Training Set Optimization and Classifier Performance in a Top-Down Diabetic Retinopathy Screening System

J Wigdahl*[1, 2], C Agurto[1, 2], V Murray[1], S Barriga[1], P Soliz[1]

[1] VisionQuest Biomedical, Albuquerque, NM-87106
[2] Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM-87131

ABSTRACT

Diabetic retinopathy (DR) affects more than 4.4 million Americans age 40 and over. Automatic screening for DR has shown to be an efficient and cost-effective way to lower the burden on the healthcare system, by triaging diabetic patients and ensuring timely care for those presenting with DR. Several supervised algorithms have been developed to detect pathologies related to DR, but little work has been done in determining the size of the training set that optimizes an algorithm's performance. In this paper we analyze the effect of the training sample size on the performance of a top-down DR screening algorithm for different types of statistical classifiers. Results are based on partial least squares (PLS), support vector machines (SVM), k-nearest neighbor (kNN), and Naïve Bayes classifiers. Our dataset consisted of digital retinal images collected from a total of 745 cases (595 controls, 150 with DR). We varied the number of normal controls in the training set, while keeping the number of DR samples constant, and repeated the procedure 10 times using randomized training sets to avoid bias. Results show increasing performance in terms of area under the ROC curve (AUC) when the number of DR subjects in the training set increased, with similar trends for each of the classifiers. Of these, PLS and k-NN had the highest average AUC. Lower standard deviation and a flattening of the AUC curve gives evidence that there is a limit to the learning ability of the classifiers and an optimal number of cases to train on.

Keywords: Diabetic Retinopathy, Automatic Screening, Classifier Performance, Optimization

1. INTRODUCTION

Diabetes is on the rise, with an estimated 285 million cases worldwide; as of 2010, 26.8 million of these cases are in the US [1]. Diabetic Retinopathy (DR) affects more than 4.4 million Americans age 40 and over, with annual direct medical costs of approximately $500 million [2]. More than 60% of type 2 diabetics will DR during their lifetime, unless the patient is diagnosed early and commits to controlling their blood glucose levels [3].

The American Academy of Ophthalmology recommends that diabetics receive a comprehensive eye exam at least once a year. People with early stages of DR are asked to control their blood sugar, blood pressure, and blood cholesterol. Later stages of DR can be treated surgically. Proliferative retinopathy (PDR) can be treated with laser surgery. Risk of blindness can be reduced by 95% with timely treatment [4]. Macular edema can be treated using focal laser treatment. New drugs, such as Lucentis, have also been found to help these patients.

The first stages of DR are asymptomatic, leaving screening as the most viable option for detection amongst the diabetic population. Studies have shown that a highly sensitive and specific automatic DR screening system would lower costs and improve timely treatment for those with DR [5]. By screening out normal patients, ophthalmologists could focus their care on patients with retinopathy. With the number of diabetics expected to increase to 438 million by 2030 [1], the utility of an automated screening system becomes apparent.

Screening systems can be divided into two categories: bottom-up and top-down. Bottom-up approaches need to segment DR lesions in order to determine DR status while top-down approaches look at a global feature set with classification techniques to separate normal from abnormal. We have developed a top-down DR screening system [6], and in this work we present analysis of the performance of different classifiers and the effects of varying the size of the training set and the type of feature selection. Optimizing these variables will ensure the best performance of the DR screening system while using the fewest number of cases.

*jwigdahl@visionquest-bio.com; phone 1 505 508-1994; fax 1 505 508-5308; visionquest-bio.com

This paper is organized as follows. In Section 2, we explain the image processing techniques used to create the top-down system and the different types of classifiers used. Section 3 presents the results and a discussion of our findings. Conclusions are given in Section 4.

2. METHODS

2.1. DR Risk Analysis System

A schematic of our DR risk analyzer system is shown in Figure 1. The green channel of each retinal image is processed using a multi-scale AM-FM decomposition [7] based on:

$$G(k_1, k_2) \cong \sum_{n=1}^{M} a_n(k_1, k_2) cos\varphi_n(k_1, k_2)$$

where G represents… n = 1,2,…, M denote different frequency scales, a_n denotes the instantaneous amplitude functions (IA), and φ_n denotes the instantaneous phase functions. Each scale is defined in terms of separable bandpass filters with similar frequency magnitude ranges. Each AM-FM component has an associated instantaneous frequency (IF) defined as $\nabla_\varphi = (\varphi_x, \varphi_y)$. For our system, we use three AM-FM estimates: (i) IA, (ii) IF magnitude, (iii) IF angle, and a 5-scale filterbank consisting of frequency scales defined as: (i) high (*H*), (ii) medium (*M*), (iii) low (*L*), (iv) very low (*V*), (v) ultra-low (*U*), and (vi) lowpass filter (*F*). AM-FM estimates are computed for 13 combinations of these scales (CoS). Thus a single image is divided into 39 new images (3 AM-FM estimates and 13 CoS).

Each of the 39 AM-FM representations is divided into 202 regions of interest (ROI). A 32-bin histogram of the values in each ROI is determined to approximate its probability density function. For each CoS, the 202 ROI histograms are grouped into 30 clusters using the k-means unsupervised learning method. This transforms the features from ROI histograms into cluster counts per CoS. The number of features extracted per green channel image is 1170 (39CoS x 30clusters). Generally, these features are highly correlated and the high dimensionality requires a feature selection step. In these experiments we look at classification performance using the whole feature set, principal component analysis (PCA), and partial least squares (PLS) for feature selection. These features are then fed into a classifier. The classifiers chosen for these experiments are PLS, Support vector machines (SVM), k-Nearest Neighbors (k-NN), and Naïve Bayes. These will all be looked at in further detail in the following sections. For an in-depth look at the entire system, we refer to Murray et al [6].

2.2. Data Set

This system was trained using images collected as part of our DR screening efforts with our collaborating clinics in New Mexico and San Antonio, TX. The database consists of images from 745 patients. For each patient, one macula-centered image and one optic-disc-centered image was collected per eye using a Canon CR-1 Mark II non-mydriatic retinal camera with a 45 degree field of view. Only the macula-centered images (N=1490) were used in this paper. Grading of the images was performed by two separate optometrists. Disagreements between the two graders were adjudicated by a certified retinal grader. Images were graded on a scale of 0-4, with 0 being no DR findings, 1: mild non-proliferative DR (NPDR), 2: moderate NPDR, 3: severe NPDR, and 4: PDR. One hundred and fifty of the cases were graded as positive for DR (1-4) and 595 cases were graded negative for DR, but may present with other, non-DR pathologies.

2.3. Experimental Design

Experiments were run using a fixed number of abnormal cases (50, 100, 150) and a varying number of normal cases with increments of 50 (100, 150, 200, …, 595). For each experiment we separated the data into 70% for training and 30% for testing. Ten runs of each experiment were performed with the training and testing sets randomized for each run. The same training and testing sets were used across all classifiers to create comparable results. Each experiment was run using the entire feature set, a reduced feature set using PLS for feature selection, and a reduced feature set using PCA for feature selection.

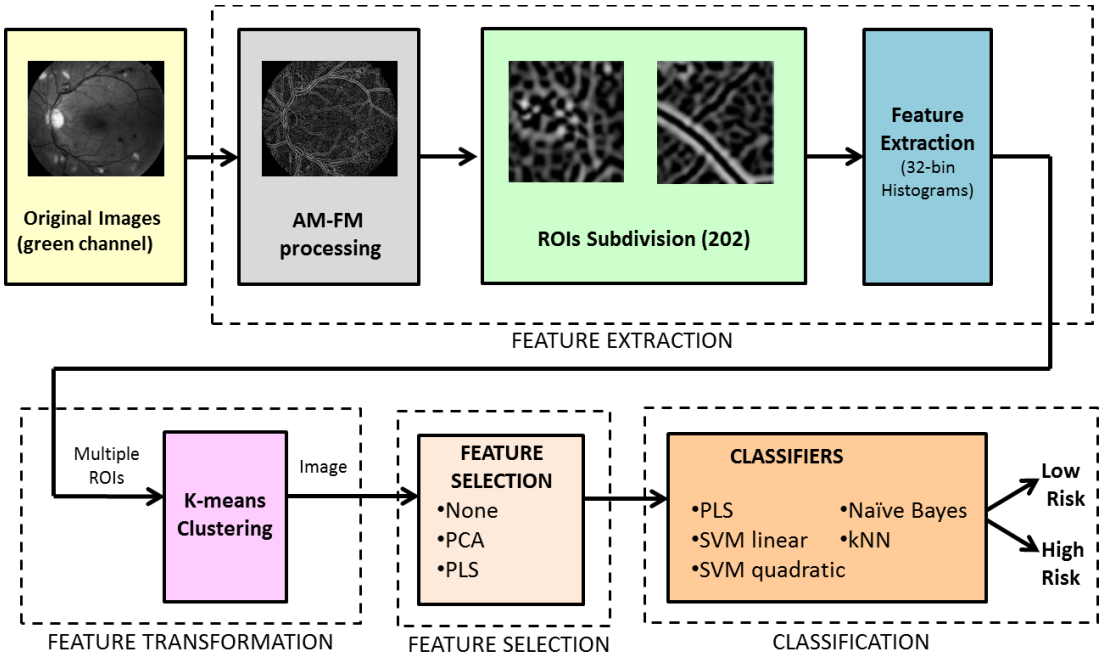

Figure 1. DR screening system schematic. The green channel image is processed using AM-FM and then broken down into 202 ROIs. The features that are extracted are the 32-bin histograms of each ROI. These features are clustered and the cluster counts are used for classification.

2.4. Classifiers and Feature Selection Methods

a) Partial Least Squares (PLS)

PLS is a regression technique [8] that allows classification using the weights associated with each of the components on the regression matrix.

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

PLS generates regression weights $\boldsymbol{\beta}$ that allows the calculation of the dependent matrix $\boldsymbol{y}$. Setting a threshold for $\boldsymbol{y}$ will separate samples into two classes. $\boldsymbol{X}$ is the matrix of extracted AM-FM features and $\boldsymbol{\varepsilon}$ is a vector of residuals. To estimate $\boldsymbol{\beta}$, the least squares solution is given by the normal equations $\boldsymbol{\beta} = (X^T X)^{-1}(X^T y)$ [9]. The AM-FM features we have extracted are highly correlated, making $X^T X$ nearly singular and a unique solution to the normal equation nonexistent. The feature selection step reduces X to a smaller, uncorrelated set of observations.

b) Support Vector Machines (SVM)

SVM are a set of supervised learning methods [10] that separate classes based on the construction of a hyperplane that maximizes the distance between the nearest training data point. Mathematically this can be shown as

$$u = \mathbf{w} \cdot \mathbf{x} - \mathrm{b}$$

where **w** is the normal vector to the hyperplane, **x** is the input vector, and *b* is and offset parameter. Hyperplanes were originally used for linear classification, but the idea was extended to non-linear classification problems using kernels. In this paper we use the linear and quadratic kernels, which can all be derived from

$$k(x_i, x_j) = (x_i \cdot x_j)^n$$

where k is the kernel function, x_i and x_j are data points, and n is the degree of the polynomial kernel function. The Bioinformatics toolbox from Matlab was used for the SVM experiments.

c) Naïve Bayes

This classifier is based on a priori probabilities and a strong assumption that the features being used are independent [11]. It is based on Bayes theorem which states

$$P(w_j|\mathbf{x}) = \frac{p(\mathbf{x}|w_j)P(w_j)}{p(\mathbf{x})}$$

where **x** is the feature vector and w_j is the state of nature. This equation gives us the independent feature model. To create a classifier, this model, along with a decision rule, must be implemented. One such decision rule is to pick the most probable hypothesis, known as maximum a posteriori. Since the classifier is based on prior probabilities, training on proportions seen in the real world becomes important, which may not be practical for diseases that have a low prevalence rate.

d) k-Nearest Neighbor

k-Nearest Neighbor (k-NN) classifies a feature vector **x** by labeling it based on its proximity to the k nearest samples [7]. The neighbors are selected based on a predetermined distance metric. Examples of the distance metric are Euclidean distance: $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$ or Hamming Distance: $\sum_{i=1}^{k}|x_i - y_i|$. k-NN is also called memory-based classification, because the training examples need to be on hand at the time of classification and any induction is performed at run-time.

This classification method can be sensitive to the value of k chosen. A small k will be sensitive to noise in the training, while a large k (with respect to the number of samples *n*) could incorporate the votes of samples that are far away from **x**. For these experiments, we set a value of k=19, which was optimized for our experiments using 150 abnormal cases.

e) Principal Component Analysis

Principal Component Analysis (PCA) computes an orthogonal transformation of a set of observations that may be correlated into a set of linearly uncorrelated variables. These variables are calculated so that the first variable explains the most variation in the data and so on [13]. The number of components can be chosen using a threshold of the percentage of variation that should be explained. For our experiments, this percentage was set at 95% of the variation.

3. RESULTS AND DISCUSSION

Experimental results for each of the classifiers are shown in Figure 2. Each experiment consisted of 10 runs that produced an average AUC and standard deviation. Before classification, the AM-FM CoS features were normalized to zero mean and unit standard deviation using only the training information in each run.

We used two different processes to reduce the AM-FM features. Supervised feature reduction was done using PLS, and unsupervised feature selection was done using PCA. PCA was set to keep 95% of the variation, which reduced the feature set by 80%. Using PLS for feature selection reduced the feature set by 75%. Figure 2 shows the average AUC for each experiment using 50, 100, and 150 abnormal patients and either PCA or PLS for feature selection. Tables 1 and 2 give a broad summary of all the experiments run with PLS and PCA, respectively.

It can be seen from the graphs that there is similar performance among the classifiers, especially when PLS is used for feature selection. The graphs show erratic behavior when 50 abnormal cases are used, suggesting the need for more patients. For the top-performing classifiers (k-NN, PLS), the graphs for 100 and 150 abnormal patients look more like a standard learning curve with a trend for increasing average AUC, as well a decreasing trend for the standard deviation of the AUC as more cases are added to the training set. There is significant performance increase in classifier AUC from 100 normal cases to 250 normal cases. From that point on, the curves start to flatten out and the standard deviation decreases. This indicates performance stabilization can be achieved using a fraction of the whole data set. This is in agreement with what others researchers have shown [14].

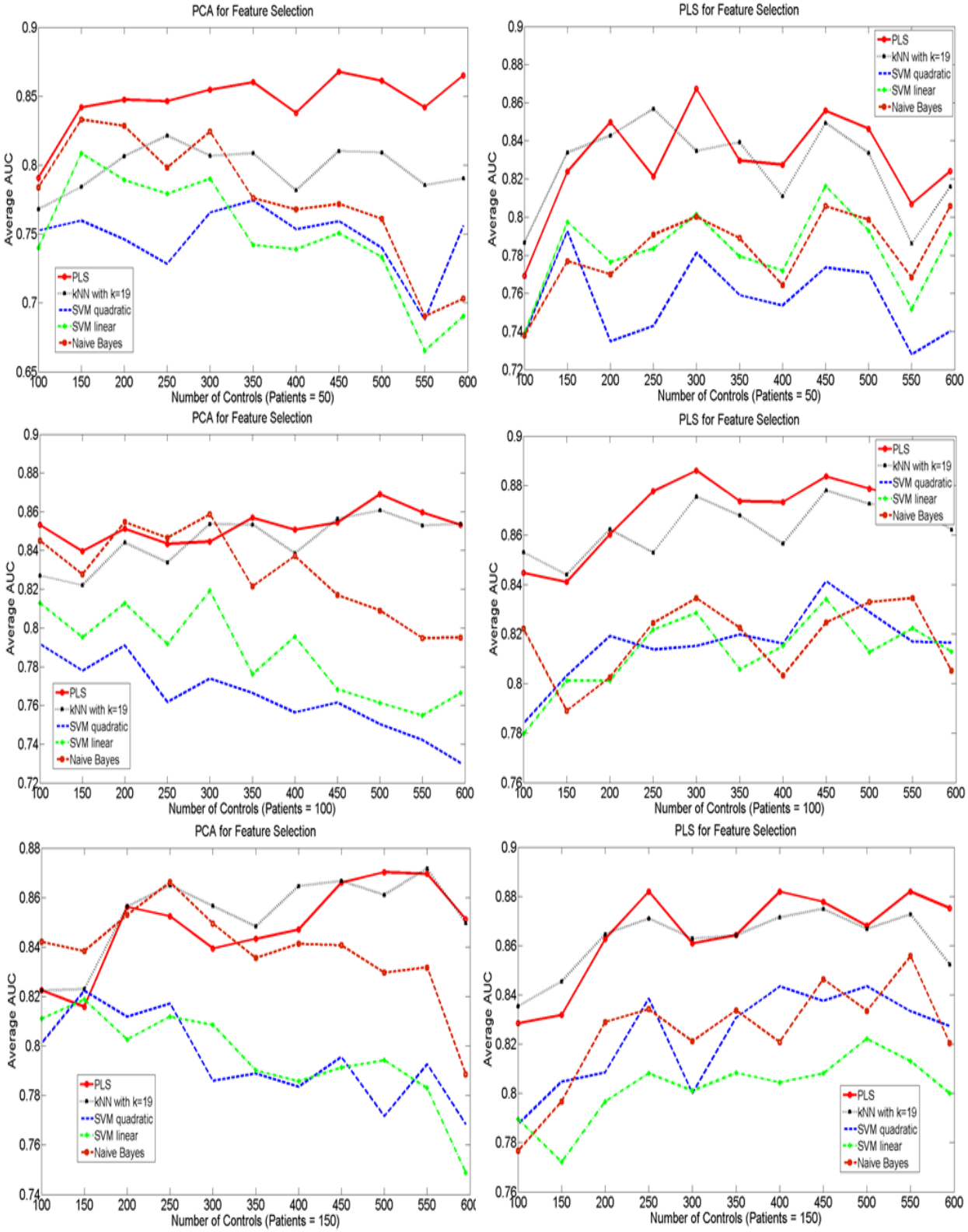

Figure 2. Average AUC for 10 runs of each experiment using 50, 100, and 150 abnormal cases. The left column used PCA for feature selection and the right used PLS.

Table 1. Summary of results using PLS for feature selection. Average AUC is listed for all the points in a line of Figure 2. Minimum and Maximum AUC as well as average standard deviation is also listed.

Classifier	150 Abnormal Cases - PLS				100 Abnormal Cases - PLS				50 Abnormal Cases - PLS			
	Avg. AUC	Min	Max	Avg. St. Dev.	Avg. AUC	Min	Max	Avg. St. Dev.	Avg. AUC	Min	Max	Avg. St. Dev.
PLS	0.87	0.83	0.88	0.02	0.87	0.84	0.89	0.03	0.83	0.77	0.87	0.04
k-NN	0.86	0.84	0.87	0.02	0.86	0.84	0.88	0.03	0.83	0.79	0.86	0.04
SVM quadratic	0.82	0.79	0.84	0.02	0.82	0.78	0.84	0.03	0.76	0.73	0.79	0.05
SVM linear	0.80	0.77	0.82	0.03	0.81	0.78	0.83	0.03	0.78	0.74	0.82	0.05
Naïve Bayes	0.82	0.78	0.86	0.03	0.82	0.79	0.83	0.03	0.78	0.74	0.81	0.06

Table 2. Summary of results using PCA for feature selection. Average AUC is listed for all the points in a line of Figure 2. Minimum and Maximum AUC as well as average standard deviation is also listed.

Classifier	150 Abnormal Cases - PCA				100 Abnormal Cases - PCA				50 Abnormal Cases - PCA			
	Avg. AUC	Min	Max	Avg. St. Dev.	Avg. AUC	Min	Max	Avg. St. Dev.	Avg. AUC	Min	Max	Avg. St. Dev.
PLS	0.85	0.82	0.87	0.03	0.85	0.84	0.87	0.03	0.85	0.79	0.87	0.04
k-NN	0.85	0.82	0.87	0.03	0.85	0.82	0.86	0.03	0.80	0.77	0.82	0.04
SVM quadratic	0.79	0.77	0.82	0.03	0.76	0.73	0.79	0.04	0.75	0.69	0.77	0.05
SVM linear	0.80	0.75	0.82	0.03	0.79	0.75	0.82	0.04	0.75	0.67	0.81	0.05
Naïve Bayes	0.84	0.79	0.87	0.02	0.83	0.79	0.86	0.03	0.78	0.69	0.83	0.04

Results show that as few as 400 total cases can give optimal performance, as long as 100 of those cases are from the abnormal class. Since we are using a top-down approach, there is no obvious association between the extracted features and the classification result. In fact, the AM-FM features that are extracted are highly correlated and thus redundant. This is one of the reasons why PLS works well with this kind of data, as it is able to find the features with the maximum variance and give them the highest weights, while diminishing the importance of features that add little to the classification. The k-NN classifier shows comparable results to PLS in all experiments, except with 50 abnormal cases and PCA. For the k-NN classifier, the number of neighbors was kept the same (k=19) for each experiment, but should be adjusted based on the size of the data set. This number was optimized for the experiments using 150 abnormal cases, but still performed better with smaller training set sizes. The SVM kernels did not perform as well as the other classifiers on the reduced feature sets; however, they performed better using PLS for feature selection. The best performance for SVM used the quadratic kernel on the entire feature set as feature selection is used when there is an issue with dimensionality, and SVM kernel-based techniques take care of this issue. This can be seen in Table 3, where SVM quadratic is a top performer when 100 or 150 cases are used. With this in mind, the best results for SVM using the entire feature set are only slightly below the performance of PLS and k-NN using feature selection. Naïve Bayes was the only classifier to perform better using PCA for feature selection. PCA can also be thought of as a compression technique in which case some information is lost (not just noise). It also had the largest AUC improvement from the full feature set to the reduced feature sets. This is attributed to the highly correlated features, a domain where Naïve Bayes performs poorly.

If we focus on the results using PLS, there is no statistical difference between using 100 and 150 abnormal cases ($p = .57$). However, those two scenarios are significantly different from the 50 abnormal case experiments ($p=.009$).

It is obvious that there should be diminished results using fewer samples, but what these results do tell us is that we will see little improvement using more than 100 abnormal cases. Bias perceived by using PLS for feature selection and classification did not materialize. While the best overall results are achieved using this combination, PLS shows its robustness in still being a top-performing classifier using PCA for feature selection or using the whole feature set, albeit at a lower AUC, than using the two-step PLS process.

Table 3. Summary of Results using the entire feature set. Average AUC is listed for all the points in a line of Figure 2. Minimum and Maximum AUC as well as average standard deviation is also listed.

Classifier	150 Abnormal Cases				100 Abnormal Cases				50 Abnormal Cases			
	Avg. AUC	Min	Max	Avg. St. Dev.	Avg. AUC	Min	Max	Avg. St. Dev.	Avg. AUC	Min	Max	Avg. St. Dev.
PLS	.85	.82	.87	.03	.85	.84	.87	.03	.85	.79	.87	.04
k-NN	.85	.82	.87	.03	.85	.82	.86	.03	.80	.76	.82	.04
SVM quadratic	.85	.81	.88	.02	.84	.81	.86	.03	.79	.76	.81	.05
SVM linear	.83	.80	.86	.02	.84	.82	.85	.03	.82	.77	.86	.04
Naïve Bayes	.74	.70	.77	.04	.72	.70	.75	.04	.71	.67	.75	.05

4. CONCLUSSIONS

This work presented the results of different feature selection methods and different classifiers in a top-down diabetic retinopathy screening system. These classifiers were trained using set amounts of abnormal cases (N=50,100,150) and a varying number of normal cases (N=100,150,200,…595). Our goal was to determine the best classification method for this problem and also find an optimal number of cases to train on while maintaining peak performance.

We have found PLS and k-NN to have better performance regardless of the number of abnormal cases used or feature selection method. SVM with a quadratic kernel shows comparable performance using 100/150 abnormal cases and the entire feature set. Naïve Bayes performed slightly worse than the other classifiers, but has better performance using PCA for feature selection. We found performance begins to stabilize around 400 total cases (with abnormal cases at 100/150), indicating a ceiling to this learning process and an optimum number of cases for training. This is important when the data collection process is time consuming and abnormal patients are difficult to collect.

Future work will focus on expanding the number of classifiers for training and in refining our database to include only normal cases that do not present with any other disease. This could lead to less variation in the normal class and better overall classification results.

ACKNOWLEDGEMENTS

This work was supported by NIH grants EY020015 and EY018280. The authors would also like to thank Sheila Nemeth for contributions to data collection and grading.

REFERENCES

1. IDF Diabetes Atlas, 4th edition International Diabetes Federation 2009. www.diabetesatlas.org/content/regional-data.

2. Rein, David B., Ping Zhang, Kathleen E. Wirth, Paul P. Lee, Thomas J. Hoerger, Nancy McCall, Ronald Klein, James M. Tielsch, Sandeep Vijan, and Jinan Saaddine. "The Economic Burden of Major Adult Visual Disorders in the United States". Archives of Ophthalmology. Vol. 124, No. 12, pp. 1754-60.

3. D S Fong, L Aiello, T W Gardner, G L King, G Blankenship, J D Cavallerano, F L Ferris III, R Klein, "Retinopathy in Diabetes," Diabetes Care January 2004 vol. 27 no. suppl 1 s84-s87

4. http://www.nei.nih.gov/health/diabetic/retinopathy.asp

5. Javitt JC, Aiello LP, Chiang Y, Ferris FL, Canner JK, Greenfield S. Preventive eye care in people with diabetes is cost-saving to the federal government: implications for health-care reform. *Diabetes Care* 1994; 17: 909-717.

6. Murray et al. "Real-Time Diabetic Retinopathy Patient Screening System Using Multiscale AM-FM Methods, *International Conference on Image Processing*, 2012

7. V. Murray, P. Rodriguez, and M. Pattichis, "Multi-scale AM-FM demodulation and image reconstruction methods with improved accuracy," *IEEE Transactionson Image Processing*, vol. 19, no. 5, pp. 1138 –1152, May 2010.

8. Matthew Barker and William Rayens, "Partial least squares for discrimination," *Journal of Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.

9. Carla Agurto, E. Simon Barriga, VictorMurray, Sheila Nemeth, Robert Crammer, Wendall Bauman, Gilberto Zamora, Marios S. Pattichis, and Peter Soliz, "Automatic detection of diabetic retinopathy and age-related macular degeneration in digital fundus images," *Investigative Ophthalmology & Visual Science*, vol. 52,no. 8, pp. 5862–5871, 2011.

10. Vapnick VN, *The Nature of Statistical learning Theory*, Springer, 1995.

11. Duda, Hart, *Pattern Classification and Scene Analysis*, Wiley Interscience, 1973

12. Cover, T.M., Hart, P.E. Nearest neighbor pattern classification. IEEE Trans. Inform. Theory, IT-13(1):21–27, 1967.

13. Jolliffe I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002, XXIX, 487 p. 28 illus. ISBN 978-0-387-95442-4

14. Margarita Sordo and Qing Zeng, "On Sample Size and Classification Accuracy: A Performance Comparision," *Lecture Notes in Computer Science*, vol. 3745, pp.193-201, 2005