CBIS-DDSM: A CURATED MAMMOGRAPHY DATA SET FOR USE IN COMPUTER-AIDED DETECTION AND DIAGNOSIS RESEARCH

Daniel L. Rubin, MD, MS
Department of Biomedical Data Science, Radiology, and Medicine (Biomedical Informatics)
Stanford University

Motivation

- Many different AI/ML applications for mammography are being developed
  - Detection of suspicious lesions (CADe)
  - Diagnosis of cancer (CADx)
- Algorithm performance evaluated on different datasets
  - Private data sets
  - Unspecified subsets of public databases
  - Variable dataset sizes
- Not possible to directly compare the performance of methods or to replicate prior results

Acknowledgements

- Rebecca Lee, PhD
- Berkman Sahiner, PhD
- Justin Kirby, John Freymann, TCIA team
- Funding support
  - NCI QIN grants
  - U01CA142555, U01CA190214, U01CA187947

Some CADe systems reported in the literature

<table>
<thead>
<tr>
<th>Authors</th>
<th>Size of Data set (Cases)</th>
<th>Public or Private Data</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>False Positives Per Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kononenko et al.</td>
<td>50</td>
<td>Public (MMID)</td>
<td>NA</td>
<td>99%</td>
<td>1</td>
</tr>
<tr>
<td>Madgwick et al.</td>
<td>56</td>
<td>Public (MAMET)</td>
<td>NA</td>
<td>81%</td>
<td>2.2</td>
</tr>
<tr>
<td>Lee et al.</td>
<td>28</td>
<td>Public (MAMET)</td>
<td>NA</td>
<td>96%</td>
<td>1</td>
</tr>
<tr>
<td>Li et al.</td>
<td>14</td>
<td>Private (MMID)</td>
<td>NA</td>
<td>99%</td>
<td>3.21</td>
</tr>
<tr>
<td>Bao et al.</td>
<td>60</td>
<td>Private (MMID)</td>
<td>NA</td>
<td>89%</td>
<td>0.61</td>
</tr>
<tr>
<td>Heret et al.</td>
<td>83</td>
<td>Private (MMID)</td>
<td>NA</td>
<td>96%</td>
<td>0.26</td>
</tr>
<tr>
<td>Wang et al.</td>
<td>103</td>
<td>Private (MMID)</td>
<td>NA</td>
<td>96%</td>
<td>3.0 per case</td>
</tr>
<tr>
<td>Theurer et al.</td>
<td>120</td>
<td>Private (MMID)</td>
<td>NA</td>
<td>96%</td>
<td>3.0 per case</td>
</tr>
<tr>
<td>Safar et al.</td>
<td>127</td>
<td>Public (DDSM)</td>
<td>NA</td>
<td>91%</td>
<td>NA</td>
</tr>
<tr>
<td>Cui et al.</td>
<td>230</td>
<td>Public (DDSM)</td>
<td>NA</td>
<td>90.8%</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Some CADx systems reported in the literature

<table>
<thead>
<tr>
<th>Authors</th>
<th>Size of Data set (Cases)</th>
<th>Public or Private Data</th>
<th>Classification Accuracy</th>
<th>At*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rakhmanov et al.</td>
<td>25</td>
<td>Private</td>
<td>88%</td>
<td>NA</td>
</tr>
<tr>
<td>Hu et al.</td>
<td>15</td>
<td>Private</td>
<td>NA</td>
<td>0.94</td>
</tr>
<tr>
<td>Paragamyan et al.</td>
<td>54</td>
<td>Public (MMID) and Private</td>
<td>95%</td>
<td>NA</td>
</tr>
<tr>
<td>Madgwick et al.</td>
<td>35</td>
<td>Public (MAMET)</td>
<td>82.1%</td>
<td>0.05</td>
</tr>
<tr>
<td>Suber et al.</td>
<td>102</td>
<td>Private</td>
<td>NA</td>
<td>0.03</td>
</tr>
<tr>
<td>Ying et al.</td>
<td>403</td>
<td>Private</td>
<td>NA</td>
<td>0.77</td>
</tr>
<tr>
<td>Gurung et al.</td>
<td>381</td>
<td>Private</td>
<td>88.6%</td>
<td>NA</td>
</tr>
<tr>
<td>Gürsell et al.</td>
<td>78, 65</td>
<td>Private, Public (MMID)</td>
<td>91.8%, 90.1%</td>
<td>NA</td>
</tr>
<tr>
<td>Que et al.</td>
<td>509</td>
<td>Private</td>
<td>77.14%</td>
<td>0.83</td>
</tr>
<tr>
<td>Cho et al.</td>
<td>200</td>
<td>Public (DDSM)</td>
<td>NA</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Standard image datasets are being developed and popular

- ImageNet (14 million images from 27 categories)
- Mixed National Institute of Standards and Technology (MNIST) database (hand-written digits)
- ChestX-ray8 and OpenI (Chest X-rays)
- DeepLesion (CT)
- Digital Database for Screening Mammography (DDSM)
DDSM is limited for evaluating AI algorithms

- Images saved in non-standard format compressed files, difficult to use
- Image metadata are fragmented, unwieldy to access/use
- ROIs are very coarse, limited value for lesion localization
- Cases not curated into training/testing subsets

CBIS-DDSM (Curated Breast Imaging Subset of DDSM)

- An updated, standardized subset of DDSM
- Curated
  - Images in standard format (DICOM)
  - Unified metadata in single files
  - Improved ROIs
  - Training/testing subsets
- Can serve as a common dataset for comparing performance of AI/ML algorithms (CADE, CADx; NB: not segmentation)

CBIS-DDSM: Images

- DDSM scanned film mammography studies
  - Source: MGH, Wake Forest Univ, Sacred Heart Hospital, WUSTL
  - Image labels: Normal, benign, and malignant (latter verified by pathology)
  - Decompressed and converted to DICOM
  - Case selection and curation by expert mammographer
  - Removal of questionable mass cases and cases containing PHI
  - Convenience images: focused crops of abnormalities based on ROI bounding box

CBIS-DDSM: ROIs

- ROIs for masses
  - Mass ROIs refined by automated segmentation
  - Mammographer verification in 188 cases (Dice 0.8 ± 0.1)
  - ROIs are binary mask images delineating the abnormality

Annotations: Parsed semantic features

- Patient age
- BI-RADS descriptors (mass shape, mass margin, calcification type, calcification distribution, and breast density)
- BI-RADS final assessment category (0 to 5)
- Rating of the subtlety of abnormality (1 to 5)
- Type of abnormality (mass or calc)
- Annotations formatted in CSV format similar to modern computer vision data sets

Summary: Preparation of CBIS-DDSM
Training/testing splits of cases

- 1,644 cases: 753 calc cases, 891 mass cases
- Separate sets of cases for training (80%) and testing (20%)
  - Enables all researchers to use same cases for these tasks
- Separate splits for mass and calc cases
  - Contain full range of label values (BI-RADS codes)
  - Equal distribution in training and testing cases

<table>
<thead>
<tr>
<th>Design Cases</th>
<th>Malignant Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcification Training Set</td>
<td>329 cases (512 abnormalities)</td>
</tr>
<tr>
<td>Calcification Test Set</td>
<td>85 cases (12 abnormalities)</td>
</tr>
<tr>
<td>Mass Training Set</td>
<td>355 cases (357 abnormalities)</td>
</tr>
<tr>
<td>Mass Test Set</td>
<td>117 cases (135 abnormalities)</td>
</tr>
</tbody>
</table>

Limitations of CBIS-DDSM

- **Limited dataset size** (1,644 cases reasonable for quantitative imaging; somewhat small for deep learning)
- **Film-screen images**: FFDM and tomosynthesis are modern techniques
  - Could produce similar dataset if such images become publicly available
- **Segmentations** of lesions are improved over original DDSM, but not all hand-drawn

Obtaining CBIS-DDSM

https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM

CBIS-DDSM is a popular dataset...

CBIS-DDSM is the 3rd most popular dataset in the time it has been in TCIA; selected in search 22,557 times since being added to TCIA

Download history since 2016

9-19 people downloading CBIS-DDSM per week
Summary

- CBIS-DDSM is a curated set of benign/malignant mammography cases
  - Annotations of mass lesions and calcifications
  - Provides data formats suitable for AI/ML development and testing
- Potentially useful as a common dataset for evaluating and comparing CADe/CADx methods
- Preferably similar dataset will be built using FFDM/tomosynthesis images if public data becomes available

Contact info:
dlrubin@stanford.edu