HALAMAN MUKA dan DAFTAR ISI

https://www.tandfonline.com/toc/teta20/30/2?nav=tocList

Getting Started | Sistem Informasi Man... | Sistem Informasi Man... | E-learning | File Manager for mbcj... | Items where Subject is... | Text Similarities : Estim... | Membuat Pecarian Au... | Other Bookmark

Taylor & Francis Online

Log in | Register | Cart

Home ▶ All Journals ▶ Journal of Experimental & Theoretical Artificial Intelligence ▶ List of Issues ▶ Volume 30, Issue 2

Journal of Experimental & Theoretical Artificial Intelligence

Enter keywords, authors, DOI, ORCID et | This Journal

Advanced search | Citation search

Publish with us
Submit an article | About this journal

Explore
Browse all articles & issues | Latest issue

Subscribe
Alerts & RSS feed | + Purchase a subscription

Journal of Experimental & Theoretical Artificial Intelligence, Volume 30, Issue 2 (2018)

Volume 30, 2018 | Vol 29, 2017 | Vol 28, 2016 | V‹ ›

See all volumes and issues

sue 5 | Issue 4 | Issue 3 | Issue 2 | Issue 1

Browse this journal

- Latest articles
- Current issue
- List of issues
- Special issues
- Open access articles
- Most read articles

Download citations | Download PDFs | Download issue | Browse by section (All) | Display order (Default)

Articles

Article

Challenges in discriminating profanity from hate speech ›

Shervin Malmasi & Marcos Zampieri

Pages: 187-202

1353 Views

1 CrossRef citations

› Most cited articles

Published online: 13 Dec 2017

Abstract | Full Text | References | PDF (1538 KB) |

14 Altmetric

Original Article

Article

Classification of ground glass opacity lesion characteristic based on texture feature using lung CT image ›

M. M. Sebatubun, C. Haryawan & B. Windarta

Pages: 203-215

Published online: 01 Dec 2017

Abstract | Full Text | References | PDF (1227 KB) |

141 Views
1 CrossRef citations
0 Altmetric

Articles

Article

Cascade heterogeneous face sketch-photo synthesis via dual-scale Markov Network ›

Saisai Yao, Zhenxue Chen, Yunyi Jia & Chengyun Liu

Pages: 217-233

Published online: 04 Dec 2017

Abstract | Full Text | References | PDF (3357 KB) |

129 Views
0 CrossRef citations
0 Altmetric

Article

P-HS-SFM: a parallel harmony search algorithm for the reproduction of experimental data in the continuous microscopic crowd dynamic models ›

Khalid Mohammad Jaber, Osama Moh'd Alia & Mohammed Mahmod Shuaib

Pages: 235-255

Published online: 05 Jan 2018

Abstract | Full Text | References | PDF (2520 KB) |

113 Views
0 CrossRef citations
0 Altmetric

Article

Modeling agent's preferences by its designer's social value orientation ›

Inon Zuckerman, Kan-Leung Cheng & Dana S. Nau

149 Views
1 CrossRef citations

Pages: 257-277

Published online: 30 Jan 2018

Abstract | Full Text | References | PDF (2460 KB) |

0
Altmetric

Article

Induced simplified neutrosophic correlated aggregation operators for multi-criteria group decision-making ›

Rıdvan Şahin & Hong-yu Zhang

Pages: 279-292

Published online: 04 Feb 2018

Abstract | Full Text | References | PDF (1501 KB) |

171
Views

1
CrossRef citations

0
Altmetric

Article

A new chaotic multi-verse optimization algorithm for solving engineering optimization problems ›

Gehad Ismail Sayed, Ashraf Darwish & Aboul Ella Hassanien

Pages: 293-317

Published online: 02 Feb 2018

Abstract | Full Text | References | PDF (2171 KB) |

372
Views

4
CrossRef citations

0
Altmetric

Article

Almost periodic cellular neural networks with neutral-type proportional delays ›

Songlin Xiao

Pages: 319-330

Published online: 08 Feb 2018

Abstract | Full Text | References | PDF (1000 KB) |

106
Views

0
CrossRef citations

0
Altmetric

Article

Normative and descriptive rationality: from nature to artifice and back ›

T. R. Besold & S. L. Uckelman

Pages: 331-344

Published online: 07 Feb 2018

Abstract | Full Text | References | PDF (1198 KB) |

338
Views

1
CrossRef citations

10
Altmetric

Notes

Note

Discrete bacteria foraging optimization algorithm for graph based problems – a transition from continuous to discrete ›

Chiranjib Sur & Anupam Shukla

Pages: 345-365

Published online: 08 Jan 2018

Ready to submit?

Start a new manuscript submission or continue a submission in progress

Go to submission site

Submission information

- Instructions for authors
- Editorial policies

Editing services

- Editing services site

About this journal

- Journal metrics
- Aims & scope
- Journal information
- Society information

Editorial board

Journal of Experimental & Theoretical Artificial Intelligence

ISSN: 0952-813X (Print) 1362-3079 (Online) Journal homepage: http://www.tandfonline.com/loi/teta20

Classification of ground glass opacity lesion characteristic based on texture feature using lung CT image

M. M. Sebatubun, C. Haryawan & B. Windarta

To cite this article: M. M. Sebatubun, C. Haryawan & B. Windarta (2017): Classification of ground glass opacity lesion characteristic based on texture feature using lung CT image, Journal of Experimental & Theoretical Artificial Intelligence, DOI: 10.1080/0952813X.2017.1409285

To link to this article: https://doi.org/10.1080/0952813X.2017.1409285

Published online: 01 Dec 2017.

JOURNAL OF EXPERIMENTAL & THEORETICAL ARTIFICIAL INTELLIGENCE, 2017
https://doi.org/10.1080/0952813X.2017.1409285

Taylor & Francis
Taylor & Francis Group

Classification of ground glass opacity lesion characteristic based on texture feature using lung CT image

M. M. Sebatubun[a], C. Haryawan[b] and B. Windarta[c]

[a]Department of Informatics Engineering, Sekolah Tinggi Manajemen Informatika dan Komputer AKAKOM, Yogyakarta, Indonesia; [b]Department of Information System, Sekolah Tinggi Manajemen Informatika dan Komputer AKAKOM, Yogyakarta, Indonesia; [c]Department of Radiological Sciences, Padangan General Hospital, Bojonegoro, Indonesia

ABSTRACT

Lung cancer causes a high mortality rate in the world than any other cancers. That can be minimised if the symptoms and cancer cells have been detected early. One of the techniques used to detect lung cancer is by computed tomography (CT) scan. CT scan images have been used in this study to identify one of the lesion characteristics named ground glass opacity (GGO). It has been used to determine the level of malignancy of the lesion. There were three phases in identifying GGO: image cropping, feature extraction using grey level co-occurrence matrices (GLCM) and classification using Naïve Bayes Classifier. In order to improve the classification results, the most significant feature was sought by feature selection using gain ratio evaluation. Based on the results obtained, the most significant features could be identified by using feature selection method used in this research. The accuracy rate increased from 83.33% to 91.67%, the sensitivity from 82.35% to 94.11% and the specificity from 84.21% to 89.47%.

ARTICLE HISTORY
Received 3 May 2017
Accepted 19 September 2017

KEYWORDS
GLCM; ground glass opacity; Naïve Bayes classifier

1. Introduction

Cancer is one of the causes of high mortality rate in the world. It is the growth and spread of abnormal cells which have typical characteristics. Cancer that has widely spread will usually cause death (National Cancer Institute, 2016). It has several types, namely breast cancer, lung cancer, blood cancer or leukemia, prostate cancer, skin cancer, etc. One type of cancers that most frequently identified is lung cancer. It becomes one of the primary causes of high mortality rate in the world. This is due to the fact that lung cancer is the most frequent type of cancer which attacks human and it is at first in a series of deadly cancers (National Cancer Institute, 2016). Mortality rate caused by lung cancer can be minimised if the symptoms and cancer cells have been detected early.

One of the techniques used to detect lung cancer is using the combination of Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) and High Resolution CT (HRCT). FDG is a radioactive substance/isotope to be absorbed by the body whose metabolism increases, as in cancer. The initial process of this FDG PET is the body injected with radioactive substances and then the organ with the abnormality will capture the substance and transmit it. Furthermore, this radioactive emission that will be captured by the sensor to be processed into images. HRCT is a process done outside the body, by utilising the power of different X-ray penetration in the human body including any abnormalities in the organs. Therefore, if FDG PET and HRCT are combined, they will be mutually reinforcing; the

CONTACT M. M. Sebatubun memey@akakom.ac.id

HRCT will show anatomical organ abnormalities and FDG PET will indicate excessive metabolism in malignant cells.

Another technique is widely known as X-ray (radiographs) or Computed Tomography (CT) scan. X-ray is the most often used technique for Lung cancer detection. X-ray image will show different results between normal and abnormal lungs. However, in some cases, lesion can't be detected because they are covered by the anatomical structure or because the poor image quality. To ensure whether the lesion is covered by the anatomical structure or not, doctor usually makes a deeper examination using CT scan because it can show the lesion more clearly. Using CT scan image, it is not only able to see the lesion more clearly, but also able to reveal the characteristics of the lesion. Lesion characteristics consist of some criteria such as pattern, shape and margin which have their own types (Li, Sone, Abe, Macmahon, & Doi, 2004). These characteristics are used to determine the level of malignancy of the lesion, but if there are mistakes in recognising the lesion characteristics, the diagnosis will be affected.

Actually, combination of FDG PET and HRCT can provide more accurate results in recognising the symptoms of the disease than CT scan. However, in Indonesia FDG PET and HRCT are still quite expensive so that only a few hospitals can provide it. In the other hand CT scan is owned by almost every hospitals. Hence, to overcome these limitations, this research will use the image of the CT scan, but this method allows mistakes when recognising lesion characteristics. To minimise those mistakes, image processing technique will be used. It can also be used as the second opinion by the radiologist.

Research about morphological characteristics in image processing is still rarely performed. But there are some previous researches conducted to recognise lesion characteristics using different methods. One of the characteristics identified is pattern. Pattern consists of three types, namely pure Ground Glass Opacity (GGO), mixed GGO and solid component (Li et al., 2004). GGO is an area with a slight increase in unclear homogeneous density located at the base of the bronchial or vascular edge on HRCT. Pathologically, GGO can be caused by partially air filling, interstitial thickening with inflammation, oedema, fibrosis, neoplastic proliferation, normal respiratory condition or increased blood volume on the pulmonary capillary. GGO is general condition and non-specific lung on HRCT. It can even occur in normal lung condition (benign) such as pneumonia, focal fibrosis and haemorrhage. Each type of GGO has different characteristics. Pure GGO is vague and has lower grey level intensity; mixed GGO has a mixture of pure GGO; solid component is an area which is more dense and white. Figure 1 displays the characteristic appearance of GGO pattern (Li et al., 2004).

Each type of GGO has different characteristics. Pure GGO is vague and has lower grey level intensity as shown in Figure 2(a), and mixed GGO has mixture of pure GGO and solid components as shown in Figure 2(b), meanwhile the solid component has an area which is more dense and white as shown in Figure 2(c) (Li et al., 2004).

There was a research conducted to identify the types of GGO pattern by searching for the features from each type of pattern based on the first order statistical features such as mean, standard deviation, skewness and kurtosis (Akram, Javed, Hussain, & Riaz, 2015; Katsumata, Itai, Kim, Tan, & Ishikawa, 2008). Another research has also been conducted using the second order statistical features

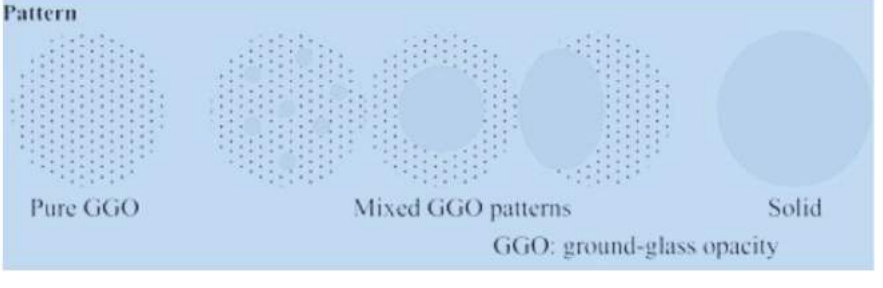
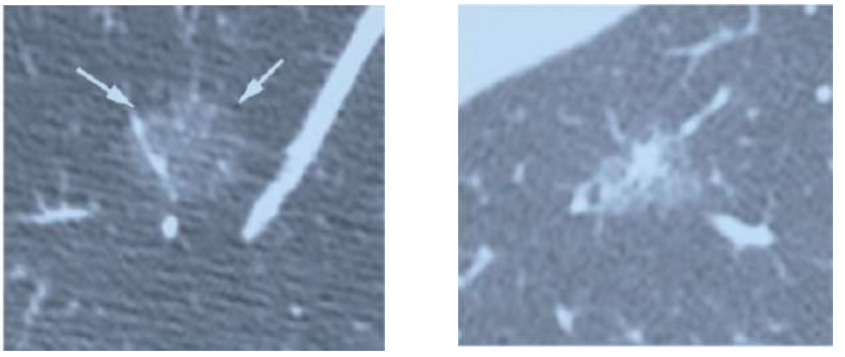
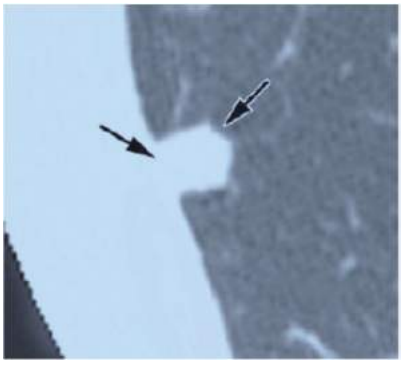


Figure 1. GGO pattern characteristics.



(a) *pure* OGG (b) *mixed* GGO



(c) *solid*

Figure 2. GGO characteristics. (a) *pure* GGO, (b) *mixed* GGO, (c) *solid.*

namely Grey Level Co-occurrence Matrices (GLCM) by calculating the four directions, namely energy, entropy, inertia and correlation (Yokota, Maeda, Kim, & Tan, 2014). GLCM is a method that can be used for texture-based feature extraction. GLCM is better than first order statistical features because it considers the relationship between two pixels of original image, while the first order statistical features do not consider the relationship of pixel neighbourhood (Nugroho et al., 2017). In addition, there are also researches that combine morphological features and grey level to distinguish the type of pattern. Morphological features considered are areas, compactness and irregularity. Grey level feature consists of mean intensity value and maximum intensity value (Bastawrous, Fukumoto, Nitta, & Tsudagawa, 2005). Recognition of GGO characteristics can be measured by the lesion texture which is light grey. Therefore, these methods are most frequently used because they are the suitable methods for texture-based feature extraction. After the feature extraction has been carried out, there is one step needed to measure how capable the feature extraction method to identify lesion characteristic. There are various classification methods that have been proposed in the previous researches, such as Artificial Neural Network (ANN) (Bastawrous et al., 2005; Katsumata et al., 2008; Yokota et al., 2014), Multilayer Perceptron (MLP) (Bhuvaneswari, Aruna, & Loganathan, 2014), Naïve Bayes Classifier (NBC) and K-Nearest Neighbour (KNN) (Kaur, Sharma, & Kaur, 2016; Sergeeva, Ryabchikov, Glaznev, & Gusarova, 2016; Yildiz, 2017; Zhou et al., 2015), Support Vector Machine and Random Forest (SVM) (Sergeeva et al., 2016), and so on. These methods have advantages and disadvantages depending on the data type used. Based on those researches, MLP provides better classification results compared to other methods (Bhuvaneswari et al., 2014), while others said that ANN is capable in providing

higher accuracy (Bastawrous et al., 2005; Katsumata et al., 2008; Yokota et al., 2014). Another research also said that KNN is able to provide higher accuracy than NBC because NBC is not effective in high dimensional feature vector (Yildiz, 2017). Based on the assumptions, this research will use NBC and add the feature reduction method for selecting the attributes or features that affect classification results significantly.

2. Literature review

Lesions in the lungs will automatically appear on the image of the CT scan image. To recognise it, the initial stage is carried out by features extraction. One of the research has detected areas of GGO by counting four statistical features from the density feature and shape features. The research used 31 thorax abnormal image datasets obtained from the Multi-row Detector CT (MDCT). The statistic feature calculated namely mean, standard deviation, skewness and kurtosis. From the fourth feature, 79% of accuracy level was obtained with 1.3 False Positives. Subsequently, shape feature was used to calculate the Minimum Directional Difference Filter (Min-DD) to reduce the level of false positives. Based on the result obtained, it was concluded that the proposed method could be used in the medical field (Katsumata et al., 2008). Other research has also detected GGO areas on 715 CT image slice containing 25 GGO nodules using some morphological features namely areas, compactness and irregularity. In addition, the research also used grey level features including mean intensity value and maximum intensity value. Classification process used template matching and obtained 92% of sensitivity with 0.76 per slice. After feature extraction, ANN method was used to reduce the level of false positive. The result show that it could be reduced from 0.76 to 0.25 per slice, but this method also reduced the sensitivity level to 84% (Bastawrous et al., 2005). Yokota et al. (2014) has also detected GGO areas on CT images based on statistical features by counting the four directions of GLCM namely energy, entropy, inertia and correlation. The proposed method applied to 31 CT images obtained from Lung Image Database Consortium (LIDC) and it gave 93% of accuracy. This was because the area of the blood vessel had been removed earlier using 3D line filter. Ye, Lin, Beddoe, and Dehmeshki (2007) have also detected GGO areas using 50 CT scan images contains 52 GGO nodules. The initial stage was pre-processing to remove the noise from the image using anti-geometric diffusion. Furthermore, geometric shape features were calculated for each pixel in lung areas which aimed to extract potential nodules. The next phase was removing the false positive area using Rule-based filtering. Based on the result, 48 nodules could be detected correctly and average level of detection accuracy was 92.3%, with the number of false positive around 12.7/scan (0.07/slice). Based on these results, the proposed method was potential for clinical applications.

Research on GGO can be used to assist the radiologist to determine the level of malignancy of the lesion but not much things can be carried out. Therefore, this study also make the recognition of the GGO characteristics by several processes, namely pre-processing, feature extraction and classification.

3. Proposed method

The data used was the data retrieved manually in 2013 until 2014 which was held at Sardjito General Hospital Yogyakarta. The amount of data obtained were 36 images of CT scan that had been classified by the radiologist. The image was taken with the thickness of 1 mm and 20 cm of scale, which was the standard used by radiologist in the diagnostic process. Out of the 36 images, there were 19 solid images and 17 images with mixed lesions. The CT image was primary lung cancer image so that there was one lesion in the lungs.

The data showed the area of the left and right lung that contained the lesions with various sizes. Besides lesion, there were also lung tissues that possessed similar characteristics with the lesions namely size, shape and colour. This research aims to identify the lesions characteristics from lung image that

can be used as the second opinion for determining whether or not a lesion is malignant. Research flow is shown in Figure 3.

3.1. Pre-processing

Before feature extraction phase, cropping process had previously been performed by a radiologist who had known the position of the lesion. Considering that, the research focused only on the lesion. Result of this process was used as input image for the feature extraction phase.

3.2. Grey level co-occurrence matrices

GLCM is one of the methods that can be used to perform texture feature extraction and it is the second order of statistical feature. Unlike the GLCM that considers the relation of neighbourhood pixels from original image, the texture measurement on the first order uses statistical calculation based on the pixel values of the original image such as variance which does not consider the relation of neighbourhood pixels. Let $f(x,y)$ is an image with size of N_x and N_y which has pixel with grey level probability (L level) and $\vec{r}$ is the offset vector. $GLCM_{\vec{r}}(i,j)$ is defined as the number of pixels with $j \in 1, \ldots, L$, which happens to offset $\vec{r}$ to pixels with $i \in 1, \ldots, L$ that can be expressed by Equation(1) (Nugroho et al., 2017) as follows.

$$GLCM_{\vec{r}}(i,j) = \#\{(x_1,y_1),(x_2,y_2) \in (N_x,N_y)X(N_x,N_y)|f(x_1,y_1) = j^{\vec{r}} = \overrightarrow{(x_2-x_1,y_2-y_1)}\} \quad (1)$$

In this case, # shows the number of set elements. Offset $\vec{r}$ is the direction or distance. Figure 4 shows the four direction of GLCM.

As an illustration, pixel neighbourhood can be selected to the right direction. One technique used to represent this relation is (1.0), which indicates the relation of two pixels lined horizontally with the pixel with value 1 and followed by pixel with value 2. Based on the composition, the number of pixels that comply this relation will be calculated. This is illustrated in Figure 5 (Nugroho et al., 2017).

Matrix in Figure 5 is named matrix framework. It needs to be processed into a symmetric matrix by adding it to the matrix transpose as shown in Figure 6 (Nugroho et al., 2017).

To eliminate the dependence on the image size, the values of GLCM elements need to be normalised until the sum is 1. Thus, Figure 6 will be as follows

$$\begin{bmatrix} \frac{4}{24} & \frac{2}{24} & \frac{1}{24} & \frac{0}{24} \\ \frac{2}{24} & \frac{4}{24} & \frac{0}{24} & \frac{0}{24} \\ \frac{1}{24} & \frac{0}{24} & \frac{6}{24} & \frac{1}{24} \\ \frac{0}{24} & \frac{0}{24} & \frac{1}{24} & \frac{2}{24} \end{bmatrix}$$

To obtain GLCM features, only a few scales are used, namely angular second moment (ASM) or energy, contrast, correlation and homogeneity. ASM is a homogeneity measure of the image and gives the numbers of squared elements. It can be calculated using Equation (2) (Nugroho et al., 2017).

$$\mathbf{ASM} = \sum_{i=1}^{L}\sum_{j=1}^{L} \mathbf{GLCM}(i,j)^2 \quad (2)$$

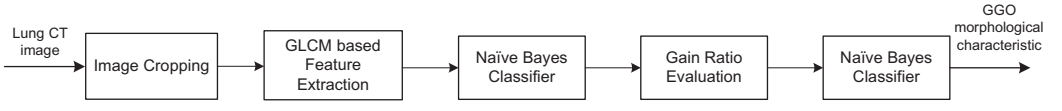


Figure 3. Research flow.

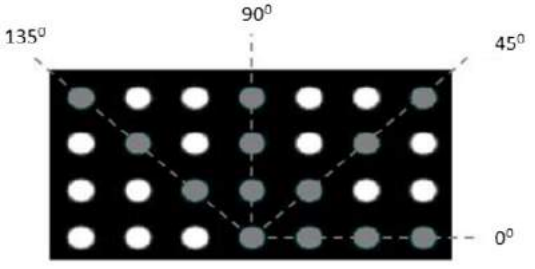


Figure 4. Four direction (0°, 45°, 90° dan 135°) of GLCM.

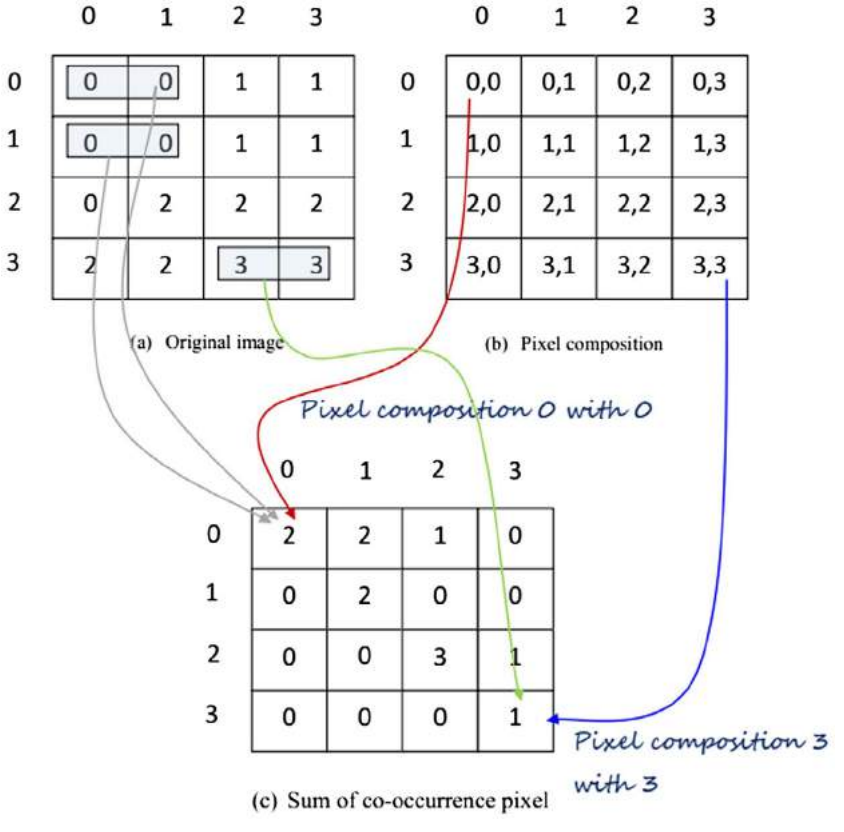


Figure 5. GLCM matrix initial determination based on two pixels.

$$\begin{bmatrix}2&2&1&0\\0&2&0&0\\0&0&3&1\\0&0&0&1\end{bmatrix}+\begin{bmatrix}2&0&0&0\\2&2&0&0\\1&0&3&0\\0&0&0&1\end{bmatrix}=\begin{bmatrix}4&2&1&0\\2&4&0&0\\1&0&6&1\\0&0&1&2\end{bmatrix}$$

Transpose → GLCM before normalized

Figure 6. Example of symmetric GLCM matrix formation.

In this case, L is the number of levels that have been used for computation.

Contrast is the measure of variety existence of grey level pixel of image calculated using Equation (3) as follows (Nugroho et al., 2017).

$$\textbf{Kontras} = \sum_{n=1}^{L} n^2 \{ \sum_{|i-j|=n} \textbf{GLCM}(i,j) \} \tag{3}$$

Correlation is a measure of linear interdependence of grey level values or how big the relation between one pixel and the neighbour pixel in the image is, calculated using Equation (4) as follows (Nugroho et al., 2017).

$$\textbf{Korelasi} = \frac{\sum_{i=1}^{L} \sum_{j=1}^{L} (i,j)(\textbf{GLCM}(i,j) - u_i' u_j'}{\sigma_i' \sigma_j'} \tag{4}$$

With :

$$\mu_i' = \sum_{i=1}^{L} \sum_{j=1}^{L} i * GLCM(i,j)$$

$$\mu_{ji}' = \sum_{i=1}^{L} \sum_{j=1}^{L} j * GLCM(i,j)$$

$$\sigma_j^2 = \sum_{i=1}^{L} \sum_{j=1}^{L} GLCM(i,j)(i - \mu_i')^2$$

$$\sigma_i^2 = \sum_{i=1}^{L} \sum_{j=1}^{L} GLCM(i,j)(i - \mu_i')^2$$

Homogeneity of inverse Different Moment (IDM) is a measure of proximity of elements distribution in GLCM which is calculated using Equation (5) (Nugroho et al., 2017).

$$\textbf{IDM} = \sum_{\mathbf{i}=1}^{\mathbf{L}} \sum_{\mathbf{j}=1}^{\mathbf{L}} \frac{\textbf{GLCM}(\mathbf{i},\mathbf{j})^2}{1 + (\mathbf{i} - \mathbf{j})^2} \tag{5}$$

3.3. Naïve Bayes classifier

In simple terms Naïve Bayes assumes that the feature value is not related to the existence of other features to the specified class variable. It assumes that each feature has a probability of contributing independently in spite of the presence or absence of other features. Although the design of Naïve Bayes looks simple, this method works quite well to solve complex problems. For some types of probability models, Naïve Bayes can be trained efficiently in supervised learning. Probability model for classification is a conditional model depending on variable C with a small number of results or classes that depend on several features variables X_1 to X_n calculated using Equation (6) (Zhou et al., 2015)

$$p(C|X_1 \ldots X_{n)} \tag{6}$$

However, if a feature has a great value, or if the number of n features is great, the model is not feasible on a probability table. Therefore, Equation (6) is reformulated as shown in the Equation (7) (Zhou et al., 2015)

$$p(C|X_1, \ldots, X_n) = \frac{p(C)p(X_1, \ldots, X_n|C)}{p(X_1, \ldots, X_n)} \tag{7}$$

In Bayesian analysis, the final classification is produced by combining two sources of information, prior of occurrence and probability of occurrence, to form the posterior probability using Bayes' rule, so that Equation (7) can be formulated into Equation (8) (Zhou et al., 2015)

$$Posterior = \frac{likelihood \times prior}{evidence} \tag{8}$$

In practice, interest is only in the numerator because the denominator does not depend on C and the feature values X_i provided. Therefore, the denominator is constant, and only need to increase the value of $p(C)p(X_1, \ldots, X_n|C)$.

Now, the assumption of conditional independence 'Naïve' will appear. It is assumed that each feature X_i is conditionally independent from the other features X_j for j = I, to category C defined in Equation (9) (Zhou et al., 2015)

$$p(X_1, \ldots, X_n|C) = \prod_{i=1}^{n} p(X_i|C \tag{9}$$

In which the probability $P(X_1|C, P(X_2|C \ldots P(X_n|C$ can be estimated with training samples. Based on this calculation, posterior probabilities of sample can be obtained for each class. Furthermore, based on the criteria of Bayesian maximum posterior, the class can be selected with the largest posterior probability as a class label.

3.4. Gain ratio attribute evaluation

The method used for data reduction is divided into two, namely, Wrapper and Filter. Wrapper model approach uses the classification method to measure the importance from a set of features. Furthermore, the feature is selected depending on the classifier model used. Filter approach precedes the actual classification process. It is independent from learning algorithm, simple computational, fast and scalable. Using filter method, feature selection process is conducted only once and then used as input for different classifier. Several feature ranking and feature selections have been used such as *Correlation-based Feature Selection (CFS), Principal Component Analysis (PCA), Gain Ratio (GR) Attribute Evaluation, Chi-square Feature Evaluation, Fast Correlation-based Feature Selection (FCBF), Information Gain, Euclidean Distance, i-test and Markov blanket filter*. Some of this filter method does not conduct feature selection but only conduct *feature ranking*. Therefore, this method is combined with the searching method to determine number of attributes (Karegowda, Manjunath, & Jayaram, 2010).

Gain Ratio method used in this study relates to Information Gain method. Information Gain is used to select test attributes on each node of the decision tree and tend to select the attributes that have great value. Variable *S* is composed of s sample data with different class of m. The expected information is required to classify the samples defined in Equation (10) (Karegowda et al., 2010).

$$I(S) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{10}$$

where p_i is the probability of random samples belong to class C_i and is estimated by s_i/s.

Attribute *A* has *v* distinct values. s_{ij} is the number of samples of class C_i in a subset S_j. S_j contains samples S that have value a_j of *A*. The entropy, or expected information based on the partitioning is part of subsets *A* defined in Equation (11).

$$E(A) = -\sum_{i=1}^{m} I(S)\frac{s_{1i} + s_{2i} + \ldots s_{mi}}{s} \tag{11}$$

The encoding information that would be gained by branching on A is defined in Equation (12).

$$Gain(A) = I(S) - E(A) \tag{12}$$

Gain Ratio is used to normalise the information gain using the values defined in Equation (13).

$$SplitInfo_A(S) = -\sum_{i=1}^{v}(|S_i|I|S|)log_2(|S_i|I|S|) \tag{13}$$

Value in Equation (13) shows the information generated by splitting the training data-set S into v partitions corresponding to v outcomes of a test on the attribute A. Gain Ratio is defined by Equation (14).

$$\text{Gain Ratio}(A) = Gain(A)/SplitInfo_A(S) \tag{14}$$

The attribute with the highest gain ratio is selected as the splitting attributes.

3.5. Measurement Index

Measurement conducted in this study is to determine the level of achievement of the processes that have been carried out. The measurements conducted are performance measurements of feature extraction and classification method. The measurements of the classification process are determined by the following value (Nugroho et al., 2017):

(1) Accuracy

The accuracy rate from classification can be obtained by counting the total of correct classification and divided by the total of different classification with the target of all classes. Accuracy is defined in Equation (15).

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{15}$$

With TP (True Positive) as the total of correct data on target classified correctly, TN (True Negative) is the total of incorrect data on target classified incorrectly, FP (False Positive) is the representation of the total of incorrect data on target classified correctly and FN (False Negative) is the representation the total of correct data on target classified incorrectly in the system. These values will appear in confusion matrix.

(2) Sensitivity

Sensitivity is a measure of the ability of the system to make predictions on data which is assumed to be true according to the TPR (True Positive Rate). Sensitivity can be formulated in Equation (16).

$$sensitivity = \frac{TP}{TP + FN} \tag{16}$$

(3) Specificity

Specificity is a contrary to the sensitivity that makes predictions to the data which considered as incorrect according to TNR (True Negative Rate). Specificity can be formulated in Equation (17).

$$specificity = \frac{TN}{TN + FP} \tag{17}$$

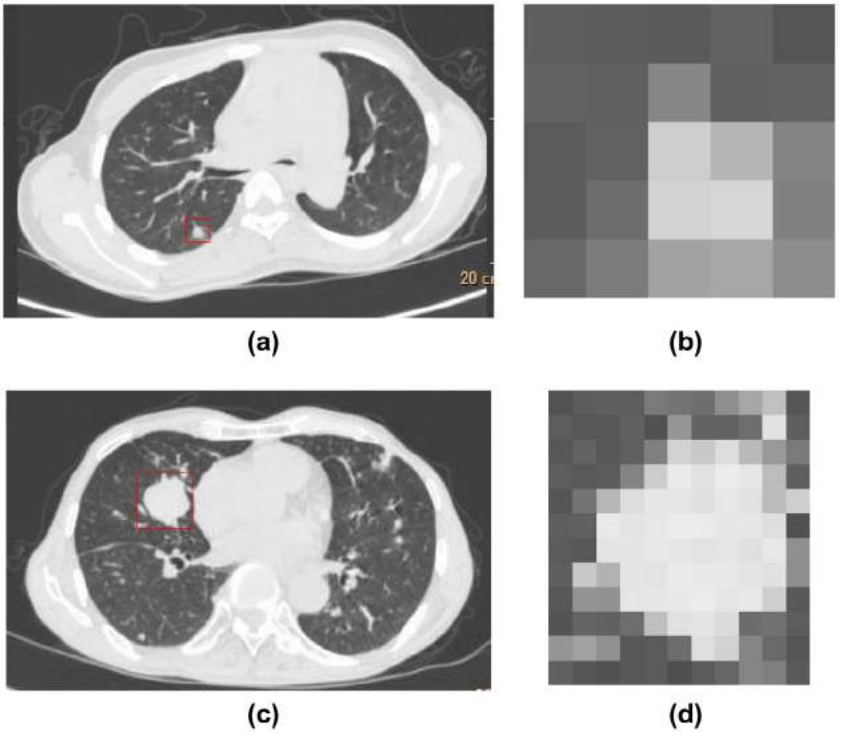


Figure 7. Original image (a) and (c), image cropping (b) and (d).

4. Result and discussion

GGO lesion characteristic recognition process using GLCM is considered as the most suitable method because it is a texture-based feature and is capable to recognise the characteristics of the object based on the density or grey level of the object. The results of the feature extraction are the numbers correspond to the measured features. Figure 7 is one of the original images that have been cropped by radiologist manually.

Figure 7(a) and (c) are the original images and Figure 7(b) and (d) are the results of image cropping. Figure 7(a) is an image with mixed characteristics and Figure 7(c) is an image with solid characteristics. Based on the data obtained, the lesion appears different size in the same characteristics or different characteristics. Therefore, the images which were processed in extraction phase have different sizes. The feature extraction phase generates a number of values in which one image has four features and each feature has four different directions. When it is calculated one image has 16 different features. The result of feature extraction was used as an input in classification process using Naïve Bayes. Table 1 is a confusion matrix of classification using Weka 3.6.

Table 2 is a confusion matrix that shown the results of classification using Naïve Bayes Classifier and obtained TP rate = 14, TN rate = 16, FN rate = 3 and FP rate = 3. It means that out of 19 solid images, Naïve Bayes was able to recognise as many as 16 solid images. The other three was recognised as mixed

Table 1. Confusion matrix of classification.

		Target	
		Mixed	Solid
Prediksi	Mixed	14	3
	Solid	3	16

Table 2. Results of feature reduction.

Ranked	Feature
0.69	Energy45
0.6	Energy90
0.593	Energy0
0.591	Energy135
0.516	Homogeneity90
0.452	Homogeneity45
0.452	Homogeneity0
0.416	Contrast135
0.397	Homogeneity135
0.396	Correlation45
0.359	Correlation135
0.322	Contrast90
0.308	Correlation90
0	Constrast45
0	Correlation0
0	Contrast0

images. Furthermore, out of 17 mixed images, Naïve Bayes was able to recognise 14 images as mixed images. Meanwhile the rest was recognised as solid images. Accuracy, sensitivity and specificity could be calculated based on the confusion matrix. For GGO recognition, the accuracy obtained was 83.33% with 82.35% sensitivity and 84.21% specificity.

Based on the features used, it was possible that there were features that highly affected the results of classification and there were also features that did not affect the classification result. Therefore, feature reduction was conducted using Gain Ratio Attribute Evaluation and it obtained a set of features starting from the most significant features to the features that are not significant. The result of feature reduction is shown in Table 2.

Based on Table 2, it can be seen that the most significant feature which can affect the results of classification is energy 45° with 0.69 significance rate/ while contrast 45°, contrast 0° and correlation 0° do not affect the results of classification. After feature reduction, the classification has been conducted using energy 45° features and the result is shown in Table 3.

Table 3 is confusion matrix that shows the classification results using Naïve Bayes Classifier and obtains TP = 16, TN = 17, FN = 1, FP = 2. It means that from 19 solid images, Naïve Bayes was able to recognise 17 images as solid images and 2 other images as mixed images. Furthermore, from 17 mixed images, Naïve Bayes was able to recognise 16 images as mixed images, while 1 other image was recognised as solid image. Based on the confusion matrix, 91.67% accuracy was obtained with 94.11% sensitivity and 89.47% specificity. It means that there was an increase in accuracy, sensitivity and specificity rate when energy45° features were used. This indicated that the result of feature reduction using Gain Ratio Attribute Evaluation was able to determine the most significant features.

The method proposed in this study was compared with the method that had been used in the previous studies. GLCM was the second order of texture-based feature extraction and considered to be able to provide better results than the first order of texture features. Therefore, GLCM was compared with the first order feature using mean intensity, standard deviation, skewness, energy, entropy and smoothness. In the next phase, these features were classified using Naïve Bayes Classifier. To improve the classification results, feature reduction using Gain Ratio Attribute Evaluation was done and obtained

Table 3. Confusion matrix after feature reduction.

		Target	
		Mixed	Solid
Prediksi	Mixed	16	1
	Solid	2	17

Table 4. Compared classification results.

	First order texture feature		Second order texture feature	
	Before feature reduction (%)	After feature reduction (%)	Before feature reduction (%)	After feature reduction (%)
Accuracy	80.56	83.33	83.33	91.67
Sensitivity	88.24	94.12	82.35	94.11
Specificity	73.68	73.68	84.21	89.47

the most significant feature namely energy feature with 0.496 significance rate. At last, classification using energy feature and the comparison results are shown in Table 4.

Based on Table 4, it can be seen that the sensitivity before feature reduction in the first order is higher than in the second order, but accuracy and specificity in the first order are lower than in the second order. After feature reduction, it can be seen that second order features provide the same sensitivity as first order but the accuracy and specificity of second order are higher than the first order.

5. Conclusion

This study proposed the second order texture-based feature extraction method which could be used to identify the characteristic of GGO lesion, which consisted of two types namely solid and mixed. After feature extraction, second phase was feature reduction that was used to find the most significant features and able to identify characteristics of lesion well. The results showed that energy45° was the most significant feature to classification result with the acquisition of 91.67% of accuracy rate, 94.11% of sensitivity and 89.47% of specificity. It also showed that the proposed feature reduction method was able to improve the classification results. The results were then compared to the first order texture feature and it obtained 83.33% of accuracy, 94.12% of sensitivity and 73.68% of specificity. This suggested that GLCM method provided better results than the first order feature.

Acknowledgement

Authors would like to express gratitude to the Dr Sardjito Public Hospital that has provided the CT scan results data. High appreciation goes to Dr Budi Windarta who has been willing to be a great resource in this study. Author also thanks STMIK AKAKOM Yogyakarta that has financially supported this study.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Akram, S., Javed, M. Y., Hussain, A., & Riaz, F. (2015). Intensity-based statistical features for classification of lungs CT scan nodules using artificial intelligence techniques. *Journal of Experimental & Theoretical Artificial Intelligence, 27*(6), 737–751. doi:10.1080/0952813X.2015.1020526

Bastawrous, H. A., Fukumoto, T., Nitta, N., & Tsudagawa, M. (2005). Detection of ground glass opacities in lung CT images using gabor filters and neural networks. *IEEE Instrumentation and Measurement Technology Conference Proceedings* (Vol. 1, pp. 17–19). Ottawa, Ont., Canada: IEEE. doi:10.1109/IMTC.2005.1604111

Bhuvaneswari, C., Aruna, P., & Loganathan, D. (2014). Performance evaluation of feature extraction and feature selection for. *Journal of Theoretical and Applied Information Technology, 61*(2), 389–394.

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management, 2*(2), 271–277.

Katsumata, Y., Itai, Y., Kim, H., Tan, J. K., & Ishikawa, S. (2008). Automatic detection of GGO candidate regions by using artificial neural networks from thoracic MDCT images. In *Proceedings of the 3rd international conference on innovative computing information and control*, ICICIC'08, 0–4. Dalian, Liaoning, China: IEEE. doi:10.1109/ICICIC.2008.177

Kaur, H., Sharma, A. S., & Kaur, S. (2016). Liver tissue classification for ultrasound images. In *Proceedings of the international conference on advances in computing, communication and automation* (p. 978–1–5090–3480–2/16). Bareilly, India: IEEE

Li, F., Sone, S., Abe, H., Macmahon, H., & Doi, K. (2004). Malignant versus benign nodules at CT screening for lung cancer: Comparison of thin-section CT findings. *Radiology, 233*(1), 793–798. doi:10.1148/radiol.2333031018

National Cancer Institute. (2016). National Cancer Institute. Retrieved from October 6, 2016, from http://www.cancer.gov/

Nugroho, H. A., Sebatubun, M. M., & Adji, T. B. (2017). Ground glass opacity lesion morphology extraction in primary lung cancer. *International Journal of Medical Engineering and Informatics*, *9*(4), 398–411. doi:10.1504/IJMEI.2017.10005939

Sergeeva, M., Ryabchikov, I., Glaznev, M., & Gusarova, N. (2016). Classification of pulmonary nodules on computed tomography scans. Evaluation of the effectiveness of application of textural features extracted using wavelet transform of image. In *Proceedings of the conference of open innovation association,* FRUCT (Vol. 2016–Septe, pp. 291–299). St. Petersburg, Russia: IEEE. doi:10.1109/FRUCT-ISPIT.2016.7561541

Ye, X., Lin, X., Beddoe, G., & Dehmeshki, J. (2007). Efficient computer-aided detection of ground-glass opacity nodules in thoracic CT images. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology,* (pp. 4449–4452). Lyon, France: IEEE. doi:10.1109/IEMBS.2007.4353326

Yildiz, K. (2017). Dimensionality reduction-based feature extraction and classification on fleece fabric images. *Signal, Image and Video Processing, 11*(2), 317–323. doi:10.1007/s11760-016-0939-9

Yokota, K., Maeda, S., Kim, H., & Tan, J. K. (2014). Automatic Detection of GGO Regions on CT Images in LIDC Dataset Based on Statistical Features. In *Proceedings of the Soft Computing and Intelligent Systems (SCIS), 2014 Joint 7th International Conference on and Advanced Intelligent Systems (ISIS), 15th International Symposium* (pp. 1374–1377). Kitakyushu, Japan: IEEE.

Zhou, X., Wang, S., Xu, W., Ji, G., Philips, P., Sun, P., & Xhang, Y. (2015). Detection of pathological brain in MRI scanning based on wavelet-entropy and naive bayes classifier. In F. Ortuño, I. Rojas (Eds.). *Bioinformatics and Biomedical Engineering. IWBBIO 2015. Lecture Notes in Computer Science*, vol 9043. Springer, Cham. doi:10.1007/978-3-319-16483-0