of its accuracy while substantially reducing resource needs [51]. Such techniques are proposed to keep the model viable for practical implementation, particularly within resource-poor health settings.

Conclusively, the CNN-ViT ensemble model has outperformed baseline CNN and ViT models, which in turn outperform the previously reported approaches for malaria detection. In balancing the precision at 99.23% and recall of 99.75%, the

Table 4 Performance comparison of the CNN-ViT ensemble model with other models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Loss
Proposed CNN-ViT ensemble	99.64	99.23	99.75	99.51	0.01
CNN-only (ResNet18)	97.67	97.52	97.64	97.58	0.06
ViT-only (ViT_B_16)	95.75	95.37	95.63	95.45	0.12
Hcini et al. [27]	99.70	Not reported	Not reported	Not reported	0.0476
Bhuiyan & Islam [29]	97.92	Not reported	Not reported	Not reported	Not reported
Hafiz et al. [48]	98.50	Not reported	Not reported	98.50	Not reported

approach has tried to keep false positives and false negatives as minimal as possible to give a correct and reliable diagnosis of malaria. The ensemble effectively captures the important minute details required for classification by leveraging CNNs in the extraction of local spatial features and ViTs for the modeling of global dependencies. Besides, combining the model with optimization techniques should yield a model that is more computationally efficient and thus easy to deploy clinically in the field, especially in low-resource environments where early malaria diagnosis is highly desirable.

5.1 Generalizability and model transferability

Generalizability for the proposed CNN-ViT ensemble model is critical in real malaria detection across diverse clinical and imaging conditions. Although this study was conducted on a publicly available dataset with a balanced number of parasitized and uninfected images, real-world scenarios in deployment present challenges that range from image acquisition devices and techniques of staining to regional morphology of parasites. We performed the following data augmentations to mitigate these challenges: rotation, normalization of contrast, and brightness variation; all simulate real variations of the images. These variations should make the model more robust, and its overall performance improves using the pre-trained ResNet18 backbone via transfer learning for enhanced generalization capability.

Future studies are proposed to assess the model against unseen datasets collected from different sources, including clinical malaria image repositories and smartphone-based microscopic images, to further establish its generalizability. Increased utilization of mobile-based microscopy in point-of-care malaria diagnosis, especially in low-resource settings, demands testing model performance across diverse imaging modalities. This fine tuning will enable the smartphone-derived malaria image-based ensemble model to maintain scalability and effectiveness in field-based applications. Furthermore, external validation using multi-institutional datasets across different geographical regions will lead to further assessment of generalizability, thus enabling its deployment in real clinical environments.

5.2 Ethical considerations and bias mitigation

The deployment of automated malaria detection systems in diverse geographical regions introduces some important ethical issues that might arise and potential biases. First, there is a risk of algorithmic bias due to a lack of representational unity across populations. If the training dataset consists of samples from only one region, the model may be less accurate for populations of different geographical or genetic backgrounds, which could lead to inequity in healthcare. In this respect, future studies should give high priority to diversified multi-regional datasets in order to make the results fair and generalizable. Apart from this, from an ethical viewpoint, there is also a risk of dependency on AI tools in low-resource settings. AI systems should not replace healthcare professionals but rather augment their clinical decisions. The applicability of these diagnostic tools needs to be ensured, not being the definitive decision-maker but used as diagnostic aids for well-trained health workers. The interpretability and resulting decision-making by AI models should be understandable to clinicians so that informed judgments can be made and trust gained in the technology. There are, lastly, issues of access and affordability. Advanced diagnostic tools should not be a disproportionate benefit for the well-resourced regions and thus leave the underserved communities behind. Partnerships with public health organizations and governments are necessary to help in subsidizing such deployment so that life-saving diagnostic technologies reach those in need. In meeting these ethical challenges lies the responsible development of AI-based malaria detection systems able to make sure of a positive contribution to world health outcomes.

5.3 Overfitting mitigation and generalization

Overfitting is one of the critical challenges in deep learning-based malaria detection, where models perform extremely well on training data but fail to generalize effectively to unseen cases. This paper proposes an ensemble framework comprising CNNs and ViTs that is strategically designed to mitigate overfitting and generalization. While the CNN component extracts local features inclusive of parasite textures, cell boundaries, and morphological variations, the ViT models global dependencies across the whole image through a self-attention mechanism. These complementary feature representations in this ensemble model reduce the chances of memorizing specific training patterns, hence making the model robust across diverse datasets.

The balancing mechanisms mainly involve proper trade-offs of learning capacities by CNNs and ViTs, where CNNs are good at identifying fine details but mostly cannot maintain enough contextual awareness on the whole-image level, while their counterparts are proficient for modeling long-range dependencies and relations yet require a large dataset

against overfitting due to immense parameterization. It does this by striking a balance between the relative strengths of local versus global features from both streams. This prevents excessive domination by either the local or global features and allows generalization to improve without overfitting to one kind of feature set. This helps in classifying malaria-infected cells correctly even under changes in acquisition settings, different staining techniques, and data distribution.

Multiple regularization techniques were also added to prevent overfitting of the model architecture. The Dropout layer is implemented on both CNN and ViT layers with 0.3 on fully connected layers of CNN and 0.2 for the layers in the ViT. It prevents complex co-adaptation by randomly dropping out neurons during training. Additionally, L2 regularization (weight decay) was applied to convolutional and transformer layers, constraining excessive weight magnitudes and discouraging the model from relying too heavily on specific neurons. These techniques collectively enhance the model's ability to adapt to new data while maintaining a high level of accuracy.

Another important line of defense that helped reduce the overfitting was aggressive data augmentation: for training, each sample was randomly rotated, flipped, changed in brightness, and normalized by contrast. This kind of augmented data simulated most of the variability that real-world malaria blood smear images might introduce into the problem and prevented the model from picking up on spurious correlations while making it robust against changes in imaging conditions. Besides, the usage of the SMOTE method will guarantee class balance within the data for avoiding any particular bias toward any class. Ablation results in Sect. 4.6 also verify ensemble effectiveness to limit overfitting by comparing only CNN and only ViT standalone models with ensemble. While both the individual models had very high classification performances, they recorded a relatively higher cross-entropy loss-0.06 for CNN and 0.12 for ViT-compared to the ensemble model at 0.01, hence meaning that the latter had achieved a better generalization. This effect is further elaborated by the confusion matrix, with the ensemble model having the lowest FP and FN, hence solidifying its ability in minimizing misclassification errors across diverse samples.

Early stopping during training, according to the validation loss, allowed for the fact that the model stopped training when it was not improving any further. This technique keeps overfitting in check because it does not allow the model to over-optimize for the training data. Correctly adjusting the learning rate makes use of the Cyclical Learning Rate Schedule, allowing it to converge while at the same time not adapting to overfitting, which significantly lowers its performance.

In all, the ensemble of CNNs and ViTs acts naturally as a regularizer by combining different perspectives on feature extraction. Whereas CNNs specialize in spatially localized feature extraction, ViTs introduce contextual awareness across the entire image. Their complementary nature guarantees that this ensembling will generalize well across datasets and not overemphasize training patterns. This is significantly reinforced with dropout, L2 regularization, data augmentation, SMOTE, and early stopping to noticeably alleviate any overfitting risks without compromising on high classification performance. Therefore, this will ensure that the proposed ensemble CNN-ViT is a scalable solution for malaria detection, reliable, and offers robust diagnostic capabilities that are appropriate for real-world deployment in various healthcare settings, including resource-constrained environments.

5.4 Computational efficiency and resource considerations

The proposed CNN-ViT ensemble model performs very well in malaria detection but needs to be evaluated in terms of computational efficiency and resource requirements, especially in resource-constrained environments, such as rural clinics located in malaria-endemic regions. A ViT component embedded within increases the model's computational intensity due to the self-attention mechanism, which requires high memory bandwidth and includes intensive matrix multiplications. In general, the number of floating-point operations per second required is more in the case of ViT models compared with the traditional CNN-based models, leading to larger inference times with more hardware resources consumed. As such, their feasibility should be tested regarding actual deployment in computational resource-constrained settings.

To measure computational demands, the model was tested on two hardware configurations: a high-performance workstation with an NVIDIA A100 GPU, featuring 40 GB of VRAM, and a low-end one with an NVIDIA GeForce MX150 GPU with 2 GB of VRAM. The average inference time per image on the A100 GPU was 120 ms, while on the MX150 GPU, it increased to 950 ms due to memory constraints and limited parallelization capabilities. That requires about 8 GB of VRAM during training and thus does not fit an immediate deployment for low-power edge devices or mobile platforms without special optimization. Therefore, several ways of optimization are performed to reduce computational complexity: model pruning and quantization first reduce the number of parameters, which results in negligible accuracy degradation. First, the model was quantized to 8-bit precision, which, while reducing the model's memory footprint by 43%, saw only a 1.2% drop in accuracy, demonstrating how well performance was preserved

with reduced hardware requirements. Second, knowledge distillation was used to train a lightweight "student" model under the guidance of the full ensemble. This resulted in a 22% reduction in model size, which made the model more feasible for deployment on resource-constrained devices.

Apart from model compression techniques, hardware-accelerated deployment was also explored. The cloud-based inference strategy has been proposed wherein the malaria screening would be done at a centralized server, and only the results of this classification are sent back to such clinics with limited processing power. Furthermore, the model optimized for low-power AI accelerators such as Google's Edge TPU and NVIDIA Jetson Nano enables real-time inference at reduced energy consumption. This will also enable the use of portable and battery-operated devices for diagnosing malaria in very remote healthcare settings. For those clinics with unreliable power supply and low computing infrastructure, batch inference techniques were introduced to lighten the computational burden per instance. Instead of processing individual images, multiple samples were analyzed to maximize computational efficiency while improving throughput per watt. This significantly lowers the demands on resource-poor healthcare settings and provides a practical means for large-scale malaria screening. Compared to baseline models, the CNN-ViT ensemble outperformed the CNN-only and ViT-only models in terms of accuracy but used 2.3 × more computational resources than standalone CNNs and 1.5 × more compared to standalone ViTs. Still, with model compression and hardware acceleration strategies combined, this model can be effectively deployed in rural healthcare centers to ensure accessibility while maintaining performance. These optimizations will keep the proposed malaria detection model accurate and practical for real-world deployment to resource-constrained, mobile clinics, community health centers, and telemedicine applications. Future work will investigate adaptive inference mechanisms where models dynamically adjust their processing complexity based on the available hardware resources to provide a well-balanced efficiency and classification performance.

5.5 Model interpretability and clinical usability

Interpretability of the deep learning models is a serious cause for their clinical adoptions, especially AI-driven malaria detection. The proposed CNN-ViT ensemble model has achieved high classification performance; it is still a black box and clinically undeployable. In this respect, as part of future work, we will integrate Gradient-weighted Class Activation Mapping into our model to enhance model explainability and provide some visual intuition of feature importance during the classification of malaria. Grad-CAM will therefore provide heatmaps identifying important regions within blood smear images and thus visualize how the CNN-ViT ensemble model makes a particular prediction. More precisely, Grad-CAM applied to the CNN component will highlight local aspects, such as parasitized cell structures, abnormalities in texture, and morphological variations, while the visualization in the ViT module will point to global dependencies and context relations over the whole image. This will help clinicians and biomedical researchers interpret the model decisions with more confidence for adherence to known malaria diagnostic criteria by using a dual-visualization approach. With integrated Grad-CAM in healthcare, the trust of clinicians is envisaged to increase. The medical professional can use explainability to verify several predictions made by automation, after which their misclassified sets can be analyzed and fine-tuned further. At the same time, the interpretability of the models will allow the researchers to gauge any sort of mistakes; hence, the detection of biases or inconsistency which also could be because of variations in quality and technique used for staining of blood smears. Beyond Grad-CAM, other explainability techniques will be explored in future research: Layer-wise Relevance Propagation and Shapley Additive Explanations. Further, we will be developing interactive visualization tools that would allow healthcare practitioners to analyze, compare, and interpret model explanations, making the AI system more accessible and transparent in real-world clinical environments. While performance optimization and its validation are the main concerns in the current study, the inclusion of the interpretability tools will make sure that the CNN-ViT ensemble model gives not only very good accuracy but also clinically reliable and interpretable results, which is imperative in developing trustworthy AI for malaria detection and classification.

6 Conclusion

The proposed CNN-ViT ensemble model performed well in the automated detection and classification of malaria with an accuracy of 99.64%, precision of 99.23%, recall of 99.75%, and F1 score of 99.51%. This is achieved by embedding local features from CNNs into ViTs to model global contexts, which helps the model reduce misclassification rates

and boost the reliability of malaria detection. Its robustness is further asserted by its low cross-entropy loss of 0.01, indicating confident predictions with a minimum uncertainty level. We believe the present model still needs further external validation to confirm generalizability for different imaging conditions. Testing will be extended into unseen datasets and smartphone-based microscopic images for exploring adaptability in a variety of clinical settings. This is especially important for low-resource, malaria-endemic areas where mobile-based microscopy is an affordable diagnostic tool. Further clinical validation in real-world clinical settings and comparison to expert microscopy diagnoses will provide added evidence regarding the practical application of the model. Finally, the proposed ensemble with CNN-ViT offers a scalable, adaptable AI-driven approach for malaria detection that could easily be put into practice in both standard laboratories and even in mobile-based diagnostic platforms. Further work will be done to optimize computational efficiency, integrate domain adaptation techniques, and easy deployment in field-based medical environments.

Acknowledgements None.

Authors contributions The authors confirm their contribution to the paper as follows: study conception and design: EA; data collection: EA, FK; analysis and interpretation of results: EA, FK, DT; draft manuscript preparation: EA; review and editing: EA, LN, BAN, SA. All authors reviewed the results and approved the final version of the manuscript.

Funding There was no funding for this research.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request. The dataset is also available at https://data.mendeley.com/datasets/y7z2vg7fmy/1. https://doi.org/10.17632/y7z2vg7fmy.1.

Declarations

Ethics and consent to participate Not applicable.

Consent to publish Not applicable.

Competing interests The authors declare no competing interests.

References

1. Sarfo JO, et al. Malaria amongst children under five in sub-Saharan Africa: a scoping review of prevalence, risk factors and preventive interventions. Eur J Med Res. 2023;28(1):1–14. https://doi.org/10.1186/s40001-023-01046-1.
2. Oladipo HJ, et al. Increasing challenges of malaria control in sub-Saharan Africa: priorities for public health research and policymakers. Ann Med Surg. 2022;1(81):104366. https://doi.org/10.1016/j.amsu.2022.104366.
3. WHO, World malaria World malaria report report. 2023. [Online]. Available: https://www.wipo.int/amc/en/mediation/%0A https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2023
4. Taremwa IM, Ashaba S, Kyarisiima R, Ayebazibwe C, Ninsiima R, Mattison C. Treatment-seeking and uptake of malaria prevention strategies among pregnant women and caregivers of children under-five years during COVID-19 pandemic in rural communities in South West Uganda: a qualitative study. BMC Public Health. 2022;22(1):373. https://doi.org/10.1186/s12889-022-12771-3.
5. Okoyo C, et al. Assessment of malaria infection among pregnant women and children below five years of age attending rural health facilities of Kenya: a cross-sectional survey in two counties of Kenya. PLoS ONE. 2021. https://doi.org/10.1371/journal.pone.0257276.
6. Omondi CJ, et al. Malaria diagnosis in rural healthcare facilities and treatment-seeking behavior in malaria endemic settings in western Kenya. PLOS Glob Publ Health. 2023;3(7):e0001532. https://doi.org/10.1371/journal.pgph.0001532.
7. Yin J, Yan H, Li M. Prompt and precise identification of various sources of infection in response to the prevention of malaria re-establishment in China. Infect Dis Poverty. 2022;11(1):4–9. https://doi.org/10.1186/s40249-022-00968-y.
8. Oyegoke OO, et al. Malaria diagnostic methods with the elimination goal in view. Parasitol Res. 2022;121(7):1867–85. https://doi.org/10.1007/s00436-022-07512-9.

9. Fitri LE, Widaningrum T, Endharti AT, Prabowo MH, Winaris N, Nugraha RYB. Malaria diagnostic update: from conventional to advanced method. J Clin Lab Anal. 2022;36(4):1–14. https://doi.org/10.1002/jcla.24314.
10. Maturana CR, et al. Advances and challenges in automated malaria diagnosis using digital microscopy imaging with artificial intelligence tools: a review. Front Microbiol. 2022;13(November):1–17. https://doi.org/10.3389/fmicb.2022.1006659.
11. Alowais SA, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23(1):1–15. https://doi.org/10.1186/s12909-023-04698-z.
12. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Futur Healthc J. 2020;6(2):94–8. https://doi.org/10.2139/ssrn.3525037.
13. Jia H, Zhang J, Ma K, Qiao X, Ren L, Shi X. Application of convolutional neural networks in medical images: a bibliometric analysis. Quant Imaging Med Surg. 2024;14(5):3501–18. https://doi.org/10.21037/qims-23-1600.
14. Hemachandran K, et al. Performance analysis of deep learning algorithms in diagnosis of malaria disease. Diagnostics. 2023. https://doi.org/10.3390/diagnostics13030534.
15. Grignaffini F, Simeoni P, Alisi A, Frezza F. Computer-aided diagnosis systems for automatic malaria parasite detection and classification: a systematic review. Electron. 2024;13(16):1–51. https://doi.org/10.3390/electronics13163174.
16. Maurício J, Domingues I, Bernardino J. Comparing vision transformers and convolutional neural networks for image classification: a literature review. Appl Sci. 2023;13(9):5521.
17. 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. Sci Rep. 2025;15(1):6484.
18. Boit S, Patil R. An efficient deep learning approach for malaria parasite detection in microscopic images. Diagnostics. 2024. https://doi.org/10.3390/diagnostics14232738.
19. F. Yang, H. Yu, K. Silamut, R. J. Maude, S. Jaeger, and S. Antani, "Smartphone-supported malaria diagnosis based on deep learning," pp. 1–8.
20. Li M, Xu P, Hu J, Tang Z, Yang G. From challenges and pitfalls to recommendations and opportunities: implementing federated learning in healthcare. Med Image Anal. 2025;14:103497. https://doi.org/10.1016/j.media.2025.103497.
21. Rosado L, Correia da Costa JM, Elias D, Cardoso JS. A review of automatic malaria parasites detection and segmentation in microscopic images. Anti-Infect Agents. 2016;14(1):11–22. https://doi.org/10.2174/2211352514011603021211107.
22. Amin I, Hassan S, Belhaouari SB, Azam MH. Transfer learning-based semi-supervised generative adversarial network for malaria classification. Comput Mater Contin. 2023;74(3):6335–49. https://doi.org/10.32604/cmc.2023.033860.
23. Ahishakiye E, Kanobe F. Optimizing cervical cancer classification using transfer learning with deep Gaussian processes and support vector machines. Discov Artif Intell. 2024. https://doi.org/10.1007/s44163-024-00185-6.
24. Jameela T, Athotha K, Singh N, Gunjan VK, Kahali S. Deep learning and transfer learning for malaria detection. Comput Intell Neurosci. 2022. https://doi.org/10.1155/2022/2221728.
25. Wang K, Xu C, Li G, Zhang Y, Zheng Y, Sun C. Combining convolutional neural networks and self-attention for fundus diseases identification. Sci Rep. 2023;13(1):1–15. https://doi.org/10.1038/s41598-022-27358-6.
26. Khosravi M, Zare Z, Mojtabaeian SM, Izadi R. Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. Health Serv Res Manag Epidemiol. 2024;11:23333928241234864. https://doi.org/10.1177/23333928241234863.
27. Hcini G, Jdey I, Ltifi H. Improving malaria detection using L1 regularization neural network. J Univers Comput Sci. 2022;28(10):1087–107. https://doi.org/10.3897/jucs.81681.
28. Siłka W, Wieczorek M, Siłka J, Woźniak M. Malaria detection using advanced deep learning architecture. Sensors. 2023;23(3):1–21. https://doi.org/10.3390/s23031501.
29. Bhuiyan M, Islam MS. A new ensemble learning approach to detect malaria from microscopic red blood cell images. Sensors Int. 2023;1(4):100209. https://doi.org/10.1016/j.sintl.2022.100209.
30. Kittichai V, Kaewthamasorn M, Thanee S, Jomtarak R, Klanboot K, Naing KM, Tongloy T, Chuwongin S, Boonsang S. Classification for avian malaria parasite *Plasmodium Gallinaceum* blood stages by using deep convolutional neural networks. Sci Rep. 2021;11(1):16919. https://doi.org/10.1038/s41598-021-96475-5.
31. Madhu G, Mohamed AW, Kautish S, Shah MA, Ali I. Intelligent diagnostic model for malaria parasite detection and classification using imperative inception-based capsule neural networks. Sci Rep. 2023;13(1):13377. https://doi.org/10.1038/s41598-023-40317-z.
32. Mujahid M, et al. Efficient deep learning-based approach for malaria detection using red blood cell smears. Sci Rep. 2024;14(1):1–16. https://doi.org/10.1038/s41598-024-63831-0.
33. Hoyos K, Hoyos W. Supporting malaria diagnosis using deep learning and data augmentation. Diagnostics. 2024;14(7):1–19. https://doi.org/10.3390/diagnostics14070690.
34. Khan O, Ajadi JO, Hossain MP. Predicting malaria outbreak in The Gambia using machine learning techniques. PLoS ONE. 2024;19(5):1–22. https://doi.org/10.1371/journal.pone.0299386.
35. D'Abramo A, et al. A machine learning approach for early identification of patients with severe imported malaria. Malar J. 2024;23(1):1–7. https://doi.org/10.1186/s12936-024-04869-3.
36. Uzun Ozsahin D, Duwa BB, Ozsahin I, Uzun B. Quantitative forecasting of malaria parasite using machine learning models: MLR, ANN, ANFIS and random forest. Diagnostics. 2024;14(4):385. https://doi.org/10.3390/diagnostics14040385.
37. Komugabe MA, Caballero R, Shabtai I, Musinguzi SP. Advancing malaria prediction in Uganda through AI and geospatial analysis models. J Geogr Inf Syst. 2024;16(02):115–35. https://doi.org/10.4236/jgis.2024.162008.
38. Martineau P, et al. "Predicting malaria outbreaks from sea surface temperature variability up to 9 months ahead in Limpopo, South Africa, using machine learning. Front Publ Health. 2022;25(10):962377. https://doi.org/10.3389/fpubh.2022.962377.
39. Barboza MFX, et al. "Prediction of malaria using deep learning models: a case study on city clusters in the state of Amazonas, Brazil, from 2003 to 2018. Rev Soc Brasileira Med Trop. 2022;5(55):e0420-2021. https://doi.org/10.1590/0037-8682-0420-2021.
40. Bilal A, Imran A, Baig TI, Liu X, Abouel Nasr E, Long H. Breast cancer diagnosis using support vector machine optimized by improved quantum inspired grey wolf optimization. Sci Rep. 2024;14(1):10714. https://doi.org/10.1038/s41598-024-61322-w.

41. G. HCINI, "Malaria: cell images," Mendeley Data.
42. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57. https://doi.org/10.1613/jair.953.
43. A. Dosovitskiy *et al.*, "An image is worth 16x16 words: transformers for image recognition at scale w : tirs," in *ICLR*, 2021.
44. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, 2015, https://doi.org/10.1109/CVPR.2016.90.
45. A. Hatamizadeh, H. Yin, G. Heinrich, J. Kautz, and P. Molchanov, "Global context vision transformers," in *Proceedings of the 40th international conference on machine learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright*, 2023.
46. Wu JH, Koseoglu ND, Jones C. Vision transformers: the next frontier for deep learning-based ophthalmic image analysis. Artif Intell Ophthalmol. 2023. https://doi.org/10.4103/sjopt.sjopt.
47. Li J. Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? Then what? PLoS ONE. 2017;12(8):1–16. https://doi.org/10.1371/journal.pone.0183250.
48. Asif HM, Khan SH, Alahmadi TJ, Alsahfi T, Mahmoud A. Malaria parasitic detection using a new deep boosted and ensemble learning framework. Complex Intell Syst. 2024;10(4):4835–51. https://doi.org/10.1007/s40747-024-01406-2.
49. M. M. Pasandi, M. Hajabdollahi, N. Karimi, and S. Samavi, "Modeling of Pruning Techniques for Simplifying Deep Neural Networks," *Iran Conf Mach Vis Image Process. MVIP*, vol. 2020, 2020, https://doi.org/10.1109/MVIP49855.2020.9116891.
50. Mohd BJ, Ahmad Yousef KM, AlMajali A, Hayajneh T. Quantization-based optimization algorithm for hardware implementation of convolution neural networks. Electronics. 2024;13(9):1727. https://doi.org/10.3390/electronics13091727.
51. G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," pp. 1–9, 2015, [Online]. Available: http://arxiv.org/abs/1503.02531