Detecting Malaria Parasitized and Malaria Cells in Thin-Blood Smear Microscopy Images

Advisor: Dr. Edward Kim, Associate Professor Department of Computer Science Drexel University
Team: Jithin Thenasseril (687), Group 20

Abstract
Malaria is a disease caused by Plasmodium parasite which infects red-blood cells (RBCs) and gametocytes which still affects parts of the developing world. Currently the gold-standard for diagnosis is manual identification and counting of the parasite in blood smear images by microscopists which is accurate but adversely impacted by large-scale screening especially in resource-strained settings. The current state-of-the-art is using Convolutional Neural Networks (CNNs) to detect infection which has an accuracy of over 95% but also high variance. The goal was to use classifiers other than CNN to reduce variance while maintaining a similar level of accuracy. Three classifiers were used in this project and the Artificial Neural Network (ANN) had the best performance with values of 92.81% and 96.25 for accuracy and F-score respectively. The accuracy was slightly lower than CNN but beat the CNN's F-score of 91.7.

Introduction
Malaria is a potentially fatal disease caused by 5 types of the Plasmodium parasite which is transmitted through the bite of the Anopheles female mosquito the most prevalent of which is the P.vivax[2]. According to the World Health Organization's 2018 Malaria report, there were 219 million reported malaria cases worldwide in 2017 and 435,000 deaths predominantly in African and Southeast Asian countries[2]. While the cases of malaria have dropped overall since 2010 due to funding of labs and public health initiatives in developing countries by the United Nations, there was a slight increase beginning 2015 (Fig. 1) attributed to warming climates[2] which allows mosquitoes to live in environments that were previously unsuitable and reproduce in larger numbers.

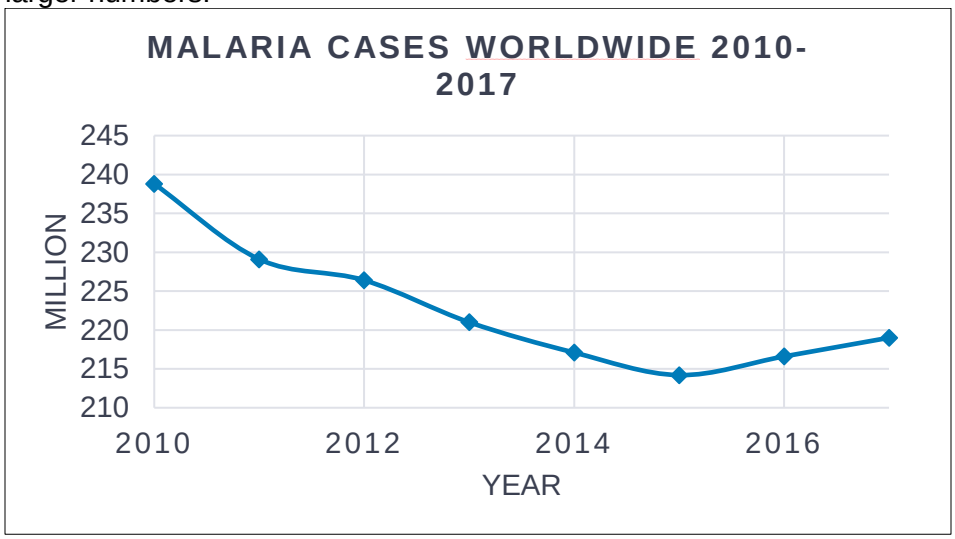

Fig. 1. Number of worldwide malaria cases from 2010-2017 [2]. There was a slight uptick beginning 2015 due to warming climate and limited resources for disease screening.

Currently, the best method for diagnosing malaria is manual identification and counting of malaria infected cells and malaria cells themselves in thin-blood smear microscopy images by trained microscopists. While this method is highly accurate, the accuracy can be adversely impacted by large-scale screening especially in resource-strained settings [1]. Researchers believe that the negative impact on accuracy could be mitigated by using machine learning techniques to assist in disease screening. The current state-of-the-art for computer-aided malaria diagnosis is Convolutional Neural Networks (CNNs) due to its success with many other image classification problems and while it has proven to be highly accurate (>95%), it suffers from high variance [1][4][5]. This suggests that CNN's are prone to overfitting at least for malaria image datasets. The goal of this project was to use classifiers other than CNNs to get a similar level of accuracy to CNNs but less prone to variance than the current state-of-the-art. The images that were used for this project were from the Broad Institute's Bioimage Repository (BBBC)[3] which contained 1208 wholeslide images (two are shown in Fig. 2). Wholeslide images are images of the prepared microscope slide as seen by a microscopists when looking through a microscope.

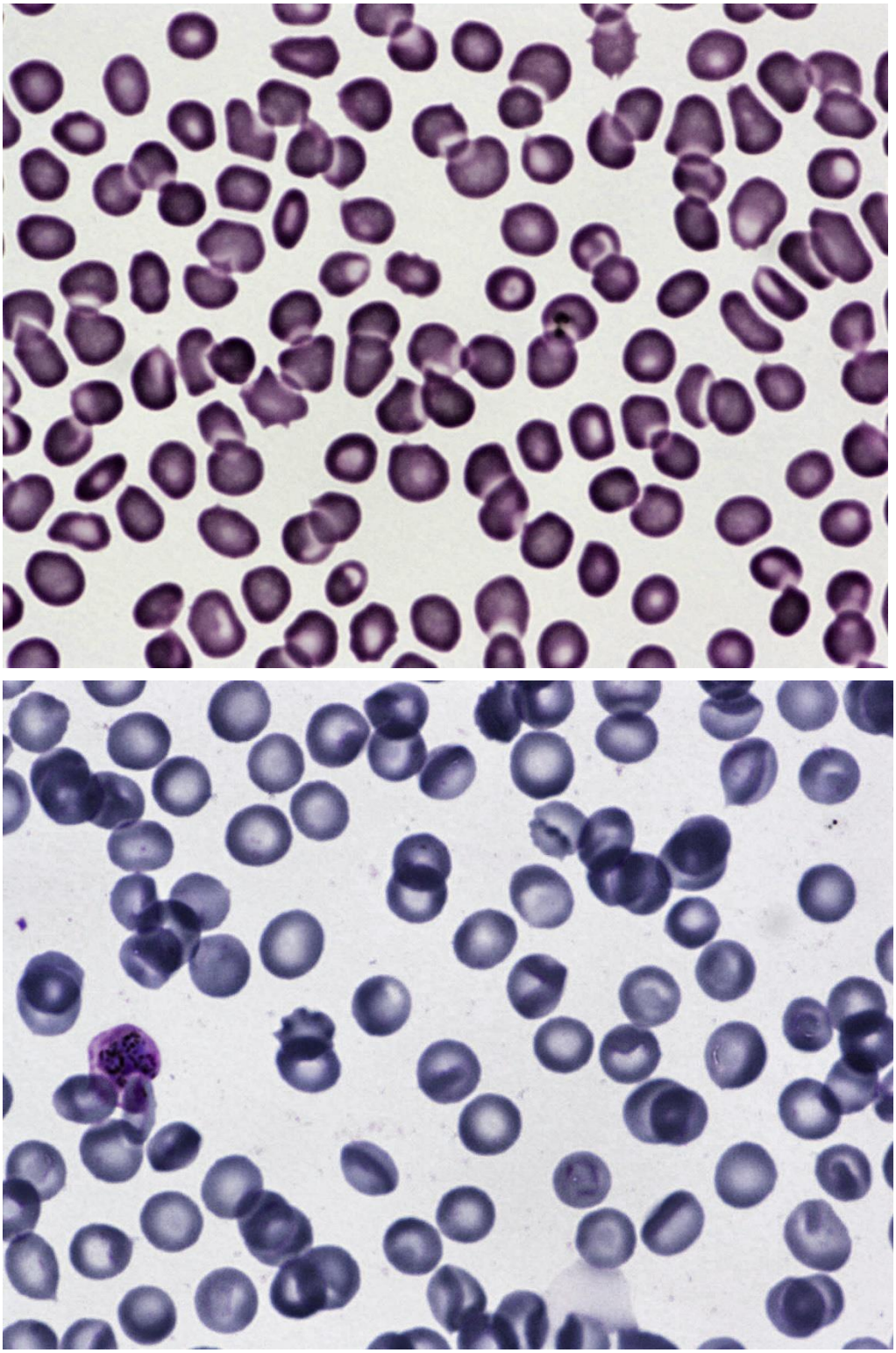

Fig. 2 sample wholeslide microscopy images from the Broad Institute Bioimage Repository [3]. The two images above are representative samples of the 1208 wholeslide images that make up the malaria dataset and show that the dataset didn't contain images of single isolated cells but a mix of multiple human and malaria parasite cells from which features would need to be extracted for classification.

The wholeslide images were a mix of uninfected red-blood cells (RBCs) (Fig. 7), uninfected while-blood cells (leukocytes) (Fig. 8), infected RBCs called rings (Fig. 5), infected gametocytes (Fig. 3) and the P.vivax malaria parasite cells in 2 lifecycle development stages called schizonts (Fig. 6) and trophozoites (Fig. 4) which would be classified as infected if detected.

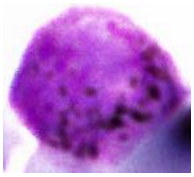
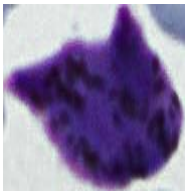

Fig. 3 (top & bottom). P.vivax infected human gametocyte

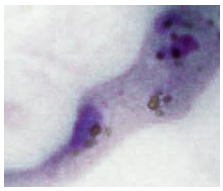


Fig. 4 (top & bottom). P.vivax trophozoite, a lifecycle stage of the P. vivax [6].

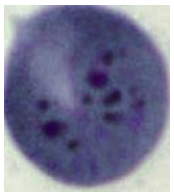
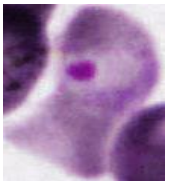

Fig. 5 (top & bottom). P.vivax infected red-blood cell (ring).


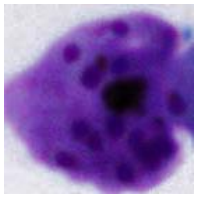

Fig. 6 (top & bottom). P.vivax schizont, a fully matured stage of the P.vivax [6].

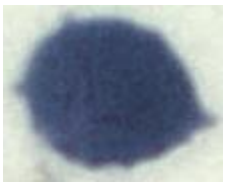
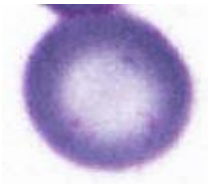

Fig. 7 (top & bottom). Uninfected red-blood cell (RBC)

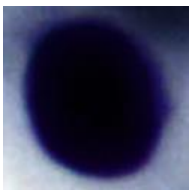
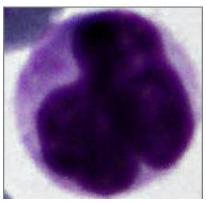

Fig. 8 (top & bottom). Uninfected white-blood cell (leukocyte)

The cell types varied in shape, size, and color not just among different types but within the same type as well as shown in Fig. 3-8 which added an additional layer of complexity because features that can help to identify a cell type within one wholeslide image may not be available in a different image. It also eliminates shape, size, and color as potential features for cell classification.

Related Work

This project used a combination of Scale Invariant Feature Transform (SIFT) algorithm, local binary pattern (LBP) algorithm, and Principal Component Analysis (PCA) on features extracted by the SIFT and LBP algorithms. PCA performed on SIFT features alone will be referred to as PCA-SIFT and PCA performed on features extracted by SIFT on LBP images will be referred to as PCA-LBP-SIFT in this paper. While the literature researched for this project used the individual feature extractors on some combination of classifiers, they didn't use the unique combination of SIFT, LBP, and PCA extracted and selected features on Random Forest Classifier, Support Vector Machine (SVM) with linear and radial basis function (RBF) kernels, and a linear artificial neural network (ANN) with hidden ReLu layers and cross-entropy loss. For example, the research conducted by Das, Ghosh, Pal et. al[8] used LBP images on Bayesian Learning machines and SVM which yielded an accuracy of 84%. More importantly, they focused solely on erythrocyte (red-blood cell) size, shape, and texture to classify wholeslide images as infected while size and shape was ruled out as significant features for classification due to deformation and varying size of the cells in the BBBC malaria dataset used for this project. The most popular classifier by far were versions of CNNs that relied on the network itself to perform feature extraction and labeling. The research conducted by Rajaraman, Jaeger, Antani used an ensemble of classifiers called VGG-19 and Squeezenet which are CNN classifiers and didn't rely on separate feature extractor algorithms like SIFT[1] and succeeded in creating a model that had higher accuracy and lower variance than the state-of-the-art CNNs but there were 2 reasons for not using CNN for this project: 1) the dataset that Rajaraman et al. used were images of single cells and therefore only required models to classify the cell type rather than relying on feature extractors to extract cell features from wholeslide images followed by cell classification 2) the classifier ensemble they created would require the microscopists to create and provide isolated cell images for classification. If the microscopists has the time to isolate cells, then they can classify it themselves without requiring machine learning altogether. If a trained microscopists isn't available and cell isolation is done by a lab technician untrained in malaria identification, then I can see the value in having a computer-aided tool capable of cell classification. However, the goal of this project was to create a model that could easily be deployed in the field and reduce the burden in resource-strained areas and two of the strained resources are time and trained lab technicians. In addition, they did acknowledge that their ensemble classifier requires powerful computational equipment which may not be available or feasible to deploy in certain areas. They did however come up with a solution: a web-based app to which isolated image cells could be uploaded for classification by remote computers. However, one of the issues faced in developing parts of the world is unreliable telecommunications and electric infrastructure. For these reasons, I didn't consider their research to be state-of-the-art since their objective was different from other literature and this project, namely create a computer-aided tool that could replace microscopists trained in malaria cell identification and reduce time and burden undertaken by lab technicians. Lab technicians would still be

required to prepare thin-blood smear slides and photograph them but this requires relatively little training and time compared to isolating individual cells. Poostchi, Silamut, Maude et al.[6] aggregated several research papers that used a combination of several feature extractors and classifiers on thin-blood smear malaria microscopy images which includes all of the techniques used for this project. However, none of the research papers they covered used the unique combination of feature extractors and classifiers that were used in this project.

Experimental Results

The first step towards classifying the wholeslide images was to identify if a given wholeslide image contained an infected red-blood cell (RBC) called a ring, an infected gametocyte, or a schizont or trophozoite which are the Plasmodium parasite cells themselves at different stages of development responsible for causing malaria. Ideally, there would be a linear separation of the infected cells that a learning machine could use so the red, green, blue (RGB) characteristics of the infected cells shown in Fig. 3-6 were plotted to check for linear separation. A 3D plot of the RGB characteristics as well as 2D plots for each color mix were plotted as shown in Fig. 9 and Fig. 10 respectively.

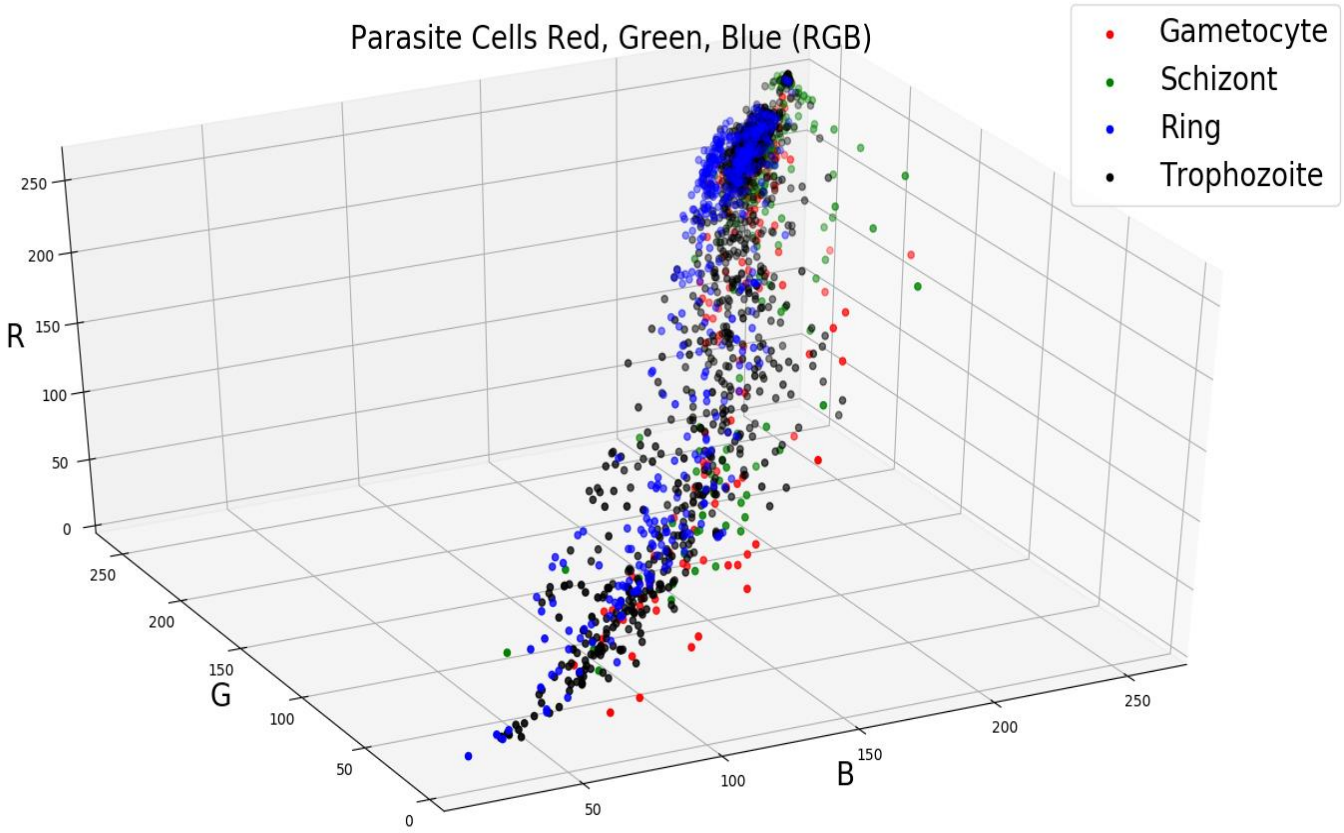

Fig. 9. Parasitized Cells RGB 3D plot. The red, green, blue (RGB) characteristics of the infected cells in Fig. 3-6 were plotted to check if there is a linear separation between the infected cells that could be used by a learning machine to classify cells. The 3D plot confirmed there is none.

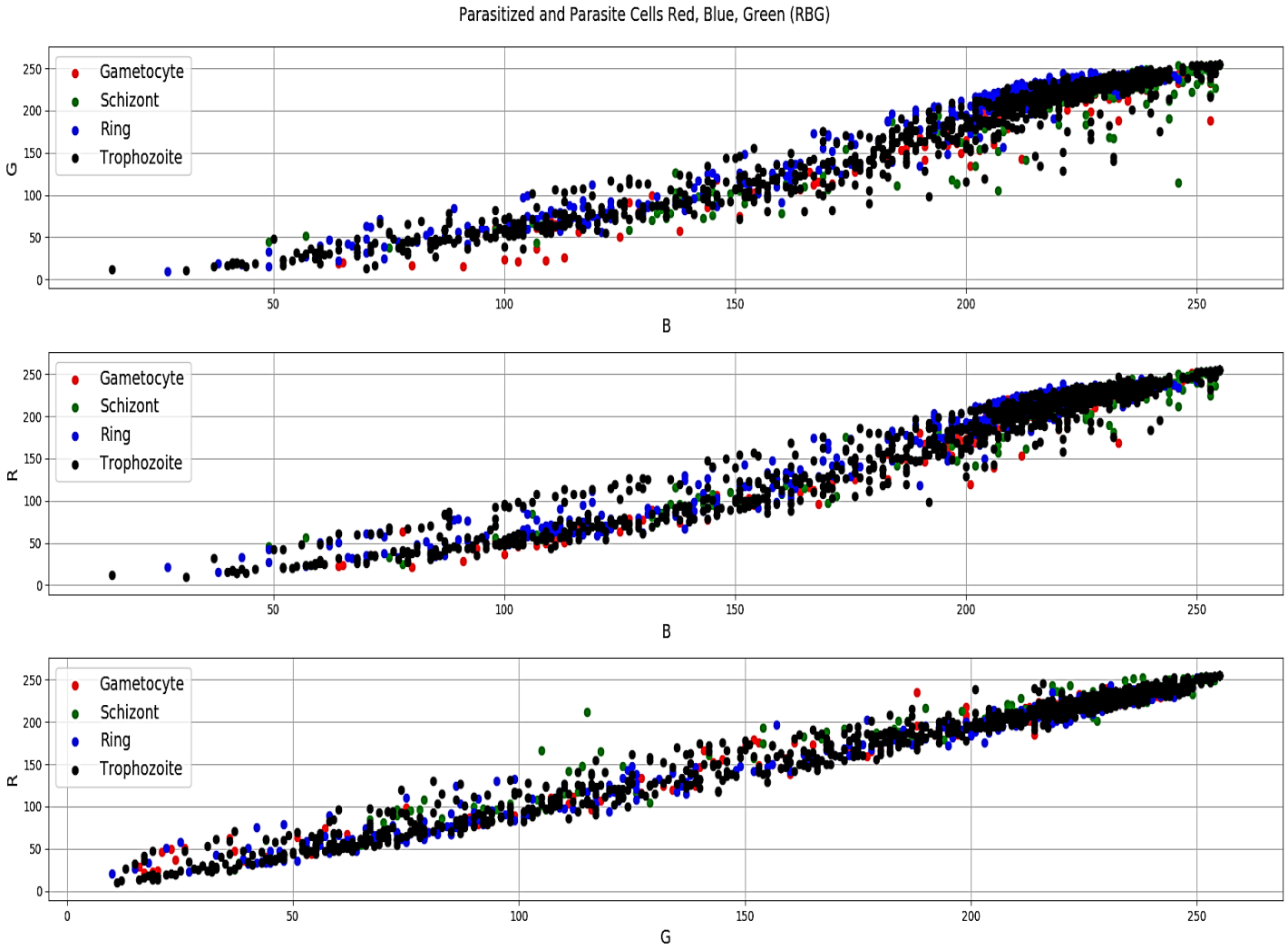

Fig. 10. Red, blue, green (RBG) 2D plots of the infected cells shown in Fig. 3-6. The 2D RGB points of the infected cells were plotted to check if there is a linear separation between infected cells that could be used for classification of the infected cells.

Fig. 9-10 demonstrated that there wasn't a linear separation that could be used for classification. Since the images were wholeslide images, a feature extractor had to be used to identify if one of the infected cells existed in the wholeslide image and if it did, classify the image as infected. The SIFT and SURF algorithms were considered for feature extraction and SIFT was chosen because it did a better job of extracting features from infected cells while ignoring uninfected RBCs, white blood cells (leukocytes), and artefacts such as stain droplets and platelets. It's also less prone to illumination and viewpoint [7] which were vital characteristics for this project since the images are light microscopy images with cells having varying translucency. SIFT was run on grayscale wholeslide images (Fig. 11) because as Fig. 3-6 demonstrate, the color can vary even if the cell type is the same and Fig. 9-10 demonstrate that RGB characteristics don't provide useful information for classification.

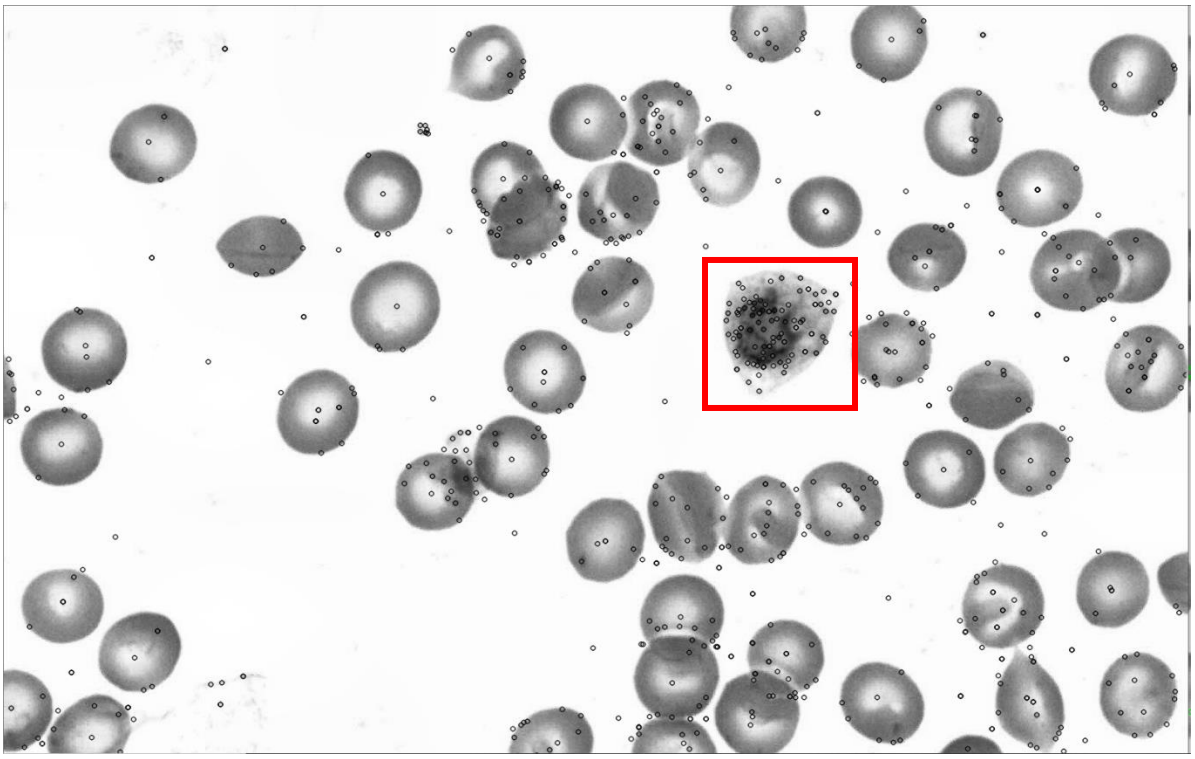

Fig. 11. SIFT calculated keypoints. The image shows as black circles the keypoints calculated by the SIFT algorithm. The algorithm did an excellent job of identifying features in an infected gametocyte which is highlighted in the red box while extracting minimal features from uninfected red-blood cells.

The SIFT algorithm outputs a 128x1 vector for each calculated keypoint that is a 3D spatial histogram which could then be passed to a learning machine for classification. The normalized histograms generated by SIFT for the infected cells shown in Fig. 3-6 are shown in Fig. 12. Principal Component Analysis (PCA) was run on the SIFT extracted features and 66 components contained 93.13% of the data for SIFT as shown in Fig. 15.

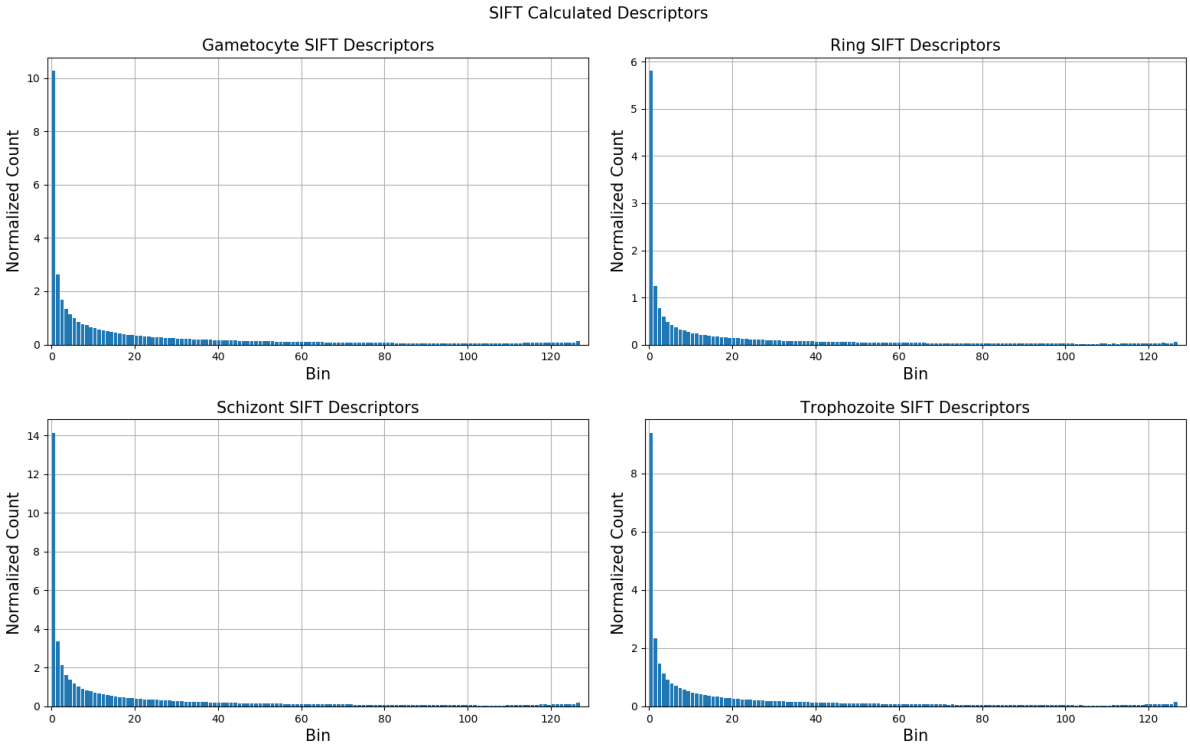

Fig. 12. Histograms of normalized SIFT calculated descriptors of infected gametocytes, rings, schizonts, and trophozoites. The histogram reveals that the infected cells are very similar but there does appear to be a way to classify the different cell types based on bins 1-20; schizonts have the largest and most dark regions, followed by gametocytes, trophozoites, and rings which is consistent with what can be confirmed visually using Fig. 3-6.

Local Binary Pattern (LBP) was also run on the wholeslide images and the features extracted using keypoint pixel locations for additional information about cell texture that may assist in classifying cells. Fig. 14 shows the output of LBP algorithm run on the image shown in Fig. 11 without the SIFT keypoints.

Fig. 14. Sample Local Binary Pattern (LBP) image. Resultant image of Fig. 11 after it's run through the LBP algorithm. The red box highlights the texture of an infected gametocyte which appears to be primarily concave and rougher than neighboring uninfected red-blood cells.

PCA was also run on SIFT-LBP extracted features and 190 of the components contained 93.43% of the information. The number of PCA components versus information gain (variance ratio) for PCA-SIFT and PCA-SIFT-LBP is shown in Fig. 15.

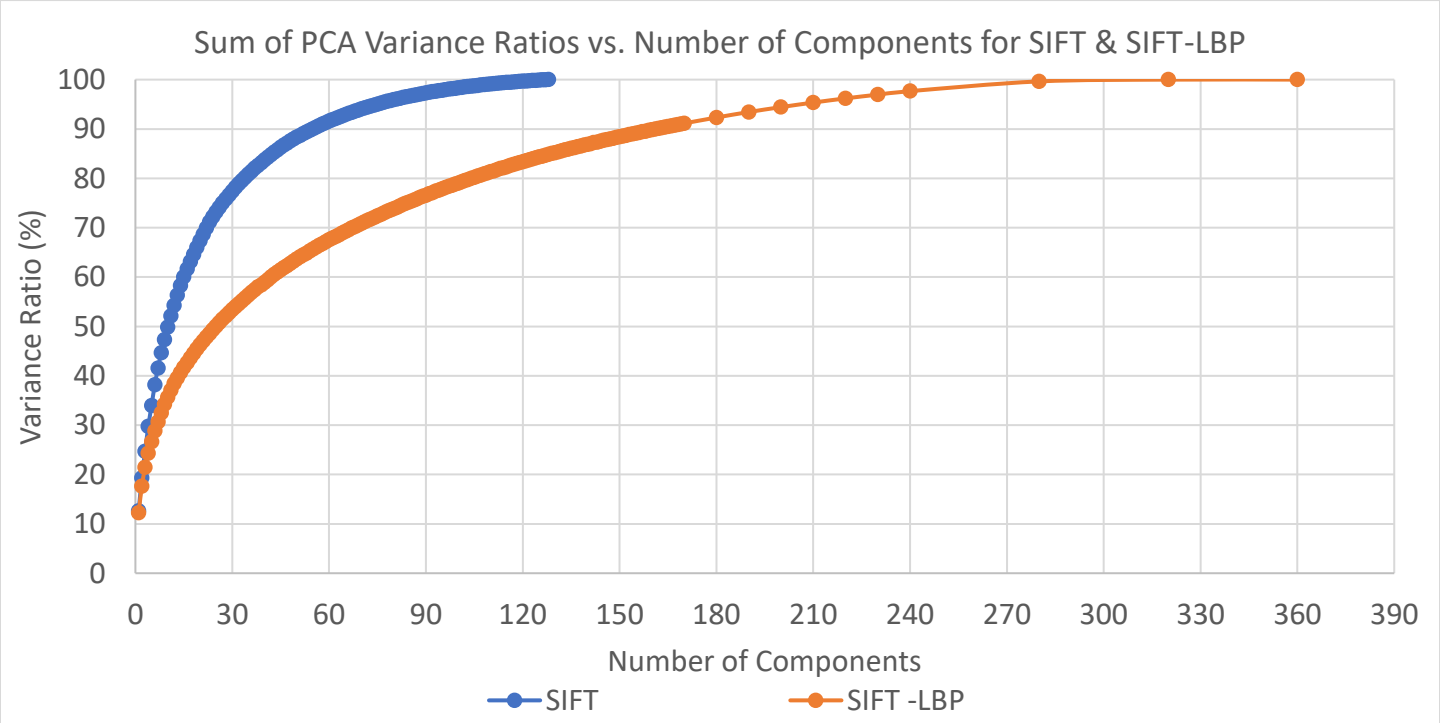

Fig. 15. Variance Ratio (%) vs. number of PCA components for SIFT and SIFT-LBP extracted features. The graph demonstrates that 66 components in SIFT and 190 components in SIFT-LBP contain approximately 93% of the information. Beyond these components, the information gain is less than 1% per component.

The SIFT, SIFT-LBP, and PCA-SIFT extracted and selected features were passed to SVM that used a radial basis function kernel (RBF) (1) but the machine failed to converge for either feature set of the training data even with multiple machines running in parallel so it was abandoned.

$$g(x) = w_0 + \sum_{i=1}^{k} w_i * \exp\left(-\frac{(x - c_i)^T (x - c_i)}{2{\sigma_i}^2}\right) \quad (1)$$

The next machine that I tried was a Random Forest Classifier on SIFT, SIFT-LBP, and PCA-SIFT-LBP extracted features. The Random Forest Classifier is simply multiple decision trees that take a sample of the training data and assign the sample to a class. A simple majority vote is then taken using the results of each tree to assign a class. The node impurity used was the Gini impurity given by (2) which was used to determine the information gain from each split.

$$G = \sum_{i=1}^{C} p_i * (1 - p_i) \quad (2)$$

Two parameters were adjusted in the Random Forest Classifier for increased accuracy, number of trees and the number of samples per leaf. The number of trees had no bearing on accuracy but Fig. 16 shows the effect of the number of samples per leaf on the accuracy of the classifier on features extracted using SIFT and SIFT-LBP. The accuracy with SIFT features alone was 73.25% and F-score was 83.29 with a Random Forest of 5 trees and 1 sample per leaf.

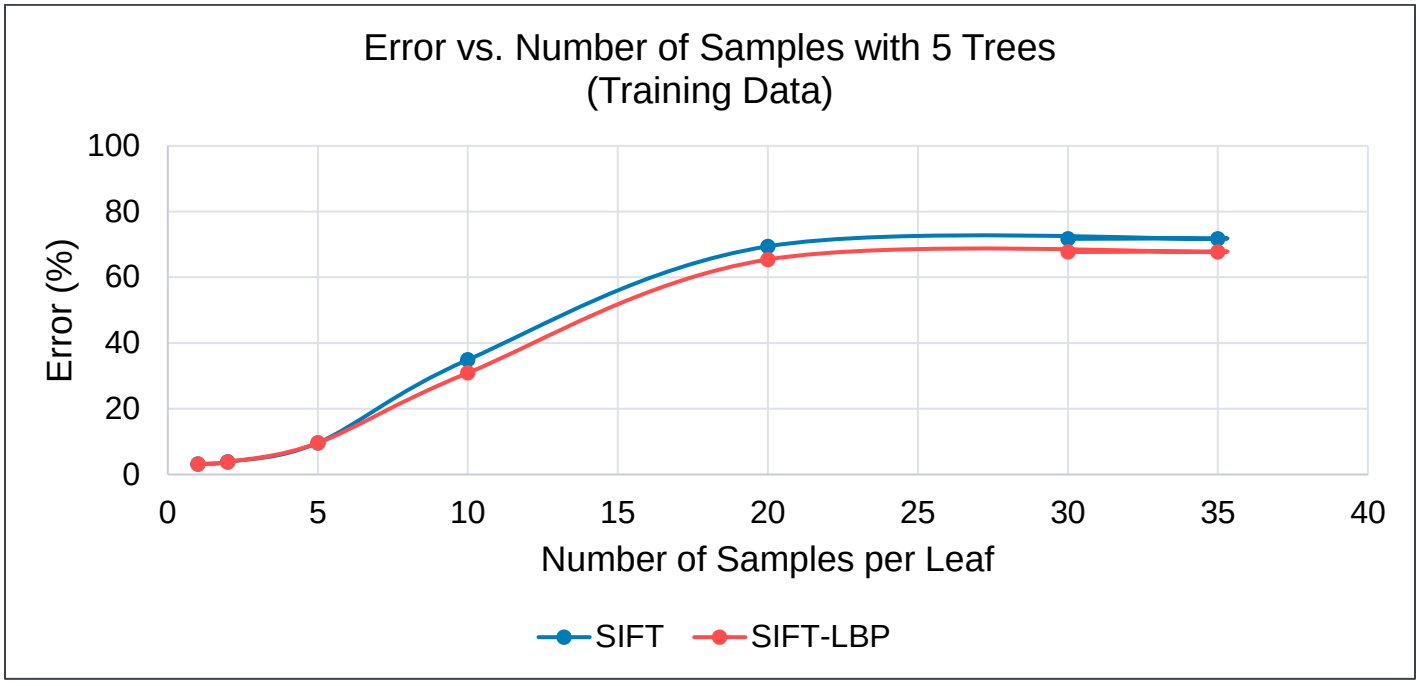

Fig. 16. Error Rate (%) vs. Number of Samples per Leaf for each tree in the Random Forest Classifier. The graph demonstrates that as the number of data samples that each leaf was able to contain decreased, the accuracy of the classifier increased because the tree would have to make more decision splits. 1 sample per leaf had the lowest error for both SIFT and SIFT-LBP extracted features at approx. 3.2 and 3.8% respectively.

The most accurate classifier that was used in this project was the linear artificial neural net (ANN) whose final version and highest accuracy consisted of a 384-node input layer with ReLu activation function, 5 hidden ReLu dense layers which had approximately half the number of nodes as the immediate previous layer, and an output layer that uses the Softmax activation function given by (3) in which *N=2* for my particular ANN. The softmax activation function proved to have higher accuracy than ReLu, sigmoid, or tanh at the output layer. The loss was calculated using negative log likelihood, which usually works well with softmax[1], which was then backpropagated through the ANN.

$$S(y_i) = \frac{\exp(y_i)}{\sum_{j=1}^{N} \exp(y_j)} \quad (3)$$

The final accuracy was on test data of the ANN was 92.81 with a F-score of 96.25 on the feature set extracted by SIFT-LBP as shown in Table 1.

Discussions and Conclusions

The results of the SVM, Random Forest Classifier, and linear ANN are summarized in Table 1 on features extracted using SIFT, combination of SIFT and LBP (SIFT-LBP), and features selected using PCA (PCA-SIFT and PCA-SIFT-LBP).

Table 1. Results of various classifiers on features extracted and selected using PCA, SIFT, and LBP algorithms on testing data. PCA-SIFT-LBP had highest accuracy for Random Forest and SIFT-LBP for linear ANN.

	SVM	Random Forest	Linear ANN
SIFT Accuracy (%)	NA	73.25	92.77
SIFT F-Score (%)	NA	83.29	96.25
SIFT-LBP Accuracy (%)	NA	78.19	**92.81**
SIFT-LBP F-Score (%)	NA	86.85	**96.25**
PCA-SIFT Accuracy (%)	NA	73.25	92.77
PCA-SIFT F-Score (%)	NA	82.94	96.25
PCA-SIFT-LBP Accuracy (%)	NA	**81.89**	92.81
PCA-SIFT-LBP F-Score (%)	NA	**89.32**	96.25

The SVM wouldn't converge with a linear or RBF kernel even after using PCA-SIFT which only passed 66 components rather than 128 which is why 'Not Applicable' (NA) is listed in Table 1. The Random Forest Classifier used Gini impurity to calculate the information gain from each split and based on Table 1, best accuracy of 81.89% was obtained using 256 PCA-SIFT-LBP extracted and selected features which is higher than accuracy obtained with SIFT-LBP. 81.89% accuracy was similar to accuracy values obtained by Das, Ghosh,Pal, et al.[8] who used SVM to get 85% accuracy. I believe the reason Random Forest was able to obtain better results with less information was because PCA selected the most relevant components that provided maximum information and eliminated the rest making it the model less prone to overfitting and allowed the classifier to learn the subtle differences in the relevant components.

Trying to tune SVM and Random Forest classifier to obtain higher accuracy taught me that perhaps the data was too complex and close together for further separation using either machine and so I turned to a linear artificial neural network (ANN). I initially only used SIFT extracted features which provided an accuracy of 92.77% as shown in Table 1 but couldn't get it higher despite changing the activation and loss functions. I hypothesized that SIFT may not be capturing all the features required for classification so I tried LBP to extract cell tecture based on success of other researchers[6]. The accuracy did improve by 0.04% to 92.81% which indicated that perhaps LBP wasn't providing a lot more information than SIFT was already extracting. SIFT-LBP proved to have the highest accuracy with linear ANN at 92.81% although PCA-SIFT-LBP was very close. The reason SIFT-LBP was chosen over PCA-SIFT-LBP even though they have the same accuracy and F-score in Table 1 was because accuracy was 10^{-5}% lower for PCA-SIFT-LBP which isn't bad considering a third of the 384 components were eliminated. Table 2 shows the cell types that the ANN had trouble classifying as infected; my hypothesis is that the infected cells in those particular images just don't have enough features for extraction although I was unable to confirm due to time constraints.

Table 2. Number of images correctly classified as infected by linear ANN based on cell type present in the image. The ANN had trouble classifying images with trophozoites and rings.

Cell Type	# of Images Correctly Classified	# of Images Present	% Accuracy
Gametocyte	136	136	100
Trophozoite	540	596	91
Ring	218	244	89
Schizont	157	157	100

The accuracy was also dependent on the activation and loss function to a certain extent. ReLu, leaky ReLu, sigmoid, and tanh were all used as activation functions with negative log likelihood, sigmoid, and tanh used as loss for backpropagation and I found that the ReLu activation function combined with softmax activation

function (3) at the output layer and negative log likelihood function for backpropagation provided the highest accuracy.

This project was not related to my current or previous research however I intend to continue trying to improve the current model using an ensemble of classifiers. If I can get it to a level that's ready for the real-world, I could try to reach out to NGOs operating in India or the Indian government to get them to use the model in their ongoing fight against malaria. This was my first machine learning project and I haven't explored all the options that SVM and Random Forest have or even all the different types of neural networks that others have built and how they would perform on this dataset so I definitely plan to improve on this project as I gain understanding of existing learning machines and the concepts we learned in class.

References:

[1] Rajaraman S, Jaeger S, Antani SK. 2019. Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images. PeerJ 7:e6977 https://doi.org/10.7717/peerj.6977 [Accessed 19 Oct. 2019].

[2] World Health Organization (2018). World Malaria Report 2018. [ebook] United Nations World Health Organization. Available at: https://www.who.int/malaria/publications/world-malaria-report-2018/en/ [Accessed 25 Nov. 2019].

[3] Data.broadinstitute.org. (2019). BBBC041: P. vivax (malaria) infected human blood smears. [online] Available at: https://data.broadinstitute.org/bbbc/BBBC041/ [Accessed 19 Oct. 2019].

[4] J. A. Quinn, R. Nakasi, P. K. Mugagga, P. Byanyima, W. Lubega, A. Andama, "Deep convolutional neural networks for microscopy-based point of care diagnostics", 2016.

[5] F. B. Tek, A. G. Dempster, I. Kale, "Parasite detection and identification for automated thin blood film malaria diagnosis", Journal of Computer Vision and Image Understanding, vol. 114, no. 1, pp. 21-32, January 2010.

[6] M. Poostchi, K. Silamut, R. J. Maude, S. Jaeger, and G. Thoma, "Image analysis and machine learning for detecting malaria," Translational research: the journal of laboratory and clinical medicine, Apr-2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5840030/. [Accessed: 21-Oct-2019].

[7] D. G. Lowe. "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, 2004.

[8] D. Das, M. Ghosh, M. Pal, A. Maiti and C. Chakraborty, "Machine learning approach for automated screening of malaria parasite using light microscopic images", Micron, vol. 45, pp. 97-106, 2013. Available: 10.1016/j.micron.2012.11.002 [Accessed 27 Nov. 2019].