A Machine Learning Texture Model for Classifying Lung Cancer Subtypes Using Preliminary Bronchoscopic Findings

Po-Hao Feng[1,2], Yin-Tzu Lin[1,3], Chung-Ming Lo[3,4*]

[1]School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan

[2]Division of Pulmonary Medicine, Department of Internal Medicine, Shuang Ho Hospital, Taipei Medical University, New Taipei City, Taiwan

[3]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan

[4]Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei, Taiwan

* Corresponding Author:
Prof. Chung-Ming Lo
Graduate Institute of Biomedical Informatics
Taipei Medical University
Taipei, Taiwan 11031, R.O.C.
Telephone: 886-2-2736-1661 ext. 3345
Fax: 886-2-27392914
E-mail: buddylo@tmu.edu.tw

Abstract

Purpose: Bronchoscopy is useful in lung cancer detection, but cannot be used to differentiate cancer types. A computer-aided diagnosis (CAD) system was proposed to distinguish malignant cancer types to achieve objective diagnoses.

Methods: Bronchoscopic images of 12 adenocarcinoma and 10 squamous cell carcinoma patients were collected. The images were transformed from a red-blue-green (RGB) to a hue-saturation-value (HSV) color space to obtain more-meaningful color textures. By combining significant textural features ($p<0.05$) in a machine learning classifier, a prediction model of malignant types was established.

Results: The performance of the CAD system achieved an accuracy of 86% (19/22), a sensitivity of 90% (9/10), a specificity of 83% (10/12), a positive predictive value of 82% (9/11), and a negative predictive value of 91% (10/11) in distinguishing lung cancer types. The area under the receiver operating characteristic curve was 0.82.

Conclusions: On the basis of extracted HSV textures of bronchoscopic images, the CAD system can provide recommendations for clinical diagnoses of lung cancer types.

Keywords: lung cancer, bronchoscopy, computer-aided diagnosis, color texture

Introduction

Despite the development of new medications and surgical therapies, lung cancer is by far the most common cancer and the leading cause of cancer deaths worldwide [1,2]. Non-small-cell lung cancer (NSCLC) occurs in nearly 85% of lung cancer patients [3-5], and 80% of NSCLCs are adenocarcinomas (ACs) and squamous cell carcinomas (SCCs). In clinical settings, advanced stages of adenocarcinoma (AC) and squamous cell carcinoma (SCC) require different treatment choices. In Asian countries, patients with advanced lung ACs comprise higher proportions, up to 50% compared to 10%~15% in Western countries, and such cases are related to mutations of the epidermal growth factor receptor (EGFR) and show a good response to EGFR-tyrosine kinase inhibitors (EGFR-TKIs); but the same is not true for SCCs [6,7]. For advanced SCCs, cytotoxic chemotherapy and/or radiotherapy have been standard treatments in the past decade. But recently developed immune checkpoint blockade therapy has become a new standard treatment for advanced SCC with high expression of programmed death ligand 1 (PD-L1) [8]. Because of the diversity of treatment choices when determining the best treatment for lung cancer patients, confirmation of the correct histologic type is the first step, followed by a molecular diagnosis panel of the lung cancer. Clinically, using an immunohistochemical (IHC) panel to confirm the

correct histological type is the most frequently employed method, but it takes a couple of days to get results.

In lung cancer patients with endobronchial lesions, a bronchoscopic biopsy is a rapid and safe way for tumor tissue acquisition, and with adequate tumor tissues, a correct pathologic diagnosis can be made. During the initial diagnosis period, a white-light bronchoscopic image might help make a diagnosis of the histologic type and facilitate subsequent cancer staging and treatment. However, one limitation of bronchoscopic examinations is the variance among different observers [9]. Usually a pathologic diagnosis and final molecular diagnosis may take up to 2 weeks. However, some patients have more-severe symptomatic disease and require immediate treatment. If computer-aided diagnosis (CAD) can provide adequate accuracy through bronchoscopic images before the pathologic diagnosis, it could help clinical physicians make decisions for prompt and adequate treatment of severely symptomatic patients.

With the development of image processing and machine learning, CAD systems are regarded as a practical way to establish objective diagnostic suggestions. By extracting quantitative image features and combining them with machine learning, each case is given a likelihood of being different types [10,11]. In the previous literature, a CAD system was only developed to classify normal mucosa and lung tumors using white-light bronchoscopy [12]. A diagnostic accuracy of 80% was achieved. Based on the

success of that CAD, this study first explored predictions of different lung cancer subtypes using bronchoscopy.

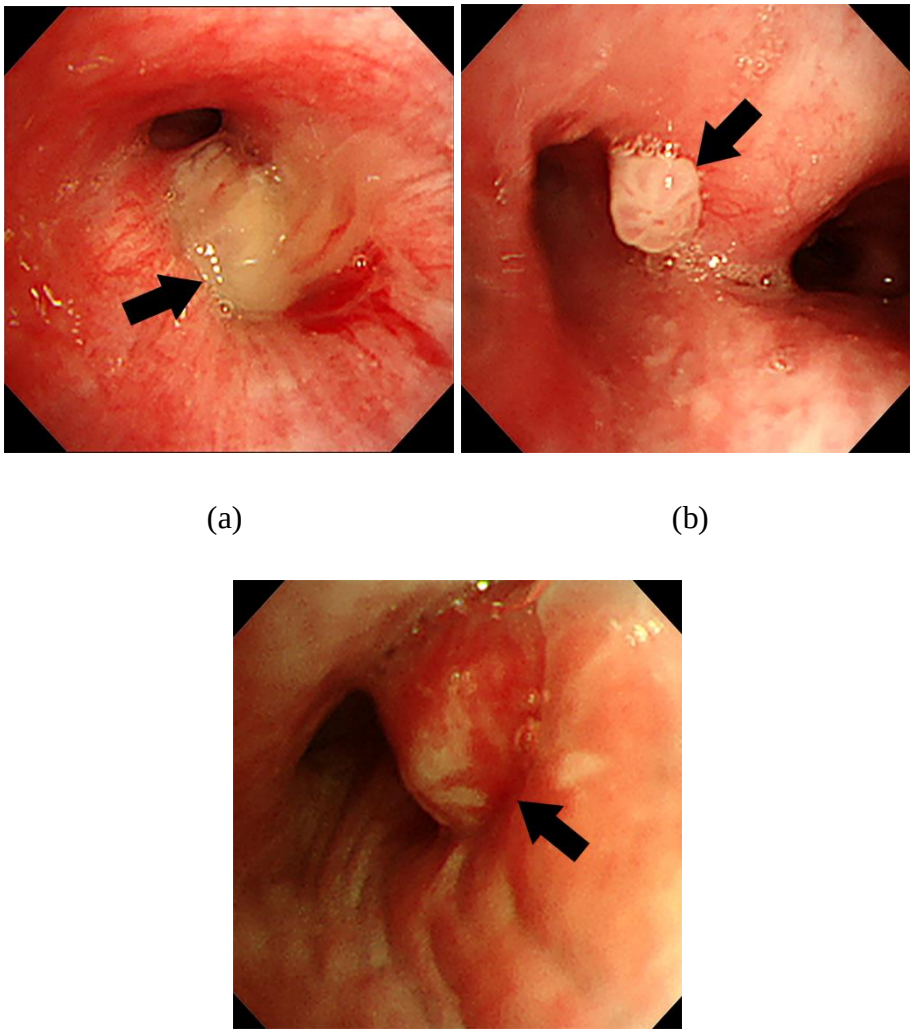
The previous literature mentioned that color of a tissue is important information for making medical diagnoses, such as when analyzing skin erythema and skin segmentation, and evaluating a wound's status [13,14]. The success of pattern recognition highly depends on how to interpret the color information of bronchoscopy and distinguish different normal tissues from lung cancers. Similar to human perception and color describing, the hue-saturation-value (HSV) color space intuitively translates colors using hue, saturation, and brightness and has been successfully used in medical applications [13,14].

In this study, quantitative texture features were extracted from HSV color channels to distinguish differences between ACs and SCCs. Under a situation with various lighting changes in bronchoscopy, separating color components from intensities would make feature extraction more robust [15]. Using a machine learning classifier, the interpreted diagnostic information was transformed into probabilities to evaluate the types of observed lung cancers. Development of such a CAD system can provide more-objective recommendations for recognizing lung cancer types using bronchoscopy and facilitate clinical physician decision-making in some circumstances.

Materials and Methods

Patient information

The institutional review board of the hospital approved this study. Informed consent was waived because we used retrospective data. From September 2015 to April 2017, 70 patients at Shuang Ho Hospital were examined using bronchoscopy (BF-F260, Olympus Optical, Tokyo, Japan). Of the 70 cases, 34 patients were found to have positive findings of neoplastic changes, while the other 36 patients had normal mucosal changes. The final pathologic diagnosis of 34 patients was made by a pathologist and confirmed by a lung cancer tumor board as having neoplastic lesions, and these data were collected as a diagnostic reference. Bleeding and blood clots caused by the bronchoscopic procedure can affect the interpretation. Thus, among the 34 patients, bronchoscopic images of 22 cases of lung cancers without bleeding were enrolled. Other neoplastic lesions of limited size, including two SCCs, two unknown carcinomas, and one tracheal tumor, were excluded. In total, 12 AC patients (aged 42~83 years) and 10 SCC patients (aged 50~90 years) were enrolled. Examples of endobronchial AC and SCC tumors are demonstrated in Fig. 1.



(a) (b)

(c)

Fig. 1. Examples of (a) an adenocarcinoma and (b and c) two squamous cell carcinomas shown on bronchoscopic images with the black arrows pointing at tumor edges.

Multi-channel features

Conventional CAD systems, such as computed tomography, ultrasound, and magnetic resonance imaging, quantify gray-scale features in medical images for

diagnoses [16,17]. These CAD systems extract brightness variations of lesion tissues and compare brightness levels between lesions and background tissues to classify benign and malignant tumors with substantial accuracy. In this study, lung cancer tissues presenting a color appearance in bronchoscopy were analyzed. Multi-channel features were extracted from three color channels for tissue characterization. First, endobronchoscopic images were converted from the unintuitive red-green-blue (RGB) to the hue (H)-saturation (S)-value (V) (HSV) color space, which is better for interpreting color properties. Then, the textures in the H, S, and V channels were individually extracted to establish a diagnostic model.

HSV transformation

Using bronchoscopy to detect lung cancers depends on the color appearance of tissues. Diagnostic results are dramatically related to the color space used for interpreting color information. In this experiment, the inherent RGB color space was transformed into an HSV composition[18]. HSV is one of the successful color spaces proposed to imitate human perception. The hue of tissues can present an abnormal appearance which can be recognized by physicians to distinguish them from other normal tissues; saturation may indicate the level of abnormalities from normal tissues; and value indicates the property of luminance. The color space of HSV channels is shown as Fig. 2. Figure 3 shows the transformed HSV images from Fig. 1.

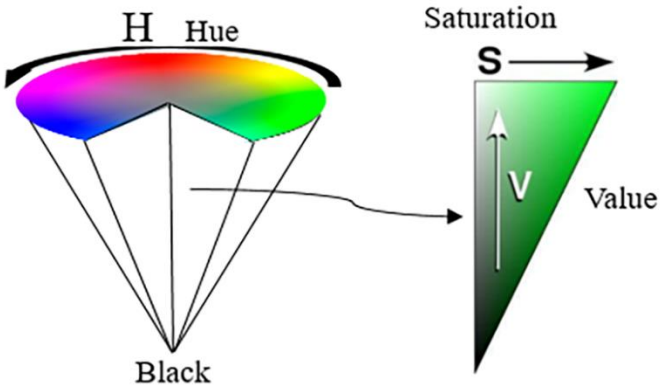
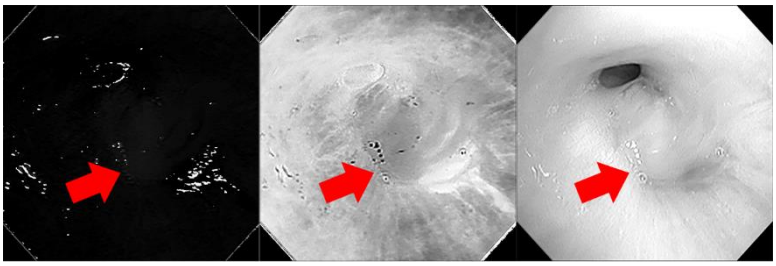
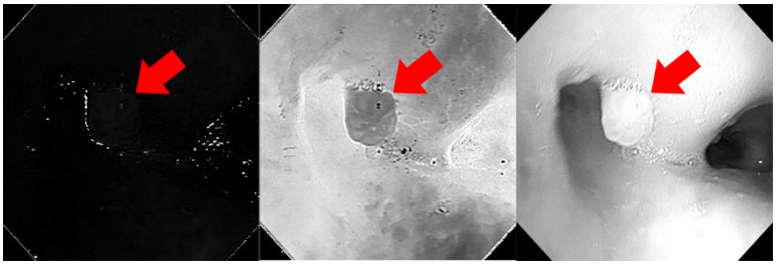


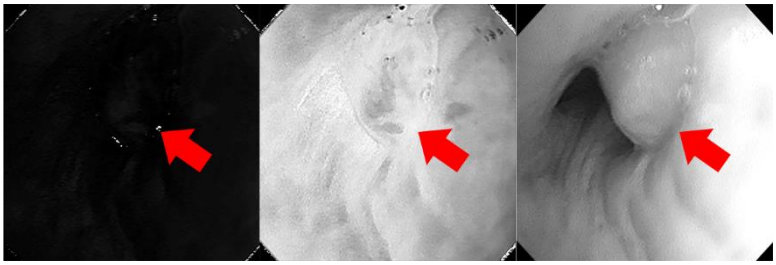
Fig. 2. HSV is a color space containing three channels: hue (H), saturation (S), and value (V).



(a)



(b)



(c)

Fig. 3. Transformed hue(left)-saturation(middle)-value(right) (HSV) images of Fig. 1. (a) HSV of Fig. 1a; (b) HSV of Fig. 1b; (c) HSV of Fig. 1c with the red arrows pointing at tumor edges.

Textural features

Textural features are widely extracted from medical images for pattern recognition [16,17]. Generally, texture information is analyzed using gray-scale values. By combining multiple texture statistics, benign and malignant tumors can be recognized by a classifier. Given the success of previous studies [16,17], this study used HSV color channels to extract textural features to classify malignant types. After delineating tumor contours by bronchoscopy, tissues inside the delineated area were characterized.

Color textures of tumors in endobronchoscopic images were interpreted by analyzing correlations between pixel values. In each of the H, S, and V channels, a gray-level co-occurrence matrix (GLCM) [19] as second-order statistics described the joint frequencies of pair-wise combinations. After scanning a pixel and its adjacent pixels, co-occurrence matrices $P=[p(i,j|d,\theta)]$ were formed to show the frequencies of two eight-connected adjacent pixels at a distance, d, and direction, θ, with respective pixel values of i and j. The parameters of $d=1$ and four offset directions, $\theta=0°$, 45°, 90°, and 135°, were used in the experiment (Fig. 4). These four directions were combined

into a single matrix to achieve rotation invariance. The derived statistics listed below describe the 14 GLCM textural features: autocorrelation, contrast, correlation, cluster prominence, cluster shading, dissimilarity, energy, entropy, homogeneity, difference variance, difference entropy, information measure of correlation, inverse difference normalized, and inverse difference moment. Upon the statistical measurement of GLCM, local intensity variations that form the specified pattern of certain tissues can be quantified. Local appearances, including smoothness, regularity, and homogeneity, need to be interpreted in medical image examinations to differentiate various tissues.

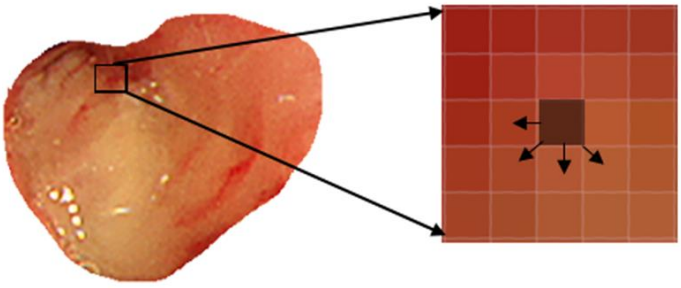


Fig. 4. Spatial correlations between a pixel and its neighboring pixels were analyzed in four directions of 0°, 45°, 90°, and 135° with distance=1.

Statistical analysis

Textural features extracted from multiple channels were evaluated if they were statistically significant in distinguishing lung cancer types. The Kolmogorov-Smirnov test[20] was used to determine if textural features were normally distributed. According to test results, Student's *t*-test[20] and the Mann-Whitney U-test[20] were respectively used to test normally and non-normally distributed features. For features with a *p* value of <0.05, there was a statistically significant difference between malignant types. In constructing the prediction model, textural features were grouped in a logistic regression classifier.

Logistic regressions are especially appropriate for establishing models involving disease states such as malignancy estimation and type classification and therefore are widely used in the health sciences. A logistic regression model generates a predicted value of the outcome variable as a sum of products with coefficients trained from statistical data. A regression model is helpful in answering questions about domain areas such as medicine, because the coefficient of each predictor variable explicitly expresses the relative contribution of that variable to the outcome variable.

Based on backward elimination [21], one feature with a minimum predictive residual error sum of squares was eliminated each time. In the end, the most relevant subset features with the lowest error rate were combined for a tumor diagnosis. Then, a leave-one-out cross-validation method [22] was used to validate the diagnostic performance. If

K cases were collected in the experiment, the training and testing executed *K* iterations. In each iteration, one case was selected from the dataset and removed in order to test the result trained by the remaining *K*-1 cases. Using the biopsy-proven pathology as the gold standard, tumors were assigned a probability value as the likelihood of being AC or SCC in the binary logistic regression [23]. Probabilities regarded as prediction results were determined according to the quantitative image features of tumors.

$$P = \frac{1}{1 + e^{-(c_1 \times x_1 + c_2 \times x_2 + \cdots + c_n \times x_n + c_0)}} \quad (1)$$

In the equation, various diagnostic features: $\{x_1, x_2, \cdots, x_n\}$ were combined using their coefficients c_i and the constant c_0. After classification, five performance indices including the accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated for evaluation. A receiver operating characteristic (ROC) curve was used to show tradeoffs between the sensitivity and specificity. The resulting area under the ROC curve, AUC, was calculated using ROCKIT software (C. Metz, University of Chicago, Chicago, IL, USA). The logistic regression and significance testing were carried out using SPSS software (vers. 16 for Windows; SPSS, Chicago, IL, USA).

Results

After transforming the original image composition from RGB to HSV, 14 GLCM textural features were extracted from individual channels. Then, 42 textural features in total were evaluated as to whether they could distinguish malignant types. As a result, the correlation, cluster prominence, cluster shading, and difference variance in the S channel and cluster prominence and cluster shading in the V channel had significant *p* values of <0.05 as shown in Table 1. After backward elimination used in the logistic regression, the most relevant features, including correlation (S), difference variance (S), and cluster shading (V), were combined to achieve an accuracy of 86% (19/22), a sensitivity of 90% (9/10), a specificity of 83% (10/12), a PPV of 82% (9/11), and an NPV of 91% (10/11). The AUC was 0.82. Among 10 SCC cases, only one was misclassified with a probability of 2% (cases having probabilities of >50% were regarded as being an SCC). Misclassified examples are illustrated in Fig. 5a (misclassified SCC) and 5b, which is the corresponding delineated tumor area of Fig. 5a.

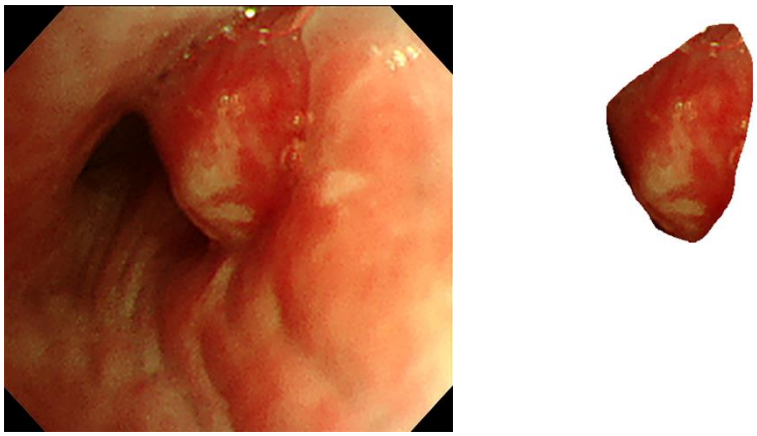
Table 1. Significant textural features in the hue-saturation-value (HSV) color space and corresponding *p* values evaluated using Student's *t*-test

Feature	AC	SCC	*p* value
	Mean±SD	Mean±SD	
Correlation (S)	0.987±0.004	0.979±0.004	<0.001*

Cluster prominence (S)	1589.937±818.998	738.032±433.412	<0.01*
Cluster shading (S)	130.865±66.141	64.554±36.635	<0.01*
Difference variance (S)	20.257±9.692	12.559±6.827	<0.05*
Cluster prominence (V)	298.707±206.022	143.151±80.896	<0.05*
Cluster shade (V)	-32.370±20.906	-16.407±8.691	<0.05*

* A *p* value of <0.05 indicates a statistically significant difference.

AC, adenocarcinoma; SCC, squamous cell carcinoma; S, saturation; V, value; SD, standard deviation.



(a) (b)

Fig. 5. A misclassified squamous cell carcinoma (SCC). (a) The appearance of an SCC shown in bronchoscopy and (b) the corresponding delineated tumor area.

Discussion

Bronchoscopy is a widely used imaging examination to detect and take biopsies of endobronchial lesions of lung cancer. The limitation is that physicians cannot classify

carcinomas solely according to bronchoscopic images. This study presents a rapid and accurate assessment of malignant types via a quantitative textural analysis using bronchoscopic images. Based on color textures in the HSV color space, the proposed CAD system can generate more-objective diagnoses with less inter-observer variability. With respect to the diagnostic accuracy, the CAD system achieved an accuracy of 86% (19/22), a sensitivity of 90% (9/10), a specificity of 83% (10/12), a PPV of 82% (9/11), and an NPV of 91% (10/11). With the help of this CAD system, clinical physicians could have rapid and more-reliable results before a final pathologic diagnosis is available, and this could facilitate decision making by physicians for more-severe and urgent lung cancer patients.

As to the few previously published articles, one CAD system obtained an 80% accuracy in classifying normal mucosa and tumors using white-light bronchoscopy [12]. The difference between that study and ours is that our study extracted textural features from HSV, a more-perceptual color space and achieved a higher accuracy (86% vs. 80%). In observations from the experiment results, significant textural features in distinguishing AC and SCC were correlation, cluster prominence, cluster shading, and difference variance from the saturation channel, and cluster prominence and cluster shading from the value channel (Table 1). Hue features did not achieve significance, which revealed that the primary colors of both malignant types were similar. Under

such a situation, the saturation channel had the most significant features and thus was selected for the prediction model. The saturation may indicate the level of abnormalities from normal tissues. Complementary to cluster shading in the value channel, tissue brightness was also considered. This performance suggests that using a texture analysis of bronchoscopic images can provide more diagnostic information between image pixels which might not be easily observed by human beings. Differentiating different lung cancer types via a bronchoscopic CAD is a promising method. With respect to diagnostic results, the misclassified SCC in Fig. 5 had a heterogeneous tissue composition. According to the significant features (difference variance), SCCs tend to have lower values than ACs, which may cause SCC cases to be misclassified. The underlying biological structure related to the heterogeneous patterns between ACs and SCCs should be evaluated in future studies.

A previous study used radiomic CT features to classify tumor histologic subtypes [24]. In that study, 440 features were extracted from gray-scale CT volumes which differed from our color features. The number of features and volumes may require substantial computation time compared to our 42 features from bronchoscopic images. For clinical use, using bronchoscopy or CT depends on the status of patients and available physicians and equipment. In a situation requiring a rapid diagnosis, time is

critical for physicians. Since both CAD systems were well-established, exploring the accuracy of combining them would be a potential future study.

There are several limitations in this study. First, this was a retrospective study, and there might have been some selection bias during the process. Second, the case number was small in this study. Nevertheless, this is a preliminary experiment regarding the use of quantitative image features and machine learning techniques to classify different lung cancer subtypes via bronchoscopic images. In the future, a larger prospective study comparing physicians and the CAD system should be conducted [25]. More cases should be added in future studies to enhance the generalization of the proposed method. Simultaneously, correlations between image findings and pathology findings can also be explored. Another challenging but very practical task is automatic tumor detection. After more-quantitative features are implemented later, the CAD system can be equipped with both lung cancer detection and diagnostic features. Then, whether the CAD system can improve physicians' performance is also a relevant topic for clinical use.

Conclusions

In this study, we demonstrate that a CAD system can help distinguish lung cancer subtypes using bronchoscopic images. The CAD system can provide rapid and precise results for clinical physicians before the final pathologic results are available, and provide a greater time window for physicians to make decisions for more-severe and urgent lung cancer patients. This will certainly improve the quality of care, and result in better patient care.

Acknowledgements

The authors thank the Ministry of Science and Technology, Taiwan (MOST 106-2221-E-038-018 and 107-2221-E-038-017), Taipei Medical University, and Shuang Ho Hospital (105TMU-SHH-02-4) for financially supporting this study.

References

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012 [published online ahead of print 2014/09/16]. *International journal of cancer.* 2015;136(5):E359-386.

2. Fitzmaurice C, Allen C, Barber RM, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study [published online ahead of print 2016/12/06]. *JAMA oncology.* 2017;3(4):524-548.

3. Sher T, Dy GK, Adjei AA. Small cell lung cancer [published online ahead of print 2008/03/05]. *Mayo Clinic proceedings.* 2008;83(3):355-367.

4. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship [published online ahead of print 2008/05/03]. *Mayo Clinic proceedings.* 2008;83(5):584-594.

5. SEER Cancer Statistics Review, 1975–2014. National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/csr/1975_2014/, based on November 2016 SEER data submission, posted to the SEER web site, April 2017. https://seer.cancer.gov/csr/1975_2014/.

6. Midha A, Dearden S, McCormack R. EGFR mutation incidence in non-small-cell lung cancer of adenocarcinoma histology: a systematic review and global map by ethnicity (mutMapII) [published online ahead of print 2015/11/27]. *Am J Cancer Res.* 2015;5(9):2892-2911.

7. Mok TS, Wu YL, Thongprasert S, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma [published online ahead of print 2009/08/21]. *N Engl J Med.* 2009;361(10):947-957.

8. Reck M, Rodriguez-Abreu D, Robinson AG, et al. Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer [published online ahead of print 2016/10/11]. *N Engl J Med.* 2016;375(19):1823-1833.

9. Van Coillie FMB, Gardin S, Anseel F, Duyck W, Verbeke LPC, De Wulf RR. Variability of operator performance in remote-sensing image interpretation: the importance of human and external factors. *International Journal of Remote Sensing.* 2014;35(2):754-778.

10. Lee H, Chen Y-PP. Image based computer aided diagnosis system for cancer detection. *Expert Systems with Applications.* 2015;42(12):5356-5365.

11. Castellino RA. Computer aided detection (CAD): an overview. *Cancer Imaging.* 2005;5(1):17-19.

12. Benz M, Rojas-Solano JR, Kage A, Wittenberg T, Munzenmayer C, Becker HD.

Computer-Assisted Diagnosis for White Light Bronchoscopy: First Results. *CHEST.*138(4):433A.

13. Lee CH, Kim JS, Park KH. Automatic human face location in a complex background using motion and color information. *Pattern recognition.* 1996;29(11):1877-1889.

14. Sobottka K, Pitas I. Extraction of facial regions and features using color and shape information. Paper presented at: Pattern Recognition, 1996., Proceedings of the 13th International Conference on1996.

15. Zheng Z, Yang J, Yang L. A robust method for eye features extraction on color image. *Pattern Recognition Letters.* 2005;26(14):2252-2261.

16. Hsieh KL-C, Chen C-Y, Lo C-M. Radiomic model for predicting mutations in the isocitrate dehydrogenase gene in glioblastomas. *Oncotarget.* 2017;8(28):45888.

17. Hsieh KL-C, Chen C-Y, Lo C-M. Quantitative glioma grading using transformed gray-scale invariant textures of MRI. *Computers in Biology and Medicine.* 2017;83:102-108.

18. Oliveira RB, Mercedes Filho E, Ma Z, Papa JP, Pereira AS, Tavares JMR. Computational methods for the image segmentation of pigmented skin lesions: a review. *Computer methods and programs in biomedicine.* 2016;131:127-141.

19. Haralick RM, Shanmuga.K, Dinstein I. Textural Features for Image Classification.

IEEE Trans Syst Man Cybern. 1973;Smc3(6):610-621.

20. Field AP. *Discovering statistics using SPSS, 3rd ed.* Los Angeles: SAGE Publications; 2009.

21. Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of clinical epidemiology.* 2004;57(11):1138-1146.

22. Alpaydin E. *Introduction to machine learning.* Cambridge, Mass: MIT Press; 2004.

23. Hosmer DW. *Applied logistic regression. 2nd edition.* New York: Wiley; 2000.

24. Wu W, Parmar C, Grossmann P, et al. Exploratory study to identify radiomics classifiers for lung cancer histology. *Frontiers in oncology.* 2016;6:71.

25. Hsieh KL-C, Tsai R-J, Teng Y-C, Lo C-M. Effect of a computer-aided diagnosis system on radiologists' performance in grading gliomas with MRI. *PloS one.* 2017;12(2):e0171342.