# BANCO DE DADOS *NoSQL* PÚBLICO DE NÓDULOS PULMONARES PARA AUXÍLIO À PESQUISA E DIAGNÓSTICO DO CÂNCER DE PULMÃO

José Raniery Ferreira Junior, Marcelo Costa Oliveira

Instituto de Computação (IC), Universidade Federal de Alagoas (UFAL)
Laboratório de Telemedicina e Informática Médica (LaTIM)
Hospital Universitário Prof. Alberto Antunes (HUPAA)
e-mail: jrfj@ic.ufal.br

**Resumo:** Devido à dificuldade no diagnóstico do câncer de pulmão, é necessária a integração de ferramentas computacionais ao processo de detecção de lesões, diagnóstico da patologia e interpretação de imagens médicas. Dessa forma, esse artigo apresenta um banco de dados não-relacional público, orientado a documentos, de nódulos pulmonares identificados e classificados por especialistas, caracterizados por Atributos de Textura 3D, com o intuito de assistir ferramentas de auxílio computadorizado ao diagnóstico de câncer de pulmão e a pesquisa em detecção e classificação de nódulos pulmonares. A base de dados atualmente se encontra com 994 exames, 2.434 nódulos e 32.101 imagens, sendo 18.844 tomografias computadorizadas e 13.257 nódulos segmentados.
**Palavras-chave:** Câncer de Pulmão, Auxílio Computadorizado ao Diagnóstico, *NoSQL*, Banco de Dados Orientado a Documentos, MongoDB.

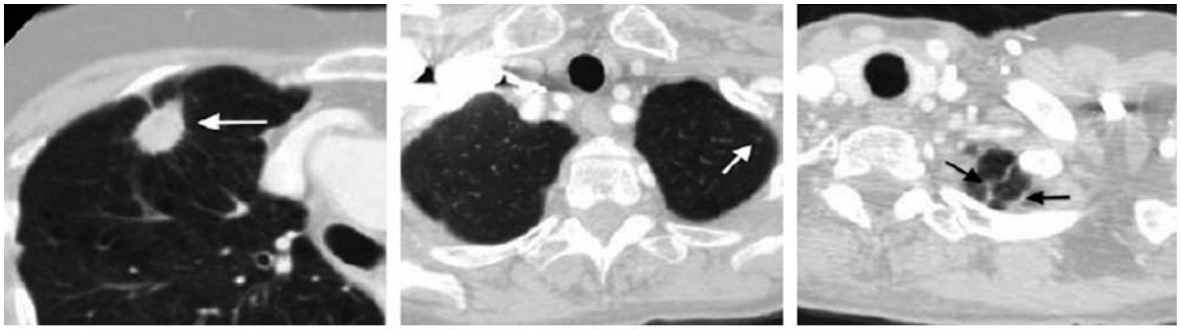
***Abstract:*** *Due to the difficulty to diagnose the lung cancer, it is necessary to integrate the computer-based tools with the lesion detection, pathology diagnosis and image interpretation processes. That been said, this paper presents a public non-relational document-oriented database of pulmonary nodules, identified and classified by specialists, characterized by 3D Texture Attributes, to assist lung cancer CAD (Compute-Aided Diagnosis) systems and pulmonary nodule detection and classification research. The developed database is now with 994 exams, 2,434 nodules and 32,101 images, 18,844 of them are CT (Computerized Tomography) scans and 13,257 are segmented nodules.*
***Keywords:*** *Lung Cancer, Computer-Aided Diagnosis, NoSQL, Document-oriented Database, MongoDB.*

## Introdução

Segundo o INCA, o câncer de pulmão é o mais comum de todos os tumores malignos, apresentando aumento de 2% ao ano na sua incidência mundial. No Brasil foi responsável por 21.867 mortes em 2010, dentre os tipos de câncer é o que mais fez vítimas [1]. O diagnóstico do câncer de pulmão é uma tarefa desafiadora, pois os nódulos são pequenos, apresentam baixo contraste e normalmente estão inclusos em estruturas anatômicas complexas. Além disso, imagens médicas são extremamente complexas por natureza e o seu diagnóstico consiste em uma tarefa minuciosa que deve ser realizada por especialistas qualificados. Contudo, mesmo o especialista mais experiente é condicionado pela capacidade humana de análise e sofre influências de fatores externos (e.g. ruído e luminosidade) e internos, como o nível de treinamento e condições psicológicas (e.g. fadiga e pressa). Portanto, esse especialista é tendente a erros de detecção (falha em detectar um câncer) ou erro de interpretação (falha em classificar corretamente um câncer detectado) [2]. Logo, é necessário integrar o auxílio computadorizado ao processo de diagnóstico de doenças e interpretação de imagens. O objetivo do diagnóstico auxiliado por computador (*Computer-Aided Diagnosis* - CAD) é melhorar a acurácia diagnóstica, assim como aprimorar a consistência na interpretação do diagnóstico em imagens, mediante o uso da sugestão de resposta diagnóstica fornecida por algum computador [3]. Por fim, a adoção de ferramentas computacionais no auxílio ao diagnóstico não é mais uma opção, mas uma necessidade [4].

O Instituto Nacional do Câncer (*National Cancer Institute* - NCI) dos Estados Unidos criou um projeto entre várias instituições americanas com o objetivo de estimular o desenvolvimento de métodos CAD pela comunidade de pesquisa em imagens médicas. Esse projeto resultou em um grande repositório de imagens de várias modalidades e tipos de câncer [5]. Uma das coleções de imagens desse repositório é formada por tomografias computadorizadas de câncer de pulmão, chamada de *Lung Image Database Consortium* (LIDC). O LIDC consiste de 1.018 exames e 244.527 imagens torácicas com lesões identificadas e classificadas por radiologistas, que tem como objetivo ajudar no desenvolvimento, treinamento e avaliação de métodos CAD para detectar e diagnosticar o câncer de pulmão [6].

Porém, o LIDC é uma coleção de imagens médicas não organizada em esquema de banco de dados, logo, existe uma descentralização da correlação entre imagens, dados dos exames e a classificação dos nódulos realizada pelos especialista. Portanto, gerenciar e utilizar os dados do LIDC ainda é um problema. Além disso, o LIDC não contém informações fornecidas por descritores de imagens. Logo, existe a necessidade de extrair atributos referentes aos *pixels* e armazenar esses



(a) Nódulo ≥ 3mm de diâmetro. (b) Nódulo < 3mm de diâmetro. (c) Não-nódulo ≥ 3mm de diâmetro.

Figura 1: Exemplos de lesões (apontadas pelas setas) de acordo com suas classificações [6].

atributos, além dos dados dos exames e as imagens em uma base de dados capaz de integrar todas essas informações. Porém, um importante aspecto computacional deve ser levado em conta no gerenciamento desse banco de dados: desempenho. Usuários de sistemas CAD precisam de uma resposta rápida do *software* [7]. Logo, é necessária a utilização de tecnologias que façam uso de armazenamento de grandes volumes de dados com alto desempenho. Com o intuito de conseguir aliar alto desempenho ao armazenamento de grandes volumes de informações, foram criados os Bancos de Dados *NoSQL* ou Não-Relacionais. Além de alta performance, bancos *NoSQL* tem como vantagens o alto *throughput* de dados, alto poder de escalabilidade e maior simplicidade no projeto de esquema de dados [8].

Os objetivos desse trabalho são: (1) desenvolver uma grande coleção de dados sobre nódulos pulmonares caracterizados por Atributos de Textura 3D, utilizando as imagens públicas do LIDC; (2) armazenar essa coleção em um banco de dados *NoSQL* e (3) promover a pesquisa reprodutível, disponibilizando o banco de dados publicamente com o intuito de ajudar a comunidade de engenharia biomédica no desenvolvimento, treinamento e avaliação de ferramentas de auxílio computadorizado ao diagnóstico do câncer de pulmão e na pesquisa em detecção e classificação de nódulos pulmonares.

## Materiais e métodos

As imagens de Tomografia Computadorizada (TC) no padrão DICOM utilizadas neste trabalho são provenientes do projeto público LIDC (*Lung Image Database Consortium*). As lesões em cada imagem foram identificadas por 4 radiologistas do próprio LIDC e classificadas em três categorias (Figura 1):

- Nódulos ≥ 3mm de diâmetro;
- Nódulos < 3mm de diâmetro;
- Não-nódulos ≥ 3mm de diâmetro.

Cada exame do LIDC possui um arquivo XML que descreve as regiões de interesse (*Region of Interest* - ROI) das lesões. Cada ROI possui um identificador da imagem original. No caso dos nódulos ≥ 3mm de diâmetro, cada ROI possui associado as posições cartesianas *(x,y)*, obtidas manualmente pelo especialista. Após a identificação, os nódulos foram classificados segundo as características subjetivas de calcificação, estrutura interna, lobulação, malignidade, margem, esfericidade, espiculação, sutileza e textura. Para cada característica foi atribuído um valor de 1 a 5 pelo radiologista, sendo 1 a menor probabilidade de ser maligno e 5 a maior probabilidade de ser maligno. No caso dos nódulos < 3mm de diâmetro e não-nódulos ≥ 3mm de diâmetro, apenas o centro de massa da lesão foi marcado. Os não-nódulos < 3mm de diâmetro não são classificados, pois não possuem marcação de região de interesse, nem posição de centro de massa.

No banco desenvolvido nesse trabalho, consideramos as lesões identificadas pelo radiologista que detectou o maior número de nódulos em cada exame. As lesões identificadas pelos outros 3 especialistas foram descartadas, com o objetivo de evitar redundância na identificação de nódulos. As imagens que não possuem lesões também foram descartadas, visto que elas não carregam informações a respeito dos nódulos. Os nódulos ≥ 3mm de diâmetro foram designados como *bignodules*, os nódulos < 3mm de diâmetro como *smallnodules* e os não-nódulos ≥ 3mm de diâmetro como *nonnodules*.

Os *bignodules* foram segmentados usando as marcações feitas pelo especialista (Figura 2). *Smallnodules* e *nonnodules* não foram segmentados, pois não possuem as marcações de contorno da lesão, apenas o centro de massa. Após a segmentação manual, foram extraídos os Atributos de Textura 3D, a partir da Matriz de Coocorrência, propostos por Haralick et al. (1973), com distância de 1 e 2 *pixels*, e orientações $\theta = 0º, 45º, 90º$ e $135º$ (sentido anti-horário) [9]. A utilização de nove atributos e quatro orientações angulares permitiu a criação de um vetor de características para cada imagem com 36 posições. Os Atributos de Textura (ATs) usados foram energia, entropia, matiz, momento da diferença inverso, contraste, proeminência, correlação, variância e homogeneidade.

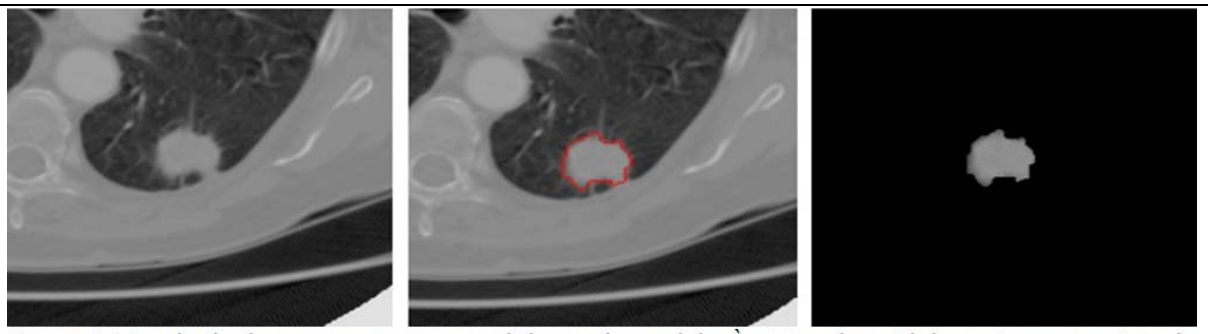


Figura 2: Resultado da segmentação manual de um *bignodule*. À esquerda, nódulo na imagem original. No centro, um nódulo identificado e marcado pelo radiologista. À direita, o mesmo nódulo segmentado.

A abordagem *NoSQL* utilizada no trabalho foi Orientada a Documentos, devido à sua consolidação como a abordagem não-relacional mais robusta e por possuir uma estrutura adequada ao armazenamento das informações do XML do LIDC [10]. Um banco de dados orientado a documentos é formado por um conjunto de coleções. Cada coleção possui um conjunto de documentos. Cada documento é formado por um conjunto de pares chave-valor, listas ou outros documentos.

O SGBD (Sistema de Gerenciamento de Banco de Dados) utilizado foi o MongoDB, devido ao grande poder de processamento paralelo, alta performance na recuperação dos dados do banco, grande escalabilidade em instâncias de BDs e por possuir a especificação GridFS, necessária para armazenar as imagens no banco desenvolvido [8]. A unidade básica de armazenamento do MongoDB é o documento JSON (*JavaScript Object Notation*), porém ele é capaz de persistir documentos no formato BSON (*Binary* JSON), por razões de eficiência e desempenho. Ele também é capaz de gerar um identificador, chamado *ObjectId*, que garante a unicidade dos documentos no banco de dados. Através da convenção GridFS, o MongoDB consegue armazenar e recuperar arquivos arbitrários em formato binário. O GridFS faz uso de duas coleções para armazenar seus arquivos: *files* e *chunks*. A coleção *files* guarda os meta-dados do arquivo a ser armazenado e a coleção *chunks* armazena os dados no formato BSON do arquivo através da chave *data*.

## Resultados

O banco de dados desenvolvido possui duas coleções principais: *exams* e *images*. A coleção *exams* contém os dados referentes aos exames. A coleção *images* armazena as imagens originais, os nódulos segmentados e os meta-dados das imagens, através do sistema GridFS do MongoDB. A coleção *images* é subdividida nas coleções *files* e *chunks*. A coleção *images.files* armazena os meta-dados (e.g. tamanho em *bytes* e a data de *upload*) e a coleção *images.chunks* armazena os dados binários das imagens.

Cada documento da coleção *exams* possui a chave *examId* com um identificador do exame e a chave *readingSession* que contém três listas: uma com todos os *smallnodules*, uma com todos os *nonnodules* e a última com todos os *bignodules*. Documentos das listas *nonnodules* e *smallnodules* possuem o identificador *noduleID;* a lista de regiões de interesse contida na chave *roi*, onde cada região possui a chave *originalImage* com o identificador *ObjectId* da imagem na coleção *images.files;* a chave *edgeMap* com a posição do centro de massa, entre outros. Os documentos da lista *bignodules* possuem as seguintes propriedades: o identificador *noduleID;* as características do nódulo (e.g. chave *calcification* possui o valor da calcificação); o vetor de Atributos de Textura 3D contido na chave *textureAttributes;* a lista de regiões de interesse (chave *roi*), onde cada região possui a chave *originalImage* com o identificador *ObjectId* da imagem original na coleção *images.files,* a chave *noduleImage* com o identificador *ObjectId* da imagem do nódulo segmentado na coleção *images.files*, a chave *edgeMap* com as marcações do nódulo na imagem original; entre outras informações.

As imagens originais foram armazenadas na coleção *images* no padrão DICOM e as imagens dos nódulos segmentados foram armazenadas no formato PNG. Foi assegurado que as intensidades dos *pixels* do nódulo na imagem original permanecessem inalterados na imagem do nódulo segmentado.

Atualmente, o banco de dados desenvolvido encontra-se com o seguinte *status*: 994 exames, 2.434 *bignodules*, 2.935 *smallnodules*, 5.007 *nonnodules* e 32.101 imagens, sendo 18.844 tomografias computadorizadas e 13.257 nódulos segmentados manualmente. As tomografias podem apresentar mais de uma lesão e os nódulos podem não possuir a mesma quantidade de cortes. Os arquivos referentes ao banco de dados estão disponíveis no endereço *http://bit.ly/1rke5mS* (verificado em 10/06/2014). O *download* desses arquivos pode ser feito por qualquer estudante, professor e/ou pesquisador. Para fazer uso do banco, basta realizar a operação *restore* do MongoDB dos arquivos disponibilizados na página anterior. O acesso ao banco pode ser feito através do MongoDB *Shell*, da API oficial disponível em diversas linguagens de programação ou por qualquer ferramenta de gerenciamento de bancos de dados em MongoDB.

## Discussão

A conversão e o armazenamento dos exames e imagens do LIDC em formato de documento JSON no

MongoDB permitiram um melhor gerenciamento das informações contidas inicialmente no arquivo XML, visto que tanto os dados dos exames como as imagens foram centralizadas em um mesmo local. Anteriormente, a coleção de imagens do LIDC não estava organizada em um banco de dados que integrasse as informações dos exames, a classificação dos nódulos e as imagens. Após a nossa implementação, as informações do nódulo e do exame, as marcações e classificações feitas pelo radiologista, as tomografias originais e as imagens com os nódulos segmentados, foram armazenados em uma única base no MongoDB. Isso facilitou o manuseio dos dados e das imagens, devido a centralização das informações.

O trabalho desenvolvido também contribuiu com a expansão do LIDC, com o acréscimo dos Atributos de Textura 3D, que permitiram a caracterização inicial do nódulo em relação a distribuição e intensidade dos *pixels* nas imagens.

A tecnologia *NoSQL* possui fraco acoplamento, logo, o desenvolvimento do banco de nódulos possui grande potencial para ser aplicado ao contexto de *Big Data*, pois a base desenvolvida poderá ser integrada ao prontuário eletrônico do paciente e outras bases de dados em saúde, podendo inclusive ser integrado a projetos de recuperação de imagens baseada em conteúdo. Com essa integração, facilitada pela utilização do MongoDB, será possível inferir novas informações que possam auxiliar no diagnóstico precoce do câncer de pulmão.

Como o foco do trabalho são os nódulos pulmonares, a exclusão das imagens do LIDC que não possuem lesões e o armazenamento das imagens que as possuem, junto dos nódulos segmentados, possibilitaram uma redução de 86% no número de imagens e uma economia de 92% de *gigabytes* em espaço de armazenamento em disco. Além disso, a segmentação permitiu uma melhor visualização do nódulo, visto que a região de interesse foi evidenciada.

A efetividade da base desenvolvida tem sido comprovada nos algoritmos em desenvolvimento pelo nosso laboratório, em diferentes aplicações CAD, como recuperação de nódulos similares, identificação de atributos relevantes, detecção de lesões pulmonares, entre outros [2] [11]. A disponibilização da base de nódulos publicamente garante que o mesmo banco de imagens possa ser usado por vários pesquisadores servindo de *testbed* em diferentes projetos de pesquisa.

Atualmente o banco de dados está disponível para usuários intermediários, como desenvolvedores de sistemas e pesquisadores com conhecimento em programação. Uma interface gráfica encontra-se em desenvolvimento para que usuários finais possam ter acesso ao banco de imagens com maior facilidade.

Esse artigo apresentou o desenvolvimento de um banco de dados *NoSQL*, orientado a documentos, de nódulos pulmonares em tomografias computadorizadas. A base de imagens está disponível para *download* com intuito de servir de apoio ao desenvolvimento, treinamento e avaliação de ferramentas de auxílio computadorizado ao diagnóstico do câncer de pulmão e à pesquisa em detecção e classificação de nódulos pulmonares. Todos os nódulos foram identificados por um radiologista do LIDC, segmentados manualmente através de marcações realizadas pelo mesmo radiologista e tiveram os Atributos de Textura 3D extraídos.

**Referências**


[1] INCA (2014). Instituto Nacional de Câncer. www.inca.gov.br. [Online; acessado em 10-06-2014].

[2] Oliveira, M. C. and Ferreira, J. R. (2013). A bag-of-tasks approach to speed up the lung nodules retrieval in the BigData age. In Proceedings of the 15th International Conference on E-Health Networking, Application & Services (HealthCom), pages 632–636. IEEE.

[3] Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics, 31(4-5):198–211.

[4] Azevedo-Marques, P. M. d. (2001). Diagnóstico auxiliado por computador na radiologia. Radiologia Brasileira, 34(5):285–293.

[5] Armato, S. G., Roberts, R. Y., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., McLennan, G., Engelmann, R. M., Bland, P. H., Aberle, D. R., Kazerooni, E. A., et al. (2007). The lung image database consortium (lidc): Ensuring the integrity of expert-defined "truth". Academic radiology, 14(12):1455–1463.

[6] Mclennan, G., Bidaut, L., Mcnitt-gray, M. F., Meyer, C. R., and et al. (2011). The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. Medical Physics, 38:915–931.

[7] Scholl, I., Aach, T., Deserno, T. M., and Kuhlen, T. (2011). Challenges of medical image processing. Computer Science-Research and Development, 26(1-2):5–13.

[8] Tiwari, S. (2011). Professional NoSQL. John Wiley and Sons, Inc.

[9] Haralick, R. M., Shanmugam, K., and Dinstein, I. H. (1973). Textural features for image classification. Systems, Man and Cybernetics, IEEE Transactions on, (6), 610-621.

[10] Strauch, C., Sites, U., and Kriha, W. (2011). NoSQL databases. Stuttgart Media University.

[11] Ferreira Junior, J. R., Oliveira, M. C. and Freitas, A. L. (2014). Performance Evaluation of Medical Image Similarity Analysis in a Heterogeneous Architecture. In Proceedings of 27th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2014). 159-164. IEEE.