A New Image Data Set and Benchmark for Cervical Dysplasia Classification Evaluation

Tao Xu[1(✉)], Cheng Xin[1], L. Rodney Long[2], Sameer Antani[2], Zhiyun Xue[2], Edward Kim[3], and Xiaolei Huang[1]

[1] Computer Science & Engineering Department, Lehigh University, Bethlehem, PA, USA
tax313@lehigh.edu
[2] Communications Engineering Branch, NLM, Bethesda, MD, USA
[3] Computing Sciences Department, Villanova University, Villanova, PA, USA

Abstract. Cervical cancer is one of the most common types of cancer in women worldwide. Most deaths of cervical cancer occur in less developed areas of the world. In this work, we introduce a new image dataset along with ground truth diagnosis for evaluating image-based cervical disease classification algorithms. We collect a large number of cervigram images from a database provided by the US National Cancer Institute. From these images, we extract three types of complementary image features, including Pyramid histogram in L*A*B* color space (PLAB), Pyramid Histogram of Oriented Gradients (PHOG), and Pyramid histogram of Local Binary Patterns (PLBP). PLAB captures color information, PHOG encodes edges and gradient information, and PLBP extracts texture information. Using these features, we run seven classic machine-learning algorithms to differentiate images of high-risk patient visits from those of low-risk patient visits. Extensive experiments are conducted on both balanced and imbalanced subsets of the data to compare the seven classifiers. These results can serve as a baseline for future research in cervical dysplasia classification using images. The image-based classifiers also outperform results of several other screening tests on the same datasets.

1 Introduction

Cervical cancer ranks as the second most common type of cancer in women aged 15 to 44 years worldwide [1]. Among death cases caused by cervical cancer, over 80% occurred in less developed regions. Therefore, there is a need for lower cost and more automated screening methods for early detection of cervical cancer, especially those applicable in low-resource regions. Screening procedures can help prevent cervical cancer by detecting cervical intraepithelial neoplasia (CIN), which is the potentially precancerous change and abnormal growth of squamous cells on the surface of the cervix. According to the WHO system [1], CIN is divided into three grades: CIN1 (mild), CIN2 (moderate), and CIN3 (severe).

C. Xin—Co-first author

L. Zhou et al. (Eds.): MLMI 2015, LNCS 9352, pp. 26–35, 2015.
DOI: 10.1007/978-3-319-24888-2_4

Lesions in CIN2/3+ require treatment, whereas mild dysplasia in CIN1 only needs conservative observation because it will typically be cleared by an immune response in a year. Thus, in clinical practice one important goal of screening is to differentiate CIN1 from CIN2/3 or cancer (denoted as CIN2/3+ [2]).

The most widely used cervical cancer screening methods today include the Pap test, HPV testing, and visual examination. Digital Cervicography, a non-invasive visual examination method that takes a photograph of the cervix (called a cervigram) after the application of 5% acetic acid to the cervix epithelium, has great potential to be a primary or adjunctive screening tool in developing countries because of its low cost and accessibility in resource-poor regions. However, one concern with Cervicography is that the overall effectiveness of Cervicography has been questioned by reports of poor correlation between visual lesion recognition and high-grade disease as well as disagreement among experts when grading visual findings. To address the concern and investigate the feasibility of using images as a screening method for cervical cancer, we conjecture that computer algorithms can be developed to improve the accuracy in grading lesions using visual (and image) information. This conjecture is inspired and encouraged by recent successes in computer-assisted Pap tests such as the ThinPrep Imaging System (TIS) [3], FocalPoint [4], and the work by Zhang et al. [5]; these computer-assisted Pap tests apply multi-feature Pap smear image classification using SVM and other machine learning algorithms, and they have been shown to be statistically more sensitive than manual methods with equivalent specificity.

In this work, we describe our efforts of building a dataset of multiple features extracted from cervigram images along with patient diagnosis ground truth based on worst histology. We also present some baseline results of applying seven classic machine-learning algorithms to differentiate patient visits that are high-risk from those visits that are low-risk, using cervigrams. We train binary classifiers to separate CIN1/Normal and CIN2/3+ images. All the classifiers are trained and tested on the same datasets, with a uniform parameter optimization strategy. They are then compared by ROC curves and other evaluation measures. Moreover, we compare the performance of cervigram based classifiers with Pap tests and HPV tests results on the same datasets.

2 The Image Data Set for CIN Classification

Here we introduce a dataset for image-based CIN classification, built from a large medical data archive collected by the National Cancer Institute (NCI) in the Guanacaste project [6]. The archive consists of data from 10,000 anonymized women, and the data is stored in the Multimedia Database Tool (MDT) developed by the National Library of Medicine [7]. In the archive, each patient typically had multiple visits at different ages. During each visit, multiple cervical screening tests including Cervicography were performed. The Cervicography test produced two cervigram images for a patient during her visit and the images were later sent to an expert for interpretation.

In our dataset, we collected 345 positive (CIN2/3/cancer) patient visits and 767 negative (CIN1/Normal) patient visits from NLM's MDT. The ground truth

diagnosis (i.e. the CIN grade) for each patient visit is based on the Worst Histology result of the visit. Multiple expert histology interpretations were done on each biopsy; the most severe interpretation is labeled the Worst Histology for that visit in the database. Then, for each patient visit, we take the pair of cervigram images for that visit, and extract three types of features from the images: the Pyramid histogram in L*A*B* color space (PLAB) feature, the Pyramid Histogram of Oriented Gradients (PHOG) feature, and the Pyramid histogram of Local Binary Patterns (PLBP) feature. The PLAB feature captures color information; the PHOG feature encodes edges and gradient information; and the PLBP feature extracts texture information. More details about the PLAB-PHOG-PLBP features and their extraction process can be found in [8]. For each image, after feature extraction, the total length of the concatenated PLAB-PHOG-PLBP feature is 2,538. Note that there are two images from each patient visit, which are visually similar but not identical. We have to avoid using one image for training while the other image is being used for testing. Thus we construct two separate image datasets, D1 and D2, and randomly assign one image of a visit to D1 and assign the other image from the same visit to D2. D1 and D2 are used separately in experiments, and each set contains 345 images from positive visits and 767 images from negative visits.

Our image dataset along with the ground truth diagnosis for each image can be used as a new image feature benchmark to evaluate automated cervical dysplasia (i.e. CIN) grading or classification algorithms.

3 Seven Classifiers for Comparison

On the cervigram image benchmark datasets introduced above, we compare seven classic machine learning methods, including random forest (RF), gradient boosting decision tree (GBDT), AdaBoost, support vector machines (SVM), logistic regression (LR), multilayer perceptron (MLP), and k-Nearest Neighbors (kNN). Some of them, such as SVM, have been widely used in the field of medical image analysis [9–12], while others, like random forest and GBDT, are witnessing applications only in the recent few years [13]. In the literature, there have been other works that aim to compare classifier performances on benchmark datasets. Morra et al. [9] compared AdaBoost with SVM while Osareh et al. [10] compared SVM with neural networks. In both papers, the comparisons were done between two classifiers. In the work by Wei et al. [11], more classifiers were studied, but excellent ensemble methods like RF and GBDT were not included. In this paper, we conduct a comprehensive comparison of seven popular classifiers. Next, we will briefly introduce each of them.

Random Forest (RF) is an increasingly popular machine learning method [14]. It builds an ensemble of many decision trees trained separately on a bootstrapped sample set of the original data. Each decision tree grows by randomly selecting a subset of candidate attributes for splitting at each node. We optimize parameters for RF by searching the number of trees in {10, 100, 200, 500, 1000, 2000} and

searching the subset size of features for node splitting among {'sqrt', 100, 200, 500, 1000, 2000} where 'sqrt' is the square root of the whole feature size.

Gradient Boosting Decision Tree (GBDT) is a kind of additive boosting model which, in general, can be expressed as function (1)

$$f(x) = \sum_{m=1}^{M} \beta_m b(x; \gamma_m) \tag{1}$$

where β is called expansion coefficient, serving as the weight of the tree in each iteration, and $b(x; \gamma)$ are usually simple basic functions, e.g. decision tree, characterized by parameters γ. Details for the training process of GBDT can be found in [14]. We optimize the parameters for GBDT by searching the number of trees among {10, 100, 200, 500, 1000, 2000} and the learning rate in {1, 0.1, 0.01, 0.001, 0.0001}.

Adaboost is a classic boosting tree model [15]. It has the form $H(x) = \sum_t \alpha_t h_t(x)$, which can be trained by minimizing the loss function in a greedy fashion. An optimal weak classifier h_t is selected for each training iteration t. We use shallow decision trees (i.e. stumps) as the weak learners. In the final strong classifier $H(x)$, the weight of the weak classifier $h_t(x)$ is α_t, which is inversely proportional to the classification error of $h_t(x)$. To optimize parameters for AdaBoost, we search the depth (d) of each decision tree in {1, 2, 3, 4} and the number of weak classifiers from 10 to the whole feature size with an increment of 120/d.

Multilayer Perceptron (MLP) is a feedforward neural network. MLP uses layerwise connected nodes to build the architecture of the model. Each node(except for the input nodes) can be viewed as a neuron with a nonlinear activation function. In this paper, we use the sigmoid function(2) as the activation function,

$$\sigma(x) = \frac{1}{1 + exp(-(w * x + b))} \tag{2}$$

where the weight vector w and bias vector b in each layer pair are trained by the Back Propagation algorithm. We also introduce L2 regularization weight decay to prevent overfitting. We optimize hyperparameters for MLP by searching the hidden layer size in {2, 3}, the hidden unit size in {0.0625*m, 0.125*m, 0.25*m} where m is the feature size 2538, and searching the weight decay strength among {0.0005, 0.0001, 0.00001, 0.0}.

Logistic Regression is a type of probabilistic statistical classification model. For the binary classification problem, with labeled sample set $\{(x_i, y_i)\}_{i=1}^{N}$, it computes the positive probability by (3) and the model parameter θ is trained to minimize the cost function(4).

$$P_1(x_i) = \frac{1}{1 + exp(-\theta^T * x_i)} \tag{3}$$

$$L(\theta) = -\frac{1}{N}[\sum_{i=1}^{N} y_i log P_1(x_i) + (1 - y_i) log(1 - P_1(x_i)] \tag{4}$$

In our experiments, we use the batch gradient descent algorithm with L2 regularization to train the model. The strength of regularization is searched from 10^{-5} to 10^5, with an increment of 1 for the exponent.

Support Vector Machines (SVM) is one of the most widely used classifiers in medical image analysis [2,5,9,10]. It performs classification by constructing a hyperplane in a high-dimensional feature space. It can use either linear or non-linear kernels, and its effectiveness depends on the selection of kernel, the kernel's parameters, and the soft margin parameter C. Linear SVM is widely used because it has good performance and fast speed in many tasks. In this paper, we also choose to use the linear SVM; we did try nonlinear kernels such as the radial basis functions (RBF) but they are time consuming and did not improve performance in our task. For linear SVM, we need to optimize the parameter C. Let $C = 2^m$, we search m in the range [-8, 9] with a step increment of 1.

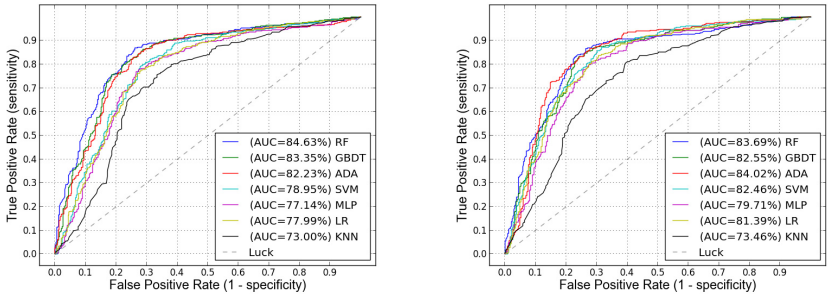
k-Nearest Neighbors (kNN) is one of the simplest classifiers, which classifies a new instance by a majority vote of its k nearest neighbors. In this paper, we use the Euclidean distance metric to find the k nearest neighbors. We search the optimal k value for our task in the range [1, 50] with a step increment of 1.

4 Experiments

In Section 2, we described the construction of two cervigram image datasets, D1 and D2, where each one contains 345 images from positive (CIN2/3+) patient visits and 767 images from negative (CIN1/normal) patient visits. Note that the datasets are imbalanced, i.e. they contain more negative cases than positive cases. Since many classification methods assume a balanced distribution of classes and require additional strategies to handle imbalanced data, we apply undersampling to the negative visits and randomly choose 345 negative visits from each dataset. The resulting two balanced datasets, D_1^{bal} and D_2^{bal}, use all 345 positive visits and the randomly selected 345 negative visits.

We conduct experiments to compare the seven classifiers described in Section 3, on the two balanced datasets D_1^{bal} and D_2^{bal}, and on the two larger imbalanced datasets, D_1 and D_2. The classifier implementations we use are from well known open source libraries. Our Random Forest, GBDT, and LR classifiers are implemented with scikit-learn [16]; the MLP classifier is provided by pylearn2 [17]; the SVM is offered by Libsvm [18]; the AdaBoost is provided by Appel et. al. [15]; and the kNN classifier is provided by the implementation in MATLAB.

We perform the same ten-round ten-fold cross validation using these seven classifiers. On each dataset, we randomly divide the samples (cervigrams) into



(a) ROC curves on D_1^{bal} (b) ROC curves on D_2^{bal}

Fig. 1. ROC curves on balanced datasets D_1^{bal} and D_2^{bal}.

ten folds. In the ten rounds, we rotationally use one fold for testing and nine folds for training. On the training set, we use a uniform strategy, Exhaustive Grid Search [18], to search for the optimal parameters of each classifier. Three cross validations are used in the parameter searching process. The exact parameters and search ranges for each classifier are discussed in the Section 3.

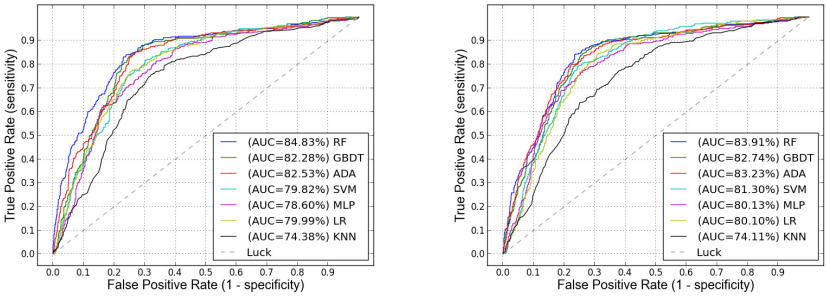
The results of the ten rounds are used to draw ROC curves. We compare different classifiers by analyzing their ROC curves, areas under ROC curves (AUC), and accuracy, sensitivity and specificity values at the point where the probability threshold is 0.5. We also compare the results of our image-based classifiers with several other screening tests results, obtained for the same visits that are used to construct our datasets.

4.1 Results on Balanced Datasets

In our first set of experiments, we compare seven classifiers on the balanced dataset D_1^{bal} and D_2^{bal}. The comparison results are shown in Fig. 1 as ROC curves and in Table 1 with overall AUCs, and accuracy, sensitivity and specificity values at the default probability threshold 0.5. The ROC curves illustrate that the three ensemble-tree models— RandomForest (RF), GBDT, and AdaBoost— outperform other classifiers. AUCs in Table 1 also show that the ensemble-tree models have a better overall performance. At the 5% significance level, there is no difference between RandomForest, GBDT and AdaBoost. On D_1^{bal}, for instance, the p value is 0.0708 by paired t-test between RF (1st rank) and AdaBoost (3rd rank). However, these three ensemble-tree classifiers are significantly better than all other classifiers. On D_1^{bal}, the p value is 0.0062 and $1.7191 * 10^{-4}$, by paired t-test between RF (1st rank) and SVM (4th rank), and between RF and kNN (lowest rank), respectively. We conjecture that the ensemble-tree models perform best because they are more robust to over-fitting than other models such as SVM and MLP when dealing with scalar data sets that are not too large.

Table 1. Overall AUC and accuracy (accu), sensitivity (sensi) and specificity (speci) at the default threshold on the balanced dataset D_1^{bal} and the imbalanced dataset D_1

	D_1^{bal}				D_1			
Classifier	AUC(%)	accu(%)	sensi(%)	speci(%)	AUC(%)	accu(%)	sensi(%)	speci(%)
RF	**84.63**	**80.00**	**84.06**	**75.94**	**84.83**	**78.24**	**67.54**	**83.05**
GBDT	83.35	78.55	82.03	75.07	82.28	77.07	62.61	83.57
AdaBoost	82.23	76.81	77.68	75.94	82.53	76.44	57.97	84.75
SVM	78.95	74.78	76.52	73.04	79.82	74.37	46.67	86.83
LR	77.99	74.20	76.23	72.17	79.99	75.45	54.20	85.01
MLP	77.14	75.27	77.78	72.75	78.60	76.53	59.13	84.35
kNN	73.00	70.87	75.07	66.67	74.38	71.67	48.12	82.27



(a) ROC curves on D_1 (b) ROC curves on D_2

Fig. 2. ROC curves on imbalanced datasets D1 and D2.

4.2 Results on Imbalanced Datasets

We also conduct the same ten-round ten-fold experiments on the imbalanced datasets D1 and D2. The results are shown in Fig. 2 and Table 1. One clear difference between results on the imbalanced datasets and those on the balanced datasets is that, at the same default threshold, all seven classifiers give higher specificity values and lower sensitivity values on the imbalanced dataset (see Table 1, right column). This is expected since in the imbalanced datasets, there are more negative samples than positive samples, thus when penalizing equally errors on samples from any class and training to minimize the overall classification error, the classifiers trained on the imbalanced data become biased to the class with a majority of samples. Interestingly, since higher specificity is a desired property for a clinical test meant for screening, training classifiers on the imbalanced dataset (which more closely reflect the true underlying patient distribution) can be beneficial.

Moreover, Fig. 2 shows that the overall ROC curves and AUCs on the imbalanced datasets are similar to that on the balanced datasets. Although more samples are used to train classifiers on the imbalanced datasets, the overall performance by the classifiers did not seem to improve.

4.3 Cervigram Based RandomForest (RF) vs. Pap and HPV Tests

In this experiment, we first compute the average result of our image-based classifier RF to represent its visit-level performance on balanced and imbalanced datasets, respectively. We then compare the visit-level result of RF with Pap and HPV tests results, which are available for the same visits that are used to construct our datasets. As illustrated in Table 2, on both datasets the image-based RF classifier outperforms every single Pap test or HPV test at specificity around 90%.

Table 2. Comparing visit-level sensitivity (sensi) and specificity (speci) of image-based RF classifier with that of Pap tests and HPV tests.

	Balanced dataset		Imbalanced dataset	
Method	sensi(%)	speci(%)	sensi(%)	speci(%)
Alfaro ThinPrep	20.69	81.82	20.69	85.27
Cytyc ThinPrep	49.55	88.46	49.55	89.77
Costa Rica Pap	39.42	88.12	39.42	89.31
Hopkins Pap	36.00	97.11	36.00	97.13
HPV16	33.82	94.19	33.82	92.94
HPV18	08.16	97.97	08.16	98.17
Cervigram based RF	51.00	90.00	49.00	90.00

5 Conclusions

In this paper, we present a new benchmark dataset for evaluating cervical dysplasia classification or grading algorithms. Both image features and ground truth diagnosis are included in the dataset. It is our intention to publish[1] the original datasets D1 and D2, sample images and the source code for PLAB-PHOG-PLBP image feature extraction. We will also add information from other screening tests such as Pap and HPV and expand the size of the dataset in the future.

In our experiments, we adopt a uniform experimentation and parameter optimization framework to compare seven classic machine learning algorithms in terms of their performance in classifying an image into either CIN1/Normal (i.e. low-grade lesion/healthy) or CIN2/3+ (i.e. high-grade lesion/cancer). The reported results can serve as a baseline for future comparisons of automated cervical dysplasia classification methods. From the results, we find that ensemble-tree models—Random Forest, Gradient Boosting Decision Tree, and AdaBoost—outperform other classifiers such as multi-layer perceptron, SVM, logistic regression and kNN, on this task. This finding is consistent with the conclusion in other works [19]. Another finding is that, training and testing on the larger imbalanced dataset (containing more negative samples) give similar overall performance (measured by AUC and accuracy) to that on the balanced dataset (with equal

[1] Download from http://www.cse.lehigh.edu/~idealab/cervitor

number of negative and positive samples). However, the results on the imbalanced dataset have higher specificity than sensitivity whereas the results on the balanced dataset have higher sensitivity.

References

1. WHO: Human papillomavirus and related cancers in world. In: ICO Information Centre on HPV and Cancer Summary Report, August 2014
2. Kim, E., Huang, X.: A data driven approach to cervigram image analysis and classification. In: Color Medical Image analysis, Lecture Notes in Computational Vision and Biomechanics, vol. 6, pp. 1–13 (2013)
3. Biscotti, C.V., Dawson, A.E., et al.: Assisted primary screening using the automated thinprep imaging system. AJCP **123**(2), 281–287 (2005)
4. Wilbur, D.C., Black-Schaffer, W.S., Luff, R.D., et al.: The becton dickinson focalpoint gs imaging system: Clinical trials demonstrate significantly improved sensitivity for the detection of important cervical lesions. AJCP **132**(5), 767–775 (2009)
5. Zhang, J., Liu, Y.: Cervical cancer detection using SVM based feature screening. In: Barillot, C., Haynor, D.R., Hellier, P. (eds.) MICCAI 2004. LNCS, vol. 3217, pp. 873–880. Springer, Heidelberg (2004)
6. Herrero, R., Schiffman, M., Bratti, C., et al.: Design and methods of a population-based natural history study of cervical neoplasia in a rural province of costa rica: the guanacaste project. Rev. Panam. Salud Publica **1**, 362–375 (1997)
7. Jeronimo, J., Long, L.R., Neve, L., et al.: Digital tools for collecting data from cervigrams for research and training in colposcopy. Journal of Lower Genital Tract Disease **10**(1), 16–25 (2006)
8. Xu, T., Kim, E., Huang, X.: Adjustable adaboost classifier and pyramid features for image-based cervical cancer diagnosis. In: International Symposium on Biomedical Imaging (ISBI) (2015)
9. Morra, J.H., Tu, Z., Apostolova, L.G., et al.: Comparison of adaboost and support vector machines for detecting alzheimer's disease through automated hippocampal segmentation. Medical Imaging **29**, 30–43 (2010)
10. Osareh, A., Mirmehdi, M., Thomas, B., Markham, R.: Comparative exudate classification using support vector machines and neural networks. In: Dohi, T., Kikinis, R. (eds.) MICCAI 2002, Part II. LNCS, vol. 2489, pp. 413–420. Springer, Heidelberg (2002)
11. Wei, L., Yang, Y., Nishikawa, R.M., Jiang, Y.: A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. Medical Imaging **24**, 371–380 (2005)
12. Timoner, S.J., Golland, P., Kikinis, R., Shenton, M.E., Grimson, W.E.L., Wells III, W.M.: Performance issues in shape classification. In: Dohi, T., Kikinis, R. (eds.) MICCAI 2002, Part I. LNCS, vol. 2488, pp. 355–362. Springer, Heidelberg (2002)
13. Alexander, D.C., Zikic, D., Zhang, J., Zhang, H., Criminisi, A.: Image quality transfer via random forest regression: applications in diffusion MRI. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part III. LNCS, vol. 8675, pp. 225–232. Springer, Heidelberg (2014)
14. Hastie, T., Tibshirani, R., Friedman, J., et al.: The elements of statistical learning, vol. 2. Springer (2009)

15. Appel, R., Fuchs, T., Dollr, P., Perona, P.: Quickly boosting decision trees pruning underachieving features early. In: ICML (2013)
16. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
17. Goodfellow, I.J., Warde-Farley, D., Lamblin, P., et al.: Pylearn2: a machine learning research library (2013). arXiv:1308.4214
18. Chang, C., Lin, C.: LIBSVM: a library for support vector machines (2001)
19. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems? The Journal of Machine Learning Research **15**(1), 3133–3181 (2014)