Malaria Detection from Blood Smear Images using Vision Transformer and Deep Learning Technique

Akriti Singh
Department of Computer Science & Engineering
Amity University Uttar Pradesh, India

Syed Wajahat Abbas Rizvi
Department of Computer Science & Engineering
Amity University Uttar Pradesh, India

Divya Srivastava
Department of Computer Science & Engineering
Amity University Uttar Pradesh, India

ABSTRACT
Malaria diagnosis by use of automated systems in place of interpreted microscopic blood smear images can be used to increase the screening throughput and reduce diagnostic result variation. Four deep learning models are compared and analysed on a balanced dataset of 27,560 labelled cell images (13,780 Parasitized, 13,780 Uninfected), which were divided into training (80) and test (20) datasets, in this study. Two tailor-made convolutional neural networks (CNNs), Mo-bileNetV2 with transfer learning, and a Vision Transformer (ViT) are the models. The initially trained CNN, which was trained on 64x64 and most common types of augmentation, had a test accuracy of 96.06%. Mo-bileNetV2 was tested at 93.56% test accuracy on 128x128 images and has been demonstrated to be a lightweight alternative whilst being pretrained on ImageNet and a custom head. A deeper CNN, the second, was trained with 20 epochs with learning rate scheduling and regularization methods, yielding an accuracy of 94.32% on the test set, as well as on classification, and the accuracy during the validation varied over the training. ViT model was trained on 224x 224 images and optimized with Adam production which provided the best results 97.59 test accuracy. Finally, the great-est accuracy was achieved by attention-based models, yet by use of a CNN architecture, models were still competitive, and with less computationally costly, could permit repartitioning of computational resources and make a viable diagnostic choice in low-resource contexts. It is a complete reference to the classification of malaria cells and allusions to the choice of the model in situations of biomedical image classification.

Keywords
CNN, ViT, Mobilenet, Transfer learning, Transformer, Models

1. INTRODUCTION
Malaria remains a serious health problem throughout the world, and sub-Saharan Africa is the area in which it is experienced most. As a matter of fact, the area is the cause of approximately 95 percent of all cases of malaria and almost 96 percent of deaths associated with the disease [1, 2]. Children under the age of five and pregnant women are the most vulnerable groups and face the most adverse consequences most of the time [3]. Due to this, early and correct diagnosis gets very crucial. Unluckily, the currently existing standard practices, such as microscopy and rapid diagnostic tests (RDTs), are not problem-free. They tend to rely on trained specialists, sometimes are influenced by the human factor, and do not always work well in the case of parasites in the blood which are very low [4, 5]. The above limitations affirm the fact that there is a great demand to have the superior diagnostic tools that are reliable, fast, and can be used even in areas with limited resources [6].Over the last several years, Artificial Intelligence (AI) and Machine Learning (ML) have begun to have their significant role in medical image analysis. As an example, Convolutional Neural Networks (CNNs) are especially effective at detecting spatial features in images, so it is an excellent candidate to detect malaria parasites in blood smears in a blood smear image [7, 8, 9]. Vision Transformers (ViTs) on the other hand harness self-attention to identify wider areas of relationships within an image that can serve to supplement the local feature extraction of CNNs [10]. The downside however is that ViTs are typically more memory-intensive and computationally expensive to operate well, making their application in low-resource settings difficult [11, 12]. The combination of CNNs and ViTs can thus be a decent compromise, as it would be efficient and accurate in diagnosis. In our project, we have created Malaria Diagnosis System (MDS) which will be largely based on CNNs to detect parasites in blood smear images. The primary rationale was to minimize the dependence of the process on expert analysis, but provide reliable results. We trained and tested the system on 27,558 images which had an accuracy of 94.32. This result demonstrates that this method might be quite beneficial to practice, particularly in a healthcare environment with a restricted number of resources and skilled employees.

2. RELATED WORK
A number of scholars attempted to employ AI and deep learning to assist in detecting malaria. Other earlier approaches were more conventional and applied machine learning algorithms like Support Vector Machines (SVMs) to isolate infected and uninfected cells [13]. As an example, Dong et al. [14] ran some popular CNNs such as LeNet, AlexNet, and GoogleNet using a relatively small dataset of approximately 2,565 images. Among them, GoogleNet provided the highest accuracy, which was cited at 98.13. Liang et al. [15] operated slightly in different idea by combining AlexNet features with the use of SVM classifier and their methodology offered more accuracy, sensitivity and specificity in comparison to utilising CNNs alone. A different study conducted by Bibin et al. [16] constructed a six-layer deep belief net-work which attained an accuracy of 96.4% on approximately 4,100 red blood cell pictures.

To differentiate between parasitized and healthy cells, Shaik et al. [17] employed the CNN consisting of three convolution layers and two fully connected layers. Their model achieved an accuracy of 95.7, and interestingly, ResNet-50 features performed better than other sets of features that they experimented with. Prasad et al. [18] then proceeded to design an even deeper 19-layer CNN and VGG-16 transfer learning as well, allowing them to achieve a top accuracy of 97.77%. The hybrid CNN design can be highly effective as demonstrated by AOCT-NET [19] and some other transfer learning models with

approximately 18 layers. Some subsequent works also attempted pretrained models including VGG-16 and VGG-19, and custom CNNs and the vast majority of them always achieved high accuracy between about 95% and 98% on various datasets [20, 21].

3. METHODLOGY

3.1 Dataset Description

The data required in this research was obtained on Kaggle and consists of 27,560 pictures altogether, half of which are parasitized and the other half uninfected cells (Figure 1).

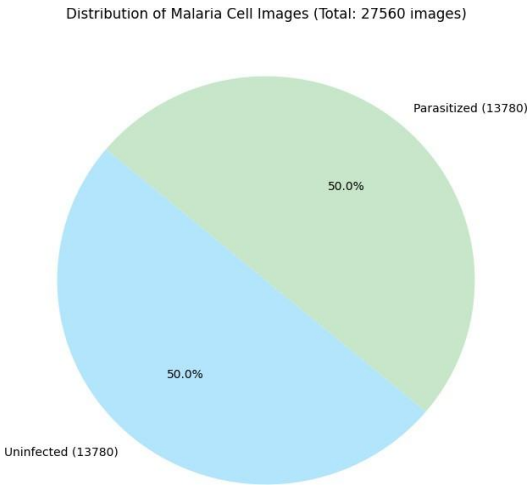

Figure 1. Demonstrates the distribution of the cell images of malaria by classes indicating equal distribution of parasitized and uninfected cells

In order to train and test the model, we separated the data into 22,048 images (80 percent) to train and 5,512 images (20 percent) to test (Figure 2). With this, the model possessed sufficient data to learn with and at the same time had a separate data to evaluate without bias. A similar balance of classes was also applied in the two subsets.

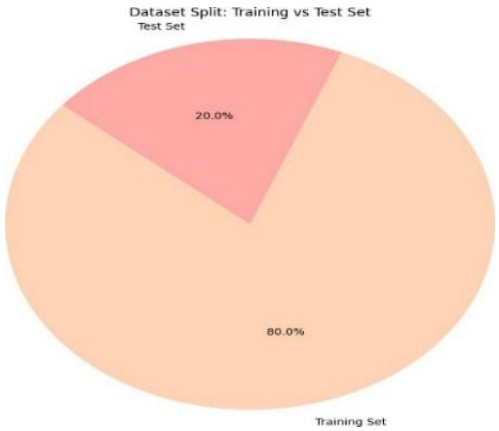

Figure 2. illustrates the division of training and test sets with a great focus on 80/20 division when training and assessing the models

3.2 Data Preprocessing

The data we have used in the current study is the one on Kaggle and contains about 27,560 red blood cells images. Approximately one-half get parasitized, and the remaining half are not parasitized. To train and test, we divided the data into 22,048 and 5,512 images, respectively, which approximately is 80/20, where a portion of the data was separated to test the performance of the model on unseen data. The balance of the classes was the same in both sets (Figure 1 represents the total distribution of classes and Figure 2 represents the division between training images and test images).

We scaled the images to the demands of all the models before feeding them to the models. The MobileNetV2 and Custom CNN models had 60x60 pixels whereas the Vision Transformer (ViT) required 256x256 pixels. We also clustered the images to the range of 0 to 1 which somehow makes the models train faster.

In order to boost the strength of the models and prevent overfitting, we used data augmentation such as horizontal and vertical flips, rotations, zooms, and shifts. We also randomized the pictures per epoch such that the model receives a somewhat new batch every time. A validation set was selected randomly about 25 percent of the training set to monitor the performance in the course of training. In general, the purpose of this preprocessing was to provide every model with clean, balanced, and augmented images, all of which contributed to the better learning and more reliable generalization of the four models to detect malaria.

3.3 Model Selection

In this paper, we have selected four deep learning models to calculate malaria-infected and uninfected cells. We trained two individually constructed CNNs, a transfer learning model, MobileNetV2, and a model based on trans- former, the Vision Transformer. Each model had approximately 20 epochs of training and we monitored the validation performance to observe the level of generalization. All the models have their results presented in Figures 3 to 6 in accurate and loss curves.

3.3.1 Custom CNN(First)

The original CNN that we created was quite shallow consisting of convolutional, pooling, and dense layers. Its training was done on 60x60 pixels images. The test accuracy of this model was determined to be 96.06 indicating it was good in distinguishing between infected and uninfected cells. The console output in the training process is presented as shown in Figure 3, and the accuracy, and loss per epoch which provides an approximation of how consistently the model was learning.

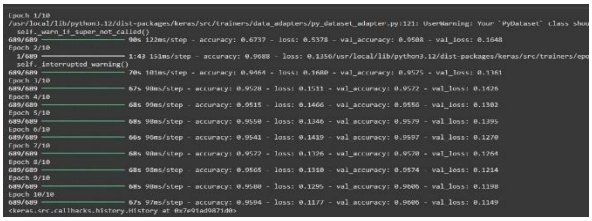

Figure 3: CNN-1 Training Results

3.3.2 MobileNetV2

Using a lightweight pretrained model, MobileNetV2, fine-tuning was done on malaria dataset to create a model that is efficient. Similar to the rest of the models, it would work with images that had been resized to 60x 60 pixels. This model did attain a test accuracy of 93.56, slightly less than the custom CNNs but also fairly competitive. Figure 4 displays the training progress, the accuracy and loss in each epoch, which provided an idea of how the transfer learning process allowed the model to remain steady despite the small size of the image.

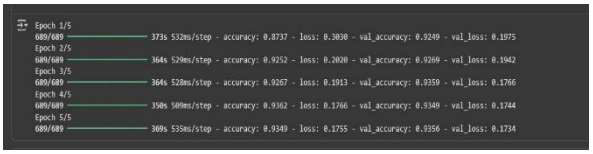

Figure 4: MobileNetV2 Training Results

3.3.3 Vision Transformer(VIT)

The Vision Transformer (ViT) was used to test how a transformer-based model would do on malaria image classification. ViT takes global relationships into consideration in images, and therefore the input images were downsized to 256x256 pixels. This model performed the highest of all the four with a test accuracy of 97.59%. The results of the training are also presented in figure 5, and the accuracy and loss per epoch, it is evident that the model was better learned with minimal overfitting and indicates that ViT can be indeed highly effective on the task.

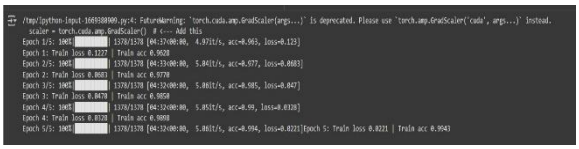

Figure 5: ViT Training Result

3.3.4 CustomCNN(Second/Deeper)

To enable the model to extract more desirable features as compared to the initial CNN, we developed a more advanced custom CNN by incorporating additional convolutional and pooling layers. It was also trained on 60x60 images. This model achieved a test accuracy of 94.32, indicating that richer architectures can be used, but it was not generalizing quite as well as the original CNN. Figure 5 displays the training outcomes in each epoch, including the accuracy and loss, which provide one with the idea of the way the model changed throughout training.

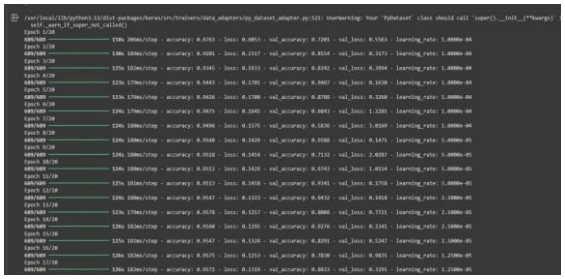

Figure 6: Deeper CNN Training Results

3.4 Training

The models were trained on the cleaned malaria data to approximately 20 epochs, with mini-batches being used in updating the models. The Adam optimizer was selected and the initial learning rate is 1e-4.

We did not maintain the rate constant, but instead we decreased it slowly with training so that the models could change smoothly, accelerating quickly in the beginning and being more cautious in the later stages. Our total training set was 22,048 and test set was 5,512. Out of the training part, we randomly removed approximately 25% of the training as validation, which served the purpose of us ensuring that the models were in fact learning some useful patterns, and not merely memorising the data.

With the continuation of the training, the accuracy and the loss on the training set improved as a rule with every epoch. However, the accuracy of validation was not always on a straight line, and occasionally varied, which demonstrates that the issue is not entirely straight.

The Vision Transformer model was able to achieve good and consistent accuracy on validation relatively early when compared to MobileNetV2, which was more slow to improve but steady. All in all, it was demonstrated that all models could identify the most important differences between the parasitized and uninfected cells.

To simplify this understanding, we also provided graphs of training and validation accuracy and loss of each model. These plots provide a better understanding of the rate at which each network converged as well as how stable the training process was and also the likely ones to overfit.

3.5 Model Evaluation

We tested the models with the held-out test set containing 5,512 microscopic images not used in the training and validation process.

This assisted in establishing a reasonable and objective evaluation of the ability of the models to generalize. The test set was balanced, having an equal number of

parasitized and uninfected cells and the same preprocessing steps namely, resizing, normalization, and augmentation were used in keeping things consistent with the training data.

Accuracy was largely used as a measure of model performance and that reflects the percentage of correctly classified images. ViT had the highest test accuracy of 97.59 and the first Custom CNN of 96.06. The deeper Custom CNN achieved 94.32% and MobileNetV2 was a little bit lower at 93.56% which is rather competitive, still. Figure 7 shows a bar chart of all the four models.

It indicates clearly that transformer-based learning worked the best, and also proves that convolutional models can still be used to classify malaria cells.

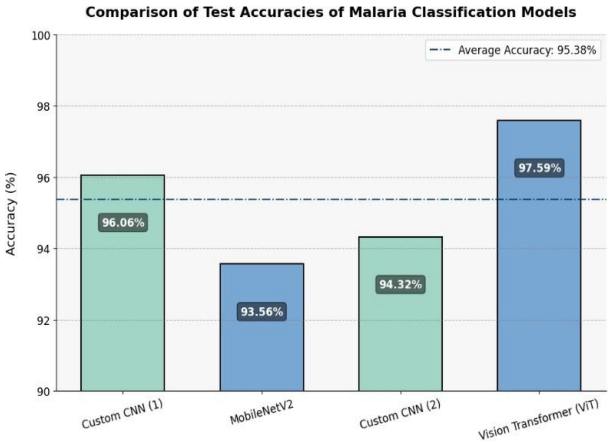

Figure 7: Bar chart comparing the final test accuracies of the four models (Custom CNN-1, MobileNetV2, Custom CNN-2, and Vision Transformer).

4. RESULT

The four models did not do the same. Table 1 contains their results. Vision Transformer (ViT) produced the best accuracy of almost 97 percent. This is most likely due to the fact that it is able to view the image as a whole and identify the patterns that extend to greater areas. The two CNN models were also fairly performing. The former controlled approximately 96 percent and the deeper CNN had just slightly lower at

approximately 94 percent. It also achieved approximately 94% with MobileNetV2 that is supposed to be light and fast. It was faster and sacrificed some precision in comparison with others. On the whole, these findings indicate that CNNs remain quite useful in the task, although trans-formers, such as ViT, can take the performance a bit further.

Detailed classification measures of ViT model are presented in Table 2. The precision, recall and F1-score of both classes, i.e. parasitized and uninfected are quite close to each other i.e. approximately between 0.97 and 0.98 which is indicative of the fact that the model is recognizing both classes correctly and in balanced manner. The test accuracy is approximately 0.98, which is comparable to the high accuracy in Table 1

Table 1: A comparison of the performance of the four evaluated models

Model	Key Architecture Details	Test Accuracy (%)
CNN-1 (Custom)	3 Conv layers + 2 FC layers, ReLU + MaxPool	96.06
Mobile NetV2	Depthwise separable convs, inverted residuals	93.56
CNN-2 (Deeper)	5 Conv layers + 2 FC layers, dropout added	94.32
ViT	Transformer encoder blocks, patch embeddings	97.59

Comparing the macro and weighted averages, it is apparent that the model does not discriminate against one of the classes. Such findings indicate that the ViT model, specifically, can be useful in distinguishing between parasitized and uninfected cells and thus is a desirable choice to implement automated malaria detection in a real-life setting.

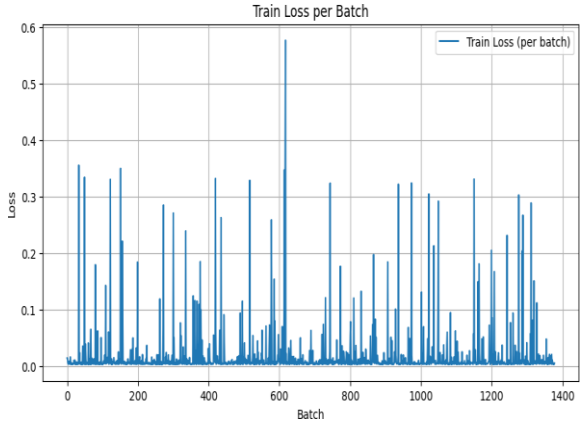

Figure 8. (Training Loss per Batch)

Then, we attempted to estimate the effect of the model on new data, as we tested the model on the test set. The test loss at the conclusion of every epoch was plotted in Figure 9. The loss was almost consistently low, although it did not follow a straight path, a few leaps here and there, and a giant spike appearing around epoch 310. This informs us that the model was successful in picking up training information extraordinarily well but it was not as steady once on new information. Such a behaviour is typically an indication of some overfitting or simply some difficulty in generalizing correctly.we have checked the overfitting behaviour of the model by applying it to the hidden test data.

Figure 9 presents the test loss at every epoch. In general, the loss remained relatively low, yet had obvious fluctuations and certain spikes, particularly around epoch 310. Such a variability indicates that the model did well on the training data, but its performance on new data was somewhat less consistent, which is indicative of possible overfitting or inability to generalize.

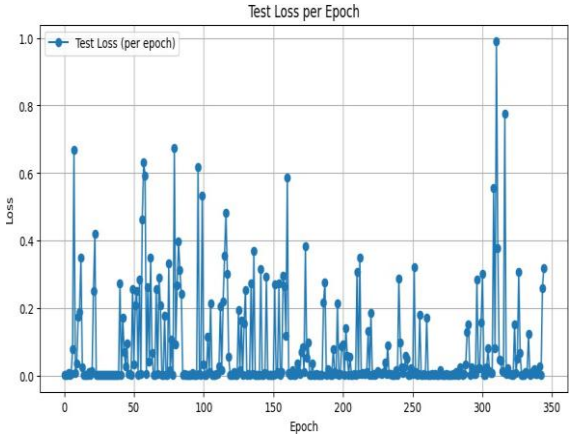

Figure 9. (Test Loss per Epoch)

5. CONCLUSION & FUTUREWORK

This project attempted to demonstrate that malaria can be detected using the blood smear images with the help of deep learning. We have employed some models, two CNNs which we trained, MobileNetV2, and a Vision Transformer (ViT). The ViT classified most accurately with a 97 percent accuracy. The CNNs were nearer yet slightly low. Transformers appear to perform a little better, however, CNNs are not bad. The outcomes also indicate that this type of system would save significant time. Doctors do not need to look through slides over time which is time consuming and tiresome. It is also able to provide more consistent results. This may be of great assistance in areas where hospitals lack man power and equipment. Tools such as this could make patients receive treatment within a shorter time.

Bigger datasets: More images from different labs, stains, and setups could make the models work better on new data .Explainability: Using Grad-CAM or attention maps could help doctors see why the model made certain predictions. Smaller models: Making lightweight versions that can run on phones or small devices would help in rural areas. More classes: Right now it's only healthy vs infected. It could be extended to detect different malaria species like *falciparum* or *Various*-world testing: The models need to be tried in hospitals or clinics to see if they actually work outside the lab.

6. REFERENCES

[1] World Health Organization, *World Malaria Report 2022*. Geneva: WHO, 2022.

[2] World Health Organization, "Malaria," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/malaria. [Accessed: Oct. 3, 2025].

[3] M. Desai, F. O. ter Kuile, F. Nosten, R. McGready, K. Asamoa, B. Brabin, and R. D. Newman, "Epidemiology and burden of malaria in pregnancy," *Lancet Infect. Dis.*, vol. 7, no. 2, pp. 93–104, 2007.

[4] A. Moody, "Rapid diagnostic tests for malaria parasites," *Clin. Microbiol. Rev.*, vol. 15, no. 1, pp. 66–78, 2002.

[5] L. B. Ochola, P. Vounatsou, T. Smith, M. L. H. Mabaso, and C. R. Newton, "The reliability of diagnostic techniques in the diagnosis and management of malaria in the absence of a gold standard," *Lancet Infect. Dis.*, vol. 6, no. 9, pp. 582–588, 2006.

[6] C. K. Murray, R. A. Gasser Jr., A. J. Magill, and R. S. Miller, "Update on rapid diagnostic testing for malaria," *Clin. Microbiol. Rev.*, vol. 21, no. 1, pp. 97–110, 2008.

[7] G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[8] A. Esteva, B. Kuprel, R. A. Novoa, *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017.

[9] S. Rajaraman and S. Antani, "Modality-specific deep learning model ensembles toward improving malaria parasite detection in thin blood smear images," *PeerJ*, vol. 6, p. e4568, 2018.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–21.

[11] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, H. Shen, and L. Shao, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 200:1–200:41, 2022.

[12] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. ICML*, 2021, pp. 10347–10357.

[13] D. K. Das, M. Ghosh, M. Pal, A. K. Maiti, and C. Chakraborty, "Machine learning approach for automated screening of malaria parasite using light microscopic images," *Micron*, vol. 45, pp. 97–106, 2013.

[14] Y. Dong, Z. Jiang, H. Shen, W. D. Pan, L. A. Williams, V. V. Reddy, W. H. Benjamin, A. W. Bryan, "Evaluations of deep convolutional neural networks for automatic identification of malaria infected cells," in *Proc. IEEE EMBC*, 2017, pp. 3158–3161.

[15] Z. Liang, A. Powell, I. Ersoy, M. Poostchi, K. Silamut, K. Palaniappan, P. Guo, M. A. Hossain, A. Sameer, R. J. Maude, S. Jaeger, G. Thoma, and S. Antani, "CNN-based image analysis for malaria diagnosis," in *Proc. IEEE BHI*, 2016, pp. 493–496.

[16] D. Bibin, M. S. Nair, and P. Punitha, "Malaria parasite detection from peripheral blood smear images using deep belief networks," *IEEE Access*, vol. 5, pp. 9099–9108, 2017.

[17] N. S. Shaik, T. G. Sai, P. B. Krishna, and S. S. Basha, "Automatic detection of malaria infected cells using deep convolutional neural network," in *Proc. ICCIDS*, 2018, pp. 129–135.

[18] K. Prasad, S. Prasad, and S. S. Durbha, "Malaria disease recognition using CNN architectures," *Int. J. Pure Appl. Math.*, vol. 118, no. 20, pp. 1569–1574, 2018.

[19] C. Han, *et al.*, "AOCT-NET: Automatic detection of malaria parasites using hybrid CNN architectures," *Comput. Methods Programs Biomed.*, vol. 197, p. 105725, 2020.

[20] M. Poostchi, K. Silamut, R. J. Maude, S. Jaeger, G. Thoma, and S. Antani, "Image analysis and machine learning for detecting malaria," *Transl. Res.*, vol. 194, pp. 36–55, 2018.

[21] S. Rajaraman, S. Antani, M. Poostchi, K. Silamut, M. A. Hossain, R. J. Maude, S. Jaeger, and G. Thoma, "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," *PeerJ*, vol. 6, p. e4568, 2018.