VOL. 10, NO 20, NOVEMBER, 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

# LUNG CANCER DETECTION USING SUPERVISED CLASSIFICATION WITH CLUSTER VARIABILITY ON RADIOGRAPHS DATA

Noreen Kausar[1], Brahim Belhaouari Samir[2] and Ramil Kuleev[2]
[1]College of Science, Alfaisal University, Riyadh, Saudi Arabia
[2]Department of Computer Science, Innopolis University, Kazan, Russia
E-Mail: noreenkausar88@yahoo.com

**ABSTRACT**

Performance enhancement for disease diagnostic systems has been utmost challenging aspect of providing further treatments or proceeds surgeries without any possible delays. In recent times, various data mining techniques are being applied as the ratio of lung cancer is increasing enormously in recent years and require significant developments in its accurate detection at a possible early stage to cure the patients from further suffering. Developing a diagnostic system for lung cancer demands efficiency in processing and classification of X-rays of normal and cancerous cases. In this work, robust computer aided diagnostic system is proposed by utilizing modified clustering based classifiers such as Support Vector Machine (SVM) and k- Nearest Neighbors (k-NN) with optimized processing techniques for feature processing and selection of suitable features to enhance system's performance in terms of accuracy, sensitivity and specificity. Overall, this work has proved to have a maximum detection rate with respect to earlier techniques applied. In future this approach will be implemented for determining the region of interest (ROI) and classifying the severity of cancer cases as mild, moderate or critical.

**Keywords:** lung cancer, statistical energy based selection, support vector machine (SVM), feature extraction, radiograph processing, principal components, laplacian algorithm.

## 1. INTRODUCTION

Cancer is considered foremost cause of mortality globally with 14 million new cases and 8.2 million deaths among other infectious and cardiac disease. In the United States of America (USA) around $125 billion have been utilized for cancer cures and can even rise to $156 billion by 2020. Lung cancer is the most common cancer type among men, as diagnosed in 2012 with mainly cases from Asia and Africa [1]. Figure-1 provides the facts regarding lung cancer incidence and mortality per 100,000 cases.

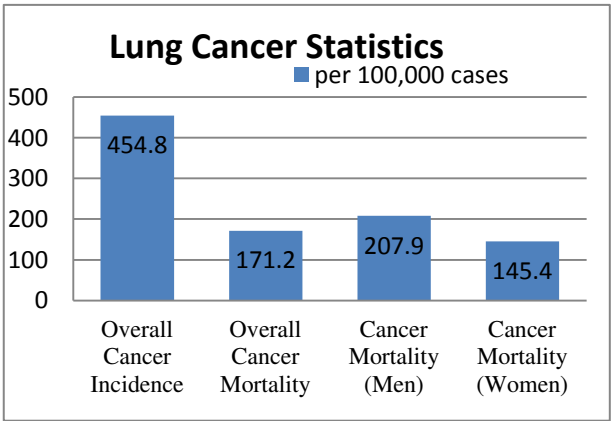



**Figure-1.** Incidence and mortality statistics for lung cancer.

As per the WHO statistics, in Russia lung cancer is leading the death rate in both genders as shown in Figure-2 with respect to other cancer diseases such as stomach cancer, oral or liver cancer etc. [2] which can spread much readily because of imbalanced living measures or inappropriate medication.

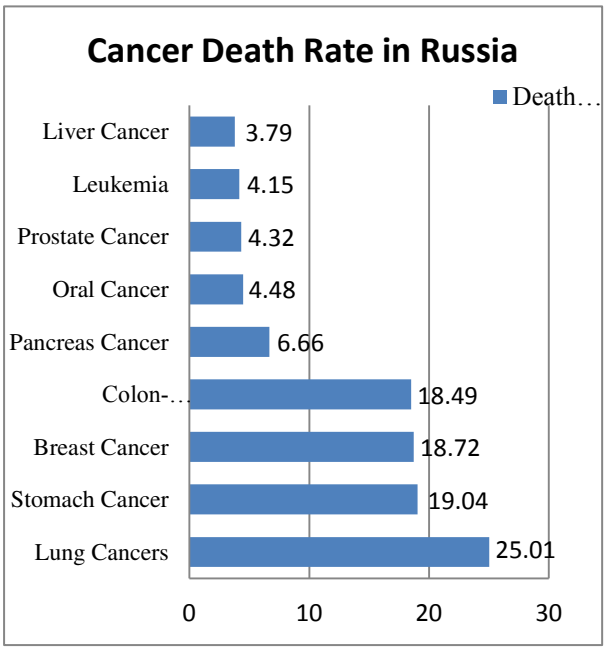



**Figure-2.** WHO cancer death rate statistics for Russia.

These details presented in Figure 1 and 2 are from the most recent data of these primary sources: WHO, World

Bank, United Nations Educational, Scientific and Cultural Organization (UNESCO), Central Intelligence Agency (CIA) and individual country databases for global health and causes of death [2].

To overcome causalities at an early stage of lung cancer from further damages in term of untreatable tumor, specified diagnostic system should be designed which are capable enough to detect the cancerous symptoms and assist medical specialists to start appropriate medication or required surgeries. In recent years, with the advancements in the medical engineering, many computational algorithms have been implemented to classify different radiographs data as normal or cancerous. To develop such system for lung cancer, chest radiographs such as X-rays can be easily incorporated as input because of their low cost and less examination time but demands much efficiency of the system to detect nodules after suitable filtering and processing in order to get desired accurate results.

In this paper, an embedded clustered based supervised classification is proposed which is applied on the optimized features extracted and processed from X-rays dataset from Japanese Society of Radiological Technology (JSRT). The paper is organized as follows:

Existing techniques and applied algorithms are discussed in Section 2, overview of JSRT is provided in Section 3, system design and implementation is mentioned in Section 4, results are explained in Section 5 followed by conclusion and future work in Section 6.

## 2. EXISTING TECHNIQUES AND APPLIED ALGORITHMS

In recent years, various computer-aided diagnosis (CAD) systems are designed to enhance the performance factor for the detection and segmentation of chest radiographs. Ahmad *et al*. performed experiments on JSRT by taking normal and abnormal images of equal number and determined two region of interest from each image and applied texture function along with a fisher coefficient to identify best features. The accuracies they achieved are: 93.59% using linear discriminant analysis, 98.7 using nonlinear discriminant analysis and 85% with principal component analysis. Overall, 98% accuracy was scored by artificial neural network. They also used classification error probability and accuracy correlation coefficient, but managed to get 92% accuracy [3].

Orban *et al*. worked on processing the radiograph images to remove the ribs and collar bone from the lung images to enhance the detection rate by applying constrained sliding band filter for intensity measure with Support Vector Machine (SVM) on suitable textual features and scored 61% sensitivity by minimizing possible false alarms [4] whereas an accuracy of 83% was achieved with neural network by integrating with entropy and energy based feature processing techniques [5].

Udeshani *et al*. proposed denoising and segmentation method for X-rays data by using various techniques to train neural network classifier and achieved 88% with feature processing techniques where as 96% with pixel based method [6].

In 2011, Samulski *et al.* observed the significant variation in the diagnosis of lung cancer by comparing conventional CAD with modified interactive CAD which out formed with low false rate. Location response operating characteristic analysis (LROC) was used for analysis [7]. For lung cancer identification, CT scan based imaging resource have also been equipped enough to detect the nodules with better sensitivity and less false positive rate in clinical examination [8].

Xu *et al*. [9] worked on automatic lung field segmentation by enhancing the accuracy rate of active shape model and minimizing its processing time to perform lung segmentation. The accuracy of left lung from JSRT dataset was 92.52, sensitivity 89.7% and specificity 97.2% in 0.38 seconds, whereas the accuracy, sensitivity and specificity for right lung were 95.5%, 91.2% and 97.6% in 0.35 seconds.

Apart from diagnosing tumor from lung nodule, some researchers have worked for lung lesions from chest radiographs by integrating textual and photometric based techniques for attribute processing along with effective classification algorithms [10, 11].

Enhancing the image quality by modifying specific attributes using spatial domain algorithms which eliminates irrelevant details of the image along with noise and increases brightness by appropriate resizing and filtering to have better image quality for clinical practice [12].

Jaeger *et al*. [13] worked on chest radiographs for lung segmentation and classification from local hospital based datasets among which, one from US and another from China to diagnose Tuberculosis with an accuracy of 78.3% on first and 84% on second dataset which is better than the radiologist results on the same dataset.

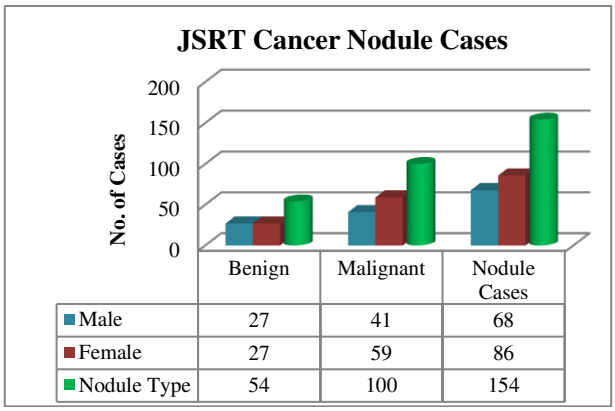
In 2014, an optimized approach for identifying the edges of lung images extracted from X-rays based on content based retrieval and masking images with similarity measures and energy function [14]. The maximum accuracy acquired was 95.4% on JSRT dataset.

In another work, 95.8% accuracy was achieved in comparison to other datasets for lung segmentation by utilizing an unsupervised approach based on fuzzy C-means clustering and thresholding with Gaussian derivatives filtering for locating lung regions [15].

Apart from classification of lung nodules, recently various methods are also being used for enhancing sharpness, equalizing histogram without harming the radiographs sensitivity and achieve better results [16].

## 3. JSRT RADIOGRAPH DATASET FOR USING LUNG CANCER

For designing computer-aided diagnosis systems for lung cancer, Japanese Society of Radiological Technology (JSRT) X-rays dataset is used for classification and comparison with earlier approaches. Overall, it includes 247 chest radiographs among which 100 are malignant and 54 benign, whereas 93 are normal X-ray images [17] gathered from fourteen medical centers. Figure-3 provides the distribution of abnormal (lung cancer) cases among male and female.




| | Benign | Malignant | Nodule Cases |
|---|---|---|---|
| Male | 27 | 41 | 68 |
| Female | 27 | 59 | 86 |
| Nodule Type | 54 | 100 | 154 |

**Figure-3.** Gender wise distribution of JSRT cancer cases.

The matrix size is 2048x2048 with 0.175mm pixel in high resolution and wide density range of 12 bit and 4096 gray scales.

## 4. SYSTEM DESIGN AND PROPOSED METHODOLOGY

In this paper, a comparative classification approach has been integrated with modified image based features filtering and selection to increase system detection rate by minimizing false alarms in terms of false positive and false negative rates. The system design of this proposed ensemble approach is presented in Figure-4.

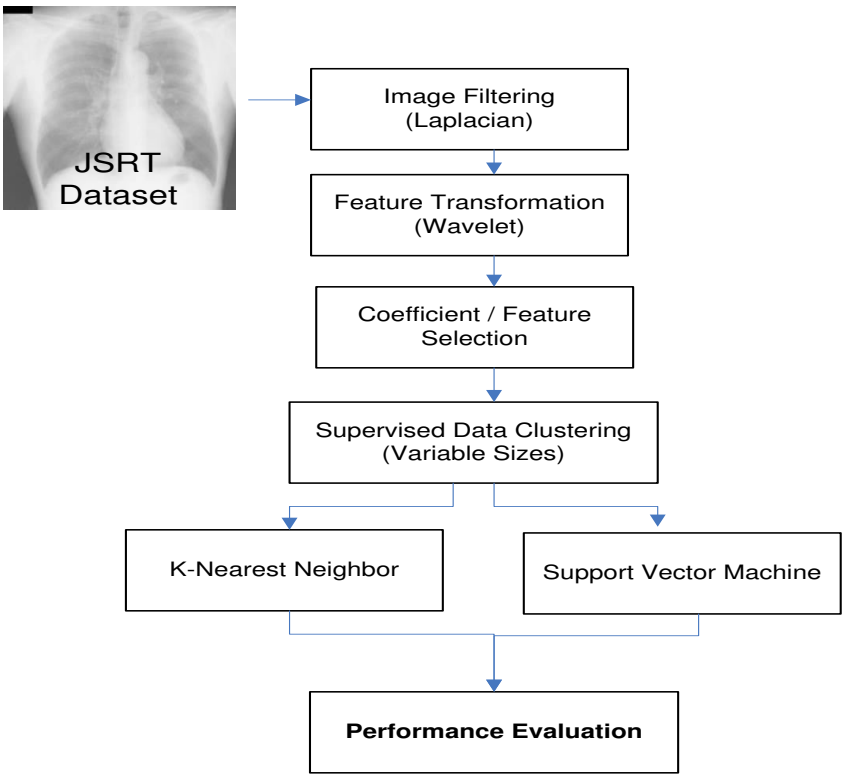



**Figure-4.** Hybrid design of lung cancer diagnostic system.

Figure-4 shows various phases involved in designing an optimized CAD system for lung cancer. The dataset used is based on radiograph images taken from JSRT. After that the images are filtered using Laplacian filter to enhance the contrast and sharpen for better visibility of the lung nodules with less noise and disturbance. Once the images are processed, extraction and selection of sensitive features are done for supervised clustered based classification. The phases are explained as below:

**Image filtering and equalization**

The dataset images are in raw form which need sufficient pre-processing phase. Firstly histogram equalization is used which balances the intensity level and adjusts the contrast of the images to achieve a flat histogram [18].

After a repeated process of intensity equalizing, the images are filtered using the Laplacian method to further omit the odd regions of the images.

**Feature transformation and selection**

For signal decomposition, wavelet transformation is used which is considered better than curvelet transformation based on scaling image dimension. Such transformation helps in storing significant details which are useful for classifying the images by utilizing the signal's time-frequency content [19].

The next phase is to reduce the dimensionality of the newly generated coefficients from the wavelet transformation. Here two algorithms Supervised Principal Component Analysis (SPCA) and statistical energy based metric selection are applied. Both are modified enough to eliminate the large number of features produced from the image and figure out few high variant attributes which can be classified to determine the normal as well the cancerous images [20]. The steps involved in the computation of principal components are provided in Figure-5.

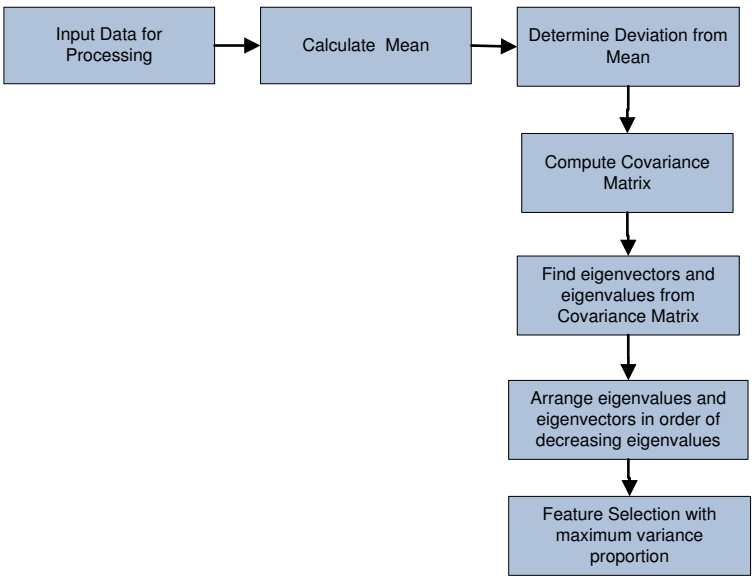



**Figure-5.** Steps involved in principal component analysis.

As shown in the Figure-5, first the dimension is reduced by using covariance matrix and arranging associated eigenvectors and eigenvalues. Subsequently, energy of all the features is calculated by adjusting specific threshold to eliminate the ones which are below the mentioned scale.

The benefits of using principal components is that they transforms the features in to reduced dimension, but keeps all the important details required by the classifier to identify the possible classes [21]. In this work, PCA is modified as supervised it means the class information is also included while producing the principal components (PCs) which eventually become more helpful with the following criteria.

- PCs are linear combinations of the original attributes.
- PCs are orthogonal to each other.
- PCs capture the maximum amount of variation in the data.

Figure-6 shows the newly transformed matrix from SPCA, which have reduced the attributes from 20018 to 247 by eliminating the remaining which did not have enough proportion of the total variance.

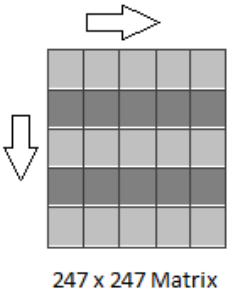



**Figure-6.** SPCA transformed matrix for JSRT dataset.

**Supervised clustering**

Once the features are selected, prior to classification, supervised clustering is applied to the processed features to highlight the possible dissimilar groups available in the dataset. In this approach, varying sizes are used, ranging from two which is default for carrying binary classification up to 25-30 depending upon the dataset. Clustered based classification helps in increasing the detection rate of the patterns because it is applied on the clusters separately, which have maximum similarity index and enhances the precision of the system.

Clustering is done by incorporating distance measures to assess the patterns and arrange them in different clusters based on their coordinates [22]. Each cluster has its own centroid (center) which is used as an index measure during the classification in a 5x2 cross validation [23]. The distance measures used in this work is squared Euclidean. To optimize every cluster size, the respective sum of distance and number of iterations is incorporated to analyze the best possible clustering. Figure-7 shows the segmentation of principal components from JSRT dataset in variable cluster sizes from 2 to 5. Different colors are used to identify possible number of classes. The black cross encircled denotes the centroid of each cluster which can be seen in the exact center of the clusters.

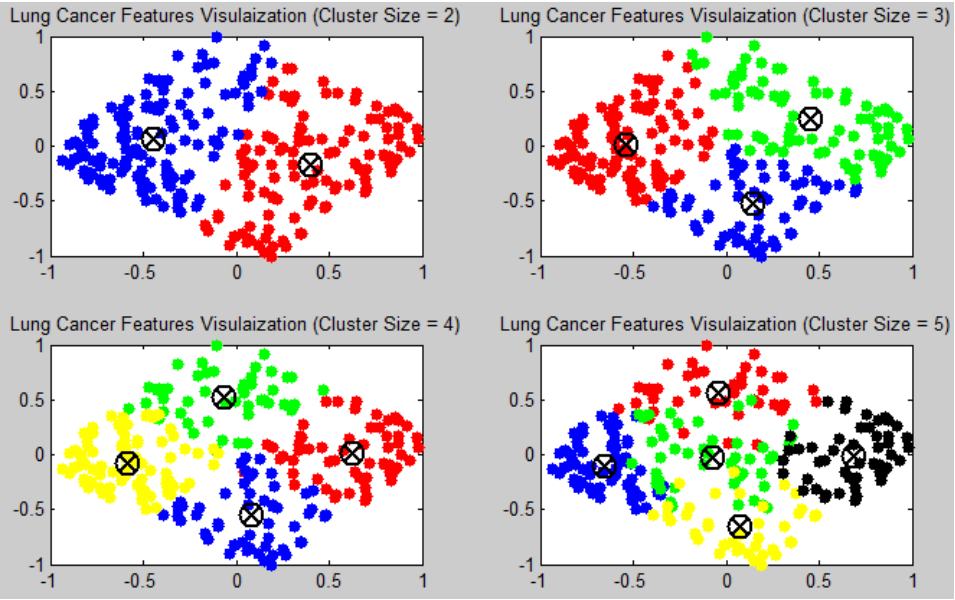



**Figure-7.** Feature visualization for variable cluster sizes of JSRT dataset.

**Ensemble classification with optimized parametric tuning**

In this ensemble approach binary classification is used in combination with clustering, which enhances the input data representation in the form of groups of patterns having similar behavior irrespective of their class labels for Support Vector Machine (SVM) classification where as k-NN takes clusters which have patterns of same class labels. The classifier architecture is based on the cross validation method which means there will be an iterative procedure of training and testing to ensure that the system is quantified with misclassifications which eventually decreases the false alarms.

For SVM, 5x2 cross validation is used, whereas 10 fold is used for k-NN. In SVM training and testing samples are selected by randomizing the whole dataset each time during the process. There are 5 iterations with 2 fold, which means in each iteration the selected train and test data are used once in a fold and then interchanged with each other to complete the other fold to precede the second iteration and so on. In k-NN, sequential fold are used irrespective of the iterations.

In k-NN no such parametric optimization is done, but for SVM, non-linear mapping is selected which means the use of kernel function to transform the input data into higher dimension where they can be linearly represented and can be classified with maximum accuracy rate[24]. To

achieve best possible classification where both classes are separated as much possible, various hyper planes are drawn and the one which has a maximum margin from both sides is selected as shown in the Figure-8.

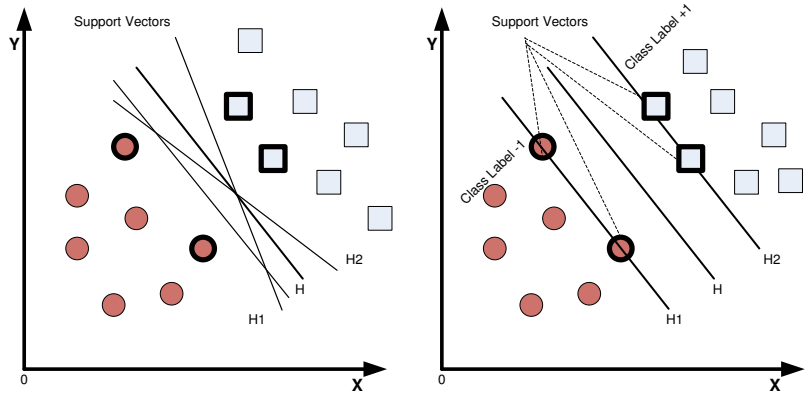



**Figure-8.** (a) SVM classification with possible number of hyper planes.
(b) Linear SVM classification with optimal hyper plane.

Red circles and blue squares are used to represent abnormal (cancerous) and normal classes respectively. The patterns which lie on the margin line of either class are called support vectors. The kernel used for this work is default radial basis function (RBF)[25]. The equation for RBF kernel and maximum margin classification are as below:

$$K(x, x') = \exp(-\gamma \left\| x^T - x \right\|^2)$$

$$margin = \underset{X \epsilon D}{\text{argmind}}(\text{X}) = \underset{X \epsilon D}{\text{argmin}} \frac{|X.w + b|}{\sqrt{\sum_{i=1}^{d} W_i^2}}$$

The class labels (-1 for abnormal and 1 for normal) can be represented as follows:

$$for Y_i = +1; \text{wx}_\text{i} + \text{b} \geq 1$$
$$for Y_i = -1; \text{wx}_\text{i} + \text{b} \leq 1$$
$$for alli;\ y_i\ (w_i + b) \geq 1$$

## 5. EXPERIMENTAL RESULTS AND COMPARATIVE ANALYSIS

In this section, the classification results of designed algorithms are presented. Clustered SVM and K-nn are implemented and applied under different circumstances. SVM is used to classify features reduced and selection by SPCA, whereas K-nn is used for features extracted and nominated by iterative statistical energy and metric procedure. The configuration settings for both classifiers are given in Table 1.

**Table-1.** Configuration setting for proposed classifiers.

| Classification measures | SVM | k-NN |
|---|---|---|
| Number of features used | 247 | 291 |
| Classification method | 5x2 cross validation | 10 fold cross validation |
| Classification type | Labeled Binary Classification (-1 abnormal; 1 normal) | Unlabeled Binary Classification |
| Ensemble approach used | SPCA, K-means, SVM | Statistical Metric, K-means, K-nn |
| Software used | Matlab, LibSVM | Matlab |
| System specification | Model: Aspire M3800<br>Processor: Intel(R) Core(TM)2 Quad CPU Q8300 @2.50GHz 2.50 GHz<br>RAM: 4.00 GB<br>System Type: 32-bit Operating System | |

The comparison of the approaches is based on the performance factor like accuracy which is the measure of detecting normal and abnormal (CAD) patterns correctly. [26]. The equation for calculating accuracy is given below.

$$\text{Accuracy} = \frac{TP + TN}{TN + TP + FN + FP} \%$$

where

True Positive (TP) means an abnormal (CAD) is detected as an abnormal (CAD).

True Negative (TN) means a normal is detected as normal.

False Positive (FP) means an abnormal (CAD) is detected as normal.

False Negative (FN) means a normal is detected as abnormal (CAD).

**Performance of Clustered-SVM**

The SVM based classification is performed on the principal components extracted by SPCA, which are further arranged in clustered form of different sizes with default distance measure. For each cluster, it requires a number of iterations to get minimum sum of distances which means that the patterns within the cluster are similar enough and on least possible distance. In this work, cluster sizes are optimized from default size 2 up to size 32 in order to configure related parameters in such manner to get maximum detection rate. The result of Clustered-SVM is provided in Table-2. The time required for each clustered based processing and related cluster parameters are also presented which helps in analyzing the variant clustering modifications on the classifier architecture.

**Table-2.** Classification measures for clustered SVM on JSRT dataset.

| **Number of clusters** (k) | **Sum of distances** | **Number of iterations** | **Accuracy** (%) | **Processing time** (%) |
|---|---|---|---|---|
| 2 | 91.75 | 4 | 87.8 | 18.03 |
| 3 | 70.22 | 6 | 87.3 | 14.96 |
| 4 | 55.84 | 8 | 88.9 | 13.34 |
| 5 | 46.38 | 7 | 89.2 | 13.99 |
| 7 | 38.68 | 14 | 91.50 | 14.21 |
| 10 | 29.47 | 13 | 92.27 | 14.50 |
| 12 | 29.27 | 9 | 92.89 | 14.07 |
| 14 | 23.91 | 15 | 93.25 | 19.46 |
| **15** | **23.78** | **11** | **94.66** | **16.85** |
| 16 | 21.16 | 10 | 93.59 | 22.81 |
| 17 | 20.38 | 16 | 93.89 | 23.50 |
| 18 | 19.50 | 11 | 94.63 | 18.52 |
| 19 | 18.66 | 13 | 94.66 | 20.56 |
| 20 | 17.95 | 8 | 94.32 | 25.51 |
| 23 | 15.29 | 24 | 92.26 | 27.88 |
| 24 | 15.82 | 10 | 92.28 | 21.33 |
| 25 | 14.43 | 10 | 93.26 | 22.04 |
| 30 | 12.88 | 12 | 94.75 | 31.05 |
| 32 | 12.18 | 8 | 93.47 | 37.11 |

**Performance of Cluster k-NN**

In this phase, k-NN is used as a binary classifier for the clustered data prepared from coefficients selected by statistical energy and metrics and the cluster size is set to 2. As its supervised approach so the training and the testing samples are set as 60% and 40% respectively of the total dataset. The approach used for k-NN is 10 fold, which means the process of classification will be span on 10 iterations and the resultant accuracy is determined after taking the mean of accuracies achieved by all the iterations. Table-3 presents the average accuracy of k-NN classifier.

**Table-3.** Classification measures for Cluster k-NN on JSRT dataset.

| **10 Fold cross validation** | **Accuracy** (%) |
|---|---|
| Iteration I | Between 92-98% for K-fold validation<br>Average Accuracy: 97.2% |
| Iteration II | |
| Iteration III | |
| Iteration IV | |
| Iteration V | |
| Iteration VI | |
| Iteration VII | |
| Iteration VIII | |
| Iteration XI | |
| Iteration X | |

Overall, k-NN achieved maximum accuracy in comparison with SVM but lacks in some perspectives when it comes to time cost, processing efficiency and variant clustering integrations. Next section highlights all the possible limitations of k-NN and significance of SVM based diagnostic system for lung cancer.

**Comparative discussion**

Support Vector Machine and k-Nearest Neighbor, both are supervised, binary classifier and parametric based algorithms which can be integrated and modified as per the dataset and set dimension. In this work, K-means clustering is ensemble with both classifiers to enhance their performance on the dataset by identifying the normal and cancer patterns from transformed, dimensionally reduced and clustered data attributes. Figure-9 provides the comparison of accuracy rates and processing time for different cluster sizes applied with SVM classifier. The best accuracy was reached when the cluster size was 15 and it took 16.85 seconds to complete the whole process of clustering the data, randomizing the data and dividing into equal train and test patterns and perform 5x2 cross validations and finding the mean of their performance measures.

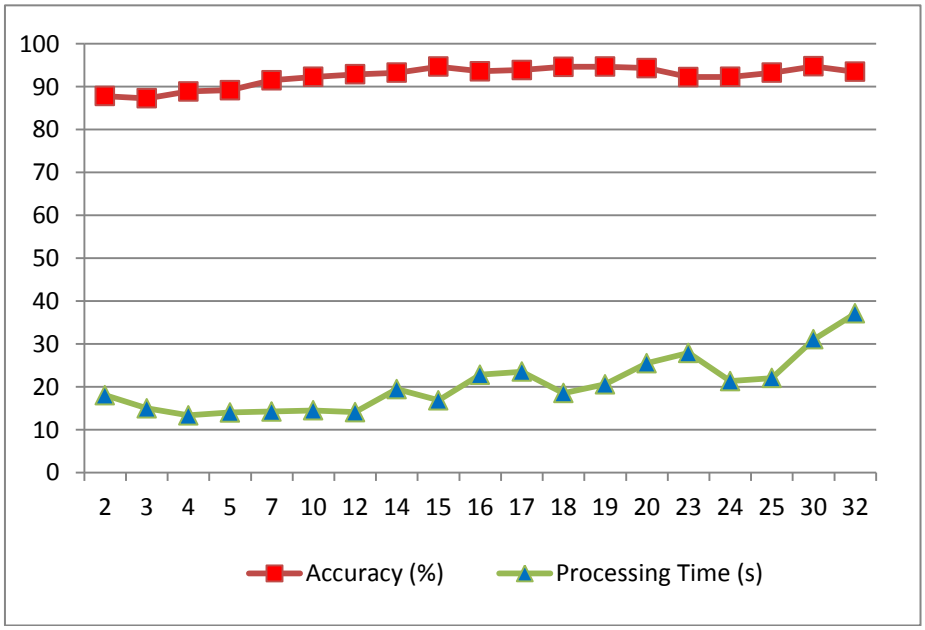



**Figure-9.** Accuracy rate and processing time comparison of different cluster sizes in SVM classification.

On the other hand, k-NN enhanced the accuracy little bit than SVM but suffers in terms of optimization measures associated with the classification which is as below:

- SVM achieved consistent performance with variable clustering sizes than k-NN which worked on clustering size based on the number of classes which is 2 (default) in case of binary classification.
- SVM costs less classification architecture and processing time and met desired accuracy in seconds where as k-NN took more than 10 minutes which is not preferable in terms of real-time environment.
- SVM used same proportion of training and testing with respect to k-NN which demands more patterns to be involved in training.
- SVM has easily achieved the accuracy on randomized data patterns, whereas k-NN takes normal and abnormal in procedural manner and maintains clustering size based on classes which means one cluster for each class only where as SVM is independent in choosing similar patterns from both classes and does not give favor to any respective class or pattern.

## CONCLUSIONS AND FUTURE WORK

Diagnosis of lung cancer really needs to be efficient enough and it depends upon the performance of designed computer-aided system. This work focused on the ensemble approach of clustering and supervised classifier for detection of cancer alignments where as the SPCA and statistical metrics performed an optimized feature selection and transformation into few variant attributes. To enhance the acquired accuracy, parameter tuning is also involved to get the best from the combination of applied techniques and algorithms. SVM performed the classification with less features and minimum time cost to achieve 94% accuracy, whereas k-NN utilized more features, considerable time cost and reach approx 97.2% accuracy rate. Overall, the ensemble system based on SVM classifier managed to overcome earlier drawbacks in terms of time and processing consumption. Related parameters such as cluster size also affect SVM processing and performance and they need further optimization to be more effective for determining possible cancer cases. In future, other categories of radiographs can be explored to enhance performance of clinical based disease diagnosis.

## ACKNOWLEDGEMENTS

The work has been supported by the Russian Ministry of education and science (agreement: 14.606.21.0002, ID: RFMEFI60614X0002).

## REFERENCES

[1] National Cancer Institute. Retrieved 14 January, 2015, from http://www.cancer.gov/about-cancer/what-is-cancer/statistics.

[2] World Health Organization. Retrieved 15 January, 2015, from http://www.worldlifeexpectancy.com.

[3] Ahmad M. S., Naweed M. S. and Nisa M. 2009. Application of Texture Analysis in the Assessment of Chest Radiographs. International Journal of Video and Image Processing and Network Security (IJVIPNS), 9(9).

[4] Orbán G., Horváth Á. and Horváth G. 2010. Lung Nodule Detection on Rib Eliminated Radiographs. In P. Bamidis and N. Pallikarakis (Eds.), XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010 (Vol. 29, pp. 363-366): Springer Berlin Heidelberg.

[5] S.A. Patil D. and Kuchanur M. B. 2012. Lung Cancer Classification Using Image Processing International Journal of Engineering and Innovative Technology (IJEIT) 2(3).

[6] Udeshani K. A. G., Meegama R. G. N. and Fernando T. G. I. 2011. Statistical Feature-based Neural Network Approach for the Detection of Lung Cancer in Chest X-Ray Images. International Journal of Image Processing (IJIP), 5(4).

[7] Samulski M. R. M., Snoeren P. R., Platel B., van Ginneken B., Hogeweg L., Schaefer-Prokop C., *et al.* (2011). Computer-aided detection as a decision assistant in chest radiography.

[8] Armato S. G., 3rd, McLennan G., Bidaut L., McNitt-Gray M. F., Meyer C. R., Reeves A. P., *et al*. 2011. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys. 38(2): 915-931.

[9] Xu T., Mandal M., Long R., Cheng I. and Basu A. 2012. An edge-region force guided active shape approach for automatic lung field detection in chest radiographs. Comput Med Imaging Graph. 36(6): 452-463.

[10] Xu T., Cheng I., Long R. and Mandal M. 2013. Computer-Aided Detection of Acinar Shadows in chest Radiographs. ICTACT Journal on Image and video Processing. 3(4).

[11] Ramaraju D. P. V. and Praveen S. 2014. Classification of lung tumour Using Geometrical and Texture Features of Chest X-ray Images. International Journal for Research in Applied Science and Engineering Technology (IJRASET), 3.

[12] Tarambale M. R. and Lingayat N. S. 2013. Spatial Domain Enhancement Techniques for Detection of Lung Tumor from Chest X-Ray Image. International Journal of Application or Innovation in Engineering and Management (IJAIEM) 2(8).

[13] Jaeger S., Karargyris A., Candemir S., Folio L., Siegelman J., Callaghan F., *et al*. 2014. Automatic Tuberculosis Screening Using Chest Radiographs. Medical Imaging, IEEE Transactions on, 33(2), 233-245.

[14] Candemir S., Jaeger S., Palaniappan K., Musco J. P., Singh R. K., Zhiyun X., *et al*. 2014. Lung Segmentation in Chest Radiographs Using Anatomical Atlases With Nonrigid Registration. Medical Imaging, IEEE Transactions on. 33(2): 577-590.

[15] Wan Ahmad W. S., WM W. Z. and Ahmad Fauzi M. F. 2015. Lung segmentation on standard and mobile chest radiographs using oriented Gaussian derivatives filter. Biomed Eng Online. 14(1): 015-0014.

[16] Alavijeh F. S. and Mahdavi-Nasab H. 2015. Multi-scale Morphological Image Enhancement of Chest Radiographs by a Hybrid Scheme. J. Med Signals Sens. 5(1): 59-68.

[17] Shiraishi J., Katsuragawa S., Ikezoe J., Matsumoto T., Kobayashi T., Komatsu K., *et al*. 2000. 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. AJR Am J Roentgenol. 174(1): 71-74.

[18] Al-Absi H. R. H., Samir, B. B. and Sulaiman S. 2014. A Computer Aided Diagnosis System for Lung Cancer based on Statistical and Machine Learning Techniques. Journal of Computers. 9(2): 425-231.

[19] Mallat S. G. 1989. A theory for multiresolution signal decomposition: the wavelet representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 11(7): 674-693.

[20] Kausar N. 2015. Hybrid Approach of Clustered-SVM for Rational Clinical Features in Early Diagnosis of Heart Disease. Malaysia University of Science and Technology, Selangor.

[21] Kausar N. 2013. On The Provisioning of Extending the Efficiency of Intrusion Detection Through Optimal SVM's Kernel. Universiti Teknologi Petronas, Tronoh.

[22] Wang J. 2010. Consistent Selection of the Number of Clusters via Crossvalidation. Biometrika 97(4): 893-904.

[23] Kausar N., Abdullah A., Samir B. B., Palaniapan S. and Alghamdi B. S. 2015. Ensemble Clustering Algorithm with Supervised Classification of Clinical Data for Early Diagnosis of Coronary Artery Disease, Journal of Medical Imaging and Health Informatics.

[24] Belhaouari S. B. and Abaza R. 2011. Gas Identi cation by Using a Cluster-k-Nearest-Neighbor (Vol. 3). Perth, Australia: IPCSIT.

[25] Chang Y. and Lin C. 2008. Feature Ranking Using Linear SVM. JMLR: Workshop and Conference Proceedings. 3: 53-64.

[26] Kausar N., Belhaouari Samir B., Sulaiman S. B., Ahmad I. and Hussain M. 2012. An Approach towards Intrusion Detection Using PCA Feature Subsets and SVM, 2012 International Conference on Computer and Information Science (ICCIS). 2: 569-574.