Oncology Medical Image Database (OMI-DB):

A curated platform for facilitating big data analytics in healthcare

James A. Leighs[1], M. Patel[1] and M. D. Halling-Brown[1]

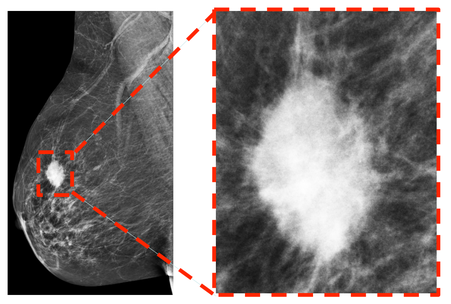
[1]Scientific Computing, Royal Surrey County Hospital, Guildford, UK

Introduction

Routine use of imaging in healthcare is continuously increasing and has been implemented in national screening programs such as Breast Cancer. Medical imaging collects a wealth of invaluable information, representing an estimated dataset of over 55 petabytes that currently remains largely unexploited. Presented here is an Oncology Medical Image Database (OMI-DB), developed as a platform for facilitating analytical healthcare research. This is an area with huge potential, particularly when placed within the context of recent cancer statistics, such as the likelihood that 1 in 2 people will be diagnosed with cancer. OMI-DB is a unique resource given the number of cases and volume of additional curated information. The data contained holds great potential for aiding machine learning studies, such as for predicting the results of invasive pathological tests.

Image Database

OMI-DB was designed purely to support medical research and contains anonymised unprocessed and processed images, annotations, pathology data and expert-determined ground truths describing regions of interest (ROIs), as shown in the example below. Currently OMI-DB holds mammographic cases but can easily be expanded to other modalities. A fully automated image retrieval system has been developed, collecting data from 4 breast screening sites in London, Cambridge, Dundee and Guildford.



(left) an example mammogram from OMI-DB along with an expert-marked ROI

OMI-DB Statistics

# Patients recruited	8579
# Centres recruiting	4
# Cancer cases	6358
# Cancer cases marked	3912
# Benign cases	541
# Normal Cases	1680
# Images	142,056

Feature Extraction

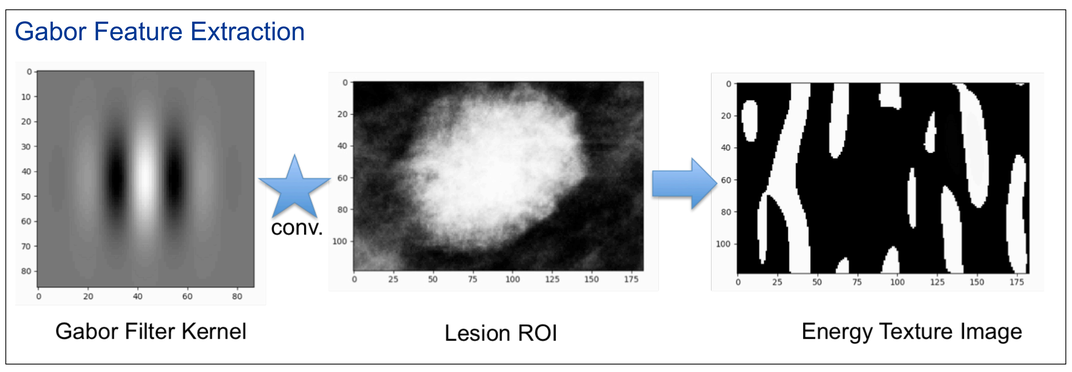
FeaTPy Toolkit:

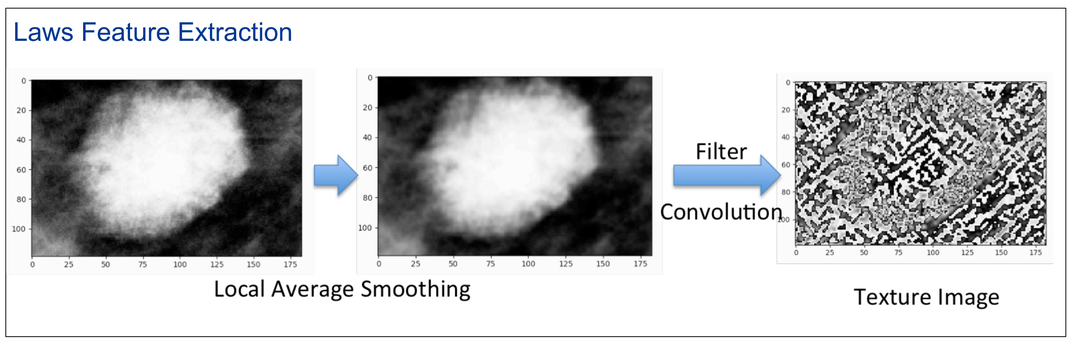
A custom Python library has been written to extract first, second and higher order quantitative imaging features from images received by OMI-DB. The feature classes including Gabor, Laws and Run Length features, produce the 321 individual features summarised in the table below. The feature classes have been engineered to extract data from a variety of scales and orientations.

Texture Features:

Six feature classes extract data from images, of which two perform matrix convolution, Gabor and Law's. These higher order texture features are often used in image recognition and classification. Both construct mathematical filters, before convoluting them with the original image. The below examples show some of the different types and scales of data extracted using these features.

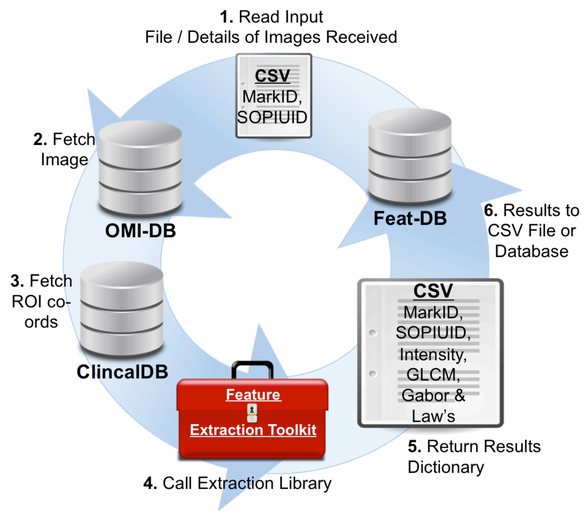
Intensity Stats	Gabor	Laws	Run Length	GLCM	LBP
Mean	Mean	Levels	Short Run Emphasis	Contrast	Distances = 4
Median	Standard Deviation	Edges	Long Run Emphasis	Dissimilarity	
Variance & Range	Skewness	Spots	Gray Level Non-uniformity	Homogeneity	
Skewness & Kurtosis	Frequencies = 7	Ripples	Run Length Non-uniformity	Energy	
RMS	Angles = 7	Additional Combinations = 6	Run Percentage	Correlation	
Entropy			Low Gray Level Run Emphasis	ASM	
Minimum			High Gray Level Run Emphasis	Angles = 4	
Maximum			Directions = 2	Distances = 5	
1st, 10th, 90th & 99th Percentile					
14	**147**	**22**	**14**	**120**	**4**





Pipeline:

The custom feature extraction toolkit has been implented within a data extraction pipeline for automated and on-demand feature extraction. When new data is received by OMI-DB, the feature pipeline is run and the full set of image features are extracted from each image and corresponding lesion ROI.



(above) a schematic of the automated feature extraction pipeline (1) image details received (2) image data is transferred from OMI-DB (3) ROI details are queried from a local clinical database (4) FeaTPy feature extraction toolkit is called (5) feature data is returned (6) features are written to Feat-DB or local CSV file

Feat-DB:

FeaTPY currently extracts 321 features from each image and ROI, including an add-on to adjust ROI size and accept CAD segmentation co-ordinates. There are >100,000,000 features contained within Feat-DB., including quantitative imaging features, additional pathological data and other annotations.

Conclusions

The main focus has been on developing automated image collection and feature extraction, resulting in a large, curated database of images and associated feature data. Preliminary efforts to analyse potential for predicting results of pathological tests have showed promise. Law's and Gabor features have shown to be the most significant biomarkers. Two diagnostic/prognostic indicators have been investigated. Disease grade, a measure of the rate at which the cancer is likely to spread, has been predicted with accuracies up to 65%. Molecular subtype, a diagnostic classifier used to target personalised treatment plans, has been predicted up to 82% accuracy. Further work is ongoing, aiming to both improve classification rates and predict further pathological results. Success in this area has potential benefits in prioritsing and streamlining patient care by reducing wait times for results. Impacts of this could be significant, given early diagnosis is seen as one of the most important factors for survival.

Acknowledgements

OMI-DB development was undertaken as part of the OPTIMAM project and supported by CR-UK

Corresponding Author Email: mishalpatel@nhs.net