Systematic Evaluation and Validation

Dr. Debdoot Sheet
Assistant Professor
Department of Electrical Engineering
Indian Institute of Technology Kharagpur

www.facweb.iitkgp.ernet.in/~debdoot/

Contents

- Datasets and baselines
- Prospective vs. retrospective experiments
- Bias and variance
- Sample sufficiency
- Evaluating segmentation
- Evaluating classification
- Receiver operating characteristics
- Folded Cross-validation

Datasets and Baselines

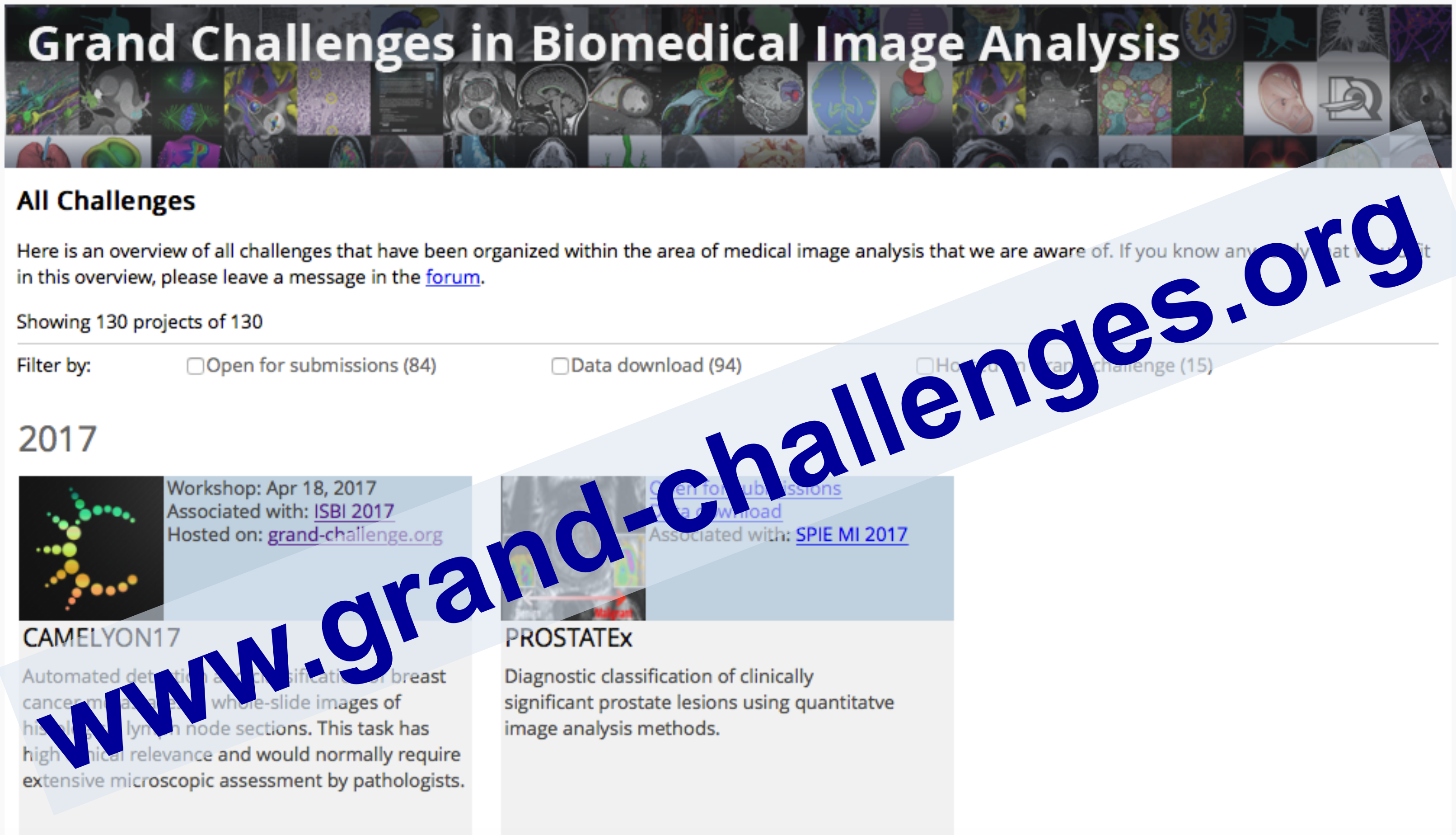
Playground

- Standard evaluation
  - Anonymized
  - Same modality or set of modalities
  - Cohort collected
    - Similar disease
    - Similar genetic makeup
    - Similar age-group and demography
  - Equal number of subjects per pathological group
    - Control / normal may be absent
- Follow-up / prognosis data
  - Similar number for all cases used

Prior Record Holders

- Literature reviews
  - Performance of methods employing same dataset
  - Re-implementation of other methods on standard dataset
  - Same set of metrics used for evaluation
    - Comparing computation times - similar hardware platform (acceleration if any) to be used

Datasets



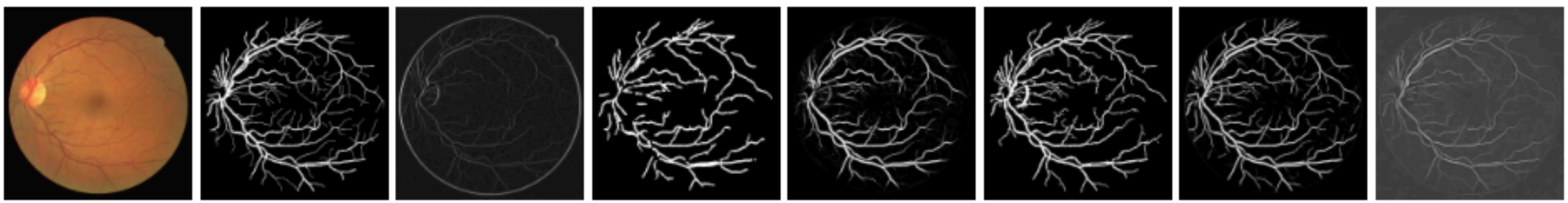
www.grand-challenges.org

Benchmarking on a Dataset

DRIVE: Results Browser

Next Prev Go to 1 Magnification factor: 0.2 ☑display soft classifcation when available

Display the following: ☑input ☑gold standard ☐human observer ☑Chaudhuri ☑Jiang ☑Niemeijer ☑Perez ☑Staal ☑Zana



Results for case 1.

Displayed	Sensitivity	Specificity	Accuracy	Az
1. Input				
2. Gold standard				
3. Chaudhuri	0.276	0.997	0.903	0.950
4. Jiang	0.714	0.949	0.918	
5. Niemeijer	0.719	0.972	0.939	0.944
6. Perez	0.796	0.961	0.939	
7. Staal	0.778	0.971	0.946	0.967
8. Zana	0.773	0.975	0.949	0.942

http://www.isi.uu.nl/Research/Databases/DRIVE/browser.php

Experiments

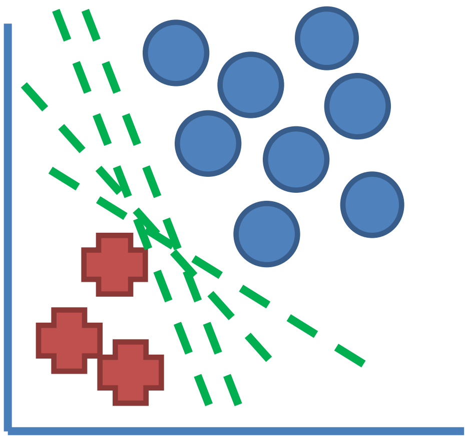
Prospective

- Plan data collection before event occurs
  - Pharmaceutical / drug trials
  - Controlled animal model trials
- Used to test a certain hypothesis
- Pros
  - Class balance
- Cons
  - Regulatory approvals
  - Costly to perform

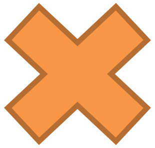
Retrospective

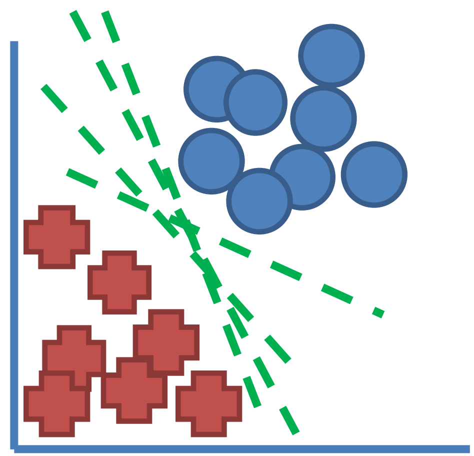
- Data collected as event occurs
  - Epidemic data
  - Imaging modality efficacy studies
- Used to form a hypothesis from observations
- Pros
  - Less expensive and generally free of cost
  - Regulatory approvals hassle free
- Cons
  - Class imbalance

Sampling Issues

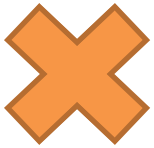


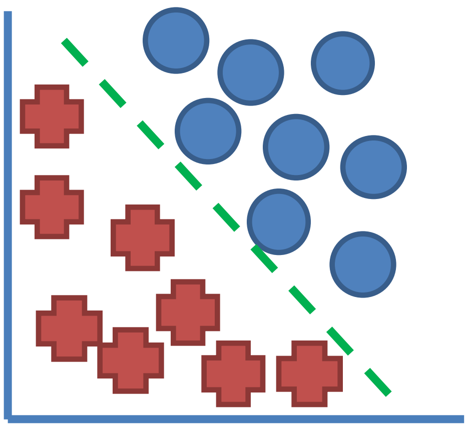
Biased





Unbiased
low variance





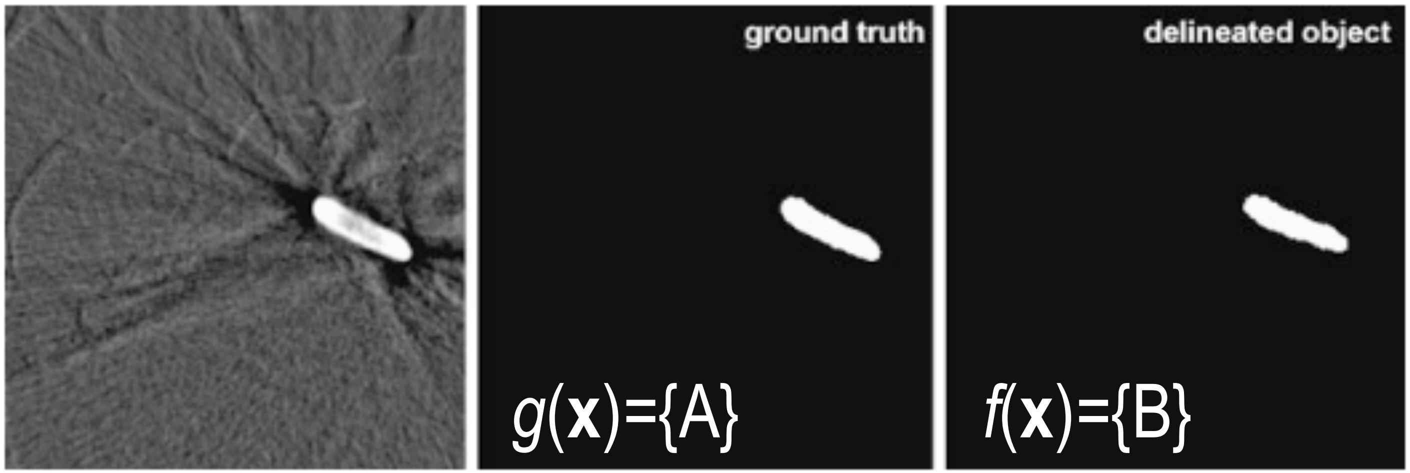
Unbiased
high variance

Ensuring Sample Sufficiency

- Sample sufficiency?
  - Myth for medical image datasets
  - Normal or healthy
    - More cases and samples
  - Abnormal or diseases
    - Rarer a diseases – lesser the samples
    - Require high performance for rare diseases

- Solution
  - Data augmentation during training
  - How?
    - Replicate samples for the weaker class
    - Use rotations, affine transformation, etc.
    - Restrict use of warping on images or applying intensity transformations.

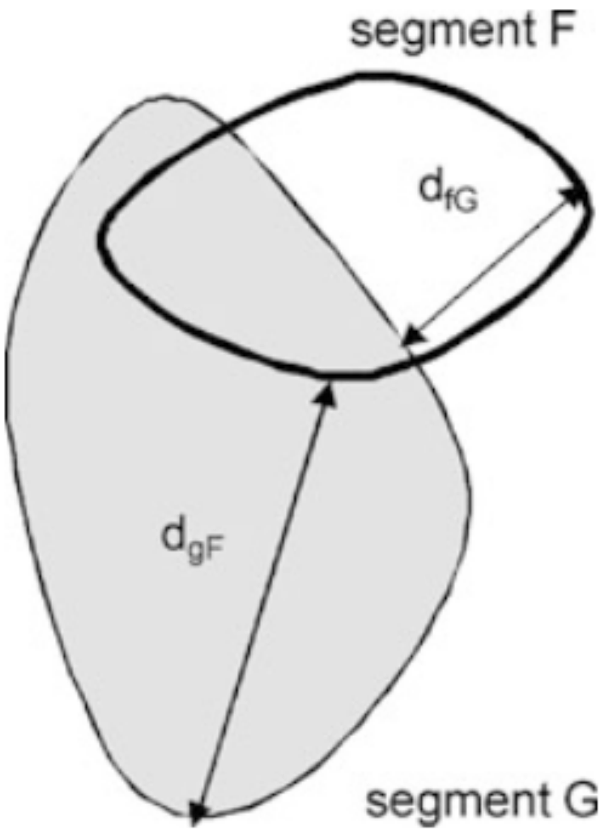
Segmentation



$$O = \frac{|B \cap A^c|}{|A|} \qquad U = \frac{|A \cap B^c|}{|A|} \qquad D = \frac{2|A \cap B|}{|A| + |B|} \qquad J = \frac{|A \cap B|}{|A \cup B|}$$

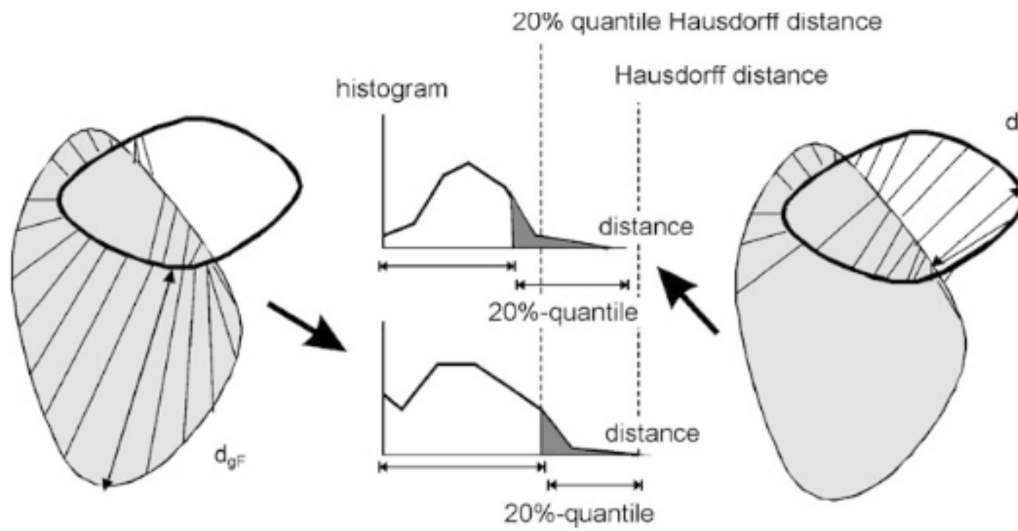
Oversegmentation | Undersegmentation | Dice coefficient | Jaccard coefficient

Segmentation



Hausdorff distance

$$H = \max\left(\inf_{f\in F} d(f,G), \inf_{g\in G} d(g,F)\right)$$

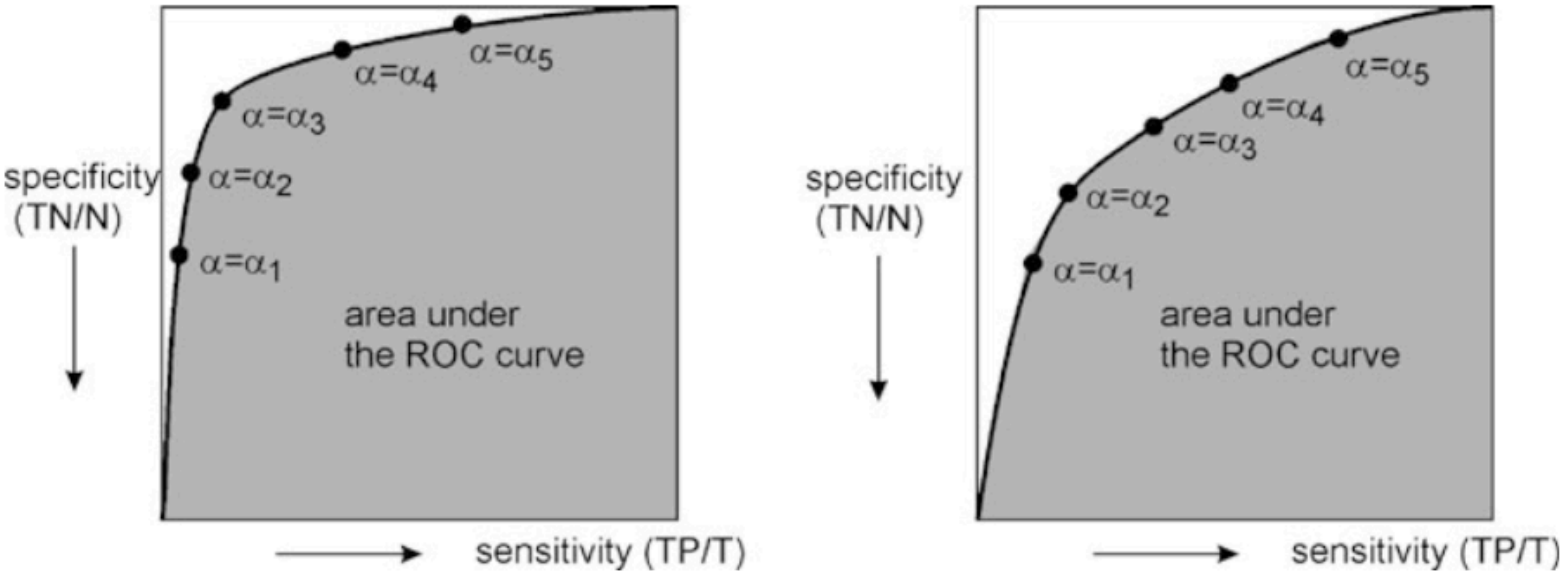


Classification

		Prediction	
		P	N
Ground Truth	P	TP	FN
	N	FP	TN

- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- $Sensitivity = \frac{TP}{TP+FN}$
- $Specificity = \frac{TN}{TN+FP}$
- $Precision = \frac{TP}{TP+FP}$
- $F - score = \frac{2TP}{2TP+FP+FN}$

Receiver Operating Characteristics



Folded Cross-validation

- Folding
  - Creating non-overlapping sample sets
  - Class 0 – N_0 Samples
  - Class 1 – N_1 Samples
  - Folds – k
  - Training samples per fold = $\frac{N_0+N_1}{k}(k-1)$
- Divide samples in k number of bags
  - K-th bag will contain
    - $\frac{N_0}{k}$ samples of Class 0
    - $\frac{N_1}{k}$ samples of Class 1
  - No bags will overlap

Take home message

- K.D. Toennies, *Guide to Medical Image Analysis* [Chap. 13], Advances in Computer Vision and Pattern Recognition, Springer-Verlag, 2012.
- J. Kalpathy-Cramer, H. Mueller, "Systematic Evaluations and Ground Truth", T.M. Deserno (ed.), Biomedical Image Processing, Biological and Medical Physics, Biomedical Engineering, Springer-Verlag, 2011.