Multimodal Entity Coreference for Cervical Dysplasia Diagnosis

Dezhao Song*, Edward Kim, Xiaolei Huang, Joseph Patruno, Héctor Muñoz-Avila, Jeff Heflin, L. Rodney Long, and Sameer Antani

***Abstract*—Cervical cancer is the second most common type of cancer for women. Existing screening programs for cervical cancer, such as Pap Smear, suffer from low sensitivity. Thus, many patients who are ill are not detected in the screening process. Using images of the cervix as an aid in cervical cancer screening has the potential to greatly improve sensitivity, and can be especially useful in resource-poor regions of the world. In this paper, we develop a data-driven computer algorithm for interpreting cervical images based on color and texture. We are able to obtain 74% sensitivity and 90% specificity when differentiating high-grade cervical lesions from low-grade lesions and normal tissue. On the same dataset, using Pap tests alone yields a sensitivity of 37% and specificity of 96%, and using HPV test alone gives a 57% sensitivity and 93% specificity. Furthermore, we develop a comprehensive algorithmic framework based on Multimodal Entity Coreference for combining various tests to perform disease classification and diagnosis. When integrating multiple tests, we adopt information gain and gradient-based approaches for learning the relative weights of different tests. In our evaluation, we present a novel algorithm that integrates cervical images, Pap, HPV, and patient age, which yields 83.21% sensitivity and 94.79% specificity, a statistically significant improvement over using any single source of information alone.**

***Index Terms*—Cervical dysplasia, cervical image analysis, disease classification, entity coreference, patient case retrieval.**

I. Background and Significance

THE COMBINATION of screening and diagnostic procedures has led to the sharp decline of cervical cancer death rates in Western countries. However, in areas that lack laboratories and trained personnel for conducting screening, diagnostic, and follow-up tests, cervical cancer is still one of the leading causes of death in middle-aged women. In 2008, an estimated 275 100 women died from cervical cancer, and nearly 90% of the deaths occurred in developing parts of the world [1]. Consequently, there is a need for less expensive and more automated screening methods [2], [3], especially those applicable in low-resource regions. Digital Cervicography is a noninvasive visual examination method performed by taking a photograph of the cervix (called a cervigram) after the application of 5% acetic acid to the cervix epithelium. It has a low cost and is widely accessible in resource-poor regions. However, in the literature, 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. Encouraged by recent developments in computer-assisted diagnosis such as automated Pap tests, in this paper, we investigate computer algorithms to improve the accuracy in early detection of cervical cancer using cervigrams and study the potential of further improvement by combining images with other clinical test results such as Pap and HPV.

A. Clinical Methods for Cervical Cancer Screening and Diagnosis

Cervical cancer afflicts an estimated 12 200 women in the U.S. [4] and 529 800 women worldwide [1] every year. It can be cured if detected during its early stages and treated appropriately. Screening can prevent cervical cancer by detecting Cervical Intraepithelial Neoplasia (CIN), also known as cervical dysplasia. CIN is classified in grades: CIN1 (mild), CIN2 (moderate), and CIN3 (severe). This disease grading is the basis for follow-up treatment and management of the patients. In clinical practice, one of the most important goals of screening is to differentiate CIN1 from CIN2/3 or cancer (denoted as CIN2/3+ in this paper), since those lesions in CIN2/3+ will require treatment, whereas mild dysplasia in CIN1 can be observed conservatively because it will typically be cleared by an immune response in a year or similar timeframe.

The Pap test is the most widely used cervical cancer screening method [5]. It involves collecting a small sample of cells from the cervix and examining it under a microscope for squamous and glandular intraepithelial lesions (SIL). The result of a Pap test can be either normal or abnormal. Pap tests are effective, but nevertheless require a laboratory infrastructure and trained personnel to evaluate the samples. Furthermore, it is well

Manuscript received May 26, 2014; revised July 27, 2014; accepted August 16, 2014. Date of publication August 27, 2014; date of current version December 24, 2014. This research was supported 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). *Asterisk indicates corresponding author.*

*D. Song was with the Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA 18015 USA. He is now with Research and Development, Thomson Reuters, Eagan, MN 55122 USA (e-mail: dezhao.song@thomsonreuters.com).

E. Kim is with the Department of Computing Sciences, Villanova University, Villanova, PA 19085 USA.

X. Huang, H. Muñoz-Avila, and J. Heflin are with the Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA 18015 USA (e-mail: xih206@lehigh.edu).

J. Patruno is with the Department of Obstetrics and Gynecology, Lehigh Valley Health Network, Allentown, PA 18105 USA.

L. R. Long and S. Antani are with the Communications Engineering Branch, National Library of Medicine, Bethesda, MD 20894 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2014.2352311

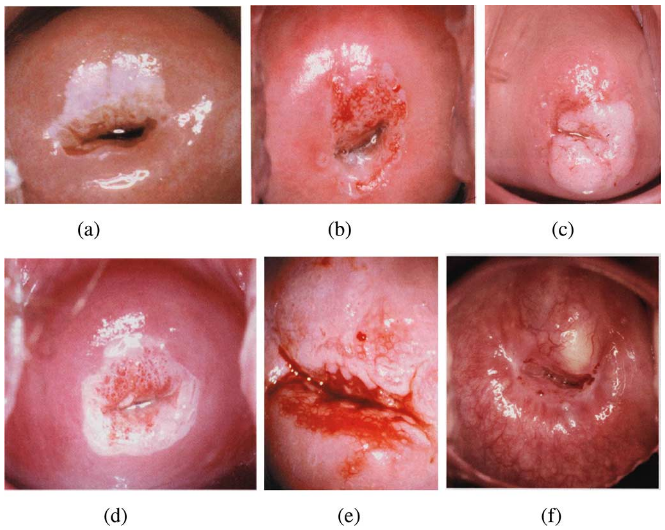

Fig. 1. Representative images of different visual features present in photographs of the cervix. (a) Acetowhite epithelium of the anterior lip of the cervix. Biopsy revealed CIN2. (b) Cobblestone appearance of mosaic in a high-grade lesion. (c) High-grade lesion with dense acetowhite epithelium and an irregular surface contour. (d) High-grade lesion with coarse punctation on the anterior lip of the cervix. (e) Very coarse, irregular mosaic vessels. (f) Large nabothian cyst and distinct fine normal, branching vessels of the cervical squamous epithelium. Images courtesy of [40]. (a) Acetowhite epithelium. (b) Cobblestone mosaic. (c) Irregular surface. (d) Coarse punctation. (e) Mosaic vessels. (f) Nabothian cyst.

known in the literature that Pap tests suffer from low sensitivity (20%–85%) in detecting CIN $2/3+$ [6]–[14].

The automated Pap test is an alternative to the conventional Pap test (Pap smear). According to some studies in [15]–[17], higher sensitivities have been achieved with automation-assisted Pap tests: 79%–82% by using the ThinPrep Imaging System [15] and 81%–86% by adopting the Becton Dickinson FocalPoint GS Imaging System [16]. Although studies regarding the significance of the difference between automated Pap test and conventional Pap test are inconclusive [18]–[22], automated Pap tests do provide several benefits. In terms of specimen adequacy, evidence indicated a lower proportion of unsatisfactory slides for automated Pap test than for conventional cytology, e.g., 0.4% versus 1.1% in The Netherlands ThinPrep Versus Conventional Cytology (NETHCON) trial [21] and 2.6% versus 4.1% in the New Technologies for Cervical Cancer Screening (NTCC) trial [23]. Also, in the case of an equivocal cytology result, the automated Pap test permits HPV testing without the need for another patient sample [21], [24].

The HPV test is another screening method that has been used in conjunction with the Pap test either as an additional test or when Pap test results are inconclusive. It has been well established that cervical dysplasia are caused by persistent infection with certain types of human papillomavirus (HPV), thus DNA tests to detect HPV strains associated with cervical cancer (i.e., HPV test) can be used for screening and triage of cervical abnormalities. The sensitivity of the HPV test in detecting CIN $2/3+$ lesions varied from 66% to 100% and the specificity varied from 61% to 96% [6]. However, the HPV test is not recommended as a primary screening method, because of its relatively high false positive rate, particularly in younger women [25].

An abnormal Pap test result may also lead to a recommendation for Colposcopy of the cervix, during which a doctor examines the cervix in detail through a magnifying device. If an area of abnormal tissue is seen, the doctor may decide to remove a small sample of tissue from that area (i.e., biopsy) and send it to a lab to be examined under a microscope. CIN can be diagnosed by biopsy. Being a diagnostic procedure and often accompanied by biopsy, Colposcopy is more costly than screening methods such as Pap and HPV tests.

Digital Cervicography [26]–[30] is another visual examination method; it takes a photograph of the cervix (called a cervigram) after applying 5% acetic acid to the cervix epithelium. In the literature, Cervicography has been shown to effectively increase the sensitivity of Pap test in detecting invasive cancer [31] and high-grade (CIN2–3) lesions in patients with previous atypical squamous cells of undetermined significance (ASCUS) or low-grade squamous intraepithelial lesion (LSIL) Pap result [32]. But questions remain regarding its overall effectiveness because studies find poor correlation between visual lesion recognition and disease [33] as well as disagreement among experts for grading visual findings [34].

B. Computer Methods for Cervigram Image Analysis and Interpretation

Because Digital Cervicography is noninvasive and has low cost, it has the potential to be a widely accessible automated screening method for cervical cancer in resource-poor regions. Being a photographic test, it can also support mobile applications. The goal of cervigram image analysis is to explore these possibilities and to develop a computer algorithm for cervical dysplasia diagnosis by automatically interpreting a cervigram. Some of the most important visual observations in cervigrams include the Acetowhite region, Mosaicism, Punctation, Atypical Vessel, polyps, cyst, etc.; see Fig. 1 for some example images of such observations. The identification of these different characteristics within a cervigram could help with diagnosis. 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 Expectation-Maximization (EM) [35], or by Gaussian-Mixtures Model (GMM) [27], [36], [37]. After detecting the cervix Region of Interest (ROI), further image classification tasks are performed. In [27], 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 [38], the authors use texture features to recognize important vascular patterns found in cervix images. Similarly, [39] uses a filter bank of texture models for recognizing punctation and mosaicism. In summary, most of these existing approaches attempt to characterize the different areas or tissue regions within a cervigram image. However, as these characteristic regions in cervigrams have high variability in color, texture, shape, and size, it is difficult to identify them with high accuracy and most of the existing algorithms do not scale well to large datasets.

C. Overview of Our Approach

The objective of this paper is to evaluate whether a photographic test (Cervicography) can be used as an adjunctive screening method to better detect high-grade $(\mathrm{CIN}2/3+)$ cervical lesions through computer-assisted image interpretation and to evaluate the effectiveness of a multimodal framework that integrates images and other textual/numeric clinical test results (including Pap, HPV) to perform disease classification.

To examine whether computer interpretation of cervigrams will lead to better diagnosis, we first develop a computer algorithm that interprets cervigrams based on color and texture. Our approach is different from previous approaches that explicitly attempt to segment or characterize the different tissue regions within a cervigram image. Given a new cervigram, we develop a data centric system that is able to find similar cervigrams in a collection of expertly labeled cervigrams. The expert labels, including the cervix region boundary as well as boundaries for abnormal tissue regions if any, from these similar cervigrams can assist in locating the cervix region of interest and classifying disease patterns in the new unlabeled cervigram. Furthermore, we develop an algorithmic framework based on multimodal Entity Coreference for combining various clinical tests (e.g., Pap, HPV, pH value) and image analysis results to perform cervical disease classification. Our framework enables the efficient evaluation of the performance of various combinatory tests.

II. Methodology

In this section, we first introduce the source and format of our experimental data and explain how we select patient cases for evaluation. We then present an overview of the proposed cervical disease classification framework. Next, we describe in detail each component of our system: how we compute data similarity based upon patient clinical test results, how we compute image similarity, and how we combine clinical data and image similarities for patient classification. At the end of this section, we present our method and metrics for experimental evaluation.

A. Data and Materials

We carry out our research on 60 000 digitized uterine cervix images collected by the National Cancer Institute (NCI) in a longitudinal multiyear study of the evolution of lesions related to cervical cancer. Through the NLM MDT (U.S. National Library of Medicine Multimedia Database Tool) [41], we can access these digital cervix images as well as clinical, cytologic, and molecular information at multiple examinations of 10 000 women who participated in NCI's Guanacaste study [42]. The women can be categorized as follows: patients with invasive cancer, patients without cervical lesion at enrollment but later developed disease at follow-up, and healthy women who never developed any pathological changes in the cervix. Some statistics about the dataset are shown in Table I. Fig. 2 shows sample images from the dataset that represent different cervical dysplasia grades; the resolution for these images is 2891 by 1973 pixels.

Since our goal is to study the potential of using cervigrams as an adjunct screening test, and also evaluate different combinations of multiple screening tests, we use data only from the Pap test (i.e., cytologic data), the HPV test, and images (i.e., cervigrams). We also consider patient age and pH value in some experiments. The "gold standard" ground truth against which we evaluate our disease classification method is histologic data obtained from microscopic evaluation of tissue samples taken during biopsy.

TABLE I
DATASET STATISTICS

Dataset	Category	Number of All Patients	Used for Experiments
Guanacaste	<CIN2	7669	140
	CIN2	62	60
	CIN3	70	70
	Cancer	10	10

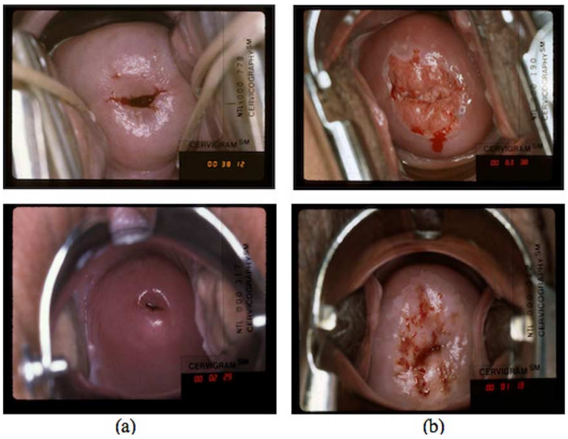

Fig. 2. Sample cervigram images from the Guanacaste dataset. Column (a) images of patients from normal or CIN1 category. Column (b) images of patients from $\mathrm{CIN}2/3+$ category.

We evaluate our proposed cervical disease classifiers on 280 randomly selected participants from this Guanacaste database. We had to choose an unbalanced number of patient cases for the four categories because only 10 cancer cases are available in the entire Guanacaste dataset. However, since we are performing a binary classification, i.e., classifying patient cases into one of the following two categories: $<\mathrm{CIN}2$ and $\mathrm{CIN}2/3+$, we do have an equal number of patient cases in these two classes: 140 cases in $<\mathrm{CIN}2$, and 140 cases in $\mathrm{CIN}2/3+$.

B. System Overview

Fig. 3 demonstrates the main components of our proposed system for cervical dysplasia disease classification.

1) Data Converter. The raw data is stored in a relational database and the data needs to be converted and represented into a hierarchical format required by our algorithm.
2) Similarity Calculator. Clinical cases are composed of multiple kinds of data including not only the cervigrams but nonimaging data such as other test results and patient history. Therefore, it is important to combine these multimodal data sources for reliable patient classification. In our system, the calculation of patient case similarity has two sub-components: the data-level similarity involving numeric and symbolic data such as Pap and HPV test results, and image similarity involving cervigrams. We calculate a final similarity between two patient cases by taking the linear combination of the two component similarities.

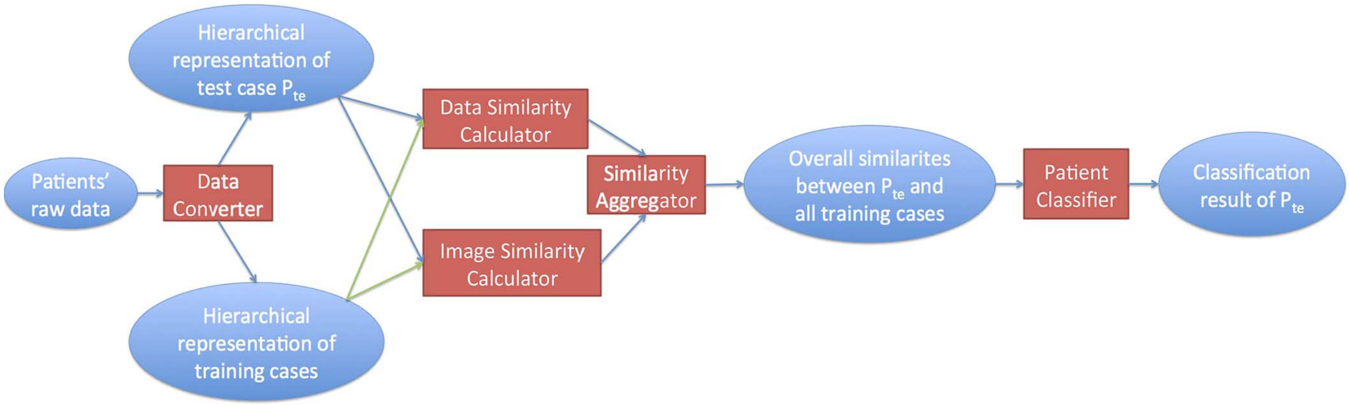
Hierarchical representation of test case P_te
Patients' raw data
Data Converter
Hierarchical representation of training cases
Data Similarity Calculator
Image Similarity Calculator
Similarity Aggregator
Overall similarites between P_te and all training cases
Patient Classifier
Classification result of P_te

Fig. 3. Overall architecture for proposed cervical disease classification system.

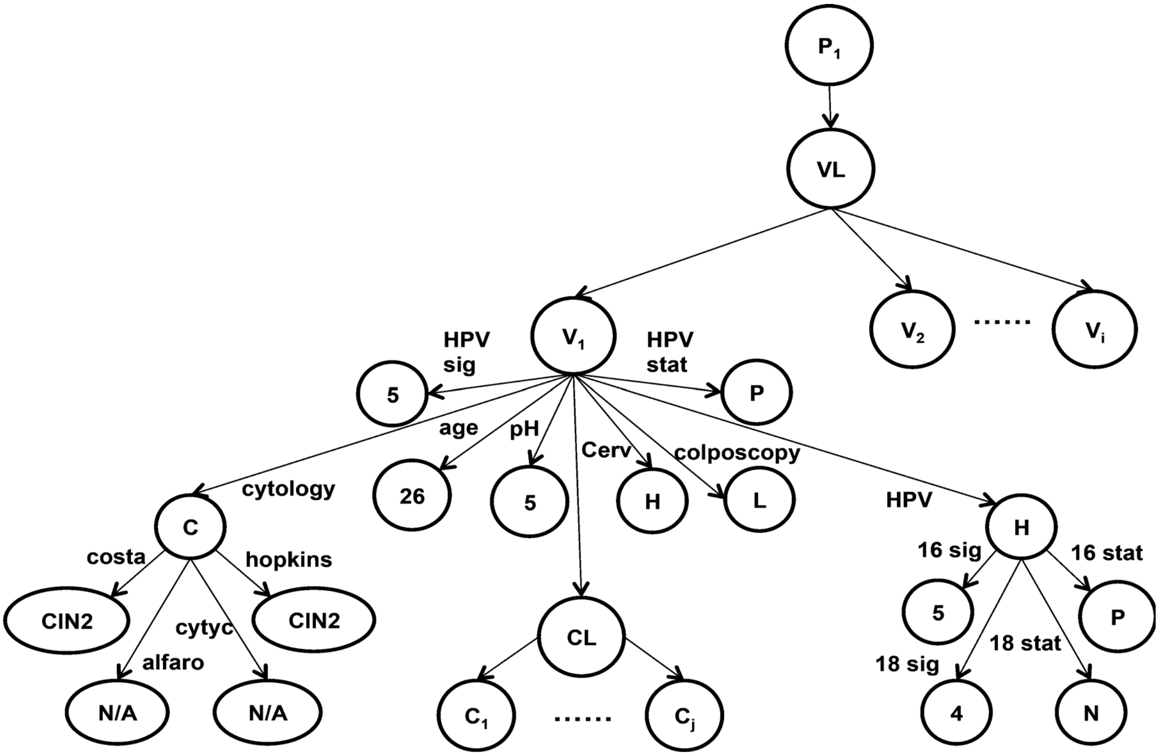

Fig. 4. Transformed hierarchical representation of patient data.

3) Patient Classifier. This is the multimodal aggregation scheme that translates patient case similarity to patient disease classification. We explored different aggregation methods, such as nearest neighbor (i.e., maximum aggregation) and majority voting. And we finally adopt an approach that retrieves the most similar cases from a training database and let the top-cluster training cases to vote to determine the disease grade of a test patient case.

The data structures and algorithms that we propose are motivated by the challenges presented in the data records. In the database, each patient may have data from multiple clinic visits, and the number of visits differs from patient to patient. Although there are cytology, HPV results and images of almost all patients, the types of cytology and HPV tests may differ, and there may be missing data for some visits. There is also significant variability in the images due to illumination, image acquisition setup, and specular reflection, among other factors. Thus, a key challenge to our system is that it has to be able to handle highly unbalanced data, and measure similarity between patients regardless of differences in number of visits and available information from each visit.

C. *Computing Data Level Similarity*

In this section, we introduce how we convert the raw data from relational database records to a hierarchical format and we then formally present our approach for measuring data-level patient similarity.

As shown in Fig. 4, Patient P_1 has several visits denoted by $VL = \{V_1, V_2, \ldots, V_i\}$. Take the first visit V_1 of patient P_1 as an example. V_1 stores some basic information about the patient such as age at this visit, and it is also associated with information from some simple tests such as pH value, cervigrams (Cerv), and colposcopy impression. Furthermore, V_1 has two complex clinical tests, C (Cytology) and H (HPV), which are further expanded to have some other simple test results (e.g., HPV16 and HPV18 are child nodes of H). We use simple and complex to refer to the type of the test result (whether it has subtest results or not), rather than the complexity of the procedure of the test itself. Finally, CL denotes a node that expands to a list of cervical images (i.e., cervigrams), $C_1, \ldots, C_j$, that were taken of the patient P_1 during visit V_1.

We adapt our entity coreference algorithm [43] (Algorithm 1) to compute the similarity of two patients by taking into account their clinical test results. In Algorithm 1, G is a function, retrieving the set of chains for a given patient; *comparable* checks the comparability of two chains; the method l is used to get the leaf node of a chain; and *Sim* computes the similarity between two leaf nodes.

The essential idea for comparing two patients is what we call a *bag-of-chains* approach that compares chains between patients, such as *patientA* and *patientB*. In the hierarchical representation of a patients data (Fig. 4), a chain is the path from the root to a leaf node. For each chain c of *patientA*, we compare its leaf node to the leaf node of every comparable chain of *patientB* and choose their average similarity score, denoted as *chain_score*. We need to repeat the process for every chain of *patientA*. The final data similarity for a pair of patients is the average of all chain similarity scores between them.

Algorithm 1 $\boldsymbol{Data\_Sim(G(a), G(b))}$**,** $\boldsymbol{a}$ **and** $\boldsymbol{b}$ **are two patients**

```
score ← 0;
count ← 0;
for $c \in G(a)$ do
  if $\exists c' \in G(b)$, $comparable(c, c')$ then
    //$t$ refers to the clinical test represented by $c$
    $chain\_score = Average_{c' \in G(b), comparable(c,c')} Sim(t, l(c), l(c'))$;
    $score \leftarrow score + chain\_score$;
    $count++$;
  end if
end for
if $count > 0$ then
  $score \leftarrow \frac{score}{count}$;
end if
return $score$
```

One key problem to solve is that, for a chain of *patientA*, we need to find all its comparable chains from *patientB*. In this work, we say two chains are comparable if they represent the same type of clinical test determined by the type of edges in the hierarchy. For example, we can have a chain of *patientA* with the following sequence of edges: $\text{Chain\_1} = (\text{has\_visit\_}list \rightarrow \text{has\_visit} \rightarrow \text{has\_}hpv \rightarrow \text{has\_hpv16})$. A comparable chain from *patientB* will need to have the same sequence of edges in the chain. Although only leaf nodes in the chains will be used for computing the data similarity, the edges in the tree structure are needed for determining which chains between two patients are comparable. For example, *Chain_1* of *patientA* as shown above is not comparable to the following Chain_2 of *patientB*, because their edges at depth levels 3 and 4 are indicating different clinical tests and thus not comparable: $\text{Chain\_2} = (\text{has\_visit\_list} \rightarrow \text{has\_visit} \rightarrow \text{has\_cytology} \rightarrow \text{has\_hopkins})$.

In datasets with unbalanced data, as is in our case, this approach has the advantage of measuring the similarity of one patient to another patient by computing the accumulative similarity between their "comparable chains" which have leaf nodes of the same kind of test data, while ignoring all other chains that do not have counterparts for comparison. Any mismatch between two patients that might be caused by information incompleteness cannot simply be treated as a real-world mismatch. We address the missing data problem by not applying penalties to missing data, i.e., if no chains of *patientB* are comparable to *chainA* of *patientA*, we simply ignore *chainA* and do not apply any penalty on the similarity score between *patientA* and *patientB*.

When computing the similarity between the leaf nodes of two comparable chains (i.e., the *Sim* function in Algorithm 1), the comparison can be either between two numeric values (e.g., patient age, pH value and HPV signal strength), or between two strings (e.g., cytology result and HPV status). The similarity between two numeric values is computed

$$\text{Sim}_{\text{numeric}}(n_1, n_2) = 1 - \frac{|n_1 - n_2|}{\max(n_1, n_2)} \tag{1}$$

where n_1 and n_2 are two numeric values and the function max returns the maximum between them. The numeric values that we currently handle are all positive numbers.

As for string values, in the cervical cancer domain or the more general clinical care domain, syntactically different strings could be semantically similar or vice versa. For example, "Positive0-Probable Normal" and "Positive1-LSIL" are two syntactically distinct strings but they represent two similar results of a clinical test of cervical cancer; "CIN 1" and "CIN 2" are similar in their syntactic representation whereas clinically they are two very different disease stages of cervical dysplasia. Therefore, instead of adopting traditional string matching algorithms (e.g., Jaccard and Edit distance) [43]–[48], we utilize domain knowledge about the semantic similarity between the result strings of a test. Fortunately, such knowledge is available in the original NCI/NLM relational database. In the NLM MDT [41], the possible results of a given clinical test are indexed with some integer numbers indicating their degree or grade. For instance, in a Cytology result, "Normal" is given index 1; "ASCUS" is given index 3; "CIN1" is given index 5; "CIN2" and "CIN3" are indexed with 6 and 7 respectively. Thus, utilizing such available domain knowledge, string similarity is computed

$$\text{Sim}_{\text{string}}(t, s_1, s_2) = 1 - \frac{|\text{Index}(t, s_1) - \text{Index}(t, s_2)|}{\text{maxDist}(t)} \tag{2}$$

where t represents a particular clinical test; s_1 and s_2 are two possible results of this test in string format; $\text{Index}(t, s)$ is a function, retrieving the assigned integer index for string s of clinical test t; $\text{maxDist}(t)$ gives the maximum distance between all possible results of test t, which is computed by subtracting the smallest assigned integer index from the largest one.

The similarity measures defined in (1) and (2) provide a more semantic notion of closeness than simply checking if two numeric values or two strings are identical; and thus they allow us to calculate more accurately the similarity between patients whose clinical test results are similar but not identical.

D. Computing Image-Based Similarity

In conjunction with database records, the cervical images can provide valuable and insightful information to assist in diagnosis and disease classification. In this paper, we adopt a data-driven approach in which we match a new, unannotated image to a database of expertly labeled images and rely on the human annotation of the closest matched images to guide interpretation of the new image. In our previous work, we have explored color [49], [50] and region-based features [51], [52] for cervigram comparison, segmentation, and annotation. Through our research, we have found that some of the most effective image features that highlight important visual characteristics present in cervigrams are color and texture, with color being slightly more discriminative in disease classification [53]. Thus, we utilize a combination of color and texture features to perform cervigram image similarity analysis.

Given a new cervigram image, it is important to isolate the cervix region from the rest of the image. Typically, the other parts of the image contain irrelevant information including equipment, text, and other noncervix tissue that may be detrimental to cervix classification and retrieval. Several previous works have used the local color and position features in order to isolate the cervix region [37], [54], [55]. These methods rely on a generative model of the cervix region to identify and segment various regions in an image. As mentioned before, we take a different approach to the region of interest detection problem. Our approach is data driven; we rely on an expertly labeled database of 939 cervigram images with their delineated rectangular regions of interest in order to find a suitable bounding box for the region of interest in a new cervigram image. The labeling of the region of interest in the new test image is generated automatically based on the bounding boxes of top matching training images in the expertly labeled database.

In order to utilize our image database, we need to compute a matching score between two images. Using the matching score, we can sort the similarity of the images to a new image and find the top $(\mathrm{k} = 20)$ most similar cases as shown in Fig. 5. For this initial image match process, we utilize a texture-based image feature, PHOG [56] (pyramid histogram of oriented gradients). PHOG has been one of the most effective features in image matching and retrieval. A PHOG descriptor represents local image shape and its spatial layout. The shape correspondence between two images can be measured by the distance between their PHOG descriptors using a spatial pyramid kernel. In this way, a strong match requires not only similar local patch appearance but also spatial correspondence over multiple scales. A spectrum of spatial correspondences can be represented in PHOG. At the coarsest level, the descriptor is a global edge or orientation histogram. At the finest level, the descriptor enforces tile (i.e., spatial bin) correspondence. Since our goal is to find the most similar cervigram images in terms of cervix shape and position, this texture feature encodes our ideal characteristics. To extract the PHOG descriptors from a cervigram image, we first compute the gradient response using a Sobel edge filter. We then split the image into rectangular regions, increasing the number of regions at each level. If we use an eight bin orientation histogram over four levels, the total vector size of our PHOG descriptor for each image is 680 bins; please see Fig. 6 for an illustration of this process.

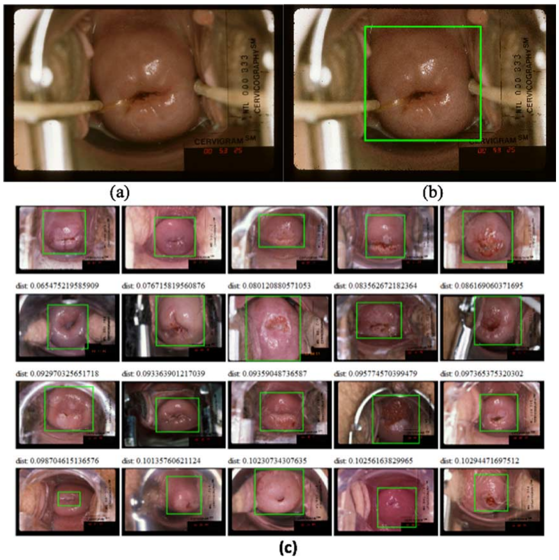

Fig. 5. Example of the computation of the cervix region of interest given a new image (a). Top 20 most similar images (c) are found in our expertly labeled image database and their bounding boxes are averaged to obtain the result (b). (a) Unlabeled input image. (b) Computed ROI. (c) Top 20 most similar images to the input image and their highlighted ground truth regions of interest.

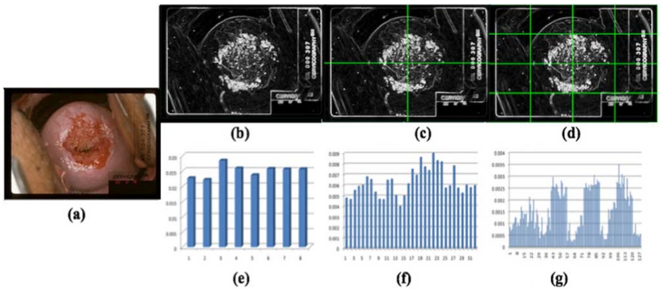

Fig. 6. Example of the PHOG feature extracted at multiple levels. Edges of the input image are computed by a Sobel edge filter and partitioned into a pyramid of regions (b)–(d). Eight orientation bins are extracted from each rectangle and concatenated into the PHOG feature vector represented in (e)–(g). Over four levels, the total vector size of the PHOG feature is 680 bins. (a) Input image. (b) $\mathrm{level} = 0$. (c) $\mathrm{level} = 1$. (d) $\mathrm{level} = 2$. (e) 8 bins at level 0. (f) 32 bins at level 1. (g) 128 bins at level 2.

Given the feature vector representation from individual images, the dissimilarity between two images can be computed via the measure as shown

$$\chi^2(s, q) = \frac{1}{2} \sum_{k=1}^{K} \frac{\left[h_s(k) - h_q(k)\right]^2}{h_s(k) + h_q(k)} \tag{3}$$

where h_s and h_q are the PHOG feature vectors of images s and q, and k is the vector index ranging from 1 to $K = 680$. To obtain a similarity score between two images, we normalize the

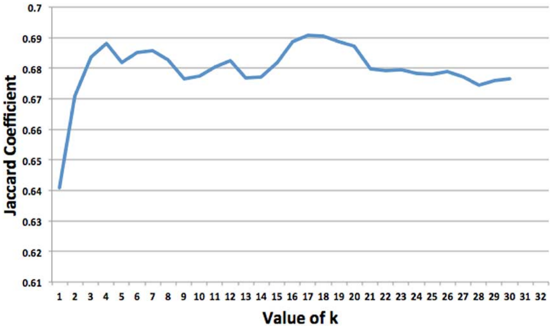

Fig. 7. Empirical justification for $\mathrm{k} = 20$ for our ROI computation.

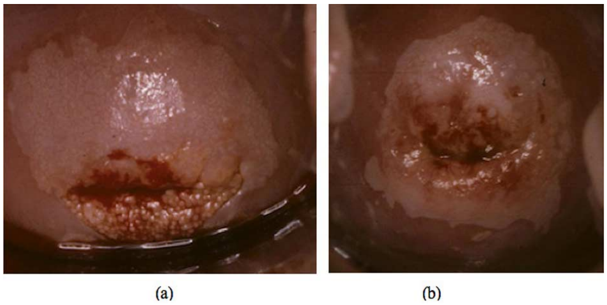

Fig. 8. More sample cervigrams of patients from $\mathrm{CIN2/3+}$ category. (a) Large acetowhite region with mosaicism. (b) Large and thick acetowhite region with punctation. Note that these are cropped and resized images; the original images are much larger (2891 by 1973 pixels).

above χ^2 measure to a value between 0–1, and the similarity is defined as

$$\mathrm{Sim}_{\mathrm{image}}(s, q) = 1 - \chi^2(s, q). \tag{4}$$

We use a database of images that are labeled with ground truth bounding boxes of the cervix region of interest in order to compute the bounding rectangle for our input image. Using a simple K-NN (k-nearest neighbor) technique and the similarity results, we can find the $k(= 20)$ most similar cervigram images to the input image. Then, we average the ROI bounding box coordinates from the k top matched cervigrams, after normalizing the ROI coordinates based on upon image size. The average coordinates serve as the bounding box ROI for our unlabeled input image. An example of a computed ROI can be seen in Fig. 5. In this example, Fig. 5(a) is the input image, and Fig. 5(b) contains the ROI obtained from matching with the set of images in our database, Fig. 5(c). The value of k used in our K-NN selection is chosen by gradient ascent of the Jaccard coefficient (intersection divided by the union of bounding boxes) on a sample of 50 ground truth images. The value of k is varied between 1 (closest match only) to 30. The higher the coefficient, the better the match of the ROI to a human annotated bounding box. Fig. 7 is an empirical justification for our selection of k.

Next, we describe how we use both texture and color features in the cervix ROI to compute cervix region similarity. As background information, the cervix is covered by a thin player of tissue (called the epithelium). A normal cervix has a glistening surface with smooth epithelium. A diseased cervix, on the other hand, typically exhibits various patterns of abnormal epithelium, such as Acetowhite, vascular structures, mosaicism, and punctation. Some examples can be seen in Fig. 2, where a normal/CIN1 cervix has a smooth appearance [Fig. 2(a)] whereas images of high-grade cervical lesion or cancer contain more complex texture due to pathological features [Fig. 2(b)]. In Fig. 8, we show several more images from the $\mathrm{CIN2/3+}$ category to demonstrate textures due to acetowhite, mosaicism, and punctation. Thus it is critical for an image comparison algorithm to capture the absence or presence of abnormal lesion patterns by extracting texture features from cervix ROI areas. Some commonly used texture features for segmentation and recognition include moment-based texture feature [57], GIST [58], [59], SIFT [60], PHOG [56], among others. As demonstrated before, we have found the PHOG feature to be effective in characterizing cervigram texture. Thus, within the ROI image area, we rerun our PHOG feature extraction on the internal region. This gives our method a more precise numerical description of the cervix area, without encoding a significant amount of background noise. The internal region PHOG feature has similar parameters as before (an 8 bin histogram over four levels, total of 680 bins). Each of the 8-bin histograms comprising the PHOG feature is range normalized to have a magnitude between [0, 1]. In this way, the variability within the ROI sizes (i.e., difference in the number of pixels within ROI areas) is accounted for. Furthermore, the PHOG feature is scale-invariant, thus two cervix regions of interest with similar texture but different scales still have high similarity.

Color also plays an important role in cervical lesion identification and classification. One of the most important visual features on the cervix that have relevant diagnostic properties include the presence of Acetowhite regions, or the whitening of potentially malignant cervical regions with the application of dilute acetic acid. The perceived thickness of an Acetowhite region is also relevant to cervical lesion grading, e.g., Fig. 1(c). Furthermore, there is usually greater color variance in images of high-grade lesion or cancer. Thus, we extract color-histogram features from the internal region of a detected cervix ROI to enable color-based image matching. We convert the pixel colors in a cervigram ROI into the perceptually uniform color space. A property of this color space is that a small change in the color value corresponds to about the same small change in visual appearance. The PLAB feature is also scale-invariant thus the scale of the region of interest does not affect color similarity computation. For each channel ($\mathrm{L}*$, $\mathrm{a}*$, or $\mathrm{b}*$) of the color space, we extract a 128-bin histogram. Concatenating the three channels together yields a total feature vector of 384 bins.

To compute the similarity between the cervix regions of interest of two images, we measure the numerical similarity between the ROI feature vectors. Each of the histograms comprising the PHOG or PLAB feature has been range normalized and then we fuse all the texture and color features into a 1064 bin histogram to represent a ROI. In this final representation of our feature vector, there are several multimodal fusion methods, each with their own benefits and drawbacks. In the early fusion approach, the color and texture feature vectors are combined early and used as a single representative feature vector in the image similarity computation. This method can capture the interactions between modalities; however, early fusion can be problematic with heterogeneous data, scales, and length distributions, as is the case with our PHOG texture feature and three-channel color feature. We employ a late fusion technique where each modality's similarity is first computed independently and then combined in a weighted distance measurement. The first 680 bins of our 1064 bin histogram (corresponding to PHOG texture feature) has full weight (1.0) in the similarity metric computation and each of the next 128 bin chunks (corresponding to three color-channel features) contributes one-third of the weight towards the similarity metric. Thus the final similarity measurement treats the texture and color features equally and also gives each color channel equal weight. This late fusion method alleviates the problems that early fusion has with heterogeneous data and is more extensively studied in [61].

And as aforementioned, we can use the normalized measure [(4)] to obtain a similarity score between two cervix ROIs. Since each patient typically has multiple cervigrams taken at multiple visits, the overall image similarity between two patients is measured by computing the average similarity between the ROIs of all pairs of images of the two patients. That is, the image-based patient similarity is defined

$$\mathrm{Image\_Sim}(a,b) = \frac{1}{N} \sum_{s \in I(a), q \in I(b)} \mathrm{Sim}_{\mathrm{image}}(s,q) \quad (5)$$

where a, b are two patients, the function $I(p)$ returns the set of cervigram images for patient p, N is the total number of image pairs, and the similarity score between two images $\mathrm{Sim\_image}(s,q)$ is defined in (4). Our image-based patient similarity score can now be used in conjunction with our data-level similarity score to measure the aggregated similarity between patients.

E. Patient Classification by Aggregating Image and Data Similarity

In this section, we describe how to augment patient data that are traditionally used in clinical testing with the cervigram image data. Our hypothesis is that the aggregation of these two sources of data should significantly improve the sensitivity and overall accuracy of the classifier in detecting high-grade cervical lesions compared to using either type of data alone.

1) Aggregating Data and Image Similarity: For combining these two heterogeneous types of data, we define an aggregated similarity metric over the data similarity $\mathrm{Data\_Sim}(G(a), G(b))$ (computed by Algorithm 1) and the image-based similarity $\mathrm{Image\_Sim}(a,b)$ (described in Section II-D). The aggregated similarity metric $\mathrm{sim}(a,b)$ for patients a and b is defined

$$\begin{aligned}\mathrm{sim}(a,b) = \alpha \times Data\_Sim(G(a), G(b)) \\ +(1-\alpha) \times \mathrm{Image\_}Sim(a,b). \quad (6)\end{aligned}$$

The parameter α is a weighting factor that represents how important Data Similarity is in the aggregation process.

In order to determine the value for α in (6), one approach is to manually assign weights and find the weight that gives us the best performance. In contrast, we could also automatically learn the optimal weights for combining Data Similarity and Image Similarity. In this paper, we employ a gradient-based learning approach [62], [63]. Specifically, we separate some of our data as validation data and use these data to find the optimal weights for Data Similarity and Image Similarity. We start with an initial α value (0.0), and then keep incrementing α value as long as the accuracy on the validation data does not drop significantly and current α value equals or is below 1.0. (In our current implementation, we keep trying the next α value as long as the accuracy does not drop more than 2%.) And among all the tested α values, we assign the value that produces the highest accuracy on the validation data to be our weight for Data Similarity. We then apply the learned weight to our testing data for classification.

One potential drawback of this gradient-based approach is that the process may fall into a local maximum of accuracy. In our evaluation later (Section IV), we perform multiple experiments where we start with different initial values for α in (6) and show that the achieved accuracies by using different initial values are very similar.

2) Determining Weights for Different Clinical Tests Within Data Similarity: In order to compute Data Similarity, one could manually assign weights to different clinical tests; however, approaches that can automatically learn the relative importance of different tests are preferred. In this paper, we employ an information gain-based learning approach to automatically calculating the weights for different clinical tests, i.e., Cytology (C), HPV (H), Age (A), and pH value (P), in order to compute Data Similarity. We treat each clinical test as a feature and compute the information gain [64] of different features with respect to the class label of the training samples. We then use the computed information gain values as the relative weights of the clinical test features. The Data Similarity is then calculated as the weighted average of similarities from all clinical test results.

3) Patient Classification: Our classification task is a binary classification task: whether a new patient p_n will be classified as <CIN2 (Negative) or CIN2/3+ (Positive). Conceptually our patient repository can be seen as a case base (CB) of cases, where each case has the form (p, c) where c is the class (i.e., Negative or Positive) of the patient p.

Algorithm 2 Classification of a new patient p_n given a case base CB

```
1) $CB_n \leftarrow \emptyset$
2) for each $(p, c) \in CB$ do
3)    $CB_n \leftarrow CB_n \cup (sim(p, p_n), c)$
4) end for
5) $CL \leftarrow KMeansCluster(CB_n)$
6) $tC \leftarrow topCluster(CL)$
7) return majorityVote($tC$)
```

We combine lazy and eager learning approaches [65] for our classifier as shown in Algorithm 2. First we initialize an auxiliary case base CB_n (Step 1). Then, CB_n is filled with pairs, in the format of (sim, c), of similarities between each patient p in the case base CB and the new patient p_n as well as the class label for p (Steps 2–4). We then apply K-means clustering on CB_n; the clusters are grouped by the similarity scores (Step 5). Finally, we return the class that occurs the most amongst cases in the top cluster (Steps 6 and 7). If there is a tie, a random selection is done among the classes that most frequently occur in the top cluster.

Note that K-means clustering was applied to the computed similarities between a testing case and all the training cases. Specifically, for a testing case, we compute its overall similarity (i.e., a weighted combination of Data Similarity and Image Similarity) to all training cases; thus, we have a list of similarity values (floating values). We then apply K-means to these similarity values to find the top cluster, i.e., the cluster that has the highest similarity values. After this, we do majority vote on training cases in the top cluster to determine the label of this testing case.

As for the size of CB, it includes all the training patients. In our experiments, we have 280 patient cases in total, and we divide them into 10 folds (28 cases in each fold). When we apply learning to determine the α value [in (6)] for combining Data and Image Similarity, we use 1 fold of cases as development/validation data to learn the parameter, use 1 fold for testing, and put the remaining 8 folds (224 cases) in the CB for training. When we do not use learning to set α but adopt a default value (e.g., setting $\alpha = 0.2$), we use 1 fold for testing and have the remaining 9 folds (252 cases) in the CB for training.

III. Experiment

A. Evaluation Metrics

As aforementioned, we evaluate our proposed system in a binary classification scenario, i.e., we classify a patient to be either $<$CIN2 (Negative) or CIN2/3+ (Positive). We measure the accuracy, sensitivity and specificity of our proposed multimodal patient classifier (see Algorithm 2). The definitions for these metrics are given as follows:

$$\text{Accuracy} = \frac{|\text{correctly classified patient cases}|}{\text{test cases}|} \tag{7}$$

$$\text{Sensitivity} \frac{= |\text{true positive}|}{|\text{true positive}| + |\text{false negative}|} \tag{8}$$

$$\text{Specificity} = \frac{|\text{true negative}|}{|\text{true negative}| + |\text{false positive}|} \tag{9}$$

where *true positive* refers to the set of patients who fall into the class "Positive" and are correctly classified; *false negative* refers to the set of patients who fall into the class "Positive" but are misclassified as "Negative"; *true negative* and *false positive* are similarly defined.

Following a standard of evaluating machine learning systems, we perform a 10-round 10-fold cross validation on our dataset of 280 patient cases (Table I). In each round, we randomly divide the patient cases into 10 folds; in a rotational manner, we use 1 fold for testing and the nine remaining folds for training; the testing result for the round is the average of the testing results for each of the 10 folds. The final testing result is the average accuracy/sensitivity/specificity of the ten rounds.

We also test the statistical significance between the results of our proposed system and other systems on our dataset. In this paper, we compare each pair of systems through the ten rounds and perform a two-tailed t-test on the two sets of results from the systems.

B. Multimodal Entity Classifier versus Data/Image-Only

In this experiment, our goal is to examine the effectiveness of different types of information in the cervical cancer patient classification task, including Cytology, HPV, patient age, pH value, and cervigrams (digital images). We first test the individual effectiveness of Cytology, HPV and cervigram, i.e., only using one of the three types of information for classification, and compare their performance. Furthermore, we perform classification by combining different types of information, e.g., using Cytology, HPV, age, pH, cervigram together, and then compare the classification accuracy using these combinatory tests with that of using a single type of information.

In the Guanacaste dataset used in our experiments, the possible values for Cytology include Normal, Rctive, ASCUS, Koil. Atyp, CIN 1/2/3, Micrinv Cncr and Inv Cancer. There are two components to HPV: 1) HPV Signal, which is a floating value ranging from 0.0 to 5.0, and 2) HPV Status, which can be either Negative or Positive. Patient age is a numeric value ranging from 15 to 100. pH value is another numeric value ranging from 1.0 to 14.0. When computing data-level similarity, (2) is used to calculate similarity between numeric-value features (such as HPV signal, age, and pH value), and (2) is used to calculate similarity between textural/string-value features (such as Cytology and HPV status). Equation (3) is applied to compute image-based similarity between patients cervigrams. Then data-level and image-based similarities are aggregated according to (6). By retrieving similar patient cases from an annotated database based on aggregated similarity, a new patient can be classified following Algorithm 2.

1) Manual Weight Assignment for Integrating Data and Image Similarity: In our first experiment, we treat all clinical test results equally to compute Data Similarity. Then we manually assign the weights for integrating Data and Image Similarity, i.e., manually varying the value for α in (6) between [0, 1], to find the weights that give the best accuracy. Please see all the "Manual" columns in Table II for results from this experiment.

In general, compared to systems that use textual/numeric data-only or use images-only for patient classification, our proposed system that aggregates the data and image similarities significantly improves accuracy over systems that use fewer information sources. Please note that the performance numbers in Table II are the average accuracy/sensitivity/specificity from 10-round 10–fold cross validation using 280 patient cases.

First of all, the Image-Only (I) patient classification gave overall accuracy 81.93%, sensitivity 74.14% and specificity 89.71%. These results are better than classification using Cytology (C) alone, HPV (H) alone, even both Cytology and HPV (C+H). This demonstrates the great potential of using computer-assisted interpretation of photographic images as an adjunctive screening and diagnosis test for cervical cancer. Then, by integrating multiple clinical tests and images, the overall best accuracy was 86.86% and it was achieved by applying multimodal patient classification using the combination of Cytology, HPV, pH, and images $(\mathrm{C} + \mathrm{H} + \mathrm{P} + \mathrm{I})$. In comparison, using clinical data-only $(\mathrm{C} + \mathrm{H} + \mathrm{A} + \mathrm{P})$, the accuracy was 80.07%, using C+H+P only achieves an accuracy of 78.79%, and using image-only (I), the accuracy was 81.93%. The results here are statistically significant with 95%

TABLE II
Performance of Multimodal (Both Clinical Data and Image), Data-Only and Image-Only Classifications (C: Cytology; H: HPV; I: Image; A: Age; P: pH) (AC: Accuracy; SE: Sensitivity; SP: Specificity) (Manual: Manually Determine Weights for Data and Image Similarity, and Treat All Clinical Tests Within Data Similarity Equally; IG: Automatically Learn the Weights for Different Clinical Tests With Information Gain; IG + Gradient: Utilize IG and Also Adopt Gradient-Based Learning Approach to Automatically Determine the Weights for Data and Image Similarity)

System	Manual			IG			IG+Gradient		
	$AC(\%)$	$SE(\%)$	$SP(\%)$	$AC(\%)$	$SE(\%)$	$SP(\%)$	$AC(\%)$	$SE(\%)$	$SP(\%)$
C	66.36	36.79	**95.93**	61.70	25.16	**98.21**	N/A	N/A	N/A
H	74.99	56.54	93.43	75.06	56.71	93.42	N/A	N/A	N/A
I	81.93	74.14	89.71	N/A	N/A	N/A	N/A	N/A	N/A
H+I	85.89	**80.21**	91.57	86.07	80.57	91.57	87.79	82.79	92.82
C+I	83.04	71.29	94.79	83.21	71.43	95.00	84.93	74.14	95.71
C+H	76.25	58.64	93.86	76.86	60.14	93.57	N/A	N/A	N/A
C+H+I	85.71	77.07	94.36	86.21	78.36	94.07	88.29	81.43	95.14
C+H+P	78.79	64.29	93.29	76.86	60.14	93.57	N/A	N/A	N/A
C+H+P+I	**86.86**	**80.21**	93.50	86.21	78.36	94.07	88.29	81.43	95.14
C+H+A	79.32	65.93	92.71	79.54	67.79	91.29	N/A	N/A	N/A
C+H+A+I	86.57	80.00	93.14	**87.43**	**82.00**	92.86	**89.00**	**83.21**	94.79
C+H+A+P	80.07	68.93	91.21	79.54	67.79	91.29	N/A	N/A	N/A
C+H+A+P+I	86.14	**80.21**	92.07	**87.43**	**82.00**	92.86	**89.00**	**83.21**	94.79

confidence. This demonstrates the effectiveness of combining both data and image similarities for patient classification. For both accuracy and sensitivity, a two-tailed t-test on the results between "$\mathrm{C+H+P+I}$" and "$\mathrm{C+H+P}$" or "I" gave a P value of 0.0001. Although no significant difference was observed between "$\mathrm{C + H + P + I}$" and "$\mathrm{C + H + P}$" on specificity, both of them have higher specificity than "I", with a P value of 0.0001.

Furthermore, we also show how the accuracy, sensitivity and specificity of $\mathrm{C+H+P+I}$ change by adopting different weights to Data Similarity, as demonstrated in Fig. 9. One can see that, starting with zero weight to Data Similarity and by gradually increasing the relative weight of Data Similarity, we gradually achieve better performance for all three metrics and the best results are achieved by assigning a value of 0.2 to 0.3 for Data Similarity. By only using Image ($\alpha = 0.0$) or Data ($\alpha = 1.0$), the performance is not as good as by integrating both types of information.

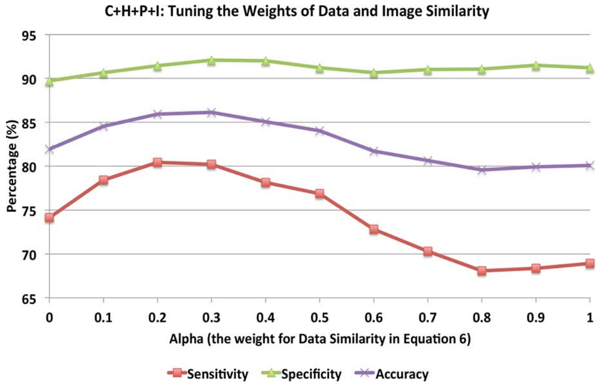

Fig. 9. Accuracy, sensitivity, and specificity by manually varying the weight for data similarity in (6).

2) Automatic Learning of Weights: Instead of treating all clinical tests equally and manually assigning weights to Data Similarity and Image Similarity, here, we present our evaluation results by adopting an information gain based approach for learning relative weights between data terms and a gradient-based approach for learning relative weights between Data Similarity and Image Similarity.

Please see all the "IG" columns in Table II for results using the information gain-based approach to learning relative weights between clinical tests (i.e., data terms). One can see that, using information gain-based learning, $\mathrm{C + H + A + I}$ achieved the best accuracy of 87.43%, which is even higher than the best accuracy under "Manual" (achieved by $\mathrm{C+H+P+I}$). Moreover, compared to the highest sensitivity of "Manual" (achieved by $\mathrm{C+H+P+I}$), $\mathrm{C+H+A+I}$ using information gain gives a sensitivity that is 1.79% higher than the best under "Manual". Comparing $\mathrm{C+H+A+I}$ to itself across "Manual" and "IG," for both accuracy and sensitivity, the improvements by IG are significant with P values of 0.0017 and 0.0003, respectively. Also, among the 12 systems in Table II, by using information gain-based learning method, 8 of these systems were able to achieve higher accuracy than the corresponding systems under "Manual." The results here demonstrate the effectiveness of utilizing information gain for learning the weights of different clinical tests. One interesting observation is that, in this experiment, C+H+A+P+I has exactly the same performance as $\mathrm{C + H + A + I}$. This can be explained by the fact that the calculated weight (information gain) for P (pH) is 0, indicating that P does not provide any added value for our classification task.

Please note that, for results shown under the "IG" columns in Table II, we automatically learn the weights for different clinical tests but still manually assign the weights between Data and Image Similarity. Fig. 10 shows how the accuracy, sensitivity, and specificity of $\mathrm{C+H+A+I}$ change by varying the weight for Data Similarity [i.e., α in (6)]. Similar to the results in Fig. 9, the best accuracy was achieved by setting the weight of Data Similarity to 0.2.

Next, rather than manually tuning the weights for Data and Image Similarity, we discuss the results of employing our gradient-based method for learning such weights. As discussed in Section II-E1, we need to separate some data out as validation data so that we can learn the optimal weight from the validation data. In our current experiments, we use 1 fold (i.e., 28 patient cases) for validation, one fold for testing, and use the other eight folds for training. Using our gradient-based learning approach, the learned optimal weight of Data Similarity is 0.176, which is

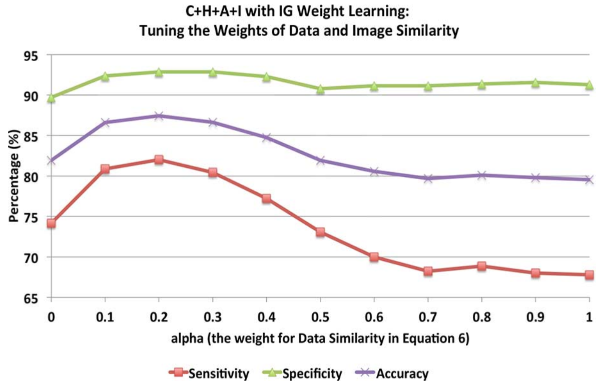

Fig. 10. Accuracy, sensitivity, and specificity after information gain based weighting of data terms and by varying the weight for Data Similarity in (6).

generally consistent with our manual tuning results (Figs. 9 and 10).

Please see all the "$\mathrm{IG + Gradient}$" columns in Table II for results using both gradient-based learning (to optimize α) and information gain based learning (to combine data terms). Because the gradient-based learning method is designed to integrate the two high level similarities: Data Similarity and Image Similarity, systems that only use Data or Image Similarity are not affected, including C, H, I, $\mathrm{C + H, C + H + P, C + H + A, C + H + A + P}$; for these systems, we put "N/A" as their performance under "$\mathrm{IG + Gradient}$" columns.

From Table II, we can see that by adopting both information gain and gradient-based learning approaches, $\mathrm{C+H+A+I}$ now has the best accuracy at 89.00% and also the highest sensitivity at 83.21%. Comparing $\mathrm{C+H+A+I}$ to itself between "IG" and "$\mathrm{IG + Gradient}$," a two-tailed t-test shows that the difference on both accuracy and specificity are statistically significant with a P value of 0.0001; the difference on sensitivity is also significant with a P value of 0.0313. The experimental results here demonstrate the effectiveness of adopting both learning-based approaches for automatic weight learning. It also shows that the best performance is achieved not by using all cues but by selecting the best subset of cues, i.e., $\mathrm{C + H + A + I}$ (Cytology, HPV, patient age, and image) under "IG+Gradient" columns.

Furthermore, we also compare $\mathrm{C + H + A + I}$ to other systems that also adopt "IG+Gradient". For accuracy, statistically, the differences between $\mathrm{C + H + A + I}$ and $\mathrm{C + I/H + I}$ are significant with P values of 0.0001 and 0.0012, respectively; although $\mathrm{C + H + A + I}$ has higher accuracy than $\mathrm{C + H + I}$, the difference is not significant with a P value of 0.0530. For sensitivity, we have similar results and the differences between $\mathrm{C + H + A + I}$ and $\mathrm{C + H + I/C + I}$ are significant with P values of 0.0212 and 0.0001, respectively. We think the results here verify the benefits of using multiple types of information together for performing our cervical cancer classification task. In Table II, for "$\mathrm{IG + Gradient}$", although $\mathrm{C + H + A + I}$ also has better sensitivity than $\mathrm{H + I}$ (0.42% higher), the results are not significant with a P value of 0.4335.

TABLE III
PERFORMANCE COMPARISON BETWEEN OUR PROPOSED METHOD AND SEVERAL OTHER PUBLISHED ALTERNATE APPROACHES
(AC: ACCURACY; SE: SENSITIVITY; SP: SPECIFICITY)

Computerized System	$AC(\%)$	$SE(\%)$	$SP(\%)$
Multi-Modal (our proposed method)	89.00	83.21	94.79
Kim et. al, Majority Vote [53]	75*	73	77
Kim et. al., SVM [53]	75.5*	75	76
DeSantis et. al. [66]	71.3*	95	55
Chang et. al. [67]	82.39*	72	83
ThinPrep [17]	81.36∼95.75*	79∼82	98∼99
BD FocalPoint GS Imaging System [16]	N/A	81∼86	85∼95

C. Comparison to Alternate State-of-the-Art Systems

In addition to comparing between our own systems, in this section, we compare our best system ($\mathrm{C + H + A + I}$ by using $\mathrm{IG+Gradient}$) to several other published alternate approaches. We summarize the comparison in Table III[1].

We first compare our proposed system multimodal to state-of-the-art imaging systems for cervical cancer detection. The systems by Kim and Huang [53] perform cervical cancer detection by analyzing cervigram images. They utilize preannotated images for automatically locating the region of interest (ROI) on the cervix; then, by extracting color and texture features from the cervix ROI, the systems were able to achieve comparable accuracy to a trained expert. However, by only using image for classification, both of the systems have lower performance than our proposed multimodal method.

Next, we compare our proposed system with imaging techniques designed for assisting the cervical cancer diagnosis process. Desantis *et al.* [66] examined the potential of using tissue spectroscopy for the diagnosis of cervical cancer. They used a prototype device made by Guided Therapeutics, Inc., Norcross, GA for taking spectroscopy measurements. Then such collected images and other relevant data (such as Pap result and patient demographic information) are processed and analyzed by a diagnostic algorithm to produce the final result, i.e., whether this patient has cancer or not. This is the most similar system to our proposed approach. Instead of only using image analysis techniques, it tries to combine different modalities together for diagnosis. This system achieved satisfactory sensitivity (95%); however, it has a high false-positive rate (i.e., low specificity), which could potentially cause patients to encounter unnecessary and costly diagnostic procedures and even treatments. Chang *et al.* [67] try to analyze the diagnostic potential of utilizing reflectance and fluorescence spectra to discriminate normal and precancerous cervical tissue. They examined different combinations of spectral features and utilized the features in classification algorithms for evaluating the diagnostic performance of different feature sets. This system achieved a similar sensitivity to our proposed multimodal system; however, the specificity of their system is significantly lower than that of our system. Thekkek and Richards-Kortum [68] summarized results from previous similar studies.

Feature selection and fusion are important aspects of classification problems since using a suitable set of features can significantly improve the final classification accuracy. Zhang *et al.* [69] designed a feature selection algorithm for choosing the most effective features for image annotation; Gehler *et al.* [70] proposed an algorithm for learning the correct weighting of different features for multiclass classification. In our system, using

[1] The accuracy values marked with an asterisk were derived from data provided in those papers.

TABLE IV
IMPACT OF DOMAIN KNOWLEDGE ON CLASSIFICATION RESULTS (AC: ACCURACY; SE: SENSITIVITY; SP: SPECIFICITY)

System	AC (%)	SE (%)	SP (%)
Multi-Modal (C+H+A+I)	**89.00**	**83.21**	94.79
Multi-Modal (C+H+A+I) no DK	87.04	79.21	**94.86**
Data-Only (C+H+A) DK	**79.54**	**67.79**	**91.29**
Data-Only (C+H+A) no DK	76.76	65.74	87.82

TABLE V
PERFORMANCE COMPARISON FOR MULTIMODAL CLASSIFICATION WITH DIFFERENT CLASSIFIERS (AC: ACCURACY; SE: SENSITIVITY; SP: SPECIFICITY.) (CLUSTER: MAJORITY VOTING BY CASES IN TOP CLUSTER; AVG: AVERAGE SIMILARITY TO CASES IN EACH CLASS; MAX: MAXIMUM SIMILARITY TO CASES IN EACH CLASS)

Classifier	AC (%)	SE (%)	SP (%)
Cluster	**89.00**	**83.21**	94.79
Avg	84.36	71.93	**96.79**
Max	85.50	82.36	88.64

automatically learned weights also greatly improved our classification accuracy.

Recent developments in industry have also led to imaging-based cervical cancer diagnosis systems, such as ThinPrep [17] and BD FocalPoint GS Imaging System [16]. The ThinPrep Imager (Cytyc) system, a computerized system for reading slides, is a new technology applied to liquid based cytology. The imager identifies 22 fields of interest most likely to contain abnormal cells, which are then examined by a cytologist. The system from BD [16] implements a similar idea. Compared to these two commercial systems on the market, our proposed multimodal was able to achieve comparable performance, except that the specificity of our system is lower than that of ThinPrep. However, there are two advantages of our system over these two commercial systems: 1) Our system is able to integrate multiple clinical tests and images to achieve better performance and also can produce a diagnosis directly from a photograph of a cervix; 2) Our system is more applicable in resource poor regions and also better suited for telemedicine.

In addition to systems that perform cancer diagnosis, algorithms were also proposed for detecting lesion regions. Alush *et al.* [71] and Park *et al.* [72] developed systems for automated lesion detection and segmentation. Yu *et al.* [73], Zhang *et al.* [74], and Gordon *et al.* [75] proposed algorithms for segmentation of cervical images. Although these works are not performing end-to-end cervical cancer diagnosis, accurately detected lesion regions and other regions of interest in cervigram images can be further analyzed to assist with the diagnosis process. In our current approach, being able to accurately recognize the region of interest (ROI) is also important for calculating image similarity in order to facilitate the final classification task.

D. Effectiveness of Domain Knowledge

In this experiment, we show that adopting domain knowledge (DK) for computing data-level string similarity [(2)], can significantly improve the results as shown in Table IV. For this experiment, we use $\mathrm{C} + \mathrm{H} + \mathrm{A} + \mathrm{I}$, the best-performer according to Table II (i.e., combining Cytology, HPV, and the patient age information together); and we adopt both information gain and gradient-based learning approaches. Since adding domain knowledge will not affect Image-Only classification, it is not compared here.

We can see that adopting domain knowledge helped to achieve significant improvements, particularly in accuracy and sensitivity for both $\mathrm{C} + \mathrm{H} + \mathrm{A} + \mathrm{I}$ and $\mathrm{C} + \mathrm{H} + \mathrm{A}$ classification. For accuracies, the differences here are statistically significant with a P value of 0.0003 between multimodal and multimodal no DK, and a P value of 0.0001 between Data-Only and Data-Only no DK. For sensitivity, we also observe results that are statistically significant: a P value of 0.0005 between the two multimodal systems, and a P value of 0.0001 between the two Data-Only systems. This verifies our assumption that in this domain, syntactically different strings could actually be semantically close to each other; therefore, it is important to capture such semantic similarity. In our current work, such semantic similarity is exploited by utilizing the index integers assigned to strings in the NLM-MDT database, assuming semantically similar test-result strings will be assigned close indices. In future work, we plan to explore how to compute semantic similarity of two strings by using some dictionaries or ontologies in the domain [76].

E. Comparing Different Classification Schemes

As presented in the section on Patient Classification, our classification scheme involves retrieving similar patient cases from a case database, performing K-means clustering on the similar cases, and adopting the class label as voted by a majority of cases in the top cluster. For K-means clustering (Step 5 of Algorithm 2), we tried different K values and found $K = 5$ to be a good choice given our training case base of size 252. Note that our training case base has a size of 224 because there are 280 cases in total and in each round of 10-fold cross validation, one fold (28 cases) is used for development, one fold (28 cases) is used for testing, and the rest eight folds (224 cases) are used for training.

Alternatively, instead of majority voting by cases in the top cluster, we could compute the average (or maximum) similarity between a test case and all training cases in each class, and then assign to the test case the class label with maximum similarity. We compared these alternatives in Table V for multimodal classification using $\mathrm{Cytology} + \mathrm{HPV} + \mathrm{Age}$ (as Data) as well as images, i.e., $\mathrm{C} + \mathrm{H} + \mathrm{A} + \mathrm{I}$, the best-performer according to Table II; and we utilize both information gain and gradient-based learning approaches.

The results show that majority voting by top cluster gives both the best accuracy and sensitivity. Statistically, on accuracy, the differences between Cluster and other classification schemes (Avg and Max) are significant with a P value of 0.0001. On sensitivity, the difference between Cluster and Avg is significant with a P value of 0.0001.

F. Summary of Results

In summary, we have developed a computer-assisted algorithm that interprets cervigrams based on color and texture. The algorithm yields 74% sensitivity and 90% specificity in differentiating $\mathrm{CIN2/3+}$ from $<\mathrm{CIN2}$, on a dataset involving 280 randomly selected patient cases. In comparison, using Pap test alone gives sensitivity 37% and specificity 96%, and using HPV test alone gives sensitivity 57% and specificity 93%, on the same dataset. When computer assistance is not used, the sensitivity for detecting $\mathrm{CIN2/3+}$ was 39%, as reported by a study that had 20 expert colposcopists visually assess digital cervical images [33].

Furthermore, Our framework enables the efficient evaluation of the performance of various combinatory tests. A novel combinatory test, which integrates multiple modalities—Pap, HPV, information derived from Cervicography images, and patient age, yields about 83% sensitivity and 95% specificity, a statistically significant improvement over any single modality or other combinatory tests derived from proper subsets of these four modalities. Our results demonstrate the potential of using computer interpretation of cervical images as an adjunctive test to Pap and HPV in cervical cancer screening.

IV. Discussion

In this paper, we presented a data-driven approach for cervical dysplasia diagnosis using images and other clinical test results. Patient data are represented in a hierarchical tree-like data structure. Patient comparison is performed through an entity coreference algorithm that compares two entities through similarity between "comparable data chains" without incurring penalty for unmatchable data chains; thus our method naturally handles unbalanced data. Compared to existing cervical image analysis methods that only perform processing or segmentation of cervigrams without patient classification [27], [35]–[37], [39], [54], [55], our cervigram image interpretation algorithm is able to produce a cervical dysplasia diagnosis (either $<\mathrm{CIN2}$ or $\mathrm{CIN2/3+}$) with high accuracy. Furthermore, our novel multimodal Entity Coreference algorithm can effectively compute the similarity between patients utilizing their hierarchical representation of heterogeneous data including cervigram images, Pap, HPV, pH and other clinical test results. Both our multimodal and image-alone classification schemes achieve similar or better sensitivity and specificity when compared to other methods for cervical disease classification [72], [77]; furthermore, while these other methods were tested on several dozen patient cases, our system is tested on a much larger set of 280 patient cases.

Regarding the cervical cancer screening application, our work has demonstrated the potential of Digital Cervicography, which produces cervigram images, as a low-cost and widely accessible screening method with reasonable accuracy, when augmented by computerized interpretation of cervigrams and a large database of expertly annotated patient cases and images. It has also shown that integrating images with other clinical information can improve the accuracy in differentiating low-grade cervical lesions from high-grade lesions and invasive cancer. By using only digital cervigram images, our proposed system achieved 74.14% sensitivity for detecting CIN2/3+ lesions; and by using images and three other clinical test results (Cytology, HPV, age), our system achieved 83.21% sensitivity. In comparison, the commonly used Pap test screening highly depends on the expertise of laboratory personnel as well as workplace infrastructure; as shown in Table VI, its sensitivity for detecting CIN2/3+ lesions varies widely in different geographic regions: 18%–20% in Germany [14], 22%–24% in Chile [7], 26%–43% in Peru [9], 42%–56% in Canada [10], 57% in Africa and India [11], 63%–86% in Costa Rica [12], and 77% in the United Kingdom [13]. As one can see, the sensitivity levels of our system match the best results reported in clinical literature, which shows the potential of using our system for cervical cancer screening and diagnosis.

TABLE VI
Comparing to Cervical Cancer Diagnosis That Uses Pap Test

System/Clinical Trial	$Sensitivity(\%)$	$Specificity(\%)$
Multi-Modal	83.21	94.79
Image Only	74.14	89.71
Schneider et. al., Germany [14]	18$\sim$20	99
Ferreccio et. al., Chile [7]	22$\sim$24	99
Almonte et. al., Peru, [9]	26$\sim$43	38$\sim$99
Mayrand et. al., Canada [10]	42$\sim$56	97$\sim$99
Ferreccio, Costa Rica [12]	63$\sim$86	88$\sim$94
Cuzick, UK [13]	77	96

TABLE VII
Patient Age Distribution in 280 Randomly Selected Patient Cases)

Category	<21	21-29	30-40	41-65	>65
<CIN2 (Negative)	0	9	48	59	24
CIN2/3+ (Positive)	1	38	53	42	6

Another interesting observation from our experimental results in Table II is that adding patient age information improved the sensitivity of the system by sacrificing some specificity and therefore enabled the system to achieve better overall classification accuracy when combining all information together. Comparing $\mathrm{C+H+A+I}$ to $\mathrm{C+H+I}$, $\mathrm{C+H+A}$ and $\mathrm{C+H}$ for all three situations (i.e., "Manual," "IG," and "$\mathrm{IG + Gradient}$"), one can see that patient age information helped to improve sensitivity and accuracy. In fact, patient age has been an important factor used in cervical cancer screening guidelines for average-risk women [78], [79]. For example, it is recommended that women aged less than 21 should not be screened; for women between 21 and 29 years old, Cytology alone should be used every three years without HPV co-test; for women between 30 and 65 years old, Cytology should be used every three years with HPV co-test every five years; and it is recommended that cervical cancer screening can stop for women aged >65 years with adequate screening history. To seek further explanation for the improvement in classification accuracy by adding patient age as a feature, we compiled statistics about patient age from our 280 randomly selected patient cases, as shown in Table VII; here one can see that the distribution of disease does differ significantly from one age group to another, thus making age a useful feature when comparing patients and performing disease classification. The best performance was obtained by $\mathrm{C + H + A + I}$ (Cytology, HPV, age, and images) with information gain and gradient-based learning approaches, which gave much better accuracy and sensitivity than only using individual tests.

In our current work, we perform automatic weight learning in two situations: 1) We use gradient-based approach to learn the weights of Data Similarity and Image Similarity; 2) In order to appropriately integrate the different clinical test results within Data Similarity, we employ an information gain-based method. Theoretically, we could utilize the gradient-based approach for learning in both situations. However, as shown in

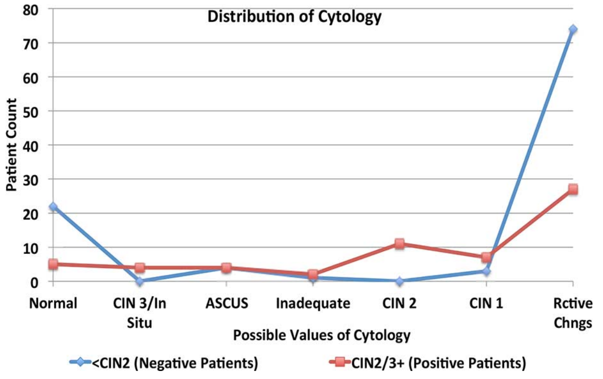

Fig. 11. Distribution of cytology for all patients.

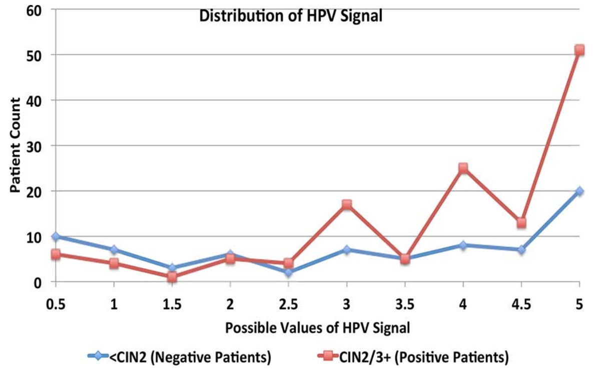

Fig. 12. Distribution of HPV signal for all patients.

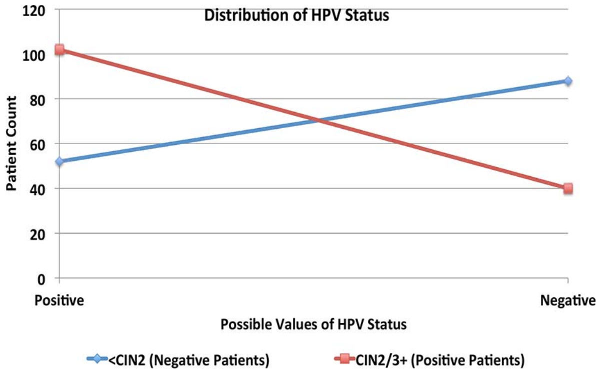

Fig. 13. Distribution of HPV status for all patients.

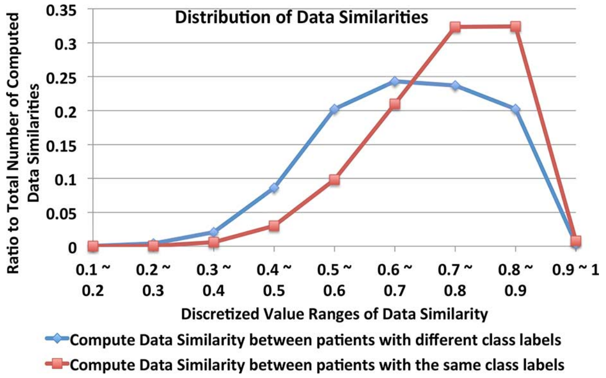

Fig. 14. Distribution of all data similarities.

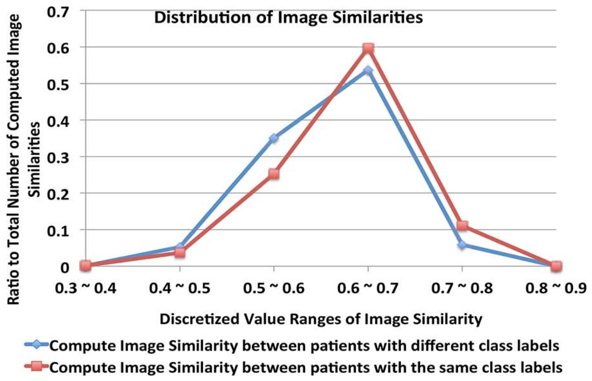

Fig. 15. Distribution of image similarities.

Figs. 11, 12, and 13, the distributions of clinical test results (i.e., data terms) have no distinct modes, and gradient/hill climbing approaches often do not work well on such data.

In contrast, as demonstrated in Figs. 14 and 15, the distributions of Data and Image Similarities are fairly smooth and have clear modes, thus a gradient/hill climbing approach was a good fit.

As discussed in Section II-E1, our gradient-based approach may fall into a local maximum. Therefore, we performed an additional experiment where we adopt different initial values for learning the optimal weights for integrating Data Similarity and Image Similarity. We utilized three different starting values for α in (6): 0.0, 0.5, and 1.0. And, we achieved accuracies of 89.00%, 89.04%, and 88.71%, respectively. Although we got different accuracies here, the differences between the three results are not statistically significant $(\mathrm{P} > 0.05)$. This shows to some extent that the gradient-based learning method is not sensitive to initialization and is effective for learning the weights to integrate Data Similarity and Image Similarity in our classification task.

One limitation of our approach is that, because of its data centric nature, it works well with typical cervigram images and typical cervical cancer patient cases, but may have difficulty with outlier images (or patient cases) that do not closely match any of our expertly annotated examples in the database. Expanding the expertly annotated case base and improving patient similarity measures are feasible remedies to this problem.

We also examined the computational complexity of our multimodal classification system. For a single image, (PHOG) feature extraction takes between 2.9–3.1 s. It takes little time (0.2 ms) to compute the image-similarity score between two images once their features have been extracted. In our experiments, feature extraction for the 939 images in the expertly labeled database is done offline and the extracted features are stored, thus reducing the run time to compute the ROI of an input test image to around 3 s. During the patient disease classification phase, we also precompute image features for all images of patients in the training dataset. Our algorithm does not require any manual interaction for a testing image. The labeling of a bounding box is only needed for images in the training database; for a new test image, the bounding box region of interest is generated automatically using the bounding box information about top matching training samples that we have stored in our training database.

Thus, for a test patient case, the run time includes the time to compute features for all its images (3 s/image), the time to compute its image similarities to patients in the training set (on average 3.96 s), and once image similarities are calculated, the multimodal entity coreference algorithm takes 33 ms to classify the patient case using all five sources of information on a laptop computer with 4 GB memory and 2.0 GHz quad-core CPU. Theoretically, the complexity of our multimodal entity coreference algorithm depends on the number of chains in a patients data tree. In the worst case, suppose a tree has p chains and each chain is comparable with all chains in the other tree, the complexity for comparing two trees is then $\mathrm{O}(p^2)$. For classification, each test case is computed against all training cases and suppose we have n training cases in total, then the complexity for comparing a test case with all training cases is $\mathrm{O}(n * p^2)$. Once the similarity scores are computed, it takes $O(n \log^n)$ time to sort the scores and obtain the top cluster of most similar cases for classification.

V. Conclusion and Future Work

In this paper, we demonstrate that computerized interpretation of cervical images and combinatory tests that integrate images with other clinical information have the potential to be used as adjunctive screening or diagnosis tests for differentiating low-grade cervical lesions (which do not need treatment) from high-grade lesions and invasive cancer. We presented an intelligent computer-assisted diagnostic system that can identify different stages of cervical lesion with fair correlation by integrating cervical image interpretation, Pap test, HPV test, and additional clinical information such as pH value and patient age. We demonstrated how adding images to traditional Pap and HPV tests can improve sensitivity of detecting high grade CIN. We also used the framework to evaluate the sensitivity and specificity of various combinatory tests, and discovered the value of other parameters such as patient age. The framework can be extended to other applications that involve multimodal classification and efficiently perform data analysis and data mining to discover unknown linkages and effective combinatory tests.

In our future work, to further improve the accuracy of our classification process, it would be interesting to explore other weighting mechanisms for integrating different types of clinical and image information. We will explore approaches that can be used to further reduce the computational cost [80]. For image analysis, additional techniques such as color calibration and illumination correction can be explored to further improve the quality of patient image similarity computation. In addition, instead of only using the clinical data available in the database, we could potentially try to bring in information from other data sources such as medical literature, apply text annotation/mining techniques to extract useful information from those sources [81]–[84], and use the extracted information as additional features for our classification task. We will also attempt to extend our system to more fine-grained multiclass disease grading instead of binary classification.

Acknowledgment

The authors gratefully acknowledge the long-standing expert advice and support they have received from Dr. M. Schiffman and Dr. N. Wentzensen of the National Cancer Institute in the use and interpretation of data from the NCI Guanacaste and ALTS Projects.

References

[1] Atlanta: American Cancer Society “Global cancer facts and figures,” 2011.

[2] D. G. Ferris, “Cervicography—An adjunct to Papanicolaou screening,” *Am. Fam. Phys.*, vol. 50, pp. 363–370, 1994.

[3] L. Denny, L. Kuhn, A. Pollack, H. Wainwright, and J. T. C. Wright, “Evaluation of alternative methods of cervical cancer screening for resource-poor settings,” *Cancer*, vol. 89, pp. 826–833, 2000.

[4] Am. Cancer Soc. “Cancer facts and figures,” 2010.

[5] Cancer facts—The pap test: Questions and answers [Online]. Available: http://www.cancer.gov/cancertopics/factsheet/detection/Pap-test, 2011

[6] R. Sankaranarayanan, L. Gaffikin, M. Jacob, J. Sellors, and S. Robles, “A critical assessment of screening methods for cervical neoplasia,” *Int. J. Gynecol. Obstetr.*, vol. 89, pp. 4–12, 2005.

[7] C. Ferreccio, M. I. Barriga, M. Lagos, C. Ibez, H. Poggi, F. Gonzlez, S. Terrazas, H. A. Katki, F. Nez, J. Cartagena, V. Van De Wyngard, D. Viales, and J. Braes, “Screening trial of human papillomavirus for early detection of cervical cancer in Santiago, Chile,” *Int. J. Cancer*, vol. 132, no. 4, pp. 916–923, 2013.

[8] M. Arbyn, R. Sankaranarayanan, R. Muwonge, N. Keita, A. Dolo, C. G. Mbalawa, H. Nouhou, B. Sakande, R. Wesley, T. Somanathan, A. Sharma, S. Shastri, and P. Basu, “Pooled analysis of the accuracy of five cervical cancer screening tests assessed in eleven studies in Africa and India,” *Int. J. Cancer*, vol. 123, no. 1, pp. 153–160, 2008.

[9] M. Almonte, C. Ferreccio, J. L. Winkler, J. Cuzick, V. Tsu, S. Robles, R. Takahashi, and P. Sasieni, “Cervical screening by visual inspection, HPV testing, liquid-based and conventional cytology in Amazonian Peru,” *Int. J. Cancer*, vol. 121, no. 4, pp. 796–802, 2007.

[10] M.-H. Mayrand, E. Duarte-Franco, I. Rodrigues, S. D. Walter, J. Hanley, A. Ferenczy, S. Ratnam, F. Coutle, and E. L. Franco, “Human papillomavirus DNA versus Papanicolaou screening tests for cervical cancer,” *N. Eng. J. Med.*, vol. 357, no. 16, pp. 1579–1588, 2007.

[11] R. Sankaranarayanan, P. Basu, R. S. Wesley, C. Mahe, N. Keita, C. C. G. Mbalawa, R. Sharma, A. Dolo, S. S. Shastri, M. Nacoulma, M. Nayama, T. Somanathan, E. Lucas, R. Muwonge, L. Frappart, and D. M. Parkin, “Accuracy of visual screening for cervical neoplasia: Results from an IARC multicentre study in India and Africa,” *Int. J. Cancer*, vol. 110, no. 6, pp. 907–913, 2004.

[12] C. Ferreccio ***et al.***, “A comparison of single and combined visual, cytologic, and virologic tests as screening strategies in a region at high risk of cervical cancer,” *Cancer Epidemiol., Biomark. Prevent.*, vol. 12, no. 9, pp. 815–823, Sep. 2003.

[13] J. Cuzick, A. Szarewski, H. Cubie, G. Hulman, H. Kitchener, D. Luesley, E. McGoogan, U. Menon, G. Terry, R. Edwards, C. Brooks, M. Desai, C. Gie, L. Ho, I. Jacobs, C. Pickles, and P. Sasieni, “Management of women who test positive for high-risk types of human papillomavirus: The HART study,” *Lancet*, vol. 362, no. 9399, pp. 1871–1876, 2003.

[14] A. Schneider, H. Hoyer, B. Lotz, S. Leistritza, R. Khne-Heid, I. Nindl, B. Mller, J. Haerting, and M. Drst, “Screening for high-grade cervical intra-epithelial neoplasia and cancer by testing for high-risk HPV, routine cytology or colposcopy,” *Int. J. Cancer*, vol. 89, no. 6, pp. 529–534, 2000.

[15] C. V. Biscotti, A. E. Dawson, B. Dziura, L. Galup, T. Darragh, A. Rahemtulla, and L. Wills-Frank, “Assisted primary screening using the automated thinprep imaging system,” *Am. J. Clin. Pathol.*, vol. 123, no. 2, pp. 281–287, 2005.

[16] D. C. Wilbur, W. S. Black-Schaffer, R. D. Luff, K. P. Abraham, C. Kemper, J. T. Molina, and W. D. Tench, “The becton dickinson focalpoint GS imaging system: Clinical trials demonstrate significantly improved sensitivity for the detection of important cervical lesions,” *Am. J. Clin. Pathol.*, vol. 132, no. 5, pp. 767–775, 2009.

[17] E. Davey, J. d'Assuncao, L. Irwig, P. Macaskill, S. F. Chan, A. Richards, and A. Farnsworth, "Accuracy of reading liquid based cytology slides using the thinprep imager compared with conventional cytology: Prospective study," *Br. Med. J.*, vol. 335, no. 7609, p. 31, Jul. 2007.

[18] M. Arbyn, C. Bergeron, P. Klinkhamer, P. Martin-Hirsch, A. Siebers, and J. Bulten, "Liquid compared with conventional cervical cytology: A systematic review and meta-analysis," *Obstet. Gynecol.*, vol. 111, no. 1, pp. 167–177, 2008.

[19] E. P. Whitlock, K. K. Vesco, M. Eder, J. S. Lin, C. A. Senger, and B. U. Burda, "Liquid-based cytology and human papillomavirus testing to screen for cervical cancer: A systematic review for the U.S. preventive services task force," *Ann. Internal Med.*, vol. 155, no. 10, pp. 687–697, 2011.

[20] H. C. Kitchener, R. Blanks, G. Dunn, L. Gunn, M. Desai, R. Albrow, J. Mather, D. N. Rana, H. Cubie, C. Moore, R. Legood, A. Gray, and S. Moss, "Automation-assisted versus manual reading of cervical cytology (mavaric): A randomised controlled trial," *Lancet Oncol.*, vol. 12, no. 1, pp. 56–64, 2011.

[21] A. G. Siebers *et al.*, "Comparison of liquid-based cytology with conventional cytology for detection of cervical cancer precursors: A randomized controlled trial," *J. Amer. Med. Assoc.*, vol. 302, no. 16, pp. 1757–1764, 2009.

[22] J. M. Roberts and J. K. Thurloe, "Comparative sensitivities of thinprep and Papanicolaou smear for adenocarcinoma *in situ* (AIS) and combined AIS/high-grade squamous intraepithelial lesion (HSIL): Comparison with HSIL," *Cancer Cytopathol.*, vol. 111, no. 6, pp. 482–486, 2007.

[23] G. Ronco, J. Cuzick, P. Pierotti, M. P. Cariaggi, P. D. Palma, C. Naldoni, B. Ghiringhello, P. Giorgi-Rossi, D. Minucci, F. Parisio, A. Pojer, M. L. Schiboni, C. Sintoni, M. Zorzi, N. Segnan, and M. Confortini, "Accuracy of liquid based versus conventional cytology: Overall results of new technologies for cervical cancer screening: Randomised controlled trial," *Br. Med. J.*, vol. 335, no. 7609, p. 28, 2007.

[24] M. Saraiya, V. Benard, and J. Miller, "Liquid-based cytology vs conventional cytology in detecting cervical cancer," *J. Am. Med. Assoc.*, vol. 303, no. 11, pp. 1034–1035, 2010.

[25] K. Hartmann *et al.*, "Screening for cervical cancer: Systematic evidence review," Agency Healthcare Res. Quality: U.S. Preventive Service Task Force 2002.

[26] A. Stafl, "Cervicography: A new method for cervical cancer detection," *Am. J. Obstet. Gynecol.*, vol. 139, pp. 815–825, 1981.

[27] S. Gordon, G. Zimmerman, and H. Greenspan, "Image segmentation of uterine cervix images for indexing in PACS," in *Prov. 17th IEEE Symp. Comput.-Based Med. Syst.*, 2004, pp. 298–303.

[28] S. Yang, J. Guo, P. King, Y. Sriraja, S. Mitra, B. Nutter, D. Ferris, R. L. M. Schiffman, and J. Jeronimo, "A multi-spectral digital cervigramtm analyzer in the wavelet domain for early detection of cervical cancer," in *Proc. SPIE Image Process.*, 2004, vol. 5370, pp. 1833–1844.

[29] M. Anderson, J. Jordon, A. Morse, and F. Sharp, *A Text and Atlas of Integrated Colposcopy*. New York: Mosby Year Book, 1991.

[30] Y. Srinivasana, D. Hernesa, B. Tulpulea, S. Yanga, J. Guoa, S. Mitraa, S. Yagneswarana, B. Nuttera, J. Jeronimob, B. Phillipsc, R. Long, and D. Ferris, "A multi-spectral digital cervigramtm analyzer in the wavelet domain for early detection of cervical cancer," in *Proc. SPIE, Med. Imag., Image Process.*, 2005, vol. 5747.

[31] D. Ferris, M. Schiffman, and L. M. S. , "Cervicography for triage of women with mildly abnormal cervical cytology results," *Am. J. Obstetr. Gynecol.*, vol. 185, no. 4, pp. 939–943, 2001.

[32] C. Eskridge, W. Begneaud, and C. Landwehr, "Cervicography combined with repeated Papanicolaou test as triage for low-grade cytologic abnormalities," *Obstetr. Gynecol.*, vol. 92, no. 3, pp. 351–355, 1998.

[33] L. S. Massad, J. Jeronimo, H. A. Katki, and M. Schiffman, "The accuracy of colposcopic grading for detection of high-grade cervical intraepithelial neoplasia," *J. Lower Genital Tract Disease*, vol. 13, no. 3, pp. 137–144, 2009.

[34] L. S. Massad, J. Jeronimo, and M. Schiffman, "Interobserver agreement in the assessment of components of colposcopic grading," *Obstetr. Gynecol.*, vol. 111, no. 6, pp. 1279–1284, 2008.

[35] W. Li, J. Gu, D. Ferris, and A. Poirson, "Automated image analysis of uterine cervical images," in *Proc. SPIE*, 2007, vol. 6514, p. 65 142P-1.

[36] G. Zimmerman-Moreno and H. Greenspan, "Automatic detection of specular reflections in uterine cervix images," in *Proc. SPIE Med. Imag.*, 2006, vol. 6144, p. 61 446E-61 446E-9.

[37] Z. Xue, S. Antani, R. Long, and G. Thoma, "Comparative performance analysis of cervix ROI extraction and specular reflection removal algorithms for uterine cervix image analysis," in *Proc. SPIE*, 2007, vol. 6512, pp. 4I1–4I9.

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

[39] Y. Srinivasan, B. Nutter, S. Mitra, B. Phillips, and E. Sinzinger, "Classification of cervix lesions using filter bank-based texture mode," in *Proc. 9th IEEE Int. Symp. Comput.-Based Med. Syst.*, 2006, pp. 832–840.

[40] B. Apgar, G. Brotzman, and M. Spitzer, *Colposcopy: Principles and Practice*. Philadelphia, PA: Saunders Elsevier, 2008.

[41] J. Jeronimo, L. R. Long, L. Neve, B. Michael, S. Antani, and M. Schiffman, "Digital tools for collecting data from cervigrams for research and training in colposcopy.," *J. Lower Genital Tract Dis.*, vol. 10, no. 1, pp. 16–25, 2006.

[42] R. Herrero, M. Schiffman, C. Bratti, A. Hildesheim, I. Balmaceda, and M. Sherman, "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, pp. 362–375, 1997.

[43] D. Song and J. Heflin, "Domain-independent entity coreference for linking ontology instances," *J. Data Inf. Quality*, vol. 4, no. 2, p. 7, 2013.

[44] C. Xiao, W. Wang, X. Lin, J. X. Yu, and G. Wang, "Efficient similarity joins for near-duplicate detection," *ACM Trans. Database Syst.*, vol. 36, no. 3, pp. 15:1–15:41, Aug. 2011.

[45] D. Song and J. Heflin, "Domain-independent entity coreference in RDF graphs," in *Proc. 19th ACM Conf. Inf. Knowl. Manage.*, 2010, pp. 1821–1824.

[46] D. Dey, V. S. Mookerjee, and D. Liu, "Efficient techniques for online record linkage," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 3, pp. 373–387, Mar. 2011.

[47] D. Song and J. Heflin, "Automatically generating data linkages using a domain-independent candidate selection approach," in *Proc. 10th Int. Semantic Web Conf.*, 2011, pp. 649–664.

[48] N. Aswani, K. Bontcheva, and H. Cunningham, "Mining information for instance unification," in *Proc. Int. Semantic Web Conf.*, 2006, pp. 329–342.

[49] X. Huang, W. Wang, Z. Xue, S. Antani, L. Long, and J. Jeronimo, "Tissue classification using cluster features for lesion detection in digital cervigrams," in *Proc. SPIE*, 2008, vol. 6914, p. 69 141Z-1.

[50] W. Wang, Y. Zhu, X. Huang, D. P. Lopresti, Z. Xue, L. R. Long, S. Antani, and G. R. Thoma, "A classifier ensemble based on performance level estimation," in *Proc. 2009 IEEE Int. Symp. Biomed. Imag., From Nano to Macro*, 2009, pp. 342–345.

[51] E. Kim, X. Huang, G. Tan, L. R. Long, and S. Antani, "A hierarchical SVG image abstraction layer for medical imaging," in *Proc. SPIE Med. Imag.*, 2010, vol. 7628, p. 762 809-762 809-9.

[52] E. Kim, S. Antani, X. Huang, L. R. Long, and D. Demner-Fushman, "Using relevant regions in image search and query refinement for medical CBIR," in *Proc. SPIE*, 2011, vol. 7967, p. 796 707-796 707-8.

[53] E. Kim and X. Huang, "A data driven approach to cervigram image analysis and classification," in *Color Medical Image Analysis*, ser. Lecture Notes in Comput. Vis. Biomechan., M. E. Celebi and G. Schaefer, Eds. Amsterdam, The Netherlands: Springer, 2013, vol. 6, pp. 1–13.

[54] S. Gordon, G. Zimmerman, R. Long, S. Antani, J. Jeronimo, and H. Greenspan, "Content analysis of uterine cervix images: Initial steps towards content based indexing and retrieval of cervigrams," in *Proc. SPIE*, 2006, vol. 6144, p. 61 444U-61 444U-8.

[55] H. Greenspan, S. Gordon, G. Zimmerman, S. Lotenberg, J. Jeronimo, S. Antani, and L. R. Long, "Automatic detection of anatomical landmarks in uterine cervix images," *IEEE Trans. Med. Imag.*, vol. 28, no. 3, pp. 454–468, Mar. 2009.

[56] A. Bosch, A. Zisserman, and X. Muñoz, "Representing shape with a spatial pyramid kernel," in *Proc. 6th ACM Int. Conf. Image Video Retriev.*, 2007, pp. 401–408.

[57] M. Tuceryan, "Moment-based texture segmentation," *Pattern Recognit. Lett.*, vol. 15, no. 7, pp. 659–668, 1994.

[58] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[59] A. Friedman, "Framing pictures: The role of knowledge in automatized encoding of memory for gist," *J. Exp. Psychol.: General*, vol. 108, no. 3, pp. 316–355, 1979.

[60] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[61] A. Divakaran, *Multimedia Content Analysis: Theory and Applications*. New York: Springer, 2009.

[62] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik, "Using contours to detect and localize junctions in natural images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[63] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi, "TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 1–15.

[64] T. M. Mitchell, *Machine Learning*, ser. McGraw Hill Series in Computer Science. New York: McGraw-Hill, 1997.

[65] D. W. Aha, "Lazy learning: Special issue editorial," *Artif. Intell. Rev.*, vol. 11, pp. 1–5, 1997.

[66] T. DeSantis, N. Chakhtoura, L. Twiggs, D. Ferris, M. Lashgari, L. Flowers, M. Faupel, S. Bambot, S. Raab, and E. Wilkinson, "Spectroscopic imaging as a triage test for cervical disease: A prospective multicenter clinical trial," *J. Lower Genital Tract Disease*, vol. 11, no. 1, pp. 18–24, 2007.

[67] S. Chang, Y. Mirabal, E. Atkinson, D. Cox, A. Malpica, M. Follen, and R. Richards-Kortum, "Combined reflectance and fluorescence spectroscopy for *in vivo* detection of cervical pre-cancer," *J. Lower Genital Tract Disease*, vol. 10, no. 2, p. 024031, 2005.

[68] N. Thekkek and R. Richards-Kortum, "Optical imaging for cervical cancer detection: Solutions for a continuing global problem," *Nature Rev. Cancer*, vol. 8, no. 9, pp. 725–731, 2008.

[69] S. Zhang, J. Huang, H. Li, and D. N. Metaxas, "Automatic image annotation and retrieval using group sparsity," *IEEE Trans. Syst., Man, Cybern., Part B*, vol. 42, no. 3, pp. 838–849, 2012.

[70] P. V. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 221–228.

[71] A. Alush, H. Greenspan, and J. Goldberger, "Automated and interactive lesion detection and segmentation in uterine cervix images," *IEEE Trans. Med. Imag.*, vol. 29, no. 2, pp. 488–501, Feb. 2010.

[72] S. Y. Park, D. Sargent, R. W. Lieberman, and U. Gustafsson, "Domain-specific image analysis for cervical neoplasia detection based on conditional random fields," *IEEE Trans. Med. Imag.*, vol. 30, no. 3, pp. 867–878, Mar. 2011.

[73] Y. Yu, J. Huang, S. Zhang, C. Restif, X. Huang, and D. N. Metaxas, "Group sparsity based classification for cervigram segmentation," in *Proc. 8th IEEE Int. Symp. Biomed. Imag.: From Nano to Macro*, 2011, pp. 1425–1429.

[74] S. Zhang, J. Huang, D. N. Metaxas, W. Wang, and X. Huang, "Discriminative sparse representations for cervigram image segmentation," in *Proc. IEEE Int. Symp. Biomed. Imag., From Nano to Macro*, 2010, pp. 133–136.

[75] S. Gordon and H. Greenspan, "Segmentation of non-convex regions within uterine cervix images," in *Proc. IEEE Int. Symp. Biomed. Imag., From Nano to Macro*, 2007, pp. 312–315.

[76] M. Batet, D. Sánchez, and A. Valls, "An ontology-based measure to compute semantic similarity in biomedicine," *J. Biomed. Inform.*, vol. 44, no. 1, pp. 118–125, 2011.

[77] J. Zhang and Y. Liu, "Cervical cancer detection using SVM based feature screening," in *Proc. MICCAI*, 2004, pp. 873–880.

[78] D. Saslow *et al.*, "American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology Screening Guidelines for the Prevention and Early Detection of Cervical Cancer," *CA: A Cancer J. Clin.*, vol. 62, no. 3, pp. 147–172, 2012.

[79] ACOG Committee on Practice Bulletins-Gynecology, "Acog practice bulletin no. 109: Cervical cytology screening," *Obstet. Gynecol.*, vol. 114, no. 6, pp. 1409–1420, 2009.

[80] D. Song and J. Heflin, "A pruning based approach for scalable entity coreference," in *Proc. 25th Int. Florida Artif. Intell. Res. Soc. Conf.*, 2012, pp. 98–103.

[81] D. Song, C. G. Chute, and C. Tao, "Semantator: Annotating clinical narratives with semantic web ontologies," in *Proc. AMIA Summits Transl. Sci.*, 2012, pp. 20–29.

[82] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. K. Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications," *J. Am. Med. Inform. Assoc.*, vol. 17, no. 5, pp. 507–513, 2010.

[83] C. Tao, D. Song, D. K. Sharma, and C. G. Chute, "Semantator: Semantic annotator for converting biomedical text to linked data," *J. Biomed. Inform.*, vol. 46, no. 5, pp. 882–893, 2013.

[84] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny, "Application of information technology: Medex: A medication information extraction system for clinical narratives," *J. Amer. Med. Inf. Assoc.*, vol. 17, no. 1, pp. 19–24, 2010.