Multi-Test Cervical Cancer Diagnosis with Missing Data Estimation

Tao Xu*[a], Xiaolei Huang[a], Edward Kim[b], L. Rodney Long[c], Sameer Antani[c]

[a]Department of Computer Science & Engineering, Lehigh University, Bethlehem, PA 18015, USA;
[b]Department of Computing Sciences, Villanova University, Villanova, PA 19085, USA;
[c]Communications Engineering Branch, National Library of Medicine, Bethesda, MD 20894, USA

ABSTRACT

Cervical cancer is a leading most common type of cancer for women worldwide. Existing screening programs for cervical cancer suffer from low sensitivity. Using images of the cervix (cervigrams) as an aid in detecting pre-cancerous changes to the cervix has good potential to improve sensitivity and help reduce the number of cervical cancer cases. In this paper, we present a method that utilizes multi-modality information extracted from multiple tests of a patient's visit to classify the patient visit to be either low-risk or high-risk. Our algorithm integrates image features and text features to make a diagnosis. We also present two strategies to estimate the missing values in text features: Image Classifier Supervised Mean Imputation (ICSMI) and Image Classifier Supervised Linear Interpolation (ICSLI). We evaluate our method on a large medical dataset and compare it with several alternative approaches. The results show that the proposed method with ICSLI strategy achieves the best result of 83.03% specificity and 76.36% sensitivity. When higher specificity is desired, our method can achieve 90% specificity with 62.12% sensitivity.

Keywords: Cervical dysplasia, automated screening, disease classification, missing data estimation

1. INTRODUCTION

Cervical cancer ranks as the second most common type of cancer in women aged 15 to 44 years worldwide [1]. The combination of screening and diagnostic procedures has led to the sharp decline of cervical cancer death rates. Screening helps prevent cervical cancer by detecting cervical intraepithelial neoplasia (CIN), which is classified in grades: CIN1 (mild), CIN2 (moderate) and CIN3 (severe). The disease grading is the basis for follow-up treatment and management of patients. In clinical practice, one of the most important goals of screening is to distinguish CIN1 from CIN2-3 or cancer (denoted as CIN2/3+), because mild dysplasia in CIN1 only needs to be observed but lesions in CIN2/3+ require treatment.

The most widely used cervical cancer screening and diagnostic methods include Pap tests, HPV tests, colposcopy, and digital cervicography [2]. Pap tests are effective, but suffer from low sensitivity in detecting CIN 2/3+ [9]. Moreover, Pap tests need the laboratory infrastructure and trained personnel to evaluate the samples. The sensitivity of HPV tests in detecting CIN 2/3+ lesions varies greatly [9]. Also, these tests have low specificity, particularly in younger women [2]. Colposcopy is a diagnostic procedure that also often involves getting a biopsy. Digital Cervicography is a low-cost and non-invasive visual examination method which includes taking a photograph of the cervix (called a cervigram) after the application of 5% acetic acid to the cervix epithelium. Interpretations based on cervigrams have been shown to effectively increase the sensitivity in conjunction with Pap tests in detecting invasive cancer [10] and high-grade (CIN2-3) lesions [11].

To investigate the potential of computer-assisted interpretation of cervigrams for early detection of cervical cancer, an automated algorithm was developed [3] to classify cervigrams into CIN1/normal or CIN2/3+. It was reported to achieve specificity of 76% and sensitivity of 75%. However, for cervicography coupled with computer-aided interpretation to be useful in clinical practice, it is important to further develop methodologies that can achieve specificity above 90% and sensitivity as high as possible. Moreover, often multiple tests are administered on a patient during her screening visit. So it would be useful to develop methods that can integrate multi-modality information from these multiple tests to achieve high sensitivity and specificity in cervical disease classification.

In this paper, we denote the multi-modality information that can be extracted from multiple tests conducted on a patient's visit as either image features or text features. We extract image features from cervigrams, and obtain text features from textual/numeric results in the patient's clinical records. We present an algorithm that integrates these image and text features to perform CIN classification for a patient's visit (i.e., differentiate CIN1/normal from CIN2/3+). Our CIN classification algorithm requires training and applying image-SVM and text-SVM classifiers which are trained on image features and text features, respectively. In our dataset, every patient's visit has at least one cervigram, but not every visit has the complete set of test results, such as Pap and HPV. Thus we face the challenge of estimating missing data in the text features. Mean/mode imputation (MMI), hot deck imputation (HDI), k-nearest neighbor imputation (KNN) and multiple imputations (MI) are frequently used for missing data estimation [7, 8]. Based on HDI, we propose two strategies to predict values for the missing data: Image Classifier Supervised Mean Imputation (ICSMI) and Image Classifier Supervised Linear Interpolation (ICSLI). We have conducted extensive experiments on a large medical dataset to evaluate our method. The results show that our CIN classification method coupled with the ICSLI strategy achieves the best result of 83.03% specificity and 76.36% sensitivity. It significantly outperforms the baseline classifier [3] which achieves the specificity of 73.48% and the sensitivity of 71.52%. When a higher specificity level is desired, our method can guarantee specificity of about 90% with 62.12% sensitivity.

2. DATA AND MATERIALS

We carry out our study on a large dataset collected by the National Cancer Institute (NCI) in the Guanacaste project [4]. To acquire this dataset of cervigrams and other clinical information of 10,000 anonymized women, we used the Multimedia Database Tool (MDT) created by the U.S. National Library of Medicine (NLM) [12]. Since the Guanacaste project is a population-based study of cervical neoplasia, a large number of women who were screened were healthy. Only 345 patient visits were diagnosed to be CIN2/3+ based on the Worst Histology of each visit (Multiple expert histology interpretations were done on each biopsy; the most severe interpretation is labeled "worst histology" for that visit in the database.) In our experiments, we construct a dataset that consists of all the 345 CIN2/3+ visits (positive samples) and 345 randomly selected CIN1/normal visits (negative samples). The information available from each visit includes cytology results (i.e. Pap test results), HPV test results, Digital Cervicography results (such as multiple cervigrams), the pH test result, colposcopy test results, and the patient age, among fields collected. We use the Worst Histology result as the ground truth for each visit. The ground truth information is only used for algorithm performance evaluation and is not used as a feature in training classifiers and algorithm development. Figure 1 illustrates the hierarchical representation of a patient's visit information.

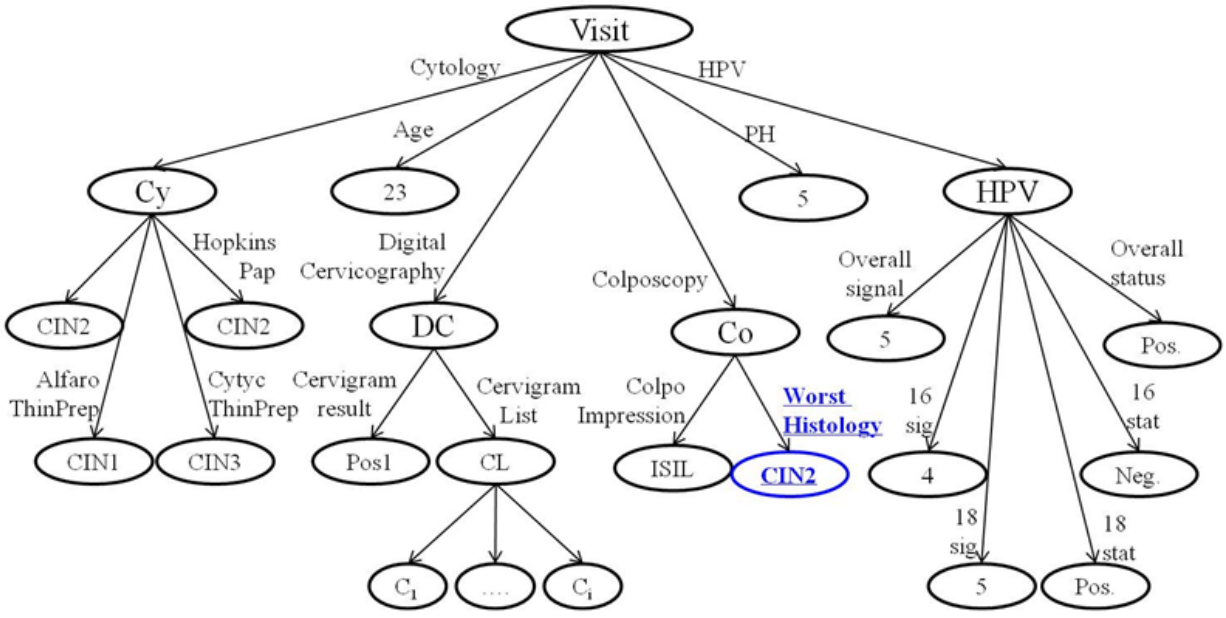


Figure 1. Hierarchical representation of data in an example patient visit.

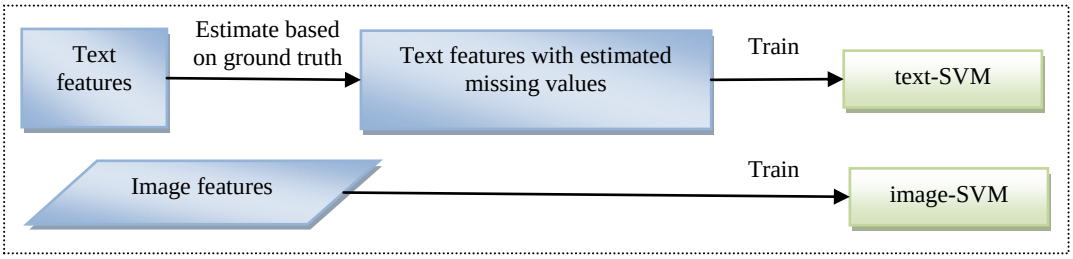
3. METHODOLOGY

Our proposed CIN classification algorithm utilizes multi-modality information to classify a patient visit to be either CIN1/Normal or CIN2/3+. Figure 2 shows an overview of our approach, which consists of a training module and a testing module.

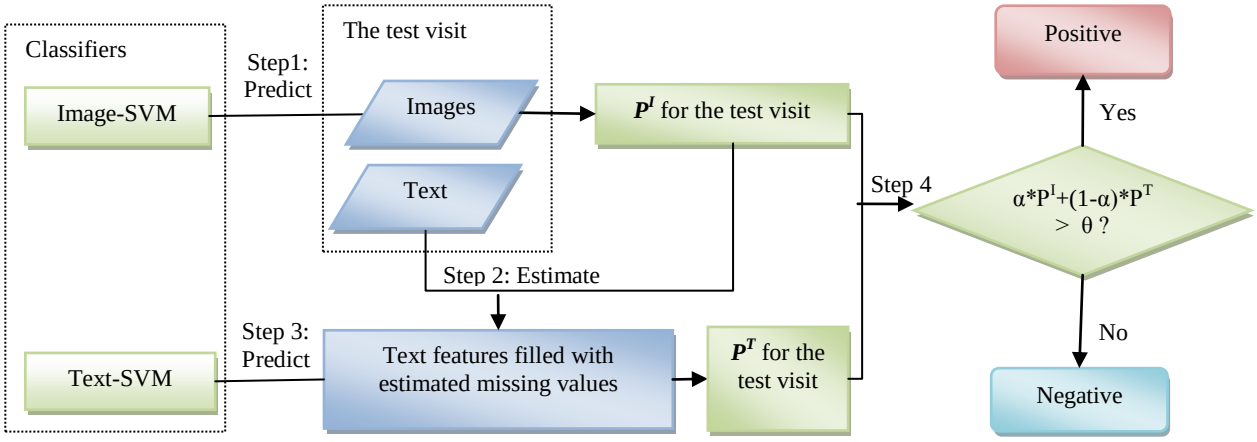
Training module: As Figure 2 (a) shows, we train an image-SVM classifier based on image features, and also train a text-SVM classifier based on text features with estimated values. The missing data in text features for a training sample are estimated based on the expected values for those features in the class that the sample belongs to, according to the ground truth label (i.e. worst histology result) of the sample.

Testing module: As illustrated in Figure 2 (b), we first use the image-SVM classifier to predict the probability of each test visit to be positive (CIN2/3+), indicated by $\boldsymbol{P^I}$. Second, based on $\boldsymbol{P^I}$, any missing data in text features are estimated and filled in adopting the ICSMI or ICSLI strategy. Third, based on text features with estimated values, we apply the text-SVM classifier to predict the probability of the test visit to be positive, denoted by $\boldsymbol{P^T}$. Finally, a proper threshold is found on the weighted sum of $\boldsymbol{P^I}$ and $\boldsymbol{P^T}$, to decide the test visit's category.

Next, we will describe details in the algorithm.



(a). The process of training Image-SVM and text-SVM classifiers



(b). The process of testing a visit by image-SVM and text-SVM classifiers.

Figure 2. Overview of the proposed algorithm (refer to text for more information)

3.1. Feature Extraction

3.1.1 Image Features

Interpretations based on cervigrams have been shown to effectively increase the sensitivity in detecting CIN2/3 when used in conjunction with other screening tests such as Pap tests [10, 11]. Some of the most important visual observations in cervigrams include the acetowhite region, and features within that region, such as mosaicism, punctation, and atypical vessels; it is also important to distinguish these possibly disease-related features from benign features such as polyps or cysts. Figure 3 shows some example images of those observations [14]. The identification of these different characteristics within a cervigram could help with diagnosis.

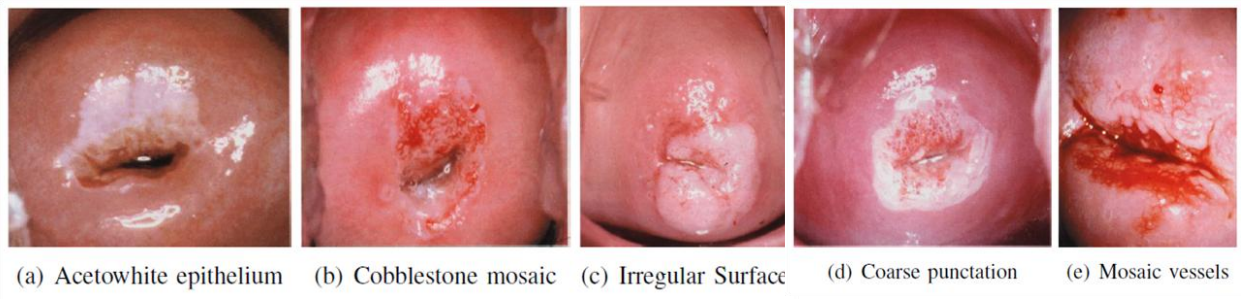


Figure 3. Representative cervigrams with different visual features

Previous works have attempted to develop computer algorithms to recognize these visual features. A common process is the detection of the region of interest (i.e., cervix region) either by color features and EM (Expectation Maximization) [5], or by GMM (Gaussian-Mixtures Model) [15, 16, 17]. After detecting the cervix Region of Interest (ROI), further image classification tasks are performed. In [15], color features and a GMM are used to classify different cervix tissue regions. They conclude that color features alone are not sufficient for cervigram image analysis, and texture features should be explored. In [18], the authors use texture features to recognize important vascular patterns found in cervix images. Similarly, [19] uses a filter bank of texture models for recognizing punctation and mosaicism. In the work by Kim et al. [3], the authors use the pyramid color histogram in L*A*B space (PLAB) and the pyramid histogram of oriented gradients (PHOG) features to encode both color and gradient information in cervigrams. Their results illustrate that the combination of color and gradient features produces a robust feature descriptor for cervigram interpretation. Thus, in this work we will use PLAB and PHOG [3] as our image features.

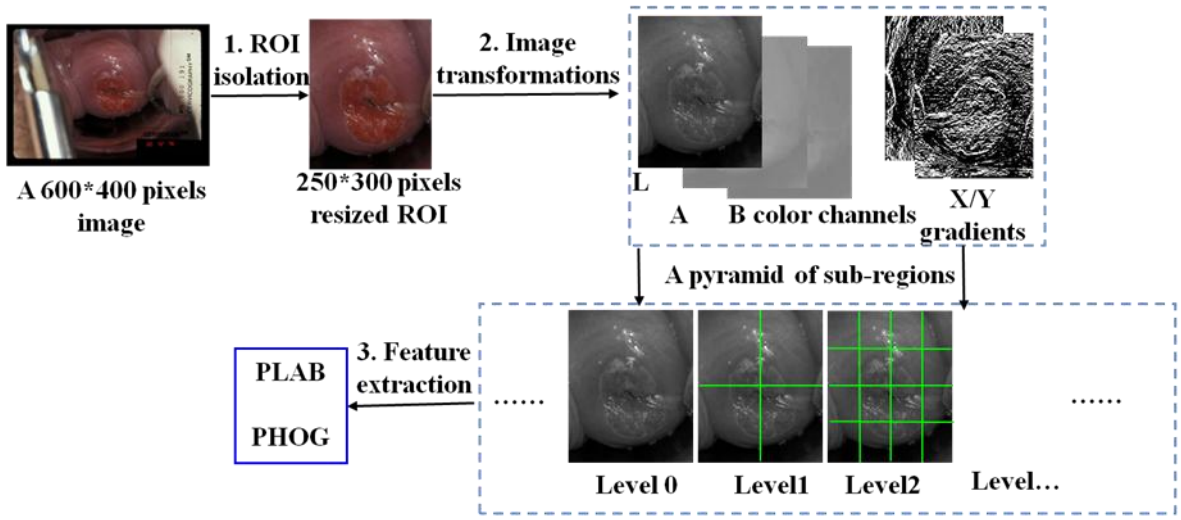


Figure 4. Image feature extraction

Figure 4 shows the process of extracting PLAB and PHOG image features. First we isolate the region of interest in cervigrams (ROIs). Then we construct a spatial pyramid for each ROI. We can construct the pyramid by splitting the ROI image patch into rectangular sub-regions, increasing the number of regions at each level; i.e., level 0 has 1 sub-region; level 1 has 4 sub-regions; level 2 has 16 sub-regions, and so forth. Histogram features can be extracted within these pyramid sub-regions. The extracted pyramid histogram encodes the statistical feature representation of abnormal characteristics in cervigrams at different positions and scales. For each channel in L*A*B color space, we extract 3 pyramid levels with a 16-bin histogram from each sub-region. Thus, the PLAB color feature has 1008 dimensions. Based on the gradient maps of the ROI, we calculate the pyramid histogram of oriented gradients (PHOG). An 8-bin orientation histogram over 4 levels is used. Hence, the total vector size of the PHOG feature is 680.

3.1.2 Text Features

For each patient visit, our dataset has 14 types of useful non-image clinical information, which consists of the results of four Pap tests (Alfaro ThinPrep, Cytyc ThinPrep, Costa Rica Pap, and Hopkins Pap), six HPV tests (overall HPV status, overall HPV signal, HPV16 status, HPV16 signal, HPV18 status, HPV18 signal), pH test, colposcopy impression, cervigram result, and age. Among those 14 types of information, five have numeric values (including three HPV signals, pH value and age) and others have string values (e.g. Pap results). The string values are indexed with integers indicating disease grades in the NLM/NLM database. Therefore, we use the integer indexes of string values and the numeric values to construct our 14 dimensional text feature, some of which have missing values.

3.2. Class Probability based on Image-SVM Classifier

Similar to the method used in [3], we train a linear Support Vector Machine (SVM) classifier on PLAB and PHOG image features [13]. We also use the SVM classifier to perform image-level classification. However, we take the process beyond image classification to the patient visit level. The main improvements are (1) all images obtained in one visit are used in our image-SVM classifier; (2) the class probability from image-SVM is also used to supervise missing data estimation for text features.

We denote images obtained during one visit as $\boldsymbol{I_j, j=1, 2, \ldots, n}$, where $\boldsymbol{n}$ is the number of images for the visit. For each image $\boldsymbol{I_j}$, we use the image-SVM to predict the probability of the images indicating a positive case, termed $\boldsymbol{p_j}$ ($\boldsymbol{p_j \in [0, 1]}$, where $\boldsymbol{0}$ represents negative (CIN1/normal) and $\boldsymbol{1}$ represents positive (CIN2/3+)). The probability of the visit being in the positive class based on image information ($\boldsymbol{P^I}$) is calculated by averaging the probabilities given by all of the $\boldsymbol{n}$ images of the visit. In our database, there are at most two cervigrams on each visit, i.e., $n \leq 2$.

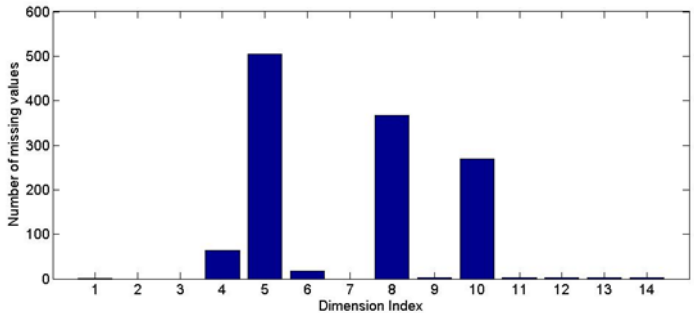


Figure 5. Number of patient visits (Y-axis) that do not have the value for each dimension of the 14D text feature (X-axis). (The meaning of each dimension of the text features is, 1: Age, 2: PH, 3: Cervigram Result, 4: Colpo Impression, 5: Alfaro ThinPrep; 6: Cytyc ThinPrep, 7: Costa Rica Pap, 8: Hopkins Pap, 9: Overall HPV status, 10: Overall HPV Signal, 11: HPV16 Result, 12: HPV16 Signal, 13: HPV18 Result, 14: HPV18 Signal)

3.3. Image Classifier Supervised Missing Data Estimation

In our dataset, every patient's visit has at least one cervigram, but not every visit has the complete set of test results, such as Pap and HPV. Thus we face the challenge of estimating missing data in text features. Figure 5 shows the number of visits that have missing data in each dimension of the 14D text feature.

To handle the missing data problem, we first respectively compute the average text feature of the negative and positive training datasets using available data; we denote these average features by $\boldsymbol{A = (a_1, a_2, \ldots, a_{14})}$ for the negative dataset and $\boldsymbol{B = (b_1, b_2, \ldots, b_{14})}$ for the positive dataset. Then, for any text feature $\boldsymbol{F}$ in the training dataset, if there is any missing data, they can be filled using values in set $\boldsymbol{A}$ or set $\boldsymbol{B}$ based on the ground truth of the corresponding visit; that is, if the visit is CIN2/3+, the missing data in $\boldsymbol{F}$ will be filled with corresponding value from set $\boldsymbol{B}$, and if the visit is CIN1/Normal, the missing data in $\boldsymbol{F}$ will be filled with corresponding value from set $\boldsymbol{A}$.

For testing, since we do not know the ground truth for a test patient's visit, we apply either the ICSMI or ICSLI strategy to estimate the values of missing test results in the text feature. More specifically, ICSMI and ICSLI are two imputation strategies that can be used to estimate missing values in the text feature, based on the outcome of the image-SVM classifier, $\boldsymbol{P^I}$.

Image classifier supervised mean imputation (ICSMI): any missing value $\boldsymbol{f_i}$ in $\boldsymbol{F}$ is estimated by either $f_i = a_i$, *if* $P^I \leq 0.5$, or $f_i = b_i$, *if* $P^I > 0.5$, where a_i, b_i are respectively taken from the sets A, B of average feature values described above

Image classifier supervised linear interpolation (ICSLI): any missing value $\boldsymbol{f_i}$ in $\boldsymbol{F}$ is estimated by Equation (1),

$$f_i = a_i + P^I * (b_i - a_i) \qquad (1)$$

3.4. CIN Classification by Integrating Image and Text information

Text features with estimated missing values are used as input for the text-SVM classifier prediction. The output of the text-SVM is the text based probability of a visit being in the positive class, termed P^T. Then, the final probability of a test visit having positive outcome, P, is computed by Equation (2),

$$P = \alpha * P^I + (1-\alpha) * P^T \qquad (2)$$

Where $\alpha \in [0,1]$ is a weight parameter; $\boldsymbol{P^I}$ and $\boldsymbol{P^T}$ are the respective image- and text-based probabilities of the visit being in the positive class.

The final class label of the test visit is given by Equation (3),

$$L(P) = \begin{cases} 1, & if\ P > \theta \\ 0, & else \end{cases} \qquad (3)$$

Where $\boldsymbol{0}$ represents negative (CIN1/normal) and $\boldsymbol{1}$ represents positive (CIN2/3+). $\boldsymbol{\theta}$ is a threshold on the final probability of the test visit having positive outcome.

4. EXPERIMENTS

We performed a ten-round ten-fold cross validation on the dataset to evaluate the performance of our algorithm for CIN classification. We randomly divided the visits into ten folds; in a rotational manner, we used one fold for testing and the remaining folds for training in each round. The final testing result is the average result of the ten rounds. We set the threshold $\boldsymbol{\theta = 0.5}$ and the weight parameter for combining image-based probability and text-based probability $\alpha = 0$. We trained and tested the method proposed in paper [3] on our dataset as the baseline for comparison.

Imputation for missing data treatment. Imputation aims to fill in missing data values with estimates [7]. We compared different imputation strategies for estimating the missing data in text features, specifically: (1) our proposed Image Classifier Supervised Linear Interpolation (ICSLI), (2) our proposed Image Classifier Supervised Mean Imputation (ICSMI), (3) Multiple Imputation (MI) [8], in which the missing data is filled in by making a random choice between picking the value from A or from B. The result is the average of 10 rounds, (4) Mean/Mode Imputation (MMI) [7], in which the missing data is filled in with the average of values from set ***A*** and set ***B***, and (5) K-Nearest Neighbor imputation (KNN) [8], in which the missing data is filled in with the average of values in k (k=10) closest neighbors. Figure 6 shows the performance comparison of these strategies using ROC curves. The comparison is done based on classification results from the text-SVM classifier only, by setting α=0. The Area Under Curve (AUC) of the proposed strategy ICSLI is 0.8325 while those of ICSMI, KNN, MI, and MMI are 0.8271, 0.7971, 0.7502, and 0.3282, respectively. Bigger AUC means better performance. Thus, it is clear that our proposed method, ICSLI, outperforms all other strategies.

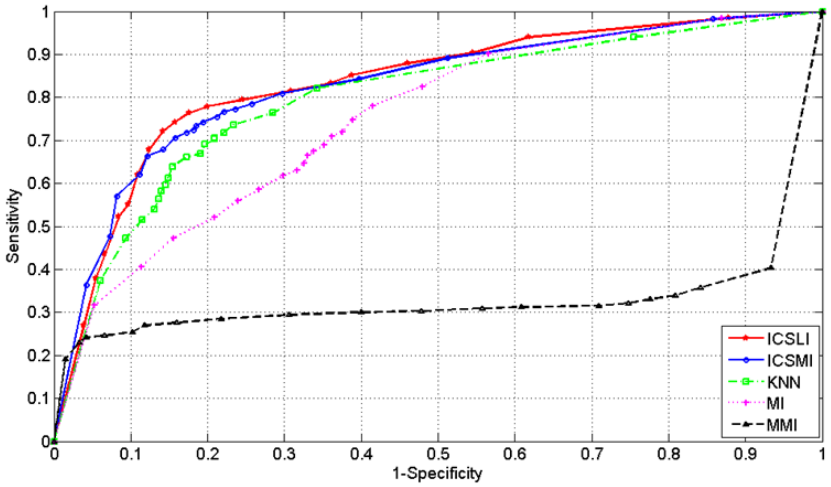


Figure 6. Comparison of imputation strategies (Text-SVM classifier only). ROC curves are plotted by changing θ.

Table 1. Results of our CIN classification algorithm when $\alpha = [0, 1]$ and $\theta = 0.5$

Weight	0.0	**0.1**	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Accuracy (%)	79.39	**79.70**	79.55	79.39	79.39	78.33	78.18	77.88	75.91	75.00	73.18
Specificity (%)	82.42	**83.03**	83.03	82.73	82.73	81.82	79.09	78.18	75.45	74.24	72.12
Sensitivity (%)	76.36	**76.36**	76.06	76.06	76.06	74.85	77.27	77.58	76.36	75.76	74.24

Image vs. text information for visit level CIN classification. We evaluate the relative influence of image and text features on the performance of our CIN classification algorithm by altering the weight parameter ***α*** in Equation (2). Table 1 summarizes the results. By using the text feature alone (***α=0***), it is still able to achieve fairly good specificity of 82.42% with 76.36% sensitivity. In the case that only image feature is used (***α=1***), specificity and sensitivity drop significantly to 72.12% and 74.24%. Thus we can conclude that the text feature plays a vital role in our algorithm. Furthermore, it demonstrates that our ICSLI strategy is an efficient way to estimate missing values in text features, and produces robust text features.

To further put our results in perspective, we compare our method with the baseline image-only classifier [3] whose specificity is 73.48% and sensitivity is 71.52%. Our method improves the specificity to 83.03% and improves the sensitivity to 77.36% when α=0.1, θ=0.5 and ICSLI strategy are used. If we change θ to 0.25, our algorithm achieves a specificity of 89.09% with a sensitivity of 62.12%. Thus, our method outperforms the image-only classifier significantly by integrating information in the text feature.

5. CONCLUSION

We propose a cervical CIN grade classification algorithm by using both image and text features to classify a patient visit. We propose two image classifier supervised imputation strategies (ICSMI and ICSLI) for estimating and filling in missing data in text features. We demonstrate that using multi-modality information can significantly improve the classification performance compared with an image-only classifier [3].

6. ACKNOWLEDGEMENTS

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC). We gratefully acknowledge the past help of Dr. Mark Schiffman and Dr. Nicolas Wentzensen of the National Cancer Institute (NCI) in obtaining access to and interpretation of the NCI ALTS and Guanacaste datasets.

REFERENCES

[1]. WHO/ICO Information Centre on HPV and Cervical Cancer, "Human papillomavirus and related cancers in the world," in Summary report, (2014).

[2]. Hartmann, K., Hall, S., Nanda, K., and et al., "Screening for cervical cancer: Systematic evidence review," in Agency for Healthcare Research and Quality: United States Preventive Service Task Force, (2002).

[3]. Kim, E., and 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, 1-13, (2013).

[4]. Herrero, R., and 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, Vol. 1, 362-375, (1997).

[5]. Li, W., Gu, J., Ferris, D., and Poirson, A., "Automated image analysis of uterine cervical images," in Proc. of SPIE Medical Imaging, (2007).

[6]. Wilbur, D.C., Black-Schaffer, W.S., Lu, R.D., and et al., "The Becton Dickinson focal point GS imaging system: Clinical trials demonstrate significantly improved sensitivity for the detection of important cervical lesions," American Journal of Clinical Pathology, 132(5), 767-775, (2009).

[7]. Batista, G.E.A.P.A., and Monard, M.C., "An analysis of four missing data treatment methods for supervised learning," Applied Artificial Intelligence, 17(5-6), 519-533, (2003).

[8]. Rubin, D.B., "An overview of multiple imputations," in Survey Research Section, American Statistical Association, 79-84, (1988).

[9]. Sankaranarayanan, R., Ga kin, L., Jacob, M., and et al., "A critical assessment of screening methods for cervical neoplasia," International Journal of Gynecology and Obstetrics, 89, 4-12, (2005).

[10]. Ferris, D., Schi man, M., and M.S., L., "Cervicography for triage of women with mildly abnormal cervical cytology results," American Journal of Obstetrics and Gynecology, 185(4), 939-943, (2001).

[11]. Eskridge, C., Begneaud, W., and Landwehr, C., "Cervicography combined with repeated papanicolaou test as triage for low-grade cytologic abnormalities," Obstetrics and Gynecology, 92(3), 351-355, (1998).

[12]. Jeronimo, J., Long, L.R., Neve, L., and 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).

[13]. Chang, C., and Lin, C., "LIBSVM: a library for support vector machines," (2001).

[14]. Song, D., Kim, E., Huang, X., and et al., "Multi-modal Entity Coreference for Cervical Dysplasia Diagnosis," IEEE Trans. Medical Imaging, (2014).

[15] Gordon, S., Zimmerman, G., and Greenspan, H., "Image segmentation of uterine cervix images for indexing in pacs," in Symposium on Computer-Based Medical Systems, pp. 298–303, (2004).

[16] Zimmerman-Moreno, G., and Greenspan, H., "Automatic detection of specular reflections in uterine cervix images," in Proc. of SPIE Medical Imaging, (2006).

[17] Xue, Z., Antani, S., Long, R., and Thoma, G., "Comparative performance analysis of cervix ROI extraction and specular reflection removal algorithms for uterine cervix image analysis," in Proc SPIE, (2007).

[18] Ji, Q., Engel, J., and Craine, E., "Classifying cervix tissue patterns with texture analysis," Pattern Recognition, vol. 33, no. 9, pp. 1561–1574, (2000).

[19] Srinivasan, Y., Nutter, B., Mitra, S., Phillips, B., and Sinzinger, E., "Classification of cervix lesions using filter bank-based texture mode," in Symposium on Computer-Based Medical Systems, pp. 832–840, (2006).