Challenges in Radiology image data analysis

KN Manjunath PhD

Faculty in Computer Science and Engineering, Manipal Academy of Higher Education (Formerly Consultant, SIEMENS Healthineers, Bengaluru), kn.manjunath@ieee.org
Presentation outline

1. Application of Medical Image Processing
2. Radiology Image Source
3. Data Collection
4. Data Analysis
5. Tools used
6. Challenges
CT findings are reported based on the Hounsfield Units (HU) – the tissue intensity, the texture and shape.
7/21/2020

Fig 1: Image header and pixel details of ajpg image and a DICOM image (Source: SIEMENS customer magazine, 2018)
Mathematics and image processing remains same, only physics, image acquisition method and image interpretation differs in all modalities.
The great challenge is validating the information from data and extracting the relevant information for analysis. Example,

Customer (Radiologist) has the requirement in inception phase
I want to know the set of images which have good contrast of the tissues.

**CAD schemes typically consist of the following key steps:**

1) Apply automated image analysis to extract a vector of *quantitative features* to characterize the relevant image content,
2) Apply a *pattern classifier* to determine the category to which the extracted feature vector may belong

**Radiomics:**
In the field of medicine, *radiomics* is a method that extracts large amount of features from radiographic medical images using *data-characterisation* algorithms

**During Image Preprocessing:**
Extracting the best projection data and the right combination of image reconstruction technique for better 2D image generation

**During Image Post processing (Summers, 2012, 10.1016/j.media.2012.02.005)**
Computer aided detection (*CAD*$_e$), and Computer Aided Diagnosis (*CAD*$_x$)

*Fig 3:* The appearance of tumor cells on axial CT image and the boundary of the large intestine on axial CT
Radiology image source (1/3)

Cancer Imaging Archive
https://www.cancerimagingarchive.net/

CT Medical Images
https://www.kaggle.com/kmader/siim-medical-images

NCBI – Medical Image Databases
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC61234/

NIH Database of 100,000 Chest X-Rays
https://nihcc.app.box.com/v/ChestXray-NIHCC

Open-Access Medical Image Repositories
http://www.aylward.org/notes/open-access-medical-image-repositories

The Berkeley Segmentation Dataset and Benchmark
https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/

Online Medical Images

UCL – Medical Image Repositories
https://www.ucl.ac.uk/child-health/support-services/library/resources-z/medical-image-repositories

DERMOFIT IMAGE LIBRARY (skin lesion images)
https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html

Cancer facts and figures (Siegel R, 2020) gives the cancer statistics of all anatomies and all geographical areas (countrywise)
Radiology image source (2/3) – NCI Dataset

• **Dataset**: More than 10TB of CT, MRI, PET and RT Objects. Maintained by National Institute of Health (NIH), National Cancer Institute (NCI), Walter Reed Army Medical center (WRAMC), and Washington School of Medicine (WSM), United States.

• **Procedure to download**
  - Install the downloader
  - Install Jdk latest
  - Select the dataset and download manifest file
  - Open manifest file, UI pops up
  - Select the folder and click download images

• **Acknowledgements**: Acknowledge the dataset providers and cite their publications

<table>
<thead>
<tr>
<th>Data Access</th>
<th>Detailed Description</th>
<th>Citations &amp; Data Usage Policy</th>
<th>Versions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Access</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Click the <strong>Download</strong> button to save a <code>.tcia</code> manifest file to your computer, which you must c NBIA Data Retriever. Click the <strong>Search</strong> button to open our Data Portal, where you can browse collection and/or download a subset of its contents.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPTAC-GBM</td>
</tr>
<tr>
<td>Created by Tracy Nolan, last modified by natasha honomichi on Jun 18, 2020</td>
</tr>
</tbody>
</table>

**Summary**

This collection contains subjects from the National Cancer Institute’s Clinical Proteomic Tumor Analysis Consortium, Glioblastoma Multiforme (CPTAC-GBM) cohort. CPTAC is a national effort to accelerate the understanding of the molecular basis of cancer through the application of large-scale proteome and genome analysis, or proteogenomics. Radiology and pathology images from CPTAC Phase 3 patients are being collect and made publicly available by The Cancer Imaging Archive to enable researchers to investigate cancer
Radiology image source (3/3) – Grand challenge DB

- Clinically proved dataset
- Freely accessible
- Register in the website for download
- Submit the results to their conference
- Best place to defend your results and for good publications

Grand challenges dataset (https://grand-challenge.org/challenges/)
Data collection (1/3) – Big data

Why is data processing so important?

Big Data demands cost effective, innovative forms of information processing for enhanced insight and decision making.

Dimensions and Challenges

- **Velocity**: Thousands of images hit the PACS server every minute. Data must be analyzed immediately for necessary doctor’s intervention.
- **Volume**: Dataset is growing in GB, PB.
- **Variety**: Structured and semi-structured data. Clinical notes, imaging, audio transcriptions, EEG.
- **Veracity and Validity**: Abnormality, wrong patient details, bias and noise in images and data.
- **Volatility**: Till what time I have to store these images (till 2025? 2050? or forever)?
- **Variability**: When the scanner halts or when patient moves (induces disturbances in images).
Data collection (2/3) – DICOM format

- Radiology images are in **DICOM** format (for CT, MRI, PET, SPECT, RT)

- Each DataElement has one row on information shown on the right.

- They are protocol 3.3 (NEMA 2020b version) standard files

- File consists of stream of bytes which are parsed into different logical entities.

- For example, Patient Name, Age, Gender, ID, DOB constitutes **Patient Module**

- Using incomplete dataset is of no use

- Classify them based on important parameters (slide 12). This helps in empirical testing.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>VR</th>
<th>VM</th>
<th>Value</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>0002.0000</td>
<td>Group Length</td>
<td>UL</td>
<td>1</td>
<td>198</td>
<td>4</td>
</tr>
<tr>
<td>0002.0001</td>
<td>File Meta Information Version</td>
<td>OB</td>
<td>0</td>
<td>00 01</td>
<td>2</td>
</tr>
<tr>
<td>0002.0002</td>
<td>Media Storage SOP Class UID</td>
<td>UI</td>
<td>1</td>
<td>1.280.10008.5.1.4.1.481.5</td>
<td>30</td>
</tr>
<tr>
<td>0002.0003</td>
<td>Media Storage SOP Instance UID</td>
<td>UI</td>
<td>1</td>
<td>1.3.6.1.4.1.4519.5.2.1.7064.0195.21467897294753319</td>
<td>64</td>
</tr>
<tr>
<td>0002.0010</td>
<td>Transfer Syntax UID</td>
<td>UI</td>
<td>1</td>
<td>1.280.10008.12</td>
<td>16</td>
</tr>
<tr>
<td>0002.0011</td>
<td>Implementation Class UD</td>
<td>UI</td>
<td>1</td>
<td>1.2.40.1.11.1</td>
<td>10</td>
</tr>
<tr>
<td>0002.0012</td>
<td>Implementation Version Name</td>
<td>SH</td>
<td>1</td>
<td>dcm4che-1.4.35</td>
<td>14</td>
</tr>
<tr>
<td>0002.0013</td>
<td>Specific Character Set</td>
<td>CS</td>
<td>1</td>
<td>ISO R 100</td>
<td>8</td>
</tr>
<tr>
<td>0002.0014</td>
<td>Instance Creation Date</td>
<td>DA</td>
<td>1</td>
<td>20011110</td>
<td>8</td>
</tr>
<tr>
<td>0002.0015</td>
<td>Instance Creation Time</td>
<td>TM</td>
<td>0</td>
<td>16344</td>
<td>8</td>
</tr>
<tr>
<td>0002.0016</td>
<td>SOP Class UID</td>
<td>UI</td>
<td>1</td>
<td>1.280.10008.5.1.4.1.481.5</td>
<td>30</td>
</tr>
<tr>
<td>0002.0017</td>
<td>SOP Instance UID</td>
<td>UI</td>
<td>1</td>
<td>1.3.6.1.4.1.4519.5.2.1.7064.0195.21467897294753319</td>
<td>64</td>
</tr>
<tr>
<td>0002.0020</td>
<td>Study Date</td>
<td>DA</td>
<td>1</td>
<td>20011120</td>
<td>8</td>
</tr>
<tr>
<td>0002.0030</td>
<td>Study Time</td>
<td>TM</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>0002.0050</td>
<td>Accession Number</td>
<td>SH</td>
<td>1</td>
<td>281949768494126</td>
<td>16</td>
</tr>
<tr>
<td>0002.0060</td>
<td>Modality</td>
<td>CS</td>
<td>1</td>
<td>RTPLAN</td>
<td>8</td>
</tr>
<tr>
<td>0002.0070</td>
<td>Manufacturer</td>
<td>LO</td>
<td>0</td>
<td>ADAC</td>
<td>8</td>
</tr>
<tr>
<td>0002.0090</td>
<td>Referring Physician's Name</td>
<td>PN</td>
<td>1</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>0002.1030</td>
<td>Study Description</td>
<td>LO</td>
<td>1</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>0002.1034</td>
<td>Manufacturer's Model Name</td>
<td>LO</td>
<td>1</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>0002.1100</td>
<td>Referenced Study Sequence</td>
<td>SQ</td>
<td>1</td>
<td>FE FF 00 E0 FF FF FF FF FF FF FF FF FF FD 00 00 00 11 18 00 00 00 31 E0 32 32 E0 120</td>
<td>16</td>
</tr>
<tr>
<td>0003.0010</td>
<td>Patient's Name</td>
<td>PN</td>
<td>1</td>
<td>HNSCC-00-0003</td>
<td>14</td>
</tr>
<tr>
<td>0003.0020</td>
<td>Patient ID</td>
<td>LO</td>
<td>1</td>
<td>HNSCC-00-0003</td>
<td>14</td>
</tr>
<tr>
<td>0003.0030</td>
<td>Patient's Birth Date</td>
<td>DA</td>
<td>1</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>0003.0040</td>
<td>Patient's Sex</td>
<td>M</td>
<td>2</td>
<td>59 45 53 20</td>
<td>4</td>
</tr>
<tr>
<td>0003.0060</td>
<td>Not in Dictionary</td>
<td>UN</td>
<td>1</td>
<td>00 00 95 20</td>
<td>4</td>
</tr>
<tr>
<td>0003.0063</td>
<td>Not in Dictionary</td>
<td>UN</td>
<td>1</td>
<td>50 66 72 20</td>
<td>4</td>
</tr>
<tr>
<td>0003.0064</td>
<td>Not in Dictionary</td>
<td>UN</td>
<td>1</td>
<td>00 00 95 20</td>
<td>4</td>
</tr>
<tr>
<td>0003.0066</td>
<td>Not in Dictionary</td>
<td>UN</td>
<td>1</td>
<td>50 66 72 20</td>
<td>4</td>
</tr>
<tr>
<td>0003.0070</td>
<td>Private Creator</td>
<td>CT</td>
<td>P</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>0003.0100</td>
<td>Not in Dictionary</td>
<td>UN</td>
<td>1</td>
<td>48 46 53 43 43 20</td>
<td>14</td>
</tr>
<tr>
<td>0003.1010</td>
<td>Not in Dictionary</td>
<td>UN</td>
<td>1</td>
<td>00 00 95 20</td>
<td>4</td>
</tr>
<tr>
<td>0003.1013</td>
<td>Not in Dictionary</td>
<td>UN</td>
<td>1</td>
<td>31 37 36 38 30 34 30</td>
<td>8</td>
</tr>
<tr>
<td>0003.1016</td>
<td>Body Part Examined</td>
<td>CS</td>
<td>1</td>
<td>HEADNECK</td>
<td>8</td>
</tr>
<tr>
<td>0003.1020</td>
<td>Software Versions</td>
<td>LO</td>
<td>2</td>
<td>9.89S.8</td>
<td>8</td>
</tr>
</tbody>
</table>
Data collection (3/3) – Methods

- **Data acquisition and validation methods** – Check for image acquisition protocols, approval of these dataset, whether single center or multi center scans, clinical trials or mass screening etc.

- **Data Format and Usage Notes** – Check whether images are in DICOM format (.dcm/.ima)

- **Data selection** – decide on samples (n in N), apply sampling techniques (structured, quota, systematic etc.)

- **Data collection** – Collection method (questionnaire, observing, first hand collection etc..)

- **Data analysis and processing**
  - **Data analysis** – Check reliability, suitability (artefact and noise), and adequacy of images
  - **Classification** – Classify them into groups based on important parameters of interest
  - **Data Validation** – DICOM images validation against Standard PS3.3 (2020b)
### Data Analysis (1/4) - Classification

#### Fig. 4: Number of datasets classified based on various CT image acquisition parameters. The classified datasets are used for empirical testing based on the parameters of interest.
Data Analysis (2/4) – Analyze the quality

Fig 4: Bad diagnostic quality in abdominal CT images (Sagital, coronal and axial views respectively (a, c, e). a) Incomplete air insufflation, c, d ) Patient too large and outside the scan field of view, c, d) Streak artifact.

- Apply image quality assessment algorithms like CNR, PSNR, MTF, Noise measurement while analyzing its diagnostic quality
- Whether the intensity values on image is enough to describe a structure?

Fig 5: Accurate tissue details in abdominal CT (Manjunath K N et. al., 2016, 10.1166/jmihi.2016.1786)
This is the most difficult part in the medical image research.

The dataset is mainly checked for type 1 and type 2 attributes as per the DICOM standards.

Even though dataset are statistically approved, still it is necessary to check the completeness as per the latest standard (PS 3.3, 2020b)

Using incomplete dataset is unethical and meaningless.

One missing DICOM tag leads to incomplete dataset.
Features:
Features are the representative candidate of the entire image (single instance learning) or entire volume (multiple instance learning). Features are selected from from *domain* perspective & from *technical* perspective.

Radiological/Technical features
What we compute from an image are,
- Mean, geometrical features, morphological,
- Shape index,
- Principal curvarture
- Principal Component Analysis,
- Morphology, texture, variation in shape, orientation,
- Surface normal overlap etc..

Clinical features
- Patient demographic details
- Family history
- Disease symptomsetc..

Ultimately all these features (irrespective of their data types) are converted to numerical before data analysis.
Tools for data analysis and visualization

- pyDICOM library (open source), Tensorflow, and R
- Accort.NET Framework (open source, C# based)
- CNTK (Microsoft Corporation, USA, C# based)
- Weka 3.8.4 (Java based)
- 3D slicer (Fedoroc et. al., Harvard University)
- MITK (Dkfz, Germany, C++ based)
- Mevislab (Fraunhofer Institute, Germany)
- Syngo FastView (SIEMENS, Erlangen, Germany) (only viewer)
- DICOM Viewer (Philips, Netherlands) (only viewer)
Tools for analysis (2/6) – Reading DICOM in Python

#Program to read the DICOM files from a directory and exporting the DICOM object into xml format

#Source:

#Modules for xml related
import xml.etree.ElementTree as ET
#installed through pip install -U cmdname
import os
import glob
import pydicom
import cv2
print(__doc__)

#For reading all the files iteratively within a given path
path = 'E:\Manjunath KN\Samples\ClivusChordoma'

#Specific type of files
filename = '*.dcm'

#Required for file name formatting
FileNameIncrementer=1
DICOMFileName='DICOMFile_'
PNG = False

#Collect the files list from the directory
files=glob.glob(path+filename)

#For each file in the directory, read the DICOM file and build the XML tree structure for file in files
for file in files:
    root=ET.Element("root")
doc=ET.SubElement(root,"doc")
    ds = pydicom.dcmread(file)
    #For each of the DICOM elements in the .dcm/.ima file, iterate and add the XML tags under the
    #DicomDataFromPython element
    for elem in ds.iterall():
        if str(elem.tag) == '(7fe0, 0010)':
            ET.SubElement(doc, "DataElement", Description=elem.description(), tag=str(elem.tag), VR=str(elem.VR), VM=str(elem.VM), Value=(ds[0x7fe0, 0x0010]))
        else:
tree=ET.ElementTree(doc)

#Write the xml tree to the formatted file name under the same directory as original DICOM files
tree.write(path+DICOMFileName+ds.SOPInstanceUID+str(FileNameIncrementer)+'.xml')
print(path+DICOMFileName+ds.SOPInstanceUID+str(FileNameIncrementer)+'.xml')

#Increment the counter which is used to format the xml file name
FileNameIncrementer+=1
<xml version="1.0"?>
  <PatientID>SD VC-125</PatientID>
  <SOPClassUID>1.2.840.10008.5.1.4.1.1.2</SOPClassUID>
  <BlobInformation>
    <BlobDetails>
      <PerpendicularLines><FirstPoint><X>82</X><Y>288</Y></FirstPoint>
      <SecondPoint><X>82</X><Y>284</Y></SecondPoint>
      <DistanceBetweenTwoPoints>0</DistanceBetweenTwoPoints></PerpendicularLines>
      <MedialAxis><Point><X>82</X><Y>286</Y></Point></MedialAxis>
      <LesionAssociatedWithBlob><Point><X>83</X><Y>286</Y></Point></LesionAssociatedWithBlob>
      <BlobDiameter>35</BlobDiameter>
      <SliceLocation>111.17001299999998</SliceLocation>
      <BlobID>9</BlobID>
      <BlobCenter><X>91</X><Y>288</Y></BlobCenter>
      <BlobBoundary><Location><X>73</X><Y>269</Y></Location></BlobBoundary>
      <Size><Width>35</Width><Height>41</Height></Size>
      <Height>41</Height></BlobBoundary>
      <BlobBoundaryPoints><Point><X>88</X><Y>268</Y></Point></BlobBoundaryPoints>
      <BlobArea>0</BlobArea>
    </BlobDetails>
    </BlobInformation>
    <ROIContour>
      #Sample Java Script Object Notation object
      {
        "glossary":
        {
          "ImageID": "WRAMC",
          "MaxHounseFieldUnit": "1320",
          "MaxHounseFieldUnit": "-1024",
          "UniqueIdentifier": E:\Manjunath KN\",
          "TubeVoltage": "120",
          "BitsAllocated": "16",
          "Rows": "512",
          "Columns": "512",
          "SizeOfPixel": "0.683594"
        }
      }
    </ROIContour>
  </ROIContour>
</xml>

Exported DICOM data (structured format) into XML and JSON format (semi structured)
1. **Accord.NET** provides statistical analysis, machine learning, image processing and computer vision methods for .NET applications ([http://accord-framework.net/](http://accord-framework.net/))

2. **Software required**
   Microsoft windows 10, Visual Studio 2017 and .NET Framework 4.7.2

3. **Installation in the project file**: Install-Package Accord.MachineLearning -Version 3.8.2-alpha

4. **After installation**, package.config in Visual Studio has the following tags with dlls loaded in the project file.

```xml
<packages>
  <package id="Accord" version="3.8.2-alpha" targetFramework="net472" />
  <package id="Accord.Math" version="3.8.2-alpha" targetFramework="net472" />
  <package id="Accord.Statistics" version="3.8.2-alpha" targetFramework="net472" />
</packages>
```

5. **Technical reports and technical publications** are available at [http://accord-framework.net/publications.html](http://accord-framework.net/publications.html)
Decision tree code is available at
http://accord-framework.net
(Source: Cesar C Souza)

All the functionalities are available through software classes, create the required instances and call the method with required parameters and plot the results on the image control.

Needs to code everything except the algorithm core logic
Steps:

- Create the training data (numerical) \((X \text{ and } Y)\) in csv format including the class label \((G)\)
- Training data is the feature extracted from the images
- Load the csv file
- Click on Create Tree
- Go to Model testing and run

Fig 7: The UI of DT showing the scatter plot of \(X\) and \(Y\)
Challenges in data analysis - conversions

- **Major challenge** is, some algorithms in some tools does not work with float and string, they expect integer only

- **DICOM** data incompleteness sometimes and compatibility checking against the DICOM standard

- **Scaling the data**: Standardization is followed when we have dataset with different units (gm, km, ltr, kv, etc.)
  E.g.: formula $z = (x - u) / s$, is used, where $z$ is the new value, $x$ is the original value, $u$ is the mean and $s$ is the standard deviation

- **Data type conversion**: Changing the string value to the numerical values.
  E.g.: In python, for mapping the country names, we use the dictionary of integer values to represent the string
  
  $d = \{ 'UK': 0, 'USA': 1, 'N': 2 \}$ and $df['Nationality'] = df['Nationality'].map(d)$
Conclusion

• $\text{CAD}_e$ and $\text{CAD}_x$ are young disciplines combining image processing, ML, Pattern Recognition and domain knowledge of medicine.

• For anything and everything, selection of right dataset and right features is most important.

• The quantum of data being produced is really challenging.

• Carefully sift the right dataset by looking at the dataset description and its validity.

• Be clear about what data you are going to process.

• You can do a comparative analysis to study the performance of the tools in an experimental setup.
References


All images shown in this presentation are properly cited and acknowledged. There is no infringement or copyright violation of any details. And these details are available in public domain
Acknowledgements