A Diagnostic Study of Content-Based Image Retrieval Technique for Studying the CT Images of Lung Nodules and Prediction of Lung Cancer as a Biometric Tool

Rajeev Dixit1*, Dr. Pankaj Kumar2 and Dr. Shashank Ojha3
1Department of Computer Science, United College of Engineering & Research, Prayagraj Dr. A.P.J. Abdul Kalam Technical University, Lucknow India,
2Department of Computer Science, Sri Ramswaroop Memorial Group of Professional Colleges Lucknow, University: Dr. A.P.J. Abdul Kalam Technical University, Uttar Pradesh, Lucknow, India
3Chest Department, Guru Shri Gorakshnath Chikitsalaya, Gorakhpur, India

*Correspondence: Rajeev Dixit; dixitrajeev19@gmail

ABSTRACT - Content Based Medical Image Retrieval (CBMIR) can be defined as a digital image search using the contents of the images. CBMIR plays a very important part in medical applications such as retrieving CT images and more accurately diagnosing aberrant lung tissues in CT images. The Content-Based Medical Image Retrieval (CBMIR) method might aid radiotherapists in examining a patient's CT image in order to retrieve comparable pulmonary nodes more precisely by utilizing query nodes. Intending a particular query node, the CBMIR system searches a large chest CT image database for comparable nodes. The prime aim of this research is to evaluate an end-to-end method for developing a CBIR system for lung cancer diagnosis.

Keywords: Medical image retrieval, Content-based image retrieval, Lung CT images.

1. INTRODUCTION

Lung cancer endures plaguing mankind as the prominent root of cancer-related fatalities. Timely recognition of pulmonary nodes, which can be an indicator of lung cancer, is the unique approach to recuperate survival rates. On a CT scan, a node manifests as a spherical mass with a diameter of up to 3cm. There are hundreds of slices (images) in CT scan for a patient, and manually processing such a large image set is time-consuming and inefficient. CAD systems are meant to help radiotherapists in the same way, but they have certain limitations because they functioning is somewhat similar to a black box. In other words, a highly effective CAD system might determine whether a patient is malignant or not based on the results of a CT scan. A CBIR system assists radiotherapists with diagnosis by allowing them to view a number of scans from the database of images that are optically comparable to the image of input query.

The CBIR system has two modes of operation and also has a high diagnostic potential. An image repository and a character base are available in offline mode. A vast number of images of mild and malicious nodes may be found inside the image library. Numerous characteristics are retrieved and saved in the character base using the character extraction procedure. The CBIR require an input query image of a mild or malignant node using an HCI in online mode. This image is subjected to a character extraction procedure, and the extracted characteristics are compared to those in the character database. One of the several distance measurements aids this comparison. Visually most comparable pictures are recovered and shown for the domain expert using an HCI based on the threshold value. This study offers a step-by-step approach for developing a CBIR system for lung cancer diagnosis. [1]

Radiotherapists utilize CT imaging as a base to detect lung illnesses since it may offer the essential info for exhibiting whether the tissues of lung are ordinary or irregular [2]. Instead of measuring whole node, the content-based image retrieval technique of pulmonary nodes assess an illustrative share of the node [3]. Furthermore, the previous analysis depends on radiotherapists to manually segment nodes. For clinical usage, the retrieval system with the least amount of user interaction is recommended. The current system's objective is to decrease user interaction while improving retrieval accuracy. The diagnostic efficiency and accuracy can be enhanced by radiotherapists are provided deep knowledge of history. The CBMIR approach is useful to solve this challenge. Considering the outcomes of the survey, a novel method for lung cancer detection and prediction will be proposed, which will aid in the prognosis of lung cancer and early recognition. As a result, the patient's chances of survival will improve.
2. CONTENT BASED MEDICAL IMAGE RETRIEVAL SYSTEM (CBMIR)

Content-Based Image Retrieval are utmost commonly utilized method in Health Image Retrieval. Shape, Texture, Color, and additional information in an image are retrieved using the Content-Based Image Retrieval technique. CBMIR is a leading research field in supercomputer vision and image processing. The primary goal of the Content-Based Medical Image Retrieval System (CBMIR) technique is to extract images quickly. As a result, the suggested CBMIR approach not only combines visual and semantic info to overwhelm their limitations but also makes use of the database's basic structure for improved retrieval efficiency.

2.1 Computed Tomography (CT) Image

Although there are other ways of diagnosing lung cancer [4], such as PET, MRI, and X-ray, the CT scan image is regarded to be the distinctive method for providing a thorough description of lung nodes. It has less distortion and noise, as well as improved clarity, than additional images, making it easier to compute and predict the texture characteristics of the image.

2.2 Dataset

The dataset of lung images was obtained from the NIH/NCI dataset consortium of lung images, which offers computed tomography images of the lung on the network-based cancer [5] imaging archive location. It is open to the public and simple to use.

3. NODE CLASSIFICATION STATEMENTS

There are primarily two research approaches for the organization of nodes in lung CT imaging. (i) 2-type cataloging of node, and (ii) four-type classification of node (built on appearance, shape, and location).
It is too regarded a complication in the categorization of lung node malice shiftiness (possibility of node enmity). The formulation is identical to equation (1), with \( y \) denoting the malice of the node (mild and enmity are signified by 0 and 1). Figure 5 depicts node instances by great malice suspected and little malice.

### 4. MAIN DATASETS

For training and testing the ground truth dataset is required developed algorithms for lung node image categorization. It is difficult to get datasets as easily as in other disciplines due to rules and patient privacy. As a result, some academics rely on local hospitals or commercial databases for their data. Few datasets available publically, such as ANODE09, ELCAP, DLCST, LIDC, JSRT, LIDC-IDRI, NELSON and are currently broadly applied to assess algorithm presentation [7]. In this part, details about these datasets are discussed.

#### 4.1 LIDC-IDRI and LIDC

The image database resource initiative and lung image database consortium (IDRI- LIDC) used thoracic CT images with marked-up annotated scratches for diagnosis and lung cancer screening. The national cancer institute (NCI) was considered as the driving force behind the creation of this database. The majority of the information for the pre-stage lung cancer investigation came from six academic institutes and eight medical imaging firms [8].

IDRI-LIDC has 1018 scans or a total of 1010 patients. There were three types of nodes in the lesion's region: (i) nodes \( \geq 3.2 \) mm, (ii) nodes \( 3.2 \) mm, and (iii) non-nodes \( \geq 3.2 \) mm.

All radiotherapist findings were compiled in an XML file to a given CT sequence. It included the elaboration of altogether nodes larger than 4.2mm (outlines and personal node evaluations), estimated 3d center of mass of nodes less than 4.2mm, and altogether non-nodes (no outline or personal valuations of features). The IDRI-LIDC had 7372 lesions labeled “node” by on slightest radiotherapist, with 2670 of them recognized as node \( \geq 3 \) mm by a minimum of one radiotherapist. Figure 6 shows an example image with shown marks. Table 1 shows the most commonly used node descriptions.

#### 4.2 Public Lung Image Database: ELCAP

In December 2003, the early lung cancer action program (ELACP) database was unrestricted for the primary time. This database is designed in association between the ELCAP and VIA research groups, and it is often used to evaluate the effectiveness of various CAD schemes [11].

There were 380 unrepeated lung node CT pictures and 50 reported reduced whole-lung CT scans in the database. The CT scans stayed acquired in a sole snort grasp with a slice thickness of 1.26 mm. The radiotherapist also gave the positions of the nodes found.

In ELCAP, the proportions of node types be located at W-16 percent, V-17 percent, J-29 percent, and P-40 percent [12]. ELCAP differed after IDRI- LIDC in 2 ways: Non-nodes were not included, and its node size was lowered [13]. The individual situation included an additional *.csv file providing the center location of each detected node, tabulated 2. A piece row represents a lung node. The 1st column contains the scan identification. The 3rd and 4th columns indicate the location of the node on this portion (position, contours, and evaluations) for nodes larger than 3mm, whereas the others provide info (location) for nodes less than 3mm.

<table>
<thead>
<tr>
<th>Event</th>
<th>Node Id</th>
<th>X loc.</th>
<th>Y loc.</th>
<th>Z loc.</th>
<th>Outline of the node</th>
<th>Maligancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>06</td>
<td>Node 001</td>
<td>195</td>
<td>291</td>
<td>36</td>
<td>((190,281),(188,281),..., (191,282), (189,280))</td>
<td>5</td>
</tr>
<tr>
<td>06</td>
<td>IL057 ,1597 ,47</td>
<td>294</td>
<td>267</td>
<td>31</td>
<td>((290,281),(291,278),... (287,278), (288,281))</td>
<td>5</td>
</tr>
</tbody>
</table>

Each row represents a node's important information. The primary column contains a component of the affected node ID. The other column contains the node's unique identification. The 3rd and 4th columns show the node's x and y coordinates. The Z location specifies the portion position where the node is outlined. The 6th column denotes the assortment of (x, y) locations that form the node's border (individual a share of synchronizes are tabulated). The last column indicates the radiotherapists' assessment of the menace of this node. The first two rows give info about (position, contours, and evaluations) for nodes larger than 3mm, whereas the others provide info (location) for nodes less than 3mm.

**Figure 5:** CT image of Lung: The green color shows the Inventive image with node, (b) the portion of CT image with node > 3.2 mm ROI (green color).

**Figure 6:** From ELCAP illustration of CT images of lung [9]: (a) the comprehensive CT pictures, (b, c) shows the slice of CT scans. The sign “1” is the position of node
### 4.3 Others Datasets

Both Japanese Societies of radiological technology collaborated Japanese radiological society to create a database of the JSTR. There were 155 nodes and 94 non-nodes with labels in the database [14]. Age of patient elusiveness grade, gender, node size, and organization were some of the other pieces of data.

Since 2003, the Netherlands-Leuven Longkanker Screenings Onderzoek (NELSON) has gathered information from 15,524 people in four institutions. Yearly CT screenings were understood first at a residential facility and subsequently at a fundamental place. They offered resources such as semi-automated measurement of lung node volume, mild vs. malicious node classification, automated lung node identification, and dissection [15].

The Automatic Node Detection 2009 (ANODE09) database contained 50 test scans and five example scans. ANODE09 served as a testing ground for node identification algorithms. Any team can submit their findings for assessment as a benchmark [17]. Researchers used databases gathered from hospitals that are operated privately or other organizations in several studies. These databases are less well-known, and around isn’t much information about them. Table 3 provides a quick overview of various datasets.

### 5. AIM OF THE STUDY

The main purpose of this research is to examine an endwise method for developing a CBIR scheme for lung cancer diagnosis.

### 6. RELATED WORK

Despite the fact that the literature on generic CBIR comprises a large number of research publications, this section gives a quick overview of the work of various researchers to CBIR systems specialized to analytical radiology. Proposal of an input query image, character extraction, computation of a resemblance amount, and a Human Computer Interface (HCI) to offer imagining to end users are all part of a typical CBIR system. Among all the processing stages, character extraction is the most important. During the character extraction phase, several characteristics such as shade, figure, and feel are mined, and the literature [18] emphasizes how these characteristics can be extracted. Another key aspect in the development of the CBIR system is the computation of resemblance measures. Academics frequently use a variety of metrics to measure similarity, including Euclidean distance, Minkowski distance, Manhattan distance, Cosine similarity, and Jaccard similarity. [19] Provides a thorough analysis of these topics. When it comes to the implication of CBIR in analytic radioscopy in over-all, and CBIR for diagnosis of lung cancer in specific, it is clear the analytic and understanding correctness is steadily improving with the introduction of CBIR devices.

There is a plethora of literature on CBIR systems for lung cancer. For analogous CT image retrieval, [20] employed graphic and conceptual resemblances amid the database's individual images as well as the query image. They tested their method using CT scans of the lungs. [21] Suggested a CBIR method for detecting the malignancy levels of lung nodes. Node density level and node lesion density heterogeneousness are the characteristics they use. Their research used CT scans of the lungs from the LIDC data collection [22]. Utilized 27 Haralick texture characteristics to build GLCM in 4 different alignments. The Mahalanobis distance degree was cast off to assess the correlation amid the question image of a node and the dataset's position node. Employed margin sharpness characteristics to retrieve related pulmonary nodes. Each node in the CT segment has a margin sharpness vector of size 12 retrieved for it [22], collected and kept in the character repository for CT images from LIDC are 2D and 3D shape-, margin-, and texture-based properties. The similarity is calculated using Euclidian, Manhattan, and Chebyshev measurements. Each node had an increased scale, calcification, sphericness, boundary, lobulation, supposition, and degree of malice. An XML parser application is built, and 760 nodes are retrieved from CT images, 375 each of mild and malignant nodes. After that, the node images are harvested to 60*60 pixels and saved in the image base. These pictures are utilized in the character abstraction and post-processing stages [23]. Proposed a model which is created on convolutional neural network (CNN) architecture with ten layers to obtain high accuracy. The model demonstrated a considerable training and validation accuracy of 94% and 92% respectively. [24] Visual Analytics for Medical...
Image Retrieval is a novel procedure for medicinal CBIR proposed in this research (VAMIR). B. [25] Optimized Fuzzy Intensification parameter constants are used to minimize overexposed and underexposed areas and offers elevated contrast improvement [26]. Medical images are pre-processed using hybrid median filter to discard noise and then decomposed using integer wavelet transform and employed modified grasshopper optimization algorithm to select the optimal coefficients for efficient compression and decompression.

7. CONCLUSION

The content-based medical image retrieval technique is a quick in addition easy way to discover and repossess CT images of lungs from a huge medical image folder. CBMIR is an approach for recovering common CT imaging signs of lung diseases (CISLs) that has lately become popular. CBMIR uses a range of methodologies to identify lung diseases early. This study presented a method for developing a CBIR system for retrieving and analyzing lung nodes.

REFERENCES


[26] N. Shyamala and Dr.S. Geetha (2022), Compression of Medical Images Using Wavelet Transform and Metaheuristic Algorithm for Telemedicine Applications. IJEEER 10(2), 161-166. DOI: 10.37391/IEEER.100219

© 2023 by the Rajeev Dixit, Dr. Pankaj Kumar, Dr. Shashank Ojha. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Website: www.ijeer.forexjournal.co.in