Biomedical Informatics 260

Medical Imaging Applications of AI

Lecture 18

Spring 2019

AI Applications

- **What are the motivations for AI applications?**
- **What are the key methods?**
- **What are the types of AI applications?**
- **What are challenges to progress?**

What are the motivations for AI applications?

Key motivations for AI applications

1. Flood of image data
   - Impacts *disease detection*
2. Variation in clinical practice
   - Impacts *diagnosis*
3. Variations in disease in people
   - Impacts *clinical prediction* and *clinical decision making*

1) Flood of image data…

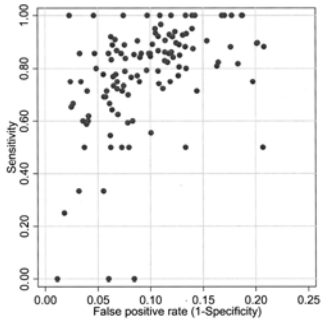
2001 2018

2) Variation in practice

- There are large variations and disparities in care
  (Institute of Medicine, 2001)
- *"Errors and variations in interpretation now represent the weakest aspect of clinical imaging* "*

*Robinson PJ. Radiology's Achilles' heel: error and variation in the interpretation of the Röentgen image. British Journal of Radiology. 1997 Jan 1;70(839):1085–98.

Variable Performance of Radiologists



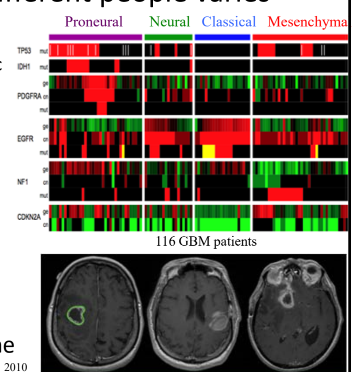
Barlow WE, Chi C, Carney PA, Taplin SH, D'Orsi C, Cutter G, et al. Accuracy of Screening Mammography Interpretation by Characteristics of Radiologists. J. Natl. Cancer Inst. 2004 Jan 15;96(24):1840–50.

3) Variation in disease among people

People (and their diseases) differ…

Disease in different people varies

- Molecular diversity
  - Heterogeneous genomic aberration landscape of individual tumors*
- Phenotypic diversity
  - Variable appearance of lesions on images
- Clinical diversity
  - Patients have different response to treatment
- Ideally we will "profile" disease for personalized medicine



The TCGA Research Network. Cancer Cell. 2010

Realizing "Precision Medicine"

- Patient care often lacks **specificity** ("One size fits all" does not usually apply in medicine)
- There are "**subtypes**" of disease (e.g., many types of "breast cancer" needing specific therapy for each type)
- Precise diagnoses based on "**electronic phenotyping**" and **molecular profiling** enables treatments that are individually tailored to each patient
- Opportunity: Leverage **Big Data** and AI methods to build **prediction models**

"Precision Health"

- A paradigm shift, focusing on **prediction and prevention**, rather than relying exclusively on diagnosis and treatment of existing disease
- **Prevents or forestalls** the development of disease
- Requires accurate methods of prediction based on **monitoring** people's health status
- Opportunity: Like precision health, leverage **Big Data** and AI methods to build **prediction models**

What are the key approaches?

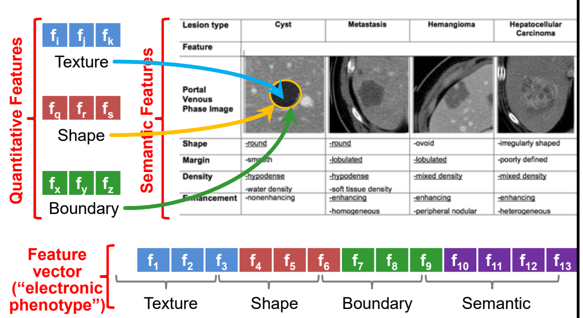
Approaches to AI in imaging

1. **Pre-defined feature capture**
2. **Unsupervised feature learning**

Pre-defined feature capture

- Use *domain knowledge to define features* extracted for learning a multivariate model
- Basis for “radiomics”
- Supervised machine learning on these features

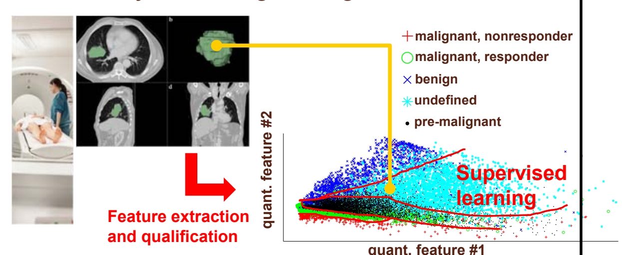
Capturing pre-defined image features



Model = f (quantitative + semantic features)

Machine learning with pre-defined image features

***Radiomics**: High-throughput extraction of quantitative image features with the intent of creating mineable databases from radiological images*



Radiomics: The process and the challenges, Magnetic Resonance Imaging, 30(9):1234-48, 2012.

Unsupervised feature learning

- *Raw image pixel data* input into a model
  - Image patch analysis
  - Deep learning } **Image data**
  - Word embeddings } **Text data**

Image patch analysis

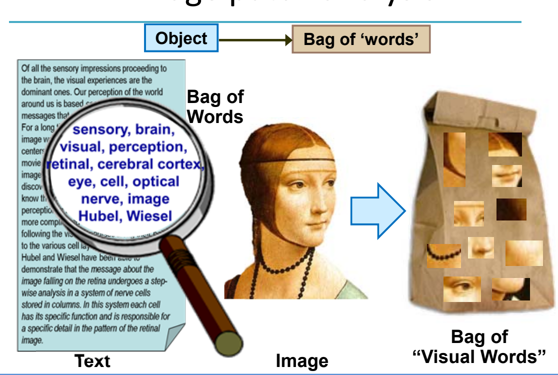
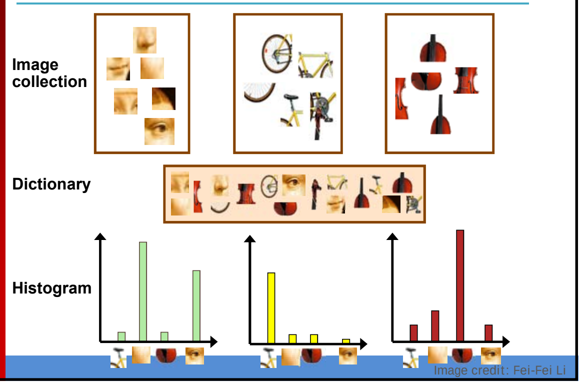


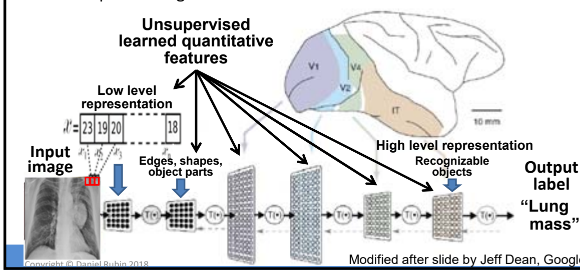
Image credit: Fei-Fei Li

Image patch analysis: Feature vectors of visual words



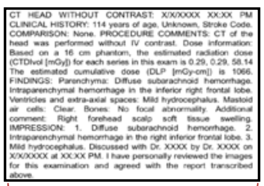
Unsupervised feature learning with images: Deep learning

- High-level abstractions of image features (hierarchical, non-linear transformations)
- Inspired by hierarchical visual processing by the brain
- Higher-level features (layers) are defined from lower-level ones, and represent higher levels of abstraction



Modified after slide by Jeff Dean, Google

Word embedding methods for classification

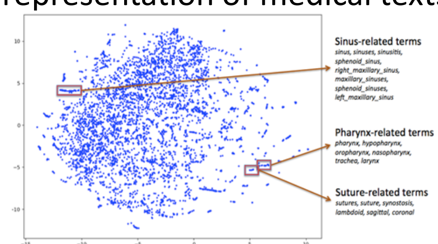
f() = y

CT HEAD WITHOUT CONTRAST: X/X/XXXX XX:XX PM CLINICAL HISTORY: 114 years of age, Unknown, Stroke Code. COMPARISON: None. PROCEDURE COMMENTS: CT of the head was performed without IV contrast. Dose information: Based on a 16 cm phantom, the estimated radiation dose (CTDIvol [mGy]) for each series in this exam is 0.29, 0.29, 58.14 The estimated cumulative dose (DLP [mGy-cm]) is 1066. FINDINGS: Parenchyma: Diffuse subarachnoid hemorrhage. Intraparenchymal hemorrhage in the inferior right frontal lobe. Ventricles and extra-axial spaces: Mild hydrocephalus. Mastoid air cells: Clear. Bones: No focal abnormality. Additional comment: Right forehead scalp soft tissue swelling. IMPRESSION: 1. Diffuse subarachnoid hemorrhage. 2. Intraparenchymal hemorrhage in the right inferior frontal lobe. 3. Mild hydrocephalus. Discussed with Dr. XXXX by Dr. XXXX on X/X/XXXX at XX:XX PM. I have personally reviewed the images for this examination and agreed with the report transcribed above.

Word embedding provides vector-based representation of text (learned using unsupervised methods);

e.g., to permit learning a classifier for document *x* being classified to label y

Word embeddings learn feature representation of medical texts



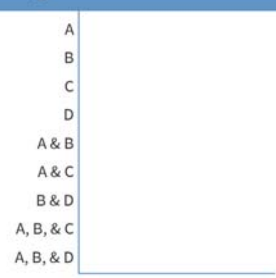
Word embedding using deep learning (4,442 words) projected in two dimensions

Unsupervised deep learning algorithms can **learn a feature representation from texts** without the need of supplying specific domain knowledge

Imon Banerjee, JDI 30:506-518, 2017

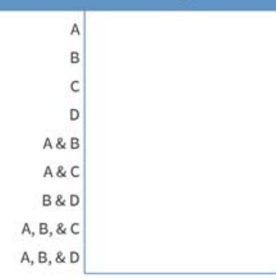
Which item(s) are used in connection with "pre-defined image features?"

A. Radiomics
B. Image patches
C. Image texture
D. Word embeddings



Which item(s) are used in connection with "unsupervised feature learning?"

A. Radiomics
B. Image patches
C. Image texture
D. Word embeddings



What are the types of AI applications?

Key clinical uses of unsupervised feature learning

1. Disease detection
2. Lesion segmentation
3. Diagnosis
4. Treatment selection
5. Response assessment
6. Clinical prediction (of treatment response or future disease)

Current application focus (1–3)

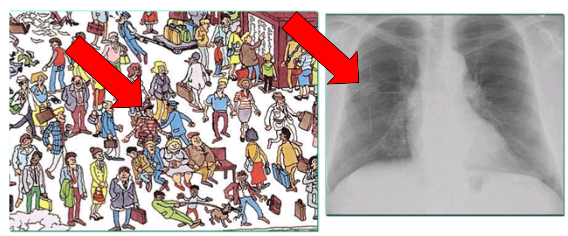
Active research area (4–6)

Key clinical uses of unsupervised feature learning

1. Disease detection
2. Lesion segmentation
3. Diagnosis
4. Treatment selection
5. Response assessment
6. Clinical prediction (of treatment response or future disease)

1) **Detection** of image abnormalities

AKA "where's Waldo?"



Detection and **segmentation**: General fully connected networks

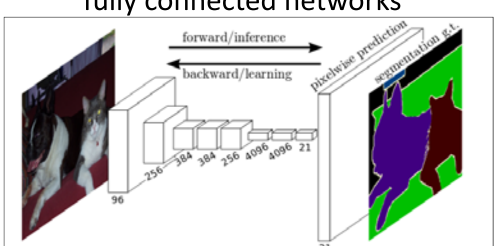
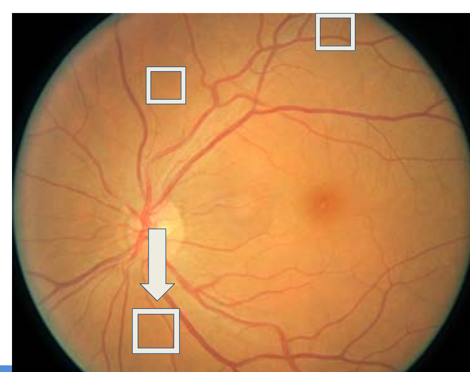


Figure 1. Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation.

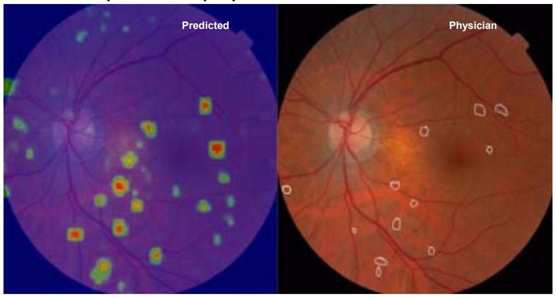
Detection/segmentation are pixel-based classification tasks

http://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Long_Fully_Convolutional_Networks_2015_CVPR_paper.pdf

Detecting retinal hemorrhages



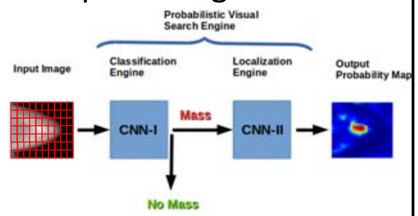
Sliding window detection of small features compared to physician manual detection



Single 224x224x3 input CNN sliding window:
Detecting any abnormal feature
Red = P~.99 **Green = P ~ 0.5** **Blue = ~ 0.01**

Detection of breast masses with deep learning

- Digital Database for Screening Mammography (DDSM)
- 2420 mass ROIs
- 80%/10%/10% training/test evaluation sets
- 256x256 patches, labeled as "mass" or "non-mass"
- Data augmentation: cropping, translation, rotation, flipping and scaling of image tiles
- Probability classification map of location (fully connected CNN)

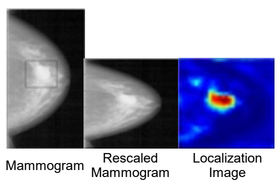


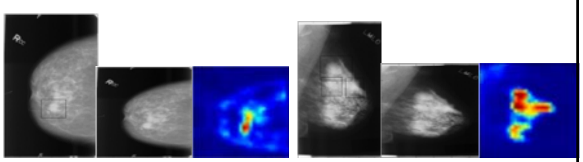
Performance:

	Number of Layers	Number of Parameters	Validation Accuracy
AlexNeT	8	60M	84%
VGG Net-16	16	140M	82%
GoogLeNeT	22	4M	85%

IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1310-1315, 2015

Examples





IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1310-1315, 2015

Key clinical uses of unsupervised feature learning

1. Disease detection
2. Lesion segmentation
3. Diagnosis
4. Treatment selection
5. Response assessment
6. Clinical prediction (of treatment response or future disease)

2) **Segmentation** of image regions

- Division of image into non-overlapping, homogeneous regions
- Segmented regions often input to other processing (e.g., feature extraction, image classification)

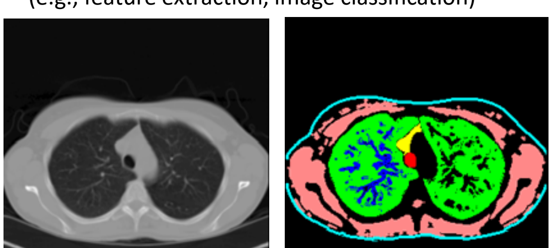
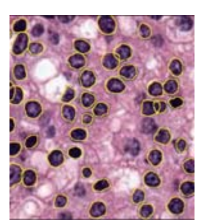
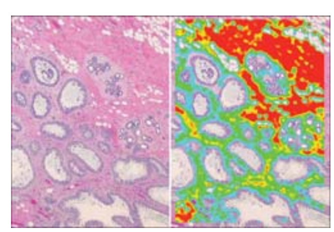


Image segmentation in pathology, different image scales



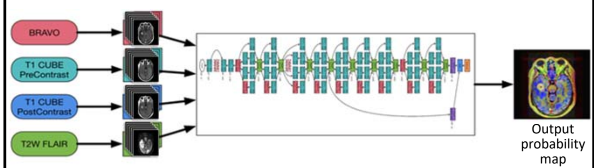
Cell/nuclear regions



Tissue regions

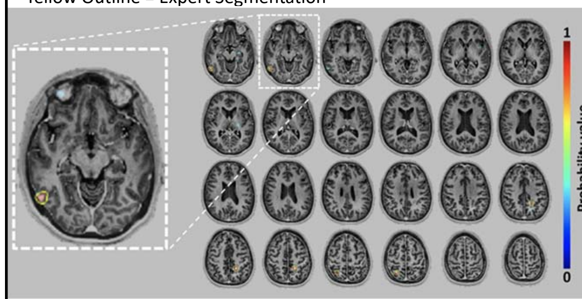
Segmentation of brain tumors using deep learning

- Modified GoogLe Net: Final layer is fully convolutional transpose layer for segmentation
- 2.5D model: 7 consecutive slices (center slice + 3 below and above) and 4 different modalities in the channel space
- Network output: Probability map of whether each voxel is a metastasis



Example case from test set

Yellow Outline = Expert Segmentation

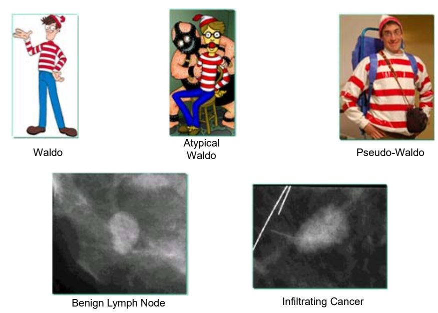


Key clinical uses of unsupervised feature learning

1. Disease detection
2. Lesion segmentation
3. Diagnosis
4. Treatment selection
5. Response assessment
6. Clinical prediction (of treatment response or future disease)

3) Diagnosis: Classification of images

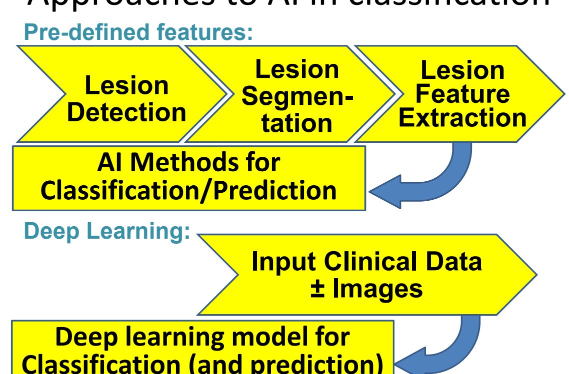
AKA "is it Waldo?"



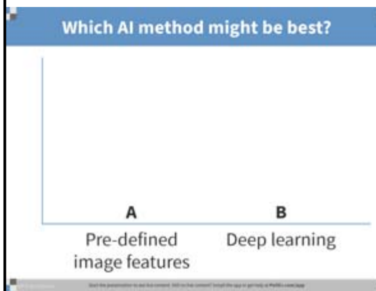
Diagnosis: Different approaches

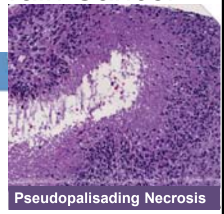
- **Pre-defined image feature extraction**
  - Domain knowledge available as to key informative features
  - Limited training data available
  - Generally slow development
  - Explainability by looking at model weights
- **Unsupervised feature learning** (deep learning)
  - Key informative features are not known
  - Lots if training data available (thousands of cases)
  - Generally fast development
  - ”Black box” difficult explainability

Approaches to AI in classification



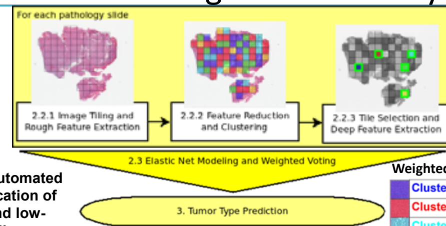
Pathology Images in Brain Cancer





- Key image features distinguish GBM from LGG are known
  - Vascular Proliferation
  - "Pseudopalisading Necrosis"
- Limited data available

Pathology classification using quantitative image feature analysis



Goal: Automated classification of high- and low-grade glioma

		Actual GBM	Actual LGG
Predicted	GBM	23	1
Predicted	LGG	0	21

- Correct classification in 44 of the total 45 tissue slices
- Accuracy of 97.78% with 95% CI 88.23%-99.94%
- NIR of 51.11% gives p-value = 3.366x10^-12

Weighted voting

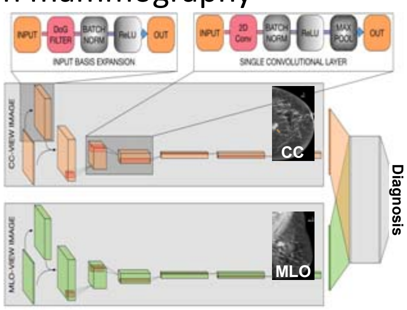
Cluster 1	+
Cluster 2	-
Cluster 3	+++
Cluster 4	++
Cluster 5	+/-

Result = ++++ GBM

Barker, J, Hoogi, A, Depeursinge, A, Rubin, D, Medical Image Analysis 30:60-71., 2016.

Deep learning: Diagnosis of masses on mammography

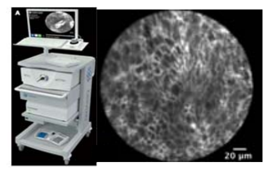
- Classify breast masses as benign vs. malignant
- Branching structure of CNN to account for two views of breast
- Predictive accuracy ~ 0.8

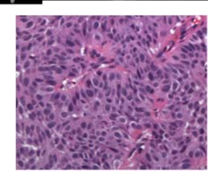


Yi D and Rubin DL, in preparation

Confocal Endomicroscopy

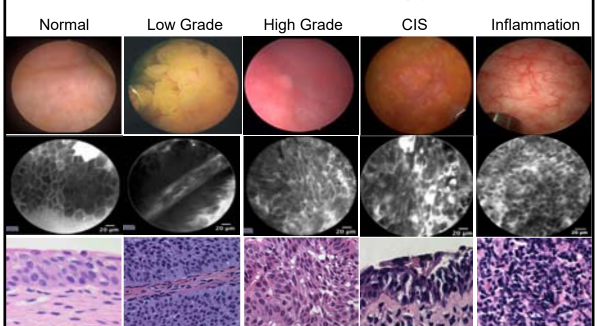
- Real-time *in vivo* microscopic images
- Used in GI and pulmonary applications
- Opportunity for "optical biopsy"
- High resolution, dynamic, sub-surface imaging recorded as ***movies***





Sonn et al. J Urol. 2009. 182(4):1299-305.

Bladder pathology



Diagnosis on movies

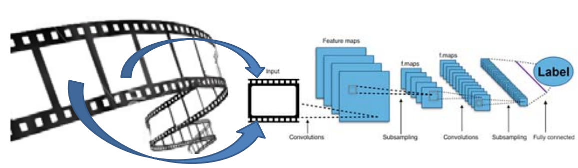
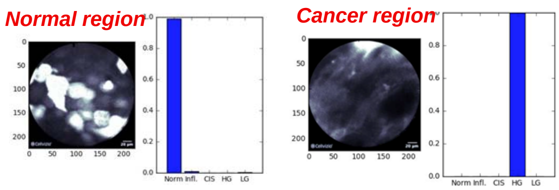


Image frame classification for real-time diagnosis of bladder cancer

Computerized interpretation during confocal endomicroscopy examination of the bladder permits localization of tumors in heterogeneous bladder lesions



Yi D, Chang TC, Liao JC, and Rubin DL, in preparation

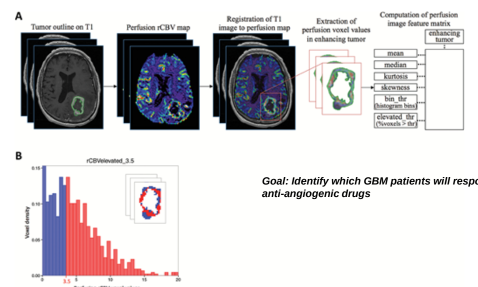
NB: How much image data train deep learning models is needed?

- Image detection/segmentation
  - This is a pixel-based classification task
  - Generally 100s of images/cases (which provides thousands of training examples!)
- Image classification
  - This is a whole image-based classification task
  - Thousands (preferably 10s or 100s of thousands) images/cases

Key clinical uses of unsupervised feature learning

1. Disease detection
2. Lesion segmentation
3. Diagnosis
4. Treatment selection
5. Response assessment
6. Clinical prediction (of treatment response or future disease)

4) Treatment selection



Goal: Identify which GBM patients will respond to anti-angiogenic drugs

Magnetic resonance perfusion image features uncover a **subgroup of GBM patients** with **poor survival** and **better response to drug treatment**

Neuro Oncol. 2016;19(7):997-1007

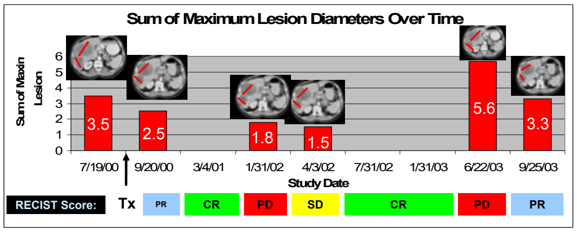
Key clinical uses of unsupervised feature learning

1. Disease detection
2. Lesion segmentation
3. Diagnosis
4. Treatment selection
5. Response assessment
6. Clinical prediction (of treatment response or future disease)

5) Response assessment

- Is the disease responding to treatment (disease getting better)
- Task: Evaluate images and determine if disease is:
  - Stable disease (SD)
  - Progressive disease (PD)
  - Partially responding to treatment (PR)
  - Completely responded to treatment (back to normal) (CR)

Treatment response assessment



"Is the patient's cancer responding to treatment?"

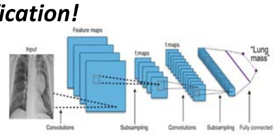
(A task for **computer reasoning**, discussed in next lecture)

Key clinical uses of unsupervised feature learning

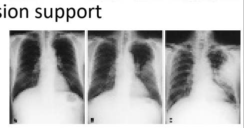
1. Disease detection
2. Lesion segmentation
3. Diagnosis
4. Treatment selection
5. Response assessment
6. Clinical prediction (of treatment response or future disease)

5) Clinical Prediction

- ***Don't confuse with classification!***
- **Classification**
  - Input: Usually **single image/ study, single time point**
  - Input usually **only images**
  - Goal: Reporting, diagnosis, decision support



- **Prediction**
  - Input: Usually **multiple images/ studies, multiple timepoints**
  - Input may include **clinical data**
  - Goal: Forecast future clinical outcomes (response to treatment, adverse events, survival)



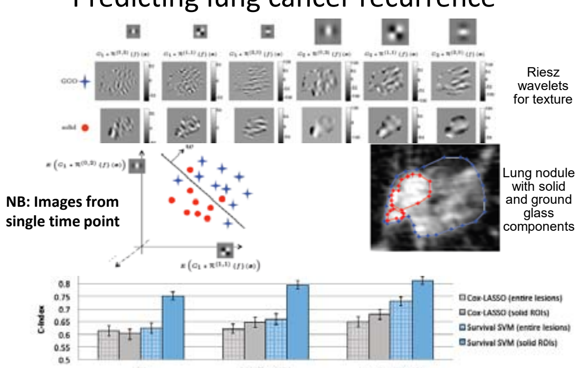
- Pre-defined features or deep learning (NB: deep learning can model **multiple timepoints**)

Prediction is key to Precision Health and Precision Medicine

- **Precision Medicine**
  - Will a patient's will disease progress?
  - Will a patient have particular good/bad outcomes?

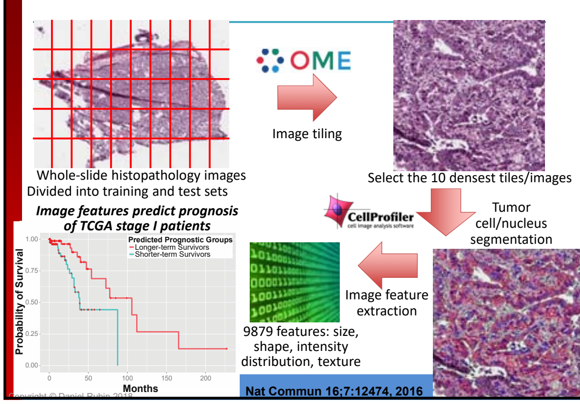
- **Precision Health**
  - Which healthy people will develop disease?
  - Can we develop custom screening for early detection or prevent disease?

Pre-defined image features (lesion texture): Predicting lung cancer recurrence



Depeursinge, et. al, and Rubin, Med Phys 42(4):2054-63, 2015

Pre-defined image features: Predicting survival



Nat Commun 16;7:12474, 2016

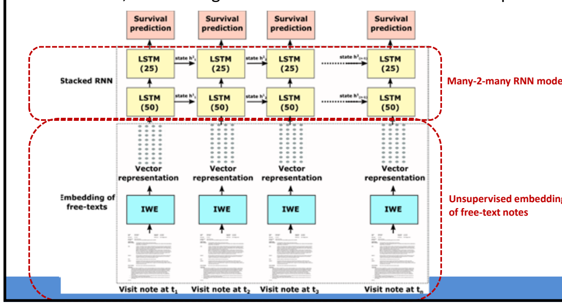
Unsupervised feature learning: Predicting patient survival

- Goal: Predict patient survival in metastatic cancer from medical records data
- Rationale:
  - Overutilization of aggressive medical treatment in patients close to the end of life
  - Physicians cannot currently accurately estimate patient life expectancy; thwarts shared patient/physician decision making
- Approach: Model incorporating **longitudinal medical records clinical notes** using **word embeddings** to represent the text

Banerjee I, Gensheimer MF, Wood DJ, Henry S, Chang D, Rubin DL. Probabilistic Prognostic Estimates of Survival in Metastatic Cancer Patients (PPES-Met) Utilizing Free-Text Clinical Narratives. AMIA Informatics 2018, arXiv preprint arXiv:1801.03058. 2018 Jan 9.

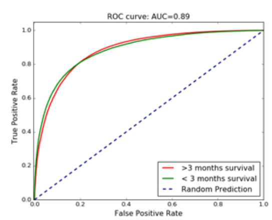
Use of word embeddings for prediction

- LSTM model; input = sequence of free-text clinical visit narratives ordered by the date of visits.
- Output = probability of short-term life expectancy (> 3 months) for each visit, considering the current and all the historic time points.

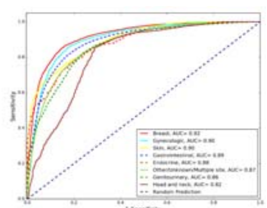


Results: Quantitative Evaluation

Tested on 1818 patients with multiple visits



Overall ROC AUC for predicting 3 month survival = 0.89 CI [0.884 - 0.897]

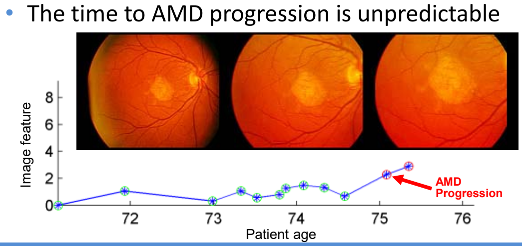


ROC for each primary cancer site

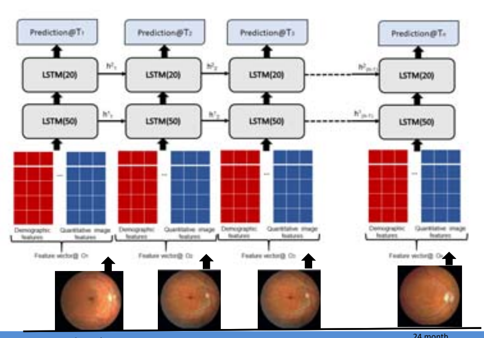
NB: State of the art prediction (based on clinician judgment) AUC approx. 0.7

Unsupervised feature learning: Disease progression

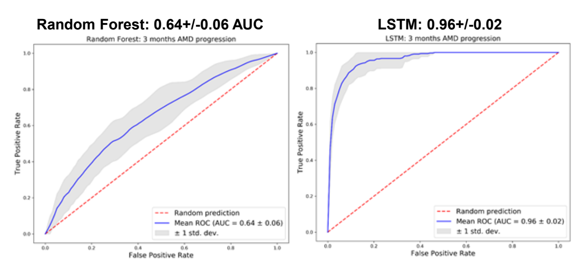
- Age related macular degeneration (AMD) changes over time
- AMD progresses in approx. 5% of patients
- The time to AMD progression is unpredictable



Prognosis of AMD Disease using longitudinal image biomarkers

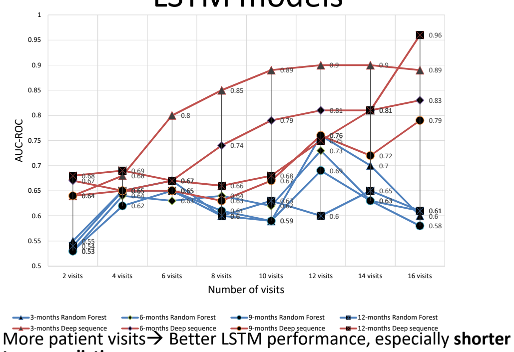


Prediction results: Short term



10-fold cross validation ROC curves

Amount of longitudinal data helps LSTM models



- More patient visits→ Better LSTM performance, especially **shorter term prediction**
- Negligible effect on performance of random forest model

Which kind of AI application analyzes a CT scans and tells a physician whether to use Treatment A or Treatment B?

Detection
Segmentation
Diagnosis
Treatment selection
Response assessment
Clinical prediction

Which kind of AI application points out a suspicous region in an image that a physician should biopsy?

Detection
Segmentation
Diagnosis
Treatment selection
Response assessment
Clinical prediction

What are the challenges to progress?

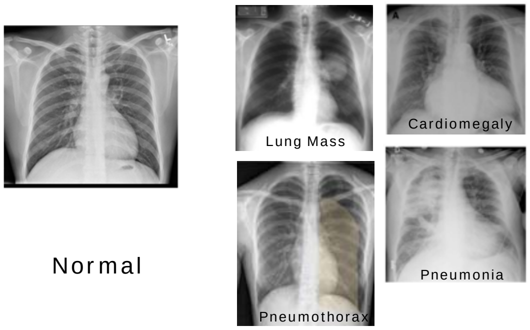
Challenges to progress

1. Data **quantity and quality**
2. Integrating **domain knowledge** into AI models
3. Leveraging data from **multiple institutions**
4. **Evaluation** of AI applications in practice and impact on clinician performance

1) Data quantity/quality issues

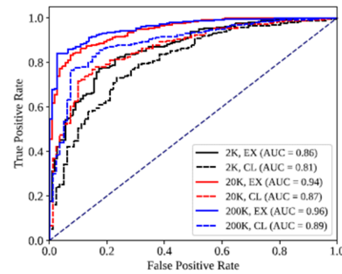
- You generally need a **lot of data**
  - Ideally 100,000+ training examples
  - NB, for segmentation you can get away with less data
- You generally need many **high quality labels**
  - Costly to obtain
- Most historical data in PACS/EMR is **not annotated**
  - Thus, difficult to leverage historical data
  - Generally hand-curation effort for each project
  - Many institutions spending $$ creating curated datasets (but not sharing…)

Binary classification task: Normal vs. Abnormal



Abnormal

Effect of data quantity and quality



EX: Expert Label (high quality)

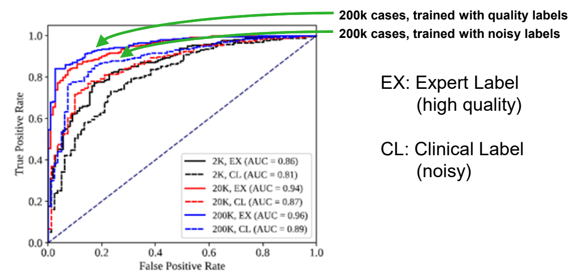
CL: Clinical Label (noisy)

Larger training set → Better CNN performance
Better quality labels → Better CNN performance

Approaches to data quantity/quality

- Data augmentation
  - Perform reasonable image transformations (rotations, flips, etc.)
- Transfer learning
  - Train a model on related task and use model weights to initialize new model to be trained
  - ImageNet very commonly used, but may not be relevant to medical imaging use cases
- Train with more data, even if labels have noise
  - "Weak learning"

Value of training with more data, even labels are noisy



EX: Expert Label (high quality)

CL: Clinical Label (noisy)

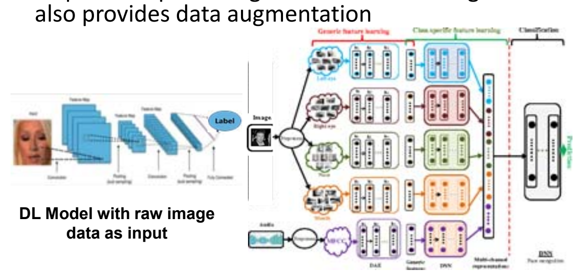
Training with <u>many cases</u> (200k) having <u>noisy labels</u> approaches maximum performance (realized with fewer (20k) cases with quality labels)

2) Integrating domain knowledge

- **Pre-defined features** capture expert knowledge about relevant image signals
- Approach: Generate images based on extraction of pre-defined features (**feature maps**)
- Benefit: **Incorporates knowledge** and also provides data augmentation

AI challenges and potential solutions

Feature extraction from images as pre-processing step for deep learning models enhances signal and also provides data augmentation



DL Model with raw image data as input

DL Model with extracted features as inputs

Deep learning for predicting future cancer risk

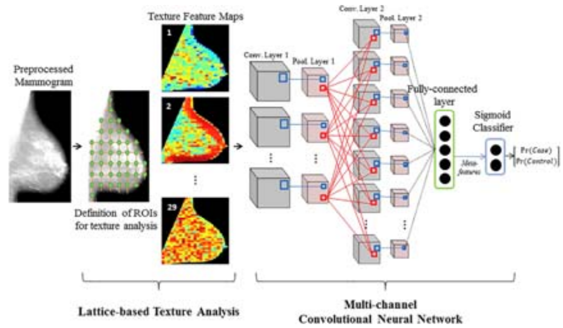


Image texture feature maps *preserve discriminative spatially-dependent features and augment data in multi-channel CNN*

Acad Radiol 25(8):977-984, 2018

3) Leveraging data from multiple institutions

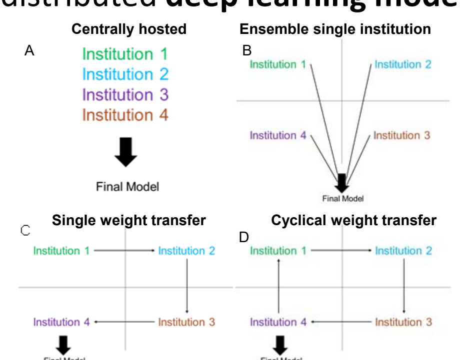
- Most AI models built with data from only **one institution**
- Data among institutions varies
  - Geographic variations in patient populations
  - Differences in imaging parameters
  - Differences in vendor equipment
  - Variations in medical practices
- Thus, AI methods may not **generalize**
- Acquiring/sharing data from **multiple institutions** is challenging!

Centralized approach to AI model development

Overcoming barriers to data sharing

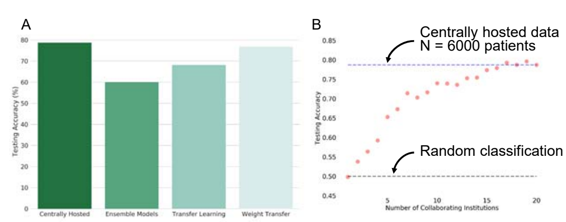
- Bring the **model to the data** instead of bring the data (centralized) to the model
- **Distributed computation** of training deep learning models

Alternative approaches for training distributed **deep learning models**



J Am Med Inform Assoc 25(8):945-954, 2018

Cyclical weight transfer has similar performance to centrally-hosted training



Results based on having 4 institutions

Accuracy increases with number of collaborating institutions

J Am Med Inform Assoc 25(8):945-954, 2018

5) Evaluating AI systems in practice

- Everything an AI system "knows" is based on the **data upon which it is trained**
- AI algorithms may not **generalize** to new data (wasn't seen before)
  - Data used to create algorithms can contain **bias**
  - Differences in **patient populations** (e.g., foreign vs. domestic
  - Differences in **equipment/parameters** for imaging
  - Rare **disorders/abnormalities** may be under-represented

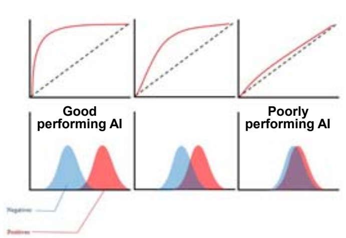
Deep learning models may not generalize

Train - Tune Site	Comparison Type*	Test Site (Images)	AUC (95% C.I.)	Acc.	Sens.	Spec.	PPV	NPV
NIH	Internal	NIH (N=22,062)	0.750 (0.721-0.778)	0.255	0.951	0.247	0.015	0.998
	External	MSH (N=8,388)	0.695 (0.683-0.706)	0.476	0.950	0.212	0.401	0.884
	External	IU (N=3,807)	0.725 (0.644-0.807)	0.190	0.974	0.182	0.012	0.999
	Superset	MSH + NIH (N=30,450)	0.773 (0.766-0.780)	0.462	0.950	0.403	0.160	0.985
	Superset	MSH + NIH + IU (N=34,257)	0.787 (0.780-0.793)	0.470	0.950	0.418	0.148	0.987
MSH	Internal	MSH (N=8,388)	0.802 (0.793-0.812)	0.617	0.950	0.432	0.482	0.940
	External	NIH (N=22,062)	0.717 (0.687-0.746)	0.184	0.951	0.175	0.014	0.997
	External	IU (N=3,807)	0.756 (0.674-0.838)	0.099	0.974	0.090	0.011	0.997
	Superset	MSH + NIH (N=30,450)	0.862 (0.856-0.868)	0.562	0.950	0.516	0.190	0.989
	Superset	MSH + NIH + IU (N=34,257)	0.871 (0.865-0.877)	0.577	0.950	0.537	0.180	0.990
MSH + NIH	Internal	MSH + NIH (N=30,450)	0.931 (0.927-0.936)	0.732	0.950	0.706	0.279	0.992
	Subset	NIH (N=22,062)	0.733 (0.703-0.762)	0.243	0.951	0.234	0.015	0.997
	Subset	MSH (N=8,388)	0.805 (0.796-0.814)	0.630	0.950	0.451	0.491	0.942
	External	IU (N=3,807)	0.815 (0.745-0.885)	0.238	0.974	0.230	0.013	0.999
	Superset	MSH + NIH + IU (N=34,257)	0.934 (0.929-0.938)	0.732	0.950	0.709	0.258	0.993

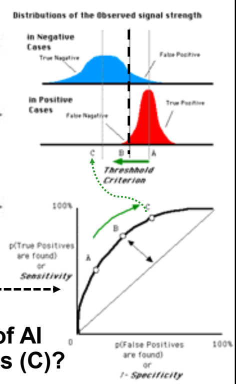
*Superset= a test dataset containing data from the same distribution (hospital system) as the training data as well as external data. Subset = a test dataset containing data from fewer distributions (hospital systems) then the training data.

Zech JR et al., Confounding variables can degrade generalization performance of radiological deep learning models, arXiv:1807.00431

AI algorithm performance: ROC Curve

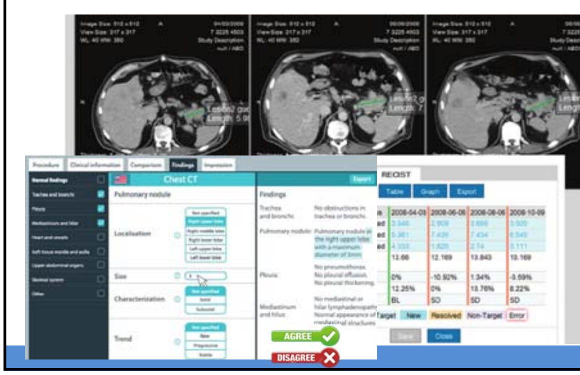


AI algorithm output is based on a particular operating point on ROC



What is the clinical impact of AI that picks up all positive cases (C)?

Toolkit for collecting AI performance metrics in the clinical workflow



Need for clinician expertise…

Physicians need to maintain their expertise to guard against becoming overly dependent on AI algorithms that may lead them astray

TESLA'S AUTOPILOT WAS INVOLVED IN ANOTHER DEADLY CAR CRASH

TESLA NOW HAS another fatality to hang on its semi-autonomous driving system. The company just reveale its Autopilot feature was turned on when a Model X SU slammed into a concrete highway lane divider and bur flames on the morning of Friday, March 23. The driver, Huang, died shortly afterwards at the hospital.

This is the second confirmed fatal crash on US roads in Tesla's Autopilot system was controlling the car. It rais familiar questions about this novel and imperfect syst which could make driving easier and safer, but relies o constant human supervision.

Summary

Motivation for AI applications is to help clinicians deal with **flood of image data**, **reduce variation in practice**, and address variations in disease for **precision medicine/health**

Types of AI applications include **disease detection**, **lesion segmentation**, **diagnosis** (classification), **treatment selection**, **response assessment**, and **clinical prediction**

Challenges to progress are **data quantity and quality**, **integrating domain knowledge** into AI models, **leveraging data from multiple institutions**, and **evaluation of AI applications** in practice

What does it mean for me?

- Awareness of clinical needs
- Think carefully about amount of data and best method
- Ideas for potentially useful medical applications

Next time:

Leveraging semantic data for image query and computerized inference

Thank you!