Supervised learning and image processing for efficient malaria detection

Michael White[0000−0001−6902−9790] and Patrick Marais[0000−0001−8747−7765]

University of Cape Town, Cape Town, South Africa
mike.james.white@icloud.com
patrick@cs.uct.ac.za

Abstract. Malaria is a devastating disease that leads to many deaths each year. Currently, most malaria diagnoses are performed manually, which is time consuming. This may result in it taking longer to diagnose patients, especially those in poor and rural areas, motivating the development of automated detection tools. Deep learning approaches, particularly convolutional neural networks (CNNs), have seen success in the existing literature. However, CNNs are computationally expensive and require significant amounts of training data, which may limit their real-world viability, especially in poorer and rural communities. Non-deep supervised techniques are largely free from these limitations but have received less attention in the existing literature. This paper differs from existing work using non-deep systems by investigating the use of RFs, adopting a more rigourous testing methodology and conducting a broader exploration of pre-processing techniques. Two non-deep supervised systems are proposed, based on random forests (RFs) and support vector machines (SVMs). The RF system performs better, having achieved an accuracy of 96.29% when tested on 20 000 images, with runtimes of less than two seconds. Testing on a small dataset of images gathered from a different source achieves similar performance, suggesting the model may generalise to different imaging conditions. The system achieves higher recall than existing non-deep approaches, and its accuracy, recall and precision are within 4% of the highest performing CNN approach.

Keywords: Supervised Learning · Image Processing · Computer-Aided Diagnosis · Machine Learning · Image Classification · Feature Extraction · Malaria.

1 Introduction

Malaria is a parasitic disease that can have devastating effects, not only for individuals who contract it, but also for their families and communities, which may suffer economic harm as a result [14]. African countries are disproportionately affected, with 92% of global infections and 93% of global deaths falling in the World Health Organisation African region. Malaria also disproportionately affects poorer and rural areas where access to diagnosis and treatment is limited.

This presents a clear problem as those who are most in need of medical assistance are less likely to receive it in time. Currently, the gold standard for malaria diagnosis is manual microscopy performed by an expert. While this is a reliable method of diagnosis, it is also costly and time consuming [15]. Part of the reason poorer and rural areas are worst affected by malaria may be attributed to a lack of access to these experts. This motivates the need for low-cost and reliable automated detection systems that minimise the time burden on medical experts.

Deep learning approaches, specifically convolutional neural networks (CNNs), have proven effective when applied to the problem of malaria diagnosis, with one example achieving over 99% accuracy [12]. However, CNNs are reliant on large sets of training data and significant computational power for training. Since imaging conditions vary across different clinics and laboratories, additional training on local blood sample data may be necessary. As such, limited amounts of labelled data may hamper widespread adoption. Moreover, the computational resources required to run CNN models may limit their application in poorer and rural areas.

Non-deep machine learning approaches have not been thoroughly investigated in the existing literature. However, these approaches do not suffer from the limitations noted above, and may therefore prove to be more suitable for use in environments with limited computing resources. This paper presents two non-deep supervised learning systems, using random forests (RFs) and support vector machines (SVMs). This paper differs from existing work using non-deep models by adopting more rigourous testing on a larger set of data, investigating RFs as a potential solution and exploring more image filtering and feature extraction approaches.

Through an initial evaluation process, feature extraction and image filtering approaches were selected to combine with the RF and SVM models. It was found that the SVM model operates significantly better on Haralick texture attributes [4], while the RF model achieves its best performance when combined with histogram extraction. Both systems showed better performance when paired with appropriate image filtering. In particular, isolating the saturation channel of an input image during pre-processing works well with both systems. The SVM system achieves even better performance when applying additional contrast and binary thresholding to this isolated channel.

The RF system outperforms the SVM system, though both improve upon the recall achieved by previous non-deep attempts. Moreover, the RF system provides comparable performance to that offered by the best existing CNN-based approach, achieving accuracy, recall and precision within 4% of that system [12]. These results are achieved when testing on 20 000 images, which is a significantly larger testing set than has been seen in the existing literature. The RF system is also computationally efficient, with an average runtime of less than two seconds on a system without a dedicated GPU. The combination of high accuracy and good computational efficiency suggests that this system may be suitable for low-resource environments such as rural clinics, where the technology is needed the most.

Section 2 examines the existing work around automated malaria diagnosis, while Section 3 discusses the design of the systems presented in this paper.

Section 4 outlines the experimental methodology that was followed. Section 5 discusses the results of the conducted experiments. Finally, Section 6 draws relevant conclusions and suggests future work.

2 Related work

There has been significant research interest around the topic of automated malaria diagnosis, and the existing literature has demonstrated varying levels of success using a wide variety of supervised learning approaches, including deep convolutional networks, transfer learning models and non-deep models.

Rajaraman et al. used CNNs to achieve accuracy of 98.6% and later improved this to 99.5% with an ensemble approach [11][12]. The improved model demonstrated precision of 99.8%. However, these metrics are given at the patient level rather than the cell level, which is more commonly used in the existing literature. Unfortunately, the study used a dataset, provided by the U.S. National Library of Medicine, that consists of images taken under the same conditions, with the same staining method and from the same archive. As such, it is not clear how well the model would generalise to images collected under different conditions. The deep convolutional architecture limits the viability of the model being deployed in areas with scarce computational resources. The ensemble approach of the later attempt exacerbates this problem, as multiple CNNs must be run to get the final result.

Transfer learning is another technique that involves using a CNN for feature extraction before performing classification with an external algorithm. Mehanian et al. [8] propose a solution for malaria diagnosis using transfer learning with a feature extraction CNN and non-deep logistic regression classifier. They achieved sensitivity of 91.6% and specificity of 94.1%. Though the results of their study are not quite as positive as Rajaraman et al., they also tested the model on substantially larger dataset that originates from 12 countries. This suggests that the model developed by Mehanian et al. may be more robust and have more real-world applicability. While the logistic regression classifier is a non-deep supervised learning technique, the reliance on the underlying CNN feature extractor suggests that the real-world application of this approach would still be limited by its computational complexity.

While deep learning approaches have been more prevalent in the existing literature, some authors have used non-deep supervised learning techniques. Diaz et al. [3] are the most successful example, having achieved 94% sensitivity and 99.7% specificity using an SVM based approach. However, the validity of these results must be questioned, as they were obtained when running the model on the full dataset used in the study, including those images used for training. Unfortunately, their study makes use of a private datset, which hinders direct comparisons. Other non-deep approaches have not seen comparable success, and many have adopted flawed evaluation methodologies [6] [17]. As such, they are not discussed in detail here.

3 System design

This section begins by discussing the software used to develop the proposed systems. The feature extraction and image filtering algorithms that were considered are then outlined. Subsequently, the final designs for the systems are presented.

3.1 Software frameworks

The systems proposed in this papers were constructed using Python 3.6 along with several machine learning and computer vision libraries. Specifically, the systems use scikit-learn [10] for the underlying machine learning models. For image processing and feature extraction tasks, two libraries were used, namely Mahotas CV [2] and OpenCV [1].

3.2 Feature extraction algorithms

Feature extraction algorithms can be applied before input is supplied to machine learning models in order to reduce the input dimensionality. By doing so, models can run much faster and, in some cases, achieve higher performance. In this paper, three approaches were considered:

Hu moments Hu [5] proposed a set of seven moment invariants, all of which are invariant to scale, rotation and translation. These features, known as the Hu moments, have been applied broadly to problems in the field of computer vision. For example, Otiniano-Rodriguez et al. [9] achieved over 90% accuracy using the Hu moments as input to an SVM classifier for sign language recognition.

Haralick texture attributes Haralick et al. [4] presented a set of 14 textural features that can be extracted from images to improve image classification accuracy. These features have since been used in a broad range of image classification tasks. For example, Roula et al. [13] used them for classification of prostatic neoplasia in microscopic images of samples taken by needle biopsy.

Histograms Histograms count the number of pixels that fall into a specified number of intensity bins, for each colour channel. Existing works have used histograms for various image classification tasks. For example, Szummer and Picard [16] used histograms as part of a feature set for indoor-outdoor image classification, resulting in 90.3% accuracy.

3.3 Image filters

Applying filters to images before they are passed to machine learning models may serve to accentuate differences in visual characteristics between positive and negative cases. This may, in turn, result in improved model performance. Five image filtering configurations were considered for use in this paper:

1. No filters
2. HSV colour space conversion
3. Saturation channel isolation
4. Saturation channel isolation with contrast

5. Saturation channel isolation with contrast and binary thresholding

Figure 1 demonstrates the effect that each combination has on a sample image of an infected red blood cell. From left to right, we first see the unfiltered image, followed by the HSV converted version. Next, the isolated saturation channel is displayed. Following this, we see the same isolated channel but with added contrast. Finally, we see the application of the threshold function after contrast boosting.

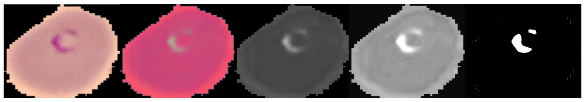

Fig. 1: Image filtering techniques.

3.4 Support Vector Machine (SVM) system

The SVM system uses the scikit-learn support vector classifier for its underlying model. This model uses a third degree polynomial kernel function, as this showed the best performance when compared to radial basis function, linear and fifth degree polynomial kernels. Gamma values between 0.001 and 1 were evaluated, as well as the automatically generated gamma value provided by scikit-learn.

A gamma value of 0.1 demonstrated the highest accuracy, but this was only an improvement of 0.02% over the automatically generated value. The automatically generated value achieved 0.51% higher recall at the cost of some precision. Typically higher recall is preferred in the context of initial screening tests, and so the automatically generated value was selected. Other hyperparameters were set to the default scikit-learn values.

The system extracts Haralick texture attributes as features. Initial evaluations showed that, in combination with the binary thresholding image filtering approach, this provided much higher accuracy than either Hu moments or histogram features. Unfortunately, the computational cost of Haralick feature extraction is much higher than that of these other approaches.

3.5 Random Forest (RF) system

The RF system uses the scikit-learn random forest classifier for its underlying model. During hyperparameter tuning, forest sizes ranging from 1 to 250 were evaluated and peak performance was seen at a value of 100. Maximum depths ranging from 10 to 1000, as well as unlimited depth, were also evaluated. Minimal performance improvements were seen above a value of 100, so it was decided to select this value. Other hyperparameters were set to the default scikit-learn values.

Histogram extraction was adopted as the feature extraction approach for this system. Initial evaluations showed that, in combination with the saturation channel isolation filter, this provided the best classification accuracy and computational performance. Haralick texture attributes resulted in slightly lower accuracy and are far more computationally expensive. On the other hand, Hu moments resulted in poor model performance, though they are not computationally expensive.

4 Experimental methodology

This section lays out the process by which the research aims were addressed. Specifically, two research questions are posed:

1. Can the proposed non-deep systems achieve accuracy within 5% of the top performing CNN approach?
2. Do the performances of models degrade when applied to differently sourced data?

The details of the datasets and computational equipment used are presented, followed by explanations of the two experiments that were conducted.

4.1 Evaluation criteria

Systems are evaluated along three evaluation criteria, namely accuracy, recall and precision. Accuracy refers to the overall ability of a system to make correct predictions, while recall refers to the ability of a system to not label positive examples as negative. In other words, recall is the ability of a system to avoid false negatives. On the other hand, precision refers to the ability of a system to not label negative examples as positive or, in other words, to avoid false positives. Specificity is another metric used by some papers, which measures the ability of a system to correctly identify negative examples. Qualitatively, this is quite similar to precision, though it is calculated differently. Typically studies will not report both specificity and precision, and so it becomes necessary to draw comparisons between these metrics.

4.2 Datasets

Two datasets are used to run the experiments detailed in this study. The first is a publicly accessible dataset, originating from a study by Rajaraman et al. [11] and made available by the U.S. National Library of Medicine (NLM). This dataset is made up of pre-cropped Giemsa-stained blood cell images, with 13 779 classified as infected and the same number classified as uninfected, for a total of 27 558 images. This NLM dataset was also used during the model tuning process.

The second is a private dataset provided to the author by PathCare Laboratory Services. This dataset was provided as Giemsa-stained blood slide images, not pre-cropped but with infected cells having been identified by a pathologist. The dataset was manually cropped into a set of 120 cell images, with 60 labelled as infected and 60 labelled as uninfected.

Images in each dataset were not of a standardised size, and so it was decided to resize all images to 50 pixel by 50 pixel squares. These measurements were decided to avoid artifacts generated from excessive upscaling or stretching in either axis, with most of the original images being about the same size or larger, and roughly square in shape.

Dataset splitting To ensure the validity of results, an iterative random hold-out approach was taken, whereby random subsets of the dataset were selected at each iteration to form the training and test sets. The test set is not used for training models so that when testing occurs, that data is unseen by the model. The metrics achieved for each iteration are aggregated to form the final result.

4.3 Computational equipment

Development, training and testing were all conducted on the same machine: a laptop running Ubuntu 18.04.1 LTS. The machine had 16GB of RAM and a four core Intel i7 CPU, with clock speeds of 2.7GHz. The machine did not have a dedicated graphics card.

4.4 Experiment design

Two experiments were conducted on the proposed systems, allowing for evaluation of various experimental hypotheses.

Experiment 1: Performance of systems on NLM data Systems were evaluated on their ability to operate on data that is part of the same dataset used for training. Results were obtained from ten iterations, each time trained on 5000 images and tested on a disjoint set of 20 000 images. Initial testing indicated that a training set of 5000 images was sufficient for models to converge, so the rest of the data could be used to form a more extensive testing set. This experiment was designed to evaluate the ability of the systems to predict infection in images collected in a similar way to the training data. It was hypothesised that the systems would outperform existing non-deep attempts, and achieve performance within 5% of that demonstrated by the top-performing CNN approach.

Experiment 2: Performance of systems on PathCare test data The systems were put through a final evaluation to test their ability to operate on datasets other than the one used for training. Results were obtained by loading models from the first experiment, pre-trained on the NLM dataset, and tested on the full PathCare dataset of 60 infected and 60 uninfected images. This experiment was designed to evaluate the generalisability of systems trained on a dataset gathered from a singular source, answering the second research question. It was hypothesised that the prediction performance on the PathCare dataset would be worse than that seen in the first experiment, as it is unlikely that the systems will generalise successfully without a broad range of collected data.

5 Results and discussion

In this section, the results of the two experiments are presented. The implications of these results are discussed, and possible explanations are laid out. Finally, limitations of the experimentation process are noted.

5.1 Experiment 1: Performance of tuned models on NLM data

The results observed when testing on the NLM dataset are encouraging. The SVM system achieved an accuracy of 93.06%, recall of 96.12% and precision of 90.58%. The best performing existing approach by Diaz et al. [3] does not report precision but reports specificity of 99.7% and recall of 94%. While the SVM system's precision is significantly lower than the specificity reported by Diaz et al., it also demonstrates a 2.12% improvement in recall. The RF system achieved an accuracy of 96.29%, recall of 96.06% and precision of 96.49%. This amounts to a 2.06% increase in recall, with a 3.68% lower precision than the specificity achieved by Diaz et al. See Figure 2 for a graphical representation of this comparison.

Medical experts argued that attaining higher recall while maintaining comparable overall accuracy is better than attaining higher specificity or precision in the context of initial screening tests [7]. This is because it is far more important to ensure that false negatives do not occur, as these could lead to patients not being given treatment when needed. Following the high sensitivity screening test, another test with higher precision may be performed to ensure overall accuracy. An automated malaria screening system could identify cells with a high probability of infection and present these to a pathologist for confirmation. This reduces the number of cells which must be checked by the expert while still ensuring diagnostic precision. In this sense, the hypothesis that the proposed systems would improve on existing supervised approaches is confirmed, though it is acknowledged that the improvement is not observed across all metrics.

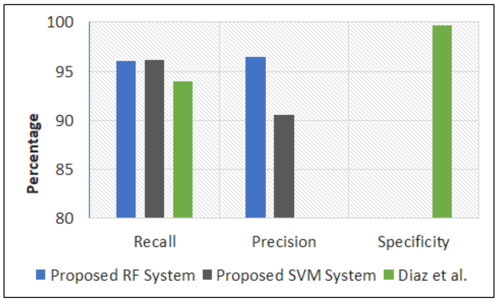

Fig. 2: Comparison of proposed systems with the current best non-deep system.

The current work also evaluates models on significantly more testing data than Diaz et al.: 20 000 blood cells compared to 12 557. Moreover, the parasitemia of the testing data used by Diaz et al. is reported as 5.6%, amounting to approximately 703 infected blood cells. On the other hand, the testing data used in this experiment was balanced, with 10 000 infected cells, possibly resulting in a more accurate reflection of the real-world performance that can be expected from the system.

As expected, the proposed systems did not achieve better results than the CNN approaches detailed by Rajaraman et al. [11][12]. However, the RF model's metrics all fell within 4%, and the SVM achieved recall within 4%. These results, particularly those of the RF system, confirm the hypothesis that the proposed systems would show performance within 5% of the best CNN-based approach. Mehanian et al. [8] achieved recall of 91.6% and specificity of 94.1% with their transfer learning model. Both the SVM and RF systems improve significantly on this recall, and the RF also shows higher precision than the specificity reported by Mehanian et al. The testing set used in this paper was significantly larger than those used by both Mehanian et al. and Rajaraman et al., though that used by the former was more diverse, with images taken under different conditions and originating from 12 countries. See Figure 3 for a graphical representation of this comparison.

Papers in the existing literature have not reported training or testing time metrics, so it is not possible to provide a comparison here. However, comparing the RF and SVM systems proposed in this paper, it is clear that the RF system is significantly less computationally expensive. The testing time of the SVM system is a factor of 236 greater than that of the RF system, while the training time is a factor of 53 greater. These runtime metrics suggest that the RF system may be more suitable for deployment in situations where computational power is limited. Table 1 shows the full results of this experiment.

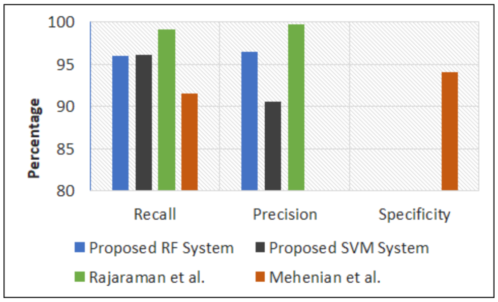

Fig. 3: Comparison of proposed systems with existing deep learning approaches.

System	Accuracy	Recall	Precision	Train Time	Test Time
SVM	0.9306	0.9613	0.9058	63.9346s	254.9055s
RF	0.9629	0.9606	0.9649	1.2108s	1.0796s

Table 1: Results of Experiment 1.

5.2 Experiment 2: Performance of tuned models on PathCare data

Both the SVM and RF systems saw slight improvements in accuracy when run on the PathCare dataset. While the precision of both decreased, by 1.03% for the SVM system and 2.74% for the RF system, each achieved 100% recall. It must be noted that these results are likely overly optimistic due to the small size of the PathCare dataset, however this encourages further evaluation of the systems on a larger PathCare dataset.

The runtime metrics of this experiment confirm the observation made in the first experiment: The RF system is more computationally efficient than the SVM system. While train times are not evaluated, as the systems were pre-trained, the test time of the SVM system is a factor of 128 greater than that of the RF system. Table 2 shows the full results of this experiment.

System	Accuracy	Recall	Precision	Train Time	Test Time
SVM	0.9417	1.000	0.8955	n/a	2.053s
RF	0.9667	1.000	0.9375	n/a	0.016s

Table 2: Results of Experiment 2.

5.3 Limitations

The small size of the PathCare dataset limits the acceptability of the results of the second experiment. While these results were very positive as an initial evaluation, further experimentation on a larger set of PathCare data would be necessary to confirm them.

The data that was made available for this research was only classified in two classes: as parasitised or non-parasitised. However, in reality there are multiple species of the malaria parasite, each with various life stages. Because the data was only classified in this binary manner, it is unclear whether the models generalise to provide similar performance for all species and life stage combinations. Unfortunately, it is time-consuming to manually label data in this manner and the labelling must be conducted by an expert pathologist. For this reason, it is difficult to acquire large enough datasets to allow for adequate training and testing of models.

Both datasets were made up of images of Giemsa-stained blood cells. It is not clear whether the systems would maintain similar performance when different staining methods are employed. Giemsa staining is the method recommended by the World Health Organisation [18], but some clinics in lower-income areas may vary from this. Large datasets of blood cells not stained with Giemsa are difficult to acquire, as it is uncommon for major pathology practices to use other staining methods.

We tested a range of values for each hyperparameter during model tuning, however, due to limited computational resources, a fairly coarse search was performed. It is possible that a finer grid search or more sophisticated intelligent search techniques may find better configurations.

6 Conclusions

The experiments presented in this paper demonstrate that non-deep supervised learning techniques may be used as an alternative to popular deep learning approaches for malaria classification. Several conclusions can be drawn from the results of these experiments.

Firstly, the proposed RF and SVM systems outperform existing non-deep supervised approaches in terms of recall, which is arguably the most important metric by which medical screening tests can be judged. Secondly, the RF system appears to be a more suitable solution, as it achieves better accuracy and is far more computationally efficient than the SVM system. Thirdly, the proposed RF system is able to achieve recall, precision and accuracy within 4% of that reported by the best existing CNN approach. This indicates that the system may be a viable alternative in situations where the high computational power required by CNN systems is not possible, such as in rural clinics. Fourthly, initial results when run on a small dataset gathered from a different source to the training data seem to indicate that the promising performance demonstrated by the proposed systems may generalise to various imaging conditions. However, further work is necessary to confirm this.

Finally, it is noted that feature extraction and image processing can significantly impact both the computational efficiency and classification performance of systems. Extracting histogram features is computationally cheap and greatly improves the accuracy of the RF model, but the SVM model saw better classification performance when operating on Haralick texture attributes. However these attributes are computationally expensive to extract, limiting the viability of the SVM system. Hu moments performed the worst for both SVM and RF models, suggesting they may be less suitable for the task of malaria detection.

On the whole, it is shown that non-deep supervised learning techniques have great promise for reducing the burden on medical professionals in performing malaria diagnosis. This may, in turn, result in much faster times to diagnosis, and allow quicker intervention, which is described by the WHO as the most important factor in preventing severe cases and deaths from occurring [19]. The higher computational efficiency of non-deep systems, such as those presented in this paper, may allow for more widespread adoption, especially in areas where extensive computational resources are not available. Thus, the value added by these systems may have a significant impact on rural and poor communities, where access to medical experts is typically limited.

6.1 Future work

The systems proposed in this paper show promise as a computationally cheap alternative to CNN-based systems, with a lower training data requirement. As such, future work to develop a fully integrated diagnosis system is warranted. Such a system should include the full computational pipeline of automated blood cell image cropping, filtering, feature extraction and classification.

Rajaraman et al. [12] improved the performance of their initial CNN approach by adopting an ensemble strategy. Similarly, future work may produce better results by using an ensemble of high performing non-deep models, such as those proposed in this paper.

7 Ethics

This paper involved the use of both a publicly-available dataset from the U.S. National Library of Medicine, as well as a private dataset acquired from PathCare Laboratory Services. Both datasets are de-identified, containing no personal or demographic information on the patients corresponding to each blood cell image. Ethics clearance was granted by both the University of Cape Town's Science Faculty Research Ethics Committee and the PathCare Research Committee.

8 Acknowledgements

Thanks are extended to PathCare Laboratory services for their contribution of blood sample data.

References

1. Bradski, G., Kaehler, A.: Learning OpenCV: Computer vision with the OpenCV library. " O'Reilly Media, Inc." (2008)
2. Coelho, L.P.: Mahotas: Open source software for scriptable computer vision. arXiv preprint arXiv:1211.4907 (2012)

3. Díaz, G., González, F.A., Romero, E.: A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images. Journal of Biomedical Informatics **42**(2), 296–307 (2009)
4. Haralick, R.M., Shanmugam, K., et al.: Textural features for image classification. IEEE Transactions on systems, man, and cybernetics (6), 610–621 (1973)
5. Hu, M.K.: Visual pattern recognition by moment invariants. IRE transactions on information theory **8**(2), 179–187 (1962)
6. Khan, N.A., Pervaz, H., Latif, A.K., Musharraf, A., et al.: Unsupervised identification of malaria parasites using computer vision. In: 2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE). pp. 263–267. IEEE (2014)
7. Lalkhen, A.G., McCluskey, A.: Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia Critical Care & Pain **8**(6), 221–223 (2008)
8. Mehanian, C., Jaiswal, M., Delahunt, C., Thompson, C., Horning, M., Hu, L., Ostbye, T., McGuire, S., Mehanian, M., Champlin, C., et al.: Computer-automated malaria diagnosis and quantitation using convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 116–125 (2017)
9. Otiniano-Rodrıguez, K., Cámara-Chávez, G., Menotti, D.: Hu and zernike moments for sign language recognition. In: Proceedings of international conference on image processing, computer vision, and pattern recognition. pp. 1–5 (2012)
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)
11. 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 **6**, e4568 (2018)
12. Rajaraman, S., Jaeger, S., Antani, S.K.: Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images. PeerJ **7**, e6977 (2019)
13. Roula, M., Diamond, J., Bouridane, A., Miller, P., Amira, A.: A multispectral computer vision system for automatic grading of prostatic neoplasia. In: Proceedings IEEE International Symposium on Biomedical Imaging. pp. 193–196. IEEE (2002)
14. Russell, S.: The economic burden of illness for households in developing countries: a review of studies focusing on malaria, tuberculosis, and human immunodeficiency virus/acquired immunodeficiency syndrome. The American journal of tropical medicine and hygiene **71**, 147–155 (2004)
15. Shillcutt, S., Morel, C., Goodman, C., Coleman, P., Bell, D., Whitty, C.J., Mills, A.: Cost-effectiveness of malaria diagnostic methods in sub-saharan africa in an era of combination therapy. Bulletin of the World Health Organization pp. 101–110 (2008)
16. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: Proceedings 1998 IEEE International Workshop on Content-Based Access of Image and Video Database. pp. 42–51. IEEE (1998)
17. Tek, F.B., Dempster, A.G., Kale, I.: Parasite detection and identification for automated thin blood film malaria diagnosis. Computer vision and image understanding **114**(1), 21–32 (2010)
18. WHO: Giemsa staining of malaria blood films (2018)
19. WHO: World malaria report 2018. World Health Organization (2018)