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

Application of Medical Image Processing → Radiology Image Source → Data Collection → Data Analysis → Tools used → Challenges

MIP applications (1/3) – Image capture

Image → Header → Matrix of intensities → Display intensities

Name : WaterLillies.jpg
Location : xyz lake
Time : 16.20
Photographer Name: Abc
Size : 100 x 100
Resolution : 200 dpi

File.jpg

10	20	40	50	80	90	80	30	20
34	54	24	10	45	78	02	34	34
32	67	12	78	29	10	09	34	89
22	87	34	69	19	93	17	39	93
38	48	10	18	19	83	84	95	98
22	74	82	19	20	48	81	85	01
38	82	38	38	75	57	29	01	28
47	57	93	69	96	37	37	91	38
47	75	29	02	27	95	57	67	69
98	85	27	85	94	74	34	12	23

LUT →

0	50	100
225 (124,21,240)	150	175
50	75	250

File.dcm/.ima

Name : Pulmo
Location : Abdomen
Pixel Size : (1mm x 1 mm)
Patient Name : xyz
Size : 100 x 100
Bits used per pixel : 16

-1024	-1000	2000	1500	1500	1500	1500	2000
3000	1500	3000	-1000	2000	3000	-1000	3000
3000	1500	-1000	3000	2000	3000	-1000	3000
3000	0	0	100	300	100	-1000	2000
0	-400	-400	2000	0	0	0	0
-1000	-1000	100	300	-1000	400	900	999
3072	3072	0	-1000	0	500	210	0
0	0	0	0	0	0	0	900

W,C →

CT findings are reported based on the Hounsfield Units (HU) – the tissue intensity, the texture and shape.

In CT image:

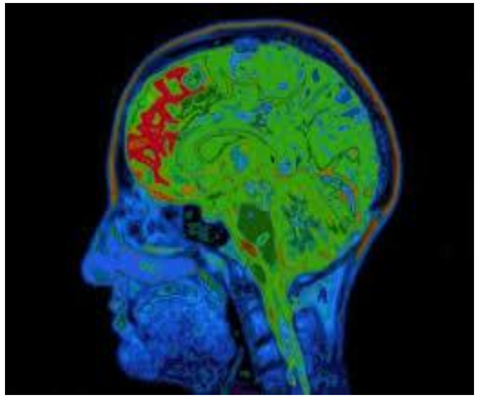
- Polyp: **+54 ± 5*HU*,**
- Lipoma (fat): **−89±10*HU*,**
- High density contrast: **+130*HU***
- Air: **−1000±10*HU* [3].**

Fig 1: Image header and pixel details of a jpg image and a DICOM image (Source: SIEMENS customer magazine, 2018)

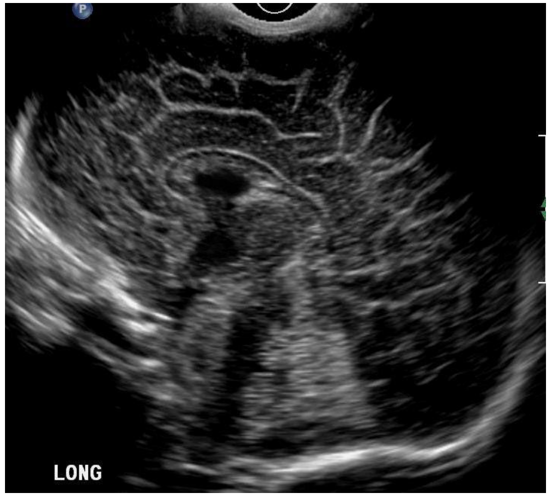
MIP applications (2/3) – Modalities

CT

MRI



PET

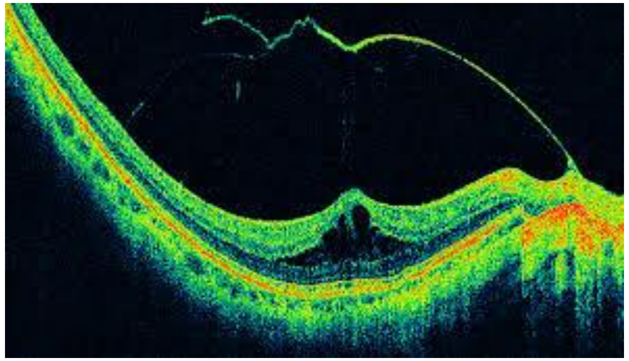


US

Fig 2: Different medical imaging modalities
(Source of CT, MRI, PET, US – SIEMENS Medical Solutions, Erlangen)

Thermogram

(Source: Matzinger Institute of Healing)



OCT

(Source: Ophthalmic Photographers' Society)

Mathematics and image processing remains same, only physics, image acquisition method and image interpretation differs in all modalities.

MIP applications (3/3) – CAD systems

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)

Freely downloadable radiology images from universities and radiology center

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
http://www.onlinemedicalimages.com/index.php/en/

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

Data Access   Detailed Description   Citations & Data Usage Policy   Versions

Data Access

Click the **Download** button to save a ".tcia" manifest file to your computer, which you must c
NBIA Data Retriever . Click the **Search** button to open our Data Portal, where you can browse
collection and/or download a subset of its contents.

Data Type	Download all or Query/Filter
Images (DICOM, 39.8 GB)	Download Search
Tissue Slide Images (SVS, 87 GB)	Download Search
Clinical Data API (JSON - more info)	Download
Discovery Study Proteomics/Clinical Data (external)	• CPTAC Data Portal (Georgetown) • Proteomic Data Commons
Genomics/Clinical Data (External)	Genomic Data Commons

- **Dataset description**

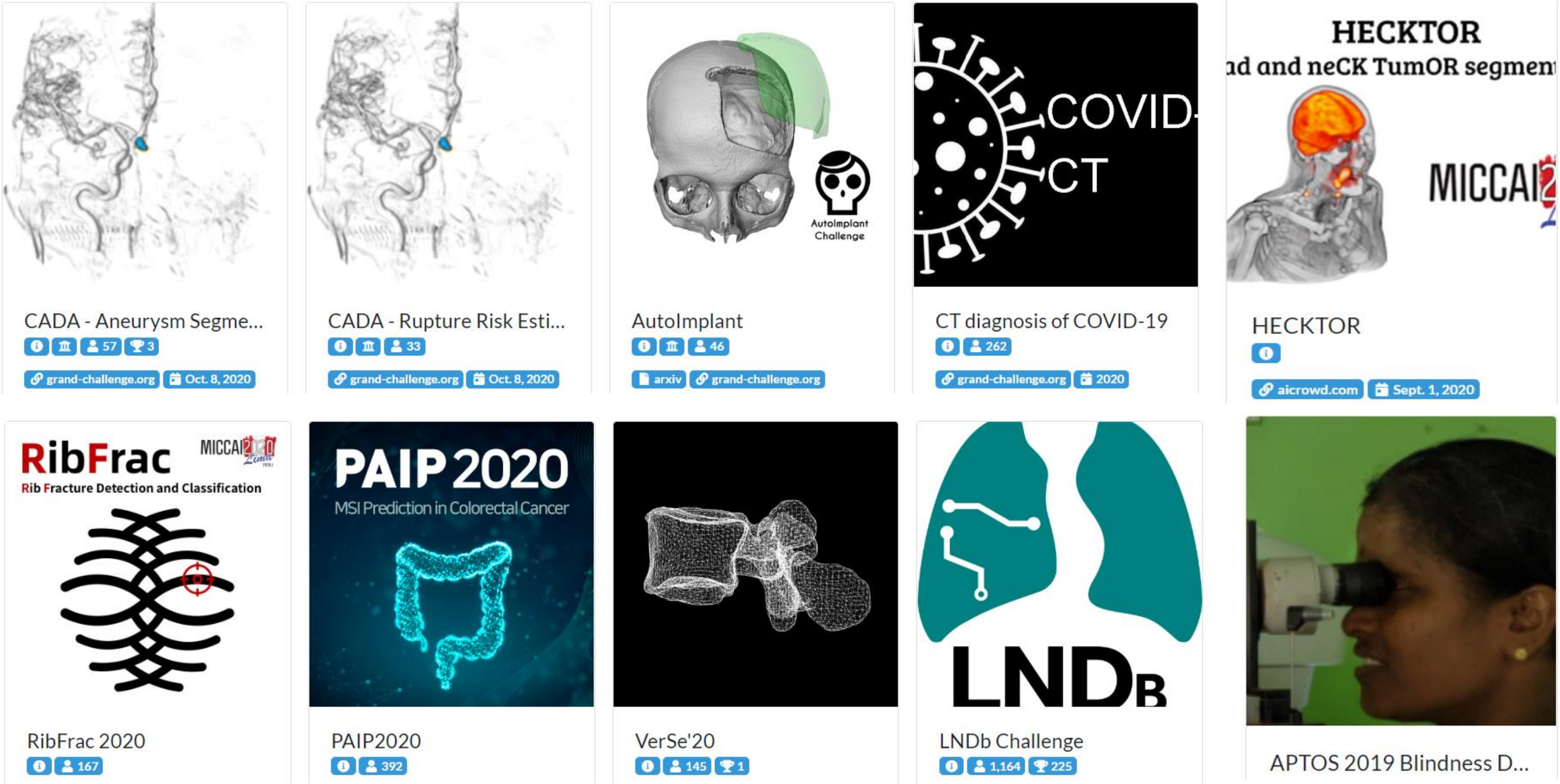
CPTAC-GBM

Created by Tracy Nolan, last modified by natasha honomichl on Jun 18, 2020

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

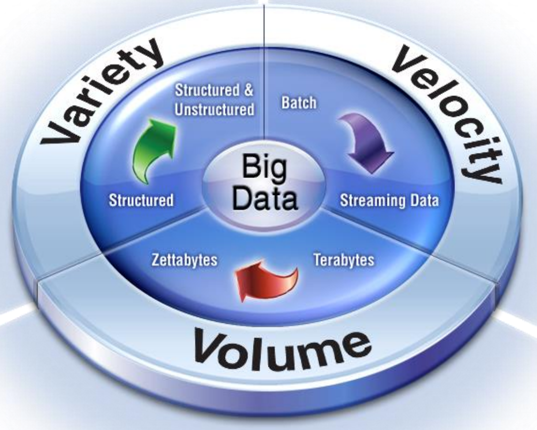
Radiology image source (3/3) – Grand challenge DB



- 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



(Image source: IBM Corp)

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 hits 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.

Tag	Description	VR	VM	Value	Length
0002,0000	Group Length	UL	1	198	4
0002,0001	File Meta Information Version	OB	1	00 01	2
0002,0002	Media Storage SOP Class UID	UI	1	1.2.840.10008.5.1.4.1.1.481.5	30
0002,0003	Media Storage SOP Instance UID	UI	1	1.3.6.1.4.1.14519.5.2.1.1706.8040.214619878292947523319	64
0002,0010	Transfer Syntax UID	UI	1	1.2.840.10008.1.2	18
0002,0012	Implementation Class UID	UI	1	1.2.40.0.13.1.1.1	18
0002,0013	Implementation Version Name	SH	1	dcm4che-1.4.35	14
0008,0005	Specific Character Set	CS	1	ISO_IR 100	10
0008,0012	Instance Creation Date	DA	1	20011012	8
0008,0013	Instance Creation Time	TM	1	163444	6
0008,0016	SOP Class UID	UI	1	1.2.840.10008.5.1.4.1.1.481.5	30
0008,0018	SOP Instance UID	UI	1	1.3.6.1.4.1.14519.5.2.1.1706.8040.214619878292947523319	64
0008,0020	Study Date	DA	1	20011120	8
0008,0030	Study Time	TM	1		0
0008,0050	Accession Number	SH	1	2819497684894126	16
0008,0060	Modality	CS	1	RTPLAN	6
0008,0070	Manufacturer	LO	1	ADAC	4
0008,0090	Referring Physician's Name	PN	1		0
0008,1030	Study Description	LO	1	RT SIMULATION	14
0008,1090	Manufacturer's Model Name	LO	1	Pinnacle3	10
0008,1110	**Referenced Study Sequence**	**SQ**	**1**	**FE FF 00 E0 FF FF FF FF 08 00 50 11 18 00 00 00 31 2E 32 2E**	**120**
0010,0010	Patient's Name	PN	1	HNSCC-01-0003	14
0010,0020	Patient ID	LO	1	HNSCC-01-0003	14
0010,0030	Patient's Birth Date	DA	1		0
0010,0040	Patient's Sex	CS	1	M	2
0012,0062	Not in Dictionary	UN	1	59 45 53 20	4
0012,0063	Not in Dictionary	UN	1	50 65 72 20 44 49 43 4F 4D 20 50 53 20 33 2E 31 35 20 41 6E	46
0012,0064	Not in Dictionary	UN	1	FE FF 00 E0 FF FF FF FF 08 00 00 01 06 00 00 00 31 31 33 31	662
0013,0010	*Private Creator*	*LO*	*1*	*CTP*	*4*
0013,1010	*Not in Dictionary*	*UN*	*1*	*48 4E 53 43 43 20*	*6*
0013,1013	*Not in Dictionary*	*UN*	*1*	*31 37 30 36 38 30 34 30*	*8*
0018,0015	Body Part Examined	CS	1	HEADNECK	8
0018,1020	Software Versions	LO	2	9.8\9.8	8

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

	SubjectID	Patient ID	DICOM validation	Metal artifact	Polyp found	Slice thickness	kVp	mA	Contrast	pixel spacing	Filter Kernel	Image dimension	Patient position	Slices	Colon segment	Age	Gender	Manufacturer	Description
2	SD VC 003M	1.3.6.1.4.1.9328.50.6.554			No	2.5	120	240		0.703125	SOFT	512,512	FFS,FFP	434,453		40	M		
3	SD VC 006M	1.3.6.1.4.1.9328.50.6.3294			Yes	2.5	120	240		0.730469	SOFT	512,512	FFS,FFP	437,436		60	M		
4	SD VC 008M	1.3.6.1.4.1.9328.50.6.4247			Yes	2.5	120	240		0.681641	SOFT	512,512	FFS,FFP	437,424		50	M		
5	SD VC 010M	1.3.6.1.4.1.9328.50.6.5503			Yes	2.5	120	240		0.755859	SOFT	512,512	FFS,FFP	457,465		70	M		
6	SD VC 012M	1.3.6.1.4.1.9328.50.6.6267			No	2.5	120	240		0.726562	SOFT	512,512	FFS,FFP	410,423		60	M		
7	SD VC 015M	1.3.6.1.4.1.9328.50.6.8321			Yes	2.5	120	240		0.789062	SOFT	512,512	FFS,FFP	420,432		70	M		
8	SD VC 016M	1.3.6.1.4.1.9328.50.6.8943			Yes	2.5	120	240		0.703125	SOFT	512,512	FFS,FFP	441,460		70	M		
180	1.3.6.1.4.1.9328.50.4.0001	1.3.6.1.4.1.9328.50.4.0001	Exception	No	Yes	1	120	280		0.78125	B30F	512,512	FFS,FFP	70,604		59	F	Siemens Sensation 64	24ACRIN_Colo_IRB2415-04
181	1.3.6.1.4.1.9328.50.4.0040	1.3.6.1.4.1.9328.50.4.0040	Exception	No	Yes	1	120	120		0.80078125	B30F	512,512	FFP	604		71	M	Siemens Sensation 16	W/O CONTRAST PRONE
182	1.3.6.1.4.1.9328.50.4.0042	1.3.6.1.4.1.9328.50.4.0042	Validated	No	Yes	1	120	60	FT, <contrast	0.5703125	B30F	512,512	FFS,FFP	508,501		52	F	Siemens Sensation 16	W/O CONTRAST PRONE
183	1.3.6.1.4.1.9328.50.4.0080	1.3.6.1.4.1.9328.50.4.0080	Validated	No	Yes	1	120	100	FT, <contrast	0.73046875	B30F	512,512	HFS,HFP	527,531		51	M	Siemens Sensation 64	Abdomen^ACRIN_COLON_USE (Adult)
184	1.3.6.1.4.1.9328.50.4.0123	1.3.6.1.4.1.9328.50.4.0123	Exception	No	Yes	1	120	120		0.859375	B30F	512,512	HFS,HFP	682,657		65	M	Siemens Sensation 16	JN/SMC 1619172 PRONE CO2 CONTRAST
185	1.3.6.1.4.1.9328.50.4.0127	1.3.6.1.4.1.9328.50.4.0127	Validated	No	Yes	1	120	240		0.9765625	B30F	512,512	FFS,FFP	620,626			M	Siemens Sensation 16	Abdomen^3_COLONOGRAPHY
186	1.3.6.1.4.1.9328.50.4.0136	1.3.6.1.4.1.9328.50.4.0136	Validated	No	Yes	1	120	120		0.78125	B30F	512,512	FFS,FFP	565,589			M	Siemens Sensation 16	CT VIRTUAL COLONSCOPY SCREENING
187	1.3.6.1.4.1.9328.50.4.0152	1.3.6.1.4.1.9328.50.4.0152	Validated	No	Yes	1	120	120	FT, <contrast	0.78125	B30F	512,512	FFS,FFP	501,501			M	Siemens Sensation 16	CT VIRTUAL COLONSCOPY SCREENING
188	1.3.6.1.4.1.9328.50.4.0156	1.3.6.1.4.1.9328.50.4.0156	Validated	No	Yes	1	120	141		0.65234375	B30F	512,512	HFS,HFP	489,517		63	F	Siemens Sensation 64	Abdomen^14_SUPINE_COLON (Adult)
189	1.3.6.1.4.1.9328.50.4.0175	1.3.6.1.4.1.9328.50.4.0175	Validated	No	Yes	1.25	120	140		0.78125	STANDARD	512,512	FFS,FFP	533,579		54	M	GE Lightspeed 16	4.6 COLONOSCOPY (ACRIN) DR.IYER
190	1.3.6.1.4.1.9328.50.4.0259	1.3.6.1.4.1.9328.50.4.0259	Validated	No	Yes	1.25	120	140		0.78125	STANDARD	512,512	FFS,FFP	550,540		84	M	GE Lightspeed 16	6.10 CT COLONOGRAPHY
191	1.3.6.1.4.1.9328.50.4.0264	1.3.6.1.4.1.9328.50.4.0264	Validated	No	Yes	1	120	120	FT, <contrast	0.61328125	B30F	512,512	HFS,HFP	391,401		58	M	Siemens Sensation 64	Abdomen 14_SUPINE_COLON (Adult)
192	1.3.6.1.4.1.9328.50.4.0269	1.3.6.1.4.1.9328.50.4.0269	Validated	No	Yes	1.25	120	140		0.78125	STANDARD	512,512	FFS,FFP	545,533		62	M	GE Lightspeed 16	4.6 COLONOSCOPY (ACRIN) DR.IYER
193	1.3.6.1.4.1.9328.50.4.0272	1.3.6.1.4.1.9328.50.4.0272	Validated	No	Yes	1.25	120	140		0.664062	STANDARD	512,512	FFS,FFP	464,477		54	F	GE Lightspeed 16	4.6 COLONOSCOPY (ACRIN) DR.IYER
194	1.3.6.1.4.1.9328.50.4.0347	1.3.6.1.4.1.9328.50.4.0347	B	No	Yes	1	120	120		0.798828125	B30F	512,512	HFS,HFP	602,616		54	M	Siemens Sensation 64	Abdomen^14_SUPINE_COLON (Adult)
195	1.3.6.1.4.1.9328.50.4.0514	1.3.6.1.4.1.9328.50.4.0514	Exception	No	Yes	1	120	280		0.859375	B30F	512,512	FFP	576		60	M	Siemens Sensation 64	Abdomen^24ACRIN_Colo_IRB2415-04 (Adult)

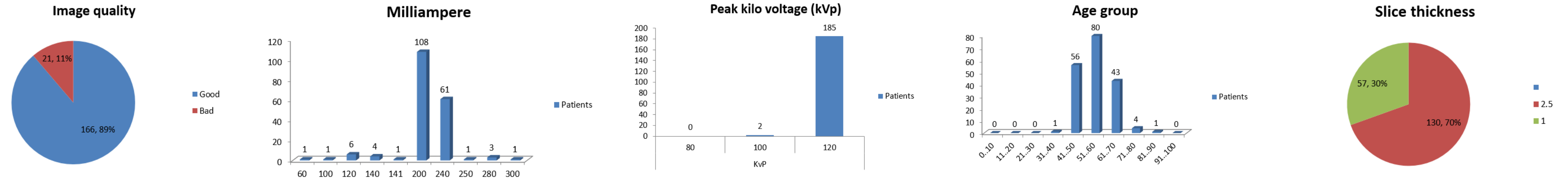


Fig. 4: Number of datasets classified based on various CT image acquisition parameters. The classified datasets are used for empirical testing based on the parameters of interest

Data Analysis (2/4) – Analyze the quality

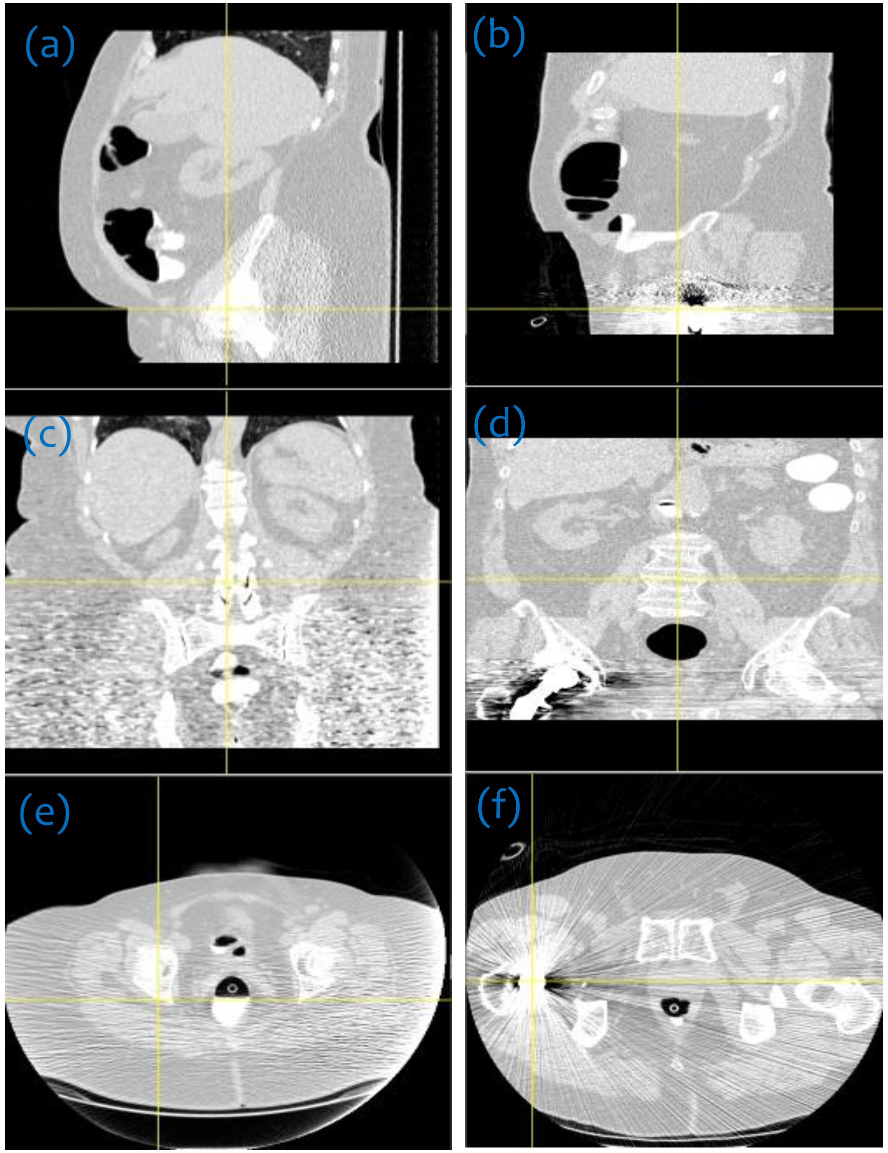


Fig 4: Bad diagnostic quality in abdominal CT images (Sagital, coronal and axial views respectively (a, c, e). a) Incomplete air insufflation, c, d) Patient too large and outside the scan field of view, c, d) Streak artifact.

- 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?

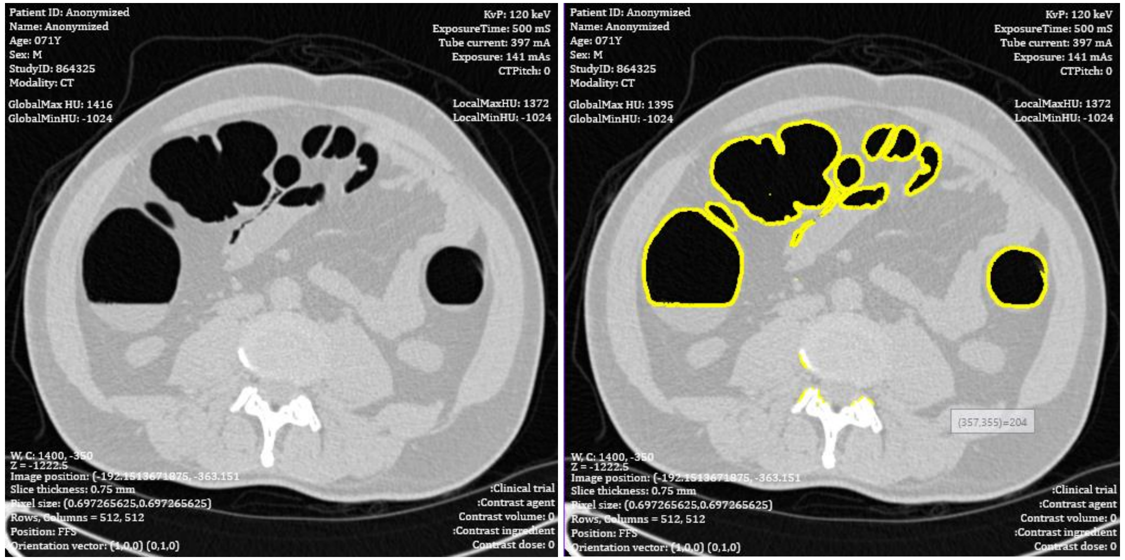


Fig 5: Accurate tissue details in abdominal CT (Manjunath K N *et. al.*, 2016,[10.1166/jmihi.2016.1786](https://doi.org/10.1166/jmihi.2016.1786))

Data analysis (3/4) – Image validation

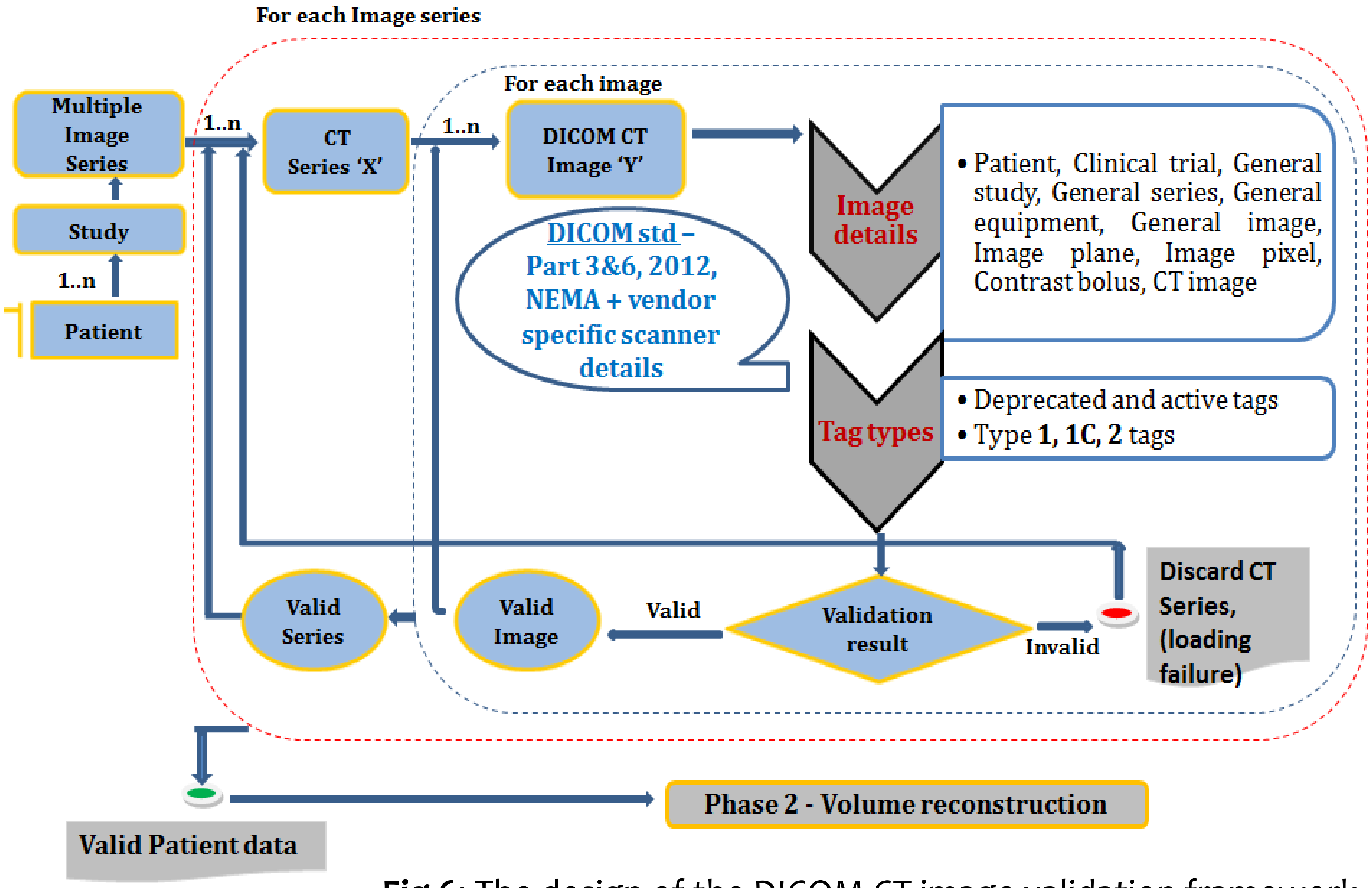


Fig 6: The design of the DICOM CT image validation framework

- 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

Data analysis (4/4) – Features extraction

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 analysis (1/6) – Open source software

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 (Frauhhofer 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:
https://pydicom.github.io/pydicom/stable/getting_started.html

#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
root=ET.Element("root")
doc=ET.SubElement(root,"doc")

#Read the DICOM file
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:
      ET.SubElement(doc,"DataElement", Description=elem.description().strip(),
      tag=str(elem.tag).strip(), VR=str(elem.VR).strip(), VM=str(elem.VM).strip(),
      Value=str(elem.value).strip())
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
```

Tools for analysis (3/6) – Exporting DICOM to XML


```xml
<?xml version="1.0"?>
<ROIContour xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
 <PatientID>SD VC-125</PatientID>
 <SOPClassUID>1.2.840.10008.5.1.4.1.1.2</SOPClassUID>
<BlobInformation>
  <BlobDetails>
   <PerpendicularLines>
    <PerpendicularLine><FirstPoint> <X>82</X> <Y>288</Y>  </FirstPoint>
     <SecondPoint> <X>82</X> <Y>284</Y> </SecondPoint>
     <DistanceBetweenTwoPoints>0</DistanceBetweenTwoPoints>
    </PerpendicularLine>
   </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>
<Size><Width>35</Width><Height>41</Height></Size><X>73</X><Y>269</Y><Width>35</Width>
    <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"
            }
        }
}
```


Exported DICOM data (structured format) into XML and JSON format (semi structured)

Tools for analysis (4/6) – In C#

1. Accord.NET provides statistical analysis, machine learning, image processing and computer vision methods for .NET applications (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.

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

Souza, C.R., " A Tutorial on Principal Component Analysis with the Accord.NET Framework ". Department of Computing, Federal University of Sao Carlos. arXiv:1210.7463. Technical Report, 2012

Tools for analysis (5/6) – Coding in C#

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

Decision tree sample code

```
// Creates a matrix from the entire source data table
double[,] table = (dgvLearningSource.DataSource as DataTable).ToMatrix(out columnNames);

// Get only the input vector values (first two columns)
double[][] inputs = table.GetColumns(0, 1).ToJagged();

// Get only the output labels (last column)
int[] outputs = table.GetColumn(2).ToInt32();


// Specify the input variables
DecisionVariable[] variables =
{
    new DecisionVariable("x", DecisionVariableKind.Continuous), new DecisionVariable("y", DecisionVariableKind.Continuous),
};

// Create the C4.5 learning algorithm
var c45 = new C45Learning(variables);

// Learn the decision tree using C4.5
tree = c45.Learn(inputs, outputs);

// Show the learned tree in the view
decisionTreeView1.TreeSource = tree;

// Get the ranges for each variable (X and Y)
DoubleRange[] ranges = table.GetRange(0);

// Generate a Cartesian coordinate system
double[][] map = Matrix.Cartesian(Vector.Interval(ranges[0], 0.05), Vector.Interval(ranges[1], 0.05));

// Classify each point in the Cartesian coordinate system
double[,] surface = map.ToMatrix().InsertColumn(tree.Decide(map));

CreateScatterplot(zedGraphControl2, surface);
```

Tools for analysis (6/6) – Coding in C#

Steps:

- Create the training data (numerical) (**X** 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

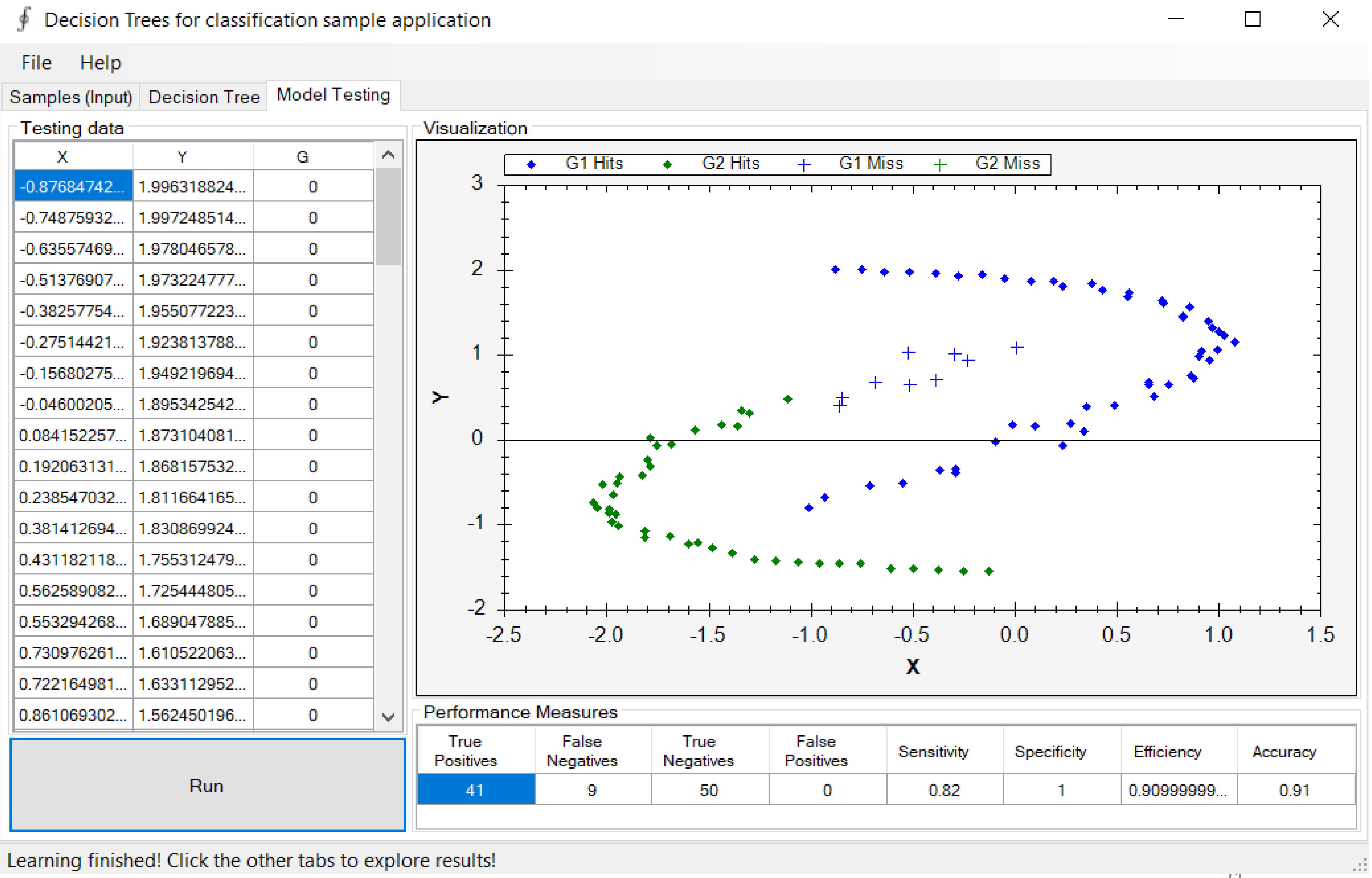


Fig 7: The UI of DT showing the scatter plot of **X** and **Y**

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

0008,0090	Referring Physician's Name	PN	1		0
0008,1030	Study Description	LO	1	RT SIMULATION	14
0008,1090	Manufacturer's Model Name	LO	1		0
0008,1110	**Referenced Study Sequence**	**SQ**	**1**	**FE FF 00 E0 FF FF FF FF 08 00 50 11 18 00 00 00 31 2E 32 2E**	**120**
0010,0010	Patient's Name	PN	1		0
0010,0020	Patient ID	LO	1	HNSCC-01-0003	14
0010,0030	Patient's Birth Date	DA	1		0
0010,0040	Patient's Sex	CS	1	M	2
0012,0062	Not in Dictionary	UN	1	59 45 53 20	4
0012,0063	Not in Dictionary	UN	1	50 65 72 20 44 49 43 4F 4D 20 50 53 20 33 2E 31 35 20 41 6E	46

- **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 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

- **CAD_e** and **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

- NCI,www.cancerimagingarchive.net,(2016).NationalCancerInstitute(NCI).[Online]Availableat:https://public.cancerimagingarchive.net/ncia/login.jsf.
- Philips Medical Systems, The Netherland (2014), In: DICOM conformance statement technical report PIIOffc.0001414. Available via Philips website. https://www.usa.philips.com/healthcare/resources/support-documentation/dicom-computed-tomography. Accessed 06 April 2014
- Siemens, PA, USA (2020). In: Clinical application guide. Available via website. https://www.healthcare.siemens.co.in/computed-tomography/options-upgrades/clinical-applications/syngo-ct-colonography. Accessed 12 June 2020
- Siegel,R.(2013).Colorectalcancerfactsandfigures2011-2013.[Online]Atlanta:AmericanCancerSociety,pp.1-5.Availableat:http://www.cancer.org.
- SiemensMedicalSolutions,(2012).Clinicalapplicationguide.[Online]Availableat:www.medical.siemens.com[AccessedMarch.2012].
- The DICOM chapter3(2012). PS3.3. [Online] Virginia: NEMAUSA, pp.390–1174. Available at: http://dicom.nema.org/standard.html[Accessed19June.2016].
- Kalender,WA. (2006).X-raycomputedtomography.Phys.Med.Biol,51(13),pp.29-43.

Acknowledgements