OPTIMAM Mammography Imaging Database (OMI-DB): A Valuable Dataset to Fuel Machine Learning Research

Mishal N. Patel, PhD, Royal Surrey County Hospital; Ken Young, PhD; Mark Halling-Brown, PhD

Background

The importance of medical imaging for clinical decision-making has been steadily increasing over recent years, with greater emphasis on the utilization of medical images for preclinical decision making as well as a drive towards quantification of imaging findings by adopting quantitative imaging features. This demand led to the design and implementation of a flexible image repository, which retrospectively collects images and data from multiple sites throughout the UK. The Oncology Medical Image Database (OMI-DB)\(^1,2\) was created to provide a centralized, fully annotated dataset for research. The database contains both processed and unprocessed images, associated data, annotations and expert-determined ground truths. Furthermore, calculated and derived data are obtained from the images resulting in a huge number of quantitative image features, properties and characteristics that can be useful for classification, CAD (Computer Aided Detection) and radiomics/radiogenomics applications.

A research dataset of this scale, which includes original normal and subsequent malignant cases along with expert derived ground truths, clinical annotations and quantitative features, is currently unique. These data provide a powerful resource for future research and has initiated new research projects, amongst which, is the quantification of normal cases by applying a large number of quantitative imaging features, with a priori knowledge that eventually these cases develop a malignancy. This paper describes OMI-DB collection systems and tools and discusses the prospective applications of having such a rich dataset for future research applications.

Evaluation

Current efforts have focused on collecting mammographic images, however the system has been designed to-be easily extended to any modality. The database contains unprocessed and processed images, associated data and expert-determined ground truths. The need to collect unprocessed images is important for many research applications. Unprocessed images are essential for studies on different types of image processing and computer aided detection. The pixel values and the noise in the unprocessed images also have a direct link back to the physics of the image formation. Once the images are processed this link is lost and cannot be recovered. Unprocessed images can also be used to study the effect of different design parameters on clinical performance\(^4,5\). The process of collection, annotation and storage (Figure 1) is fully automated and adaptable and has been described extensively elsewhere\(^1\). Currently, the associated data comprises radiological, clinical and pathological information extracted from the National Breast Screening System (NBSS). However the collection system is flexible enough to allow new data sources. A fully automated feature extraction framework has been implemented, which extracts both first and second order statistics at time of collection. Furthermore, a web-enabled, remotely accessible software application MedXViewer (www.medxviewer.com) has been developed which allows radiologists to view cases, annotate clinical features and participate in observer studies.
Currently the full collection system is deployed at three sites and other partial collection processes are in place at two other sites. The database statistics are summarized in Table 1.

**Table 1**

<table>
<thead>
<tr>
<th>OMI-DB Mammography statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>Ground Truth Defined</td>
</tr>
</tbody>
</table>

When loading the images into the repository, all relevant DICOM tags are extracted to allow a searchable index to be produced. Additionally, CAD predictions (commercial software) and breast density are determined and inserted into the database. Expert-determined ground truths are then obtained from our panel of readers (utilizing the MedXViewer software) that indicate the relevant ROI and other attributes, such as lesion type and conspicuity. Further annotations are obtained from associated data sources. For each image and ROI over 285 quantitative image features are calculated resulting in over 10M features and descriptors in the database. The quantity of additional annotations obtained is large, and includes information on screening history, previous occurrences of cancer, biopsy results and surgical procedures. Initially collection was only for 2D mammography images, however as the collection has progressed we have been able to add additional modalities where available for each case (Table 2), at present ultrasound is collected for each case along with the 2D, tomosynthesis and MRI for a small subset.

**Table 2**

<table>
<thead>
<tr>
<th>OMI-DB Mammography multimodality statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>Ground Truth Defined</td>
</tr>
</tbody>
</table>
Discussion

Figure 2 shows the collection of data plotted over time. Highlighted are the key milestones since collection began. Collection of data for the OMI-DB has been ongoing since early 2011, routine breast screening in the UK is every three years and thus during 2014 (denoted by the orange box Figure 2) the opportunity to collect sequential data became a reality. The collection system can bow retrospectively collect all prior images for new malignancies and collect new screening episodes for existing normal and benign cases. Moreover, during the collection timeline over 567 interval cancers (denoted by the red yellow in Figure 2) have been collected. Interval cancers (cancers arising between screening episodes) are a key measure to monitor screening performance. Radiological analysis of the imaging features and retrospective classification are an important educational tool for readers to improve individual performance.

Figure 2

![Image Collection Over Time](image)

**Figure 2**: Image collection over time for the OMI-DB, annotated are key developments during the last four years since commencement of collection.

This availability of previous screening events and interval cancers opens up a wealth of potential Machine learning research applications evaluating whether an abnormality could have been detected on the previous screening films via alterations to processing or perception. We have begun various research projects and collaborations dedicated to utilizing this data. These projects include perceptions studies, which challenge readers with previous normal screens of subsequent malignancies, which have been reprocessed using varying algorithms to identify if earlier detection could have been possible.

A research dataset such as the OMI-DB can provide a powerful resource for future research; the sequential normal cases can be analyzed using quantitative imaging features, with a priori knowledge that some years later these cases develop a malignancy. We have begun work extracting the imaging features for this dataset along with the associated annotations from the OMI-DB. We are using this data to build expert models in an attempt to predict high-risk patients from routine screening based on their quantitative image profile, which could facilitate early stratification when used in conjunction with experts.
An ideal future application of this rich dataset would be the comprehensive quantification of tumour phenotypes by applying the large number of quantitative image features extracted and correlating these data with gene expression profiles produce prognostic radiomic signature. Such studies could be helpful for improved personalized diagnosis, prognosis, and assessment of treatment response.

Conclusion

A valuable database have been developed which holds both processed and unprocessed mammographic images. The provision of unprocessed images enables a multitude of potential research applications. The availability of associated data and expertly determined ground truth along with computational image feature extraction can facilitate many data intensive research applications. Such as building predictive models to aid image classification, treatment response assessment as well as to identify prognostic imaging biomarkers.

References


Keywords

mammography, image database, breast cancer, machine learning