Binary Count Ratio for Lung Cancer Classification in Computerized Tomography Scan Images

Sittisak Saechueng
Faculty of Informatics,
Burapha University,
Chonburi, Thailand
sittisak.sa@buu.ac.th

Ungsumalee Suttapakti
Faculty of Informatics,
Burapha University,
Chonburi, Thailand
ungsumalee.su@buu.ac.th

***Abstract*—Accurate lung cancer classification is important for patient treatment. However, existing methods inefficiently classify lung cancer. Therefore, the binary count ratio (BCR) is proposed to enhance the accuracy of lung cancer classification. This method utilizes adaptive thresholding based on column mean to binarize CT images. The features of the BCR method are computed by using the ratio of black and white pixels to identify the lung cancer areas. The proposed BCR method captures the white pixels which are cancer areas in CT scan images. After that, the Euclidean distance method is used to classify a normal or benign or malignant class. For 1,097 images of the IQ-OTH/NCCD dataset, the proposed BCR method achieves 0.9810, 0.9618, 0.9705, and 0.9832 of precision, recall, F1-score, and classification accuracy values which are higher than the conventional methods. The proposed method is able to efficiently extract pixel areas of lung cancer, thus improving the effectiveness of lung cancer classification in CT scan images.**

Keywords—Lung cancer classification, machine learning, texture feature, binary count ratio, shape feature

I. Introduction

Lung cancer is a serious disease and causes high mortality. This cancer is caused by the rapid and uncontrolled growth of abnormal cells, which is a tumor in the lung. Tumors can be divided into two broad groups: benign and malignant [1]. In the early stage, a benign tumor begins in the lung. This tumor grows slowly and is not cancerous. When the growth of tumors is rapid, uncontrolled, and abnormal, the tumor becomes a malignant tumor or lung cancer. A malignant tumor can spread to body parts. Treatment of malignant tumors is difficult. Benign tumors in patients are detected before developing malignant tumors; the patients are completely cured. Most lung cancer is caused by smoking. Some cases occur from inhaling damaging chemicals. Early lung cancer often has no symptoms. The detection of this disease uses medical imaging. A computerized tomography scan (CT scan) is a method to obtain medical images for diagnosing lung cancer by specialized pulmonologists. Many researchers have attempted to apply machine learning and image analysis approaches to classify lung cancer. The performance of lung cancer classification depends on the feature extraction method. Feature extraction can be divided into two categories: texture-based and shape-based methods.

In the texture-based method, Ankita [2] introduced gray level co-occurrence matrix (GLCM) approach to extract features for lung cancer classification. Moreover, this approach uses a Median filter for image preprocessing and Fuzzy C-means for image segmentation. This approach is effective for classifying lung cancers. In the same way, the GLCM method introduced by Firdaus et al. [3] was used to finding area, contrast, energy, entropy, and homogeneity to extract features for lung cancer classification in CT-scan images. This method is effective. After that, Muayed et al. [4] compared the Gabor filter method with the GLCM method to extract features for classifying lung cancers. This method uses image enhancement, segmentation, and feature extraction to distinguish a normal, benign, or malignant class. Gowda et al. [5] introduced the scale-invariant feature transform (SIFT) and GLCM algorithms for classifying lung cancer in the IQ-OTH or NCCD dataset.

Despite the GLCM method, HOG is widely used to extract texture features for lung cancer classification. Ashwini et al. [6] applied multiple feature extractions consisting of GLCM, local binary pattern (LBP), and histogram of oriented gradients (HOG) to extract features for evaluating lung cancer. This method achieves high accuracy because it uses a diversity of feature extraction methods. The HOG method extracts features based on the gradient on edges but the edges of the tumor in the image are unclear. This affects the accuracy of lung cancer classification.

The texture-based method can be used to extract features for classifying lung cancers. Nonetheless, the effectiveness of those methods is insufficiently accurate because a characteristic of lung cancer is flat region which is more than texture regions.

In the shape-based methods, Nadkarni et al. [7] presented a shape feature extraction including area, perimeter, and eccentricity for detecting lung cancer in CT scan images. Extracted features are used to classify a normal or abnormal class of lung cancer. Moreover, Hoque et al. [8] used shape feature extraction consisting of area, circularity, and solidity to identify lung cancer in the early stage in CT images. Moreover, this method uses image enhancement and segmentation to increase the effectiveness of lung cancer classification. In the same way, Nawreen et al. [9] introduced lung cancer classification in CT scan images by using shape features such as area, perimeter, eccentricity, compactness, and circularity. These features are used to classify severity levels (benign and malignant) of lung cancers.

The shape-based feature extraction is simple and efficient. The feature dimension of this method is low. However, this method is sensitive to objects which are similar to tumors. This problem affects the accuracy of classification.

As mentioned in the previous paragraph, the methods for feature extraction are developed to improve accuracy. However, both texture- and shape-based feature extractions have limitations using CT images. Therefore, a binary count ratio (BCR) is proposed to enhance the efficiency of lung cancer classification. The proposed BCR method has three main steps: image pre-processing, feature extraction, and lung cancer classification.

This work was supported by Faculty of Informatics, Burapha University, Thailand.

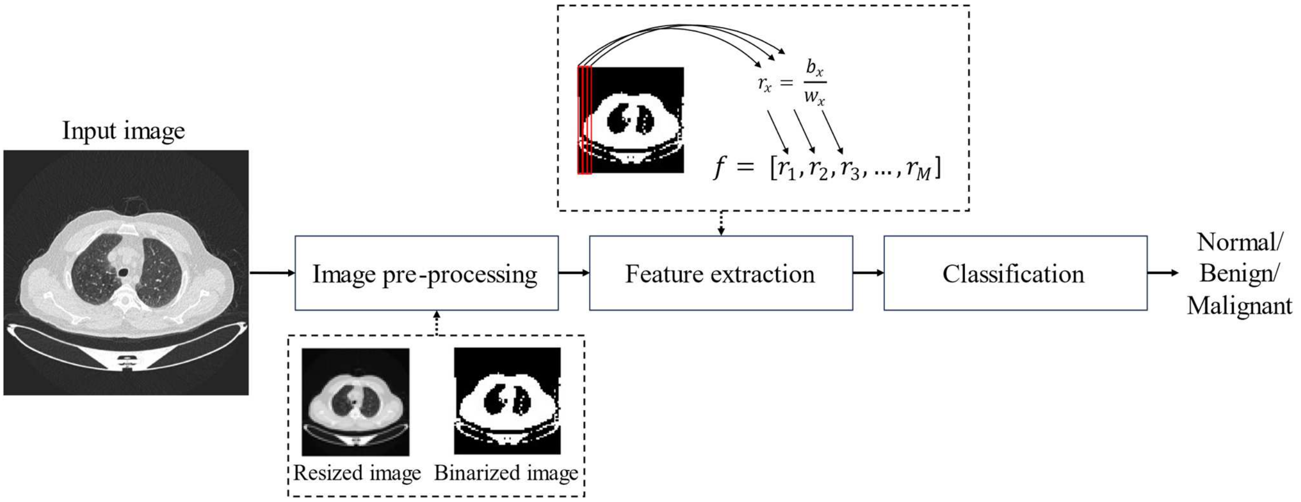

Fig. 1. Overall framework of the proposed BCR method.

In the second step, the binary count ratio features are calculated to represent the lung cancer areas. Finally, a Euclidean distance classifier is applied to classify features in lung cancer images. As a result, the proposed method achieves a high result compared to the traditional methods.

The remainder of this paper is organized as follows: Section II briefly presents the feature extraction problems. The proposed method is detailed in Section III. Section IV presents the experimental results and discussions, and the conclusion is explained in Section V.

II. Problem Formulation

In this section, we briefly describe the limitations of the existing methods of feature extraction for classifying lung cancer in CT scan images. Feature extraction can be divided into texture-based and shape-based methods.

A. *Texture-Based Method*

Texture feature extraction, such as gray level co-occurrence matrix (GLCM) [10], Gabor filter [11], and histogram of oriented gradients (HOG) [12], is widely used to capture the characteristics of lung cancer. The GLCM method examines the relationship of pixels on the spatial domain in images. After using GLCM, Haralick texture features [13], such as contrast, homogeneity, energy, correlation, and entropy, are calculated by statistical methods to reduce feature dimensions. The GLCM method extracts texture features which can be used as features for lung cancer classification. Nonetheless, texture features may not represent features of tumors.

In addition, a histogram of oriented gradients (HOG) is first introduced to detect humans in images. This method is efficient. After that, HOG is widely used for object detection and feature extraction. This method captures gradient information of objects in images. The gradients of HOG are used as texture features. HOG is applied to extract features for lung cancer classification. However, HOG is sensitive to unclear edges of images. Gradient information on the HOG method may not be sufficiently informative. This makes low effective in lung cancer classification.

Moreover, tumor regions have the characteristics of flat regions. Thus, shape features are more suitable than texture features for lung cancer classification.

B. *Shape-Based method*

Shape feature extraction is an approach to measure the shape and structure of objects in the region of interest. The approach. Examples of shape features are area and perimeter. These features are used because they can indicate the sizes of tumors. Nonetheless, shape features are low effective when objects are similar to tumors.

As previously mentioned, the problems of lung cancer classification are that the use of texture and shape-based methods is not suitable to classify lung cancers in CT scan images.

III. Proposed Method

The binary count ratio (BCR) is proposed to increase the effectiveness of lung cancer classification. The main idea of the proposed BCR method is extracting cancer features based on counting white pixels in CT images because these pixels represent cancer areas. Fig. 1 illustrates the proposed BCR method consisting of three main steps: image pre-processing, feature extraction, and classification. More details of the proposed BCR method are explained below.

A. *Image Pre-processing*

This step aims to prepare suitable CT images for extracting features. The input images of lung CT scan images are resized into *M*×*M* pixels to reduce the feature dimensions. Then each resized image is converted into a binary image *B* as given by

$$B(x,y) = \begin{cases} 1, if\ I(x,y) > T_x \\ 0, if\ I(x,y)\ \le T_x \end{cases}, \qquad (1)$$

In this step, the images are converted into binary images by means of adaptive thresholding using its column intensity means.

The adaptive threshold value T_x is computed by

$$T_x = \frac{\sum_{x=1}^{M} I(x,y)}{M}, \qquad (2)$$

where T_x is the mean intensity value of each column in a resized image; It is used as a threshold value for transforming binary images; x represents the column index of an image. y is the row index of an image; M is the resized image dimension; I is a resized image.

B. Feature Extraction

To obtain lung cancer features, the binary count ratio r_x of each column on the image is computed by

$$r_x = \frac{b_x}{w_x}, \tag{3}$$

where b_y and w_y are the number of black pixels in column x and that of white pixels in column x.

Here, b_y and w_y are calculated by

$$b_y = \sum_{y=1}^{M}(B(x,y) = 0), \tag{4}$$

$$w_y = \sum_{y=1}^{M}(B(x,y) = 1). \tag{5}$$

Finally, the binary count ratio r_x of all image columns is used as a features vector as defined in

$$\boldsymbol{f} = [r_x, r_2, r_3, \dots, r_M], \tag{6}$$

where $\boldsymbol{f}$ is a feature vector of lung cancer and M presents the total dimension of features.

The high ratio r means that the number of black pixels in an image is high which represents a normal class of the image. Otherwise, the r value is low, it represents an abnormal class consisting of a high number of white pixels.

The features based on the proposed BCR method are used because they can represent lung cancer pixels. The ratios normalize the values of the features. Normalization makes the features more robust to different sizes of images.

C. Classification

This step classifies binary count ratio features to identify lung cancer images. In this paper, the Euclidean distance is employed to predict the features of normal, benign, or malignant class because this method is a very straightforward classification to implement and interpret that quantifies the similarity or dissimilarity between features in different classes. The minimum distance is considered to deciding classes.

IV. Experimental Results and Discussion

In this experiment, the IQ-OTH/NCCD lung cancer dataset [14] is used. Images in this dataset are lung CT scan images in the sizes of 512×512 pixels. The total number of this dataset is 1,097 images consisting of 416 normal, 120 benign, and 561 malignant images as shown in Fig. 2. In this paper, the size M of the image, which is resized, is 64.

To evaluate the effectiveness of all feature extraction methods, we implemented a k-fold cross-validation, $k = 5$, in all experiments. The precision, recall, F1-score, and classification accuracy are measured. The Euclidean distance classifier is used to classify lung cancer in all feature extraction methods.

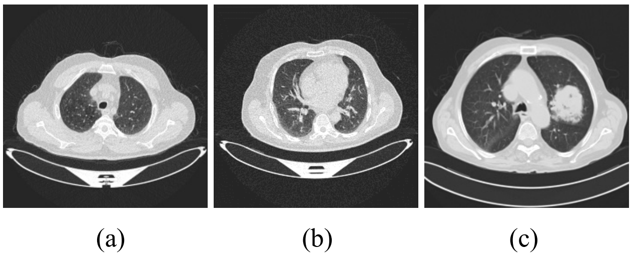

Fig. 2. Samples of CT scan images in three classes: (a) normal, (b) benign, and (c) malignant. The red rectangle is an abnormal region.

TABLE I. Effectiveness of the proposed method and the conventional methods.

Feature extraction method	Effectiveness (mean±S.D.)			
	Precision	*Recall*	*F1-score*	*Classification accuracy*
GLCM [10]	0.8282 ±0.02	0.8248 ±0.02	0.8263 ±0.02	0.8813 ±0.01
Gabor filter [11]	0.9481 ±0.01	0.9175 ±0.02	0.9308 ±0.01	0.9562 ±0.01
Shape feature extraction [7]	0.9367 ±0.02	0.9258 ±0.01	0.9309 ±0.01	0.9564 ±0.01
HOG [12]	0.9632 ±0.00	0.9417 ±0.01	0.9516 ±0.01	0.9676 ±0.00
Proposed BCR method	**0.9810** ±0.00	**0.9618** ±0.01	**0.9705** ±0.01	**0.9832** ±0.00

Bold numbers indicate the highest value.

Table I shows that the proposed method outperforms the other conventional methods in terms of precision, recall, F1-score, and classification accuracy.

When analyzing a confusion matrix of the proposed BCR method as shown in Fig.3, the proposed BCR method correctly predicts three classes which are greater than the conventional methods including GLCM [10], Gabor filter [11], shape feature extraction [7], and HOG [12] methods. The misclassification of the proposed BCR method is less than the conventional methods. Fig. 4 displays a confusion matrix of the GLCM method for classifying lung cancers. When considering the misclassification of GLCM, normal case of images is incorrectly predicted to be benign and malignant classes. Moreover, GLCM often incorrectly predicts benign and malignant classes to be normal classes. A confusion matrix of the Gabor filter method is shown in Fig. 5. The most misclassification of the Gabor filter method is that benign classes are classified to be normal classes. In the same way, shape feature extraction and the HOG method incorrectly classify benign classes to be normal classes as illustrated in Fig. 6 and 7.

It proves that the proposed BCR method can extract lung cancer features effectively because of the ability of extracting malignant tumor regions based on black and white pixels. A higher ratio of our method is that there are more black pixels than white pixels in images.

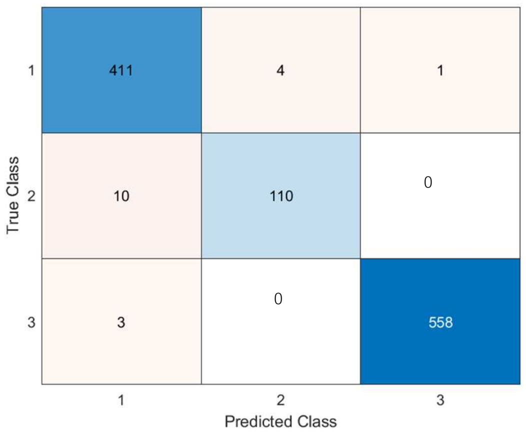

Fig. 3. A confusion matrix of the proposed BCR method.

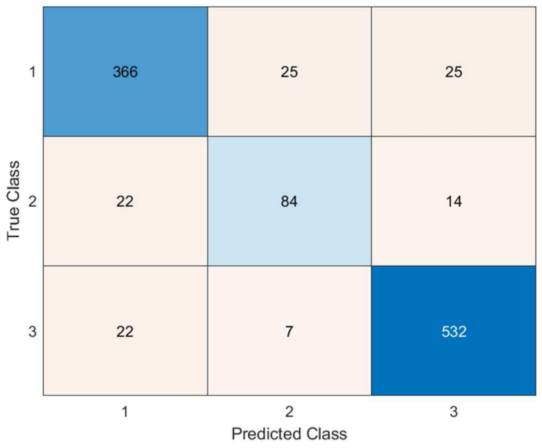

Fig. 4. A confusion matrix of the GLCM method [10].

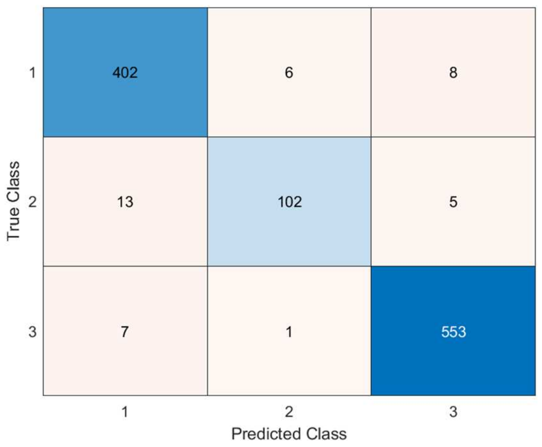

Fig. 5. A confusion matrix of the Gabor filter method [11].

This ratio can represent the normal classes of images. Otherwise, the lower ratio, which is more white pixels than black pixels, means that there are more tumors in lung CT scan images.

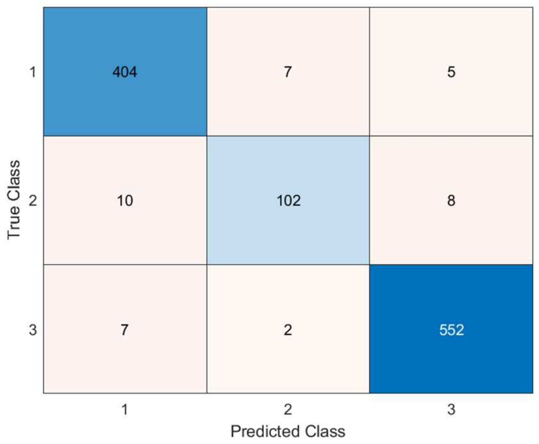

Fig. 6. A confusion matrix of the shape feature extraction [7].

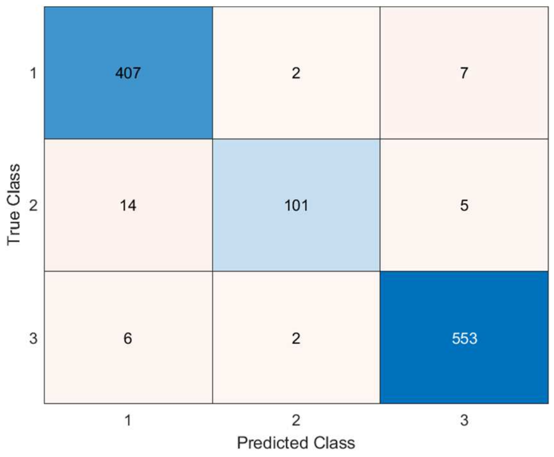

Fig. 7. A confusion matrix of the HOG method [12].

V. Conclusion

In this paper, the binary count ratio (BCR) method is proposed to improve the effectiveness of lung cancer classification in CT scan images. The main contribution of the proposed method is extracting cancer features based on counting white pixels in CT scan images because these pixels represent cancer areas. The proposed BCR method consists of three main steps: image pre-processing, feature extraction, and classification. First, CT scan images are resized to 64×64 pixels. Then the resized images are converted into binarized images by means of adaptive thresholding based on their column mean values. Second, the features of the binary count ratio are calculated to represent the lung cancer areas. Finally, a Euclidean distance is applied to classify features into a normal or benign or malignant class. The proposed method outperforms the conventional method in terms of precision, recall, F1-score, and classification accuracy. The reasons for the improvements achieved by our method are: (i) adaptive thresholds can separate cancer regions and (ii) binary count ratios can represent normal and abnormal cancer regions with black to white pixels. This leads to significant improvement in the effectiveness of lung cancer classification. In future work, the proposed method will be combined with a segmentation technique or another type of feature to increase classification accuracy.

Acknowledgment

We would like to sincerely thank Faculty of Informatics, Burapha university, Thailand for supporting finances and thank Hamdalla Alyasriy and Muayed AL-Huseiny for publishing lung cancer dataset.

References

[1] H. M. Rai and J. Yoo, “A comprehensive analysis of recent advancements in cancer detection using machine learning and deep learning models for improved diagnostics,” *J. Cancer Res. Clin. Oncol.*, vol. 149, no. 15, Aug 2023.

[2] R. Ankita, Ch. U. Kumari, M. J. Mehdi, N. Tejashwini, and T. Pavani, “Lung Cancer Image- Feature Extraction and Classification using GLCM and SVM Classifier,” *Int. j. eng. innov. technol.*, vol. 8, Sep 2019.

[3] Q. Firdaus, R. Sigit, T. Harsono, and A. Anwar, “Lung Cancer Detection Based On CT-Scan Images With Detection Features Using Gray Level Co-Occurrence Matrix (GLCM) and Support Vector Machine (SVM) Methods,” *2020 International Electronics Symposium (IES)*, Surabaya, Indonesia, 2020, pp. 643–648.

[4] A. Muayed, M. Furat, K. Enam, H. Zainab, F. Hamdalla, and F. A. Hamdalla, “Evaluation of SVM Performance in the Detection of Lung Cancer in Marked CT Scan Dataset,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 21, March 2021.

[5] M. A. S. Gowda and A. Jayachandran, "Triple SVM Integrated with Enhanced Random Region Segmentation for Classification of Lung Tumors," *Int J Adv Comput Sci Appl.*, vol. 13, no. 10, 2022.

[6] S. S. Ashwini, M. Z. Kurain, and M. Nagaraja, “Performance Analysis of Lung Cancer Classification using Multiple Feature Extraction with SVM and KNN Classifiers,” *2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, Tumkur, Karnataka, India, 2021, pp. 1–4.

[7] N. S. Nadkarni and S. Borkar, “Detection of Lung Cancer in CT Images using Image Processing,” *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2019, pp. 863–866.

[8] A. Hoque, A. K. M. A. Farabi, F. Ahmed and M. Z. Islam, “Automated Detection of Lung Cancer Using CT Scan Images,” *2020 IEEE Region 10 Symposium (TENSYMP)*, Dhaka, Bangladesh, 2020, pp. 1030-1033.

[9] N. Nawreen, U. Hany and T. Islam, “Lung Cancer Detection and Classification using CT Scan Image Processing,” *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, Rajshahi, Bangladesh, 2021, pp. 1–6, doi: 10.1109/ACMI53878.2021.9528297.

[10] P. Mohanaiah, P. Sathyanarayana and L. GuruKumar, “Image texture feature extraction using GLCM approach,” *Int. j. sci. res. publ.*, vol. 3, no. 5, pp. 1–5, 2013.

[11] D. Zheng, Y. Zhao, and J. Wang, “Features extraction using a gabor filter family,” *Proceedings of the sixth Lasted International conference, Signal and Image processing*, Hawaii, 2004.

[12] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* , San Diego, CA, USA, 2005, pp. 886-893, vol. 1.

[13] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural Features for Image Classification,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[14] H. Alyasriy and M. A. Huseiny, *The IQ-OTH/NCCD lung cancer dataset*. (2023). Mendeley Data, V4, doi: 10.17632/bhmdr45bh2.4.