

## A DECISION TREE-BASED APPROACH FOR CERVICAL SMEARS

CHING-CHENG SHEN<sup>1</sup>, HSU-HAO YANG<sup>2,\*</sup> AND YUEH-CHING CHANG<sup>1</sup>

<sup>1</sup>Graduate School of Information Management  
Vanung University  
No. 1, Van-nung Rd., Chung-li, Taoyuan 320, Taiwan  
james@mail.vnu.edu.tw

<sup>2</sup>Department of Industrial Engineering and Management  
National Chinyi University of Technology  
No. 57, Sec. 2, Zhongshan Rd., Taiping Dist., Taichung 41170, Taiwan

\*Corresponding author: yanghh@ncut.edu.tw

Received January 2011; revised June 2011

**ABSTRACT.** *Cervical smears are used to detect cervical intraepithelial neoplasia (CIN) and remain a popular method for the early detection of precancer and cervical cancer. One of the classification systems for CIN, the Bethesda system, classifies atypical cells into ASC-US (atypical squamous cells of undetermined significance) and ASC-H (atypical squamous cells: cannot exclude a high-grade squamous intra-epithelial lesion). This paper proposes a methodology involving a decision tree (DT) to identify cases of ASC-US by constructing a classification decision tree (CDT) based on samples from a Taiwanese teaching hospital. The main difference between our methodology and those of other studies is the use of a DT to identify ASC-US cases. We calculated the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) to evaluate the performance of our methodology and found a sensitivity of 86.36%, a specificity of 78.94%, a PPV of 90.47% and an NPV of 71.42%. The results indicate that our CDT is capable of detecting cases that are abnormal and positive; in contrast, detecting cases that are benign and negative remains a challenge. Moreover, our CDT performs well when predicting the cases that are abnormal and positive but performs poorly when predicting benign and negative cases. Considering both the sensitivity and PPV, we demonstrate that our CDT can help decrease the number of ASC-US cases.*

**Keywords:** Cervical smear, Pap test, Decision tree, Data mining, Cervical cancer

**1. Introduction.** Cervical cancer is the second largest cause of female cancer mortality worldwide; the estimated number of cervical cancer deaths per year is approximately 250,000 [1]. In Taiwan, cervical cancer is the third leading cause of death for females between 25 and 44 years of age and ranked second out of all cancers for female deaths in the late 1980s [2]. Cervical cancer did drop low as a cause of death in the 2000s; however, on average, the mortality rate remains at greater than 900 deaths per year. This high mortality rate prompted the Taiwanese government to promote a variety of programs over the past decade, which effectively reduced the mortality rate to 710 deaths in 2008. The programs ranged from encouraging females to undergo screening tests earlier (see details below), to encouraging regular follow-up examinations if the test results were inconclusive. Therefore, an accurate test not only reduces mortality but also decreases health-care costs.

The primary cause of cervical cancer is infection with any of a number of high-risk types of human papillomavirus (HPV). Most HPV infections regress spontaneously or are eliminated by the host immune system; however if left untreated, some infections may

lead to the development of cervical intraepithelial neoplasia (CIN) or precancer, which can lead to cervical cancer. As it usually takes more than 10 years for precancer lesions caused by HPV to develop into invasive cancer, most cervical cancers can be prevented by early detection and treatment of precancerous lesions.

CIN can be discovered through screening with the Papanicolaou test (also called Pap smear, Pap test, cervical smear or smear test), which can detect potentially precancerous changes in the ectocervix. Once detected, significant changes can be treated, and thus, cervical cancer can be prevented. Several systems have been designed for classifying precancerous conditions of the cervix based on cytology and histology. One such system, the Bethesda System (TBS), was originally developed in the 1990s. In a later improvement in 2001, TBS started classifying atypical cells as ASC-US (atypical squamous cells of undetermined significance) or ASC-H (atypical squamous cells: cannot exclude a high-grade squamous intra-epithelial lesion). This classification is also recommended by the World Health Organization (WHO) for cytological reports. Due to its indeterminate nature, an ASC-US case is commonly left untreated. However, according to previous research [3], roughly 12.7% of ASC-US cases develop into HSILs (high-grade squamous intraepithelial lesions) or SCC (squamous cell carcinoma), both of which are highly likely to eventually develop into cervical cancer. This high probability demonstrates the need to develop methods for correctly identifying ASC-US cases [4,5].

The objective of this study was to decrease the number of ASC-US cases by applying a decision tree (DT) [6]. We propose a methodology that constructs a classification decision tree (CDT) based on cervical smear samples, either positive or negative. With this CDT, we classify ASC-US samples and relabel the samples as positive if the CDT classifies them as positive. A DT is a logic structure that can model a decision analysis. Each DT starts with a root node, such as the attributes of positive samples, recursively generates branching nodes, and ends with leaf nodes specifying the samples as either positive or negative.

Before proceeding to report the details of the DT, we describe the unique features of our proposed methodology and the main advantages of the results. According to Ho et al. [7], the majority of previous studies on cervical cancer have focused on describing patient characteristics and genetic polymorphisms independently using statistical methods. To overcome the limitations of past studies, the authors used DTs to identify cervical cancer risk factors, such as demographic and environmental factors. The authors stated that their study was the first to use induction techniques to analyze risk factors for cervical cancer. Horng et al. [8] also studied environmental and host factors involved in the progression of HPV infection to HSIL and cervical cancer. In summary, others have used DTs to study cervical cancer, but their research has mainly focused on studying the risk factors involved. To the best of our knowledge, no previous research has attempted to resolve ASC-US cases using a DT. Moreover, from the perspective of managing health-care costs, the main advantage of our methodology lies in the elimination of a certain number of follow-up examinations that are not required if the ASC-US cases can be properly determined.

The remainder of this paper is organized as follows. Section 2 briefly introduces cervical smears. Section 3 describes our methodology and research framework. Section 4 describes the data and the construction of a CDT using a smaller sample size, and Section 5 discusses the empirical results. Section 6 presents the conclusions.

**2. Cervical Smears.** For a cervical smear, sample cells are collected from the outer opening of the cervix of the uterus and the endocervix using an extended-tip wooden spatula or brush. After the collection, the sample is smeared onto a glass slide, and

the slide is stained using the Papanicolaou technique, in which tinctorial dyes and acids are selectively retained by cells. Afterward, the slide is examined using a microscope to determine whether the cells are normal and to classify them appropriately, using a system such as TBS, mentioned earlier. Smears are evaluated in a laboratory by trained cytotechnicians under the supervision of a pathologist who is responsible for the reported results. Due to its convenience and simplicity, the cervical smear remains an effective and widely used method for the early detection of precancer and cervical cancer.

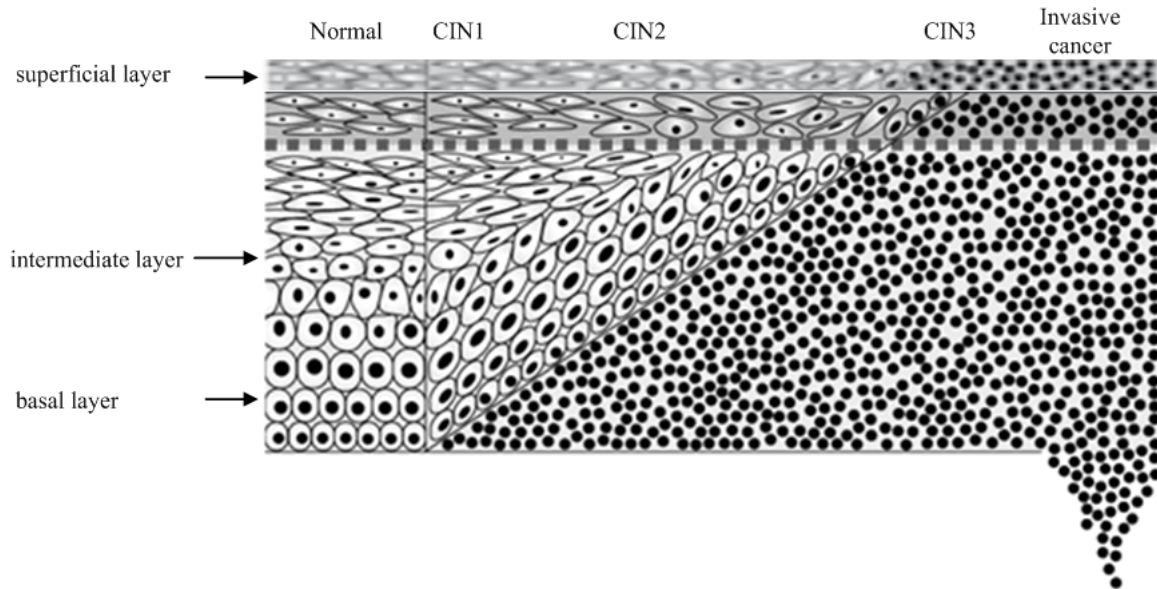


FIGURE 1. Progression from normal epithelium to invasive cancer

TABLE 1. Classifications of cervical smear according to the Department of Health of Taiwan

NEGATIVE FOR INTRAEPITHELIAL LESION AND MALIGNANCY	
Within normal limits .....	1
Reactive changes: Inflammation, repair, radiation, and others .....	2
Atrophy with inflammation .....	3
ATYPICAL SQUAMOUS CELLS	
Atypical squamous cells (ASC-US) .....	4
Atypical squamous cells – cannot exclude HSIL .....	16
LOW-GRADE SQUAMOUS INTRAEPITHELIAL LESION (LSIL)	
Mild dysplasia (CIN 1) with koilocytes .....	6
Mild dysplasia (CIN 1) without koilocytes .....	7
HIGH-GRADE SQUAMOUS INTRAEPITHELIAL LESION (HSIL)	
Moderate dysplasia (CIN 2) .....	8
Severe dysplasia (CIN 3) .....	9
Carcinoma in situ (CIN 3) .....	10
Dysplasia – cannot exclude HSIL .....	17
SQUAMOUS CELL CARCINOMA (SCC) .....	11

CIN, also known as cervical dysplasia, is the potentially premalignant transformation and abnormal growth (dysplasia) of squamous cells on the surface of the cervix. CIN is

commonly classified by grade: CIN 1, CIN 2 or CIN 3. The progression from normal epithelium to CIN 1, 2 and 3 and invasive cancer is shown in Figure 1 [9]. The vertical axis of Figure 1 represents the morphology of a given cell layer during this progression. As mentioned earlier, TBS is a system used for reporting cervical smear results. Some abnormal cervical smear results, as determined by TBS, include the following: atypical squamous cells, low-grade squamous intraepithelial lesions (LSILs) and HSILs. According to the WHO [9], LSIL is equivalent to CIN 1, and HSIL comprises the combination of CIN 2 and CIN 3. On the basis of the 2001 TBS, the Department of Health of Taiwan classifies the cervical smear as listed in Table 1. Throughout the paper, we will refer to positive samples as Items 6 and 7 (LSIL); 8, 9, 10 and 17 (HSIL) and 11 (SCC). Similarly, ASC-US samples are Items 4 and 16.

**3. Methodology.** A DT is a logic structure that can partition a large collection of records into smaller sets of records. The goal is to create a model for predicting the value of a target variable based on a set of input variables. In this tree, each node corresponds to one of the input variables and there are edges to the children of the node for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf. The tree recursively grows based on information gain; namely, at each node, the attribute with the most information gain is selected. The recursion is completed when the subset of a node has the same value as the target variable or when the splitting no longer adds value to the predictions.

The information gain is computed as follows. Given a dataset  $D$ , the *entropy* (impurity or disorder) of  $D$  is defined as:

$$Entropy(D) = - \sum_{i=1}^m p_i \log_2(p_i) \quad (1)$$

where  $p_i$  is the proportion of class  $i$  samples in  $D$  and  $m$  is the number of classes. The information gain of an attribute  $A$  is the expected reduction in entropy caused by splitting on this attribute:

$$Gain(D, A) = Entropy(D) - \sum_{j=1}^v \frac{|D_v|}{|D|} Entropy(D_v) \quad (2)$$

where  $D_v$  is the subset of  $D$  for which attribute  $A$  has value  $v$ , and  $|D_v|$  denotes the number of  $D_v$ . In this context, the entropy of the partitioned data is calculated by weighting the entropy of each partition by its size relative to the complete dataset.

A later version of the DT, C4.5 [10], contained a number of improvements. At each node of the tree, C4.5 chooses one attribute that most effectively splits the dataset into subsets by normalizing information gain. The attribute with the highest normalized information gain is the one chosen to split and is computed as follows:

$$Split(D, A) = \sum_{j=1}^v \frac{|D_v|}{|D|} \times \log_2 \left( \frac{|D_v|}{|D|} \right) \quad (3)$$

With  $Split(D, A)$ , we can define a gain ratio as follows:

$$GainRatio(D, A) = \frac{Gain(D, A)}{Split(D, A)} \quad (4)$$

In the following, we will select the variable with the highest  $GainRatio(D, A)$  to split a decision tree.

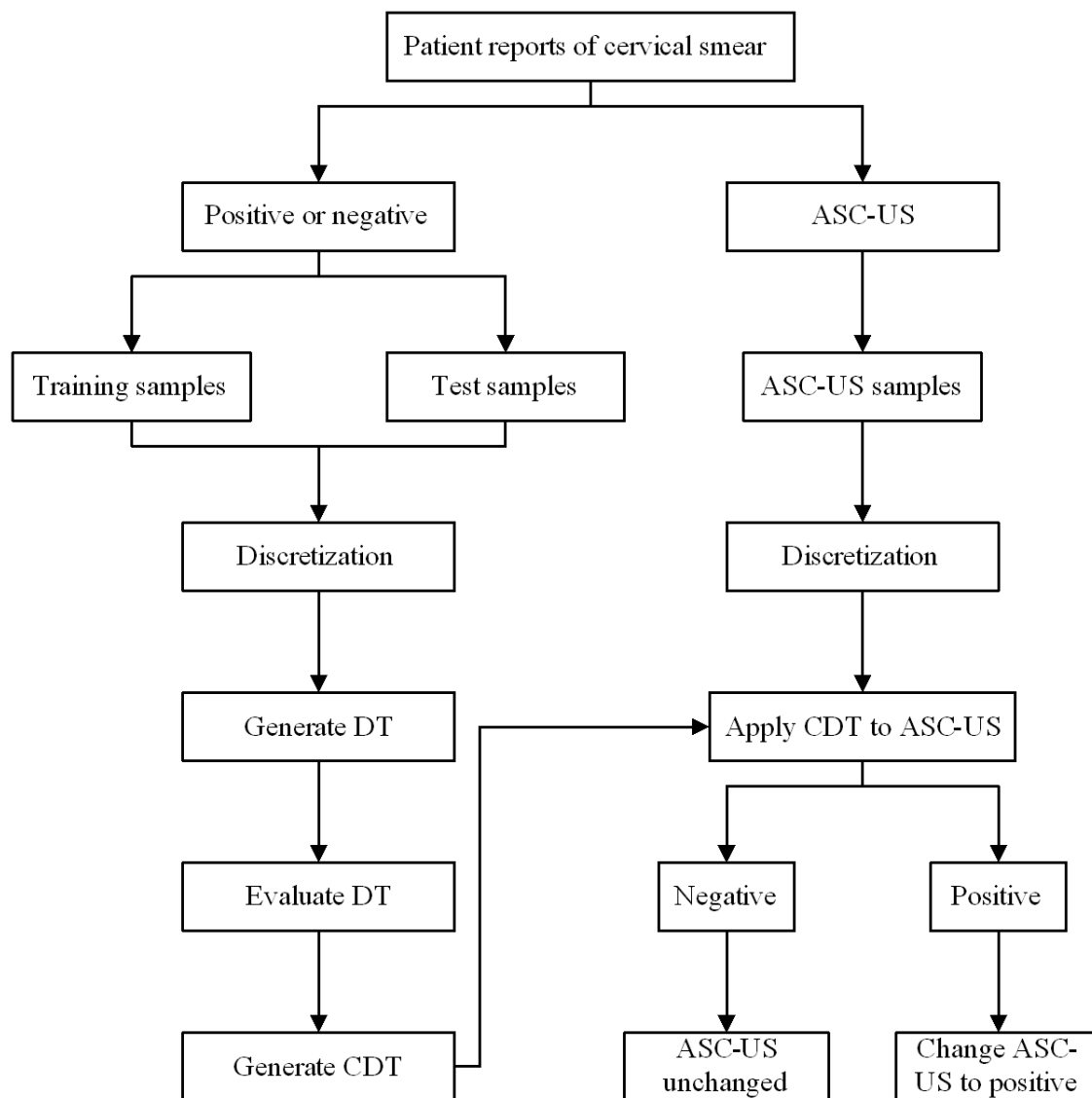


FIGURE 2. A flowchart of the methodology

As introduced in Section 1, our methodology involves constructing a CDT. A schematic of our methodology is depicted in Figure 2, but the methodology will be introduced in more detail below.

Despite the wide applications for DT, applying a DT to cervical cancer or cervical smears has occurred infrequently in reports such as those mentioned earlier [7,8]. Recall that Ho et al. [7] identified the combined patterns of cervical cancer risk factors using an induction technique and demonstrated how the DT could be used in risk analysis and target segmentation for cervical cancer management. Horng et al. [8] used a Bayesian network and four different DT algorithms, comparing the performance of these learning algorithms and claimed that the results could identify combinations of genetic factors that influence the risk associated with common complex multifactorial diseases such as cervical cancer. Readers are referred elsewhere for additional studies using DTs to investigate cancer [11-13].

The issue of cancer classification has also been addressed in the literature. Mohamad et al. [14] proposed a two-stage gene selection method to identify a smaller subset of informative genes that are most relevant for cancer classification. Later, Mohamad et

al. [15] improved their earlier study by proposing a three-stage selection method. The additional stage analyzes the frequency of the appearance of each gene in the subsets, producing a small subset of informative genes. Yeh et al. [16] proposed simplified swarm optimization for discovering breast cancer classification rules. The authors concluded that the proposed approach has potential application in hospital decision-making.

4. **Data.** We collected sample cells (referred to as samples thereafter) from one of the largest teaching hospitals in Taiwan. Samples were originally collected between 2006 and 2009, and the samples were verified by one pathologist and two cytotechnicians. Each sample was classified either as positive or negative, and the samples were divided into training samples and test samples. We now describe attribute definitions, followed by the construction of a DT using a small sample size.

4.1. **Attribute definition and clustering.** To classify a cervical smear sample appropriately, we consider the following four attributes: 1) nuclear chromatin characteristics, 2) nuclear membrane or nuclear envelope characteristics, 3) nuclear to cytoplasmic (N/C)

TABLE 2. Classes of nuclear chromatin and nuclear membrane

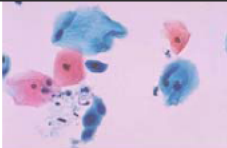
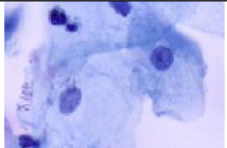
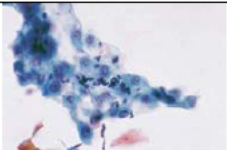
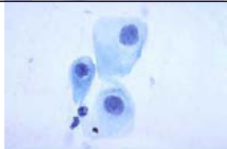

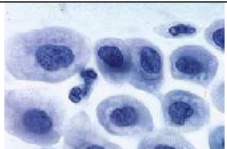
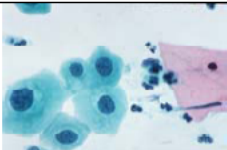

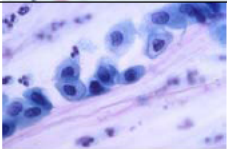
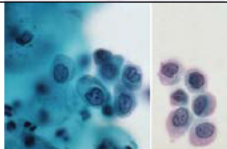
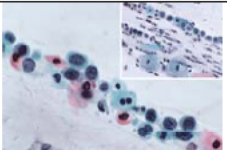
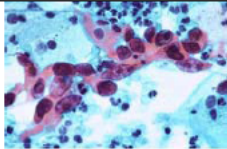
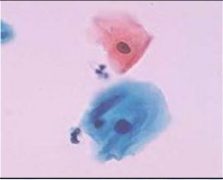
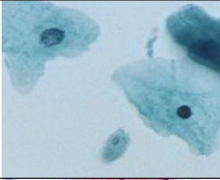
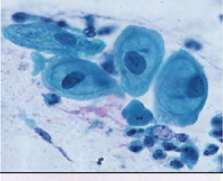
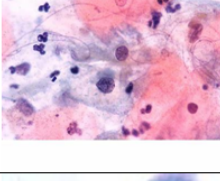
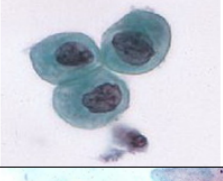
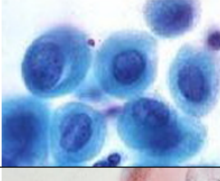
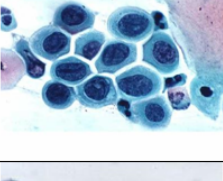
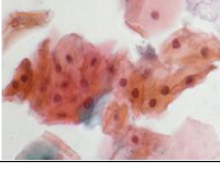
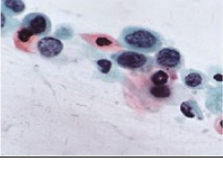
	Nuclear chromatin			Nuclear membrane	
1	Equally distributed		1	Smooth	
2	Equally distributed with small nuclei		2	Smooth and enlarged	
3	Equally distributed with dysplastic changes		3	Enlarged nuclei and nuclear contour irregularities	
4	Equally distributed and hyperchromatic		4	Slightly enlarged nuclei and binucleation	
5	Polygonal in shape with occasional nuclear contour irregularities		5, 6	Increased N/C ratios and nuclear contour irregularities	
6	Contour irregularities and a stream of mucus		7	Pleomorphism of size and shape	

TABLE 3. Classes of N/C ratio and cellular color

Nuclear to cytoplasmic (N/C) ratio		Cellular color			
1	$< 0.2$		1	Light green	
2	$\geq 0.2$ but $< 0.4$		2	Light orange	
3	$\geq 0.4$ but $< 0.6$		3	Green	
4	$\geq 0.6$ but $< 0.8$		4	Orange	
5	$\geq 0.8$ but $< 1.0$				

ratio and 4) cellular color. Nuclear chromatin is categorized into six classes according to its distribution and density. The nuclear membrane can fall into one of seven classes based on its roughness and smoothness. The nuclear to cytoplasmic ratio is a ratio of the size (i.e., volume) of the nucleus of a cell to the size of the cytoplasm of that cell. The nuclear to cytoplasmic ratio has five classes equally divide the range between zero and one. Cellular color is categorized into four classes based on cellular maturity and the type of cell. Detailed descriptions of each class are provided in Tables 2 and 3, and pictures are reported elsewhere [17]. All samples are discretized based on the attribute classes described above [18]. To enhance computational efficiency and reduce excessive branching of the DT, we cluster each attribute into three equal-depth partitions. This clustering method has been shown to be able to produce desirable results for single-layer clustering [19]. The three equal-depth clusters for each attribute are listed in Table 4, where a label such as C1 is associated with a cluster containing the value range discretized as described above.

TABLE 4. Three equal-depth clusters for each attribute

Nuclear chromatin		Nuclear membrane		N/C ratio		Cellular color	
0.000 - 1.030	C1	0.000 - 1.0035	M1	0.000 - 1.005	R1	0.000 - 1.010	L1
1.031 - 1.100	C2	1.0036 - 1.009	M2	1.005 - 1.021	R2	1.011 - 1.020	L2
$> 1.100$	C3	$> 1.009$	M3	$> 1.021$	R3	$> 1.021$	L3

4.2. **Constructing trees using small sample size.** As a preliminary step to evaluate the performance of our proposed methodology, we initially tested on a small set of samples. For this preliminary evaluation, we used several confirmed cases as training samples and generated a DT. Afterward, we applied the tree to test samples and calculated the classification accuracy to evaluate whether the tree could be used as the CDT mentioned in Section 1. The classification accuracy is the ratio of items that are classified into the correct classes. Finally, we used the CDT to test some ASC-US samples; if the sample was classified as positive, it was relabeled as positive. We simply used the classification accuracy as the criterion at this step, additional performance criteria will be introduced in Section 5.

TABLE 5. Information gain and gain ratio for each attribute using a small sample size

	Gain (D, A)	Split (D, A)	GainRatio (D, A)
Nuclear chromatin	0.226	1.214	0.1859
Nuclear membrane	0.192	1.496	0.1286
N/C ratio	0.035	1.579	0.0220
Cellular color	0.001	1.546	0.0007

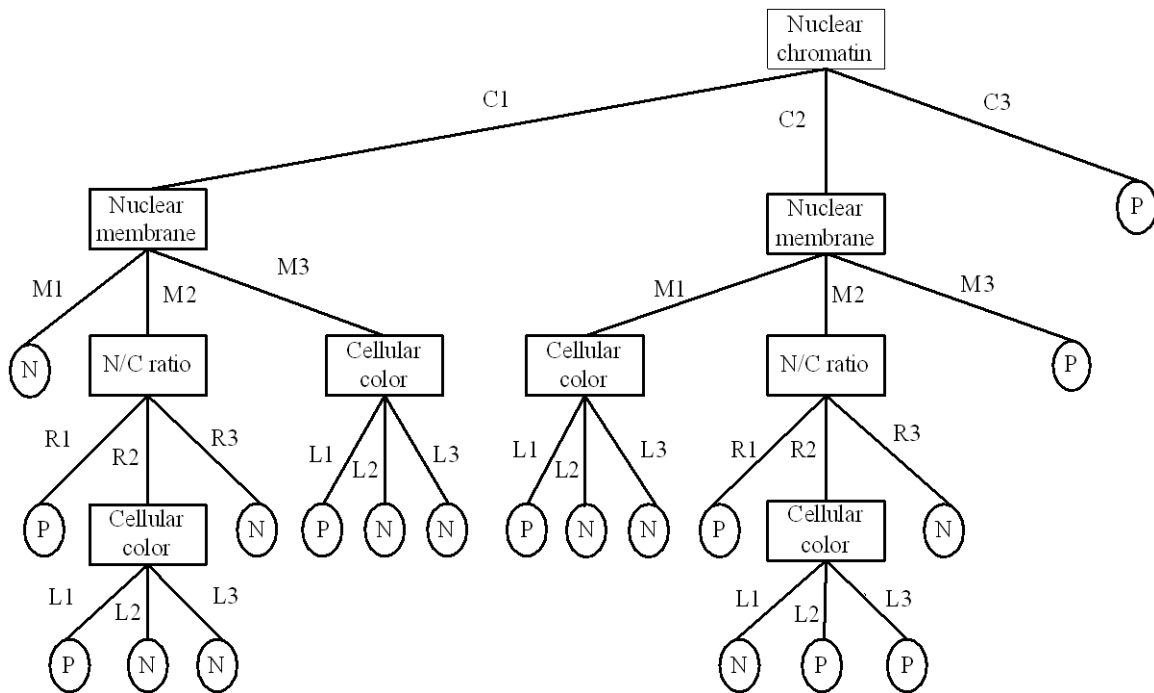


FIGURE 3. Classification decision tree (CDT) with a small sample size

After using the training samples with 29 positives and 29 negatives, we computed the gain ratio for each attribute defined in Equation (4); all gain ratios are listed in Table 5. According to Table 5, nuclear chromatin had the highest value, and cellular color had the lowest one. Therefore, the root of the CDT started with nuclear chromatin; the underlying tree is shown in Figure 3, where the label next to an edge corresponds to the cluster label defined in Table 4, and a leaf node with “P” or “N” means positive or negative, respectively. For example, consider the leftmost branch of Figure 3. The branch means that if nuclear chromatin is in cluster C1 and nuclear membrane is in cluster M1,



the sample is determined to be negative. We continued to test another 30 samples, 15 positives and 15 negatives, and the classification accuracy was found to be 93.3%, with only two positive samples misclassified as negatives. Because the accuracy is as high as 93.3%, the underlying tree can be referred to as a CDT. With the CDT in place, we proceeded to test five ASC-US samples. After using the CDT, two were identified as positive, and three remained classified as ASC-US. Despite the small sample size, the outcome was encouraging because two out of five ASC-US cases were relabeled as positive.

**5. Empirical Results.** In Section 4, we tested a smaller set of samples and found that some ASC-US cases were relabeled as positive. In this section, we describe a second test with a larger sample size and examine whether the size of the sample affects the outcome. We begin by describing the performance criteria.

**5.1. Performance criteria for the CDT.** In addition to the classification accuracy mentioned in Section 4, we used *sensitivity*, *specificity*, *positive predictive value* (PPV) and *negative predictive value* (NPV) to evaluate the performance of the CDT. Sensitivity refers to the ability of the test to correctly identify individual cases with conditions such as precancer. The higher the sensitivity, the fewer cases with precancer that will be incorrectly identified as normal. Specificity refers to the ability of the test to correctly identify individual cases without precancer. The higher the specificity, the fewer cases with normal cervix tissue will be incorrectly identified as precancerous. An ideal test would have both a high sensitivity and a high specificity. The sensitivity and specificity are computed as follows:

$$\begin{aligned} \text{sensitivity (\%)} &= \text{TP}/(\text{TP} + \text{FN}); \\ \text{specificity (\%)} &= \text{TN}/(\text{FP} + \text{TN}), \text{ where} \\ \text{TP} &= \text{true positive,} \\ \text{TN} &= \text{true negative,} \\ \text{FP} &= \text{false positive, and} \\ \text{FN} &= \text{false negative.} \end{aligned}$$

The definitions of TP, TN, FP and FN are given in Table 6, where a patient report is referred to as *benign* if it classifies a diagnosis as normal and as *abnormal* if it classifies a diagnosis as an LSIL, an HSIL or SCC. Finally, one may also want to know the likelihood of actually having the disease when the test turns out to be positive. This likelihood is referred to as the *positive predictive value* (PPV) of the test. In contrast, the *negative predictive value* (NPV) is the chance of not having the disease when the test returns a negative result.

TABLE 6. Definitions of TP, FP, TN and FN

Category	Patient report	DT report
True positive (TP)	Abnormal	Positive
False positive (FP)	Benign	Positive
True negative (TN)	Benign	Negative
False negative (FN)	Abnormal	Negative

5.2. **Construction of CDT.** The larger set of training samples contained 250 cases, 125 positives and 125 negatives; the gain ratios and the CDT derived from the samples are given in Table 7 and Figure 4, respectively. Again, nuclear chromatin had the highest gain ratio, and cellular color had the lowest. Moreover, comparing Figure 4 to Figure 3, we observe that both trees differ marginally in terms of the structure, implying that the size of the samples may not be critical as long as the size is sufficiently large. The test samples included 100 total samples, 50 positives and 50 negatives. After applying the CDT, the classification accuracy was 91.0% (91/100). In the nine misclassified cases, five were negative but misclassified as positive, and four were positive but misclassified as negative. The sensitivity and specificity were 92.0% and 90.0%, respectively. A previous study [20] has shown that the sensitivity of the Papanicolaou smear test is between 30% and 87% and that the specificity is between 86% and 100%. In comparison, our CDT generates a considerably higher sensitivity and a specificity that equals that found in the previous study.

TABLE 7. Information gain and gain ratio for each attribute using a large sample size

	Gain (D, A)	Split (D, A)	GainRatio (D, A)
Nuclear chromatin	0.358	1.186	0.3017
Nuclear membrane	0.334	1.557	0.2145
N/C ratio	0.045	1.574	0.0284
Cellular color	0.017	1.565	0.0106

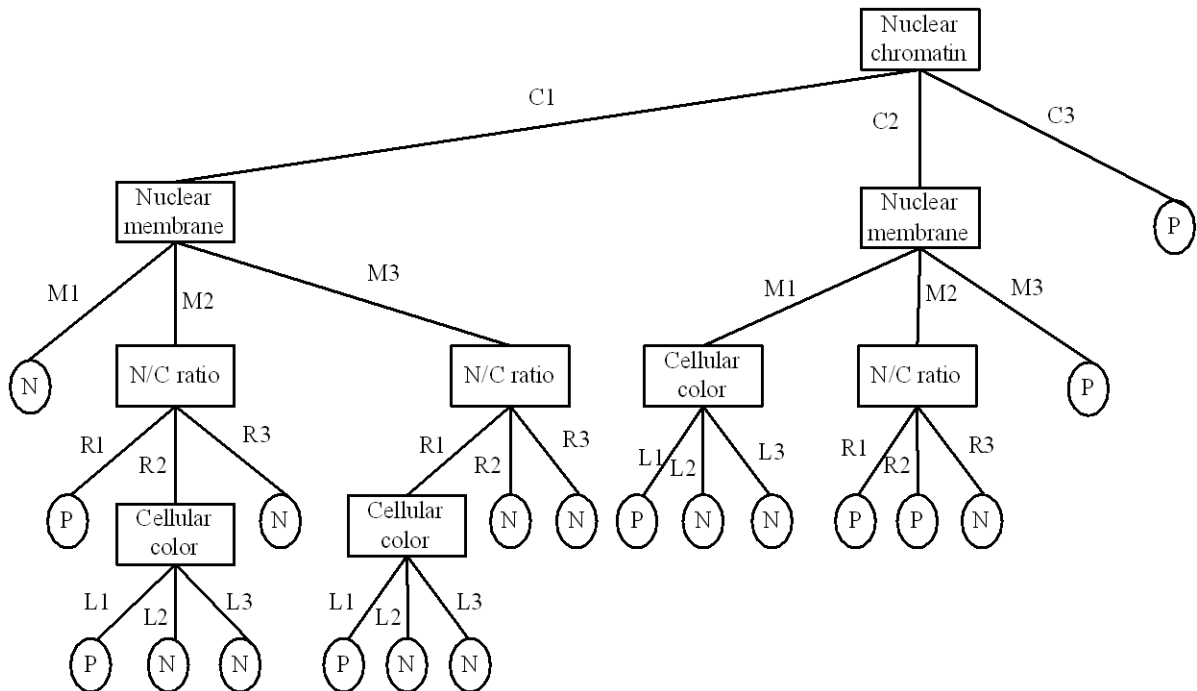


FIGURE 4. Classification decision tree (CDT) with a large sample size

5.3. **Performance of the CDT and discussion.** To evaluate how well our CDT performs, we used 63 ASC-US samples for which patient reports were available. Our discussions are based on the performance criteria introduced above, namely, sensitivity, specificity, NPV, and PPV. After applying the CDT to the 63 ASC-US samples, we classified

42 as positive and 21 as negative, indicating that the number of ASC-US cases was significantly overestimated. However, the results must be further verified against the patient reports before we can claim that the decrease is indeed an advantage of the proposed methodology. Table 8 lists the sensitivity, specificity, PPV and NPV of the CDT.

TABLE 8. Sensitivity, specificity, PPV and NPV for the CDT

CDT	Patient reports	
	Abnormal	Benign
Positive = 42	38	4
Negative = 21	6	15
Sensitivity = $38/44 = 86.36\%$		
Specificity = $15/19 = 78.94\%$		
Positive predictive value (PPV) = $38/42 = 90.47\%$		
Negative predictive value (NPV) = $15/21 = 71.42\%$		

Siddiqui et al. [5] have reported that the sensitivity range is between 62.73% and 98.04% and that the specificity range is between 73.15% and 92.22%. In this respect, the sensitivity of our CDT is relatively high, but its specificity has room for improvement. High sensitivity represents a better capability to detect the cases that are abnormal and positive, while lower specificity represents the challenge inherent to detecting cases that are benign and negative. In other words, high sensitivity enables the CDT to meet our goal of decreasing the number of ASC-US samples. With respect to PPV and NPV, the former is 90.47% and the latter is 71.42%, as shown in Table 8. Using other methods, the PPV commonly ranges from 40.01% to 96.47% and the NPV ranges from 55.87% to 99.95% [5]. By these standards, our CDT has a relatively high PPV and a somewhat low NPV. The implication is that our CDT excels in predicting cases that are abnormal and positive but poorly predicts benign and negative cases. A high PPV also helps the CDT decrease the number of ASC-US samples. In conclusion, the use of CDT is justified due to its high accuracy, high sensitivity and high NPV.

**6. Conclusions.** This paper applied a DT to determine ASC-US cases by constructing a CDT based on the cervical smear samples from a teaching hospital in Taiwan. We collected samples classified as positive or negative, divided the samples into a training set and test set and used ASC-US cases with available patient reports to verify the results. To evaluate how our CDT performed, we analyzed the sensitivity, specificity, positive predictive value and negative predictive value.

According to the empirical results, the sensitivity is 86.36%, the specificity is 78.94%, the positive predictive value is 90.47% and the negative predictive value is 71.42%. In terms of medical perspectives, our CDT has a better capability to detect cases that are abnormal and positive, whereas it is challenging for our CDT to detect cases that are benign and negative. In summary, considering the sensitivity and positive predictive value, our CDT can help decrease the number of samples classified as ASC-US. In terms of managerial implications, this paper demonstrates that the DT can be a useful tool to help identify more positive ASC-US cases simply with available cervical smear samples, without purchasing additional equipment.

Despite the encouraging results from our CDT, some concerns may be raised about the computation in our methodology. Indeed, our conclusions are based on a medium sample size and we only implemented the clustering method [19] to enhance the computational efficiency. Due to the size of this sample, we did not address the issue of computational

complexity, nor did we consider other pruning mechanisms that are widely used for enhancing the efficiency of DT. Additionally, we did not discuss how our CDT deals with missing values or outliers. In terms of these issues, our methodology leaves room for improvement.

We have also mentioned that no previous research has applied a DT to determine ASC-US cases. In this respect, our selection of the four attributes used to classify a sample, our categorization of the attributes, and our discretization of the samples was successful, but not without limitations. That is, our results are based on this selection, categorization and discretization. Future research should investigate how different attributes or categorizations of the attributes affect the results.

**Acknowledgment.** The first author was supported in part by grant number NSC 99-2410-H-238-005. The second author was supported in part by grant number NSC 99-2410-H-167-006-MY2. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

## REFERENCES

- [1] [http://www.who.int/vaccine\\_research/diseases/viral\\_cancers/en/index3.html](http://www.who.int/vaccine_research/diseases/viral_cancers/en/index3.html), 2011.
- [2] [http://www.doh.gov.tw/EN2006/DM/DM2\\_p01.aspx?class\\_no=390&now\\_fod\\_list\\_no=10864&level\\_no=2&doc\\_no=75601](http://www.doh.gov.tw/EN2006/DM/DM2_p01.aspx?class_no=390&now_fod_list_no=10864&level_no=2&doc_no=75601), 2011.
- [3] B. Limpvanuspong, S. Tangjitgamol, S. Manusirivithaya, J. Khunnarong, T. Thavaramara and S. Leelahakorn, Prevalence of high grade squamous intraepithelial lesions (HSIL) and invasive cervical cancer in patients with atypical squamous cells of undetermined significance (ASCUS) from cervical pap smears, *Southeast Asian J. Trop Med. Public Health*, vol.39, no.4, pp.737-744, 2008.
- [4] J. A. Tworek, B. A. Jones, S. Raab, K. M. Clary and M. K. Walsh, The value of monitoring human papillomavirus DNA results for Papanicolaou tests diagnosed as atypical squamous cells of undetermined significance, *Arch. Pathol. Lab. Med.*, vol.131, no.10, pp.525-531, 2007.
- [5] M. T. Siddiqui, K. Hornaman, C. Cohen and A. Nassar, ProEx C immunocytochemistry and high-risk human papillomavirus DNA testing in papanicolaou tests with atypical squamous cell (ASC-US) cytology, *Arch. Pathol. Lab. Med.*, vol.132, no.10, pp.1648-1652, 2008.
- [6] J. R. Quinlan, Discovering rules by induction from large collections of examples, in *Expert Systems in the Micro-electronic Age*, D. Michie (ed.), Edinburgh, Edinburgh University Press, 1979.
- [7] S. H. Ho, S. H. Jee, J. E. Lee and J. S. Park, Analysis on risk factors for cervical cancer using induction technique, *Expert Syst. Appl.*, vol.27, no.1, pp.97-105, 2004.
- [8] J. T. Horng, K. C. Hu, L. C. Wu, H. D. Huang, F. M. Lin, S. L. Huang, H. C. Lai and T. Y. Chu, Identifying the combination of genetic factors that determine susceptibility to cervical cancer, *IEEE Trans. Inf. Technol. Biomed.*, vol.8, no.1, pp.59-66, 2004.
- [9] [http://whqlibdoc.who.int/publications/2006/9241547006\\_eng.pdf](http://whqlibdoc.who.int/publications/2006/9241547006_eng.pdf), 2010.
- [10] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.
- [11] G. Ge and G. W. Wong, Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles, *BMC Bioinformatics*, vol.9, pp.275-287, 2008.
- [12] Y. Su, J. Shen, H. Qian, H. Ma, J. Ji, H. Ma, L. Ma, W. Zhang, L. Meng, Z. Li, J. Wu, G. Jin, J. Zhang and C. Shoul, Diagnosis of gastric cancer using decision tree classification of mass spectral data, *Cancer Sci.*, vol.98, no.1, pp.37-43, 2007.
- [13] A. Vlahou, J. O. Schorge, B. W. Gregory and R. L. Coleman, Diagnosis of ovarian cancer using decision tree classification of mass spectral data, *J. Biomedicine and Biotechnology*, vol.5, pp.308-314, 2003.
- [14] M. S. Mohamad, S. Omatu, M. Yoshioka and S. Deris, A two-stage method to select a smaller subset of informative genes for cancer classification, *International Journal of Innovative Computing, Information and Control*, vol.5, no.10(A), pp.2959-2968, 2009.
- [15] M. S. Mohamad, S. Omatu, M. Yoshioka and S. Deris, A three-stage method to select informative genes for cancer classification, *International Journal of Innovative Computing, Information and Control*, vol.6, no.1, pp.117-125, 2010.

- [16] W.-C. Yeh, W.-W. Chang and C.-W. Chiu, A simplified swarm optimization for discovering the classification rule using microarray data of breast cancer, *International Journal of Innovative Computing, Information and Control*, vol.7, no.5(A), pp.2235-2246, 2011.
- [17] NCI Bethesda system web atlas, in *The Bethesda System for Reporting Cervical Cytology*, D. Solomon and R. Nayar (eds.), New York, Springer-Verlag, 2004.
- [18] Y. C. Chang, *Increase Positive Screen Ratio of Cervical Smear Using Decision Tree*, Master Thesis, Vanung University, Taoyuan, Taiwan, 2009.
- [19] C. C. Shen and Y. L. Chen, A dynamic-programming algorithm for hierarchical discretization of continuous attributes, *Eur. J. Oper. Res.*, vol.184, no.2, pp.636-651, 2008.
- [20] L. Nanda, D. C. McCrory, E. R. Myers, L. A. Bastain, V. Hasselblad, J. D. Hickery and D. B. Matchar, Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytological abnormalities: A systematic review, *Ann. Intern. Med.*, vol.132, no.10, pp.810-819, 2000.