Genetic Algorithm Based Classification for the Lung Needle Biopsy Images

S.T. PremKumar1, M. Sangeetha2
PG Student1, Assistant Professor2
Department of Computer Science and Engineering1, 2, Adithya Institute of Technology1, 2, India
premkumarsmt@gmail.com1, sangee_ganapathy@gmail.com2

Abstract—Lung cancer is one of the most common malignant cancers that spread worldwide. Nowadays, detection of lung cancers takes place in advanced stages. Novel detecting technique by CT scanning of cancer making it possible to detection in its initial stages, when it is most curable. Cancer is an irregularly growing of cells in lungs. Normally cancer will grow from cells of the lungs, blood vessels, nerves that appear from lungs. Types of cancer which are-squamous carcinoma, adenocarcinoma, small cell cancer and nuclearatypia. Specialist conclude size and position of cancer, if it is increasing into nearby tissues region, and if it has chance to spread into lymph glands in the neck. This approaches also check for spread of cancer in lungs and its neighboring tissues. Cancer can damage the normal lungs cells by producing swelling, exerting compression on parts of lungs and increasing pressure inside the chest. Each type of cancer has individual characteristics. This technique is especially for classifying dissimilar cancerous types in the lung tissue.

Keywords: lung needle biopsy, dictionary learning, genetic algorithm, hierarchical fusion.

1. INTRODUCTION

Lung cancer studied through by doctors but its grading gives different decisions which may differ from one doctor to another. The classification and accurate determination of lung cancer grade is very vital because it influences and specifies patient’s treatment scheduling and finally their life. A novel technique multimodal sparse representation-based classification (mSRC), suggested for classifying lung needle biopsy images. In this technique data acquisition is done through the new method, the cell nuclei are mechanically segmented by itself from the input images caught by needle biopsy specimens which are obtainable in this research work, which is samples of the human thinking methods and the classification results are compared with some other computer-aided lung cancers analysis methods showing the efficiency of the proposed methodology.

The three features modalities such as texture, color and shape are extracted from the segmented cell nuclei from input images. After this procedure, mSRC goes through a testing level and training level. The training level, three discriminative subdictionaries corresponding to three features information are together knowledgeable by a genetic algorithm directed multimodal dictionary learning approach.[8] The dictionary learning is used to select highest discriminative samples and encourage large disagreement among stdissimilar subdictionaries.

In testing phase, when a novel image comes, a hierarchical fusion scheme is applied, which originally prediction of labels of all cell nuclei by fusing modalities such as color, texture and shape are predicts the label of the image by maximum popular voting. The above cell nuclei areas can be separated into five classes which consist of four cancerous classes and one regular class (non-cancer). Usually, lung cancer can be categorized into four types: squamous carcinoma (SC), adenocarcinoma (AC), small cell cancer (SCC), and nuclearatypia (NA). Fig. 1 shows several sample images of each of the four cancerous types. The results prove that the multimodal information is vital for lung needle biopsy image classification, this technique is particularly for classifying various cancerous types. The output of the proposed novel mSRC model present reasonably higher accuracy, and are more similar with other existing algorithms.

Figure 1 Biopsy images of lung cancer types.
2. LITERATURE REVIEW

The lungs are a complex tissue which contains numerous structures, such as vessels, splits, bronchi or pleura that can be situated near to lung nodules. Very simple thresholding approaches are regularly enough for the separation of solid well-circumscribed nodules, while lung nodules close to vessels or pleura need additional complex schemes exploiting geometrical and gray level features mined from nearby structures of the lungs. Various methods have been suggested to summarize the lung nodules close to vessels or pleura.

Human associated parasites create a problematic in maximum humid countries, producing death or physical and psychological illnesses. Their conclusion regularly trusts on the visual examination of microscopy images, with error rates that may collection from moderate to higher. [3] The problem has been addressed through computational image analysis, but only for a rare species and images free of fecal layers. In routine, fecal layers are aactual trial for programmed image analysis.

First method exploits ellipse matching and image foresting transform for image subdivision, multiple item descriptors and their optimalgrouping by genetic software design for object symbol, and the optimum-path forest classifier for object acknowledgment. The output demonstrations that this method is a favorable method near the completely automation of the enteroparasitosis diagnosis.

The effectiveness of sparse representations gained by learning a set of over complete basis or dictionary in the context of action recognition in videos. While the work focusses on distinguishing the movement of human, [6] physical appointments as well as the appearance of face and the suggested method is fairly general and can be used to address some other classification methods.

The suggested approach is computationally effective, highly accurate, and is strong against partial sealing, spatiotemporal scale variants, and to some range to view changes. This robustness is attained by manipulating the discriminative nature of the sparse representations joined with spatio-temporal motion descriptors. The fact that the descriptors are mined over multiple temporal and spatial determinations make them indifferent to scale variations. The descriptors being calculated locally make them robust beside blocking or other distortions. Features such as compressed sample can also develop the recognition accurateness but are highly costly computationally [8].

The presentation of the collective system is calculated in fact and the output correctness, rapidity, robustness, and an active retinal vessel segmentation method based on supervised classification using a collective classifier of boosted and bagged decision trees. In this technique 9-D feature vector which contains of the vessel map acquired from the orientation examination of the gradient vectorfield, the morphological revolution, line strength measures and the Gabor filter reaction which translates particulars to successfully handle both standard and pathological retinas.

3. RELATED WORKS

An automatic method for finding brain tumor in the MRI image by means of the simple image processing methods. [8] It is decided that the system result in better detection of presence of tumor in the image. The considerable accuracy level is detected and it mines the important features which are used to identify the class of the tumor. Performance and accurateness of the designed system is found to be better. Hence it can detect the Brain Malignant cells in real time and provide exactness detection of the class of the Brain Tumor.

The system proposed in this technique can be used in future to identify the class of any type of tumor with appropriate changes in the system. The only combination of CT, PET, and MRI etc. can increase the performance of the system. The work includes mining of the significant features for image recognition. The features mined offer the property of the textures are stored in knowledge base. Texture features or more exactly, Gray Level Co-occurrence Matrix (GLCM) features are used to differentiate between standard and irregular brain tumors [8]. Dissimilar features are enlisted as:

**Contrast**

\[ R_1 = \sum_{i,j=0}^{N-1} p(i,j) \times (i - j)^2 \]  

**Angular Second Moment**

\[ R_2 = \sum_{i,j=0}^{N-1} p_{ij}^2 \]  

One of the most vital problems in the segmentation of lung nodes in CT imaging rises from possible accompaniments happening among nodules and other lung structures, such as vessels or pleura. The problematic of vessels additions by suggesting an automatic correction technique applied to an initial rough segmentation of the lung nodule. [12]

**Inverse Difference Moment**

\[ R_3 = \sum_{i,j=0}^{N-1} \frac{p(i,j)}{1+(i+j)^2} \]  

**Entropy**

\[ R_4 = \sum_{i,j=0}^{N-1} p(i,j) \times \left[ -\ln(p(i,j)) \right] \]
Input images are occupied as input which is applied in the equitation (1), (2), (3), (4), (5). Different MRI examples are collected and given as input for the query phase. Database is a grouped database containing 70 dissimilar images characterized into 4 classes. Any automatic, computerized calculation of disease needs establishing of healthy standards against which a test subject can be equated. A high quality, carefully designed image database of healthy subjects could be of value to many clusters for creation of healthy atlases, for calculation of disease, and for calculation of the effects of both gender and healthy aged.

### Table 1 Features Extracted

<table>
<thead>
<tr>
<th>Angle</th>
<th>AS M</th>
<th>Contra st</th>
<th>Entrop y</th>
<th>IDM</th>
<th>Dissimilarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0°</td>
<td>4578</td>
<td>234189</td>
<td>-76.9528</td>
<td>3.073</td>
<td>29370</td>
</tr>
<tr>
<td>45°</td>
<td>4190</td>
<td>323601</td>
<td>-53.3858</td>
<td>3.046</td>
<td>34456</td>
</tr>
<tr>
<td>90°</td>
<td>4804</td>
<td>196481</td>
<td>-145.027</td>
<td>13.08</td>
<td>25346</td>
</tr>
<tr>
<td>135°</td>
<td>4156</td>
<td>290176</td>
<td>-51.9995</td>
<td>3.069</td>
<td>32744</td>
</tr>
</tbody>
</table>

All images were divided for the existence of disease. Images contain SC, AC, SC, and NA which goal to make this database effort efficiently. Table 1 describes the features mined from affected MRI. The features are calculated by using GLCM in four dissimilar angles (0°, 45°, 90°, and 135°). The final goal in image processing applications is to extract significant features from the image data, from which a graphic, informational, or reasonable view can be attained by the device.

**Dissimilarity**

\[ R_5 = \sum_{i,j=0}^{W-1} p(i,j) \ast |(i-j)| \quad (5) \]

An image preprocessing step goals to obtain the distinct cell nuclei from the caught images by segmenting the cell nuclei from the background. The image preprocessing step is with the following stages: smoothing the images through Gaussian kernel, segmenting the images by means of Otsu’s algorithm, and labeling the associated with cell nuclei regions. The reason of adopting Otsu’s algorithm is that the difference between the cell nuclei region and background is large enough to be simply separated see sample images in Fig.3. Fundamentally, these segmentation results can meet scientific requirements according to the pathologist’s suggestions.

**PROPOSED METHOD**

Usually, the suggested method contains three phases (Fig. 2). The first phase is the data acquisition procedure, which purpose is to extract the features for cell nuclei in lung needle biopsy images. Later this process and technique go under the rest of the two training and testing phases. In the training phase, the new idea of dictionary in the pattern recognition/computer vision, where dictionary resources a collection of elements or words or feature vectors.

The features extracted from the three modalities such as color, texture and shape of every single cell nucleus in the data acquisition procedure, which build three original subdictionaries on color, texture and shape by collecting the corresponding feature vectors of individual cell nuclei.

![Figure 2 System Architecture](image_url)

![Figure 3 Extraction of features and the label of each testing image is determined by voting on the cell-level label.](image_url)
A feature extraction step goal to extract features for distinct cell nuclei from three modalities color, texture and shape respectively. Exactly, as shown in Table 2, shape-based (9), color-based (11), and texture-based (16) features are mined. For the Fourier descriptor, the alphabet “i” in the bracket means that only use the second, third, and fourth coefficients. The first coefficient is the mean value of border coordinates, which is regularly measured as unusable in feature representation. In the texture-based features, the alphabet “j” in a bracket means that which calculates the four directions (0°, 45°, 90°, 135°) for the four features energy, entropy, contrast, and divergence, therefore totally 16 texture-based features are extracted.

Table 2 features are extracted from every cell nucleus region

<table>
<thead>
<tr>
<th>Color (11)</th>
<th>Shape (9)</th>
<th>Texture (16)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R,G,B</td>
<td>height</td>
<td>energy (4)</td>
</tr>
<tr>
<td>gray variance</td>
<td>circumference</td>
<td>contrast (4)</td>
</tr>
<tr>
<td>H,S,I</td>
<td>elongation</td>
<td>entropy (4)</td>
</tr>
<tr>
<td>gray mean</td>
<td>area</td>
<td>divergence (4)</td>
</tr>
<tr>
<td>central moment</td>
<td>width</td>
<td></td>
</tr>
<tr>
<td>feature gray</td>
<td>circularity</td>
<td></td>
</tr>
<tr>
<td>IOD</td>
<td>Fourier descriptor (3)</td>
<td></td>
</tr>
</tbody>
</table>

5. EXPERIMENTAL SETUP AND RESULT

ANALYSIS:

The proposed real-time lung cancer system is estimated in every duplication and consider supplementary under certain hardware, software constraints and population collection criteria such as tournament and necessities. The hardware platform used throughout the random roulette wheel range.

Theoretical analysis was constructed on an AMD A8 Elite Quad-Core @ 2.10 GHz processor, 4 GB of DDR3 RAM the software development has been supported out in MATLAB. The accuracy, precision, recall, F1 score, and true negative rate (TNR) are employed for calculating the multiclassification results. The accurateness is calculated as the sum of correctly categorized images separated by the amount of total images. For precision, recall, F1 score, and TNR calculate not only the class detailed values of each particular class NA, SCC, SC, NC, and AC but also the mean values by averaging totally the parallel class detailed values.

### Table 4 Comparison with related methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>F1 score</th>
<th>TNR</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>KSRC [37]</td>
<td>0.804</td>
<td>0.953</td>
<td>0.782</td>
<td>0.843</td>
<td>0.830</td>
</tr>
<tr>
<td>LapRLS [36]</td>
<td>0.657</td>
<td>0.907</td>
<td>0.533</td>
<td>0.538</td>
<td>0.625</td>
</tr>
<tr>
<td>MCMI-AdaBoost [16]</td>
<td>0.563</td>
<td>0.899</td>
<td>0.585</td>
<td>0.564</td>
<td>0.608</td>
</tr>
<tr>
<td>ESRC [13]</td>
<td>0.777</td>
<td>0.940</td>
<td>0.730</td>
<td>0.884</td>
<td>0.800</td>
</tr>
<tr>
<td>mcSVM [7]</td>
<td>0.576</td>
<td>0.921</td>
<td>0.598</td>
<td>0.577</td>
<td>0.674</td>
</tr>
<tr>
<td>mSRC</td>
<td>0.862</td>
<td>0.962</td>
<td>0.834</td>
<td>0.913</td>
<td>0.867</td>
</tr>
<tr>
<td>mSRC (rw)</td>
<td>0.866</td>
<td>0.963</td>
<td>0.846</td>
<td>0.901</td>
<td>0.881</td>
</tr>
<tr>
<td>mSRC (tn)</td>
<td>0.846</td>
<td>0.967</td>
<td>0.841</td>
<td>0.858</td>
<td>0.867</td>
</tr>
</tbody>
</table>

6. PERFORMANCE ANALYSIS

Evaluating the Hierarchical Fusion Strategy:

Notice that the output of mcSVM, KSRC, ESRC, and mSRC are by means of the hierarchical fusion strategy to take individual MCMI-AdaBoost and LapRLS without using hierarchical fusion strategy, since both MCMI-AdaBoost and LapRLS essentially belong to image-level analysis techniques. In MCMI-AdaBoost, the space between each two images is calculated by Hausdorff distance. In LapRLS, the k-
means clustering algorithm is applied to approximately partition the training cell nuclei into some clusters.

6.2 Evaluating the Performance of SRC:
The same hierarchical fusion strategy is assumed in mcSVM, KSRC, ESRC, and mSRC. The only variance among ESRC and mcSVM is the classifiers used for the single-modal classification. ESRC uses SRC though mcSVM uses SVM. Conversely, ESRC acquires better classification performance than mcSVM, which authorizes the efficiency of SRC on single-modal classification.

7. CONCLUSION
The novel technique is suggested mSRC for categorizing the lung needle biopsy images. mSRCgoal is to rise the classification performance, which exactly for the images of various cancerous types. The genetic techniquegoal is to select the topmost discriminative samples for every single distinct modality as well as to assure the huge diversity among different features. From the perception of experimental practice, misclassifying a cancerous image as a standard or normal one will be considerably more serious than misclassifying a standard or normal image as a cancerous one. Forthcoming work will examine how to implement the technique on the image set through various class relations, i.e., considering the ratio of malignant nuclei in every image. Similarly the multimodal data is broadly obtainable in medical image exploration due to the numerous data acquisition methods.

REFERENCE
[6]. E.I. Papageorgiou, C.D. Stylios, P.P. Groumpos, An integrated two-level hierarchical decision making system based on fuzzy cognitive maps,


[11]. John Wright, Student Member, IEEE, Allen Y. Yang, Member, IEEE, Arvind Ganesh, Student Member, IEEE, S. Shankar Sastry, Fellow, IEEE, and Yi Ma, Senior Member, IEEE, “Robust Face Recognition via Sparse Representation”, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 31, NO. 2, FEBRUARY 2009.


PREMKUMAR S.T obtained his B.E degree in Computer Science and Engineering from Anna University, 2011, Coimbatore, India and pursuing his M.E degree in Computer Science and Engineering at Anna University, Chennai, India. He has published 7 research articles in various conference proceedings. His research area of interest is Image processing and specialization on classification for lung cancer from lung needle biopsy images.
E-mail: premkumarsmt@gmail.com

M. SANGEETHA obtained her BE degree in Computer Science and Engineering from Anna University, Coimbatore, India. She obtained ME in computer science and engineering from Sri Krishna College of Engineering and Technology, Coimbatore, India. Her research interest focus on Data mining and image processing.
E-mail: sangee_ganapathy@gmail.com