Journal – ICON (Integrating Concepts) ISSN 2456-6071 Vol. 4, Issue 3, May 2019

AUTOMATIC TUBERCULOSIS DETECTION BASED ON MULTIPLE INSTANCE LEARNING CLASSIFIER FROM THE X-RAY IMAGES

Ms.M.Anitha M.E.(CSE), Dr.G.Nallasivan Professor/CSE,
PSN College of Engineering & Technology, Melathediyoor, Tirunelveli, Tamil Nadu 627152.

ABSTRACT:

The major advantage of multiple-instance learning (MIL) applied to a computer-aided detection (CAD) system is that it allows optimizing the latter with case-level labels instead of accurate lesion outlines as traditionally required for a supervised approach. In previous work, a MIL-based CAD system can perform comparably to its supervised counterpart considering complex tasks such as chest radiograph scoring in tuberculosis (TB) detection. However, despite this remarkable achievement, the uncertainty inherent to MIL can lead to a less satisfactory outcome if analysis at lower levels (e.g., regions or pixels) is needed. This issue may seriously compromise the applicability of MIL to tasks related to quantification or grading, or detection of highly localized lesions. In this project, we propose to reduce uncertainty by embedding a MIL classifier within an active learning (AL) framework. To minimize the labeling effort, we develop a novel instance selection mechanism that exploits the MIL problem definition through one-class classification. We adapt this mechanism to provide meaningful regions instead of individual instances for expert labeling, which a more appropriate strategy is given the application domain. In addition, and contrary to usual AL methods, a single iteration is performed. To show the effectiveness of our approach, we compare the output of a MIL-based CAD system trained with and without the proposed AL framework. The task is to detect textural abnormalities related to TB. Both quantitative and qualitative evaluations at the pixel level are carried out.

I - INTRODUCTION

Tuberculosis (TB) is an infectious disease caused by the bacillus M. Tuberculosis. According to the World Health Organization (WHO) Global Tuberculosis Report 2014, TB remains one of the world's deadliest communicable diseases. In 2013, an estimated 9.0 million people developed TB, of these 1.1 million (13%) were HIV-positive. According to the report, 1.5 million died from the disease, 360 000 of whom were HIV-positive. The rates of TB mortality are slowly declining each year and it is estimated that 37 million lives were saved between 2000 and 2013 through effective diagnosis and treatment. However, TB is now present in all regions of the world, increasingly as drug resistant variants.

TB commonly affects the lung, but can also affect other organs such as bones and other soft tissues. It spreads through the air when people with active TB cough, sneeze, or otherwise expel infectious bacteria. According to the WHO, the largest impact of the increase in TB burden relates to inadequate control in developing nations, particularly those with high endemic rates of HIV infection where opportunistic coinfections in immunocompromised HIV/AIDS patients have exacerbated the problem [2, 3]. TB is most

Prevalent in sub-Saharan Africa and Southeast Asia, where, in addition to HIV, widespread poverty and malnutrition reduce resistance to the disease. The African Region accounts for about four out of every five HIV-positive TB cases and TB deaths among people who were HIV positive [1]. Several skin tests based on immune response are available for determining whether an individual has been exposed to TB. However, skin tests are not an indicator of active disease, and are also affected by vaccinations. A definitive test for diagnosing TB is the identification of the bacteria in a clinical sputum or pus sample, which is the current gold standard [1, 3].

Patients exposed to TB can be characterized clinically into two groups: (1) those with active TB disease; and, (2) those who have been exposed but do not have active disease. This latter group may harbor small colonies of active bacilli sheltered within their bodies, particularly in their lungs, and are thus at risk for activation of TB, if their physical health state is compromised. This happens, for example, when immune surveillance and control systems fail in populations with immunocompromised states such as HIV/AIDS. Major radiographic manifestations of active pulmonary TB [8] include the following, some of which are shown in Figure 1:

• Air space consolidation: lobar opacity, often reported as pneumonia or pneumonitis;

• Miliary pattern: Fine granular sandy or seed-like appearance throughout the entirety of both lungs,

Journal – ICON (Integrating Concepts) ISSN 2456-6071 Vol. 4, Issue 3, May 2019

reported as diffuse bilateral infiltrates, sometimes (correctly) referred to as micronodular [9];

• Cavity formation: a finding with a detectible radio-dense rim, which may be continuous or discontinuous, is differentiated from a mass as it has some central complete or relative radiolucency;

• Bronchiectasis or enlargement of airways can appear as tubular rings or cylinders of irregular diameter extending radially from the lung hila, with or without central radiolucency;

• Lymph node enlargement or adenopathy: smoothly marginated well-defined mass in close anatomic proximity to the mediastinum. These are silhouetted against the margins of normal structures such as the predictable pulmonary arteries and veins, which may facilitate their detection. Also a specific feature of an enlarged lymph node is frequently seen in patients with TB [10];

• Thickened Pleura: a finding that can be easily detected through careful survey of the periphery of the lungs just adjacent to ribs superiorly and laterally, and along the diaphragm inferiorly;

• Pleural effusions: loss of visualization (indistinctness) of the lateral costophrenic and medial cardiophrenic sulci on PA radiographs.

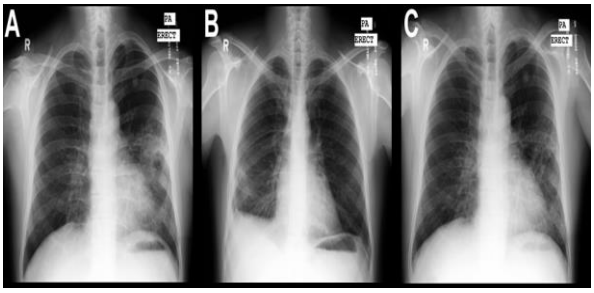


Figure 1. Example CXRs with manifestations of TB. CXR A and C in have infiltrates in both lungs. CXR B is a good example of pleural TB, which is indicated by the abnormal shape of the costophrenic angle of the right lung.

III - RELATED WORK

This section compares the different methods which had been already used for Object tracking and its merits and demerits are given.

Melendez J et al., (2015) investigated an alternative approach, namely multiple-instance learning (MIL), that does not require detailed information for optimization. We have applied MIL to a CAD system for tuberculosis detection. Only the case condition (normal or abnormal) was required during training. Based upon the well-known miSVM technique, we propose an improved algorithm that overcomes miSVM's drawbacks related to positive instance underestimation and costly iteration.

Breuninger M et al., (2014) had evaluated a method on chest radiographs of patients with symptoms suggestive of pulmonary tuberculosis enrolled in two cohort studies in Tanzania. All patients were characterized by sputum smear microscopy and culture including subsequent antigen or molecular confirmation of Mycobacterium tuberculosis (M.tb) to determine the reference standard. Chest radiographs were read by the software and two human readers, one expert reader and one clinical officer.

Monde Muyoyeta et al., (2014), proposed a technique to determine the sensitivity and specificity of a Computer Aided Diagnosis (CAD) program for scoring chest x-rays (CXRs) of presumptive tuberculosis (TB) patients compared to Xpert MTB/RIF (Xpert). Consecutive presumptive TB patients with a cough of any duration were offered digital CXR, and opt out HIV testing. CXRs were electronically scored as normal (CAD score ≤60) or abnormal (CAD score>60) using a CAD program. All patients regardless of CAD score were requested to submit a spot sputum sample for testing with Xpert and a spot and morning sample for testing with LED Fluorescence Microscopy-(FM).

Maduskar et al., (2013), proposed a retrospective analysis on 161 subjects enrolled in a TB specimen bank study. CXRs were analyzed using CAD4TB, which computed an image abnormality score (0-100). Four clinical officers scored the CXRs for abnormalities consistent with TB. We compared the automated readings and the readings by clinical officers against the bacteriological and radiological results used as reference. We report here the area under the receiver operating characteristic curve (AUC) and kappa (κ) statistics.

III - PROPOSED SYSTEM

The proposed work is introducing the multiple instance learning algorithm for segmenting the TB affected area i chest radiography images. The conventional classification or clustering technique are requires the huge amount of human (experts) intervention as well. The automatic machine learning algorithms are implemented in order to deliver the maximum efficiency with reduced human intervention. This work is proposing a multiple instance learning technique for segmenting The TB affected areas which reduces the Relabeling or correcting task for the human. The basic block diagram fo the work is given in the following figure 3.2

The multiple instance learning is nothing but the combined approach of Clustering and classification approaches. Both merits and demerits of these technique are combined together and forming an

Journal – ICON (Integrating Concepts) ISSN 2456-6071 Vol. 4, Issue 3, May 2019

efficient and automated technique for TB segmentation.

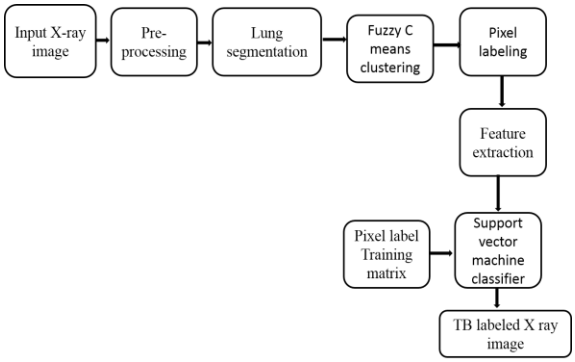


Figure 2 Block diagram for the proposed work

The proposed work is having the following main contribution over the TB segmentation with the help of MIL.

Image Preprocessing
Lung segmentation
Fuzzy C Means Clustering & labelling
Texture Feature extraction
Support vector machine classification

This section is devoted to explain the blocks of the proposed work in briefly and the multiple instance learning is implemented with two technique in this work. That is Fuzzy C means Clustering and Support vector machine classification. The concept of multiple instance learning is elaborated below.

(i) Multiple Instance Learning

In machine learning, multiple-instance learning (MIL) is a variation on supervised learning. Instead of receiving a set of instances which are individually labeled, the learner receives a set of labeled bags, each containing many instances. In the simple case of multiple-instance binary classification, a bag may be labeled negative if all the instances in it are negative. On the other hand, a bag is labeled positive if there is at least one instance in it which is positive. From a collection of labeled bags, the learner tries to either (i) induce a concept that will label individual instances correctly or (ii) learn how to label bags without inducing the concept.

Take image classification for example in Amores (2013). Given an image, we want to know its target class based on its visual content. For instance, the target class might be "beach", where the image contains both "sand" and "water". In MIL terms, the image is described as a bag X={X_{1},..,X_{N}}} X = {X_1,..,X_N}, where each {X_{i}} X_{i} is the feature vector (called instance) extracted from the corresponding i-th region in the image and N is the total regions (instances) partitioning the image. The bag is labeled positive ("beach") if it contains both "sand" region instances and "water" region instances.

(ii) Preprocessing

Chest X-ray radiographies are degraded during the process of imaging due to image transmission and image digitization by noise and existence of extra-cranial tissues in X-Ray images such as Nodules, Bones.

Pre-processing is a procedure to eliminate these noises and extra-cranial tissues from the medical images and alters the heterogeneous image into homogeneous image. Though there are lots of filters which have been used for filtering the images, some of them corrupt the miniature details of the image and some conventional filters will process the image incessantly (smoothing) and consequently harden the edges of the image. Hence, the proposed pre-processing steps namely De-noising and skull stripping provide better Image clarity. The proposed work is initiated with the gray scale conversion and where the dimension is reduced from three color band to single color band. After gray scale conversion, the gray image is processed for the contrast adjustment technique. It is done with the help of Adaptive histogram equalization technique.

(iii) Lung segmentation

The Proposed work is implemented with the Extended-maxima transform based region growing method for lung detection and some other binary morphological operations. An image can have multiple regional maxima or minima but only a single global maxima or minima. Determining image peaks or valleys can be used to create marker images that are used in morphological reconstruction. The following figure 3.3 illustrates the Maxima and minima in 1-D signal path.

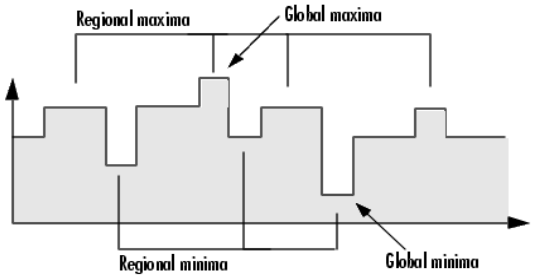


Figure 3 Illustration of Maxima and minima detection

The image processing toolbox in MATLAB can provides direct functions for these maximum and minima detection. The functions accept a grayscale image as input and return a binary image as output. In the output binary image, the regional minima or maxima are set to 1; all other pixels are set to 0.

Journal – ICON (Integrating Concepts) ISSN 2456-6071 Vol. 4, Issue 3, May 2019

Mathematical morphology is a powerful tool for image analysis, which was developed about forty years ago. Unlike other tools (e.g., Fourier methods), morphological operators relate directly to shape. When used appropriately, morphological operations can simplify images by preserving their essential shapes and eliminating noise.

(iv) Fuzzy C-Means clustering & labelling

The system utilizes the authentic size of the image to perform high quality image segmentation which causes high-resolution image data points to be clustered. Therefore utilize the FCM algorithm for clustering image data by considering that it has ability to cluster immensely colossal data and additionally outliers' payments are utilized expeditiously and efficiently. Because of starting points engendered arbitrarily, one of the local minima leads to erroneous clustering results hence FCM is arduous to reach global optimum. To evade this phenomenon, the proposed system utilizes adaptive pillar algorithm, which is very robust and superior for initial clusters optimization for FCM by deploying all centroids far discretely among them in the data distribution. This algorithm is inspired by the cerebration process of determining a set of pillars' locations to make a stable house or building. In the proposed adaptive pillar FCM we find out the average mean of the data point instead of grand mean in the previous algorithm. The average mean based initial centroid point selection can improve the performance of the clustering than the grand mean based existing method. Locate two, three, and four pillars, to withstand the pressure distributions of several different roof structures composed of discrete points. It is inspiring that by distributing the pillars as far as possible from each other within a roof, as number of centroids among the gravity weight of data distribution in the vector space the pillars can withstand the roof's pressure and stabilize a house or building. Therefore, this algorithm designates positions of initial centroids in the farthest accumulated distance between them in the data distribution.

(v) Texture Feature extraction

In machine learning, pattern recognition and in image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.

When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a features vector). This process is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

Feature extraction involves reducing the amount of resources required to describe a large set of data. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power, also it may cause a classification algorithm to overfit to training samples and generalize poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.

The best results are achieved when an expert constructs a set of application-dependent features, a process called feature engineering. Nevertheless, if no such expert knowledge is available, general dimensionality reduction techniques may help.

(VI) Support Vector Machine Classifier

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection. More formally, a support vector machine constructs a hyperplane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

SVM Algorithm

Classifying data is a common task in machine learning. Suppose some given data points each belong to one of two classes, and the goal is to decide which class a new data point will be in. In the case of support vector machines, a data point is viewed as a p-dimensional vector , and we want to know whether we can separate such points with a (p-1)-dimensional hyperplane. This is called a linear classifier. There are many hyperplanes that might classify the data. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes. So we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the maximum-margin hyperplane and the

Journal – ICON (Integrating Concepts) ISSN 2456-6071 Vol. 4, Issue 3, May 2019

linear classifier it defines is known as a maximum margin classifier; or equivalently, the perceptron of optimal stability.

IV – SIMULATION RESULTS

The proposed work is tested on the X-Ray images that are collected form the database. The different kind of chest X-Ray images are shown in the following figure 4.2

Figure 4 database chest X-ray images

Figure 5 test chest X-ray image – RGB image

Figure 6 adaptive histogram image

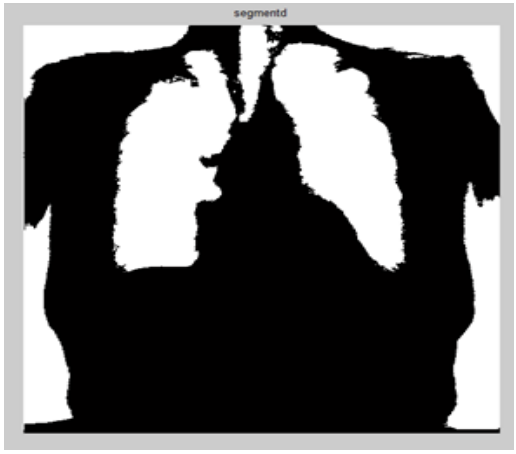


Figure 7 Binary X ray image

Figure 8 Binary Lung image

Journal – ICON (Integrating Concepts) ISSN 2456-6071 Vol. 4, Issue 3, May 2019

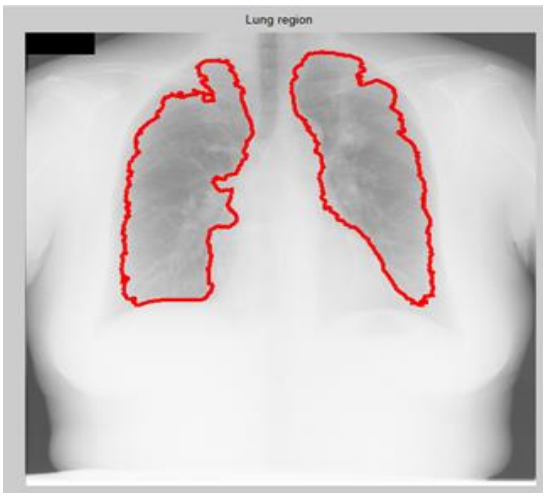


Figure 9 Segmenting the Lungs boundary

Figure 10 Lung ROI

Figure 11 FCM result after clustering

.

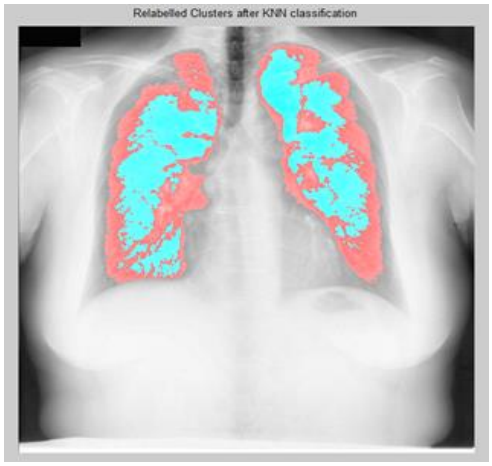


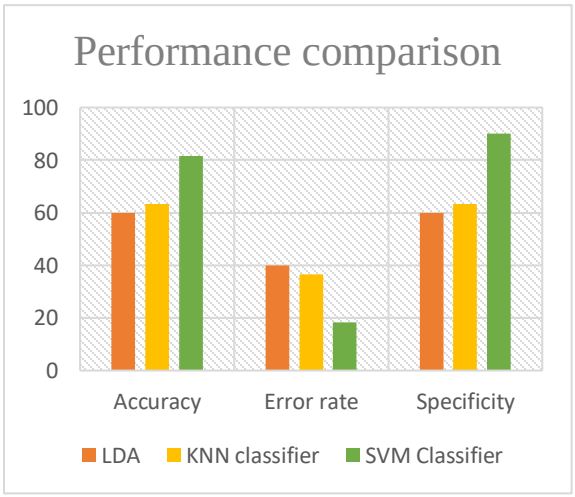
Figure 12 Pixel based Classification results

PERFORMANCE EVALUATION

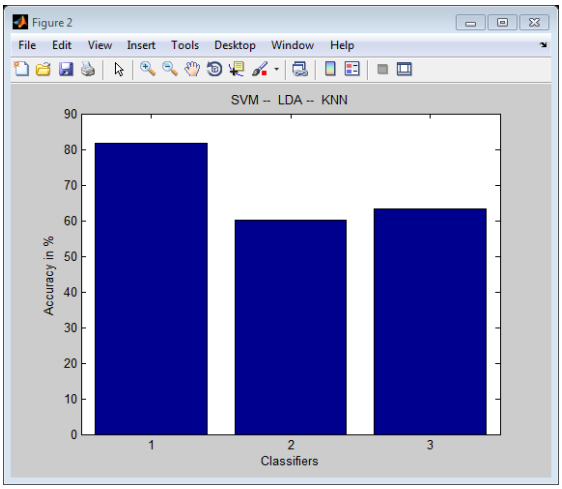
In this work, Totally 30 X-ray lung images are collected from the UC Irvine Machine Learning Repository database. The following images are the collected images for this proposed work.

Table 4.2 Parameter results

Image name	Accuracy	Error rate	Specificity
LDA	60.0	40.0	60.0
KNN classifier	63.33	36.67	63.33
SVM Classifier	81.667	18.33	90.0



Journal – ICON (Integrating Concepts) ISSN 2456-6071 Vol. 4, Issue 3, May 2019



In statistics, a receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm and can be calculated as (1 − specificity). The following graphical representation depicts about the higher TPR for the proposed SVM classification.

V- CONCLUSION

Traditional segmentation techniques are facing lot of problems especially it needs experts involvement as well. Fully automatic segmentation techniques are notable for the future implementation which is based on machine learning algorithms. More specifically multiple instance learning algorithms are addressing a new way of fully automatic segmentation. This project is introducing non-invasive Computer Aided Diagnosis technique to screen the TB affected areas in chest radiograph. This automatic computer diagnosis can able to identify the abnormality in lung based on active leaning & multiple instance learning. The Lung region form the X-ray images are segmented with the help of region growing method and Fuzzy C means clustering is applied for the initial grouping of lung pixels. The texture features are extracted from each clusters. Furthermore the active learning such as support vector machine algorithm is responsible for normal and abnormal pixel classification. The dataset are collected from the JSRT web database. The proposed work is implemented on MATLAB R2014a where the classification accuracy achieved maximum as 81.6667% when compared to existing classification.

REFERENCES

[1] Melendez, Jaime, et al. "On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis." IEEE transactions on medical imaging 35.4 (2016): 1013-1024.

[2] Van't Hoog, A. H., et al. "High sensitivity of chest radiograph reading by clinical officers in a tuberculosis prevalence survey." The International Journal of Tuberculosis and Lung Disease 15.10 (2011): 1308-1314.

[3] Muyoyeta, Monde, et al. "The sensitivity and specificity of using a computer aided diagnosis program for automatically scoring chest X-rays of presumptive TB patients compared with Xpert MTB/RIF in Lusaka Zambia." PloS one 9.4 (2014): e93757.

[4] Breuninger, Marianne, et al. "Diagnostic accuracy of computer-aided detection of pulmonary tuberculosis in chest radiographs: a validation study from sub-Saharan Africa." PloS one 9.9 (2014): e106381.

[5] Muyoyeta, Monde, et al. "The sensitivity and specificity of using a computer aided diagnosis program for automatically scoring chest X-rays of presumptive TB patients compared with Xpert MTB/RIF in Lusaka Zambia." PloS one 9.4 (2014): e93757.

[6] Melendez, Jaime, et al. "A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays." IEEE transactions on medical imaging 34.1 (2015): 179-192.

[7] Xu, Yan, et al. "Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering." Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.

[8] Wang, Shijun, et al. "Seeing is believing: Video classification for computed tomographic colonography using multiple-instance learning." IEEE transactions on medical imaging 31.5 (2012): 1141-1153.

[9] Zhang, Dan, et al. "Interactive localized content based image retrieval with

Journal – ICON (Integrating Concepts) ISSN 2456-6071 Vol. 4, Issue 3, May 2019

multiple-instance active learning." Pattern Recognition 43.2 (2010): 478-484.

[10] Lesniak, Jan, et al. "Computer aided detection of breast masses in mammography using support vector machine classification." SPIE Medical Imaging. International Society for Optics and Photonics, 2011.

