Journal of Imaging

MDPI

Article

An Empirical Evaluation of Convolutional Networks for Malaria Diagnosis

Andrea Loddo *, Corrado Fadda and Cecilia Di Ruberto

Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124 Cagliari, Italy; corradofadda1996@gmail.com (C.F.); dirubert@unica.it (C.D.R.)
* Correspondence: andrea.loddo@unica.it

Abstract: Malaria is a globally widespread disease caused by parasitic protozoa transmitted to humans by infected female mosquitoes of Anopheles. It is caused in humans only by the parasite Plasmodium, further classified into four different species. Identifying malaria parasites is possible by analysing digital microscopic blood smears, which is tedious, time-consuming and error prone. So, automation of the process has assumed great importance as it helps the laborious manual process of review and diagnosis. This work focuses on deep learning-based models, by comparing off-the-shelf architectures for classifying healthy and parasite-affected cells, by investigating the four-class classification on the Plasmodium falciparum stages of life and, finally, by evaluating the robustness of the models with cross-dataset experiments on two different datasets. The main contributions to the research in this field can be resumed as follows: (i) comparing off-the-shelf architectures in the task of classifying healthy and parasite-affected cells, (ii) investigating the four-class classification on the *P. falciparum* stages of life and (iii) evaluating the robustness of the models with cross-dataset experiments. Eleven well-known convolutional neural networks on two public datasets have been exploited. The results show that the networks have great accuracy in binary classification, even though they lack few samples per class. Moreover, the cross-dataset experiments exhibit the need for some further regulations. In particular, ResNet-18 achieved up to 97.68% accuracy in the binary classification, while DenseNet-201 reached 99.40% accuracy on the multiclass classification. The cross-dataset experiments exhibit the limitations of deep learning approaches in such a scenario, even though combining the two datasets permitted DenseNet-201 to reach 97.45% accuracy. Naturally, this needs further investigation to improve the robustness. In general, DenseNet-201 seems to offer the most stable and robust performance, offering as a crucial candidate to further developments and modifications. Moreover, the mobile-oriented architectures showed promising and satisfactory performance in the classification of malaria parasites. The obtained results enable extensive improvements, specifically oriented to the application of object detectors for type and stage of life recognition, even in mobile environments.

Keywords: computer vision; deep learning; image processing; malaria parasites detection; malaria parasites classification

check for updates

Citation: Loddo, A.; Fadda, C.; Di Ruberto, C. An Empirical Evaluation of Convolutional Networks for Malaria Diagnosis. *J. Imaging* **2022**, *8*, 66. https://doi.org/10.3390/jimaging8030066

Academic Editor: Reyer Zwiggelaar

Received: 12 February 2022
Accepted: 4 March 2022
Published: 7 March 2022

1. Introduction

Malaria is a globally widespread disease caused by parasitic protozoa transmitted to humans by infected female mosquitoes of Anopheles. In 2019, there were an estimated 229 million malaria cases worldwide, with an estimated 409,000 deaths due to malaria. Of them, 94% of malaria cases and deaths occurred in Africa [1]. In this context, children under five years of age are the most vulnerable group accounting for 67% (274,000) of all malaria deaths worldwide. Parasites of the genus Plasmodium (P.) cause malaria in humans by attacking red blood cells (RBCs). They spread to people through the bites of infected female Anopheles mosquitoes, called "malaria vectors". Five species of parasites cause malaria in humans: *P. falciparum*, *P. vivax*, *P. ovale*, *P. malariae* and *P. knowlesi*. *P. falciparum* and

P. vivax are the two posing the most significant threat [1,2]. The former is most prevalent in Africa, while *P. vivax* is predominant in the Americas. Malaria plasmids within the human host have the following life stages: ring, trophozoite, schizont and gametocyte. The World Health Organization (WHO) defines human malaria as a preventable and treatable disease if diagnosed promptly. Still, the diagnosis must be made promptly, as the worsening illness can lead to disseminated intravascular thrombosis, tissue necrosis and spleen hypertrophy [1,3–5].

Blood cell analysis using peripheral blood slides under a light microscope is considered the gold standard for the detection of leukaemia [6–9], blood cell counting [10–14] or the diagnosis of malaria [15–17]. Manual microscopic examination of peripheral blood smears (PBS) for malaria diagnosis has advantages such as high sensitivity and specificity compared to other methods. However, it requires about 15 minutes for microscopic examination of a single blood sample [18], and the quality of the diagnosis depends solely on the experience and knowledge of the microscopist. It is common for the microscopist to work in isolation without a rigorous system to ensure the quality of the diagnosis. In addition, the images analysed may be subject to variations in illumination and staining that can affect the results. In general, the manual process is tedious and time-consuming, and decisions dictated by misdiagnosis lead to unnecessary use of drugs and exposure to their side-effects or severe disease progression [19,20].

This work investigates the classification of malaria parasites using transfer learning (TL) to distinguish healthy and parasite-affected cells and classify the four *P. falciparum* stages of life. Moreover, the robustness of the models has been evaluated with cross-dataset experiments on two very different public datasets.

In this paper, transfer learning will be introduced by explaining how it works and discussing the pretrained networks selected to perform the comparative tests. The experiments are divided into (i) binary, (ii) multiclass and (iii) cross-domain classification. In the latter, networks trained on datasets from different domains were used to see if this improves accuracy over results obtained in a single domain.

The rest of the manuscript is organised as follows. Section 2 presents the literature on computer-aided diagnostic (CAD) systems for malaria analysis. Section 3 illustrates the datasets, methods and experimental setup. The results are presented and discussed in Section 4 and, finally, in Section 5, the findings and directions for future works are drawn.

2. Related Work

Several solutions for the automatic detection of malaria parasites have been developed in recent years. They aim to reduce the problems of manual analysis depicted in Section 1 and provide a more robust and standardised interpretation of blood samples while reducing the costs of diagnosis [15,21,22], mainly represented by CAD systems. They can be based on the combination of image processing and traditional machine learning techniques [23–25], and also deep learning approaches [16,26–28], especially after the proposal of AlexNet's convolutional neural network (CNN) [29].

Since malaria parasites always affect the RBCs, any automatic malaria detection needs to analyse the erythrocytes to discover if they are infected or not by the parasite and, further, to find the stage of life or the type.

Among the more recent and classical solutions not employing CNNs, Somasekar et al. [23] and Rode et al. [25] proposed two malaria parasite segmentation methods. The first one used fuzzy clustering and connected component labelling followed by minimum perimeter polygon to segment parasite-infected erythrocytes and detect malaria, while the second one is based on image filtering and saturation separation followed by triangles thresholding.

Regarding the CNN-based approaches, Liang et al. [26] proposed a novel model for the classification of single cells as infected or uninfected, while Rajaraman et al. [27] studied the accuracy of CNN models, starting from pretrained networks, and proposed a novel architecture trained on a dataset available from the National Institutes of Health (NIH). They found that some pre-existing networks, by means of TL, can be more efficient

than networks designed ad hoc. In particular, ResNet-50 obtained the best performance. Subsequently, they further improved through an ensemble of CNNs [28]. Rahman et al. [30] also exploited TL strategies using both natural and medical images and performed an extensive test of some off-the-shelf CNNs to realise a binary classification.

Some other techniques not explored in this work are based on the combination of CNN-extracted features and handcrafted ones [31–33] or the direct use of object detectors [34]. For example, Kudisthalert et al. [33] proposed a malaria parasite detection system, based on the combination of handcrafted and deep features, extracted from pretrained AlexNet. Abdurahman et al. [34] realised a modified version of the YOLOV4 detector. Moreover, they generated new anchor box sizes with a K-means clustering algorithm to exploit the model on small objects.

Finally, a recent focus has been posed on mobile devices, which enable a cheaper and quicker diagnosis in the underdeveloped areas of the world, where more expensive laboratories do not exist. As an example, Bias et al. [24] realised an edge detection technique based on a novel histogram-based analysis, coupled with easily accessible hardware, focused on malaria-infected thin smear images.

The work in [30] is the most similar to the approach here proposed. In particular, they compared different off-the-shelf networks for a binary classification using two datasets, one is the Malaria Parasite Image Database for Image Processing and Analysis (MP-IDB) [35] and another is composed of synthetic and medical images. The task faced, however, is a binary classification. On the entire MP-IDB, they reported 85.18% accuracy with a fine-tuned version of VGG-19.

In summary, the main difference between our work and the state-of-art is that here an extended set of off-the-shelf CNNs on two very different public datasets have been exploited with a dual purpose: detect healthy and unhealthy RBCs and distinguish the various stages of life. Finally, it is the first baseline provided for the stages of life classification on the MP-IDB.

3. Materials and Methods

In this section, the datasets, the techniques and the employed experimental setup are described.

3.1. Datasets

Two well-known benchmark datasets were used: the National Institutes of Health (NIH) [27], proposed for malaria detection, and MP-IDB [35], a recently proposed dataset for malaria parasite types and stages of life classification.

3.1.1. NIH

The NIH is a public PBS images dataset from healthy individuals and malaria-affected patients. Image acquisition was performed at the Lister Hill National Center for Biomedical Communications from Giemsa-stained blood samples obtained from 150 patients infected with *P. falciparum* and 50 healthy patients. The resulting dataset consists of 27,558 images, uniformly subdivided between infected and healthy cells. The infected cells contain a ring-shaped parasite. Figure 1 shows a healthy and sick RBC extracted from NIH.

3.1.2. MP-IDB

The MP-IDB consists of four malaria parasite species, *P. falciparum*, *P. malariae*, *P. ovale* and *P. vivax*, represented by 122, 37, 29 and 46 images, respectively, for a total amount of 229. The images have been acquired with 2592 × 1944 resolution and 24-bit colour depth. Moreover, every species contains four distinct life stages: ring, trophozoite, schizont and gametocyte. Figure 2 shows four examples of the types of malaria parasites included in MP-IDB.

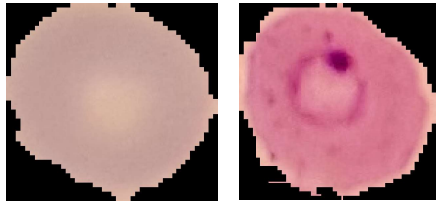
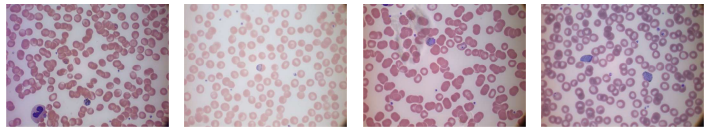


Figure 1. Example of RBCs, healthy (**left**) and sick (**right**), included in the NIH dataset.



P. falciparum. *P. malariae*. *P. ovale*. *P. vivax*.

Figure 2. Example of the four types of malaria parasites of MP-IDB.

3.2. Classification Pipeline

3.2.1. Deep Learning

In this work, deep learning approaches have been used as classifiers. In particular, eleven different off-the-shelf CNN architectures have been evaluated. All of them have been pretrained on the well-known natural image dataset ImageNet [36], and then adapted to this medical image task, as proposed in [37], following a transfer learning and fine-tuning procedure.

AlexNet [29] and VGG-16 [38] are pretty simple and similar architectures, composed of 8 and 16 layers, respectively. Nevertheless, they are widely used for transfer learning and fine-tuning [37], as they have gained popularity for their excellent performance in many classification tasks [29]. GoogLeNet [39] and Inceptionv3 [40] are both based on the inception layer; in fact, Inceptionv3 is a variant of GoogLeNet, using 140 levels, 40 more than GoogLeNet. The 3 ResNet architectures have 18, 50, 101 layers for ResNet-18, ResNet-50 and ResNet-101, respectively, based on residual learning. They are easier to optimise even when the depth increases considerably [41]. The building block of ResNet inspired DenseNet-201. It deals with the vanishing-gradient problem by introducing a connectivity pattern between layers. It is comprised of multiple densely connected layers, and their outputs are connected to all the successors in a dense block [42]. ShuffleNet [43], SqueezeNet [44] and MobileNetV2 [45] are lighter networks. In particular, the last two are oriented to real-time executions and mobile device usage.

Regarding the transfer learning strategy, the approach used in [37] was followed. All CNN layers were retained except for the last fully connected one. It was replaced with a new layer, initialised and set up to accommodate new categories according to the classification strategy exploited (two classes in the binary one and four in the multiclass one).

3.2.2. Image Preprocessing

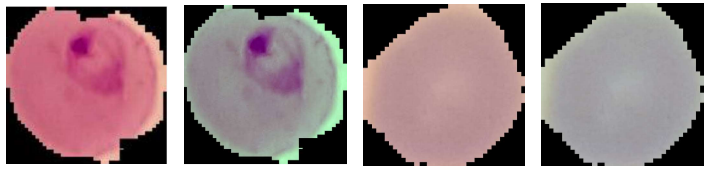
All the images have a uniform background, even if different according to the dataset. Indeed, NIH contains only single-cell images surrounded by a black background, while the images of MP-IDB have several blood components with their actual plasma background. Both datasets are organised into classes, healthy or sick, for NIH and the four malaria types for MP-IDB. As it can be seen from Figures 1 and 2, there are many colour variations between the images of both datasets. This condition is undoubtedly due to the different acquisition conditions and from the status of the blood smear [35]. Therefore,

it was considered appropriate to apply a preprocessing step to realise a colour balance. Furthermore, for MP-IDB, we created a single image for each parasite from the full-size images. The preprocessing step was designed mainly to adjust the colour components of the images and applied to all the RGB channels, using a colour-balancing technique, through Equation (1), where C_{out} is the processed component, C_{in} is the component to be processed, m_{im} is the average of the average intensities of the three channels and, finally, m_c is the average intensity of the component to be processed. This procedure was carried out on all three channels of the RGB image.

$$C_{out} = C_{in} \frac{m_{im}}{m_c} \quad (1)$$

As far as MP-IDB is concerned, the work was also oriented to generate single-cell images from the full-size and, more specifically, the Falciparum class. From now on, we refer to this dataset as MP-IDB-Falciparum-Crops (MP-IDB-FC). MP-IDB contains 104 patient-level images with the corresponding ground truths of the infected cells. As the desired classification was at the cellular level, individual cells had to be extracted from the images and labelled using the class references in the file names. The Falciparum class presents the following image distribution per stage: 1230 rings, 18 schizonts, 42 trophozoites and 7 gametocytes. In Figures 3 and 4, the preprocessing step results are shown.

Finally, to overcome the class imbalance issue in the dataset produced, a further augmentation step was applied and described below.



Sick RBC before. Sick RBC after. Healthy RBC before. Healthy RBC after.

Figure 3. Examples of NIH images, before and after preprocessing step.

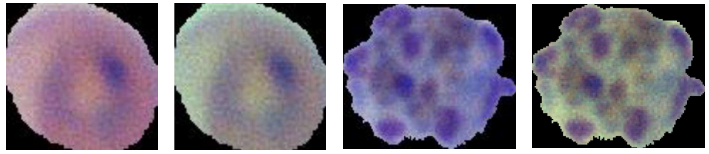


Figure 4. Examples of MP-IDB-FC images, before and after preprocessing step.

3.2.3. Data Augmentation

Different data augmentation techniques, such as flipping, shifting and rotating, are used to overcome the problem of a limited dataset and to produce further copies of the original image to give the algorithm more generalisation capability and reduce the error rate [29,46–48].

MP-IDB-FC presented an unbalanced distribution of images per class; therefore, we proposed an offline data augmentation to oversample the underrepresented classes. The applied geometric transformations are random rotations between $-90°$and $90°$, random translations between -10 and 10 pixels on the X-axis and Y-axis, and 50% reflection around the X-axis and Y-axis.

3.2.4. Experimental Setup

As previously introduced, the images to be classified are microscopic blood smear images. More specifically, NIH contains single-cell images representing healthy or ring stage malaria parasite-affected RBC, while in the case of MP-IDB-FC, the images represent a single cell containing one of the four stages of *P. falciparum*.

In the case of NIH, the dataset was divided into three parts, one for the training (80%), one for the validation (10%) and 10% for testing. A stratified sampling procedure to keep the splits balanced was used, as NIH classes are well balanced.

As for MP-IDB, in order to consider the class imbalance and to have a sufficient number of samples for the training process while preserving a sufficient number of samples for performance evaluation, the dataset was split first into two parts, namely training and testing set, with 80 and 20% of images, respectively. A total of 10% of the training set was used for validation due to the small number of images. Naturally, the subdivision of the sets was carried out before oversampling to avoid that train or validation images fell into the test set, thus compromising the results. The subdivision was made to ensure that 20% of each class made up the test set, thus avoiding the circumstance where no class images with few elements were in the test set. Oversampling and undersampling were then adopted to increase the number of images to 100, 200 or 300 per class in the training set. Moreover, the splits were not created randomly but by taking the images in lexicographic order from each class to further ease reproducibility. All the experiments have been conducted on a single machine with the following configuration: Intel(R) Core(TM) i9-8950HK @ 2.90 GHz CPU with 32 GB RAM and NVIDIA GTX1050 Ti 4GB GPU. Finally, no randomisation strategy has been applied to make all the experiments on both datasets reproducible.

This work represents a baseline for further investigation and searching for the best architecture for our purpose. The selected CNNs have been employed without any modification to their architecture. In particular, after empirical evaluation, Adam algorithm was adopted, which performed better than the other solvers. In addition, the maximum number of epochs was set to 10 due to the number of images.

As mentioned in Section 3.2.1, a fine-tuning process was applied on all the CNN architectures exploited on both datasets. The hyperparameters defined in Table 1 were used for all networks to evaluate potential performance variations. Furthermore, the regularisation factor L_2 was set to avoid a possible overfitting during the training phase.

Table 1. Hyperparameters settings for CNNs fine-tuning.

Params	Value
Solver	Adam
Max Epochs	10
Mini Batch Size	32
Initial Learn Rate	1×10^{-4}
Learn Rate Drop Period	10
Learn Rate Drop Factor	0.1
L_2 Regularisation	0.1

4. Experimental Results

Three different experiments were conducted, according to the classification purpose:

- Binary classification on the NIH dataset (healthy vs. sick);
- Multiclass classification on the MP-IDB-FC dataset (four stages of life);
- Multiclass cross-dataset classification on both datasets.

The results obtained in the analysis of each experiment were performed using the confusion matrix. The confusion matrix metric used in this study is *Accuracy*. The formula of this metric is given in Equation (2). The variables used in the equation are True Positive

(*TP*), False Positive (*FP*), True Negative (*TN*), and False Negative (*FN*), parameters of the confusion matrix used to calculate the metrics [49,50].

$$Accuracy = \frac{TP + TN}{TP + TF + FP + FN} \quad (2)$$

4.1. Binary Classification Performance on NIH

To determine the training options to use, test trials were carried out. From them, it was mainly verified that:

- Extending the training phase beyond ten epochs did not improve accuracy, as the network stored individual image features rather than class features, and overfitting compromised the results;
- The ideal learning rate was 1×10^{-4}. The accuracy increased too slowly for smaller values, and for larger ones, it did not converge to a specific value;
- Empirically, Adam was found as the best solver.

Table 2 shows that almost all the networks have an accuracy value close to the average. The standard deviation of the collected data is solely 0.16%. This aspect could be because the dataset used has many valuable images for training the network. In particular, ResNet-18 recorded the highest accuracy value, confirming the high performance expressed in [27]. MobileNetV2, SqueezeNet and ShuffleNet recorded average values, which is an important result since they are networks designed for mobile use.

Table 2. Accuracy and training times of CNNs on NIH with preprocessing step.

Network	Accuracy (%)	Time (min)
AlexNet	97.35	10
DenseNet-201	95.86	5145
ResNet-18	97.68	21
ResNet-50	97.61	82
ResNet-101	97.24	391
GoogLeNet	96.73	111
ShuffleNet	97.39	33
SqueezeNet	97.21	16
MobileNetV2	97.31	210
Inceptionv3	96.70	151
VGG-16	97.31	322
Avg.	97.25	–

4.2. Multiclass Classification Performance on MP-IDB-FC

The multiclass classification on MP-IDB-FC was designed to determine the life stage of the parasite: ring phase, adult trophozoite, schizont and gametocyte.

Like the binary classification, comparative tests with the same training set, validation set and test set were carried out to determine which networks perform best and allow comparison between them. We created three datasets with 100, 200 or 300 images per class in the training set. We refer to these sets as D1, D2 and D3, respectively. They were constructed by oversampling with augmentation of the remaining classes.

Table 3 shows the performance in this experiment. Each test was cross-validated five times; then, we reported the mean accuracy and standard deviation considering each of the five folds. The most notable result is that the average performance from D2 to D3 worsens. However, DenseNet-201 and GoogLeNet are the only networks to benefit from increasing the dimensionality of the training set.

Table 3. Average accuracy and standard deviation computed on the same test set, after training with 100, 200 and 300 images (D1, D2, D3, respectively).

	Accuracy ± Standard Deviation (%)		
Network	**D1**	**D2**	**D3**
AlexNet	83.59 ± 12.90	89.74 ± 2.56	89.23 ± 2.15
DenseNet-201	94.15 ± 3.34	95.74 ± 2.56	99.40 ± 0.40
ResNet-18	90.77 ± 6.88	95.38 ± 2.81	89.74 ± 5.44
ResNet-50	91.79 ± 4.50	94.36 ± 3.80	86.15 ± 3.89
ResNet-101	94.87 ± 2.81	92.31 ± 2.56	92.31 ± 3.89
GoogLeNet	92.82 ± 3.34	92.31 ± 2.56	93.85 ± 1.40
ShuffleNet	92.82 ± 2.81	91.28 ± 3.89	89.74 ± 4.10
SqueezeNet	90.26 ± 6.88	88.72 ± 7.82	88.21 ± 8.02
MobileNetV2	87.18 ± 3.34	84.62 ± 5.90	79.49 ± 8.42
Inceptionv3	93.85 ± 3.80	92.31 ± 2.56	84.62 ± 2.56
VGG-16	93.85 ± 3.34	92.31 ± 4.18	87.18 ± 4.50
Avg.	91.42	91.64	88.75

4.3. Cross-Dataset Classification Evaluation

The cross-domain classification was carried out to evaluate the CNNs robustness.

4.3.1. MP-IDB-FC Classification with NIH Models

Firstly, two different multiclass classifications were realised on MP-IDB-FC by:

- Training on NIH and testing on MP-IDB-FC (Exp1);
- Training on NIH + fine-tuning on MP-IDB and testing on MP-IDB-FC (Exp2).

When used for training or fine-tuning, the split on MP-IDB-FC was 50% for training (with 10% for validation) and 50% for testing. Every test was evaluated with five-fold cross-validation. Therefore, we report the average accuracy of the five folds and the standard deviation. This cross-domain experiment tested whether the networks trained on the NIH dataset could be employed on the MP-IDB-FC dataset. It is helpful to point out that the significant difference is that, on the one hand, MP-IDB-FC has parasite crops and not healthy RBCs and, on the other hand, NIH contains only ring-stage parasites. For this reason, the objective of Exp1 was to discriminate between rings and the remaining stages of life, while Exp2 aimed to expand the knowledge of the NIH pretrained models with new information on the stages of life.

The results depicted in Table 4 show that when the target domain differs excessively from the source domain, it is hard to directly apply the models trained with NIH to MP-IDB-FC (Exp1), even if the task seemed feasible, as ring-stage parasites were contained in both datasets. Conversely, Exp2 shows that using the CNNs first trained on NIH and then fine-tuned on MP-IDB-FC led to an improvement in average accuracy. The information about healthy RBCs provided with NIH training does not affect the overall result. In addition, the standard deviation is under 4% for all the networks, leading to satisfactory performance stability.

4.3.2. *P. vivax* Classification Using *P. falciparum* Data

The last experiment aimed to investigate the possibility of classifying the stages of life of *P. vivax* using the information on *P. falciparum*. So, a different dataset was created and composed of the crops of *P. vivax* parasites, referred to as MP-IDB-VC. Even in this case, three different evaluations were conducted by:

- Training on MP-IDB-VC and testing on MP-IDB-FC (Exp3);
- Training on MP-IDB-VC and testing on MP-IDB-VC (Exp4);
- Training on MP-IDB-FC, fine-tuning and testing on MP-IDB-VC (Exp5).

Table 4. Cross-dataset experiments for multiclass classification on MP-IDB-FC.

Network	Exp1 (%)	Exp2 (%)
AlexNet	42.71	88.21 ± 3.44
DenseNet-201	49.73	97.45 ± 1.40
ResNet-18	68.23	94.87 ± 2.56
ResNet-50	68.23	92.31 ± 3.63
ResNet-101	69.10	94.87 ± 2.29
GoogLeNet	68.23	89.23 ± 3.80
ShuffleNet	32.23	95.90 ± 1.40
SqueezeNet	68.23	87.18 ± 3.14
MobileNetV2	57.12	89.74 ± 2.29
Inceptionv3	41.84	93.85 ± 3.44
VGG-16	53.59	94.87 ± 2.81
Avg.	60.61	92.59

Trophozoites, schizonts and gametocytes greatly vary between the two types, while the ring stages are pretty similar.

As it can be seen from Table 5, the classification of *P. falciparum* stages of life employing models trained *P. vivax* produced dreadfully low results due to the differences between all the stages except rings. On the other hand, using same-domain models (Exp4) had satisfactory results. Exp5 demonstrates that the fine-tuning strategy on the models pretrained on *P. falciparum* improved the accuracy, as already happened in Section 4.3.1 Exp2. In this task, DenseNet-201 provided the best performance, being the only CNN to overcome 85% and outperforming the average of 10% in both cases.

Table 5. Cross-dataset experiments for multiclass classification on MP-IDB-VC.

Network	Exp3 (%)	Exp4 (%)	Exp5 (%)
AlexNet	30.16	77.14 ± 9.31	71.43 ± 11.29
DenseNet-201	72.94	88.89 ± 5.56	87.10 ± 1.15
ResNet-18	30.16	82.86 ± 3.91	82.86 ± 6.93
ResNet-50	55.56	78.57 ± 8.74	78.57 ± 5.05
ResNet-101	57.14	82.86 ± 3.91	81.43 ± 3.91
GoogLeNet	28.57	68.57 ± 3.91	74.28 ± 3.91
ShuffleNet	58.73	82.86 ± 3.91	82.86 ± 3.91
SqueezeNet	31.75	80.00 ± 5.98	80.00 ± 5.98
MobileNetV2	34.92	67.14 ± 3.91	74.28 ± 6.39
Inceptionv3	57.14	78.57 ± 5.05	78.57 ± 5.05
VGG-16	49.21	77.14 ± 3.19	81.43 ± 3.91
Avg.	46.02	78.60	79.35

5. Conclusions

The results obtained in this work support the importance of deep learning in haematology. This work aimed to demonstrate that pretrained off-the-shelf networks can offer high accuracy for diagnosing malaria utilising transfer learning however showing several limitations of this approach. Several comparative tests were developed using a selection of pretrained networks differentiated by size, depth and the number of parameters. In particular, using the NIH dataset, it is possible to distinguish a healthy from an infected erythrocyte with an accuracy of over 97%. Small networks such as SqueezeNet and ShuffleNet performed well, consolidating a possible development of software for malaria diagnosis in small devices such as smartphones. On the other hand, MP-IDB has highlighted some critical issues: deep learning is not very effective when the dataset used for training is unbalanced. Some classes of parasites in the dataset have a small number of images. Nevertheless, the oversampling, augmentation and preprocessing methods still allowed us to exceed 90% accuracy on the test set for distinguishing the four life stages of the *P. falciparum* parasite. Finally, the cross-domain experiments have highlighted some critical points in classifying data from heterogeneous domains. It was counterproductive to apply the models trained with NIH to MP-IDB-FC, but the use of the CNNs firstly trained on NIH and fine-tuned on MP-IDB-FC led to an improvement in average accuracy. This aspect also applies to using the *P. vivax* dataset as the target domain, as most of the classes deviate too much from the corresponding *P. falciparum* classes. However, using both types of parasites as source domains produced better results than training on *P. vivax* only. In general, the extensive experimentation has highlighted how DenseNet-201 offers the most stable and robust performance, offering itself as a crucial candidate for further developments and modifications.

Among the possible developments of this work, we aim to propose a framework able to detect malaria parasites from blood smear images and classify different species of parasites and different stages of life, mainly focusing on high variation data. We also plan to use domain adaptation algorithms to improve cross-domain performance.

Author Contributions: Conceptualisation, A.L. and C.D.R.; methodology, A.L. and C.D.R.; investigation, A.L., C.F. and C.D.R.; software, A.L., C.F. and C.D.R.; writing—original draft, A.L. and C.D.R.; writing—review and editing, A.L. and C.D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the material used and developed for this work is available at the following GitHub repository. All the models trained for the different experimentations can be found at this repository.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Plasmodium	*P.*
RBC	Red Blood Cell
PBS	Peripheral Blood Smears
CAD	Computer-Aided Diagnostic
CNN	Convolutional Neural Network
TL	Transfer Learning
MP-IDB	Malaria Parasite Image Database for Image Processing and Analysis
NIH	National Institutes of Health

References

1. WHO. 2021. Available online: https://www.who.int/news-room/fact-sheets/detail/malaria (accessed on 13 September 2021).
2. Stanford Healthcare. 2021. Available online: https://stanfordhealthcare.org/medical-conditions/primary-care/malaria/types.html (accessed on 13 September 2021).
3. ScienceDirect. 2021. Available online: https://www.sciencedirect.com/topics/neuroscience/malaria (accessed on 10 September 2021).
4. WHO. 2021. Available online: https://www.who.int/health-topics/malaria#tab=tab_1 (accessed on 10 September 2021).
5. Centers for Disease Control Prevention. Available online: https://www.cdc.gov/malaria/about/biology/index.html (accessed on 10 September 2021).
6. Vogado, L.H.; Veras, R.M.; Araujo, F.H.; Silva, R.R.; Aires, K.R. Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification. *Eng. Appl. Artif. Intell.* **2018**, *72*, 415–422. [CrossRef]
7. Toğaçar, M.; Ergen, B.; Cömert, Z. Classification of white blood cells using deep features obtained from Convolutional Neural Network models based on the combination of feature selection methods. *Appl. Soft Comp. J.* **2020**, *97*, 106810. [CrossRef]
8. Mondal, C.; Hasan, M.K.; Jawad, M.T.; Dutta, A.; Islam, M.R.; Awal, M.A.; Ahmad, M. Acute Lymphoblastic Leukemia Detection from Microscopic Images Using Weighted Ensemble of Convolutional Neural Networks. *arXiv* **2021**, arXiv:2105.03995.
9. Huang, Q.; Li, W.; Zhang, B.; Li, Q.; Tao, R.; Lovell, N.H. Blood Cell Classification Based on Hyperspectral Imaging with Modulated Gabor and CNN. *IEEE J. Biomed. Health Inf.* **2020**, *24*, 160–170. [CrossRef] [PubMed]

10. Di Ruberto, C.; Loddo, A.; Puglisi, G. Blob Detection and Deep Learning for Leukemic Blood Image Analysis. *Appl. Sci.* **2020**, *10*, 1176. [CrossRef]
11. Di Ruberto, C.; Loddo, A.; Putzu, L. Learning by Sampling for White Blood Cells Segmentation. In Proceedings of the 18th International Conference Image Analysis and Processing (ICIAP 2015), Genoa, Italy, 7–11 September 2015; Volume 9279, pp. 557–567.
12. Di Ruberto, C.; Loddo, A.; Putzu, L. A leucocytes count system from blood smear images Segmentation and counting of white blood cells based on learning by sampling. *Mach. Vis. Appl.* **2016**, *27*, 1151–1160. [CrossRef]
13. Di Ruberto, C.; Loddo, A.; Putzu, L. Detection of red and white blood cells from microscopic blood images using a region proposal approach. *Comput. Biol. Med.* **2020**, *116*, 103530. [CrossRef]
14. Xie, W.; Noble, J.A.; Zisserman, A. Microscopy cell counting and detection with fully convolutional regression networks. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **2018**, *6*, 283–292. [CrossRef]
15. Loddo, A.; Di Ruberto, C.; Kocher, M. Recent Advances of Malaria Parasites Detection Systems Based on Mathematical Morphology. *Sensors* **2018**, *18*, 513. [CrossRef]
16. Maity, M.; Jaiswal, A.; Gantait, K.; Chatterjee, J.; Mukherjee, A. Quantification of malaria parasitaemia using trainable semantic segmentation and capsnet. *Pattern Recognit. Lett.* **2020**, *138*, 88–94. [CrossRef]
17. Vijayalakshmi, A.; Kanna, B.R. Deep learning approach to detect malaria from microscopic images. *Multim. Tools Appl.* **2020**, *79*, 15297–15317. [CrossRef]
18. Chan, Y.K.; Tsai, M.H.; Huang, D.C.; Zheng, Z.H.; Hung, K.D. Leukocyte nucleus segmentation and nucleus lobe counting. *BMC Bioinform.* **2010**, *11*, 558. [CrossRef] [PubMed]
19. De Carneri, L. *Parassitologia Generale e Umana*; Casa Editrice Ambrosiana CEA: Milano, Italy, 1972.
20. Faust, E.C.; Beaver, P.C.; Jung, R.C. *Animal Agents and Vectors of Human Disease*; Henry Kimpton Publishers Ltd.: London, UK, 1975.
21. Jan, Z.; Khan, A.; Sajjad, M.; Muhammad, K.; Rho, S.; Mehmood, I. A review on automated diagnosis of malaria parasite in microscopic blood smears images. *Multim. Tools Appl.* **2018**, *77*, 9801–9826. [CrossRef]
22. Poostchi, M.; Silamut, K.; Maude, R.J.; Jaeger, S.; Thoma, G. Image analysis and machine learning for detecting malaria. *Transl. Res.* **2018**, *194*, 36–55. [CrossRef] [PubMed]
23. Somasekar, J.; Eswara Reddy, B. Segmentation of erythrocytes infected with malaria parasites for the diagnosis using microscopy imaging. *Comput. Electr. Eng.* **2015**, *45*, 336–351. [CrossRef]
24. Bias, S.; Reni, S.; Kale, I. Mobile Hardware Based Implementation of a Novel, Efficient, Fuzzy Logic Inspired Edge Detection Technique for Analysis of Malaria Infected Microscopic Thin Blood Images. *Procedia Comput. Sci.* **2018**, *141*, 374–381. [CrossRef]
25. Rode, K.B.; Bharkad, S.D. Automatic segmentation of malaria affected erythrocyte in thin blood films. In *International Conference on ISMAC in Computational Vision and Bio-Engineering*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 993–1002.
26. Liang, Z.; Powell, A.; Ersoy, I.; Poostchi, M.; Silamut, K.; Palaniappan, K.; Guo, P.; Hossain, M.A.; Antani, S.K.; Maude, R.J.; et al. CNN-based image analysis for malaria diagnosis. In Proceedings of the International Conference on Bioinformatics and Biomedicine (BIBM 2016), Shenzhen, China, 15–18 December 2016; pp. 493–496.
27. Rajaraman, S.; Antani, S.K.; Poostchi, M.; Silamut, K.; Hossain, M.A.; Maude, R.J.; Jaeger, S.; Thoma, G.R. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ* **2018**, *6*, e4568. [CrossRef]
28. Rajaraman, S.; Jaeger, S.; Antani, S.K. Perf. eval. of deep neural ensembles toward malaria parasite detection in thin-blood smear images. *PeerJ* **2019**, *7*, e6977. [CrossRef]
29. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; Volume 1, pp. 1097–1105.
30. Rahman, A.; Zunair, H.; Reme, T.R.; Rahman, M.S.; Mahdy, M. A comparative analysis of deep learning architectures on high variation malaria parasite classification dataset. *Tissue Cell* **2021**, *69*, 101473. [CrossRef]
31. Nanni, L.; Ghidoni, S.; Brahnam, S. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognit.* **2017**, *71*, 158–172. [CrossRef]
32. Loddo, A.; Di Ruberto, C. On the Efficacy of Handcrafted and Deep Features for Seed Image Classification. *J. Imaging* **2021**, *7*, 171. [CrossRef] [PubMed]
33. Kudisthalert, W.; Pasupa, K.; Tongsima, S. Counting and Classification of Malarial Parasite From Giemsa-Stained Thin Film Images. *IEEE Access* **2020**, *8*, 78663–78682. [CrossRef]
34. Abdurahman, F.; Fante, K.A.; Aliy, M. Malaria parasite detection in thick blood smear microscopic images using modified YOLOV3 and YOLOV4 models. *BMC Bioinform.* **2021**, *22*, 112. [CrossRef] [PubMed]
35. Loddo, A.; Di Ruberto, C.; Kocher, M.; Prod'Hom, G. MP-IDB: The Malaria Parasite Image Database for Image Processing and Analysis. In *Processing and Analysis of Biomedical Information*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 57–65.
36. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.

37. Shin, H.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* **2016**, *35*, 1285–1298. [CrossRef] [PubMed]
38. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015.
39. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
40. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
42. Huang, G.; Liu, Z.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
43. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
44. Iandola, F.N.; Moskewicz, M.W.; Ashraf, K.; Han, S.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <1 MB model size. *arXiv* **2016**, arXiv:1602.07360.
45. Sandler, M.; Howard, A.G.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
46. Wong, S.C.; Gatt, A.; Stamatescu, V.; McDonnell, M.D. Understanding Data Augmentation for Classification: When to Warp? In Proceedings of the 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA 2016), Gold Coast, Australia, 30 November–2 December 2016; pp. 1–6.
47. Shijie, J.; Ping, W.; Peiyi, J.; Siping, H. Research on data augmentation for image classification based on convolution neural networks. In Proceedings of the 2017 Chinese Automation Congress (CAC), Jinan, China, 20–22 October 2017; pp. 4165–4170.
48. Mikołajczyk, A.; Grochowski, M. Data augmentation for improving deep learning in image classification problem. In Proceedings of the 2018 International Interdisciplinary PhD Workshop (IIPhDW), Swinoujscie, Poland, 9–12 May 2018; pp. 117–122.
49. Ergen, M.B. Texture based feature extraction methods for content based medical image retrieval systems. *Bio-Med. Mater. Eng.* **2014**, *24*, 3055–3062. [CrossRef]
50. Başaran, E.; Şengür, A.; Cömert, Z.; Budak, Ü.; Çelık, Y.; Velappan, S. Normal and Acute Tympanic Membrane Diagnosis based on Gray Level Co-Occurrence Matrix and Artificial Neural Networks. In Proceedings of the 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 21–22 September 2019; pp. 1–6.