Malaria Infected Cell Detection

Anubhav Shrimal (MT18033), Vrutti Patel (MT18020)
Advisor: Dr. Richa Singh

Statistical Machine Learning (CSE 542)

Abstract

In this project our aim is to identify whether a cell is malaria infected or not. We show an in breadth & depth analysis of various features like HOG, LBP, SIFT, SURF, pixel values with feature reduction techniques PCA, LDA along with normalization techniques such as z-score and min-max over different classifiers such as Naive Bayes, SVM, XGBoost, Bagging, AdaBoost, K-Nearest Neighbors, Random Forests and compare their performance by tuning different hyper-parameters. We evaluate the performance of these classifiers on metrics such as Accuracy, Precision, Recall, F1 score and ROC.

Use Case of the Problem

- Hundreds of thousands of people die every year due to malaria majorly in the underdeveloped or developing countries due to delayed diagnostics and unavailability of specialized doctors in this field. The goal of our project is to automate the detection of malaria parasite in a given blood sample image accurately.
- It is a challenge in Computer Vision & Machine Learning to handle sensitive cases like detecting cancerous cell and classifying whether a person is suffering from a disease or not. This project is a good example of solving such issues and can be extended to other use cases or domains as well.

Literature Review

In-depth explanation about malaria and how it is detected and diagnosed in real life is explained in paper [1] so that machine learning technique which automates the detection of malaria is close to the real world practise and gives better results. It is a survey paper. In this paper [2] the authors give an in depth review of the how malaria detection can be done using machine learning and clever image pre-processing. In paper [3] a technique to morph the cell images is described to have better boundary detection and it is then compared with different techniques such as Naive bayes and neural networks. The authors in [4] show the effects of image transformations such as segmentation to get better classification results. The proposed approach in paper [5] includes a preprocessing step to correct luminance differences, segmentation technique using the normalized RGB color space to classify pixels as erythrocyte or background followed by an Inclusion Tree representation that structures the pixel information into objects.

Dataset Description

- The dataset consists of 27,558 cell images; 13,780 images of infected and uninfected cells each and is taken from the official NIH Website.
- Link: https://ceb.nlm.nih.gov/repositories/malaria-datasets/
- Fig 1 shows visualization of the dataset used.

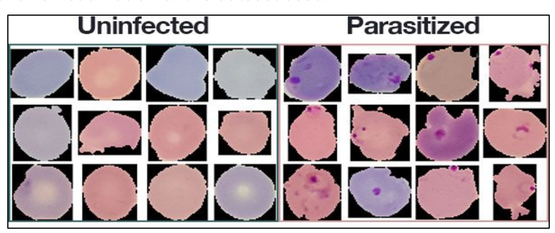

Fig 1. Dataset of uninfected and Malaria infected cells. (Source: Google Images)

Proposed Algorithm

- Different combinations of feature sets were used, some of which are shown in Table 1 & 2 (**Ugly Duckling Theorem**) many other combinations were tried.
- Evaluated with different classifiers, model parameters were varied using **Grid Search** to find the best parameters (**No Free Lunch Theorem**).
- In PCA, number of components were preserved using **Elbow method** over variance of PCA projected data (Fig. 4).

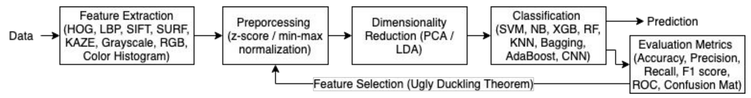

Fig 2. Training Pipeline

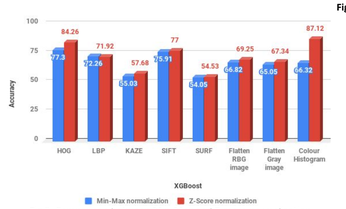

Fig 8. Comparison between Min-Max and Z-score normalization

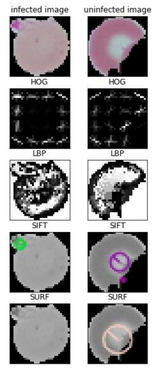

Fig 3. Feature Visualization

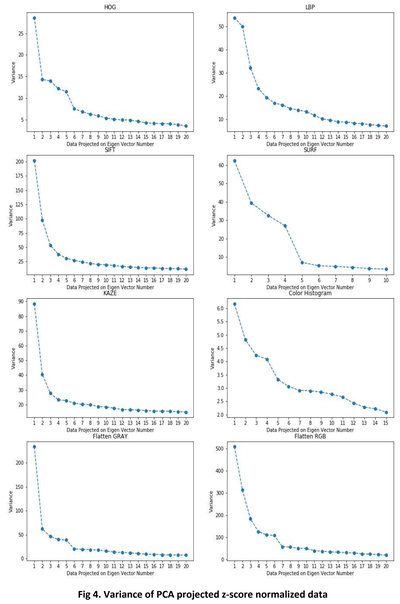

Fig 4. Variance of PCA projected z-score normalized data

Evaluation Metrics

- Receiver Operating Characteristic (ROC), Accuracy, Precision, Recall and F1-score.

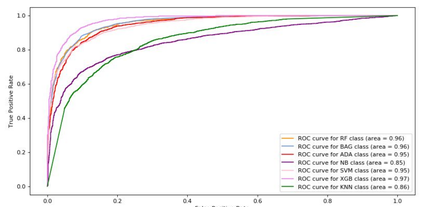

Fig 5. ROC of PCA reduced data

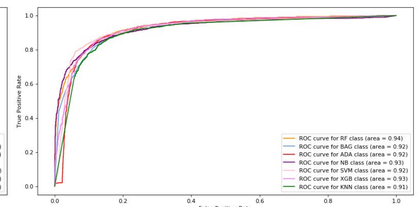

Fig 6. ROC of LDA reduced data

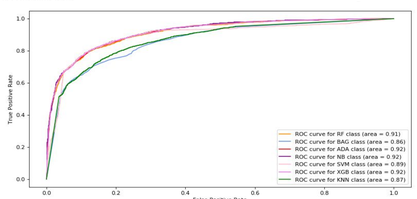

Fig 7. ROC of LDA on PCA reduced data

Results

Features / Classifiers		RF	BAG	ADA	NB	SVM	XGB	KNN
PCA (45 components): HOG(10), LBP(10), Color histogram(10), SIFT(5) and flatten RGB(10)	Accuracy	88.57	89.13	87.48	66.68	87.02	***91.49***	77.64
	Recall	88.37	89.59	86.38	42.24	85.16	***90.59***	72.27
	Precision	88.77	88.83	88.38	82.94	88.52	***92.3***	81.07
	F1 score	88.57	89.21	87.37	55.97	86.81	***91.44***	76.42
LDA (4 components): HOG, LBP, Color histogram, and SIFT	Accuracy	86.48	85.46	86.5	85.14	***86.82***	86.07	85.07
	Recall	86.06	84.89	85.34	81.86	***87.78***	85.16	85.21
	Precision	86.85	85.94	87.44	***87.69***	86.19	86.81	85.05
	F1 score	86.46	85.41	86.38	84.67	***86.98***	85.98	85.13
LDA on PCA combined features (1 component)	Accuracy	83.6	76.91	83.8	82.17	83.65	***83.83***	79.34
	Recall	86.06	76.66	***86.7***	75.75	85.02	86.11	79.42
	Precision	82.09	77.15	82.02	***87.01***	82.81	82.42	79.38
	F1 score	84.03	76.9	***84.3***	80.99	83.9	84.22	79.4

Table 1.Comparing various classifiers with different feature sets over Accuracy/Recall/Precision/F1 score

Good Features			Bad Features		
Feature	No. of Features	Accuracy RF	Feature	No. of Features	Accuracy RF
HOG	324	***86.43***	KAZE	2048	60.52
PCA HOG	10	69.97	PCA KAZE	10	58.48
LDA HOG	1	***83.44***	PCA SURF	5	54.78
LBP	1024	71.49	PCA Gray	6	67.68
PCA LBP	10	72.19	LDA Gray	1	65.89
LDA LBP	1	68.11			
PCA SIFT	5	76.03			
Color Hist	512	***94.73***			
PCA Color Hist	10	75.35			

Table 2. Good and Bad features on the basis of Accuracy on Random Forest classifier

Interpretation of Results

- Z-score normalization gave better accuracy than min-max normalization (Fig. 8).
- Features were said to be bad because of close to random accuracy i.e. no differentiating capability.
- Naive Bayes though gives good precision, performs poorly on infected class (recall).
- **XGBoost on PCA projected feature set (HOG, LBP, Color Hist, SIFT & RGB) gave the best metric scores** because boosting methods learn for misclassified data as well and XGB parameters (regularization, gradient descent) help learn better.
- AUC for ROCs of uninfected class show that the trained models are able to differentiate well.
- Table 2. shows the bad features which are close to random in classification (KAZE).

Conclusion

Compared and contrasted over different classifiers and feature extraction, reduction techniques. We found that XGBoost on PCA projected feature set gave the best results because of boosting methods.

References

1. Poostchi M, Silamut K, Maude R, Jaeger S, Thoma G (2018) “Image analysis and machine learning for detecting malaria” Transl Res 194
2. Jan Z, Khan A, Sajjad M, Muhammad K, Rho S, Mehmood I (2017) “A review on automated diagnosis of malaria parasite in microscopic blood smears images” Multimedia Tools Appl 77:1–26
3. Das DK, Maiti AK, Chakraborty C (2015) “Automated system for characterization and classification of malaria-infected stages using light microscopic images of thin blood smears” J Microsc 257(3):238–252
4. Suryawanshi MS, Dixit V (2013) “Improved technique for detection of malaria parasites within the blood cell images” Int J Sci Eng Res 4:373–375
5. Gloria Daz, Fabio A. Gonzlez, and Eduardo Romero. 2009. “A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images” J. of Biomedical Informatics