Multiclass Classification Application using SVM Kernel to Classify Chest X-ray Images Based on Nodule Location in Lung Zones

Mohd Nizam Saad[1], Zurina Muda[2], Noraidah Sahari[2], Hamzaini Abd. Hamid[3]
[1]*School of Multimedia Technology and Communication, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia.*
[2]*Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.*
[3]*Radiology Department, National University Medical Center Malaysia, Bandar Tun Razak, Kuala Lumpur, Malaysia.*
nizam@uum.edu.my

***Abstract*—Support Vector Machine (SVM) has long been known as an excellent approach for image classification. While many studies have reported on its achievement, yet it still weak to handle multiclass classification problem because it is originally designed as a binary classification technique. It is challenging task to transform SVM to solve multiclass problems like classifying chest X-ray images based on the lung zone location. Classified X-ray images improved image retrieval hence reducing time taken to assessed back the images. Realizing this difficulties, therefore, we proposed an application method for multiclass classification using SVM kernel to classify chest X-ray images based on nodule location in lung zones. The multiclass classification experiment is performed using four popular SVM kernels namely linear, polynomial, radial based function (RBF) and sigmoid. Overall, we obtained high classification accuracy (>90%) for three classifiers that are RBF, polynomial and linear kernel while sigmoid kernel classifier is only moderately good at 82.7% accuracy. Besides, values in the confusion matrices revealed that the RBF and polynomial classifiers managed to classify test data into all classification classes. Conversely, classifiers based on linear and sigmoid kernel have missed at least one classification class. Since each classifier work differently based on their kernel types, we noticed that it is better to view them as a complimentary rather than treating them as competing options. This condition also revealed that we can modify the original SVM classification method to handle multiclass classification problem.**

***Index Terms*— SVM Kernel; Image Classification; Chest X-ray Image.**

I. Introduction

The Chest X-ray (CXR) has been recorded to be the most medical image produced among others. It comprises almost one third of all radiology images produced in hospital [15]. The image is vastly produced because it is the easiest and cheapest radiology procedure to make, yet it is very important to diagnosis any abnormalities at the early stage of the treatment.

CXR is used to identify unusual objects found in chest anatomies such as the lung, mediastinum and ribs. For the lung, the most common radiology procedure done using CXR image is to detect the lung nodule (or pulmonary nodule). Typically, radiologist detect lung nodules by scanning the lung area which consumed the most space in CXR and normally located at the center of chest. To ease the detection process, radiologists divided the lung area into six parts namely Left Upper Zone (LUZ), Left Middle Zone (LMZ), Left Lower Zone (LLZ), Right Upper Zone (RUZ), Right Middle Zone (RMZ) and Right Lower Zone (RLZ). These zones; as depicted in Figure 1, are formed by dividing the lung area into two horizontal divisions and three vertical divisions (2 x 3 division) [16]. The lung zones help radiologist to approximate the nodule closer so that further diagnosis routine can be carried out to help the patient.

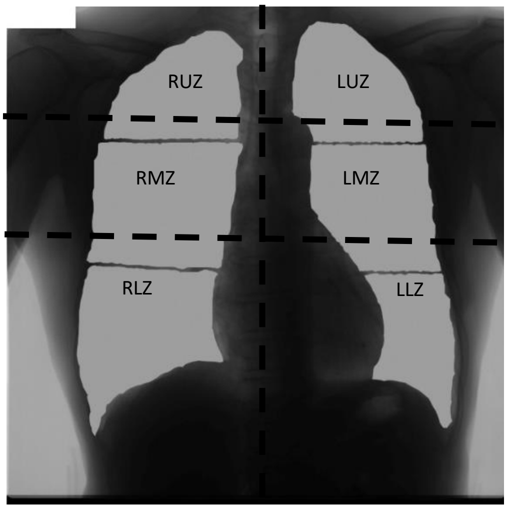

Figure 1: The lung zones in CXR image

In computer-aided diagnosis system, the lung nodule location can be traced by their coordinate in the image. Meanwhile, the presence of lung zones further assist radiologist to detect the nodule. Ideally, any CXR images whose nodule location have already been identified should be grouped together based on location feature to ease image retrieval. This ensures effective image search and access to the image repository. However, before such privilege can be achieved, these images need to be classified so that images with common features can be grouped accordingly.

In the field of image processing, image classification refers to the process of relating image attributes to the known features and the algorithm used to influence the classification is known as image classifiers [1]. These classifiers; which are driven by machine learning algorithm will build up predictive models to map the image features into predefined groups and classes. So far the models are represented either as classification rules, decision trees or mathematical functions [2] (such as SVM).

SVM starts as a part of statistical learning theory and later being modified to become a supreme method for image classification [3]. SVM has already been applied to classify image features for various usages like recognizing image features from multimodal devices [4], categorizing and pre-filtering image to reduce search space [5] and annotating image automatically based on specified image features [6]. Although SVM is popular among researchers for image classification, yet it still has loophole when dealing with multiclass classification. This due to its nature as a binary classification method where it is originally designed to classify features only between two classes (-1 ,1) at one time [7]. Meanwhile in reality, images like the CXR contain many features depending on how these features are extracted. Even the common low level image features contain three features that are color, texture and shape [8]. Therefore, using SVM for multiclass classification would potentially problematic with strategies to reduce the multiclass problem to a set of binary problems are typically sought to extend its basic binary approach [9].

Realizing the limitations as discussed in the prior paragraph, we propose an application method for multiclass classification using SVM to classify the CXR images based on the nodule location in the lung zones. By occupying the propose method, an improved SVM image classifier can be produced which is powerful enough to classify CXR images. As a mean for sharing our experience in working with the method, the rest of the paper is layout as follow. Section 2 explain SVM binary classifier, Section 3 discusses multiclass classification method for classifying CXR images based on the nodule location in lung zones. Section 4 discusses about the image dataset for the study while Section 5 elaborate the experiment that we have conducted. Section 6 presents the result based on the experiment and finally, Section 7 conclude all the works that we have done in the study.

II. SVM Binary Classifier

SVM is a supervised machine learning method that has been known as a high effectiveness approach for image features classifications. By using its kernel mapping such as linear, polynomial, radial based function (RBF) and sigmoid, SVM can classify both linear and non-linear data. Meanwhile, for researchers, this method can be found and occupied in almost every image processing tools to assist them classifying image features based on certain attributes. The advantage of SVM over other classifiers is that it achieves optimal class boundaries by finding the maximum distance between classes [10]. To achieve the objective, an SVM classifier will find a hyperplane from a training set of sample that can separate the largest portion of sample of the same class from all other sample. Figure 2 shows simple features classification using SVM.

In Figure 2, the aim is set to separate features in two classes that are solid and open diamond representing the classes of y_i=+1 and y_j=-1 respectively with linear hyperplane. The support vector is encircled and lie on the two planes, P1 and P2. The optimal separating hyperplane lies between and parallel with P1 and P2. By separating the features into two identical classes as shown in the figure, access to each class can be made easy and this can be very useful for further image processing tasks.

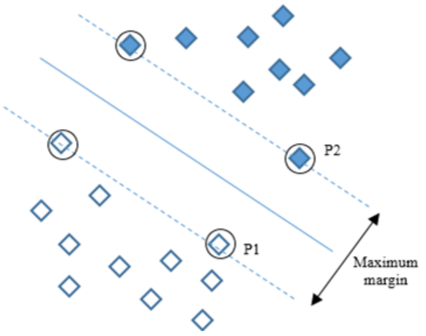

Figure 2: Simple features classification with SVM

It is a dream for researchers to extract and group image features as good as those shown in Figure 2. However, in reality, image features do not sit properly into separated groups because most of them are mixed altogether. Moreover, as highlighted in the prior section, image features normally exist in numerous classes and separating them with such a linear style would not be very effective. Let us consider group of features as shown in Figure 3. In this figures, there are five classes of objects that are trapezium, triangle, diamond, rectangle and pentagon.

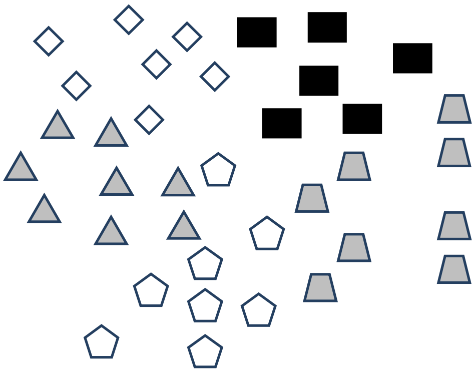

Figure 3: Mixed features poses classification problem with SVM binary classifier

Based on the condition of shape classes in Figure 3, it is almost impossible to separate them with single SVM binary classifier. Almost all features are mixed with others although some remain in their class. This problem is completely different than those shown in Figure 2 and this is what usually discovered during image features classification. Therefore, an excellent multiclass classifier is essential in to classify mixed image attributes into groups of homogeneous class.

III. SVM Multiclass Classification for the CXR Images Based on the Lung Nodule Position

The multiclass classification problem refers to assigning each of the observations (CXR images) into one of the predefined K classes [11]. There are two suggested approaches to transform the binary SVM classifier into a multiclass classifier i.e. one-against-all and one-against-one [9]. The first approach can be considered as the most common technique applied to solve multiclass problem with binary SVM classifier [12]. In this approach, each of the binary classifier is trained to separate one class from the rest. For instance, let say that we want to classify the image color into three primary colors that are red, green and blue. By using one-against-all, classification would be effected by classifying red color features against non-red color (green and blue), green against non-green (red and blue) and blue against non-blue (red and green). Later, all classification result will be calculated and analyzed.

On the other hand, the one-against-one approach involves constructing a series of classifiers or machines for each pair of classes. This approach is a bit different than the first one because it requires n(n-1)/2 classifiers to be applied to each pair of classes. For instance, let us consider a multiclass classification problem to classify image color into secondary image color; CMYK that include four color classes cyan, magenta, yellow and black. By using the given formula, we need six (derived from 4(4-1)/2) series of classifiers to solve the classification problem. The six classifiers include C and M, C and Y, C and K, M and Y, M and K, and Y and K. Afterwards, a strategy to handle each paired classes derived like max-win technique should be applied. In order to ease reader understanding towards the multiclass classification method, Figure 4 illustrates the classification process as recommended by [10].

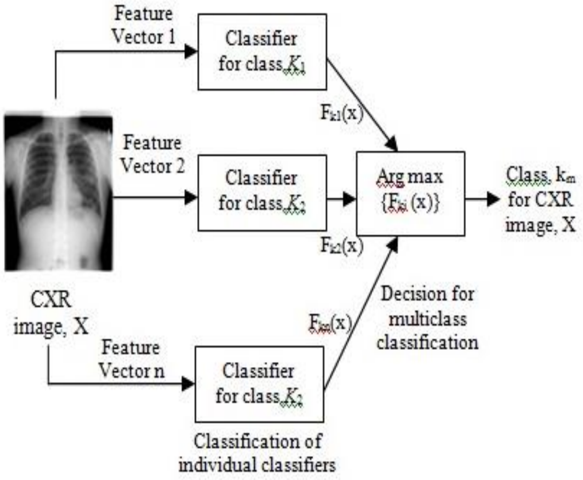

Figure 4: CXR classification process

In Figure 4, the image features must be extracted and categorized based on well predefined classes from the image source before the multiclass classification can be done. They [10] added that the complete classification task is done in two levels process; the base level consists of multiple binary classifiers while the second level fuses the decisions from the base level classifiers.

We can start the classification process by creating a classifier to handle each feature class. The number of the classifier depends on the number of the features classes created. Later, decision for the multiclass classification which can be presented by several techniques like classification accuracy and confusion matrix can be fused. The output from the classification process will be the classified image features where this features can be used for other image processing tasks.

Based on explanation regarding the multiclass classification from the prior paragraph, we considered the following approach to classify the nodule location in lung zones for the CXR images. First, we will use the nodule coordinate in the CXR image as the image features space. In other words, this coordinate will be the classification input to determine where the nodule is located in the lung zones. Secondly, since the nodule is probably distributed in one of the six lung zones (i.e. it might be in LUZ, LMZ, LLZ etc.), therefore, these zones will be treated as the classification classes for our image classification application. Thirdly, we assumed that there is no significant different between one-against-all and one-against-one approach when applying SVM to solve multiclass classification problem because each approach was designed to solve the same problem. Unless we are using different machine learning method (like kNN and Bayes Naïve) or different image dataset than the comparison might be meaningful. Instead, we will create classifiers from all SVM kernels and then compare which of these classifiers performed the best to solve the image multiclass classification problem.

IV. Image Dataset

The CXR images used in this study were downloaded from a public chest radiograph dataset of Japan Society of Radiological Technology (JSRT) [13]. There are 247 CXR images available in the dataset which are clustered into two parts that are images with lung nodule (154 images) and images without lung nodule (93 images). The images were scanned from films and defined with standard resolution of 2048 x 2048 pixels with 12 bit gray levels. JSRT also provides a text file (.txt) for additional information regarding these images which contain useful information like the patient age, gender, diagnosis and location of the nodule. During the classification experiment, we used only CXR images with lung nodule. Figure 5 shows an example of CXR image with lung nodule taken from JSRT dataset.

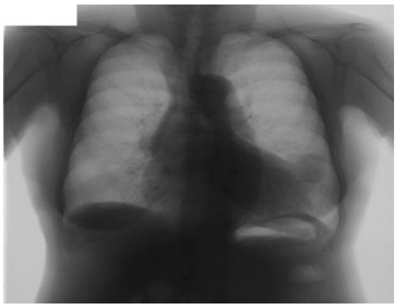

Figure 5: CXR image with nodule from JSRT dataset

V. Experiment

The aim of the experiment is to classify CXR images based on the nodule location in the lung zones using SVM multiclass classification. List of nodule coordinates (154 coordinates) were taken from the text file downloaded together with the image dataset from the JSRT website. These coordinates are scaled down to quarter because the original is too big as they are referring to the actual nodule location in the CRX images. Thus, reducing the coordinate and the image would be the best way to manipulate the images. We used Matlab as the tool to execute the multiclass classification with additional SVM library called LIBSVM [14].

There are six tasks needed to run multiclass classification [7]. First, the classification data must be transformed into SVM package which means the data must be transformed into real numbers. Since our classification data are in the form of nodule coordinate (x-axis and y-axis), therefore there is no need to do data transformation. Secondly, data scaling task must be performed in order to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. In their paper, they [7] recommended that the linear scaling for each attributes is set to the range of [-1,+1] or [0,1]. Using this recommendation, we divided the nodule coordinates by 1000 so the numerical range falls between the range of zeros to one [0, 1]. For example, the original coordinate for the first CXR image is (1634,692), reducing the coordinate to quarter yield (409,173) then applying the data scaling makes the coordinate to become (0.409,0.173).

The third task in multiclass classification is to select the kernel type. There are four famous kernels in SVM classification that include linear, polynomial, RBF and sigmoid [7]. These kernels can be attained by the following model:

Linear: $K(x_i,x_j) = x_i^T x_j$
Polynomial: $K(x_i,x_j) = (\gamma x_i^T x_j + r)^d, \gamma>0)$
RBF: $K(x_i,x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma> 0)$
Sigmoid: $K(x_i,x_j) = \tanh(\gamma x_i^T x_j + r)$

In this experiment, we have created classifiers for all SVM kernels to solve the same multiclass classification problem. The idea to use all kernels is to seek the best classifier that can provide the highest classification accuracy. Once this kernel is identified, it will be used to classify the CXR images based on the nodule location in lung zone.

The fourth task is to separate the list of coordinates into training and test data via cross-validation. By separating the coordinates into two groups, the accuracy prediction on the training data can be obtained and this reflects the performance on classifying the test data. In this study, the separation is done based on random sub-sampling where the JSRT dataset is divided into equal halves between the training and test data. Therefore, there are 77 nodule coordinates for both groups. After the separation, the fifth task is to acquire the best parameter to train the training data. For this, we executed three-fold cross-validation procedure to determine the best value for C (cost) and G (gamma) value. In our experiment, we found out that the best value would be C=8and G=4. The rest of other related parameters for LIBSVM was set to default. Readers are advised to read details specification of these parameters in [14] and [15]. In both articles, each parameter is explained in detailed and the default values for these parameters are also mentioned. Finally, the multiclass classification test is executed to the test data. In our experiment, the execution is done in Matlab. The final task is straightforward where the SVM classifiers classified the nodule coordinates in the test data group. Later the classification result can be obtained with the nodule coordinates are clustered into six lung zones based on their location in the image.

VI. Result and Discussion

We presented the classification output for each kernel type in the form of classification accuracy and confusion matrices. Briefly, these matrices enable us to analyze the test data group (for 77 CXR images) in order to indicate the pattern of class allocation. To add details for these matrices, at the end of each confusion matrices, we also calculate the average classification accuracy for each kernel (in Matlab, the LIBSVM classification function displays detailed classification accuracy for each class, but in this paper, we just count the accuracy average for each class). According to[14], the classification accuracy can be derived using the following equation:

$$Accuracy = \frac{\#\ correctly\ predicted\ data}{\#\ testing\ data}\ x\ 100\% \qquad (1)$$

Table 1, 2, 3 and 4 represent the confusion matric for linear, polynomial, RBF and sigmoid kernel respectively. At the bottom for each matric, we presented the average classification accuracy for each kernel.

Table 1
Confusion matrix for linear kernel

	LUZ	LMZ	LLZ	RUZ	RMZ	RLZ
LUZ	5	0	0	0	1	0
LMZ	4	0	4	0	4	0
LLZ	0	0	9	0	0	0
RUZ	0	0	0	9	1	0
RMZ	0	0	0	3	8	6
RLZ	0	0	0	0	0	23

Average classification accuracy = 91.8%

Table 2
Confusion matrix for polynomial kernel

	LUZ	LMZ	LLZ	RUZ	RMZ	RLZ
LUZ	9	3	0	0	0	0
LMZ	0	9	1	0	2	0
LLZ	0	1	6	0	0	0
RUZ	0	0	0	5	5	0
RMZ	0	0	0	0	17	0
RLZ	0	0	0	0	1	18

Average classification accuracy = 92.0%

Table 3
Confusion matrix for RBF kernel

	LUZ	LMZ	LLZ	RUZ	RMZ	RLZ
LUZ	8	1	0	0	0	0
LMZ	0	11	2	0	0	0
LLZ	0	0	9	0	0	0
RUZ	0	0	0	8	0	0
RMZ	0	0	0	2	14	3
RLZ	0	0	0	0	0	19

Average classification accuracy = 96.1%

Table 4
Confusion matric for sigmoid kernel

	LUZ	LMZ	LLZ	RUZ	RMZ	RLZ
LUZ	0	6	0	3	1	0
LMZ	0	3	0	0	7	0
LLZ	0	0	0	0	0	9
RUZ	2	0	0	1	9	0
RMZ	0	0	0	7	2	10
RLZ	0	0	0	0	0	17

Average classification accuracy = 82.7%

Overall, from the range of classification undertaken, the highest average classification accuracy was obtained from the RBF kernel at 96.1%. This value beats the other type of SVM kernels with the polynomial kernel managed to achieve at 92.0%, followed by linear kernel at 91.8% and sigmoid kernel at 82.7%. We expected that all SVM kernels were able to classify the nodule coordinate very accurately (>90%), but the result turns out that the sigmoid kernel is only moderately good to classify the coordinate. Furthermore, we are surprised that the linear kernel managed to achieve high accuracy because the test data group is in the form of non-linear.

Meanwhile based on the values in confusion matrices shown in Table 1 until Table 4, we can say that each kernel has its own strength and weakness in classifying the test data. For instance, the confusion matrices for the polynomial and RBF show that both kernel classifiers managed to classify the test data for all lung zones although it is not entirely perfect. However, the linear kernel classifiers missed to classify one zone i.e. LMZ (see Table 1) while the sigmoid is even bad where it missed to classify two zones i.e. LUZ and LLZ (see Table 4). Since the four kernel operated in different ways, they may be viewed as complimentary source of information rather than competing options. This may make them useful candidates for use in a consensual or ensemble-based approach to image classification [9]. But if we were given the chance to pick the most appropriate SVM kernel type to classify our nodule coordinate in lung zones then we definitely will choose the RBF kernel. The RBF is chosen because it is a nonlinear (Gaussian) kernel and able to map the nodule coordinate into higher dimensional space. Therefore, we hope that it can handle cases when relation between class labels and the coordinate is nonlinear.

VII. Conclusion

In this paper, we have shared our experience in applying multiclass classification with SVM kernels for CXR images based on nodule location in lung zones. The classification is successfully done using four SVM kernels that are linear, polynomial, RBF and sigmoid. In the classification experiment, the nodule coordinates were used as the classification input while the lung zones become the classification labels. The source of images and nodule coordinate were taken from JSRT image dataset. Additionally, during the experiment, we have divided the source data into two equal halves for the training and test data group.

Overall, the classification accuracy percentage shows high achievement for three SVM kernels namely RBF, polynomial and linear while the sigmoid performance is moderately good. We expected that all classifiers are able to score high (>90%) but the result turn out to be the other way. Probably, this condition happened because we did not change any parameters in each kernel as the experiment used only the default value available in the LIBSVM library. In the future, we hope to perform further test for each kernel classifiers and change for their related parameters accordingly to achieve high classification accuracy. We are also keen to test classifiers performance that are developed by other machine learning method such as kNN, Bayes Naïve and Decision Tree. It would be very interesting to find out which classifiers performed the best upon others based on the same dataset that we use in this experiment. Later, the selected classifiers can help us on classifying related image features so that it would be a good input for other image processing tasks.

Acknowledgment

This study was financially supported by the FRGS/1/2014/ICT05/UKM/02/2 grant from the Ministry of Education Malaysia.

References

[1] Anthony, G., Gregg, H. and Tshilidzi, M. Image Classification Using SVMs: One-against-One Vs One-against-All. Proceedings of the 28th Asian Conference on Remote Sensing. (2007).

[2] Lakshmi, S. V. and Prabakaran, T. E. Performance Analysis of Multiple Classifiers on KDD Cup Dataset using WEKA Tool 8. (2015).

[3] Vapnik, V. N,Statistical Learning Theory, Wiley, New York, (1998).

[4] Kumar, A. and Zhang, D. (2006). Personal recognition using hand shape and texture. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society, 15(8), 2454–61.

[5] Rahman, M. M., Bhattacharya, P. and Desai, B. C.(2007) A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. IEEE Transactions on Information Technology in Biomedicine, 11(1), 58–69.

[6] Wu, H., Zhang, H. and Li, C. Medical image classification with multiple kernel learning. Proceedings of the Second International Conference on Internet Multimedia Computing and Service, ICIMCS '10,ACM, (2010) 189–192; New York.

[7] Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2010)A Practical Guide to Support Vector Classification (Vol. 101).

[8] Wang, H. H., Mohamad, D., and Ismail, N. A. Semantic Gap in CBIR : Automatic Objects Spatial Relationships Semantic Extraction and Representation. International Journal Of Image Processing. 4(3)(2010) 192–204.

[9] Foody, G. M., and Mathur, A. (2004)A relative evaluation of multiclass image classification by support vector machines. IEEE Transactions on Geoscience and Remote Sensing,42(6) , 1335–1343.

[10] Zhang, D., Islam, M. M., and Lu, G. (2012) A review on automatic image annotation techniques. Pattern Recognition, 45(1), 346–362.

[11] Mueen, A., Zainuddin, R., and Baba, M. S. Automatic multilevel medical image annotation and retrieval. Journal of Digital Imaging.21(3) (2008) 290–5.

[12] Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.,Doi, K. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. American Journal of Roentgenology. 174(1) (2000) 71–74

[13] Chang, C.-C., & Lin, C.-J.(2011) LIBSVM: A Library for Support Vector Machines. ACM Trans. Intell. Syst. Technol., 2(3), 27:1–27:27.

[14] Chang, C.-C., & Lin, C.-J. (2015) LIBSVM -- A Library for Support Vector Machines.

[15] Tao, Y., Peng, Z., Krishnan, A., and Zhou, X. S. (2011) Robust learning-based parsing and annotation of medical radiographs. IEEE Transactions on Medical Imaging, 30(2), 338–50.

[16] Mohd Nizam Saad, Muda, Z., Sahari, N., and Hamid, H. A. Spatial Features Terms for Describing Lung Nodule Location in Chest X-Ray Images. In 13th International Conference on Intelligent Software Methodologies, Tools, and Techniques, (2014) Langkawi, Malaysia.