Malaria Infected Cell Detection

1st Anubhav Shrimal
M.Tech CSE (MT18033)
IIIT Delhi
New Delhi, India
anubhav18033@iiitd.ac.in

2nd Vrutti Patel
M.Tech CSE (MT18020)
IIIT Delhi
New Delhi, India
vrutti18020@iiitd.ac.in

***Abstract*—Our aim is to build and compare various machine learning models to classify a given cell image as uninfected or infected by malaria parasite. The classification of images is done using various methodologies such as Bayesian classifier, PCA and LDA to reduce dimensionality, Ensemble learning techniques: Bagging and Boosting using Decision Tree as weak classifier, CNN and Pre-trained CNN and also compare their performance on different evaluation metric to give detailed analysis of which method performs better and why.**

***Index Terms*—Parasitized images, Uninfected images, Naive Bayes, PCA, LDA, Bagging, Boosting, Decision tree, Weak Classifier, CNN**

I. Literature Review

In-depth explanation about malaria and how it is detected and diagnosed in real life is explained in paper [3] 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 as the authors have summarized different ways to obtain cell images with or without malaria parasites like light microscopy, binocular microscopy and so on. Various preprocessing techniques to enhance blood smear samples such as noise reduction, improving contrast and ways to do this. Different segmentation techniques like otsu thresholding, clustering, watershed, hough transform etc. Types and ways of feature computation like on the basis of color, texture or morphology and finally numerous classification techniques for supervised as well as unsupervised learning to detect malaria infected cells from images. Authors have cited various papers to support all the above methods.

In paper [1] the authors presents the evaluation of a color segmentation technique, based on standard supervised classification algorithms. They have implemented four different algorithms - K Nearest Neighbors, Naive Bayes, Support Vector Machine (SVM) and Multi Layer Perceptron (MLP) with different color spaces - RGB, normalized RGB, HSV and YCbCr. They have compared the results on the basis of F-score and inferred that all the algorithms are able to identify the uninfected cell images with a higher score as compared to parasitized cell images. The best performance for both classes is given by KNN classifier with normalized RGB color space and SVM classifier with YCrCb color space.

Paper [2] used the dataset of blood samples stained with giemsa so the authors were able to extract color histogram, granulometry, gradient and flat texture features after preprocessing data and used this features instead of just pixel value from the images which gives only color information as done in [1]. This features were given as input to SVM, nearest mean (NM), KNN, 1-NN, and Fishers linear discriminant classifiers. The results were compared by accuracy and precision.

As a preprocessing step staining variation has been removed from peripheral blood smear images, impulse noise has been reduced, Erythrocytes are segmented using marker controlled watershed algorithm and textural and morphological features have been extracted and trained on bayesian classifier in paper [4]. 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. Then the classification process which identifies infected erythrocytes using a trained bank of classifiers. Average sensitivity and specificity are reported to evaluate the results.

II. Dataset Used

The dataset used is Malaria-dataset downloaded from https://ceb.nlm.nih.gov/repositories/malaria-datasets/. The dataset contains 27,558 images in total and has 2 classes, Paracitized and Uninfected (13,780 images each). Each image belongs to either one of the classes and is an image of a blood smears which either does or does not contain the malarial parasite. Images are of varying sizes and are resized to same size for training and testing purposes. The dataset is split into 90:10 Training-Testing ratio. The training dataset is further split into 80:20 Training-Validation set.

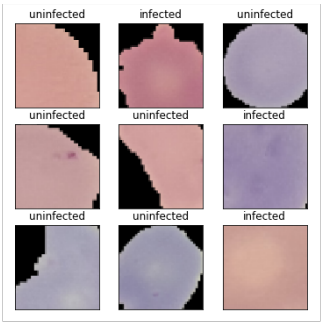


Fig. 1: Dataset Visualization

III. Tasks Completed

We have completed the tasks that were mentioned in our midterm milestone in the proposal along with PCA and LDA dimensionality reduction techniques. Following are the tasks that have been completed so far:

A. Naive Bayes

It is a simple classifier and is based on the hypothesis that features are conditionally independent of each other. Images are read in RBG color space, pixel value indicating color information are the features for classifier. Dataset is split randomly in training and testing data keeping the same size for all images. 5-fold cross validation is performed to get the best bayesian model. Testing data is classified by the bayseian model and accuracy, precision, recall F1-score, confusion matrix and ROC are reported for all the approaches followed.

B. Dimensionality Reduction over Naive Bayes

- PCA: This technique is applied to get top 40 features having maximum variance. A better separation of data is obtained by projecting data in the direction of top features.
- LDA: As PCA was not able to capture class information, dimensionality of features is reduced through LDA which gives just 1 feature as the problem is a 2 class problem but it separates data on the basis of within and between class distance giving better results.

C. Convolutional Neural Network implemented from Scratch

The input image size used is 64x64. We have applied RandomCrop and horizontal flips as image transformations. We created a CNN architecture from scratch with 5 Convolution layers, 3 Max-Pooling layers and Fully Connected (FC) layers of dimensions 500, 100 and 2. Activation function used is ReLU. Dropout of 0.2 is added in between FC layers. The optimizer used is Adam with a learning rate of 0.001 with Cross Entropy Loss function. The model was trained for 20 epochs.

IV. Tasks Remaining

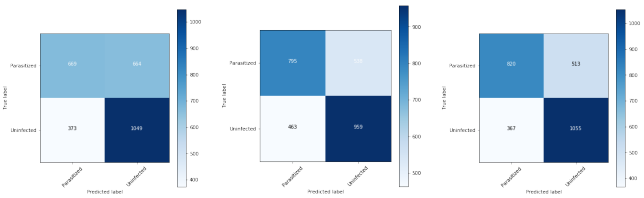
The following tasks will be covered in the final milestone:

- Bagging
- Boosting
- Pre-Trained deeper CNN architecture such as ResNet34 and ResNet50
- Model Comparisons and analysis

V. Results

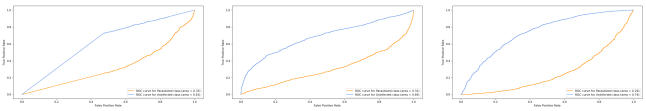
TABLE I: Accuracy on 5 fold cross validation for Naive Bayes

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy	62.97	64.12	63.07	64.46	62.96



(a) Complete data (b) PCA reduced (c) LDA reduced

Fig. 2: Confusion Matrices for Naive Bayes



(a) Complete data (b) PCA reduced (c) LDA reduced

Fig. 3: ROC for Naive Bayes

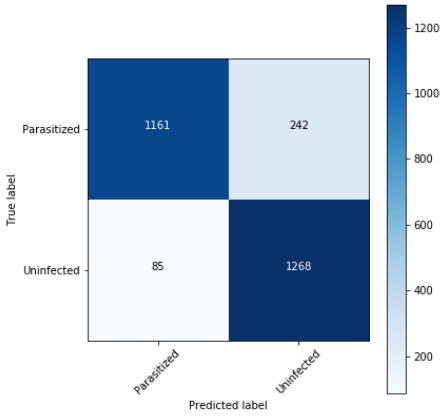


Fig. 4: Confusion Matrix for CNN scratch implementation

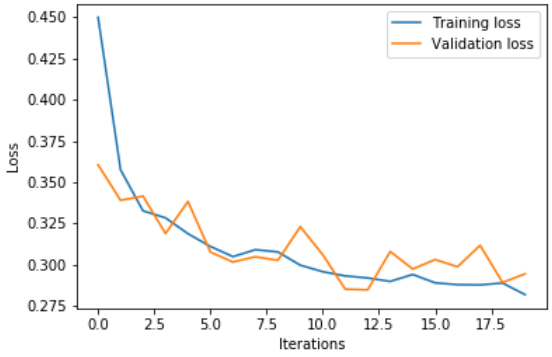


Fig. 5: Training and Validation Loss for CNN

TABLE II: Model Comparison using Accuracy, Precision, Recall and F1-Score (Positive Class: Parasitized)

	Accuracy	**Precision**	**Recall**	**F1-Score**
Naive Bayes	62.36	64.2	50.18	56.33
Naive Bayes on PCA	63.66	63.19	59.63	61.36
Naive Bayes on LDA	68.05	69.08	61.51	65.07
CNN from scratch	88	93.2	82.8	87.7

VI. Inferences

Naive Bayes classifier performs well for uninfected cell but not for parasitized cell as it can be seen from the confusion matrix the number of TN are much more as compared to number of TP. So uninfected cell has a major role in accuracy. This performance improves to some extent on applying PCA and gets better with LDA.

CNN implemented from scratch is able to perform much better in comparison to Naive Bayes. This is because CNN is able to capture spatial information and patterns and is able to keep only the important information for classification. Confusion Matrix shows that both the classes are classified reasonably well. The training and validation loss tells that the model is generalized well over the data and is not under or over fitting.

References

[1] Daz, Gloria & Gonzlez, Fabio & Romero, Eduardo. (2007). Infected Cell Identification in Thin Blood Images Based on Color Pixel Classification: Comparison and Analysis. 4756.812-821.10.1007/978-3-540-76725-1_84.

[2] Malihi, L., Ansari-Asl, K. & Behbahani, A. (2013). Malaria parasite detection in giemsa-stained blood cell images. 2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP), 360-365.

[3] Poostchi, Mahdieh & Silamut, Kamolrat & Maude, Richard & Jaeger, Stefan & Thoma, George. (2018). Image analysis and machine learning for detecting malaria. Translational Research. 194.10.1016/j.trsl.2017.12.004.

[4] Das, D.K., Ghosh, M., Pal, M., Maiti, A.K., & Chakraborty, C. (2013). Machine learning approach for automated screening of malaria parasite using light microscopic images. Micron, 45, 97-106.

[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 42, 2 (April 2009), 296-307. DOI: https://doi.org/10.1016/j.jbi.2008.11.005