J Digit Imaging (2016) 29:716–729
DOI 10.1007/s10278-016-9894-9

Cloud-Based NoSQL Open Database of Pulmonary Nodules for Computer-Aided Lung Cancer Diagnosis and Reproducible Research

José Raniery Ferreira Junior[1] · **Marcelo Costa Oliveira**[1] · **Paulo Mazzoncini de Azevedo-Marques**[2]

Published online: 20 July 2016

Abstract Lung cancer is the leading cause of cancer-related deaths in the world, and its main manifestation is pulmonary nodules. Detection and classification of pulmonary nodules are challenging tasks that must be done by qualified specialists, but image interpretation errors make those tasks difficult. In order to aid radiologists on those hard tasks, it is important to integrate the computer-based tools with the lesion detection, pathology diagnosis, and image interpretation processes. However, computer-aided diagnosis research faces the problem of not having enough shared medical reference data for the development, testing, and evaluation of computational methods for diagnosis. In order to minimize this problem, this paper presents a public nonrelational document-oriented cloud-based database of pulmonary nodules characterized by 3D texture attributes, identified by experienced radiologists and classified in nine different subjective characteristics by the same specialists. Our goal with the development of this database is to improve computer-aided lung cancer diagnosis and pulmonary nodule detection and classification research through the deployment of this database in a cloud Database as a Service framework. Pulmonary nodule data was provided by the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI), image descriptors were acquired by a volumetric texture analysis, and database schema was developed using a document-oriented Not only Structured Query Language (NoSQL) approach. The proposed database is now with 379 exams, 838 nodules, and 8237 images, 4029 of them are CT scans and 4208 manually segmented nodules, and it is allocated in a MongoDB instance on a cloud infrastructure.

Keywords Lung cancer · Pulmonary nodule · Lung Image Database Consortium · Image Database Resource Initiative · Computer-aided diagnosis · Computer-aided detection · 3D texture analysis · NoSQL · Document-oriented nonrelational database · MongoDB · Cloud computing · Database as a Service · Reproducible research

Abbreviations

3DTA	3D texture attributes
API	Application programming interface
BSON	Binary JavaScript Object Notation
CAD	Computer-aided diagnosis
CBIR	Content-based image retrieval
CR	Computed radiography
CT	Computed tomography
DBaaS	Database as a Service
DBMS	Database management system
DICOM	Digital Imaging and Communications in Medicine
DX	Digital radiography
FDA	Food and Drug Administration
IDRI	Image Database Resource Initiative
GLCM	Gray-level co-occurrence matrix
GUI	Graphical user interface
HIPAA	Health Insurance Portability and Accountability Act
IP	Internet protocol
JSON	JavaScript Object Notation

✉ José Raniery Ferreira Junior
jrfj@ic.ufal.br

[1] Lab of Telemedicine and Medical Informatics, University Hospital Prof. Alberto Antunes, Institute of Computing, Federal University of Alagoas, Av. Lourival Melo Mota, Cidade Universitária, 57072-900 Maceió, Alagoas, Brazil

[2] Center of Imaging Sciences and Medical Physics, Internal Medicine Department, Ribeirão Preto Medical School, University of São Paulo, Av. dos Bandeirantes, Monte Alegre, Ribeirão Preto, São Paulo, Brazil

LIDC	Lung Image Database Consortium
NCI	National Cancer Institute
NLST	National Lung Screening Trial
NoSQL	Not only Structured Query Language
NSCLC	Nonsmall cell lung cancer
PR	(DICOM) PResentation state
PT	Positron emission tomography
QIBA	Quantitative Imaging Biomarkers Alliance
RDBMS	Relational Database Management System
RIDER	Reference Image Database to Evaluate Therapy Response
SEG	(DICOM) SEGmentation
SR	(DICOM) Structured Report document
ROI	Region of interest
TCIA	The Cancer Imaging Archive
XaaS	Everything as a Service
XML	eXtensible Markup Language

Introduction

Lung cancer is the most common cause of cancer-related deaths, with a 5-year overall survival rate of only 15 % [1]. Detection and measurement of pulmonary nodules are important considering that a nodule may be a manifestation of a malignant cancer [2]. Lung cancer diagnosis is a challenging task for radiologists, because the nodules can be small and attached to other complex lung structures (Fig. 1) [3]. Moreover, medical image interpretation can be a challenge even for experienced specialists [4, 5]. An option to aid radiologists in the lung cancer diagnosis process is to integrate the computer-based assistance to imaging interpretation. The purpose of computer-aided dagnosis (CAD) is to improve the accuracy and consistency of medical image diagnosis through computational support used as reference [4]. Recently, detection and differential diagnosis CAD systems have been developed to reduce the expense and to improve the capability of radiologist in interpretation of medical images and differentiation between benign and malignant tissues [6].

However, a big problem of CAD research is its strong need for medical reference data repositories, particularly in the context of image-based diagnosis [7]. A large open (publicly available) reference imaging database would enable a better collaboration between physicians, radiologists, clinical researchers, and other healthcare workers [8]. Hence, a standardized data center would improve the development, testing and evaluation of CAD methods. An organized collection of anonymized clinical images alone would provide a valuable resource and eliminate database composition as a source of variability that hinders the appropriate comparison of different CAD methods [9]. The importance of employing a

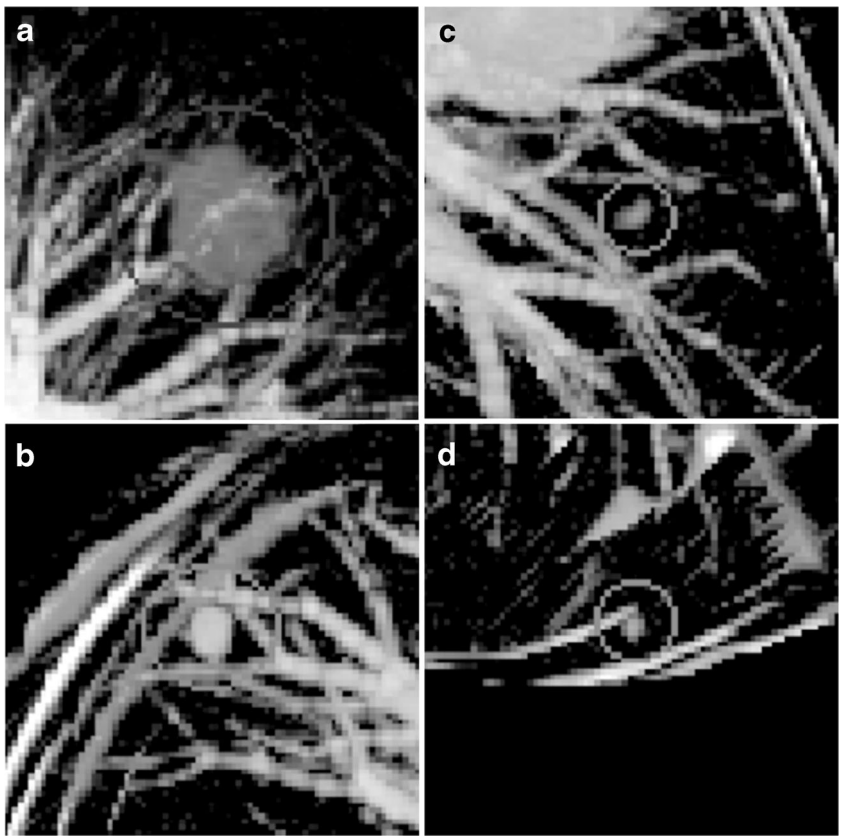

Fig. 1 Maximum intensity projection renderings of pulmonary nodules of different sizes [44]. **a**, **b** Juxtavascular and juxtapleural nodules, respectively. **c**, **d** Isolated nodules. **a** Vessel-connected with 20.1 mm. **b** Pleural-connected with 8.3 mm. **c** Isolated with 6.2 mm. **d** Isolated with 5.7 mm

standardized database is that it allows cross-validation by different CAD method implementations through the use of a single image resource to compare their results with the same testbed and hence avoid using different repositories that could lead to inaccurate output.

In order to assist lung cancer diagnosis and image-based CAD research, some public databases were created (Table 1). The National Lung Screening Trial (NLST) is a randomized multicenter study comparing low-dose helical computed tomography (CT) with chest radiography [10]. The NLST performed 73,118 studies in 26,254 patients and archived 21,082,502 images in a website. However, images were not annotated with lesion attributes, which restricts its contribution. The NSCLC-Radiomics and NSCLC-Radiomics-Genomics datasets consist of 422 and 89 nonsmall cell lung cancer (NSCLC) patients, respectively. For those patients, CT scans and clinical and survival data are available. For the NSCLC-Radiomics-Genomics collection, there are gene expression profiles to improve NSCLC characterization [11]. However, lesion delineations are not publicly available to download. The Reference Image Database to Evaluate Therapy Response (RIDER) is a targeted data collection used to generate an initial consensus on how to harmonize data collection and analysis for quantitative imaging methods applied to measure the response to drug or radiation therapy [12]. Other lung imaging projects employed phantom images, but they performed CT studies on a few number of subjects and used artificial images [13–15]. The National Cancer Institute (NCI) established a consortium of American institutions to develop a well-characterized repository of thoracic images [16]. This initiative resulted among other projects in the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [9]. In this work, we focused on the LIDC-IDRI project due to its high number of subjects, studies, and images, which is important for robust evaluation tests; its main medical imaging modality is CT, which is the most indicated exam to the early detection of some diseases, such as lung cancer, because it measures accurately size, location, and depth of tumors [17]; it does not use imaging phantom, so CT scans are not artificial; it has experienced radiologists' lesion marks, which can be used as a "gold standard" to nodule automatic detection and segmentation methods; it has radiologists' different characteristic ratings to each nodule, which is important to nodule classification methods; and based in our knowledge, it is the most important and used lung cancer image collection for CAD developers and researchers.

LIDC-IDRI project is a publicly available reference database for the medical imaging research community. The goal of LIDC-IDRI is to provide an essential medical imaging research resource to spur CAD development, validation, and dissemination in clinical practice [9]. LIDC-IDRI consists of CT scans with marked-up annotated lesions. Its image collection has associated radiologist annotations, including nodule outlines and subjective nodule characteristic ratings. It is now composed of 1010 patients, 1308 studies, and 244,527 images [18]. Despite its huge contribution to the lung cancer CAD research community, the LIDC-IDRI project has some limitations, which are listed as follows:

- *Database schemalessness*: LIDC-IDRI is a medical image collection; thus, it is not organized in a database schema. Hence, there is a decentralized correlation between images, studies, and nodule information (radiologists' marks and classifications). A centralized database schema would improve the manipulation and management of LIDC-IDRI data.
- *Absence of image descriptors*: LIDC-IDRI collection does not contain image descriptors. Several CAD methods use image descriptors extracted not only from single images, but preferably from volumetric images, to characterize lesions and diseases [5]. Therefore, pixel and voxel attribute extraction is important to computer-assisted diagnostics.

Table 1 Lung cancer and pulmonary nodule image collections

Collection	Anatomical location	Number of subjects	Number of studies	Number of images	Imaging modalities
NLST [10]	Chest	26,254	73,118	21,082,502	CT
NSCLC-Radiomics (Lung1 in [11])	Lung	422	422	51,195	CT
NSCLC-Radiomics-Genomics (Lung3 in [11])	Lung	89	89	13,482	CT
RIDER Lung PET-CT [12]	Lung	244	275	269,511	PT, CT
RIDER Lung CT [12]	Chest	32	46	15,419	CT
Phantom FDA [13]	Lung (Phantom)	4	46	872,521	CT
QIBA CT-1C [14]	Lung (Phantom)	1	3	69,258	CT, PR, SEG, SR
Lung Phantom [15]	Lung (Phantom)	1	1	237	CT
LIDC-IDRI [9]	Chest	1,010	1,308	244,527	CT, CR, DX

- *Local storage*: LIDC-IDRI data is located in a web server, and users have to download it via a graphical web interface in order to use it. Cloud storage for the LIDC-IDRI data would eliminate download and local storage processes, and all information could be accessed on-demand.

There are some technological solutions that can improve LIDC-IDRI's management (use and maintenance), functionality, and access. Currently, there is a need for integrated cloud solutions for the storage of massive heterogeneous data volume that has been created over the last years [8]. A big hospital produces several terabytes of medical data each year, bringing the total production of the European Union or the USA, for instance, to thousands of petabytes [19]. Moreover, data storage in cloud platforms is necessary because they provide data availability [20], which is important to guarantee that data is available and accessible whenever it is required. However, an important computational aspect must be taken into consideration on to the storage of massive data volume: database design flexibility. Big heterogeneous data represents a complex hierarchical data structure, which can make the use and maintenance of a database in a traditional data model difficult [21]. Not only Structured Query Language (NoSQL) or nonrelational databases were created toward reaching high performance in large data volume storage in the current "Big Data Age" [3, 22]. The term Big Data is referred as data that is too large, too dispersed, and too unstructured to be handled using conventional software [23], and by those, they mean Relational Database Management System (RDBMS). Given that NoSQL databases can represent heterogeneous data models without the complications imposed by relational databases [24], a NoSQL database is the most adequate approach for LIDC-IDRI's rich hierarchical data structure. Besides large data volume storage and full schema design flexibility, NoSQL databases also have the advantages of providing high performance, high throughput, and horizontal scalability [25].

Image analysis allow objective and precise quantitative imaging descriptors that could potentially distinct differences of tumors and may have diagnostic power and thus clinical significance across different diseases [11]. Therefore, several image features has been used to characterize pulmonary nodules, e.g., texture, shape, form, border, margin, and density [1, 3, 11, 26, 27]. In the medical domain, texture descriptors become particularly important as they can potentially reflect the fine details contained within an image structure [5]. Moreover, volumetric data should be represented because multidimensional images can increase representation accuracy [28]. Despite that, volumetric texture analysis is not an accurate method to characterize lung cancer, but it can be used as an initial step to filter, retrieve, and classify pulmonary nodules [3].

Our main goal in this paper is to minimize LIDC-IDRI's limitations by proposing an integrated nonrelational database schema to store its image collection. We also propose the development of a traditional volumetric texture analysis to characterize objectively and quantitatively LIDC-IDRI's pulmonary nodules. Finally, we aim at promoting the reproducible research by deploying the proposed integrated NoSQL database in a cloud infrastructure for on-demand database queries. As a result, all data (images, exam and nodule information, 3D texture attributes, and radiologist annotations) can be centralized in a single cloud datastore to the medical imaging informatics community, for the development, training, and evaluation of computer-aided lung cancer diagnosis and pulmonary nodule detection and classification research.

The remainder of this paper is organized as follows: the sections LIDC-IDRI Collection, Document-Oriented NoSQL Database, Cloud Database as a Service, and 3D Texture Analysis present the background for the Materials and Methods. The section Implementation describes database implementation details. The section Networks for Query Performance Tests presents the networks used on query performance tests. The sections Results and Discussion present the results and discussion of this work, and the section Conclusion concludes this paper.

Materials and Methods

LIDC-IDRI Collection

Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) project can be divided in three main processes: image interpretation, nodule characteristic assessment and data recording. Image interpretation process required a radiologist to review each scan of a CT study using a computer interface and outlines lesions considered to be a nodule with greatest in-plane dimension in the range 3–30 mm regardless of presumed histology. Consequently, such lesions could be a primary lung cancer, metastatic disease, a noncancerous process, or indeterminate in nature. Each nodule outline was meant to be a localizing "outer border" so that, in the opinion of the radiologist, the outline itself did not overlap pixels belonging to the nodule (Fig. 2) [9].

In the nodule characteristic assessment process, each reader was asked to subjectively set an integer value to nine different characteristics, according to the LIDC-IDRI documentation [29]. Nodule characteristics are as follows:

- *Subtlety*: nodule subtlety on a 1–5 scale, which 1 is extremely subtle and 5 is obvious.
- *Internal Structure*: internal structure on a 1–4 scale, which 1 is soft tissue, 2 is fluid, 3 is fat, and 4 is air.
- *Calcification*: internal calcification appearance on a 1–6 scale, which 1 is popcorn, 2 is laminated, 3 is solid, 4 is noncentral, 5 is central calcification, and 6 is absent.

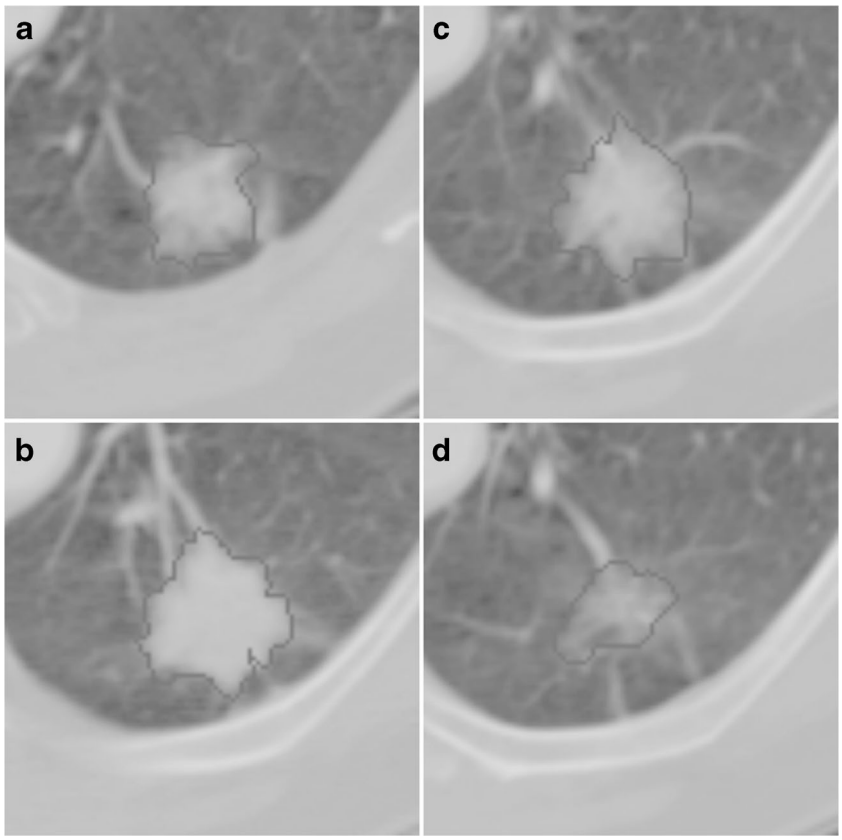

Fig. 2 Pulmonary nodule sample of a 4-slice volume, with LIDC-IDRI radiologist's marks. **a** Cropped CT scan 1 of the nodule. **b** Cropped CT scan 2 of the nodule. **c** Cropped CT scan 3 of the nodule. **d** Cropped CT scan 4 of the nodule

- *Sphericity*: shape appearance in terms of its roundness/sphericity with three values (1, 3, or 5), which 1 is linear, 3 is ovoid, and 5 is round.
- *Margin*: nodule margin with two values (1 or 5), which 1 is poorly defined and 5 has sharp margin.
- *Lobulation*: nodule lobulation with two values (1 or 5), which 1 has no lobulation and 5 has marked lobulation.
- *Spiculation*: nodule spiculation with two values (1 or 5), which 1 has no spiculation and 5 has marked spiculation.
- *Texture*: nodule internal texture with 3 values (1, 3, or 5), which 1 is nonsolid or ground glass opacity, 3 is part solid or mixed, and 5 is solid texture;
- *Likelihood of malignancy*: nodule likelihood of malignancy (assuming 60-year-old male smoker) on a 1–5 scale, which 1 is highly unlikely for cancer, 2 is moderately unlikely for cancer, 3 is indeterminate likelihood, 4 is moderately suspicious for cancer, and 5 is highly suspicious for cancer.

For last, in the data recording process, CT study information, nodule classifications, and outline Cartesian coordinates were recorded in a eXtensible Markup Language (XML) file. This XML file describes all data regarding to a single study, including all nodules that the CT scans may have and the regions of interest (ROI) with the unique identifier for each image slice. XML file and all CT scans from a single exam were stored in a folder, and all folders from all exams were uploaded to a web server located at The Cancer Imaging Archive (TCIA) website [30]. Anyone can access TCIA website and download all LIDC-IDRI data by clicking on Search Images and then selecting the LIDC-IDRI collection.

Document-Oriented NoSQL Database

NoSQL (or simply nonrelational) is a term for all databases that do not follow the popular and well-established RDBMS principles. It represents a class of products and a collection of diverse, and sometimes related, concepts about data storage and manipulation [22]. NoSQL databases have been around since the late 1960s, but it became popular with the diffusion of the Big Data concept [31]. The term "Big Data" refers to large, diverse, complex, longitudinal, or distributed datasets that are too difficult to be managed by the RDBMS [32]. Relational databases work best with structured data, which readily fits in well-organized tables [31] and that is not the case of LIDC-IDRI unstructured data.

Nonrelational databases have a wide variety of storage approaches, e.g., key-value storage, column-oriented (opposite approach to the row-oriented RDBMS), document-oriented, graph database, and object-oriented database. For the purposes of this work, we focus only on the document-oriented approach due its consolidation as the most robust NoSQL

approach [25], and for having a more suitable data structure to the storage of LIDC-IDRI heterogeneous data.

A document-oriented database is composed by a set of collections. Each collection has a set of documents and each document is composed by key-value pairs, lists, or embedded documents [22]. JavaScript Object Notation (JSON) is one of the most used document implementations because it is simple, intuitive, and human-readable (Fig. 3) [24].

One of the most used Database Management Systems (DBMS) for document-oriented datastores is the open source MongoDB. Its storage unit is JSON document, but MongoDB represents JSON documents in binary-encoded format called Binary JavaScript Object Notation (BSON), for efficiency and performance purposes [22]. MongoDB is also capable of generating a globally unique identifier for objects called ObjectID, a 12-byte BSON primary key that guarantees document uniqueness and integrity in a database [33].

GridFS is a specification that enables MongoDB to store and retrieve arbitrary files in binary format. GridFS uses two collections for the file storage: files and chunks. The file collection stores metadata (e.g., length key stores the size of the file in bytes, and uploadDate key stores the date the file was first stored). GridFS divides the file into chunks (blocks, fragments of information) and stores each of those as a separate document in the chunks collection (data key stores the binary data of each chunk document) (Fig. 4) [33].

Cloud Database as a Service

Cloud computing, or on-demand computing, was created in order to shift the location of an infrastructure to the network to reduce the costs associated with the management of hardware and software resources [34]. Cloud is a large pool of easily usable and accessible virtualized resources, such as services [20]. There are different categories of cloud services that are delivered and consumed in real-time over the Internet, which can summarized in a Everything as a Service (XaaS) model, where X is software, platform, infrastructure, hardware, data, etc. (Fig. 5) [35].

```
{
    title : "Journal of Digital Imaging",
    topics : ["CAD","DICOM","CBIR"],
    publisher : {
        name : "Springer",
        country : "US"
    }
}
```

Fig. 3 A JSON document example of a fictitious *journals* collection. Figure shows a document with one key-value pair (*title*), one list (*topics*), one embedded document (*publisher*, which has two key-value pairs—*name* and *country*), and their respective values

Database as a Service (DBaaS) is a paradigm for data management in which a third-party service provider hosts "database as a service" providing its customers seamless mechanisms to create, store, and access their databases at the cloud [36]. By using the DBaaS model, users access data using the hardware and software provided by a cloud provider instead of their own computing infrastructure. Furthermore, the responsibility of database management (backup, administration, restoration, etc.) befall the provider. Therefore, DBaaS paradigm reduces user's computational problems, e.g., database management, limited local storage space, and data availability.

a

```
{
  "_id" : <ObjectId>,
  "length" : <num>,
  "chunkSize" : <num>
  "uploadDate" : <timestamp>
  "md5" : <hash>

  "filename" : <string>,
  "contentType" : <string>,
  "aliases" : <string array>,
  "metadata" : <dataObject>,
}
```

b

```
{
  "_id" : <ObjectId>,
  "files_id" : <ObjectId>,
  "n" : <num>,
  "data" : <binary>
}
```

Fig. 4 GridFS data model for the *files* and *chunks* collections [33]. **a** Data model of the *files* collection. Documents have nine key-value pairs that describe file metadata. **b** Data model of the *chunks* collection. Documents have four key-value pairs that describe file chunks data

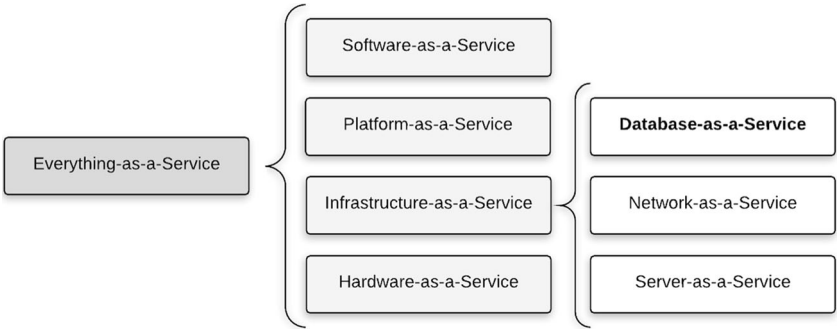

Fig. 5 Cloud computing services

3D Texture Analysis

Gray-level co-occurrence matrix (GLCM) is a technique to extract information from second-order texture and its use to classify images is already well established [37]. Despite the fact that the GLCM-based texture analysis may be dated as method, it is still relevant to scientific literature and recent works have used texture features as image descriptors, using bidimensional or tridimensional analysis [38–40]. Texture analysis over a pulmonary nodule slice was performed in a previous work [41]. Its mean precision was 15 %, using a content-based image retrieval (CBIR) system, for the first ten similar nodules. However, 3D texture analysis improved pulmonary nodule retrieval with a mean precision of 73 %, for the first ten similar nodules, as attested in [3]. Therefore, a volumetric texture analysis can improve nodule characterization.

The GLCM method obtains from a single image the occurrence probability of a pixel pair with intensity *i,j* and spacing between the pixels of *x* and *y* in the dimensions *x* and *y*, respectively, given a distance *d* and orientation [37]. Calculation of the GLCM in a volume of images extends the evaluation of the probability function to the rectangular *Z*-axis, in order to study between-slices joint probabilities on an image volume composed of multiple slices (Fig. 6) [42]. Second-order histogram statistics are applied to the GLCM producing the texture attributes. Texture attributes used in this work were suggested by Haralick et al. [43] and are listed as follows:

$$\text{Energy} = \sum_{i,j} C(i,j)^2, \tag{1}$$

$$\text{Entropy} = -\sum_{i,j} C(i,j)\,\log C(i,j), \tag{2}$$

$$\text{Inverse difference moment} = \sum_{i,j} \frac{1}{1+(i\ \ j)^2} C(i,j), \tag{3}$$

$$\text{Shade} = \sum_{i,j} \left(i+j\ \ \mu_x\ \ \mu_y\right)^3 C(i,j), \tag{4}$$

$$\text{Inertia} = \sum_{i,j} (i\ \ j)^2 C(i,j), \tag{5}$$

$$\text{Promenance} = \sum_{i,j} \left(i+j\ \ \mu_x\ \ \mu_y\right)^4 C(i,j), \tag{6}$$

$$\text{Correlation} = \sum_{i,j} \frac{(i\ \ \mu_x)\left(j\ \ \mu_y\right)}{\sqrt{\sigma_x \sigma_y}} C(i,j), \tag{7}$$

$$\text{Variance} = \sum_{i,j} (i\ \ \mu)^2 C(i,j), \tag{8}$$

$$\text{Homogeneity} = \sum_{i,j} \frac{C(i,j)}{(1+|i\ \ j|)}, \tag{9}$$

where $C(i, j)$ are the elements from the GLCM, μx and μy are the mean, and σx and σy are the standard deviation, obtained by the following equations:

$$\mu_x = \sum_i i C_x(i), \tag{10}$$

$$\mu_y = \sum_j j C_y(j), \tag{11}$$

$$\sigma_x = \sum_i (i-\mu_x)^2 \cdot \sum_j C(i,j), \tag{12}$$

$$\sigma_y = \sum_j \left(j-\mu_y\right)^2 \cdot \sum_i C(i,j), \tag{13}$$

$$C_x(i) = \sum_j C(i,j), \tag{14}$$

$$C_y(j) = \sum_i C(i,j), \tag{15}$$

Thus, a characteristic vector can be obtained by means of the calculation of the nine attributes (Eqs. 1–9) applied to the co-occurrence matrices in orientations 0°, 45°, 90°, and 135°, for instance. In this case, an image volume can be associated with a 36-dimension texture feature vector.

Implementation

In this work, we aim at integrating all data from the LIDC-IDRI collection plus texture nodule characterization in a

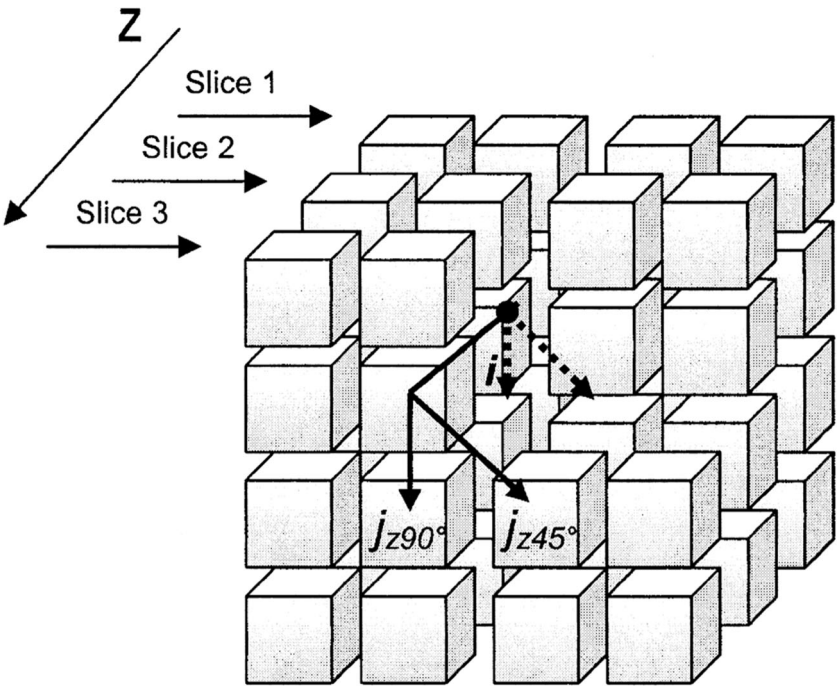

Fig. 6 GLCM calculation over a three-slice image volume [42]. Between-slices joint relationships have 1 pixel and slice distances in 45° and 90°

single cloud-based database. Therefore, a document-oriented approach was used in order to create the centralized nonrelational database schema for the LIDC-IDRI data (LIDC-IDRI Collection and Document-Oriented NoSQL Database sections). Moreover, we deployed this proposed database in a cloud DBaaS infrastructure for on-demand database queries, hence eliminating the obligation to download and to store in a local workstation (Cloud Database as a Service section).

MongoDB was the DBMS used to store the proposed database, due to its high parallel processing power, high performance on data retrieval [24], and for having the GridFS specification, necessary to the image storage. The DBaaS platform used in this work was Morpheus,[1] a cloud database as service that enables users to provision, deploy, and host MongoDB databases. Morpheus was configured on a MongoDB Instance v2.6.3, 5GB of storage, 1 shard and 3 boxes with 1 GB RAM per box.

In order to extract image descriptors and characterize lung lesions, 3D texture analysis was applied to manually segmented pulmonary nodules. All nodules had lesion images segmented using the radiologist's marks (Fig. 7). After the manual segmentation, texture attributes were extracted from the voxels using 3D GLCM (3D Texture Analysis section). Distance between the pixel pairs and between volume slices was 1 and the angle orientations were 0°, 45°, 90°, and 135°. Texture attributes used in this work were energy, entropy, inertia, homogeneity, correlation, shade, promenance, variance, and inverse difference moment. Those nine descriptors and four angle orientations resulted in the creation of 36 3D texture attributes (3DTA) for each pulmonary nodule.

An algorithm was implemented in Java program- ming language to convert the LIDC-IDRI collection to a document-oriented model, to insert the 3DTA to the document data structure and to migrate the database to the cloud service. The implemented algorithm had the following sequential steps:

1. *Image storage*: CT scans that contain nodule(s) and segmented nodule images were stored in a MongoDB database with the GridFS specification. Thus, images were converted to document BSON format, an ObjectID was generated for each image saved, and each image is uniquely identified in the database.
2. *XML file data conversion*: first, XML tags data were extracted; then, exams and nodules data were separated, for information optimization purposes, and converted to JSON documents. For last, exams and nodule documents were inserted in the MongoDB database.
3. *Slice-image mapping*: each slice had its CT scan unique identifier modified to the ObjectID of its image stored in the previous database. Segmented nodule image's ObjectID was also embedded in the document. Therefore, each slice has two ObjectIDs regarding to the original CT scan and to the segmented nodule image;
4. *3D texture attribute embedding*: each 3DTA was converted to a key-value pair, and the 36 3DTA key-value pairs were embedded in a document. Finally, each 3DTA document was inserted in its nodule document.

[1] Available at www.morpheusdata.com/ [Online; accessed on June 14, 2016]

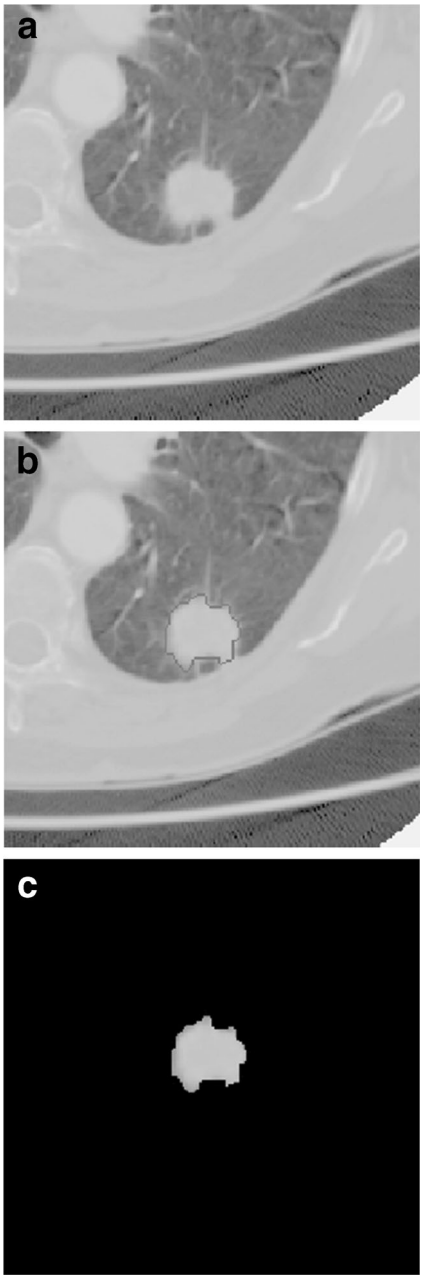

Fig. 7 Manual segmentation process of a pulmonary nodule from the LIDC-IDRI. **a** CT scan with a pulmonary nodule. **b** Pulmonary nodule that was marked by a LIDC-IDRI's radiologist. **c** Manual segmentation output of a pulmonary nodule

5. *Database migration*: the entire database was migrated to the Morpheus DBaaS platform in order to enable cloud storage and remote access.

Networks for Query Performance Tests

Performance tests on different data queries were executed in order to measure the time that is necessary to access documents on the developed cloud-based database. Three networks on uncontrolled environments were used in order to create different real scenarios, where a user could search document(s) on the nonrelational database using a heterogeneous network. Latency, download, and upload speed tests for all networks were calculated by the Speedtest tool[2] on a server in San Francisco, CA, and are listed in Table 2.

Network A presents the best mean speed values, while *Network C* presents the worst. *Network A* reaches about 3.8× more download speed and 29.6× more upload speed than *Network B*. In comparison to *Network C*, *Network A* reaches about 22.5× more download speed and 47.4x more upload speed. *Network B* reaches about 5.8× more download speed and 1.6× more upload speed than *Network C*.

Results

The developed database has four main document collections (Fig. 8): *exams*, *nodules*, *images.files*, and *images.chunks*. The *exams* collection stores CT studies and reading session data. The *nodules* collection stores nodules information, including radiologist's classification, 3D texture attributes, and ROI's data (image identifiers, lesion marks, etc.). The *images.chunks* and *images.files* are GridFS collections that store image binaries and image metadata. We model a one-to-many relationship from *exams* to *nodules* because an exam may have several nodules and a nodule belongs to only one exam. Furthermore, we model a many-to-many relationship from *nodules* to *images.files* because a nodule may have several images and an image may have several nodules. GridFS sets automatically a one-to-many relationship from *images.files* to *images.chunks* because an image file is divided into several binary chunks (Document-Oriented NoSQL Database section).

A document from the *exams* collection has an ObjectID which is stored on the *_id* key. It also has a *path* key which stores the absolute path of the exam's folder. The rest of the key-value pairs correspond to the response header, reading session, and LIDC-IDRI data.

A document from the *nodules* collection has its own ObjectID stored on the *_id* key and the ObjectID of its relative exam stored on the *examID* key. Each radiologist characteristic rating is stored on a key-value pair (calcification on the *calcification* key, for instance). It also has each 3DTA on a key-value pair on the *textureAttributes* embedded document, e.g., *energy0* key stores energy at 0° orientation attribute. Documents from the *nodules* collection have a list of regions of interest stored on the *rois* key. Each document from the *rois* list has an *originalImage* key that stores the ObjectID of its CT scan on the *images.files* collection, and a *noduleImage* key that stores the ObjectID of its segmented nodule image on

[2] Available at www.speedtest.net [Online; accessed on June 14, 2016]

Table 2 Networks speed test results

	Download speed rate (Mbps)	Upload speed rate (Mbps)	Latency rate (Mbps)
Network A	48.01 ± 5.53	29.90 ± 4.62	156.8 ± 0.60
Network B	12.50 ± 1.34	1.01 ± 0.03	167.0 ± 4.09
Network C	2.13 ± 0.47	0.63 ± 0.07	180.7 ± 2.75

All rate values are the mean of ten executions of the speedtest tool

the *images.files* GridFS collection. Finally, the *edgeMaps* list on each document from the *rois* list stores all Cartesian coordinates (*xCoord* and *yCoord* keys) of the radiologist's mark.

For the performance evaluation, seven functions were executed on those previous networks: database connection; query for one nodule in the *nodules* collection; query for one exam in the *exams* collection; query for one image in the GridFS collection; query for all nodules in the *nodules* collection; query for all exams in the *exams* collection; and query for all images in the GridFS collection. Table 3 describes the performance tests results of those functions.

Database connection presented to be the fastest and most stable function on all three networks. All cloud search queries for single documents (nodule, exam, and image) took less than 2 s considering the uncontrolled network environments. The highest time value was reached by *Network C*, the slowest network, on nodule document query, and it took 1.67 s to search for a nodule. Queries that retrieve the entire *exams* collection took less than 3 s to finish. The highest mean value for all exams query was 2.84 s, reached by *Network C*. Furthermore, queries that search for all nodules or images require much more time than exam queries. For instance, *Network A* retrieved the entire nodules collection from the database in at least 7.72 s, while *Network B* and *Network C* needed 4.95× and 6.11× more time to search for all nodules, respectively. In order to retrieve all images from the GridFS collection, *Network A* needed at least 2.06 s×, while *Network B* and *Network C* needed 2.03× and 5.34× more time, respectively.

Current database status is 379 exams, 838 nodules, and 8237 images, 4029 of them are CT scans and 4208 manually segmented nodules. CT sections used in this work can have more than one lesion. Data reading can be done via MongoDB Shell, Application Programming Interface (API), or any MongoDB database management tool (Fig. 9). Database connection settings are as follows: *readonly* is the user name with read-only privileges, *gH@h6NL38V* is the user's password, *162.252.108.127* is the Internet Protocol (IP), *12279* is the port and *publicDB* is the database name (Fig. 10).

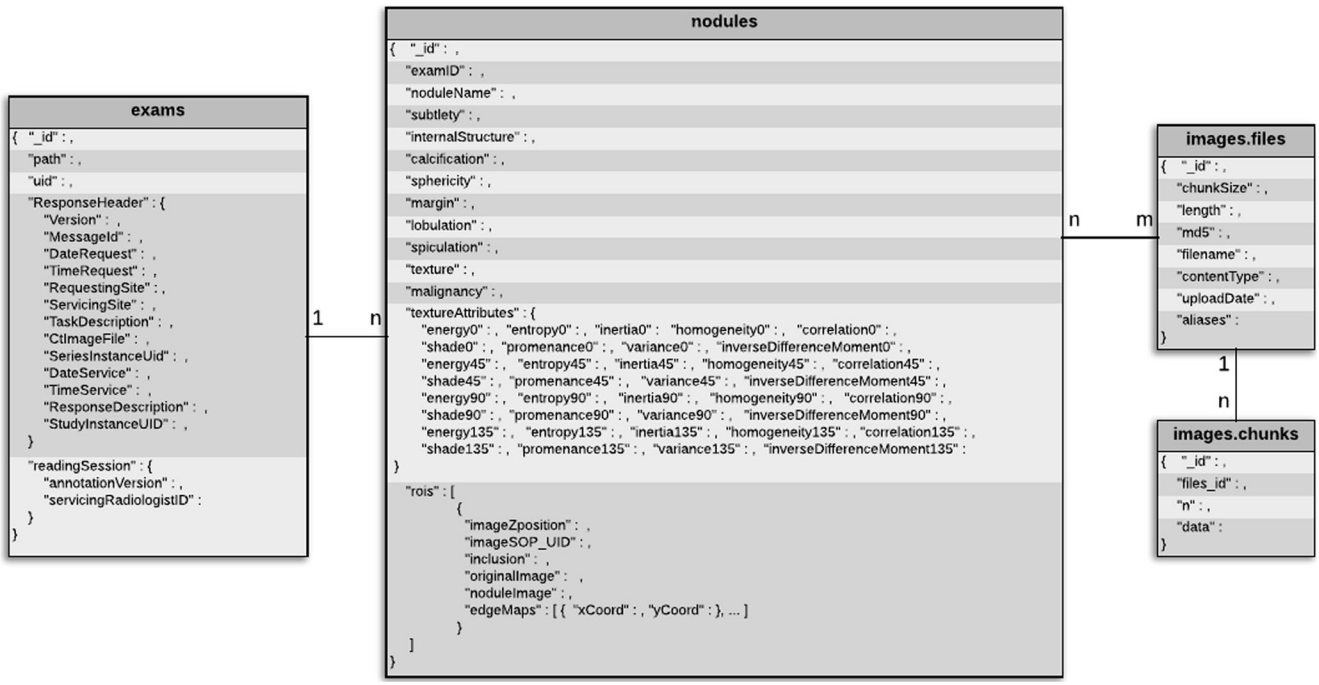

Fig. 8 Data model of the developed document-oriented database of pulmonary nodules. Documents from the *exams* collection have three key-value pairs and two embedded documents (*RespondeHeader* with thirteen key-value pairs and *readingSession* with two key-value pairs). Documents from the *nodules* collection have twelve key-value pairs, one embedded document (*textureAttributes*, which has 36 key-value pairs) and one list of documents (*rois*, which each document has five key-value pairs and one list of documents—*edgeMaps* with two key-value pairs each). Documents from the *images.files* collection have eight key-value pairs, and documents from the *images.chunks*, four key-value pairs

Table 3 Database queries performance tests results

	Connection Time (s)	One nodule Query time (s)	One exam Query time (s)	One image Query time (s)	All nodules Query time (s)	All exams Query time (s)	All images Query time (s)
Network A	0.11 ± 0.01	0.93 ± 0.16	0.16 ± 0.01	0.49 ± 0.01	8.68 ± 0.75	0.57 ± 0.23	2.61 ± 0.36
Network B	0.17 ± 0.01	0.96 ± 0.14	0.17 ± 0.01	0.51 ± 0.01	14.76 ± 2.54	0.81 ± 0.08	6.02 ± 1.07
Network C	0.16 ± 0.02	1.13 ± 0.23	0.24 ± 0.12	0.59 ± 0.04	59.64 ± 8.27	2.84 ± 0.58	14.97 ± 2.08

All time values are the mean of ten executions for each function

Discussion

This work highlighted the importance of LIDC-IDRI project to the medical imaging science and its potential contribution to computer-aided lung cancer diagnosis and pulmonary nodule CAD research. However, the LIDC-IDRI image collection has some computational issues that limit its management, use, manipulation, maintenance, functionality, and access. Nevertheless, our proposed implementation minimized those problems using a cloud-based NoSQL database approach.

Regarding to the proposed nonrelational schema, the JSON document conversion and the storage of exams and images on MongoDB allowed a better management of LIDC-IDRI data which was initially in the XML file. After our implementation, nodule and exam information (including extracted texture features), radiologist's marks and classifications, CT scans, and segmented nodule images were stored in a single MongoDB database, which facilitated data and image manipulation due to information centralization.

This work also contributed to the expansion of LIDC-IDRI collection due to two factors. Our first contribution is the application of 3D texture analysis, which allowed an initial nodule volumetric characterization. The texture feature vector can be used in differential diagnosis CAD methods to classify pulmonary nodules in terms of potential malignancy for instance. It also can be used in detection CAD methods to reduce the number of false-positive findings. Due to the fact that volumetric texture analysis is not the most accurate nodule descriptor, we encourage other researchers to add their implementation and image attributes to the implemented database. Our second contribution is the nodule image segmentation process (Fig. 7), which optimized nodule visualization for

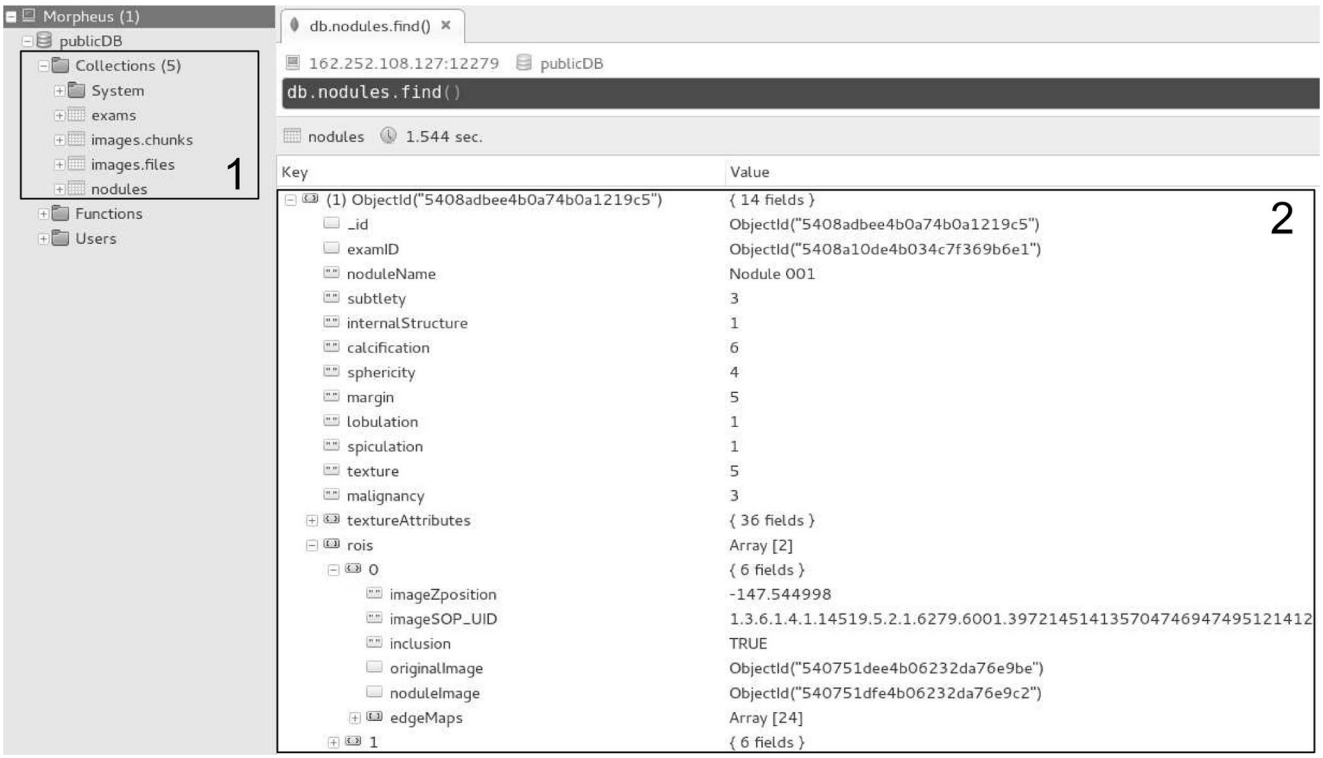

Fig. 9 Data visualization performed by a MongoDB database management software. Component 1 illustrates all collections from the database. Component 2 illustrates a document from the *nodules* collection

Fig. 10 Database settings illustrated by a MongoDB database management software. **a** Connection settings. **b** Authentication settings

inexperienced researchers. Moreover, in the native LIDC-IDRI implementation, developers had to search for each point of the lesion boundary in the XML file to manually segment the nodule. As result of the image segmentation process, nodules are set to be used in the comparison of automatic segmentation algorithms and in feature extraction methods.

Another important aspect should be taken into consideration in the development of the open cloud-based health database: patient information security. However, an anonymization process was previously performed by the LIDC-IDRI, in order to remove all protected health information contained within the Digital Imaging and Communications in Medicine (DICOM) headers of the images, in accordance with Health Insurance Portability and Accountability Act (HIPAA) guidelines [9]. Therefore, the proposed database did not need to focus on data security, because LIDC-IDRI is on public domain.

Cloud storage allowed remote and on-demand data access, which was not possible in the native LIDC-IDRI web server storage. According to the native LIDC-IDRI implementation, users have to download the entire exam folder to read a single CT scan, for instance. With our implementation, it is the user's decision whether he wants to access a single exam, nodule or image document, or an entire collection. However, performance tests on uncontrolled network environments indicated that cloud databases are not the most appropriate storage approach for slow networks. For such cases, in order to solve this problem, users can query the entire collection with a single connection to the cloud database (this query performs a temporary data download) and then have to save it to a local database. Hence, it would eliminate bad network connections on the cloud and reduce query time. Nevertheless, more performance tests, especially on controlled network environments, should be performed toward increasing precision and robustness of query performance evaluation.

To our knowledge, there is no work in literature that presents a cloud-based NoSQL database schema for the LIDC-IDRI data collection. With our implementation, database is allocated in the cloud, which enhance data remote access and users do not need to download the entire database for single queries. Moreover, our implementation allows a better database maintenance due to the schema flexibility provided by the document-oriented approach.

Database publicly distribution allows that the same reference database can be used by several researchers, serving as testbed by different research projects. Database effectiveness has been proved on algorithms developed by different CAD applications, e.g., similar nodule retrieval, identification of relevant image descriptors, and lesions (semi-)automatic detection.

One limitation of this work is that the database is available only for intermediary users—system developers and researchers with programming skills, for instance. A graphical user interface (GUI) is been developed in order to final users and beginner researchers can have access to nodule information, CT scans and segmented nodule images easier.

Moreover, the implemented database has read-only data access to the public user for data integrity purposes. Read-and-write privileges could not assure the accuracy and consistency of data over a database. Therefore, in order to alter documents or collections, by adding other image descriptors for instance, users need to retrieve the entire database (data retrieval performs a temporary data download), save in a local workstation, make all necessary changes and store the updated database on a cloud DBaaS infrastructure.

Another problem is the imaging interpretation issue. Medical image interpretation is a recognized challenge in radiology, and it has been studied for several years. Significant

interobserver variation has been documented in numerous studies, and it happened due to various aspects, e.g., time constraints, readers' perceptual errors, lack of training, or fatigue [5]. Therefore, nodule detection and manual segmentation may be subject to interobserver and intraobserver variability.

The last limitation is that the developed database could not store all LIDC-IDRI data, due to the lack of free public cloud infrastructures that deliver large storage space. A private cloud infrastructure has been developed to host cloud projects, like the proposed in this work. Furthermore, continuous extension of the database is envisaged to take place in an open productive research community, either by completing the remainder of LIDC-IDRI data or adding more pulmonary nodule CT studies with associated radiologist annotations, nodule markups, and feature ratings.

Conclusion

This paper presented the development of an open cloud-based nonrelational document-oriented database of pulmonary nodules characterized by a volumetric textural analysis. All nodule cases are CT studies provided by the public LIDC-IDRI initiative, which has marked-up annotated lesions, including nodule outlines and subjective nodule characteristic ratings. The proposed database was deployed in a cloud DBaaS infrastructure in order to improve the collaboration of lung cancer CAD researchers.

The developed database has high potential to be integrated to several different CAD systems. For instance, a CBIR software can provide CAD support by allowing radiologists to find images from a database, like the one this paper proposes, that are similar to the images they are interpreting. This reference database can also be used in image-based diagnosis teaching at medical schools. Radiology residents could practice nodule detection and identification skills by comparing their own marks to the LIDC-IDRI's radiologist marks.

This nodule database also has high potential to be applied to the Big Data context, due to the fact that NoSQL technology has low coupling, by integrating it to the electronic patient record and other health databases. Therefore, with those integrations facilitated by the use of MongoDB, it will be possible to infer new information that can aid early diagnosis of lung cancer.

Acknowledgments We thank the Brazilian institutions *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) and *Fundação de Amparo à Pesquisa do Estado de Alagoas* (FAPEAL) for the financial support in the form of a master scholarship (grant number 20130603-002-0040-0063).

References

1. Wu H, Sun T, Wang J, Li X, Wang W, Huo D, Lv P, He W, Wang K, Guo X: Combination of Radiological and Gray Level Co-occurrence Matrix Textural Features Used to Distinguish Solitary Pulmonary Nodules by Computed Tomography. J Digit Imaging 26(4):797–802, 2013
2. Reeves A, Chan A, Yankelevitz D, Henschke C, Kressler B, Kostis W: On Measuring the Change in Size of Pulmonary Nodules. IEEE Trans Med Imaging 25(4):435–450, 2006
3. Oliveira M, Ferreira J: A Bag-of-Tasks Approach to Speed Up the Lung Nodules Retrieval in the BigData age. E-Health Networking, Application & Services, DOI: 10.1109/HealthCom.2013.6720753, October 12, 2013.
4. Doi K: Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential. Comput Med Imaging and Graph 31(4–5):198–211, 2007
5. Akgul C, Rubin D, Napel S, Beaulieu C, Greenspan H, Acar B: Content-Based Image Retrieval in Radiology: Current Status and Future Directions. J Digit Imaging 24(2):208–222, 2011
6. Jalalian A, Mashohor S, Mahmud H, Saripan M, Ramli A, Karasfi B: Computer-Aided Detection/Diagnosis of Breast Cancer in Mammography and Ultrasound: a review. Clin Imaging 37(3): 420–426, 2013
7. Deserno T, Welter P, Horsch A: Towards a Repository for Standardized Medical Image and Signal Case Data Annotated with Ground Truth. J Digit Imaging 25(2):213–226, 2012
8. Tsymbal A, Meissner E, Kelm M, Kramer M: Towards Cloud-Based Image-Integrated Similarity Search in Big Data. Biomedical and Health Informatics, DOI: 10.1109/BHI.2014.6864434, June 4, 2014.
9. Armato S, McLennan G, Bidaut L, McNitt-Gray M, Meyer C, Reeves A, Zhao B, Aberle D, Henschke C, Hoffman E, et al: The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. Med Phys 38:915, 2011
10. Aberle D, Berg C, Black W, Church T, Fagerstrom R, Galen B, Gareen I, Gatsonis C, Goldin J, Gohagan J, et al: The National Lung Screening Trial: overview and study design. Radiology 258(1):243–253, 2011
11. Aerts H, Velazquez E, Leijenaar R, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D, et al.: Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach. Nature Communications, 5, 2014.
12. The Cancer Imaging Archive (TCIA). RIDER Collections. Available at http://wiki.cancerimagingarchive.net/display/Public/RIDER+Collections Accessed 23 February 2015.
13. Gavrielides M, Kinnard L, Myers K, Peregoy J, Pritchard W, Zeng R, Esparza J, Karanian J, Petrick N: A Resource for the Assessment of Lung Nodule Size Estimation Methods: database of thoracic CT scans of an anthropomorphic phantom. Optics Express 18(14): 15244–15255, 2010
14. Das M, Ley-Zaporozhan J, Gietema H, Czech A, Muhlenbruch G, Mahnken A, Katoh M, Bakai A, Salganicoff M, Diederich S, et al: Accuracy of Automated Volumetry of Pulmonary Nodules Across Different Multislice CT Scanners. Eur Radiol 17(8):1979–1984, 2007
15. The Cancer Imaging Archive (TCIA). Lung Phantom Image Collection. Available at http://wiki.cancerimagingarchive.net/display/Public/Lung+Phantom Accessed 23 February 2015.
16. Armato S, Roberts R, McNitt-Gray M, Meyer C, Reeves A, McLennan G, Engelmann R, Bland P, Aberle D, Kazerooni E, et al: The Lung Image Database Consortium (LIDC): Ensuring

the integrity of expert-defined "truth". Acad Radiol 14(12):1455–1463, 2007
17. Sluimer I, Schilham A, Prokop M, Ginneken B: Computer Analysis of Computed Tomography Scans of the Lung: a survey. IEEE Trans Med Imaging 25(4):385–405, 2006
18. Lung Image Database Consortium and Image Database Resource Initiative. The Cancer Imaging Archive. Available at http://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI Accessed 02 February 2015.
19. Montagnat J, Breton V, Magnin I: Using Grid Technologies to Face Medical Image Analysis Challenges. Biomedical Computations on the Grid, DOI: 10.1109/ccgrid.2003.1199418, May, 2003.
20. Vaquero L, Rodero-Merino L, Caceres J, Lindner M: A Break in the Clouds: Towards a Cloud Definition. ACM SIGCOMM Computer Communication Review 39(1):50–55, 2008
21. Wei-ping Z, Ming-Xin L, Huan C: Using MongoDB to Implement Textbook Management System Instead of MySQL. Communication Software and Network, DOI: 10.1109/iccsn.2011.6013720, May 29, 2011.
22. Tiwari S: Professional NoSQL. John Wiley and Sons, Inc., 2011.
23. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers A: Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, pages 1–137, 2011.
24. Banker K: MongoDB in Action. Manning Publications Co., 2011.
25. Strauch C, Sites U, Kriha W: NoSQL Databases. Stuttgart Media University, 2011.
26. Choi W, Choi T: Automated Pulmonary Nodule Detection Based on Three-Dimensional Shape-Based Feature Descriptor. Comput Methods Programs Biomed 113(1):37–54, 2014
27. Erasmus J, Connolly J, McAdams H, Roggli V: Solitary Pulmonary Nodules: Part I. Morphologic Evaluation for Differentiation of Benign and Malignant Lesions 1. Radiographics, 20(1):43–58, 2000.
28. Kumar A, Kim J, Cai W, Fulham M, Feng D: Content-Based Medical Image Retrieval: A Survey of Applications to Multidimensional and Multimodality Data. J Digit Imaging 26(6):1025–1039, 2013
29. Lung Image Database Consortium and Image Database Resource Initiative. LIDC-IDRI Documentation: Anno-tated XML File. Available at http://wiki.cancerimagingarchive.net/download/attachments/3539039/annotated xml file Mar% 202010.rtf?version = 1&modificationDate = 1319224566198&api = v2 Accessed 02 February 2015.
30. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, et al: The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging 26(6):1045–1057, 2013
31. Leavitt N: Will NoSQL Databases Live Up to Their Promise? Computer 43(2):12–14, 2010
32. Liu L: Computing Infrastructure for Big Data Processing. Frontiers of Computer Science 7(2):165–170, 2013
33. MongoDB Inc. MongoDB Manual. Available at http://docs.mongodb.org/manual Accessed 02 February 2015.
34. Hayes B: Cloud Computing. Communications of the ACM, 51(7), 2008.
35. Rimal B, Choi E, Lumb I: A Taxonomy and Survey of Cloud Computing Systems. INC, IMS and IDC, DOI: 10.1109/NCM.2009.218, August 27, 2009.
36. Hacigumus H, Iyer B, Mehrotra S: Providing Database as a Service. Data Engineering, DOI: 10.1109/ICDE.2002.994695, March 1, 2002.
37. Oliveira M, Cirne W, Marques P: Towards Applying Content-Based Image Retrieval in the Clinical Routine. Future Generation Computer Systems 23(3):466–474, 2007
38. Dhara A, Mukhopadhyay S, Dutta A, Garg M, Khandelwal N: A Combination of Shape and Texture Features for Classification of Pulmonary Nodules in Lung CT Images. J Digit Imaging, 1–10, 2016.
39. Han F, Wang H, Zhang G, Han H, Song B, Li L, Moore W, Lu H, Zhao H, Liang Z: Texture feature analysis for computer-aided diagnosis on pulmonary nodules. J Digit Imaging 28(1):99–115, 2015
40. Kaya A, Can A: A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics. J Biomed Inform 56:69–79, 2015
41. Lam M, Disney T, Raicu D, Furst J, Channin D: BRISC - An Open Source Pulmonary Nodule Image Retrieval Framework. J Digit Imaging 20(1):63–71, 2007
42. Ghoneim D, Toussaint G, Constans J, Certaines J: Three Dimensional Texture Analysis in MRI: A Preliminary Evaluation in Gliomas. Magn Reson Imaging 21(9):983–987, 2003
43. Haralick R, Shanmugam K, Dinstein I: Textural Features for Image Classification. IEEE Transactions on Systems, Man and Cybernetics, (6):610–621, 1973.
44. Mehdi A, Vassili K, Eduard S, Vahid T: A Comprehensive Framework for Automatic Detection of Pulmonary Nodules in Lung CT Images. Image Analysis & Stereology 33(1):13–27, 2014