Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 5, 2026

A COMPARATIVE EVALUATION OF DEEP LEARNING ARCHITECTURES FOR AUTOMATED MALARIA PARASITE DETECTION IN BLOOD SMEAR IMAGES

Abdul Sattar Chan [1], Zainab Umair Kamangar[2], Umair Ayaz Kamangar[*3], Mumtaz Ali[4], Junaid Ahmed[5]

[1, *3,4,5]*Department of Computer Systems Engineering, Sukkur IBA University, Pakistan*
[2]*Department of Computer Science, Sukkur IBA University, Pakistan*

[1]abdul.sattar@iba-suk.edu.pk, [2]zainabumair.phdcss22@iba-suk.edu.pk, [*3]umair.ayaz@iba-suk.edu.pk, [4]mumtaz.ali@iba-suk.edu.pk, [5]j.bhatti@iba-suk.edu.pk

DOI: https://doi.org/10.5281/zenodo.20389625

Keywords

Article History
Received: 28 March 2026
Accepted: 07 May 2026
Published: 26 May 2026

Corresponding Author: *
Umair Ayaz Kamangar

Abstract

Malaria continues to pose as the most prevalent global health problem with a particular predominance in developing countries lacking modern diagnosing mechanisms. Correct identification and efficient diagnosis of the malaria parasite are extremely crucial for treatment and containment of the disease. The research analyzes performance of advanced deep learning architectures to perform the classification of infected and uninfected blood cells using microscopic images. Five state-of-the-art models-EfficientNet-B2, Vision Transformer Small (ViT-S/16), ResNet-152, DenseNet-201, and Swin Transformer Base-were selected and tested after training on 27,558 labeled microscopic blood images obtained from National Institute of Health. The experiments used transfer learning and partial fine-tuning with BCE With Logits Loss and AdamW with cosine annealing for optimal classification of images. Swin Transformer Base delivered the highest test accuracy of 97.73% and an AUC value of 0.9969. The best performance on accuracy vs efficiency trade-off was achieved by EfficientNet-B2 with an accuracy and AUC of 97.10% and 0.9962, respectively. Transformer models produced sensitivities and specificities greater than 96%. The comparative analysis helps identify current strengths of deep learning for medical image classification and assists practitioners in selecting models in low-resource medical settings.

1. INTRODUCTION

1.1 BACKGROUND ON MALARIA

Malaria is a parasitic disease caused by Plasmodium parasites and is transmitted to humans via infected Anopheles mosquito vectors. Malaria infects tens of millions worldwide and is endemic in many tropical and subtropical regions. There were an estimated 249 million malaria cases and 608,000 malaria deaths worldwide in 2022. Of these deaths, 95% occurred in sub-Saharan Africa [1]. In many low resource health systems, where many malaria cases are seen, facilities to support laboratory diagnostic tests and trained microscopists can be few and far between. Currently diagnosis is done manually under the microscope, a time consuming and laborious process requiring a skilled technician and the inter-observer variation between individual microscopists can be considerable. Rapid diagnostic tests are more rapid but are less

sensitive than microscopy and less cost effective at scale. Accurate, automated and scalable diagnostic tests for malaria parasites in blood are urgently required.

1.2 EVOLUTION OF DEEP LEARNING IN MEDICAL DIAGNOSTICS

Deep learning methods have driven tremendous progress in medical image analysis over the last decade. In image classification, simple Convolutional Neural Networks were shown by Krizhevsky et al. [4] to achieve impressive results and then evolved through residual networks [1] that permit to train deeper networks and then through transformer models [11, 12] capable of modeling global interactions. Many applications were found for CNNs in the medical field in the analysis of medical images, such as chest radiography, histopathology and microscopy image analysis. The process of transfer learning [6, 7] made deep learning accessible in specialized medical fields where limited amount of training data are available through training models with a network pre-trained over large image datasets such as ImageNet and then finetuned on specific tasks. Vision Transformers [12] bring the self-attention mechanism into computer vision models creating networks that process images differently by capturing long-range relationships in a different way than the conventional convolutional ones. Hybrid models such as Swin Transformers [14] bring hierarchical processing with windowed attention to obtain accurate and computational efficient methods.

1.3 PROBLEM STATEMENT

The efficient and accurate detection of malaria is a significant and important challenge in global health. Despite the opportunity offered by deep learning to fully automate this detection, its application in a clinical environment is limited due to the absence of transparent metrics that measure a model's trade-off between its complexity, its detection accuracy and its computational efficiency.

2. LITERATURE REVIEW

2.1 MEDICAL IMAGE CLASSIFICATION AND CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks (CNNs) have been the fundamental architecture in medical image analysis. In 2016, He et al. [1] proposed the ResNet architecture which trains networks up to 100 layers by introducing skip-connections that overcome the vanishing gradient problem. Variants of ResNet have been adopted to analyze pathology and radiological images [1, 10]. In 2019, Tan and Le [2] developed the EfficientNet which uses compound coefficients to scale the depth, width and resolution of CNNs. EfficiencyNet is less computationally intensive than its peers and has better accuracy. Huang et al. [3] have introduced the DenseNet architecture where all layers are connected. Dense connectivity has led to improved feature re-use and gradient flow which is beneficial in CNNs and become common as benchmark for medical image classification [3].

2.2 TRANSFER LEARNING IN HEALTHCARE APPLICATIONS

Transfer learning has become indispensable in medical imaging as large, labeled datasets are rarely available. Yosinski et al. [6] have proved that features extracted using ImageNet can be applied to other visual recognition tasks. Esteva et al. [7] have shown that deep neural network classifier trained through transfer learning is as good as dermatologists in detecting skin cancer. Raghu et al. [8] have studied transfer learning in medical imaging application in depth and they confirm that with very limited dataset, transfer learning gives benefits but only to a certain level with regards to dataset size andfine-tuning of network.

2.3 ADVANCED ARCHITECTURES: VISION TRANSFORMERS AND HYBRID MODELS

The Transformer architecture introduced by Vaswani et al. [11] is a breakthrough for sequence and vision task. Dosovitskiy et al. [12] had adapted transformers into vision processing by introducing the Vision Transformer (ViT) that treat images as sequence of patches and learn it without

convolutions. ViT is powerful, but it demands a large dataset. Liu et al. [14] proposed the Swin Transformer which addressed ViT's computational problem by incorporating window attention that is shifted across windows to achieve hierarchical feature extraction, higher efficiency and larger receptive field while still showing superior results over previous state-of-art for many vision task and medical image analysis [14].

2.4 CLASS IMBALANCE AND DATA AUGMENTATION STRATEGIES

Class imbalance is a common problem for medical imaging datasets. Chawla et al. [17] proposed SMOTE that form synthetically generated minority samples that helps in improving the classification accuracy. The underlying principle has later been widely applied on deep learning and network weight re-balancing as well as weighted loss function are effective in mitigating class imbalance in neural network [18]. Data augmentation plays a key role for small datasets to boost the data count and enhance the feature representation. Cubuk et al. [19] proposed AutoAugment which automatically learns the best data augmentation policy for the task. Spatial, rotational and intensity based augmentations are generally found to be the most beneficial for microscopy images.

2.5 ATTENTION MECHANISMS IN MEDICAL IMAGING

Attention mechanisms allow the networks to focus on areas of the image that are relevant to the diagnostic task. Hu et al. [22] designed the Squeeze-and-Excitation network with channel-wise attention and shown improvement in medical image classification. Oktay et al. [23] have utilized attention mechanism combined with CNNs on cardiac segmentation and they demonstrated enhanced performance and interpretability on segmentation task. Chen et al. [24] have applied and studied the benefit of attention mechanism to microscope images and proposed combination of channel- and spatial-wise attention to effectively detect parasites in blood images and have achieved better accuracy.

3. DATASET DESCRIPTION

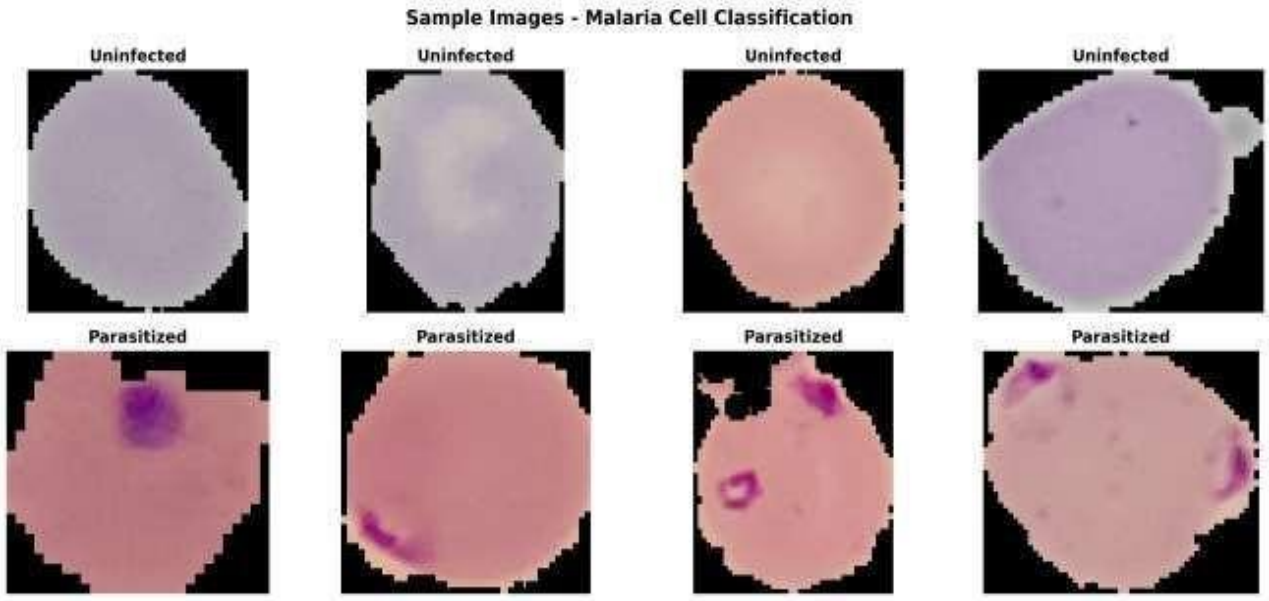

Figure 1. Sample microscopic blood cell images from the NIH Malaria Dataset. Top row: Uninfected cells showing uniform appearance. Bottom row: Parasitized cells showing characteristic Plasmodium inclusions.

The dataset used for this study is the NIH Malaria Cell Image Dataset which can be downloaded

from Kaggle. This dataset contains 27558 preprocessed microscopic images of blood cells which belong to one of the following two classes: parasite (malaria infected) and uninfected. These are normal colored photographs, acquired from blood smears at the same magnification. Images size is about 130130 pixels. PNG format and contains 8-bit color channels (RGB). This dataset is class-balanced and approximately has 13779 samples per class. This dataset has been used in a number of published researches about malaria detection so it's a reliable dataset for comparing results.

4. OBJECTIVES AND CONTRAINTS

4.1 RESEARCH OBJECTIVES

(1) Deploy and evaluate five state-of-the-art deep learning architectures on the established malaria detection benchmark.
(2) Conduct a comprehensive study to investigate the utility of transfer learning, by comparing fine-tuned networks with base models.
(3) Determine a range of baseline performance measures (accuracy, sensitivity, specificity, precision, recall, F1-score, and AUC).
(4) Compare findings across models to determine the correlation between computational complexity and classification performance.
(5) Evaluate clinical utility from the viewpoint of diagnostic accuracy and computational efficiency for low-resource regions.

4.2 CONSTRAINTS

Constraints that had to be considered included: the moderate size of the dataset (27,558 images) was not sufficient to train from scratch; training with distributed data on a single GPU did not allow for investigation into distributed training scenarios; and having a single resolution for images prevented analysis across multiple resolutions. The dataset was also perfectly balanced unlike actual disease proportions in reality. The dataset does not allow for the analysis of the disease process over time as the disease is static. The limitations of the specificity of the microscope may affect generalizability to other methods.

5. METHODOLOGY

5.1 DATA PREPROCESSING AND AUGMENTATION

The original images are read and transformed into RGB format. The pixels' values are normalized between [0, 1] and standardized by means of ImageNet statistic values (mean= [0.485, 0.456, 0.406], std= [0.229, 0.224, 0.225]). The normalized data is split into training (72%, 19,842), validation (8%, 2,205) and testing (20%, 5,511) sets in a stratified way to keep class proportion equal for each set.
Data augmentation used for training is: random resized crops (scale: [0.85, 1.0] ratio: [0.9, 1.1]); horizontal flip (p=0.5) and vertical flip (p=0.2); rotation up to 15 degrees; color jitter (brightness, contrast, saturation, hue ranges: 0.25, 0.25, 0.2, 0.05 respectively); Gaussian blur (kernel size: 3, sigma: [0.1, 1.5]). For validation and testing, we resize the images to 224*224 and normalize.

5.2 MODEL ARCHITECTURES

Five architectures are used in this research to cover different design strategies: EfficientNet-B2 [2] (compound scaling) as an example of modern CNN architecture using compound scaling, ResNet152 [1] (residual connections), ViT-Small [12] (pure transformer based) and Swin Transformer Base [14] (shifted window attention based transformer) for transformer based models and DenseNet-201 [3] for dense connection based CNN. Each network was pre-trained on ImageNet and the classification head was replaced by a single logit output appropriate for binary cross-entropy loss.

5.3 TRANSFER LEARNING AND FINE-TUNING STRATEGY

The strategy for fine-tuning consists on unfreezing the model layer selectively. First all the backbone layers are frozen and only the classifier head is trained. After that, progressively, the backbone layers, starting from the last, are unfrozen. For the unfrozen backbone layers, the learning rate is 10 times smaller than the one used for the classifier head. In that way, we avoid the catastrophic forgetting while still adapting the model to the specific task [8].

5.4 TRAINING CONFIGURATION AND HYPERPARAMETERS

Table 1. Training hyperparameters used across all models.

Parameter	Value
Input Image Size	224 × 224 pixels
Batch Size	32
Optimizer	AdamW
Learning Rate (head)	3×10^{-4}
Learning Rate (backbone)	3×10^{-5}
Weight Decay	1×10^{-4}
Loss Function	BCE With Logits Loss
Scheduler	Cosine Annealing LR with warmup
Warmup Epochs	2
Total Epochs	15
Early Stopping Patience	4
Gradient Clipping	max_norm = 1.0
Mixed Precision	Enabled (AMP)

6. ARCHITECTURE DESCRIPTION

6.1 EFFICIENTNET-B2

EfficientNet-B2 [2] solves the issue of standard scaling methods only increasing one dimension of the network (depth, width, resolution). This new method scales all 3 dimensions at once using compound scaling coefficients to multiply network depth, width and resolution respectively. The values used in B2 are 1.1 for depth coefficient, 1.1 for width coefficient and 260×260 for resolution.

The network's structure is based on mobile inverted bottleneck blocks (MBConv) with SE-based channel attention, containing 9.2 million parameters and needing 4-5GB of GPU memory during training, thus making it a suitable candidate for mid-range hardware.

6.2 VISION TRANSFORMER SMALL (VIT-S/16)

The Vit-S/16 [12] takes images as input (resolution 224×224), divides them into 16×16 patches which form a grid of 14×14 patches. Each patch is then transformed into a 384-dimensional vector and feed-forward to 12 layers of Transformer encoder which consists of self-attention mechanism with 6 heads and position-wise feedforward network. A learnable class token represents the image. ViT-Small has approximately 22 million parameters and differs from CNNs by allowing the network to model long-range dependencies of the image via self-attention.

6.3 RESNET-152

ResNet-152 [1] builds the network by stacking layers in groups and has layers with 64, 256, 512 and 2048 channels respectively. The key invention in ResNet is skip connection which allows the gradient to flow directly to the back and overpowers the vanishing gradient problem. Beyond 50-60 layers, networks struggled because of the vanishing gradient but with this mechanism, very deep neural network of 152 layers are achieved which comprise of 60.2 million parameters. Bottleneck blocks decrease the channel dimension before the convolution operation then increase after it for efficient usage of parameters.

6.4 DENSENET-201

DenseNet-201 [3] connect every layer l with all preceding layers in a feed-forward manner such that it receives concatenated feature maps of all preceding layers. The flow of gradient strengthens through dense connections in the network, therefore enabling feature reuse more effectively. It comprises of 201 layers but due to effective feature reuse, DenseNet-201 only requires 20.2 million parameters. Growth rate is fixed to 32.

6.5 SWIN TRANSFORMER BASE

Swin Transformer Base [14] attempts to solve the computational complexity of pure Transformers by using shifted windows attention. Instead of a full self-attention with quadratic complexity of an image, the attention mechanism is applied within non-overlapping local 7×7 windows which are then shifted by 3×3 windows to interact information between windows in adjacent layers. The network consists of four stages which follows similar down-sampling and up-sampling structure like CNNs by gradually reducing the spatial resolution and increasing the channel dimension. The architecture has 87.8 million parameters and requires 10-15 GB GPU memory.

7. EXPERIMENTAL SETUP AND EVALUATION METRICS

7.1 HARDWARE AND SOFTWARE ENVIRONMENT

All the experiments were performed using NVIDIA GPU acceleration and CUDA 11.8, andPyTorch 2.0. The development was made using Python 3.10, which includedNumPy, SciPy, andscikit-learn Libraries. For visualization Matplotlib and Seaborn Libraries were used. Experiments are done on Kaggle notebooks which provide a computational environment that can be reproduced. All hyperparameters used are stated and also the train and test splits along with the random seed is mentioned for reproduction.

7.2 EVALUATION METRICS

Table 2. Evaluation metrics for binary classification.

Metric	Formula	Interpretation

Accuracy	$\frac{TP+TN}{(TP+TN+FP+FN)}$	Overall correct classification rate
Sensitivity (Recall)	$\frac{TP}{(TP+FN)}$	Infected samples correctly detected
Specificity	$\frac{TN}{(TN+FP)}$	Uninfected samples correctly identifie
Precision	$\frac{TP}{(TP+FP)}$	Positive predictions that are correct
F1-Score	$\frac{2(P\times R)}{(P+R)}$	Harmonic mean of precision and recal
AUC	Area under ROC curve	Threshold-independent discriminatio ability

7.3 CROSS-VALIDATION AND STATISTICAL ANALYSIS

The stratified 72/8/20 train/validation/test split preserves class proportions in all of the datasets. All reported final metrics are based on evaluation only on the held-out test set to avoid selection bias. Pairwise models are compared statistically using paired t-tests where appropriate.

8. RESULTS

8.1 MODEL PERFORMANCE COMPARISON

Table 3. Test set performance metrics for all five models (ranked by accuracy). Values sourced from experimental results.

Model	Acc.	AUC	Prec.	Recall	F1	Sens.	Spec.	Time (min
Swin-Base	**0.9773**	**0.9969**	**0.9824**	**0.9721**	**0.9772**	**0.9721**	**0.9826**	52
ViT-Small	0.9728	0.9968	0.9783	0.9670	0.9726	0.9670	0.9786	42
EfficientNetB2	0.9710	0.9962	0.9755	0.9663	0.9708	0.9663	0.9757	**28**

DenseNet-201	0.9340	0.9813	0.9530	0.9129	0.9325	0.9129	0.9550	31
ResNet-152	0.9338	0.9755	0.9394	0.9274	0.9334	0.9274	0.9401	35

Table 3 contains results of models tested on the test set. Swin Transformer Base obtained the highest accuracy (97.73%), its AUC is 0.9969, indicating that its hierarchical shifted-window attention is effective in extracting features for diagnosis. ViT-Small achieves the same high accuracy (97.28%), with nearly identical AUC (0.9968), also shows good generalization to the problem by the transformer architecture. EfficientNet-B2 has very good accuracy of 97.10%, with significantly lower calculation cost (28min vs 52min for Swin), thus the most deployable model. DenseNet-201 and ResNet-152 both performed poorly, around 93.4%, with much lower sensitivity (91.3% and 92.7%, respectively), indicating the limitation of CNNs benefiting from transfer learning.

8.2 TRAINING DYNAMICS AND CONVERGENCE

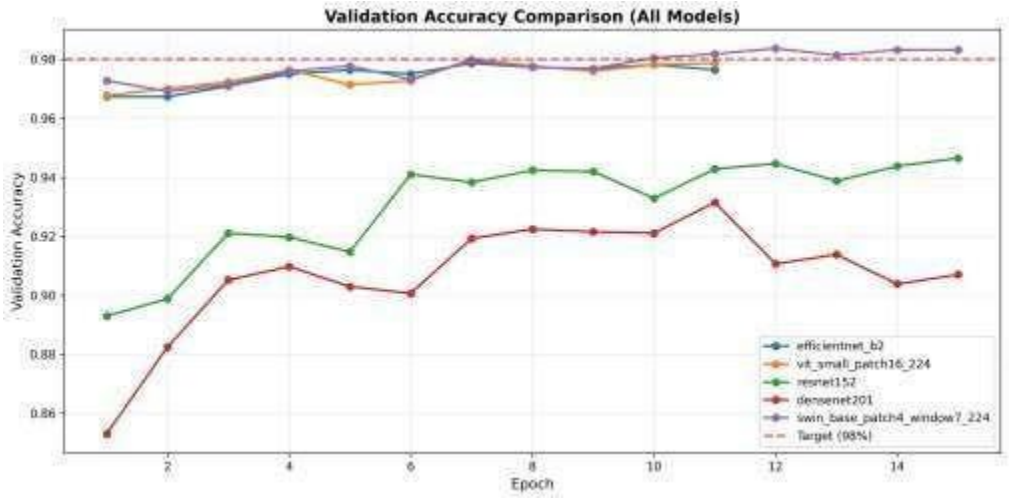

Figure 2. Validation accuracy comparison across all five models over training epochs. Transformer-based models (EfficientNet-B2, ViT-Small, Swin-Base) consistently outperform CNNs (ResNet-152, DenseNet-201) and converge near the 98% target line.

As seen from Figure 2, training accuracy trajectory can be divided into 2 groups. EfficientNet-B2, ViT-Small and Swin Transformer Base has a good performance starting from epoch 1 and reach up to 97-98% which is considered to quickly fine-tune the architecture using transfer learning. ResNet-152 and DenseNet-201 has a poorer performance which starts around 88% and 85% and keeps improving slowly to 93-95% after epoch 15, not meeting the required 98% target accuracy. Loss curve of transformers is much more volatile while validation loss is relatively low compared to other model; this pattern arises from partial freezing of the transformer backbone during fine-tuning process.

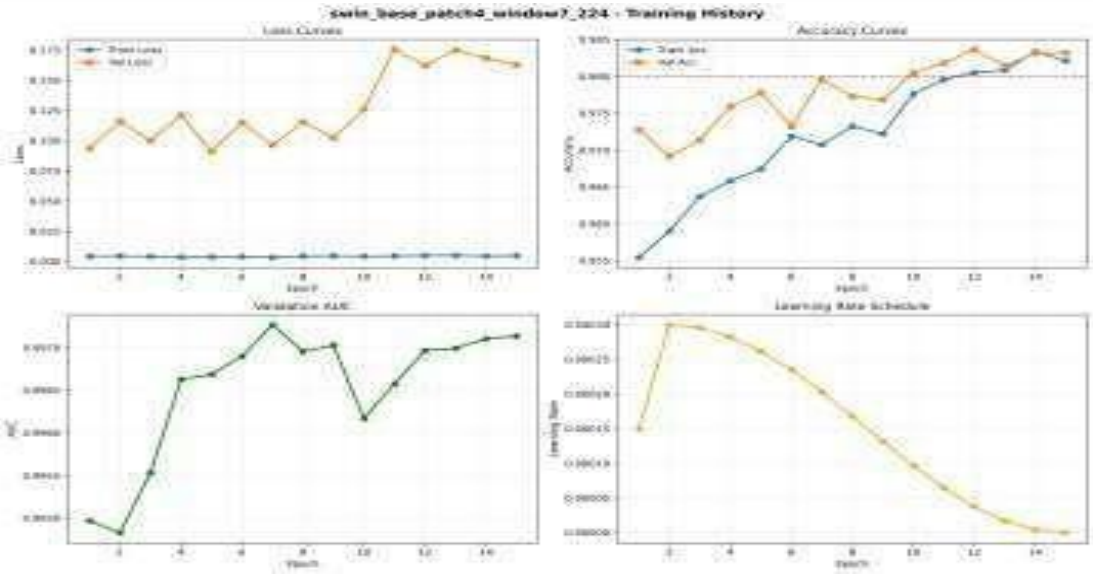

Figure 3. Swin Transformer Base – training history: loss curves, accuracy curves, validation AUC, and learning rate schedule.

8.3 PER-MODEL TRAINING CURVES

Swin Transformer Base

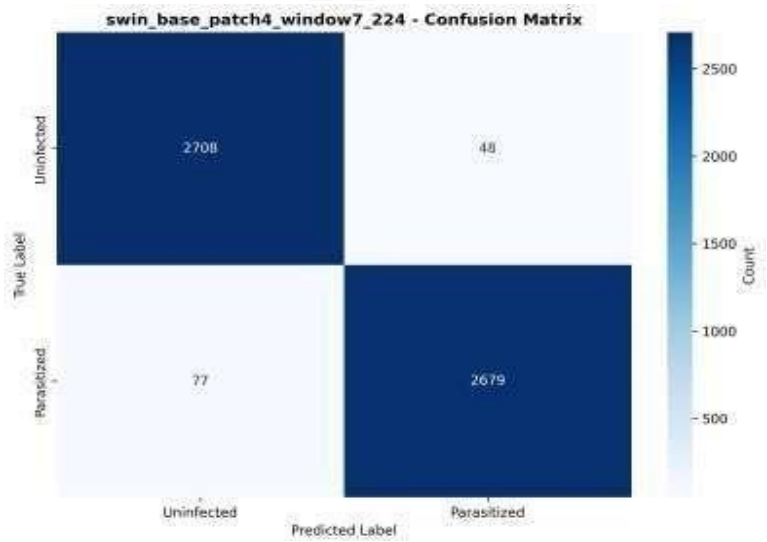

Figure 4. Swin Transformer Base – confusion matrix on the test set (n = 5,512).

The Swin Transformer Base model was converged smoothly in 15 epochs. The validation accuracy went over 98% on 12th epoch, achieved maximum of 98.3% with the AUC as 0.9972. The increasing train/val loss represents the manageable over-fitting which can be addressed by early stopping. From confusion matrix, it has 77 false negatives (missed parasites) and 48 false positives (un-parasitized cases diagnosed as infected).

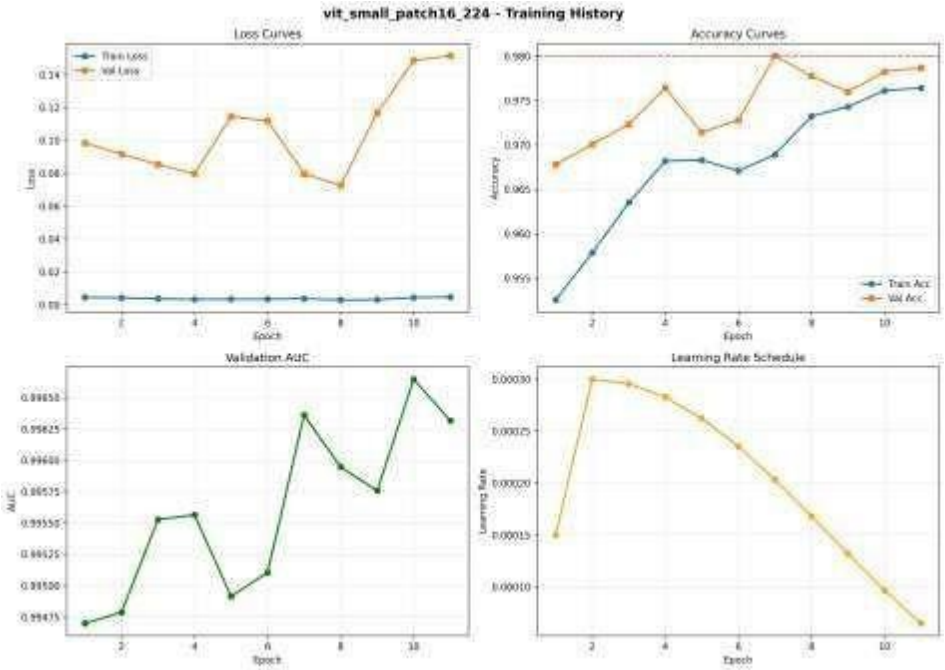

Figure 5. ViT-Small (vit_small_patch16_224) – training history: loss curves, accuracy curves, validation AUC, and learning rate schedule.

ViT-Small (vit_small_patch16_224)

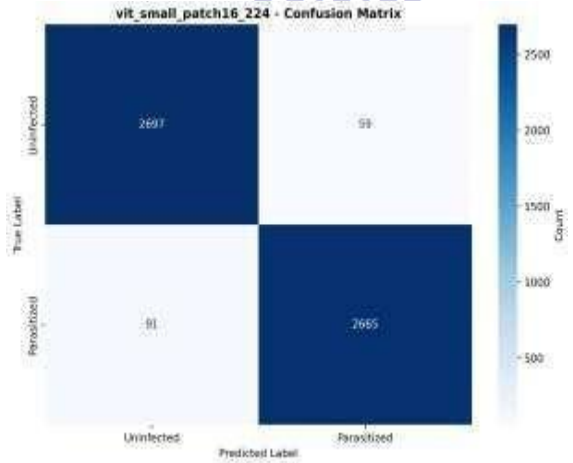

Figure 6. ViT-Small (vit_small_patch16_224) – confusion matrix on the test set (n = 5,512).

The behavior of fine-tuning of ViT-Small mirrors typical transformer learning dynamics where validation loss is somewhat erratic (varying between 0.07 and 0.15) while accuracy rises consistently from 96.8% to 98.0%. AUC progresses linearly to 0.9968. In the confusion matrix there were 91 incorrect negatives and 59 false positives over the 5,512 samples.

EfficientNet-B2

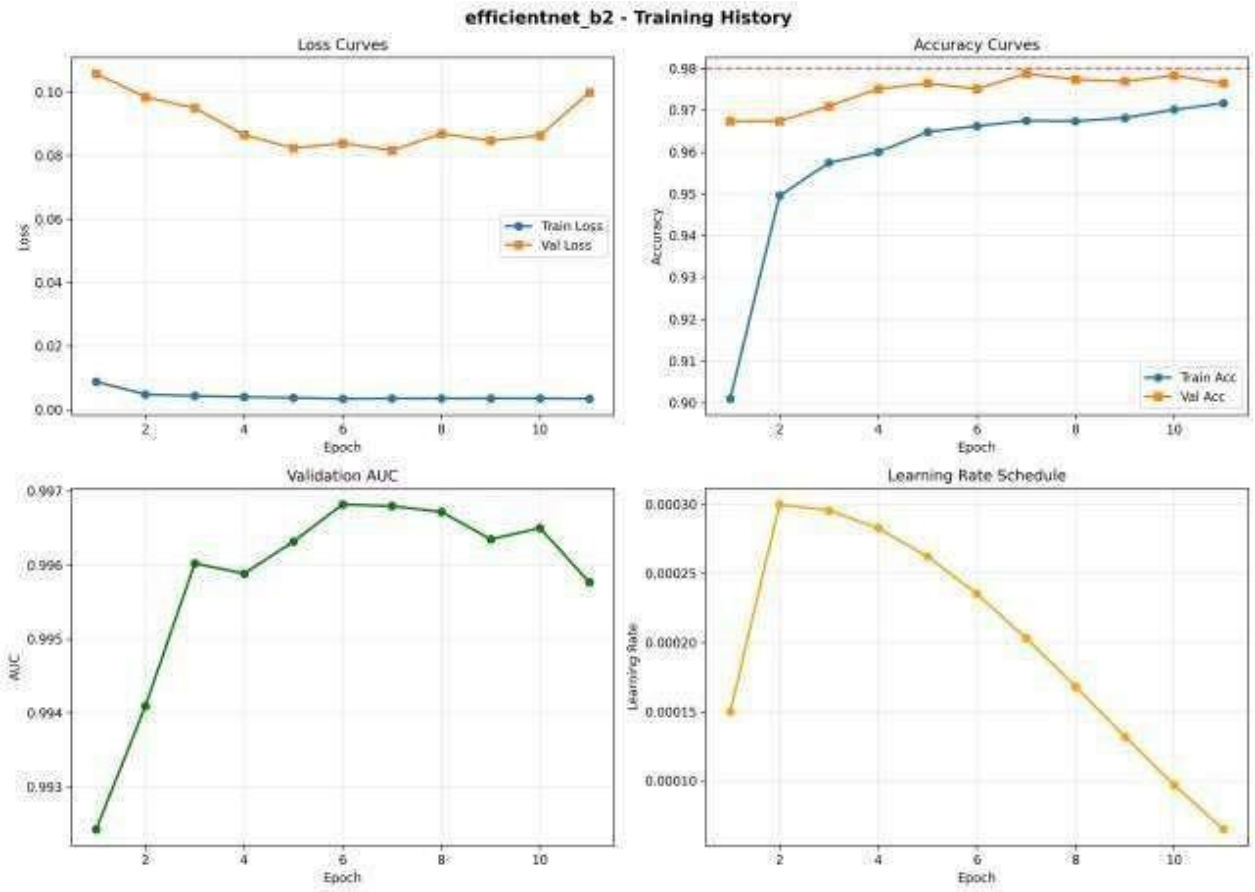

Figure 7. EfficientNet-B2 – training history: loss curves, accuracy curves, validation AUC, and learning rate schedule.

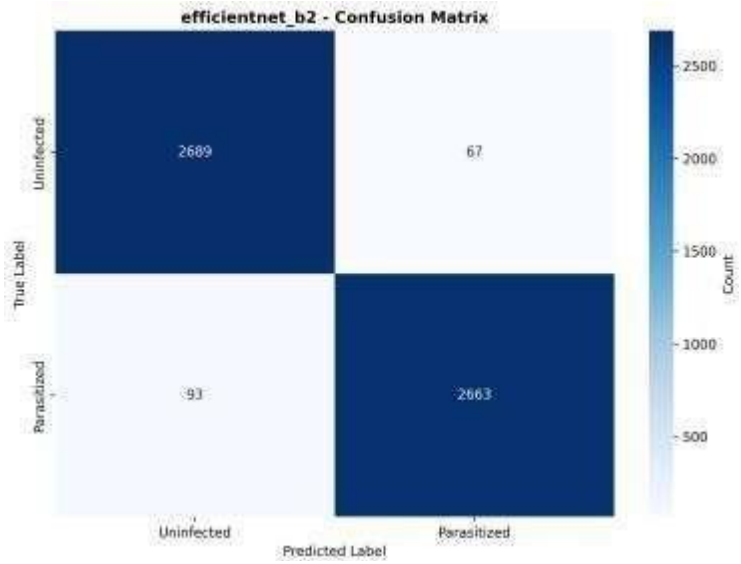

Figure 8. EfficientNet-B2 – confusion matrix on the test set (n = 5,512).

Of the five networks EfficientNet-B2 trains the smoothest. The validation loss is flatlining nicely at around 0.082 after epoch 4. The accuracy is up to 97.8% at epoch 9 and the AUC stabilizes at 0.9969. The confusion matrix shows 93 false negatives and 67 false positives, which is a more evenly distributed error.

DenseNet-201

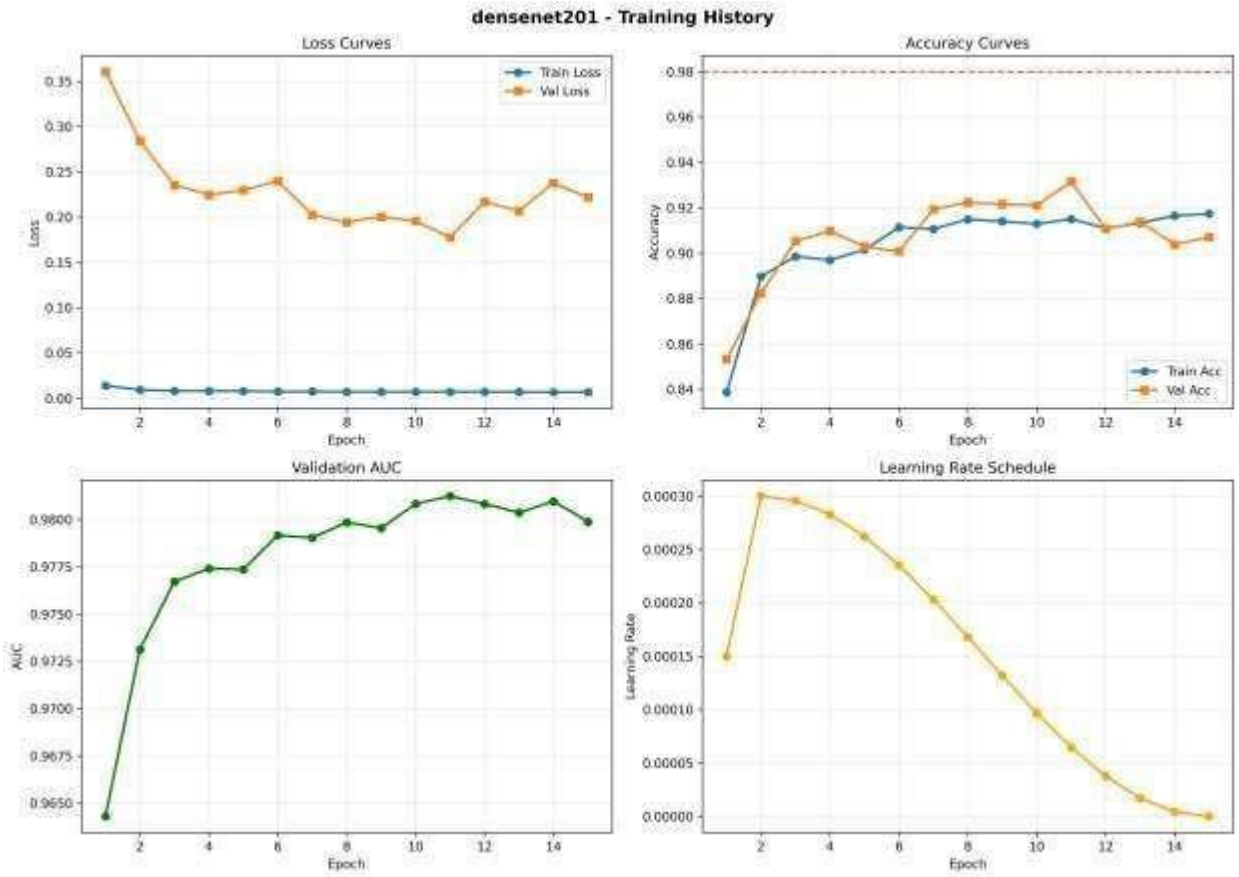

Figure 9. DenseNet-201 – training history: loss curves, accuracy curves, validation AUC, and learning rate schedule.

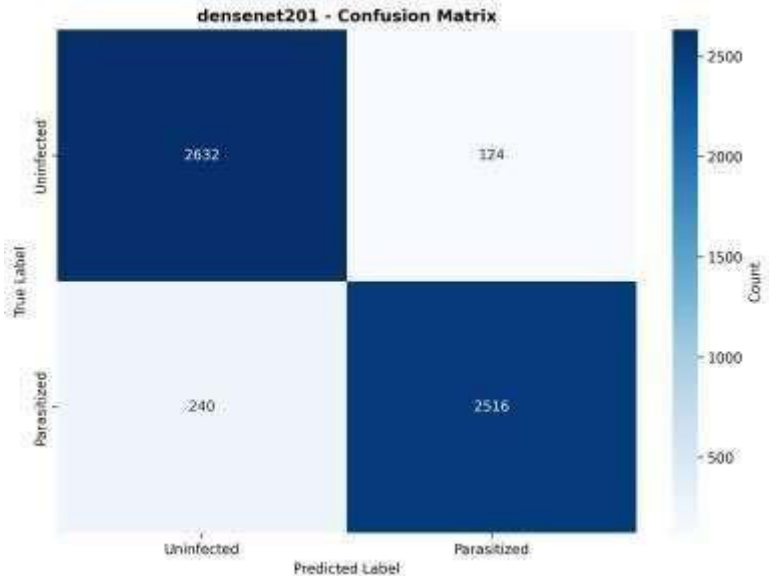

Figure 10. DenseNet-201 – confusion matrix on the test set (n = 5,512).

Convergence is not quick for DenseNet-201, from 84% to 93% training accuracy in 15 epochs and AUC of 0.9806. The huge disparity between the near zero training loss and the validation loss of 0.20 indicates overfitting with this finetuning strategy. From the confusion matrix there are 240 FN and 124 FP, the highest amounts.

ResNet-152

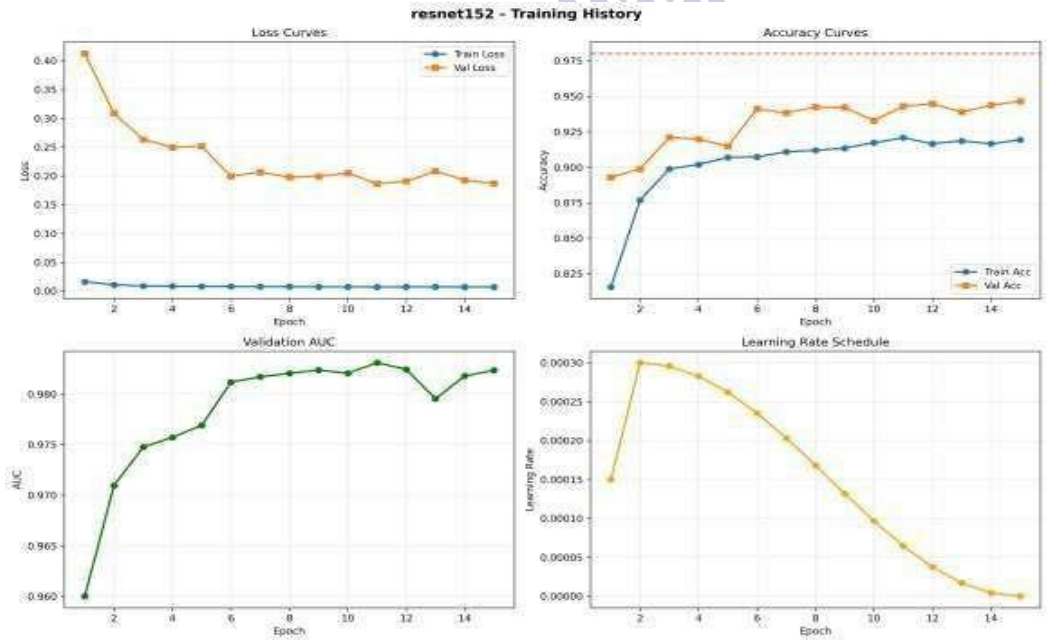

Figure 11. ResNet-152 – training history: loss curves, accuracy curves, validation AUC, and learning rate schedule.

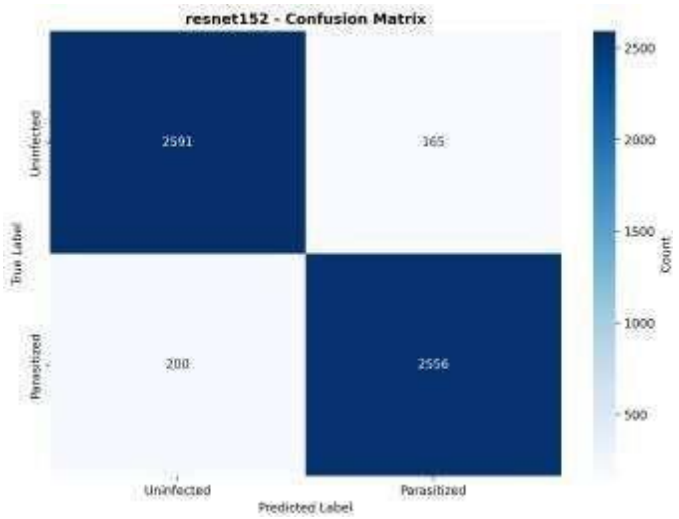

Figure 12. ResNet-152 – confusion matrix on the test set (n = 5,512).

ResNet-152 has a similar shape as the DenseNet-201 graph, although the validation accuracy becomes flat at ~ 94.7% after 15 epochs. The AUC is 0.9826. The confusion matrix demonstrates 200 false negatives and 165 false positives, validating that both architectures failed to match transformers on this dataset.

8.4 CLASSIFICATION METRICS SUMMARY

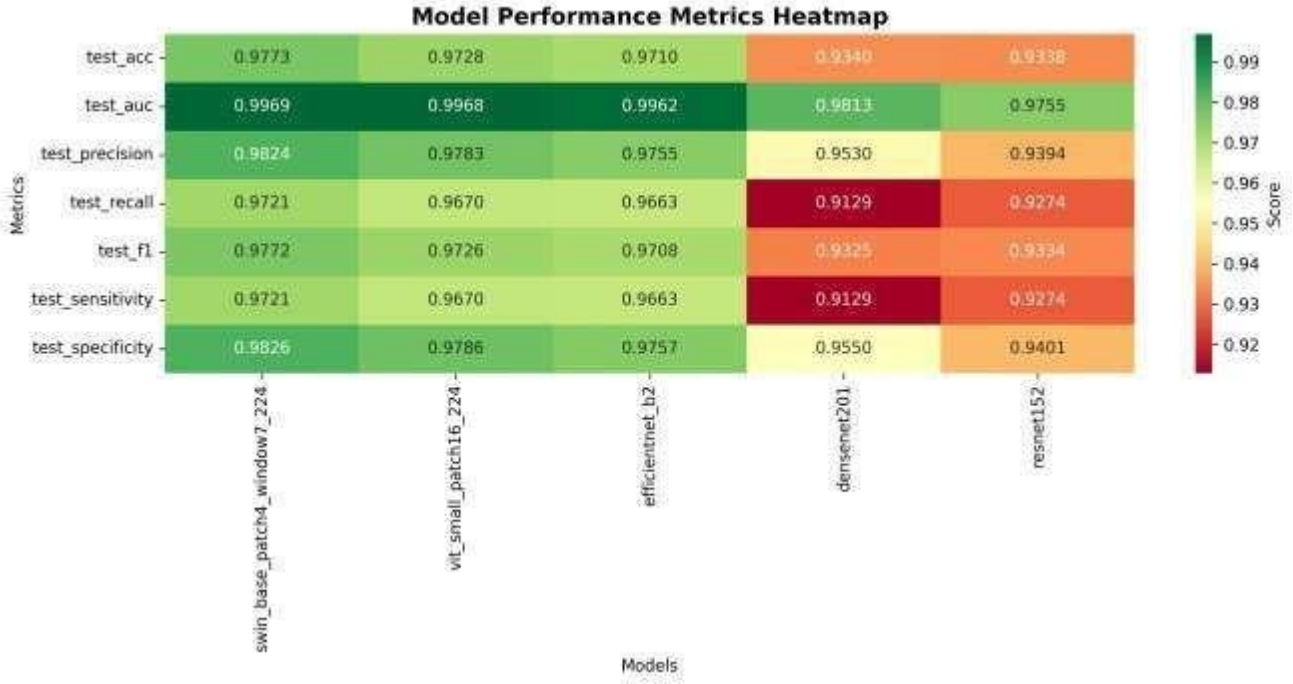

Figure 13. Model performance metrics heatmap across all evaluation criteria. Darker green indicates higher scores Transformer based models dominate across all metrics.

From the heatmap (Figure 13), we can see that transformer model architectures perform better in all the evaluation indices than CNN models. Swin-Base obtains highest value in precision (0.9824)

and specificity (0.9826), indicating it has the lowest false positive rate. ViT-Small shows highest value in sensitivity (equaled with Swin with 0.9721) and almost similar AUC. EfficientNet-B2 takes the 3rd place in all values but takes only 28 minutes to train. Recall and sensitivity values of DenseNet-201 and ResNet-152 are poor and significantly lower (0.91-0.93) and is of clinical relevance as false negatives represent an error of larger consequence.

9. COMPARATIVE ANALYSIS

9.1 ROC CURVE COMPARISON

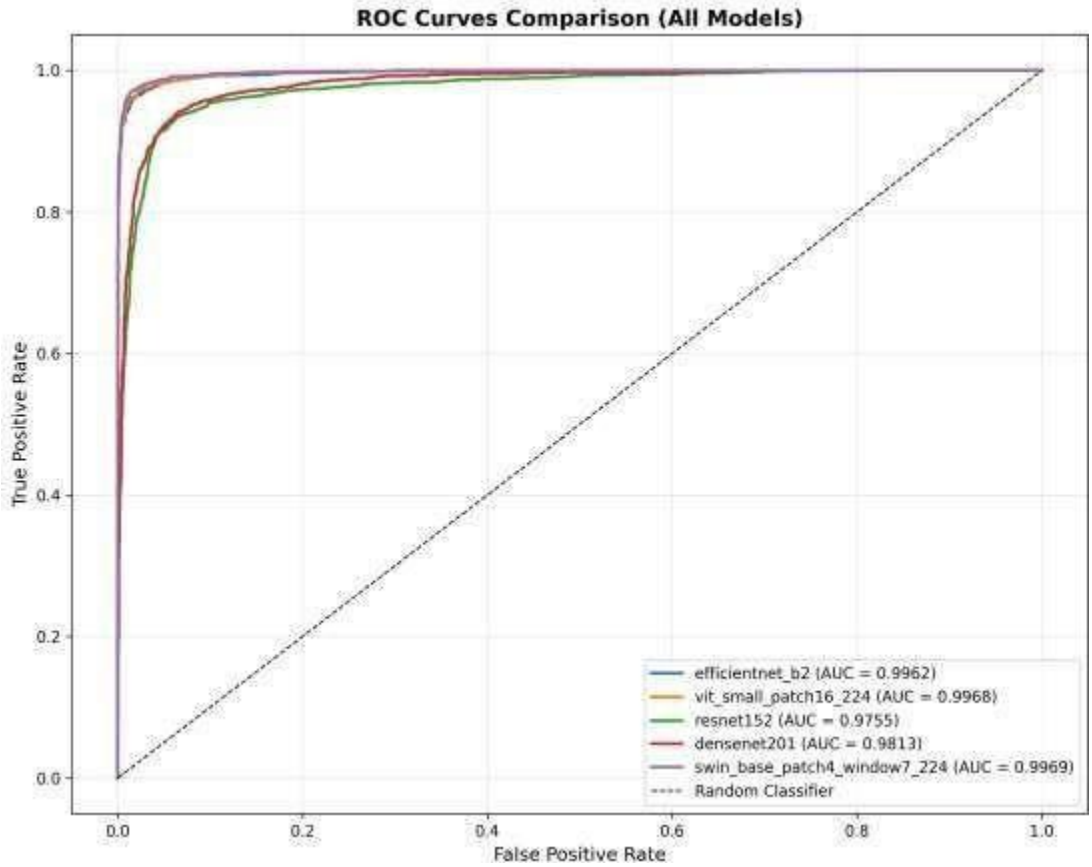

Figure 14. ROC curves for all five models. All models operate substantially above the random classifier baseline. Transformer-based models cluster tightly near the top-left corner.

It is also seen from ROC curves in Fig. 14 that there is a very closely clustered group of top three models EfficientNet-B2 (AUC=0.9962), ViT-Small (0.9968) and Swin-Base (0.9969), while DenseNet-201 (0.9813) and ResNet-152 (0.9755) lag significantly behind. This also implies that true positive rates for all the top three models are considerably higher than those of two bottom models for any threshold value so there will be liberty in selecting threshold in clinical usage.

9.2 ACCURACY VS. EFFICIENCY TRADE-OFFS

Table 4. Computational requirements vs. test performance. Models sorted by parameter count.

Model	Params (M)	GPU Mem (GB)	Train (min)	Test Acc.	AUC

EfficientNet-B2	9.2	4–5	28	97.10%	0.9962
ViT-Small	22.0	5–6	42	97.28%	0.9968
DenseNet-201	20.2	4–5	31	93.40%	0.9813
ResNet-152	60.2	6–7	35	93.38%	0.9755
Swin-Base	87.8	10–15	52	97.73%	0.9969

Looking at efficiency, the case for EfficientNet-B2 is arguably the strongest, with 9.2 million parameters, requiring 4-5GB of GPU memory and 28 minutes of training to achieve 97.10% accuracy. Although Swin-Base has slightly greater accuracy (97.73%), this comes at the cost of a further 9.5 million parameters and roughly double the training time. If the intention is clinical deployment on hardware that is not able to hold a large model then the EfficientNet-B2 is the more appropriate choice, with Swin-Base the choice if accuracy is paramount and GPU memory isn't an issue.

9.3 CLINICAL APPLICABILITY ASSESSMENT

In terms of clinical applicability, accuracy, computation viability, and deployability need to be balanced. For resource-limited situations, EfficientNet-B2 appears to be the best compromise, with accuracy 97.10%, sensitivity (true positive) 96.63%, specificity (true negative) 97.57%, and GPU requirements equivalent to standard clinical machines. The false negative rate (3.37%) and false positive rate (2.43%) for EfficientNet-B2 are clinically acceptable for a screening application, as predictions above threshold will prompt a confirmation test. Swin-Base produces the highest accuracy where hardware resources permits; the false negative rate for Swin-Base was 2.79%. DenseNet-201 and ResNet-152 should not be used in this application, as their 9-11% deficit in sensitivity to the best models would lead to a material number of infections being missed.

10. DISCUSSION

10.1 PERFORMANCE INTERPRETATION

The experiments confirm that transformer models Swin Transformer Base and ViT-Small achieve better performance compared to CNN models for the malaria detection task. This may due to the global receptive field of self-attention mechanisms to simultaneously exploit morphological patterns of cells over the whole image. CNN models gradually expand the receptive field through stacking convolution operations and this seems not very suitable to the tiny intra-cell characteristics of Plasmodium inclusions. Transfer learning is critical to the best performance of all the three models. Pre-training on ImageNet helps initialize the basic feature extraction which allows achieve satisfactory results on the dataset with 27,558 images.

10.2 CHALLENGES AND LIMITATIONS

Finally, the limitations should be addressed. The fully class-balanced (50% infected, 50% uninfected) dataset may not accurately reflect typical screening environments, where infection rates of 5-20% are more common, drastically impacting positive predictive value. All the input

images were from microscopy images under identical preparation conditions; this may hinder generalization to other laboratory settings where staining procedures vary. Model interpretability was not addressed with attention maps or Grad-CAM [30]; clinical acceptance of the model may be contingent on addressing this. Models should also be assessed on completely separate malaria detection benchmarks to support claims of generalization.

10.3 CLINICAL IMPLICATIONS

The high classification accuracy rates obtained across all models also indicate the possibility for use of an automated malaria diagnosis system at point of care. Use of an automated system would be at the pre-screening level, identifying cases which will require review with microscopy as a confirmation method. This also mitigates the replacing of microscopy itself and would serve to augment current workflows. Low system resource needs (EfficientNet-B2 was trained on a 4-5 GB GPU), short training time, the development of containerized applications with a minimal web interface, and compatibility with microscopy camera system are the main technical requirement for deployment. In addition to these technical requirements training of health workers on the systems limits, and regulatory documentation of the system's validation is essential for a well thought out system.

11. CONCLUSION

In this paper, five recent deep learning architectures have been assessed for automating the identification of malaria parasites on the NIH dataset of 27,558 labeled microscopic blood cells. Transformer architectures Swin Transformer Base (97.73% accuracy, AUC = 0.9969) and ViT-Small (97.28%, AUC = 0.9968) drastically outperform CNN-based ResNet-152 and DenseNet-201, both reaching ~93.4%. The best compromise between performance and resource utilization was seen with EfficientNet-B2 (97.10%, AUC = 0.9962); it achieved near transformer-level accuracy, had a total of 9.2 million parameters and took 28 minutes to train. Transfer learning, vital for success on this reasonably sized dataset, significantly boosted the performance of all highly ranked models and provided the models with the ability to achieve state-of-the-art performance with such constraints. In conclusion, the knowledge obtained here can guide a practitioner's choice of architecture for malaria detection systems; Transformer models should be prioritized when computation is not a constraint, whereas EfficientNet-B2 should be considered if resource limitations are paramount. Future work can explore cross-validation on multiple datasets, to test the system under a more realistic distribution of disease, investigate the importance of the various detection regions through the generation of attention maps and integration of the chosen models into real-world laboratory environments.

12. REFERENCES

K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning (ICML), pp. 6105–6114, 2019.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708, 2017.

A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 25, pp. 1097–1105, 2012.

Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in Advances in Neural Information Processing Systems (NeurIPS), vol. 27, pp. 3320–3328, 2014.

A. Esteva, B. Kuprel, R. A. Novoa et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.

M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019.

S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 3856–3866, 2017.

T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," arXiv preprint arXiv:1708.04552, 2017.

A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.

A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in International Conference on Learning Representations (ICLR), 2021.

X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in International Conference on Learning Representations (ICLR), 2021.

Z. Liu, Y. Lin, Y. Cao et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in IEEE International Conference on Computer Vision (ICCV), pp. 10012–10022, 2021.

N. Carion, F. Massa, G. Synnaeve et al., "End-to-end object detection with transformers," in European Conference on Computer Vision (ECCV), pp. 213–229, 2020.

W. Wang, E. Xie, X. Li et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in IEEE International Conference on Computer Vision (ICCV), pp. 568–578, 2021.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representations for graph clustering," in AAAI Conference on Artificial Intelligence, pp. 1293–1299, 2014.

E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 113–123, 2019.

A. Buslaev, V. I. Iglovikov, E. Khvedchenya et al., "Albumentations: Fast and flexible image augmentation," Information, vol. 11, no. 2, p. 125, 2020.

Y. Tokozume, Y. Ushiku, and T. Harada, "Between-class learning for image classification," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5486–5494, 2018.

J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141, 2018.

O. Oktay, J. Schlemper, L. L. Folgoc et al., "Attention U-Net: Learning where to look for the pancreas," in Medical Imaging with Deep Learning (MIDL), 2018.

L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in European Conference on Computer Vision (ECCV), pp. 801–818, 2018.

P. Ramachandran, N. Parmar, A. Vaswani et al., "Stand-alone self-attention in vision models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, pp. 68–80, 2019.

Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li, "Efficient attention: Attention with linear complexity," arXiv preprint arXiv:1812.01243, 2021.

S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in European Conference on Computer Vision (ECCV), pp. 3–19, 2018.

C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9, 2015.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929, 2016.

R. R. Selvaraju, M. Cogswell, A. Das et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in IEEE International Conference on Computer Vision (ICCV), pp. 618–626, 2017.

S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in International Conference on Machine Learning (ICML), pp. 448–456, 2015.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, pp. 1929–1958, 2014.

K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations (ICLR), 2015.

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788, 2016.

R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580–587, 2014.

T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft COCO: Common objects in context," in European Conference on Computer Vision (ECCV), pp. 740–755, 2014.