THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

BHI-2016 International Conference on Biomedical and Health Informatics

A COMPARISON OF FEATURE SELECTION METHODS FOR MACHINE LEARNING BASED AUTOMATIC MALARIAL CELL RECOGNITION

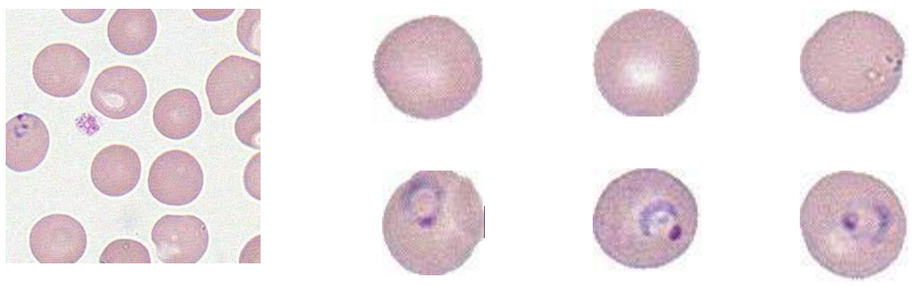
Vishnu Muralidharan, Yuhang Dong, and W. David Pan
Dept. of Electrical and Computer Engineering, University of Alabama in Huntsville, Huntsville, AL 35899, USA

Introduction

- Wholeslide Imaging: ensure accurate diagnosis of malaria using blood smears.
- Machine learning: automatic diagnosis is desirable in resource-scarce areas.
- **Objective: study SVM performance under different feature selection techniques**.

Materials and Methods

- Wholeslide images: acquired at highest resolution in DeepZoom pyramid.
- Segmented to obtain samples



http://peir-vm.path.uab.edu/debug.php?slide=IPLab11Malaria

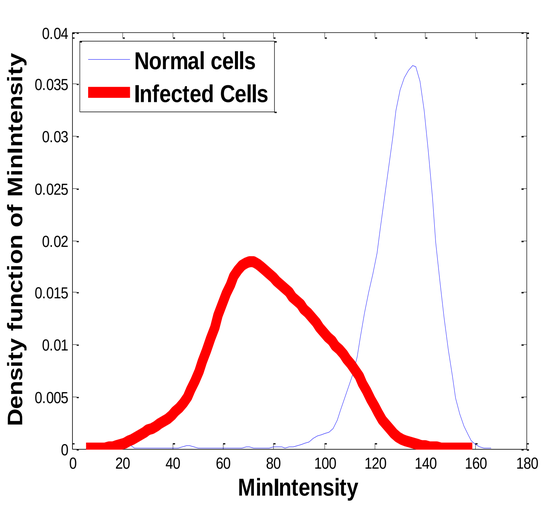
- 76 features belonging to five categories are extracted.
- Feature selection using various methods is done to select features.
- SVM classification is performed on test set.

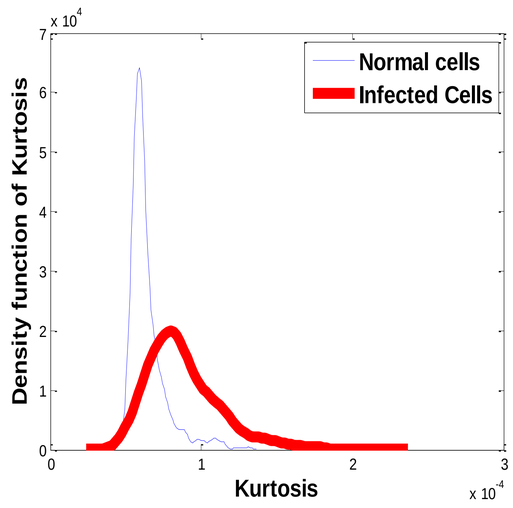
Feature Selection

- **Five different filter methods are used**.
- Study discriminating power between classes.
- Five methods:
  - Kullback-Leibler (KL) Distance

$$D_{KL}(P||Q) = \sum_x P(x) log \frac{P(x)}{Q(x)}$$

  - Bhattacharyya Distance
  - Wilcoxon Ranksum Test
  - ROC
  - F-Score
- KS Density plots show discrimination



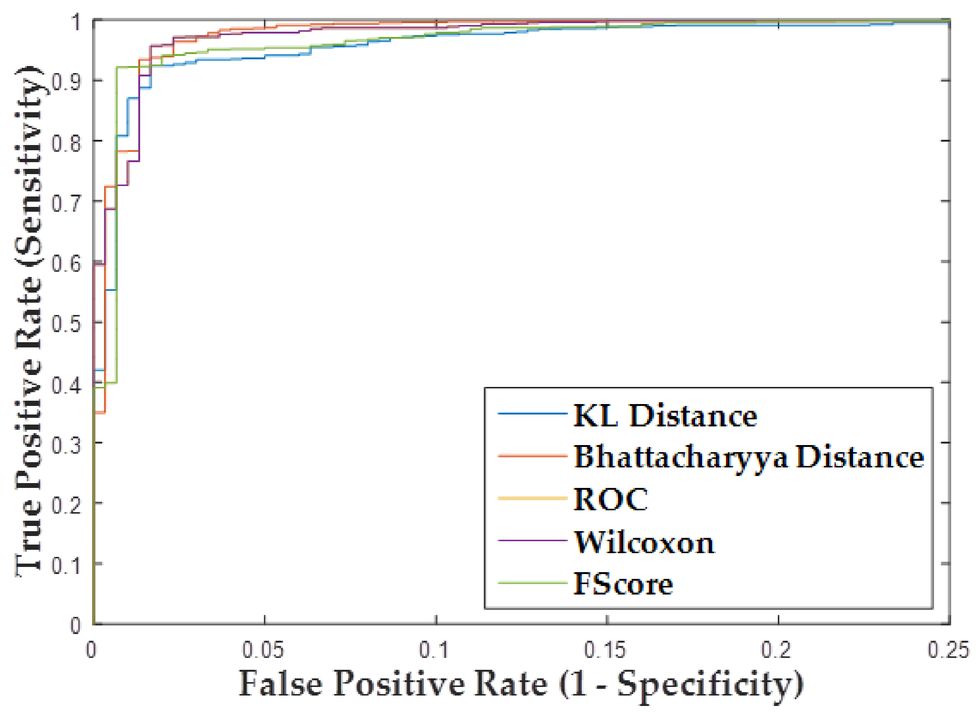


Results

Feature selection method	Classification Parameters				
	Accuracy	Sensitivity	Specificity	PPV	NPV
KL Dist.	95.5	96.43	94.61	94.50	96.50
Bhatt. Dist.	94.75	96.37	93.24	93.00	96.50
ROC	91.75	94.18	89.57	89.00	94.50
Wil. Ranksum Test	89.25	94.86	84.89	83.00	95.5
F-Score	83.25	94.63	76.49	70.50	96.00

Feature Selection method	SVM Parameters	
	Cost	Scale
KL. Dist.	8	32
Bhatt. Dist.	2	2
ROC	8	128
Wil. Ranksum Test	2	8
F-Score	2	0.5

Feature Selection	Confusion Matrices		
	Labels	Normal	Infected
KL. Dist.	Normal	189	11
	Infected	7	193
Bhatt. Dist.	Normal	186	14
	Infected	7	193
ROC	Normal	178	22
	Infected	11	189
Wil. Ranksum Test	Normal	166	34
	Infected	9	191
F-Score	Normal	141	59
	Infected	8	192



Conclusions

- **KL distance is the best measur**e.
- Models entropy and asymmetry
- Yields a classifier that is flexible and does not overfit.
- ROC curve is favorable for classification.