Classification of Pulmonary Nodules on Computed Tomography Scans. Evaluation of the Effectiveness of Application of Textural Features Extracted Using Wavelet Transform of Image

Marina Sergeeva, Igor Ryabchikov, Mark Glaznev, Natalia Gusarova
Saint Petersburg National Research University of Information Technologies, Mechanics and Optics
Saint Petersburg, Russian Federation
susandean@mail.ru, i.a.ryabchikov, markgxcv@gmail.com, natfed@list.ru

***Abstract*—In this paper, we present an efficient lung nodule classification system. The proposed system extracts various texture and morphological features from the pre-marked regions of interest containing lung nodules. The Gray Level Co-occurrence Matrix method is applied to extract the texture characteristics of the lung nodules. Additional features are extracted from the images transformed by discrete wavelet transform to extend the variety of texture features. Attribute selection is performed on the resulting feature vector to reduce the redundancy of the dataset. Resulting feature collection is used to train several classification models (Neural Network, Random Forest, K-nearest neighbors, Naive Bayes classifier, Support vector machine) which are used to distinguish regions of interest containing malignant nodules from those containing benign nodules.**

I. Introduction

Cancer is the leading cause of death nowadays and lung cancer, in particular, is responsible for more deaths than any other cancer. The reason of that high mortality rate is that lung cancer is very difficult to treat especially if diagnosed on late stages. The possibility to cure lung cancer altogether or to extend patients life to a considerable amount of time highly depends on early diagnostics. But, unfortunately, most lung cancers are recognized only in the late stages of the illness, significantly decreasing the overall lung cancer survival rate. Meanwhile, computed tomography is an important tool for early malignant lung nodules detection, but the interpretation of CT images can be a very exhausting task for radiologists. Computer-aided diagnosis (CAD) systems, therefore, can be of great help to radiologists to raise the possibility of early lung cancer recognition and reduce the probability of false diagnosis.

All computer-aided diagnosis systems developed to increase the possibility of the early lung cancer recognition have similar structure and consist of five main stages [1]:

- Lung segmentation.
- Lung nodule detection.
- Lung nodule segmentation.
- Nodule features extraction.
- Nodule classification.

Each of this stages represent a non-trivial task for which the CAD system developers are trying to offer the best solution.

This work aims to improve the lung nodule features extraction stage and to choose the most efficient machine learning algorithm for the lung nodule classification.

Under this paper, we have designed and implemented the system for lung nodules classification based on various texture and morphological features extracted from computer tomography scans (CT-scans) as well as texture features extracted from scans preprocessed by discrete wavelet transform.

A comparative study of different machine learning methods was performed on extracted data to choose the most accurate one.

II. Dataset

The dataset, created by the Lung Image Database Consortium (LIDC) [2] was used for system training and validation purpose. LIDC database consists of CT images of 1010 patients with annotations from up to 4 radiologists.

To obtain radiologists annotations of lung nodules presented on a scan, a two-phase process was applied [3]. For each nodule greater than 3 mm in size radiologists specified their opinion about various nodule characteristics and their assumption of the likelihood of nodule's malignancy.

Even though annotations are provided for more than a thousand patients, the confirmed diagnosis data is available only for 79 cases. We have made a decision to use only diagnostic data for training the classification system as it provides the more objective malignancy measure than the "likelihood of malignancy" specified by radiologists.

ISSN 2305-7254

As there can be many nodules associated with the single patient's CT-scan and each nodule can have annotations from several radiologists from 79 CT-scans the 321 lung nodule instances have been extracted.

Each CT-scan contains several slices of a human body, taken around a single axis. Every lung nodule may be present on several slices, so for feature extraction purposes the single slice on which the nodule has the biggest area is chosen. This approach bases on the assumption that the slice with the biggest nodule area reflects the morphological features of a nodule as a whole more accurately. Moreover, as the radiologist's experience suggests the nodule size is one of the most significant features in nodule analysis. The nodule with bigger size in particular is more likely to be proved malignant. The example of a pulmonary node on a CT-scan can be seen on Fig. 1.

III. Related studies

According to [4], one of the main ways to increase CAD-system efficiency is to apply image preprocessing techniques. In the [5] work the SIFT (Scale Invariant Feature Transform) algorithm is used to separate the regions of interest on the images containing lung nodules. [6] uses the fractal analysis technique to describe the nodules texture. [7] work makes use of the deep features, formed on the deep layers of the multilayer artificial neural network.

Meanwhile wavelet transform, despite the prospects of applying it to extract the specific texture features of the image [8] is rarely used in the CAD-systems. Previously published works [9-12] used wavelet-transform to preprocess the entire CT-scan to separate the ROI containing the whole lung [11] or to distinguish scans containing malignant nodules [10], [12]. In the [13] work the application of wavelet transform to preprocess small segments of the scan containing lung nodules is demonstrated.

Various computational methods are used for the classification of pulmonary nodules [4], among which machine learning algorithms, such as Artificial Neural Networks (ANN) [14], deep belief network (DBN) [15], decision trees (DT) [7] and support vector machine (SVM) [12] play significant role.

As is clear from the literature review, a common approach to the formulation of the problem of lung nodules CAD-systems is yet to be found. In the papers [12], [14] the classification problem is stated as separation of CT-scans containing nodules (both malignant and benign) from those without nodules, while [15] considers separation of images with malignant tumors from images with only benign. The work [12] analyzes the whole CT-scan while [15] work considers only the part of the scan on which lung nodule is present. The common approach is to apply machine learning algorithms to the nodule images classified by the radiologists [12], [15], [16].

All of the above makes the systematical comparison of the lung nodule classification algorithms efficiency considerably difficult. Taking the provisions stated above into account we assume that the [7] work is the most appropriate paper for the comparison with this work, as it uses the clinically approved data from the same LIDC database as in our paper and formulates the classification problem to distinguish between malignant and benign nodules.

IV. Features extraction

The lung nodule can be classified by means of the analysis of its form, size and texture on the CT-scans. Thus, in order to ensure the highest classification quality, the features used should reflect those nodule characteristics. In this work, we have used the features which efficiency was proved by several papers, including [16], [17].

A. Morphological features

- Area — the total amount of pixels on slice occupied by the nodule.
- Perimeter — the number of pixels in the boundary of the nodule on a slice.
- Irregularity Index — the metric value of roundness of the nodule shape. Equals to 1 only for a circle and is lesser than 1 for any other shape.

$$\frac{4\pi * Area}{Perimeter^2}$$

- Equivalent diameter — the scalar that specifies the diameter of a circle having the same Area as the nodule region.

$$\sqrt{\frac{4 * Area}{\pi}}$$

- Feret max diameter — the maximum possible distance between two points of the nodule region.
- Feret max diameter / equivalent diameter — the measure of the nodule shape oblongness.
- Length — the Z-axis nodule length. Computed as the maximum Z-axis distance between two slices containing the same nodule.
- Convex area — the number of pixels in convex hull [17] which is the size of the bounding box of the region. Bounding box may be imagined as the shape formed by a rope stretched around the nodule object. A convex hull is demonstrated in Fig. 2.
- Solidity — the ratio of the Area of a nodule to the Convex area.

$$\frac{Area}{Convex\ \ area}$$

B. Texture features

The extraction of texture features has been achieved by using Gray Level Co-occurrence Matrix (GLCM) [18].

GLCM is the statistical method of examining image texture

by considering the spatial relationship of its pixels. The GLCM matrix is computed by calculating how often pairs of pixel with specific gray level values occur in an image in a specified dimensional relationship. The dimensional relationship consists of a specified direction (angle) and a distance between two pixels in an image. It is shown in Fig. 3.

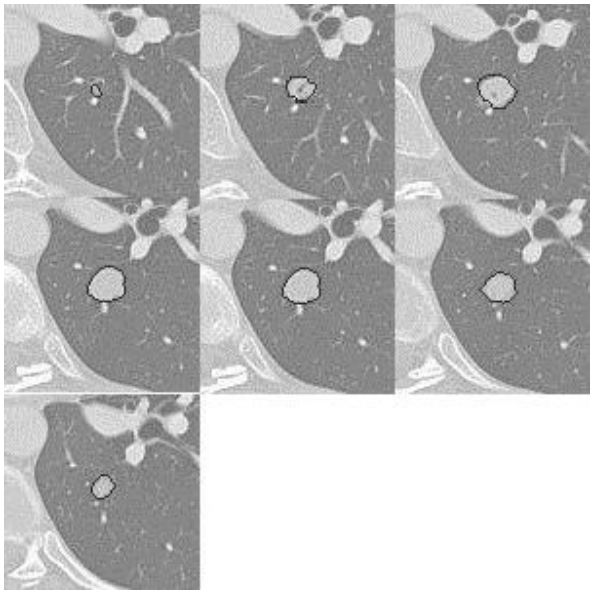

Fig. 1. Pulmonary node on a series of CT images

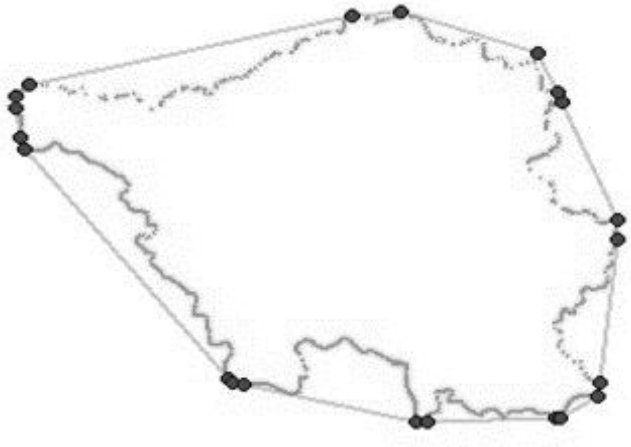

Fig. 2. Convex hull

This paper considers four directions (0, 45, 90 and 135 degrees) to afford the extracted features to capture discriminating information along various directions of target image and two distances (of 1 and 2 pixels). Thus, the following offsets represent the spatial relationships: [0 1; 0 2; -1 1; -2 2; -1 0; -2 0; -1 -1; -2 -2].

The GLCMs were calculated for all dimensional relationships (eight matrices in total). The example on Fig. 4 shows how values in the GLCM are calculated (for the offset [0 1]).

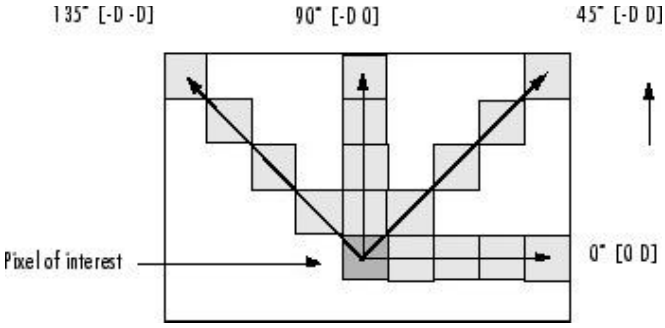

Fig. 3. Pixel shifts

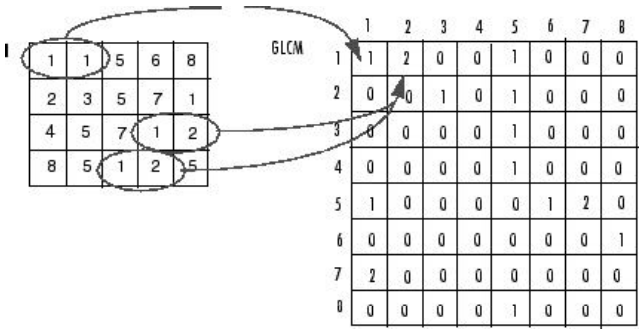

Fig. 4. GLCM formation

For each resulting matrix, the number of features was extracted. To obtain a single feature vector for each nodule image, the values of the same feature for matrices were averaged over the matrices computed with the same distance value but different directions. After this step from two resulting feature values (one for each initial distance), the smallest value was selected. Thus, for each image the following sets of features were calculated:

- Contrast — the measure of the intensity difference between two pixels in the specified dimensional relationship.

$$\sum_i \sum_j (i-j)^2 * p(i,j)$$

where i and j – gray-level values, $p(i,j)$ – the probability to find the specified pixel pair on an image.

- Correlation — specifies the measure of how correlated a pixel is to its partner in the dimensional pair.

$$\sum_i \sum_j \frac{(i-\mu_i)*(j-\mu_j)*p(i,j)}{\delta_i * \delta_j}$$

where μ_i and μ_j — are the mean values of gray-level of the two pixels in the relationship, δ_i, δ_j — standard deviations of gray-level values.

- Energy – equals to 1 for the image with the constant value of gray level.

$$\sum_i \sum_j p(i,j)^2$$

- Homogeneity — equals one for the diagonal matrix and represent the closeness to the diagonal matrix otherwise.

$$\sum_i \sum_j \frac{p(i,j)}{1+|i-j|}$$

- Entropy — the measure of the gray-level values distribution randomness.

$$-\sum_i \sum_j p(i,j) * \log(p(i,j))$$

- Third order moment (or skewness) is a measure of the asymmetry of the distribution of a probability about its mean.

$$\sum_i \sum_j (i-j)^2 * p(i,j)$$

- Inverse difference moment – sum of probabilities divided by squared difference between *i* and *j*.

$$\sum_i \sum_j \frac{p(i,j)}{1+(i-j)^2}$$

- Sum average – sum of probabilities multiplied by average of *i* and *j*.

$$\frac{1}{2}\sum_i \sum_j (i+j) * p(i,j)$$

- Variance - gives a measure of the data distribution about the mean value.

$$\frac{1}{2}\sum_i \sum_j ((i-\mu_i)^2 + (j-\mu_j)^2) * p(i,j)$$

- Cluster tendency – measures the degree to which a data set contains clusters.

$$\sum_i \sum_j p(i,j)(i-\mu_i + j-\mu_j)^2$$

- Maximum probability – the highest probability in data set.

$$\max(p(i,j))$$

V. WAVELET TRANSFORMATION

Wavelet transform is widely used for the signal encoding and for the data and image compression [19]. In particular, the famous image compression standard JPEG 2000 performs lossless compression by using wavelet transformation [20]. Moreover, the wavelet transform is an efficient tool for the analysis of the image texture [8].

Despite the prospect of applying this transform to the extraction of characteristic properties of image texture, it is not widely used by the CAD systems for this purpose. In the previously published papers [9-12] wavelet transform is applied to the whole CT-scan for the lung segmentation [11] and malignant lung nodule detection [10, 12] purposes. In another work [13] the method of preprocessing small CT-scan segments containing lung nodules by the means of wavelet transform is proposed. This technique provides the very accurate way to characterize the texture of the nodule and the surrounding areas of the lung.

Under this paper, we have applied wavelet transform to preprocess the 32*32 pixel segments of the CT-scan containing lung nodules. The size of these segments was selected considering the average size of a nodule, to ensure that the maximum number of nodules will fit into their segments. We have used the Cohen-Daubechies-Feauveau family of biorthogonal wavelets to preprocess images. Wavelet transformation represents a decomposition of the one-dimensional signal *f(x)* onto a basis of wavelet functions (1), which is usually obtained by translation and dilation of a single function Ψ – wavelet mother (2), which is localized in both spatial and frequency domain.

$$(w_a f)(b) = \int f(x)\psi^*_{a,b}(x)dx \quad (1)$$

$$\psi_{a,b} = \frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right) \quad (2)$$

When *a* and *b* are restrained to a discrete values, the discrete wavelet transform (DWT) is obtained. This transform has an implementation in which every wavelet corresponds with a high- and low-pass pair of filters.

Wavelet decomposition of an image (two-dimensional transform) can be computed by applying the low-pass filter (scaling function) and with a high-pass filter (wavelet function Ψ) to the rows and columns of an image. This transform provides 4 sub-images: LL, LH, HL and HH. Every sub-image contains information of specific scale and orientation. LL contains the coefficients of the image having the lowest frequency (or approximation coefficients), while LH, HL and HH contain the vertical, horizontal and diagonal detail information respectively. The same transform can be applied to the resulting LL sub-images recursively to obtain further decomposition.

For this work Cohen-Daubechies-Feauveau 1st level wavelet decomposition was performed on 32x32 pixel parts of original CT-scans (containing regions of interest (ROI)), producing 4 sub-images to analyze. The example is demonstrated on Fig. 5.

VI. BEST FEATURES SELECTION

From obtained sets of features, subsets were selected, in which features were mostly correlated with classes and least — with each other. Thus, redundancy of feature sets was reduced. The redundancy reduction allows to avoid the overfitting of the classification model [21] and helps to eliminate the features that have no correlation with the target variable (lung nodule malignancy measure). To find the best combinations of features, genetic algorithm was used [21]. It is one of the most popular algorithms for the feature subset selection, consistently showing accurate results.

First, the algorithm randomly forms 450 sets of features (forms population). Then in each iteration crossbreeding and mutation of individuals (feature sets) of the population performs. As a result, new individuals similar to their «parents» generate but have random deviations (mutations). Probability of mutation were set 0.033. Generated individuals get added to the current population and then selection of 450 performs. Each individual from the current population is evaluated, the best are selected and the others are removed from the population. So it makes the next "generation" of individuals. Thus, in each iteration the algorithm will try to generate better feature sets.

The indicator evaluating a set of features is a coefficient, proportional to the correlation of each feature with the class and inversely proportional to the correlation between features. To evaluate feature sets 80 random instances from the common dataset were taken.

The WEKA library [22] GeneticSearch implementation of genetic algorithm and CfsSubsetEval class for evaluating of feature sets were used.

As a result, from the first feature set features were selected:

- Feret max diameter / equivalent diameter.
- Feret max diameter.
- Length.
- Area.
- Perimeter.
- Irregularity Index.
- Equivalent Diameter.
- Homogeneity 1.
- Third order moment 1.

From the second set comprising the texture features obtained from transformed images:

- Feret max diameter / equivalent diameter.
- Feret max diameter.
- Length.
- Area.
- Perimeter.
- Irregularity Index.
- Equivalent Diameter.
- Contrast 5.
- Inverse difference moment 3.
- Sum average 1.
- Sum average 5.
- Entropy 1.
- Homogeneity 2.
- Third order moment 1.

The numbers in names of textural features say from which image they were obtained: 1 — from the original image, 2 — from preprocessed LL, 3 — LH, 4 - HL, 5 — HH.

VII. CLASSIFICATION

Then several machine learning algorithms were applied and evaluated. For training and evaluation of algorithms 241 remaining data instances were used. 10-fold cross-validation algorithm was used. The algorithm divides the data into 10 parts and performs 10 iterations. In each iteration a dataset composed of 9 parts is used for training of an algorithm and the 10-th part is used for evaluation. Then results having been obtained at all iterations are averaged.

The feature values were normalized before the application of the algorithms in order to decrease the deviation between absolute values of different features. The normalization was performed on feature value by subtracting its mean value and dividing by its standard deviation.

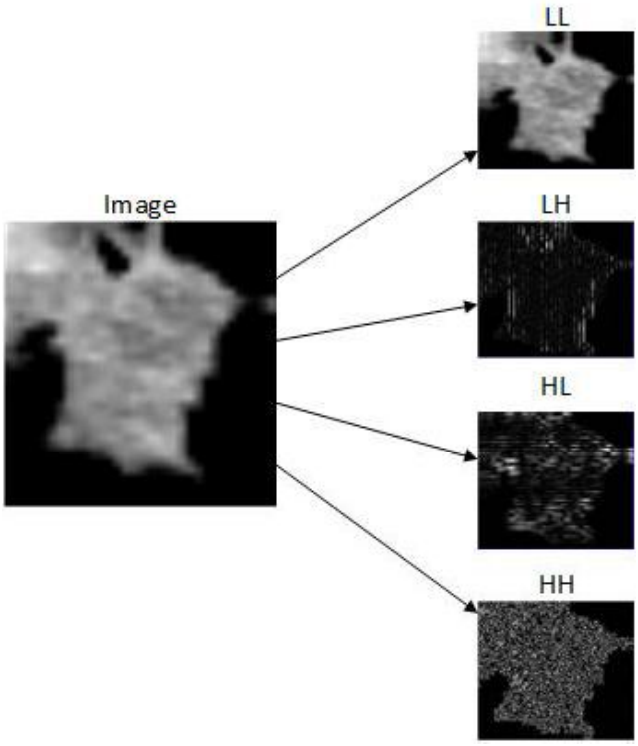

Fig. 5. First level wavelet decomposition of the ROI containing malignant nodule

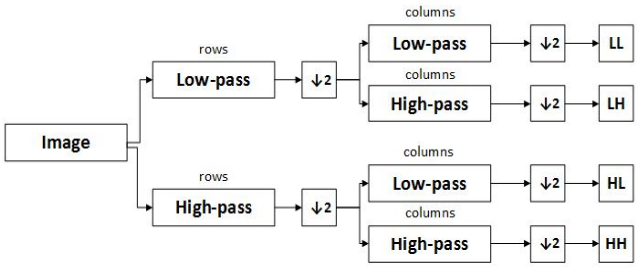

Fig. 6. Two-dimensional wavelet transform mechanics

The following algorithms are used:

- Neural network — the algorithm building a network of «neurals» where each neural calculates a value of a

nonlinear (in our case a sigmoid) function of a weighted sum of input numbers and returns a calculated value as an output. The input of neurals can be values of features (on the 1st layer of neurons) or output values of other neurals (on other layers). A single neural is a linear classificator. The goal of training an algorithm is to find optimal values of weights of neurals inputs. The example of a neural network scheme is illustrated on Fig. 7. Neural network with one hidden layer with number of nodes set to number of features plus number of classes divided by 2 were used.

- Naive Bayes classifier — classifier, based on the Bayes' theorem which characterizes the probability of an event, based on conditions that might be connected to it. The algorithm makes a «naive» guess that features of interest are not related to each other and if that is so — the algorithm is recognized as optimal.
- Random Forest — the algorithm building a number of decision trees by random sets of elements and features of a training dataset. A resulting classification algorithm applies all built decision trees to input case and averages the results.
- Support vector machines – the algorithm tries to build an optimal hyperplane separating the n-dimensional feature space into 2 parts.
- K-nearest neighbors — to classify an instance determines k instances of the training set nearest to it and defines which class dominates among neighbors. Euclidean distance is taken as a function of the distance. 5 nearest neighbors are considered.

Results obtained by each algorithm are demonstrated in Table I and Table II. Where TN — the number of instances correctly classified as benign. FN — the number of instances incorrectly classified as benign. TP — the number of instances correctly classified as malignant. FP — the number of instances incorrectly classified as malignant. Correctly Classified — the proportion of correctly classified cases of the total number of cases. Precision - the ratio of a number of cases correctly classified as malignant to a number of all cases classified as malignant (TP / (TP + FP)) (0 - the worst result, 1 - the best). Recall — the ratio of a number of cases correctly classified as malignant to a number of all actually malignant cases (TP / (TP + FP)) (0 - the worst result, 1 - the best). F-Measure - coefficient considering Recall and Precision together. It is calculated as 2*Precision*Recall / (Precision + Recall). (0 - the worst result, 1 - the best). A classifier can vary the threshold of malignancy probability when to classify a case as malignant or benign (for instance, classify a case as malignant if calculated malignancy probability is greater than 0.4 (but not 0.5) to reduce a number of cases incorrectly classified as benign), at the same, time increasing the value of one and reducing the value of another coefficient (Precision or Recall). F-Measure takes into account when increasing the value of one by decreasing the value of another. Area Under ROC. Receiver Operating Characteristic (ROC) - is a graphical plot that demonstrates the accuracy of a binary classified system (0.5 - the worst result, 0 or 1 - the best).

It can be seen that the use of other textural features extracted from transformed images increased the classification accuracy by 7 percent.

The ROC-curves for the various machine-learning techniques applied with the two subsets of features (mentioned above) are shown on the Fig. 8 and 9 and confirm the results in Table I and Table II.

The accuracy of the implemented system was compared with an accuracy of radiologists on the dataset used for training and testing the system. For each instance a radiologist specified the likelihood of malignancy on a 5-point scale (1 - least likely, 5 - the highest). If to consider the nodes having the likelihood greater than 3 as malignant, less than 3 - as benign and not to consider equal to 3, we get the results shown in Table III. It can be seen that the system showed better results than radiologists.

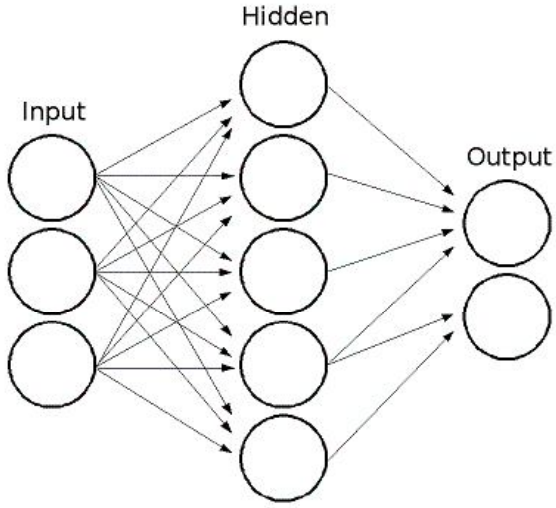

Fig. 7. Neural network scheme example

Also, the system showed greater accuracy than a system, implemented in the existing similar work at 6% [7].

VIII. Mean values of features for benign and malignant cases

In Table IV specified mean values of features for benign and malignant lung tumors.

Obtained differences in the values of features of malignant and benign nodules similar to those obtained in existing papers [16].

IX. Conclusion

The chance of full recovery, or at least, extending the life of the patient diagnosed with the lung cancer depends on the stage at which the disease was detected. Diagnosis of lung cancer - a rather difficult task, especially in the early stages. Computer diagnostics systems are able to provide doctors additional information to classify the disease, reduce the number of false diagnoses and, thus, to allow to begin treatment of the patient quicker.

Applying of textural features extracted from images of tumors preprocessed by wavelet transformation increased accuracy of the classification system by 6%.

The system of computer diagnostics created in this paper showed the greater accuracy than a system, implemented in the existing similar work at 6% [7].

From the applied machine learning algorithms used for the creation of a classification algorithm, the best results were obtained using the K-nearest neighbors with 5 nearest neighbors and Random Forest and using the Neural Network, Naive Bayes classifier and Support Vector Machine worse results were obtained, which contradicts the results of the paper [23].

ACKNOWLEDGMENT

We would like to thank the Lung Image Database Consortium for providing a great database for analysis.

REFERENCES

[1] A. El-Baz, G.M. Beache, G. Gimel'farb, K. Suzuki, K. Okada, A. Elnakib, A. Soliman and B. Abdollahi, "Computer-aided diagnosis systems for lung cancer: challenges and methodologies", *International journal of biomedical* imaging, Jan. 2013.

[2] The Cancer Image Archive (TCIA) official website, The LIDC-IDRI database description, Web: https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI.

[3] S. G. Armato, M. F. McNitt-Gray, A. P. Reeves, C. R. Meyer, G. McLennan, D. R. Aberle, E. A. Kazerooni, H. MacMahon, E. J. van Beek, D. Yankelevitz, E. A. Hoffman, C. I. Henschke, R. Y. Roberts, R. M. Engelmann, R. C. Pais, C. W. Piker, D. Qing, M. Kocherginsky, B. Y. Croft and L. P. Clarke, "The Lung Image Database Consortium (LIDC): an evaluation of radiologist variability in the identification of lung nodules on CT scans", Academic radiology, vol.14, Nov. 2007, pp. 1409-1421.

[4] I. S. Valente, P. C. Cortez, E. C. Neto, J. M. Soares, V. H. de Albuquerque and J. M. Tavares, "Automatic 3D pulmonary nodule detection in CT images: A survey", *Computer Methods and Programs in Biomedicine*, vol. 124, Feb. 2015, pp. 91-107.

[5] A. Farag, A. Ali, J. Graham, A. Farag, S. Elshazly and R. Falk, "Evaluation of geometric feature descriptors for detection and classification of lung nodules in low dose CT scans of the chest", *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium*, Mar. 2011, pp. 169-172.

[6] P. L. Lin, P. W. Huang, C. H. Lee and M. T. Wu, "Automatic classification for solitary pulmonary nodule in CT image by fractal analysis based on fractional Brownian motion model", *Pattern Recognition*, vol. 46, Dec. 2013, pp. 3279-3287.

[7] D. Kumar, A. Wong and D. A. Clausi, "Lung Nodule Classification Using Deep Features in CT Images", *Computer and Robot Vision (CRV)*, Jun. 2015, pp. 133-138.

[8] G. Van de Wouwer, P. Scheunders and D. van Dyck, "Statistical texture characterization from discrete wavelet representations", *IEEE Trans Image Processing*, vol.8, Apr. 1999, pp. 592-598.

[9] M. Assefa, I. Faye, A. S. Malik and M. Shoaib, "Lung nodule detection using multi-resolution analysis", *Complex Medical Engineering (CME)*, May 2013, pp. 457-461.

[10] X. Ye, X. Lin, J. Dehmeshki, G. Slabaugh and G. Beddoe, "Shape-based computer-aided detection of lung nodules in thoracic CT images", *IEEE Trans Biomed Engineering*, vol.56, Jul. 2009, pp. 1810-1820.

[11] O. Talakoub, J. Alirezaie and P. Babyn, "Lung segmentation in pulmonary ct images using wavelet transform", *Acoustics, Speech and Signal Processing*, vol.1, Apr. 2007, pp. 453-456.

[12] H. Madero Orozco, O. O. Vergara Villegas, V. G. Cruz Sanchez, J. Ochoa Dominquez Hde and J. Nandayapa Alfaro Mde, "Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine", *Biomedical engineering online*, vol.12, Feb. 2015.

[13] T. Osicka, "Wavelet-based Pulmonary Nodules Features Characterization on Computed Tomography (CT) Scans", *The Catholic University of America Washington*, 2008.

[14] R. Kohad and V. Ahire, "Application of Machine Learning Techniques for the Diagnosis of Lung Cancer with ANT Colony Optimization", *International Journal of Computer Applications*, vol. 113, Mar. 2015.

[15] K. L. Hua, C. H. Hsu, S. C. Hidayati, W. H. Cheng and Y. J. Chen, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique", *OncoTargets and Therapy*, vol.8, Aug. 2015, pp. 2015-2022.

[16] N. S. Lingayat and M. R. Tarambale, "A computer based feature extraction of lung Nodule in chest x-ray image", *International Journal of Bioscience, Biochemistry and Bioinformatics*, vol.3, Nov. 2013, pp. 624-629.

[17] R. Susomboon, D. S. Raicu and J. D. Furst, "Pixel-based texture classification of tissues in computed tomography", *CTI Research Symposium*, 2006.

[18] R.M. Haralick, "Statistical and structural approaches to texture", *Proceedings of the IEEE*, vol.67, May 1979, pp. 786-804

[19] I. Bocharova, *Compression for multimedia.* New York: Cambridge University Press, 2010, p. 269.

[20] J. Li, "Image Compression: The Mathematics of JPEG 2000", *Modern Signal Processing, MSRI Publications*, vol.46, 2003, pp. 185-221.

[21] T. M. Mitchell, *Machine Learning 1st Edition.* McGraw-Hill Science/Engineering/Math, Mar. 1997.

[22] The WEKA project official website, The WEKA project description, Web: http://www.cs.waikato.ac.nz/ml/weka/.

[23] P. Naresh and Dr. R. Shettar, "Early Detection of Lung Cancer Using Neural Network Techniques", *Prashant Naresh Int. Journal of Engineering Research and Applications*, vol.4, Aug. 2014.

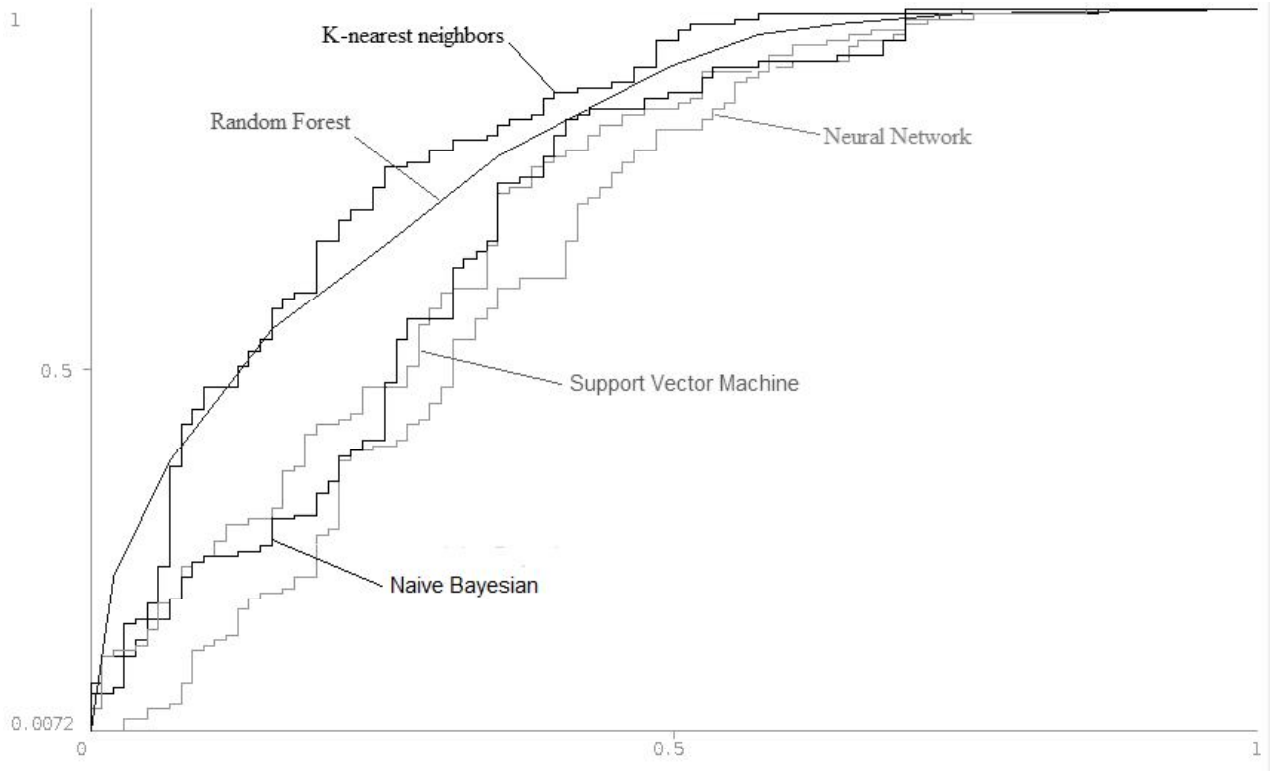

Fig. 8. ROC-curves illustrating the accuracy of classification systems based on several machine learning algorithms and the subset of features extracted from images without wavelet preprocessing

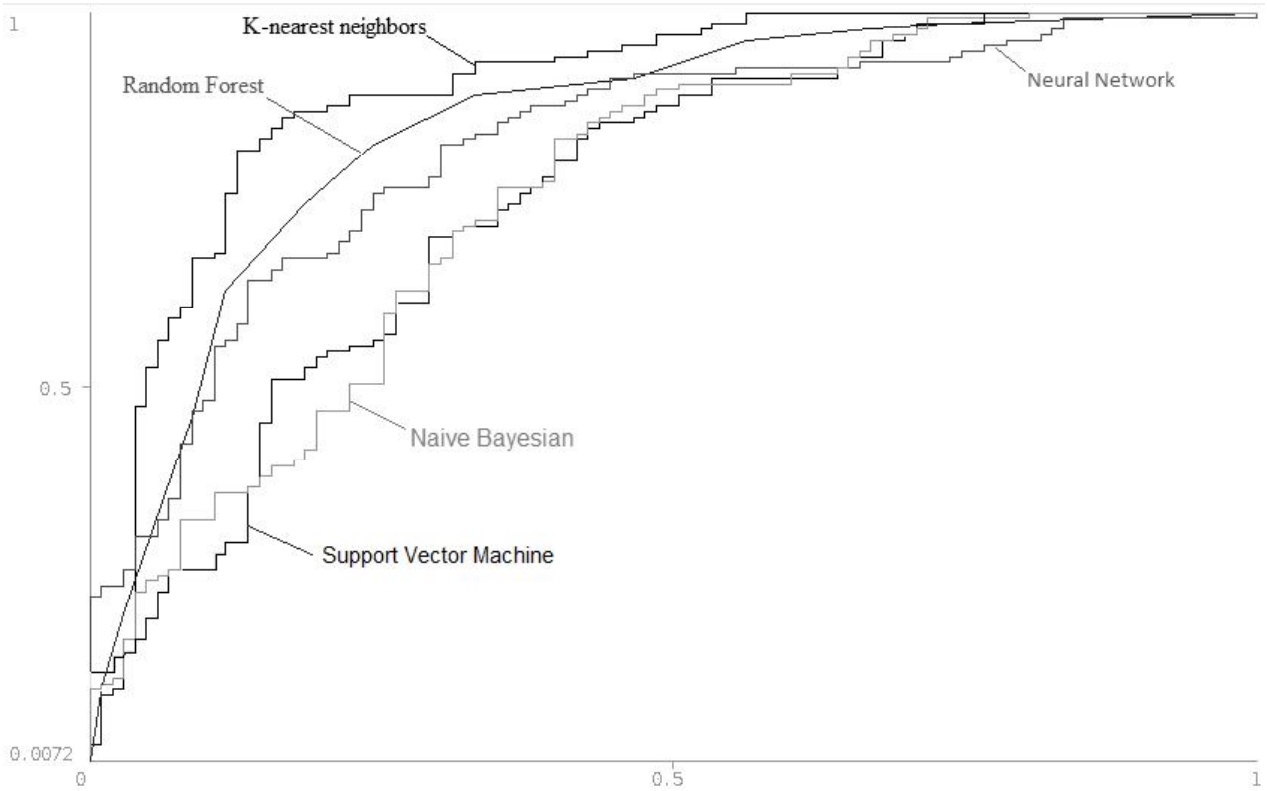

Fig. 9. ROC-curves illustrating the accuracy of classification systems based on several machine learning algorithms and the subset of features extracted from images with wavelet preprocessing

TABLE I. RESULTS OF APPLYING MACHINE LEARNING ALGORITHMS USING THE FIRST SET OF FEATURES

Classifier	TN	FN	TP	FP	Accuracy	Precision	Recall	F-Measure	Area Under ROC
KNN 5	113	35	68	25	0.751	0.731	0.66	0.694	0.823
Random Forest	119	44	59	19	0.739	0.756	0.573	0.652	0.81
Support vector machine	119	50	53	19	0.714	0.736	0.515	0.606	0.733
Naive Bayes	119	49	54	19	0.717	0.74	0.524	0.614	0.73
Neural Network	127	60	43	11	0.705	0.796	0.417	0.547	0.678

TABLE II. RESULTS OF APPLYING MACHINE LEARNING ALGORITHMS USING THE SECOND FEATURE SET WITH FEATURES EXTRACTED FROM TRANSFORMED IMAGES

Classifier	TN	FN	TP	FP	Accuracy	Precision	Recall	F-Measure	Area Under ROC
KNN 5	123	30	73	15	0.813	0.83	0.709	0.765	0.896
Random Forest	123	34	69	15	0.797	0.821	0.67	0.738	0.85
Neural network	116	34	69	22	0.768	0.758	0.67	0.711	0.822
Support vector machine	120	50	53	18	0.718	0.746	0.515	0.609	0.761
Naive Bayes	124	50	53	14	0.734	0.791	0.485	0.601	0.76

TABLE III. COMPARISON OF THE ACCURACY OF RADIOLOGISTS AND THE SYSTEM

DM	TN	FN	TP	FP	Accuracy	Precision	Recall	F-Measure
Radiologist	65	27	48	18	0.715	0.727	0.64	0.681
System	123	30	73	15	0.813	0.83	0.709	0.765

TABLE IV. MEAN VALUES OF FEATURES FOR BENIGN AND MALIGNANT TUMORS

Features	Benign	Malignant
Feret max diameter / equivalent diameter	0.815	0.778
Feret max diameter	12.71	24.85
Length	5.696	12.125
Area	95.7	373.2
Perimeter	36.65	75.58
Irregularity Index	0.7957	0.7011
Equivalent Diameter	10.14	18.72
Contrast 5	73.93	141.7
Inverse difference moment 3	0.08873	0.0701
Sum average 1	26.5	40
Sum average 5	14.21	16.87
Entropy 1	5.719	6
Homogeneity 2	0.0705	0.0682
Third order moment 1	-2440.9	-1262.4