Available online at www.ijpe-online.com

International Journal of Performability Engineering

Vol. 13, No. 4, July 2017, pp. 446-457
DOI: 10.23940/ijpe.17.04.p12.446457

Cervical Cancer Diagnosis based on Random Forest

Guanglu Sun[a,b], Shaobo Li[a], Yanzhen Cao[a] , Fei Lang[b,*]

[a]*School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China*
[b]*Research Center of Information Security & Intelligent Technology, Harbin University of Science and Technology, Harbin, 150080, China*

Abstract

Cervical cancer, with an annually increasing incidence rate, is becoming the leading cause of death among women in China. However, studies have shown that the early detection and accurate diagnosis of cervical cancer contribute to the long survival of cervical cancer patients. The machine learning method is a good substitute for manual diagnosis in the analysis of Pap smear cervical cell images, reflecting its effective and accurate classification. In the present study, a framework for cervical cancer diagnosis is presented based on a random forest (RF) classifier with ReliefF feature selection. Using preprocessing, segmentation, and feature extraction, 20 features were extracted. In the feature selection phase, 20 features were ranked according to weight using ReliefF. In the classification phase, the RF method was used as a classifier, and different dimensions of features were selected to train the classifier. To examine the efficacy of the proposed method, the Herlev data set collected at Herlev University Hospital was used, in which 917 Pap smear images were categorized into two classes: normal and abnormal. After a 10-fold cross validation, the experimental results showed that the best classification performance was obtained with the top 13 features based on the RF classifier, which were better than Naive Bayes, C4.5, and Logistic Regression. The accuracy was 94.44%, and the AUC value was 0.9804. The results also confirmed the effectiveness of cytoplasm features in the classification.

Keywords: Cervical cancer; Machine learning; ReliefF; Random forest

(Submitted on January 29, 2017; Revised on April 12, 2017; Accepted on June 23, 2017)

1. Introduction

Cervical cancer is the fourth leading cause of cancer death in women, and its incidence rate is the second highest after breast cancer. Particularly in China, the number of women suffering from cervical cancer is rapidly increasing [5,19,44]. Recently, approximately 500,000 new cases of cervical cancer have been reported annually worldwide, and approximately 28.8% of these cases have occurred in China [12]. Thus, there is much concern for the prevention of cervical cancer.

Early diagnosis and treatment of cervical cancer can improve the survival rate. There are two main methods for the screening of cervical cancer, including cytological examination and Human Papilloma Virus (HPV) testing [8,24]. Cytological examination is a widely used method. The main method in cytological examination is a Pap smear test, which is the most common method used to detect cervical cancer [44]. However, the limitation of the Pap smear test is that abnormal cells have to be manually detected among a large number of cells using a microscope. Thus, the accuracy of the analysis is not always guaranteed as a result of technical and human errors. Therefore, manual diagnosis has gradually been replaced by automated screening methods [37,42].

For the detection of cervical cancer cells, automated and semi-automated methods have been applied to select abnormal cells from the cervical cell images [7,15,37,42]. Machine learning methods are used to combine the diagnostic experience of

* Corresponding author. Tel.: +86-451-86390657.
E-mail address: langfei@hrbust.edu.cn.

pathology experts with the accurate calculation and rapid processing abilities of computer systems [3,40]. Many studies have focused on image segmentation and feature extraction, in which effective algorithms have been proposed [16,47]. In the present study, feature selection and classification are primarily discussed. Feature selection is vital for building a classification model with high predictive accuracy by selecting more effective features. It is necessary to select high-level features using ReliefF [20], which provides a ranking result to select high-level features for data with few features. The RF classifier is applied with the ReliefF feature selection method. RF [43] is used to classify cells into two classes, which is one of the most popular models in the field of bioinformatics and biomedical research. RF overcomes the overfitting problem and shows adequate noise tolerance.

The remainder of the present study is structured as follows: in section 2, methods and results of previous studies on cervical cancer diagnosis are summarized. In section 3, a detailed introduction is presented to explain the benchmark data set, the framework of the proposed method, and the methods of feature selection and classification. In section 4, evaluation metrics, the experimental setup and results, and the discussion are presented. In section 5, the conclusion is given.

2. Related Work

The method to detect cervical cancer using image analysis can be divided into five steps: image preprocessing, image segmentation, feature extraction, feature selection, and pattern classification.

Several studies have focused on image preprocessing, image segmentation, and feature extraction. Hummel [16] utilized an adaptive histogram equalization method to improve the image contrast, which was the preliminary work for image segmentation. Plissiti et al. [36] presented two methods for the segmentation of cell nuclei. One method involved the automated detection of cell nuclei in Pap smear images using morphological reconstruction, and the other method involved the combination of watershed-based and snake segmentation methods. Li et al. [25] applied K-means to coarsely segment the cytoplasm, nuclei, and background and presented an improved gradient vector flow model (R-GVF) for fine segmentation. Furthermore, many multi-dimensional feature extraction methods have been proposed [6,23,31].

Feature selection and classification are also critical issues. Martin [30] applied Fuzzy C-means and Gustafson-Kessel clustering to the classification of cervical cells and compared the performance of two methods. The results showed that Fuzzy C-means generated the best classification results and showed the highest robustness against noise. Norup [33] examined different classification methods for the classification of cervical cells. These methods included linear least square networks, K-nearest neighbor (KNN), weighted KNN (WKNN), the advanced method, which was Neuro-Fuzzy Inference for Transductive Reasoning (NFI), and the proposed method, named Nearest Class Center of gravity (NCC). The experimental results showed that the NCC method performed better among the five classifiers and achieved an overall error of 5.13%. Marinakis et al. [32] proposed an effective genetic algorithm scheme for feature selection, which was combined with a number of nearest neighbor-based classifiers. The results showed that the algorithm provided higher classification accuracy than the other algorithms used for comparison. Gençtav et al. [11] proposed an unsupervised approach for the segmentation and classification of cervical cells. The classification was posed as a grouping problem, ranking the cells based on feature characteristics modeling abnormality degrees. Chankong et al. [18] proposed the automatic segmentation and classification method for cervical cancer cells and compared this method with hard C-means clustering and watershed techniques. The results showed that the proposed method is better than its counterparts. Mbaga et al. [28] used SVM-RFE for feature selection and SVM for classification in Pap smear images. To this end, these authors compared the performance of classifiers with radial basis and polynomial kernel functions. The results showed that polynomial kernel function exhibited better performance in accuracy, sensitivity and specificity. Plissiti et al. [35] showed that the classification of cervical cells into normal and abnormal categories is more efficient when based on features extracted exclusively from the nucleus and ignoring the contingent cytoplasm features. These authors also verified that the results using only the nucleus features were better than those using both nucleus and cytoplasm features. In short, although many studies have been conducted for feature selection and classification, few studies have applied ReliefF and RF in the field of cervical cancer cell detection.

In the present study, ReliefF [39] and RF [4] were used to improve the performance of feature selection and classification, and these two methods have been demonstrated as effective and have subsequently been applied in many fields. Moore et al. [29] applied ReliefF to conduct genome-wide genetic analysis for feature selection. Kandaswamy et al. [21] used ReliefF and support vector machine for the prediction of bioluminescent proteins. Saha et al. [41] used ReliefF for automatic gesture recognition for health care, while Fuzzy KNN was used as a classification method. Peker et al. [34] applied ReliefF for feature selection, and combined this technique with a variety of other classification methods. The best result was obtained from the combination of ReliefF with random forest.

Random forest has been widely applied in fields of text categorization, language modeling, economics, finance, and medical research and achieved good results. D áz-Uriarte et al. [9] investigated the use of random forest for the classification of microarray data and proposed a new method of gene selection in classification problems based on random forest. Albert et al. [1] proposed a method for applying the random forest method for the imaging atmospheric Cherenkov telescope MAGIC. Wu et al. [45] studied the prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. Khalilia et al. [22] predicted disease risks based on random forest and obtained good results from highly imbalanced data. Lin et al. [26] combined random forest with a gray model for the identification of DNA-binding proteins. Mohan et al. [27] classified two data sets, namely SCOP and Pfam, using several machine learning algorithms trained with physicochemical parameters, and random forest performed best among all classifiers examined.

3. Method

3.1. Data Description

In the present study, the methodology was verified using the Herlev data set. The data set was collected at the Department of Pathology of Herlev University Hospital and the Department of Automation at the Technical University of Denmark, comprising 917 images of a single Pap cell [30,33]. Skilled cyto-technicians obtained the images using a microscope connected to a frame grabber. Through manual classification, the images were divided into two different classes, normal and abnormal cells. The normal cells were further divided into three sub-classes, the abnormal cells were also divided into four sub-classes, as shown in Table 1. To ensure the classification validity, every image was classified by two different cyto-technicians.

3.2. Framework

The random forest classifier was combined with the ReliefF feature selection method to generate a reliable classification framework for the cervical cell images. Figure 1 shows the framework, including preprocessing, segmentation, feature extraction, feature selection, and classification, of which feature selection and classification are the focus in this paper. The single cell images analyzed were prepared by cyto-technicians at Herlev University Hospital using the CHAMP† system for segmenting images, and the features were subsequently extracted. The feature selection and classification methods used in the present study were ReliefF and random forest, respectively.

Table 1. The distribution of different cells in the Herlev data set

Classification	Name	Number
Normal cells (242)	Superficial squamous epithelial	74
	Intermediate squamous epithelial	70
	Columnar epithelial	98
Abnormal cells (675)	Mild squamous non-keratinizing dysplasia	182
	Moderate squamous non-keratinizing dysplasia	146
	Severe squamous non-keratinizing dysplasia	197
	Squamous cell carcinoma in situ intermediate	150

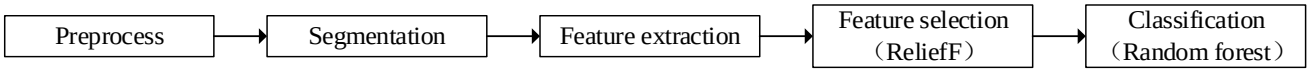

Figure 1. The framework of cell image classification

3.3. Preprocessing and Segmentation

The cell images are first preprocessed, including gray-scale and denoising. The RGB images are transformed into grayscale images by collapsing the three color channels into one channel and then passing the images through a median filter to remove the unwanted noise. The median of the intensity levels of the pixel neighborhood subsequently replaces the gray level of every pixel while preserving edge sharpness.

Cyto-technicians have divided every image into three parts: background, cytoplasm, and nuclei, using the CHAMP system. The segmentation results were assessed by cyto-technicians. The non-segmented and segmented cell images are in Figure 2.

† CHAMP is a commercial image-processing software for medical purposes by dimac-imaging, http://www.dimac-imaging.com/sider /products/champ.htm

3.4. Feature Extraction

Feature extraction converts image information into a format suitable for the classifier. Here, the chain code of the image edge representation is utilized in feature extraction as an effective coding method for edge representation. A total of 11 nucleus features and 9 cytoplasm features were extracted from a combination of the segmented and non-segmented cell images using MATLAB programs written according to Martin with MATLAB v6.5 [30], and applied in the present study. The features are shown in Table 2.

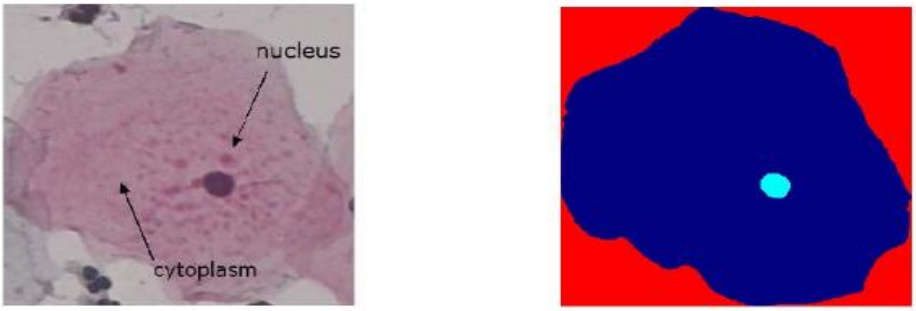

Figure 2. The original cell image and the segmented cell image

Table 2. The distribution of different cells in the Herlev data set

Nucleus features	Cytoplasm features
Nucleus area	Cytoplasm area
Nucleus brightness	Cytoplasm brightness
Nucleus short diameter	Cytoplasm short diameter
Nucleus longest diameter	Cytoplasm longest diameter
Nucleus elongation	Cytoplasm elongation
Nucleus roundness	Cytoplasm roundness
Nucleus perimeter	Cytoplasm perimeter
Maxima in nucleus	Maxima in cytoplasm
Minima in nucleus	Minima in cytoplasm
Nucleus relative position	
Nucleus/Cytoplasm ratio	

3.5. Feature Selection

Feature selection chooses an optimum subset of features through the removal of unhelpful features for classification. Typically, feature selection improves the performance of the classifier prior to classification. ReliefF is an effective feature selection method applied in many fields. This technique has low computational complexity and high robustness, which achieves the goal of reducing the dimensions of features [20].

Assume an instance space of $X=\{x_1, x_2, \ldots, x_n\}$ from the cervical cells, where n is the number of the instances, and $Y=\{y_1, y_2, \ldots, y_m\}$ is the class space, and m is the number of the classes. The algorithm of ReliefF is shown in Algorithm 1.

ReliefF randomly selects an instance x_i and subsequently searches for k of its nearest neighbors from the same class, called nearest hit H_j, and k nearest neighbors from each of the different classes, called nearest misses $M_j(C)$. This method also updates the quality estimation $W[A]$ for all attributes A depending on their values for x_i, hits H_j and misses $M_j(C)$. The weight of each feature is obtained after repeating the above process. The features are selected based on the ranking of the weight. Function $diff(A, I_1, I_2)$ calculates the difference between the values of feature A for two instances I_1 and I_2. If A is a nominal feature, it is defined as:

$$diff(A, I_1, I_2) = \begin{cases} 0 & I_1[A] = I_2[A] \\ 1 & I_1[A] \neq I_2[A] \end{cases} \quad (1)$$

If A is a numerical feature, it is defined as:

$$diff(A, I_1, I_2) = \frac{|I_1[A] - I_2[A]|}{\max(A) - \min(A)} \quad (2)$$

Algorithm 1 Feature selection of cervical cells using ReliefF

Input: cervical cell variables X and labels Y

Output: the vector W of estimations of the qualities of attributes A

1. Set all weights $W[A] := 0.0$;
2. **for** i := 1 **to** n **do begin**
3. randomly select an example x_i;
4. find k nearest hits H_j;
5. **for** each class $C \neq class(x_i)$ **do**
6. from class C fin k nearest misses M_j;
7. **for** A := 1 to a **do**
8. $W[A]:=W[A]-\sum_{j=1}^{k} diff(A, x_i, H_j)/(m \cdot k) + \sum_{C \neq class\{x_i\}} [\frac{P(C)}{1-P(class(x_i))} \sum_{j=1}^{k} diff(A, x_i, M_j(C))]/(m \cdot k)$;
9. **end**

3.6. Classification

In the diagnosis of cervical cancer, there are many methods for classification. In the present study, we use the random forest model, which has a wide application in many biological areas.

3.6.1. Random Forest Model

The RF model is a statistical learning method proposed by Breiman in 2001 [4]. This method combines the Bagging method [2] with the random subspace method [17], and contains multiple decision trees trained according to bagging. The final classification result is obtained after voting on the results of all decision trees. RF overcomes overfitting and the tolerance ability to noise and outliers.

RF is an ensemble classifier comprising multiple decision trees $\{h(x, \theta_k), k=1, 2,\ldots, q\}$, where θ_k is the parameter vector of the k-st tree, $\{\theta_k\}$ is a set of independent and identically distributed random vectors, and q is the number of the trees. Each tree casts a unit vote for the most popular class with the input sample x. The final class label set $Y=\{y\}$ of the sample set $X=\{x\}$ is determined using all decision trees. In random forest, $h_k(x)=h(x, \theta_k)$. The final classification decision is as follows:

$$H(x) = \arg\max_{Y} \sum_{k=1}^{q} I(h_k(x) = Y) \tag{3}$$

where $I(\circ)$ is an indicator function to study fractional factorial designs [46].

RF is more suitable to overcome overfitting than decision tree, as confirmed by the convergence theorem. In addition, generalization error bounds generate a theoretical upper bound for RF [4].

To construct q trees, q random variables (θ_k) should be generated. These random variables are independent and identically distributed. For any random variables, θ_k, a decision classification tree $h(x, \theta_k)$ or $h_k(x)$ can be constructed.

According to the q classifiers $h_1(x)$, $h_2(x)\ldots h_q(x)$ and the training set drawn at random from the distribution of the random vector Y,X, the definition of margin function is given as:

$$mg(X,Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j) \tag{4}$$

Margin function measures the difference between the average number classified correctly and incorrectly. The greater the value of the margin function, the more reliable the prediction results. Thus, the generalization error of the classifier can be expressed as:

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \tag{5}$$

If the number of trees is large, then it follows the law of large numbers. With the increase of the number of classes for the sequences θ_k and PE^*, the following convergence can be obtained:

$$P_{X,Y}(P_\theta(h(X,\theta) = Y) - \max_{j \neq Y} P_\theta(h(X,\theta) = j) < 0) \tag{6}$$

3.6.2. Classification of Cervical Cells Based on Random Forest

RF is a combination classifier with high dimensional data. It is often applied to obtain classifications in the field of biology. RF generates classifications in three steps.

- Step 1. The *q* sample sets are selected using bootstrap sampling from the training set of cervical cells after features extraction and selection. Every sample set has the same size.
- Step 2. The *q* classification and regression trees (CARTs) are built using the *q* sample sets. CARTs are binary decision trees.
- Step 3. The final classification results are obtained according to the vote of the *q* classification results.

4. Experiments

4.1. Experimental setup

The classification in the present study was conducted for two classes, normal and abnormal cells. The Waikato Environment for Knowledge Analysis (WEKA) [14] provides an extensive collection of machine learning algorithms, and data preprocessing tools are implemented with the default parameters for different classifiers. In the present study, we used WEKA 3.8.0, including all classifiers used in the present study. A 10-fold cross validation was adopted in all experiments. In each iteration, 9-fold was used for training, and the rest was applied for the validation of the test.

RF has two important training parameters: the number of trees n_{tree} and random features F. For selecting F, Breiman [4] suggests trying the default, F=int($\log_2 M$ +1), where M is the number of input features. In our experiments, with 20 dimensional features, 10 to 200 trees were built with a step size of 10 trees, and 100 to 2000 trees were built with a step size of 100 trees. The results showed that the classifier had the highest classification accuracy when 100 trees were built. Therefore, in our experiment, the parameters were selected with n_{tree} =100, F=int($\log_2 M$ +1).

To verify the effectiveness of the feature selection method, the weight of 20 features in the Herlev data set was ranked using ReliefF. The experiments were conducted based on different classifiers, and the performance of feature selection method was measured based on pattern variations from 1 to 20 ranking features.

In addition to RF, the other three classifiers, including Naive Bayes (NB) [10], C4.5 [38] and Logistic Regression (LR) [13], were utilized as comparative classifiers. Naive Bayes uses a probabilistic method, which multiplies the individual probabilities of each feature-value pair to realize the classification. C4.5 is an extension of ID3, which accounts for unavailable values, pruning of decision trees, the ranges of continuous attribute value, rule derivation, etc. Logistic regression measures the relationship between the categorical dependent variable and several independent variables using a logistic function to estimate probabilities.

4.2. Evaluation Metrics

The common method of evaluating a medical diagnosis classifier is to compare the diagnostic results of the classifier with the actual condition of the patients. Four metrics are typically applied to evaluate the results.

- True Negative (*TN*): The cell is diagnosed as a normal cell, and the actual condition is normal.
- False Negative (*FN*): The cell is diagnosed as a normal cell, and the actual condition is abnormal.
- True Positive (*TP*): The cell is diagnosed as an abnormal cell, and the actual condition is abnormal.
- False Positive (*FP*): The cell is diagnosed as an abnormal cell, and the actual condition is normal.

Based on these four metrics, the classification accuracy was defined to directly evaluate the entire prediction performance of the proposed method using the following formula:

$$Accuracy = \frac{TP + TN}{TN + FN + TP + FP} \times 100\% \tag{7}$$

A receiver operating characteristic (ROC) curve was also utilized to evaluate the classifiers. The ROC curve is a plotting of the True Positive Rate (TPR) as a function of the False Positive Rate (FPR). The area under the ROC curve (AUC) is a good measure of the performance of a classifier [2]. The higher the AUC value, the better the classification algorithm.

$$TPR = \frac{TP}{TP + FN} \tag{8}$$

$$FPR = \frac{FP}{FP + TN} \tag{9}$$

4.3. Results and Discussion

4.3.1. Feature Selection Results

The ranking results of feature selection are shown in Table 3. The results showed that nucleus perimeter, nucleus shortest diameter and nucleus area are the three most relevant features for classification. Table 4 shows the highest accuracies of the four classifiers with feature selection and the accuracies without feature selection. The number of features was shown when the highest accuracy was obtained. The results based on RF with the top 13 features achieved the best performance in all the experiments. Figure 3 shows the accuracies of the four classifiers with different dimensional features. Table 5 shows the value of AUC with different dimensions of features based on RF. The AUC value with the top 13 features remained the best on RF.

Table 3. The ranking results of ReliefF

No.	Feature	Average weight
1	Nucleus perimeter	0.111
2	Nucleus shortest diameter	0.11
3	Nucleus area	0.109
4	Nucleus longest diameter	0.103
5	Nucleus/Cytoplasm ratio	0.085
6	Maxima in nucleus	0.062
7	Cytoplasm shortest diameter	0.053
8	Minima in nucleus	0.05
9	Cytoplasm brightness	0.049
10	Cytoplasm longest diameter	0.047
11	Cytoplasm perimeter	0.047
12	Cytoplasm area	0.045
13	Nucleus brightness	0.044
14	Nucleus relative position	0.042
15	Maxima in Cytoplasm	0.04
16	Cytoplasm elongation	0.038
17	Minima in Cytoplasm	0.037
18	Cytoplasm roundness	0.034
19	Nucleus elongation	0.029
20	Nucleus roundness	0.027

Furthermore, the above results showed that features with high classification accuracy include both nucleus and cytoplasm features. To verify the effectiveness of different types of features, special experiments were implemented with only nucleus features. The results are shown in Table 6. The accuracies with the top 13 features were better than those with only nucleus features based on four classifiers.

Table 4. The classification results based on different classifiers and features

Method	The number of features	Accuracy rate
NB	20/16	91.71/92.04
C4.5	20/16	92.15/92.80
LR	20/13	92.35/93.13
RF	**20/13**	**92.58/94.44**

Figure 4-Figure 7 use ROC curves to show the classification performance based on different classifiers with different dimensions of features. A larger area under the ROC curves indicates better classifier performance. Table 7 shows the AUC based on NB, C4.5, LR, and RF with the top 13, 15, 17, and 19 features. RF achieved the best ROC and AUC values with different dimensions of features. The average AUC value based on RF is 0.9747.

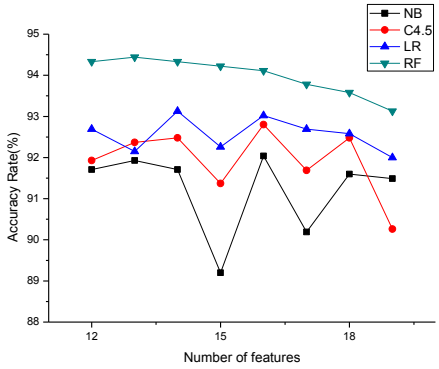

Figure 3. Accuracy of NB, C4.5, LR, and RF

Table 5. The value of AUC with different features based on RF

The number of features	The value of AUC by RF
1	0.9068
2	0.9238
3	0.9373
4	0.9396
5	0.9611
6	0.9622
7	0.9611
8	0.9589
9	0.9668
10	0.9758
11	0.9781
12	0.9793
13	0.9804
14	0.9793
15	0.9781
16	0.9770
17	0.9735
18	0.9715
19	0.9668
20	0.9622

Table 6. The accuracies with different types of features

	NB	C4.5	LR	RF
Only nucleus features	86.70%	88.75%	91.46%	94.00%
All features	91.71%	92.15%	92.35%	92.58%
The top 13 features	**91.93%**	**92.37%**	**92.15%**	**94.44%**

4.3.2. Discussion

In the feature selection phase, 20 features are ranked in descending order according to weight using ReliefF. As shown in Table 4, the classification accuracy with feature selection is better than that without feature selection based on different classifiers. As shown in Table 5, RF obtained the highest value of AUC with the top 13 features, which is higher than the value of AUC with all features, confirming that the ReliefF method is effective to improve the performance of the classification with fewer features.

The classification performance using only the nucleus features has been verified as better than that using all features [35], reaching an accuracy of 90.58% using the Fuzzy C-means model with the Herlev data set. As shown in Table 6, the classification accuracy using the nucleus features is better than that using all features based on the four classifiers. However, after feature selection, as proposed in the present study, the classification accuracies with the top 12 to 15 features are better than those using only nucleus features. These features not only include nucleus features but also the cytoplasm features. The combination of different types of features improves the performance of classification, confirming that the cytoplasm features are also helpful. Furthermore, not all nucleus features, such as Nucleus elongation and Nucleus roundness, are helpful for classification because the beneficial information for the classification has been shown in other features.

Moreover, four classifiers were selected for comparison, and the results of accuracy, ROC curve and AUC are shown. Figure 3 shows that the selected feature subsets were effective for all the classifiers. The highest accuracy was obtained using RF with the top 13 features. The results of NB, C4.5, and LR were lower than those of RF with the top 12-19 features. The results also confirmed, based on Figure 4-Figure 7 and Table 7, that RF is better than other classifiers for the classification in the present study.

Table 7. The AUC value of NB, C4.5, LR, and RF with the top 13, 15, 17, and 19 features

	13	15	17	19
NB	0.9547	0.9260	0.9363	0.9498
C4.5	0.9589	0.9485	0.9519	0.9370
LR	0.9566	0.9578	0.9622	0.9499
RF	**0.9804**	**0.9781**	**0.9735**	**0.9668**

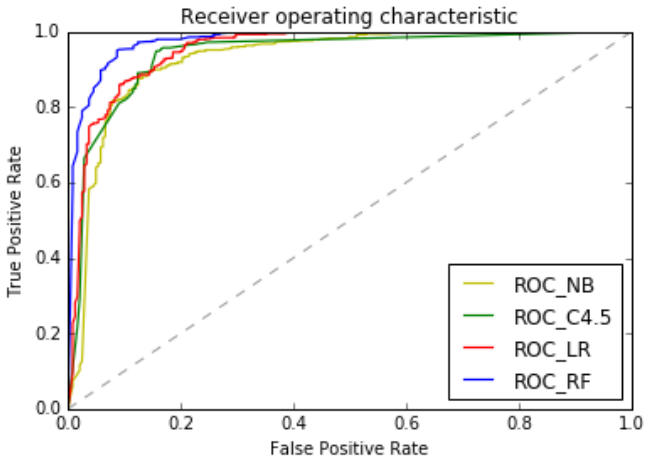

Figure 4. The ROC curves of NB, C4.5, LR, and RF with the top 13 features

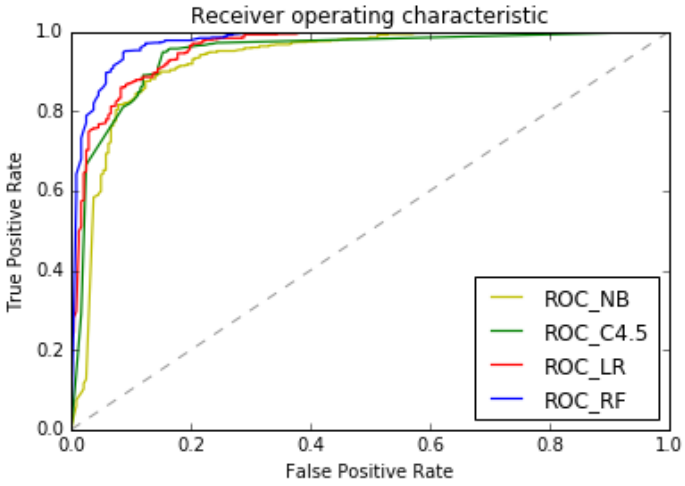

Figure 5. The ROC curves of NB, C4.5, LR, and RF with the top 15 features

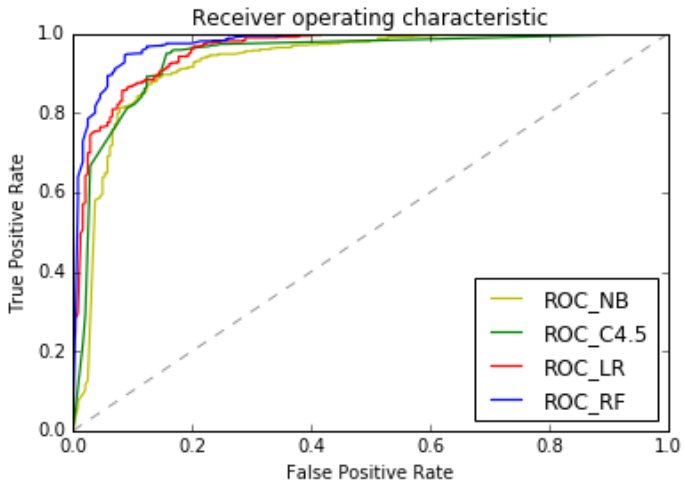

Figure 6. The ROC curves of NB, C4.5, LR, and RF with the top 17 features

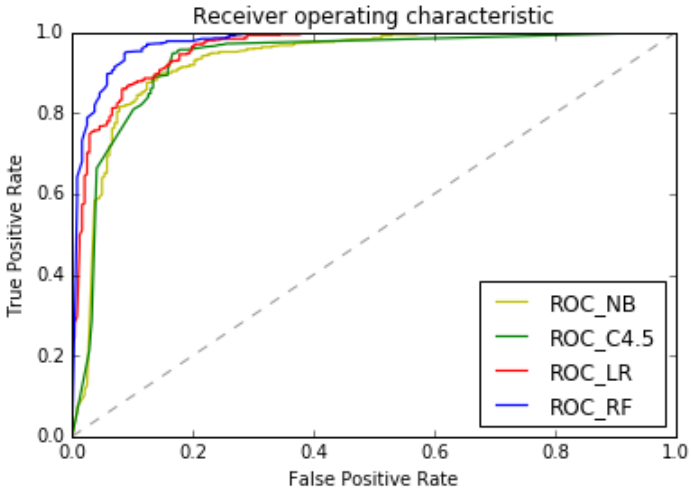

Figure 7.The ROC curves of NB, C4.5, LR, and RF with the top 19 features

5. **Conclusions**

Compared with manual diagnosis, the automated method is not only convenient and time saving but also reduces the labor intensity of doctors. The present study focuses on the feature selection and classification methods. The features are ranked according to weight using ReliefF. RF has been applied to classify cervical cells with different dimensional features. The experimental results showed the effectiveness of the proposed method compared with the results based on NB, C4.5, and LR. The effectiveness of different types of features was also analyzed based on these results, suggesting that automated analysis should include cytoplasm features.

Acknowledgements

This work was partly financially supported through grants from the National Natural Science Foundation of China (No. 60903083 and 61502123), Scientific planning issues of education in Heilongjiang Province (No. GBC1211062), and the research fund for the program of new century excellent talents (No. 1155-ncet-008). The authors thank the 3 anonymous reviewers for their helpful suggestions.

References

1. J. Albert, E. Aliu, H. Anderhub, P. Antoranz, A. Armada, M. Asensio, and J. Becker, "Implementation of the random forest method for the imaging atmospheric Cherenkov telescope MAGIC," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 588, no. 3, pp. 424-432, 2008
2. A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145-1159, 1997
3. C. Bergmeir, M. G. Silvente, and J. M. Ben ítez, "Segmentation of cervical cell nuclei in high-resolution microscopic images: A new algorithm and a web-based software framework," *Computer Methods & Programs in Biomedicine*, vol. 107, no. 3, pp.497–512, 2012
4. L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001
5. M. P. Coleman, J. Esteve, P. Damiecki, A. Arslan, and H. Renard, "Trends in cancer incidence and mortality," IARC scientific publications, 1992.
6. P. S. Chandran, N. B. Byju, R. U. Deepak, R. R. Kumar, S. Sudhamony, P. Malm, and E. Bengtsson, "Cluster detection in cytology images using the cellgraph method," *In Information Technology in Medicine and Education (ITME), 2012 International Symposium on*, vol. 2, pp. 923-927, August, 2012
7. Y. F. Chen, P. C. Huang, K. C. Lin, H. H. Lin, L. E. Wang, C. C. Cheng, and J. Y. Chiang, "Semi-automatic segmentation and classification of pap smear cells," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 1, pp. 94-108, 2014
8. L. Denny, M. Quinn, and R. Sankaranarayanan, "Screening for cervical cancer in developing countries," *Vaccine*, 2006
9. R. D áz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, pp. 1, 2006
10. R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern classification," New York: Wiley, vol. 2, 1973
11. A. Gen çtav, S. Aksoy, and S. Önder, "Unsupervised segmentation and classification of cervical cell images," *Pattern Recognition*, vol. 45, no. 12, pp. 4151-4168, 2012
12. R. T. Greenlee, T. Murray, S. Bolden, and P. A. Wingo, "Cancer statistics, 2000," *CA: a cancer journal for clinicians*, vol. 50, no. 1, pp. 7-33, 2000

13. D. W. Hosmer Jr and S. Lemeshow, "Applied logistic regression," John Wiley & Sons, 2004
14. G. Holmes, A. Donkin, and I. H. Witten, Holmes, "Weka: A machine learning workbench." in *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on* ,pp. 357-361, December, 1994
15. N. M. Harandi, S. Sadri, N. A. Moghaddam, and R. Amirfattahi, "An automated method for segmentation of epithelial cervical cells in images of ThinPrep," *Journal of medical systems*, vol. 34, no. 6, pp. 1043-1058, 2010
16. R. Hummel, "Image enhancement by histogram transformation," *Computer graphics and image processing*, vol. 6, no. 2, pp. 184-195, 1977
17. T. K. Ho, "The random subspace method for constructing decision forests," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 8, pp. 832-844, 1998
18. T. hankong, N. Theera-Umpon, and S. Auephanwiriyakul, "Automatic cervical cell segmentation and classification in Pap smears," *Computer methods and programs in biomedicine*, vol. 113, no. 2, pp. 539-556, 2014
19. A. Jemal, M. M. Center, C. DeSantis, and E. M. Ward, "Global patterns of cancer incidence and mortality rates and trends," *Cancer Epidemiology Biomarkers & Prevention*, vol. 19, no. 8, pp. 1893-1907, 2010
20. D. Kong, C. Ding, H. Huang, and H. Zhao, "Multi-label relieff and f-statistic feature selections for image annotation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2352-2359, IEEE, June, 2012
21. K. K. Kandaswamy, G. Pugalenthi, M. K. Hazrati, K. U. Kalies and T. Martinetz, "BLProt: prediction of bioluminescent proteins based on support vector machine and relieff feature selection", *BMC bioinformatics*, vol. 12, no. 1, pp. 345, 2011
22. M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC medical informatics and decision making*, vol. 11, no. 1, pp. 1, 2011
23. R. R. Kumar, V. A. Kumar, and P. N. Sharath Kumar, "Detection and removal of artifacts in cervical cytology images using support vector machine," *IT in Medicine and Education (ITME), 2011 International Symposium on*, vol. 1, pp. 717-721, 2011
24. S. Kumar, L. Jena, K. Mohod, S. Daf, and A. K. Varma, "Virtual screening for potential inhibitors of high-risk human papillomavirus 16 E6 protein," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 7, no. 2, pp. 136-142, 2015
25. K. Li, Z. Lu, W. Liu, and J. Yin, "Cytoplasm and nucleus segmentation in cervical smear images using Radiating GVF Snake," *Pattern Recognition*, vol. 45, no. 4, pp. 1255-1264, 2012
26. W. Z. Lin, J. A. Fang, X. Xiao, and K. C. Chou, "iDNA-Prot: identification of DNA binding proteins using random forest with grey model," *PloS one*, vol. 6, no. 9, pp. e24756, 2011
27. A. Mohan, M. D. Rao, S. Sunderrajan, G. Pennathur, "Automatic classification of protein structures using physicochemical parameters," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 6, no. 3, pp. 176-186, 2014
28. A. H. Mbaga, and P. Zhijun, "Pap Smear Images Classification for Early Detection of Cervical Cancer," *International Journal of Computer Applications*, vol.118, no. 7, 2015
29. J. H. Moore and B. C. White, "Tuning ReliefF for genome-wide genetic analysis." in *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pp. 166-175, April, 2007
30. L. Martin and M. Exbrayat, "Pap-smear classification" Technical University of Denmark-DTU, 2003
31. P. Malm, B. N. Balakrishnan, V. K. Sujathan, R. Kumar, and E. Bengtsson, "Debris removal in Pap-smear images," *Computer methods and programs in biomedicine*, vol. 111, no. 1, pp. 128-138, 2013
32. Y. Marinakis, G. Dounias, and J. Jantzen, "Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification," *Computers in Biology and Medicine*, vol. 39, no. 1, pp.69-78, 2009
33. J. Norup, "Classification of Pap-smear data by tranduction neuro-fuzzy methods" Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark, 2005
34. M. Peker, A Arslan, B. Sen, F. V. Celebi, and A. But, "A novel hybrid method for determining the depth of anesthesia level: Combining ReliefF feature selection and random forest algorithm (ReliefF+ RF)." in *Innovations in Intelligent SysTems and Applications (INISTA), 2015 International Symposium on*, pp. 1-8, September, 2015
35. M. E. Plissiti and C. Nikou, "Cervical cell classification based exclusively on nucleus features," *Image Analysis and Recognition. Springer Berlin Heidelberg*, pp. 483-490 ,2012
36. M. E. Plissiti, C. Nikou and, A. Charchanti, "Watershed-based segmentation of cell nuclei boundaries in Pap smear images," *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on*. pp. 1-4, 2010
37. M. E. Plissiti, C. Nikou, and A. Charchanti, "Automated Detection of Cell Nuclei in Pap Smear Images Using Morphological Reconstruction and Clustering," *IEEE Transactions on Information Technology in Biomedicine A Publication of the IEEE Engineering in Medicine & Biology Society*, vol. 15, no .2, pp. 233-241, 2011
38. J. R. Quinlan, "C4.5: programs for machine learning," *Elsevier*, 2014
39. M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF,A" *Machine learning*, vol.53, no. 1-2, pp. 23-69, 2003
40. P. Sobrevilla, E. Montseny, F. Vaschetto, and E. Lerma, "Fuzzy-based analysis of microscopic color cervical pap smear images: nuclei detection," *International Journal of Computational Intelligence and Applications*, vol. 9, no. 03, pp. 187-206, 2010
41. S. Saha, M. Pal, A. Konar, and D. Bhattacharya, "Automatic Gesture Recognition for Health Care Using ReliefF and Fuzzy kNN." In *Information Systems Design and Intelligent Applications*, pp. 709-717, 2015
42. S. N. Sulaiman, N. Ashidi, M. Isa, and N. H. Othman, "Semi-automated pseudo colour features extraction technique for cervical cancer's pap smear images," *International Journal of Knowledge-based and Intelligent Engineering Systems*, vol. 15, no. 3, pp. 131-143, 2011
43. V.Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947-1958, 2003
44. V. M. Valdespino and V. E. Valdespino, "Cervical cancer screening: state of the art," *Current Opinion in Obstetrics and*

Gynecology, vol. 18, no. 1, pp. 35-40, 2006
45. J. Wu, H. Liu, X. Duan, Y. Ding, H. Wu, Y. Bai, and X. Sun, "Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature," *Bioinformatics*, vol. 25, no. 1, pp. 30-35, 2009
46. K. Q. Ye, "Indicator function and its application in two-level factorial designs," *Annals of Statistics*, pp. 984-994, 2003
47. J Yue, Z Li, L Liu, and Z. Fu, "Content-based image retrieval using color and texture fused features," *Mathematical and Computer Modelling*, vol. 54, no. 3, pp. 1121-1127, 2011