Microaneurysm Detection Using Principal Component Analysis and Machine Learning Methods

Wen Cao, Nicholas Czarnek, Juan Shan, and Lin Li

***Abstract*— Diabetic retinopathy (DR) is an eye abnormality caused by long-term diabetes and it is the most common cause of blindness before the age of 50. Microaneurysms (MAs), resulting from leakage from retinal blood vessels, are early indicators of DR. In this paper, we analyzed MA detectability using small 25 by 25 pixel patches extracted from fundus images in the DIAbetic RETinopathy DataBase - Calibration Level 1 (DIARETDB1). Raw pixel intensities of extracted patches served directly as inputs into the following classifiers: random forest (RF), neural network, and support vector machine. We also explored the use of two techniques (principal component analysis and RF feature importance) for reducing input dimensionality. With traditional machine learning methods and leave-10-patients-out cross validation, our method outperformed a deep learning-based MA detection method, with AUC performance improved from 0.962 to 0.985 and F-measure improved from 0.913 to 0.926, using the same DIARETDB1 database. Furthermore, we validated our method on a different dataset—retinopathy online challenge (ROC) data set. The performance of the three classifiers and the pattern with different percentage of principal components are consistent on the two data sets. Especially, we trained the RF on DIARETDB1 and applied it to ROC; the performance is very similar to that of the RF trained and tested using cross validation on ROC data set. This result indicates that our method has the potential to generalize to different datasets.**

***Index Terms*— Feature representation, automated microaneurysm detection, diabetic retinopathy, random forest, support vector machine, neural network.**

I. Introduction

DIABETIC Retinopathy (DR) is a progressive disease with almost no early symptoms of vision impairment, which is the leading cause of blindness prior to the age of 50 [2], [3]. The first detectable sign of DR is the presence of microaneurysms (MAs), which result from leakage of tiny blood vessels in the retina and manifest themselves as small red circular spots on the surface of retinas. Early detection of MAs is critical for diagnosis and treatment of DR, which has led to a great deal of research towards automatic detection of MAs.

Several methods have been proposed for MA detection. Quellec *et al.* [4] explored the use of template matching in the wavelet domain. This method was further substantiated following the University of Iowa's release of the Retinopathy Online Challenge (ROC) database and subsequent competition for MA detection [5], in which the competition winner extended the wavelet domain template matching method. Ram *et al.* [6] created a clutter rejection strategy, in which successive stages of the algorithm eliminated more and more clutter, while passing most target MAs. One recent work [7] proposed a comprehensive grading system for DR based on classification of 16 features that captured shape, color, and intensity information and the features were extracted from candidate regions.

Many existing MA detection methods rely on hand-crafted features, which are often based on low-level information. Low-level information is easily susceptible to signal drift artifacts and thus prevent reliable generalization among different research sites. A recent method [1] leveraged the use of deep learning for MA detection using a Stacked Sparse Autoencoder (SSAE). Deep learning approaches often learn high-level and robust attributes directly from the raw signal input, and have been successfully applied to various classification and recognition tasks [8]–[10]. In [1], small image patches were generated from the original fundus images and used by the SSAE to learn high-level features from pixel intensities. These patches were then classified as either MA or non-MA using the high-level features learned by SSAE.

This work is inspired by the deep learning SSAE analysis performed in [1] and extends our previous MA analysis [23]. Our goal is to determine whether traditional machine learning methods can achieve similar or better performance on the same fundus image dataset by exploring the full context of the image information, especially when the size of the dataset may be too limited for reliably training deep learning

Manuscript received May 11, 2018; accepted May 13, 2018. Date of publication May 24, 2018; date of current version July 31, 2018. *(Corresponding author: Juan Shan.)*

W. Cao and J. Shan are with the Seidenberg School of Computer Science and Information Systems, Pace University, New York City, NY 10038 USA (e-mail: wc38764w@pace.edu; jshan@pace.edu).

N. Czarnek is with the Pratt School of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (e-mail: nczarnek@gmail.com).

L. Li is with the Department of Computer Science and Software Engineering, Seattle University, Seattle, WA 98122 USA (e-mail: lil@seattleu.edu).

Digital Object Identifier 10.1109/TNB.2018.2840084

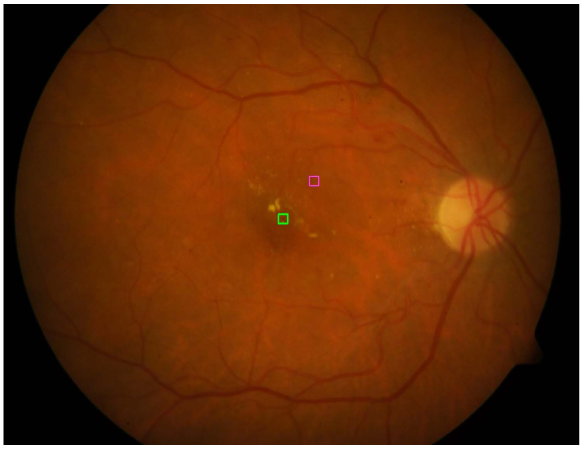


Fig. 1. An example of fundus images from DIARETDB1. The magenta square marks an example of 25 × 25 H_0 patch which contains no MA lesion while the green square marks an example of 25 × 25 H_1 patch which contains a MA lesion in the center.

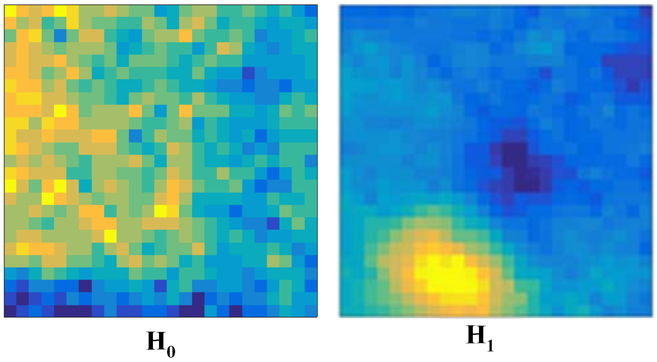


Fig. 2. Enlarged of H_0 (non-MA) and H_1 (with MA) image patches from the fundus image example in Fig. 1. Bright yellow corresponds to pixels with the highest intensities, while dark blue corresponds to pixels with the lowest intensities.

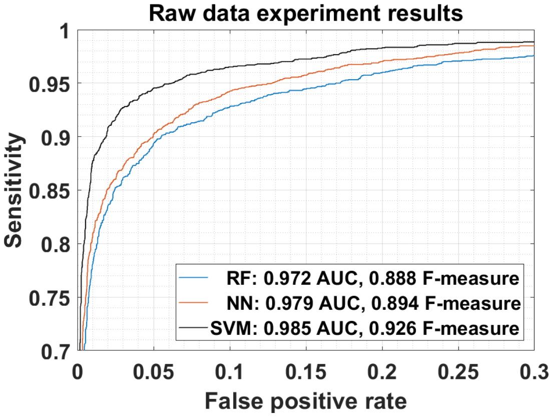


Fig. 3. ROC curves of the three classifiers using rasterized raw data of DIARETDB1.

networks. We explore the utility of raw data, extracted from the public DIAbetic RETinopathy DataBase - Calibration Level 1 (DIARETDB1) [11], and ROC [22] for MA classification using random forests (RFs) [12], neural networks (NNs) [13], and support vector machines (SVMs) [14]. The performance is measured by area under the receiver operating characteristic (ROC) curve (AUC) [15] and F-measure [16].

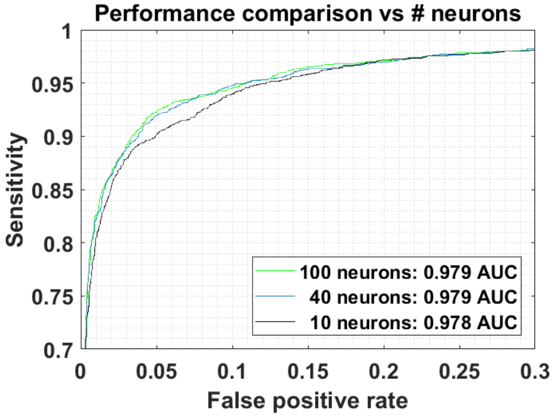


Fig. 4. ROC curves of a single hidden layer NN with different number of neurons.

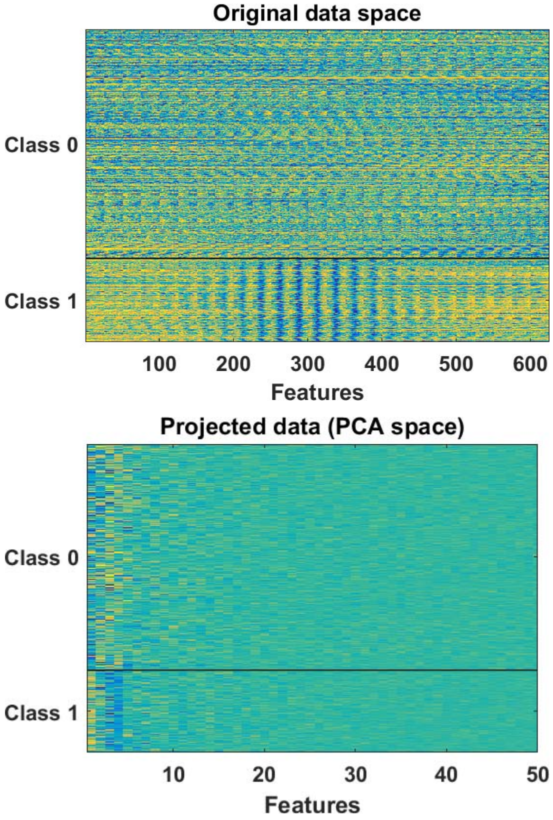
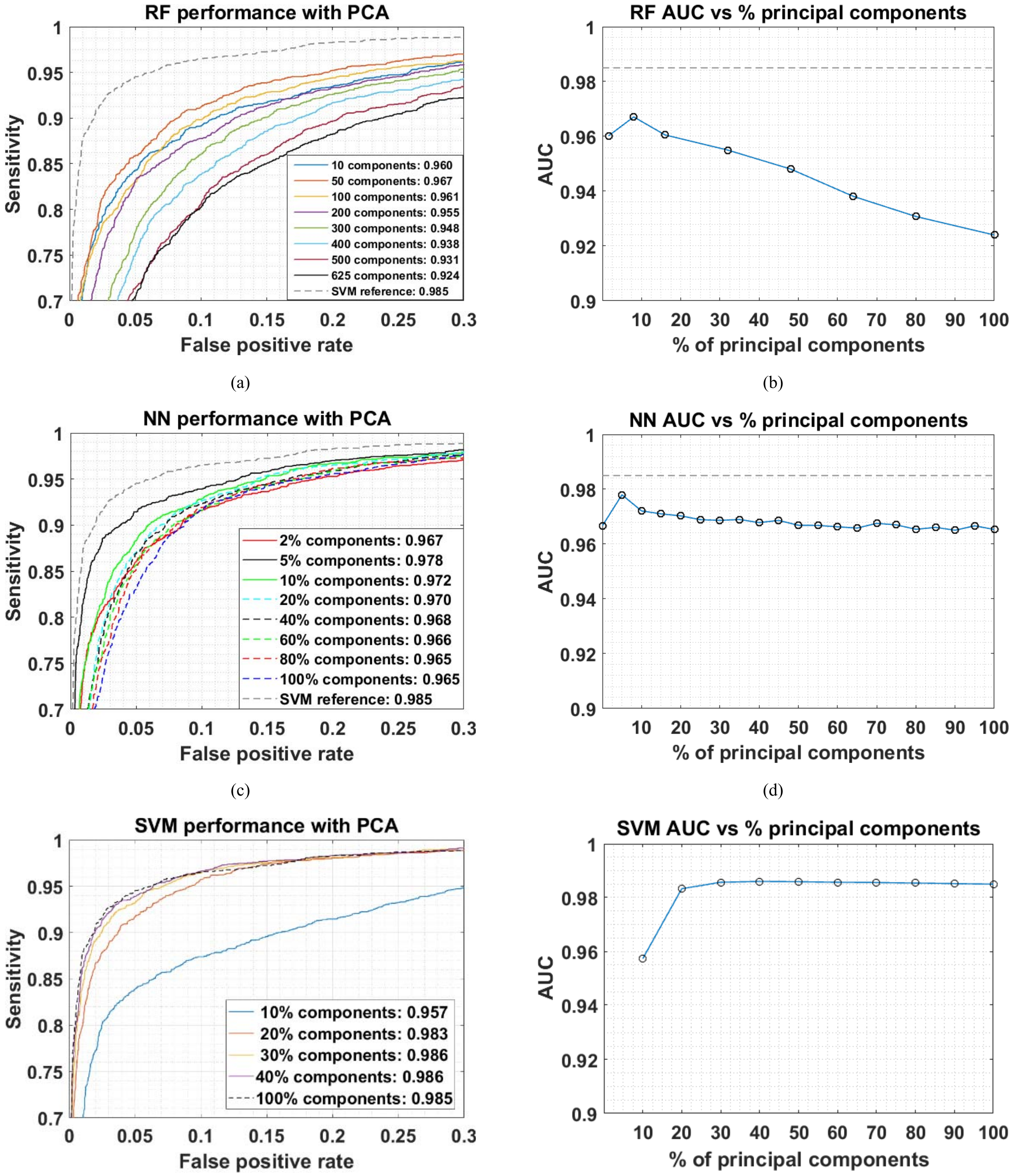


Fig. 5. Original data space after normalization (on the top) and the top 50 principal component subspace (on the bottom). Yellow corresponds to high intensity while blue corresponds to low intensity.

The rest of the paper is organized as follows. In Section II, we describe the materials and methods used in this work, including data and features, cross validation strategy, classifiers employed for MA detection and evaluation metrics. In Section III and IV, we present and analyze the experimental results on DIARETDB1 and ROC datasets respectively. Finally, in Section V, we draw conclusions and discuss future work.



(a) (b)

(c) (d)

(e) (f)

Fig. 6. Performance comparison using the RF, NN, and SVM with different amounts of principal components included for classification on DIARETDB1. The RF and NN ROCs include the SVM performance with 100% of components for reference, and the AUC plots also include the SVM AUC with 100% of components for reference (dashed lines in (b) and (d)). As shown, the SVM performance roughly saturates with 30-40% of components. However, RF and NN performance reaches the peak using ~5% of principal components before decreasing steadily.

II. Materials and Methods

A. Data

We analyzed DIARETDB1 [11] images that were acquired from 89 patients, in which each image was manually annotated by multiple experts at Kuopio hospital. 84 of the 89 patients contained at least mild non-proliferative signs of DR, while the remaining five patients were healthy with no signs of DR. From each patient image, patches were extracted and labeled as H_1 or H_0, for locations in which MAs were present or absent, respectively, based on information provided in DIARETDB1 ground truth files. Each

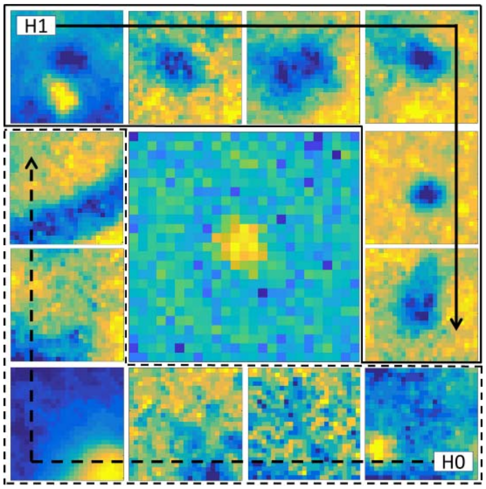


Fig. 7. Examples of H_0 (in dashed lines) and H_1 (in solid lines) image patches, surrounding an image of the feature importance (yellow is high) learned by a RF in the center block. Brighter pixel locations in the central image correspond to more important components of the image, indicating that the center of the 25 × 25 image patches contains the most important information for discrimination between MAs and non-MAs.

image patch had a size of 25×25 pixels, with an image resolution of approximately 0.015-0.02 mm/pixel. As in [1], 2182 H_1 patches were directly extracted from locations manually labeled as MA and centered at the marked ground truth H_1 locations, while 5963 H_0 patches were extracted from uniformly random locations throughout the image. An example of DIARETDB1 fundus images is shown in Fig. 1, while examples of H_1 and H_0 image patches cut from the fundus image are shown in Fig. 2. Our analysis focused on the use of the green channel from the provided RGB images, since MA lesions have the highest contrast in the green plane compared to other color planes [4], [17].

The ROC dataset [22] contained 50 patients, 28 of whom had images with roughly the same resolution and size as those found in DIARETDB1. From this data subset, we extracted features (25×25 pixel patches centered on each location) based on the ground truth provided in the dataset reference. The 28 patients had a total of 201 MAs. In order to keep the balance of the dataset and make the dataset roughly consistent with the ratio of MA patches vs non-MA patches that we extracted from DIARETDB1, we randomly extracted 588 non-MA patches (21 patches/patient) from the ROC dataset.

B. Features

Classification of MA vs non-MA patches was performed using three sets of features. The first feature set consisted of raw pixel intensities, rasterized from the image patches. The dimensionality of the feature space was 25×25, i.e., 625. Although raw pixels yielded good discriminability between patch classes, feature space reduction may be valuable to reduce both classifier training time and concerns regarding the curse of dimensionality [18]. We used two dimensionality reduction techniques to process the original feature set.

We first employed principal component analysis (PCA) [19], a common feature dimensionality reduction method. PCA projects data onto a new space in which consecutive dimensions contain less and less of the variance of the original dataspace and compresses the most important information onto a subspace with lower dimensionality than the original space. As a preprocessing step, we whitened data to zero mean and unit variance. For both PCA and whitening, only training data was used to establish the mean, variance, and principal components of the original data. We tested the feature space with 5-100% of the projected subspace using leave-10-patients-out cross-validation to inspect how many principal components are needed to generate maximum performance.

We also explored the use of the RF feature importance for dimensionality reduction. Feature importance is a measure of the performance loss that would occur if a dimension of the original data space is removed using the RF as the classifier [12]. After sorting dimensions from highest to lowest feature importance, we selected data with top 5-100% of the original space to determine the point at which performance saturated.

C. Classifiers

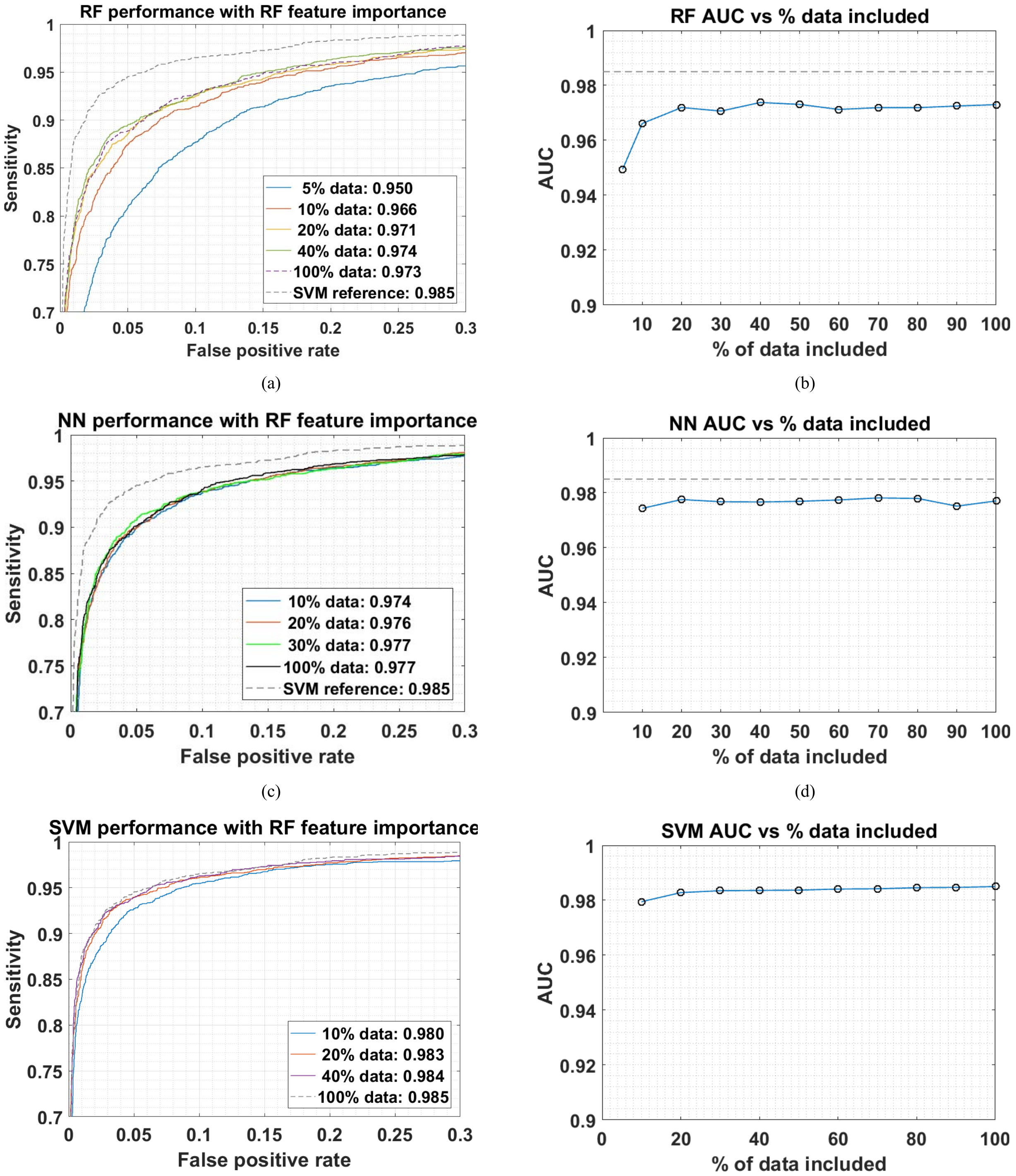
We employed three classifiers in our experiments: RFs [12] with 25 trees, single hidden layer NNs [13] with 10-100 neurons, and SVMs [14] with radial basis function kernels. RFs are quick to train and agnostic to the scale of input features. NNs are powerful classifiers, but often require large training datasets and fine tuning of classifier parameters. SVMs use kernel functions to map data into higher dimensionality, in which boundaries are learned to separate the two classes of data. The SVM was implemented with LibSVM [20] with a radial basis function kernel, while both the RF and NN were implemented using native MATLAB R2017a functions.

D. Evaluation

Leave-10-patients-out cross-validation, similar to that described in [1] was used for testing, in which data from 10 patients were held out for testing, while data from the remaining patients were used for training. It is important to note that for all experiments, all preprocessing, feature extraction, and classification were performed in a fully cross validated way to ensure that classifier training was not contaminated by test data.

We used two common metrics to evaluate performance of our experiments: the area under the receiver operating characteristic (ROC) curve (AUC) [15] and the F-measure [16]. ROC curves provide an indication of the tradeoff between classification sensitivity and specificity as the classifier confidence threshold increases or decreases. The F-measure, shown below, provides an indication of overall classification accuracy as a weighted average of precision and recall for a specified confidence threshold.

$$F\text{-}measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{1}$$



(a) (b) (c) (d) (e) (f)

Fig. 8. Performance comparison using the RF, NN, and SVM with different amounts of data included as selected by RF feature importance on DIARETDB1. The RF and NN ROCs include the SVM performance with 100% of components for reference, and the AUC plots also include the SVM AUC with 100% of components for reference. In contrast to PCA results, performance of all three classifiers maintains approximately maximum performance once AUC saturation occurs.

III. Experimental Results on DIARETDB1

A. Experiment 1: Classification of Raw Data

Fig. 3 shows the results of Experiment 1 achieved by using the raw data in cross validation with the three classifiers. As shown, the SVM achieved the highest performance, and outperformed the work in [1] in terms of both the AUC and F-measure (0.985 AUC and 0.926 F-measure compared to 0.962 AUC and 0.913 F-measure achieved by [1]), while the RF and NN outperformed the work in [1] in terms of AUC. The performance of the NN slightly increased as the number of neurons in the network increased, as shown in Fig. 4. Performance saturated with approximately 40 neurons, with little benefit gained by increasing to 100 neurons. Additional

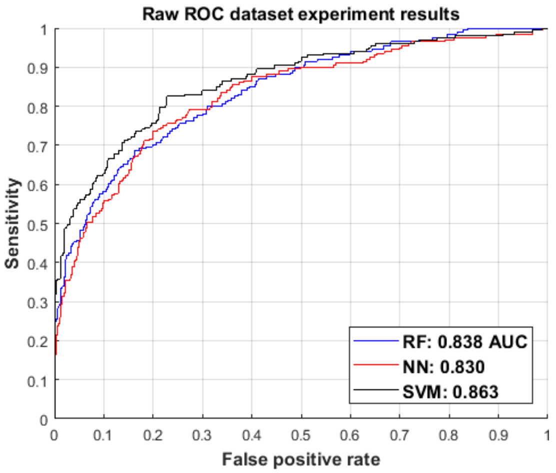


Fig. 9. ROC curves of the three classifiers using rasterized raw data from ROC dataset.

layers did not yield marked improvements, so neural network results are shown using one layer.

B. Experiment 2: Classification Using Principal Components

Fig. 5 shows a comparison between the original dataspace and a principal component subspace. As shown, a subset of the top principal components can be used to visually discriminate between Class 0 (non-MA) and Class 1 (MA). As shown, Class 1 (MA) and Class 0 (non-MA) are visually distinguishable in the original dataspace using the entire feature space. By projecting data onto the principal component axes, classes are visually separable using only the top 5-10 principal components.

The performances of all three classifiers, in terms of AUC, using various amounts of principal components are shown in Fig. 6. As shown, the SVM outperforms the RF and NN, which is similar to Experiment 1. For both the RF and NN, the performance reaches the peak with ∼5% of principal components, then decreases as the number of principal components increases. A further discussion about this performance drop is provided in Section V.

C. Experiment 3: Classification Using Features Selected by RF Feature Importance

Fig. 7 shows a montage of H_1 and H_0 examples surrounding a central image of the feature importance learned by a RF classifier. As shown, the center of the 25×25 pixel patches appear to be the most important components of the images. Selecting the input space using the RF feature importance, rather than the principal components, yields the performance shown in Fig. 8. In contrast to the PCA experiments, the performance, as indicated by AUC, increases as dimensionality increases for all classifiers. However, similar to the PCA experimental results, using only 10% of the feature space, the SVM outperformed both the NN and RF.

IV. Experimental Results on Roc Dataset

A. Experiment 1: Classification of Raw Data

Fig. 9 shows the results of Experiment 1 achieved by using the raw data in cross validation with the three classifiers. Since the dataset is smaller, we ran leave-3-patients-out cross-validation on the dataset, with all aspects of the experiment being fully cross validated. As Fig. 9 showed, the SVM achieved the highest performance 0.863 AUC, while the RF and NN achieved 0.838 and 0.830 in terms of AUC.

B. Experiment 2: Classification Using Principal Components

The difference between this experiment and experiment 1 was that we incorporated PCA to analyze the feature space. Again, PCA was used in a fully cross validated way, in that the principal components were learned using a training set, and testing data was projected onto these principal components. The performances of all three classifiers, indicated by AUC, using various amounts of principal components are shown in Fig. 10. An interesting thing is that we observed similar results and patterns as we observed on DIARETDB1. First, the SVM (best AUC = 0.870 with 40% PCA) outperforms the RF (best AUC = 0.818 with 10% PCA) and the NN (best AUC = 0.832 with 5% PCA). For both the RF and NN, the performance reaches the peak with 5% - 10% of principal components, then decreases as the number of principal components increases. For SVM, the performance becomes saturate and stable after principal components increase to 30%. The pattern is similar to the result of classification using principle components on DIARETDB1 (shown in Fig. 6).

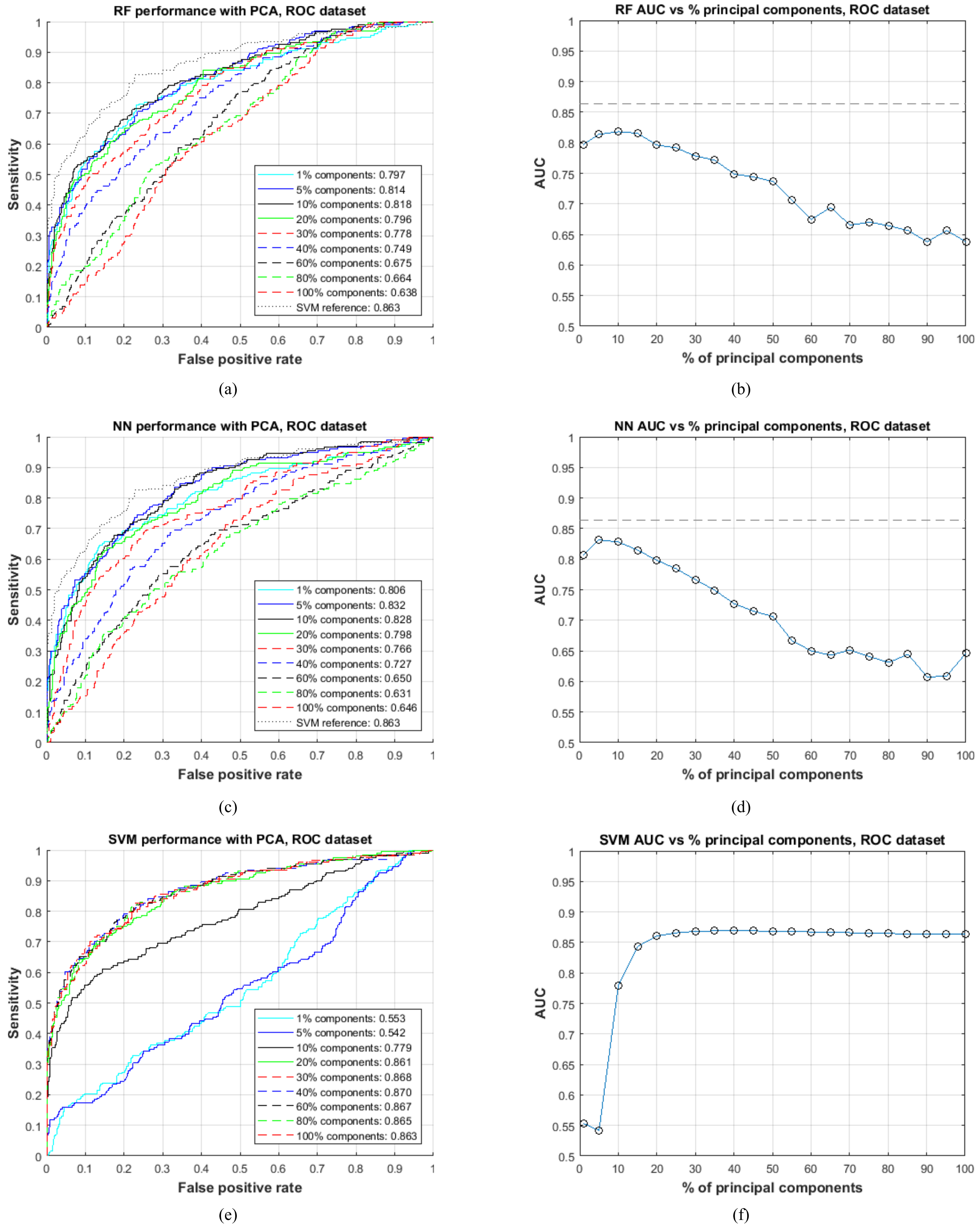
C. Experiment 3: Classification Using DIARETDB1 as Training Data and ROC as Testing Data

In this experiment, we trained a random forest classifier using the DIARETDB1 data, and then applied this classifier to the ROC dataset. The AUC achieved using this method was 0.810, close to that of using full cross-validation on the ROC dataset, which was 0.838. This result indicates that our method has the potential to generalize across different datasets.

V. Discussion and Conclusions

In this paper, we have presented several common machine learning methods to detect microaneurysms in fundus images for the diagnosis of diabetic retinopathy using the publicly available database DIARETDB1. The machine learning methods employed in this work include RF, NN, and SVM. Our methods outperformed previous work [1] that conducted deep learning using the same database, measured by both AUC and F-measure. Further validation of our method on a different dataset ROC showed similar results as we found on DIARETDB1. Our results yield a promising step towards automated early detection of microaneurysms and diabetic retinopathy.

Three interesting results were observed in our findings. First, after application of PCA to the original 625-dimensional dataspace, SVMs achieved near maximum performance using



(a) (b)

(c) (d)

(e) (f)

Fig. 10. Performance comparison using the RF, NN, and SVM with different amounts of principal components included for classification on ROC dataset. The RF and NN ROCs include the SVM performance with 100% of components for reference, and the AUC plots also include the SVM AUC with 100% of components for reference (dashed lines in (b) and (d)). As shown, the SVM performance roughly saturates with 30-40% of components. However, RF and NN performance reaches the peak using $5\% \sim 10\%$ of principal components before decreasing steadily.

only 30% of data. With the incorporation of more principal components, the performance of the SVM plateaued, while the performance of the RF and NN both decreased. Upon further investigation, we hypothesized that this performance drop is due to overtraining on uninformative, or noisy, dimensions. PCA projects most information in signals onto a subset of dimensions, or the first few principal components, in a new subspace, rendering the remainder of dimensions largely uninformative. RFs create an ensemble of decision trees, each with a random subset of dimensions selected for each node of the tree. For experiments with inclusion of more than approximately 5% of principal components (see Fig. 6), if the random subset for a given node does not include one of the informative principal components, the data partition from this node may be meaningless and lead to poor generalization performance. The NNs may be similarly overtrained to these uninformative dimensions, yielding poor performance when more dimensions are included. This result indicates that SVMs are more robust and less sensitive to noisy features than NNs and RFs in classifying MA and non-MA patches.

Second, by reducing dimensionality using RF feature importance, we were able to achieve high AUC with fewer dimensions (see Fig. 8) and maintain this high AUC as more dimensions were included. This is an important difference from the PCA experiment in that performance results are more robust and generalizable to new data. With this approach, fine tuning of the number of principal components is no longer required for any of the classifiers.

Third, when we trained the random forest on DIARETDB1 and used ROC as the testing data, the performance was similar as that achieved by the random forest trained and tested on the ROC dataset using cross-validation. Such a finding indicates that our method has the potential to generalize across different datasets.

In the future, we plan to extract and classify features from all locations within an eye to generate a two dimensional classifier confidence map, which we will then score using algorithms such as the RX anomaly detector [21]. With this new processing protocol, we plan to use a different performance metric, with results reported in terms of sensitivity vs number of false positives (false MAs) per eye, rather than sensitivity vs false positive rate. This would both provide optometrists with a more easily understandable metric and allow algorithm developers to compare performance with a metric that is more agnostic to database size, sub-selection of patches within images, etc.

References

[1] J. Shan and L. Li, "A deep learning method for microaneurysm detection in fundus images," in *Proc. IEEE 1st Int. Conf. Connected Health, Appl., Syst. Eng. Technol. (CHASE)*, Jun. 2016, pp. 357–358.

[2] R. Klein, E. K. Barbara, and S. E. Moss, "Visual impairment in diabetes," *Ophthalmology*, vol. 91, no. 1, pp. 1–9, 1984.

[3] A. K. Sjølie *et al.*, "Retinopathy and vision loss in insulin-dependent diabetes in Europe: The EURODIAB IDDM complications study," *Ophthalmology*, vol. 104, no. 2, pp. 252–260, 1997.

[4] G. Quellec, M. Lamard, P. M. Josselin, G. Cazuguel, B. Cochener, and C. Roux, "Optimal wavelet transform for the detection of microaneurysms in retina photographs," *IEEE Trans. Med. Imag.*, vol. 27, no. 9, pp. 1230–1241, Sep. 2008.

[5] M. Niemeijer *et al.*, "Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 185–195, Jan. 2010.

[6] K. Ram, G. D. Joshi, and J. Sivaswamy, "A successive clutter-rejection-based approach for early detection of diabetic retinopathy," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 664–673, Mar. 2011.

[7] M. U. Akram, S. Khalid, A. Tariq, S. A. Khan, and F. Azam, "Detection and classification of retinal lesions for grading of diabetic retinopathy," *Comput. Biol. Med.*, vol. 45, no. 1, pp. 161–171, 2014.

[8] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 609–616.

[9] H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1–9.

[10] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3361–3368.

[11] T. Kauppi *et al.*, "DIARETDB1 diabetic retinopathy database and evaluation protocol," in *Proc. Brit. Mach. Vis. Conf.*, vol. 1, 2007, p. 15.1.

[12] L. Breiman, "Random forests," *Mach. Learn.*, vol. 4, no. 1, pp. 5–32, 2001.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[15] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[16] C. J. Van Rijsbergen, *Information Retrieval*. Newton, MA, USA: Butterworth-Heinemann, 1979.

[17] A. D. Fleming, S. Philip, K. A. Goatman, J. A. Olson, and P. F. Sharp, "Automated microaneurysm detection using local contrast normalization and local vessel detection," *IEEE Trans. Med. Imag.*, vol. 25, no. 9, pp. 1223–1232, Sep. 2006.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.

[19] J. Shlens. (2014). "A tutorial on principal component analysis." pp. 1–13. [Online]. Available: https://arxiv.org/abs/1404.1100

[20] C. Chang and C. Lin, "LIBSVM," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[21] I. S. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 38, no. 10, pp. 1760–1770, Oct. 1990.

[22] M. Niemeijer *et al.*, "Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 185–195, 2010.

[23] W. Cao, J. Shan, N. Czarnek, and L. Li, "Microaneurysm detection in fundus images using small image patches and machine learning methods," in *Proc. IEEE Int. Conf. Bioinform. Biomed. (BIBM)*, Kansas City, MO, USA, 2017, pp. 325–331.