Health Care Breakout Session 2:
Identifying Skin Cancer with CNNs

Matt Engelhard

Today

- Guided discussion of “Dermatologist-level classification of skin cancer with deep neural networks” (Esteva et al., 2017)
  - Model Implementation
  - Data Preparation and Model Evaluation
  - Model Interpretation
  - Significance of Results

- Highlight several distinct problems in medicine to which CNNs have been applied

Dermatologist-level classification of skin cancer with deep neural networks

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S
Nature, 2017 Feb;542(7639):115

OVERVIEW:

- A CNN (Google's Inception v3) was trained to classify skin lesions
- 129,450 labeled images of 2,032 diseases divided into 757 classes
  - Labeled by dermatologist and in some cases by biopsy
- Performance on two tasks was tested against 21 dermatologists:
  - Task 1: keratinocyte carcinomas vs benign seborrheic keratosis
  - Task 2: malignant melanomas vs benign nevi
- CNN performance was comparable to dermatologist performance

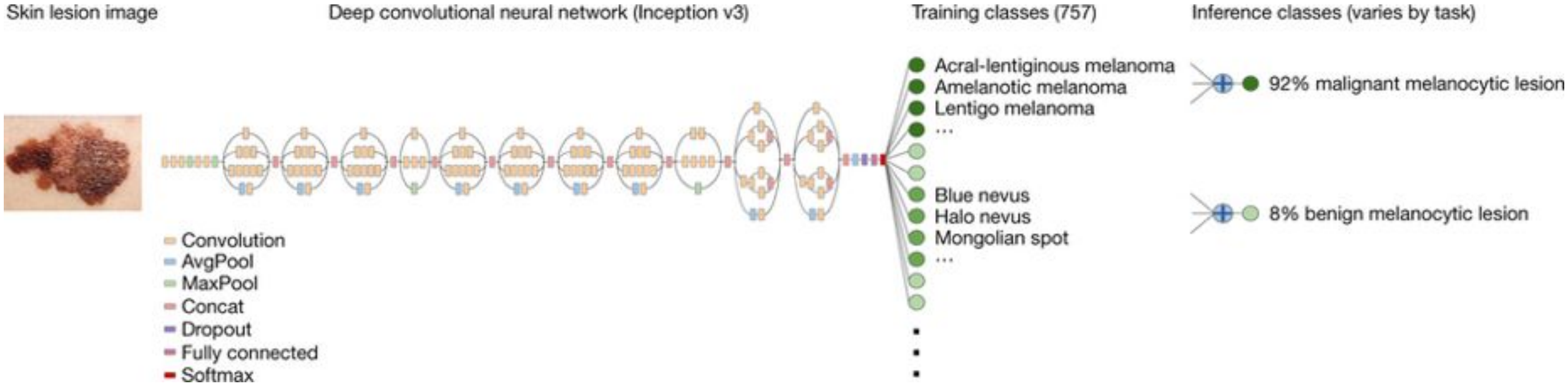
MODEL IMPLEMENTATION

Familiar Methods…

- Deep Convolutional Neural Network (Google's Inception v3)

- Trained in Tensorflow:
  - calculate the gradient of the (cross-entropy) loss on a mini-batch of images via backpropagation
  - take a step in the opposite direction (i.e. stochastic gradient descent)

Retraining the Inception v3 CNN



- Begin with weights learned for the 2014 ImageNet challenge
- Replace final classification layer
- Fine-tune ALL parameters in Tensorflow via SGD

Q: How would you begin modifying the code from today's hand-on session to solve your own image classification problem?

Inception v3 and many other models are freely available

Pre-trained Models

Neural nets work best when they have many parameters, making them powerful function approximators. However, this means they must be trained on very large datasets. Because training models from scratch can be a very computationally intensive process requiring days or even weeks, we provide various pre-trained models, as listed below. These CNNs have been trained on the ILSVRC-2012-CLS image classification dataset.

In the table below, we list each model, the corresponding TensorFlow model file, the link to the model checkpoint, and the top 1 and top 5 accuracy (on the imagenet test set). Note that the VGG and ResNet V1 parameters have been converted from their original caffe formats (here and here), whereas the Inception and ResNet V2 parameters have been trained internally at Google. Also be aware that these accuracies were computed by evaluating using a single image crop. Some academic papers report higher accuracy by using multiple crops at multiple scales.

Model	TF-Slim File	Checkpoint	Top-1 Accuracy	Top-5 Accuracy
Inception V1	Code	inception_v1_2016_08_28.tar.gz	69.8	89.6
Inception V2	Code	inception_v2_2016_08_28.tar.gz	73.9	91.8
Inception V3	Code	inception_v3_2016_08_28.tar.gz	78.0	93.9
Inception V4	Code	inception_v4_2016_09_09.tar.gz	80.2	95.2

TF-Slim Code:
Defines the model architecture

Checkpoint File:
Trained model weights

https://github.com/tensorflow/models/tree/master/research/slim#Pretrained

Q: The authors initialize their chosen convolutional neural network architecture (Google's Inception v3) to the weights learned for the ImageNet Challenge, in which everyday images are assigned to classes such as "grocery store". Why might these weights still be useful when learning to classify skin lesions?

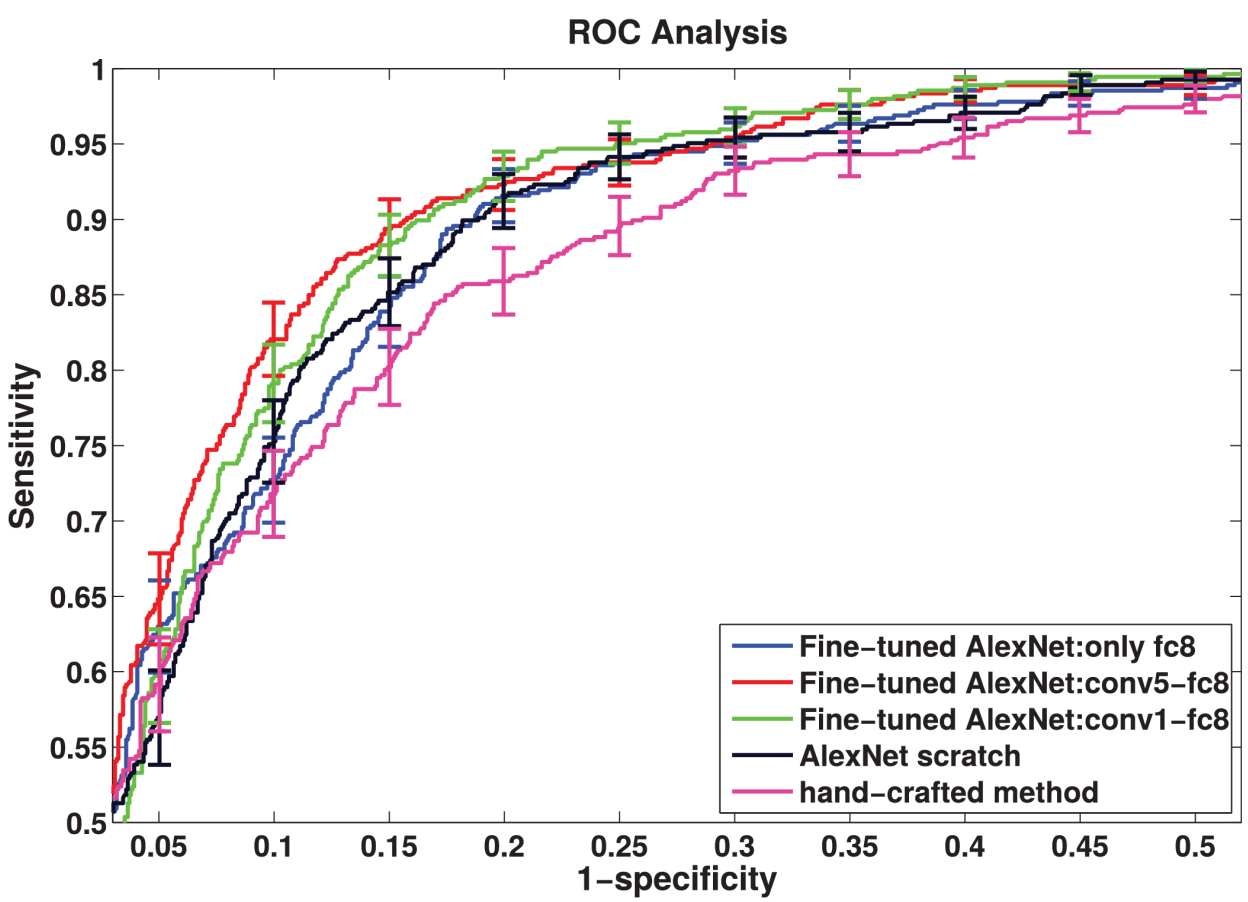
Intuition behind pre-training:

1. Old and new classification tasks may have something in common, e.g. structure in natural images
2. Previously learned weights are not likely to be worse than random initialization

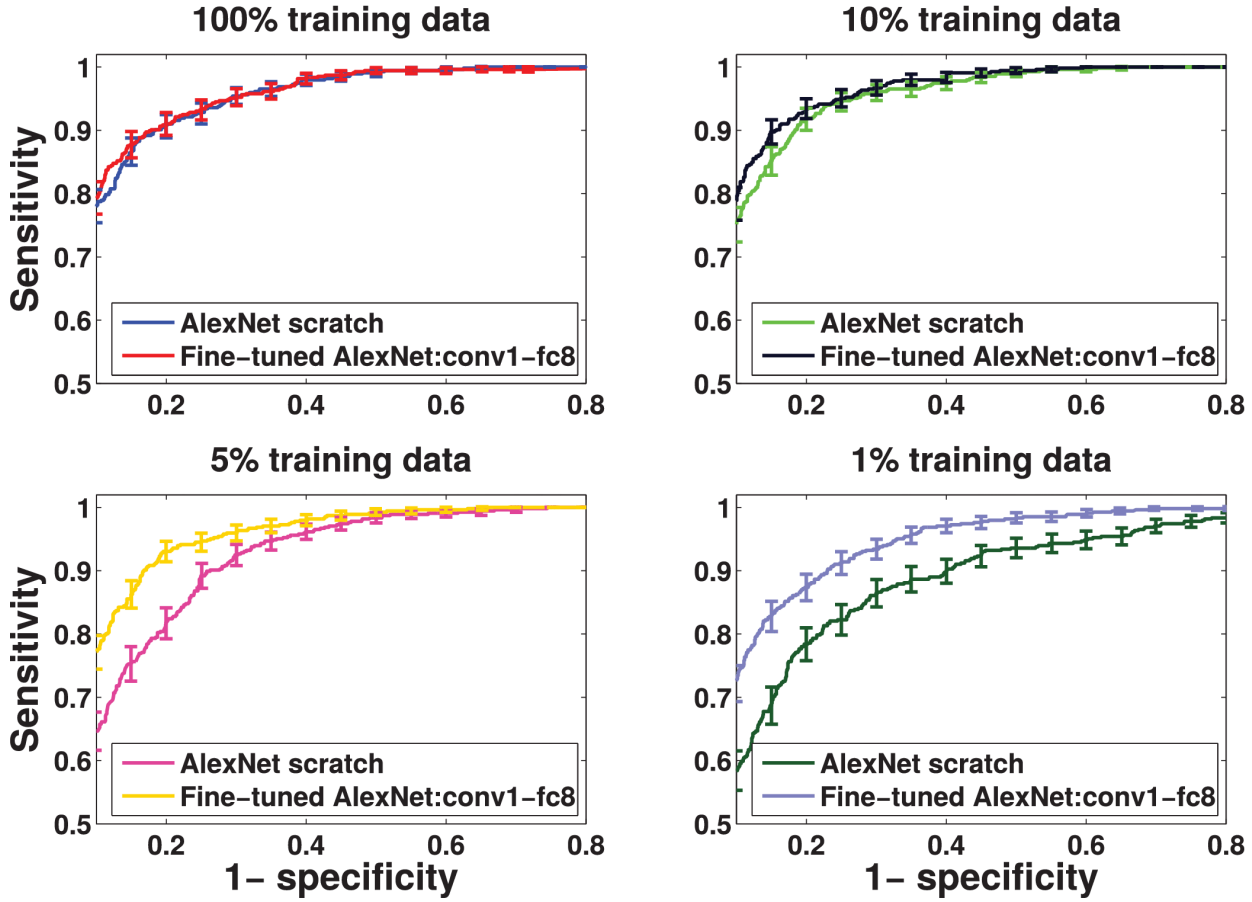
A filter that detects edges may be useful for multiple classification tasks.

https://docs.gimp.org/en/plug-in-convmatrix.html

Empirical Benefits of Pre-training



(a)



(b)

Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J. Convolutional neural networks for medical image analysis: Full training or fine tuning?. IEEE transactions on medical imaging. 2016 May;35(5):1299-312.

Benefits of Pre-training, in Brief

1) fine-tuning a pre-trained model is **at least as good as learning from scratch**

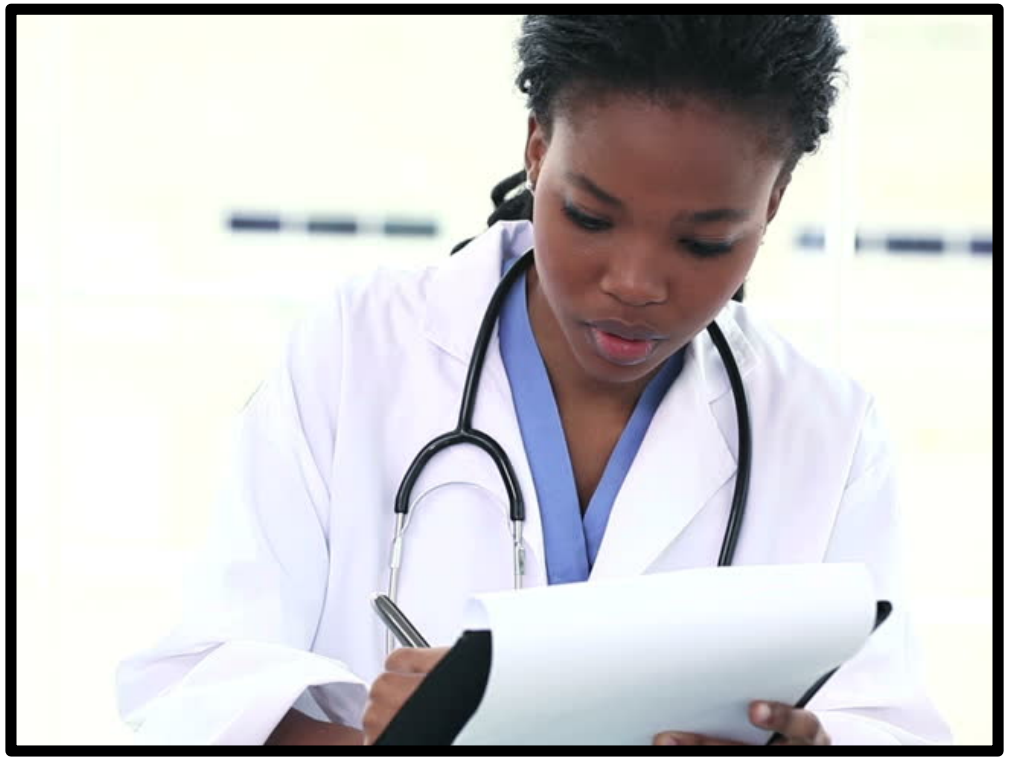
2) fine-tuning tends to work **better with smaller datasets**

3) best tuning “depth” depends on the application and size of dataset

DATA PREPARATION AND MODEL EVALUATION

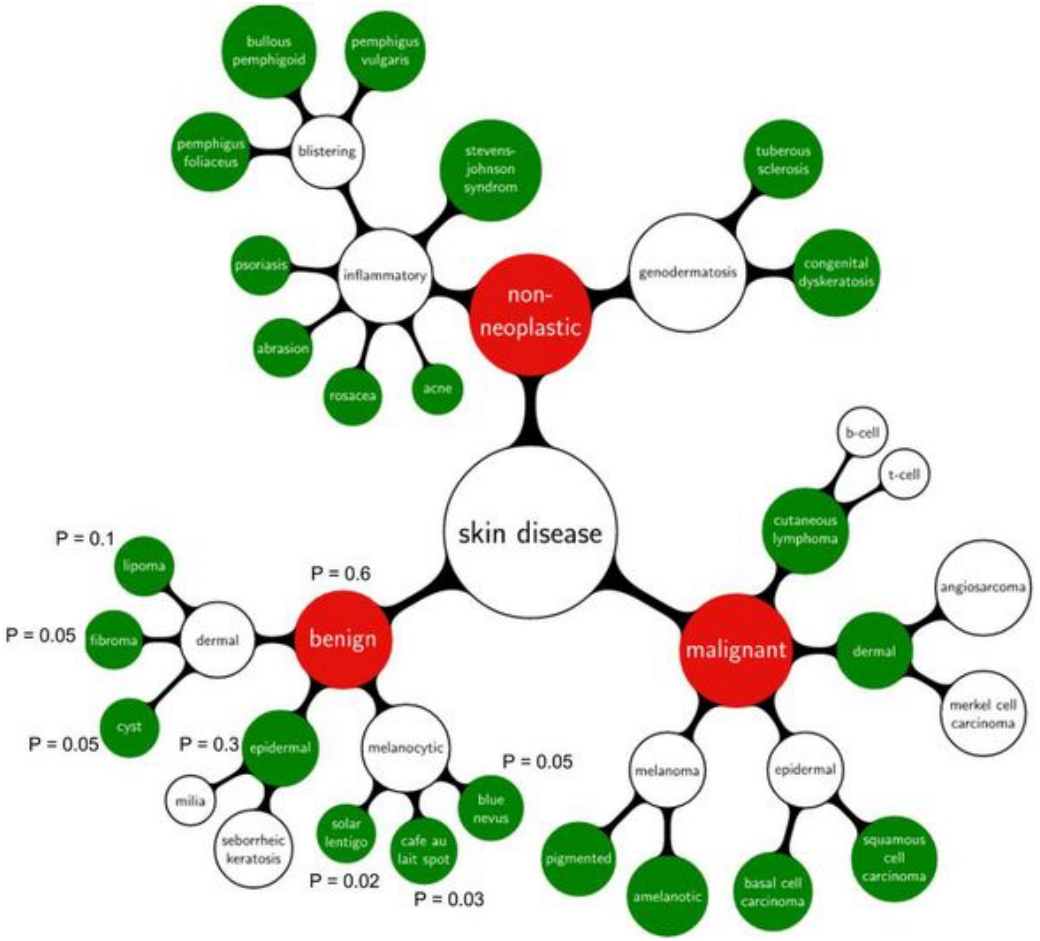
Two Types of Labels

All images: dermatologists' annotations



Some images: biopsy results

Taxonomy -> Training Classes



$$P[u] = \sum_{v \in C(u)} P[v]$$

Disease Partitioning Algorithm:

- Descend the tree until the current node contains <1000 images across all child nodes. Add these images as a distinct training class.
- Resulted in 757 training classes.
- However, validation was performed on higher-level nodes.

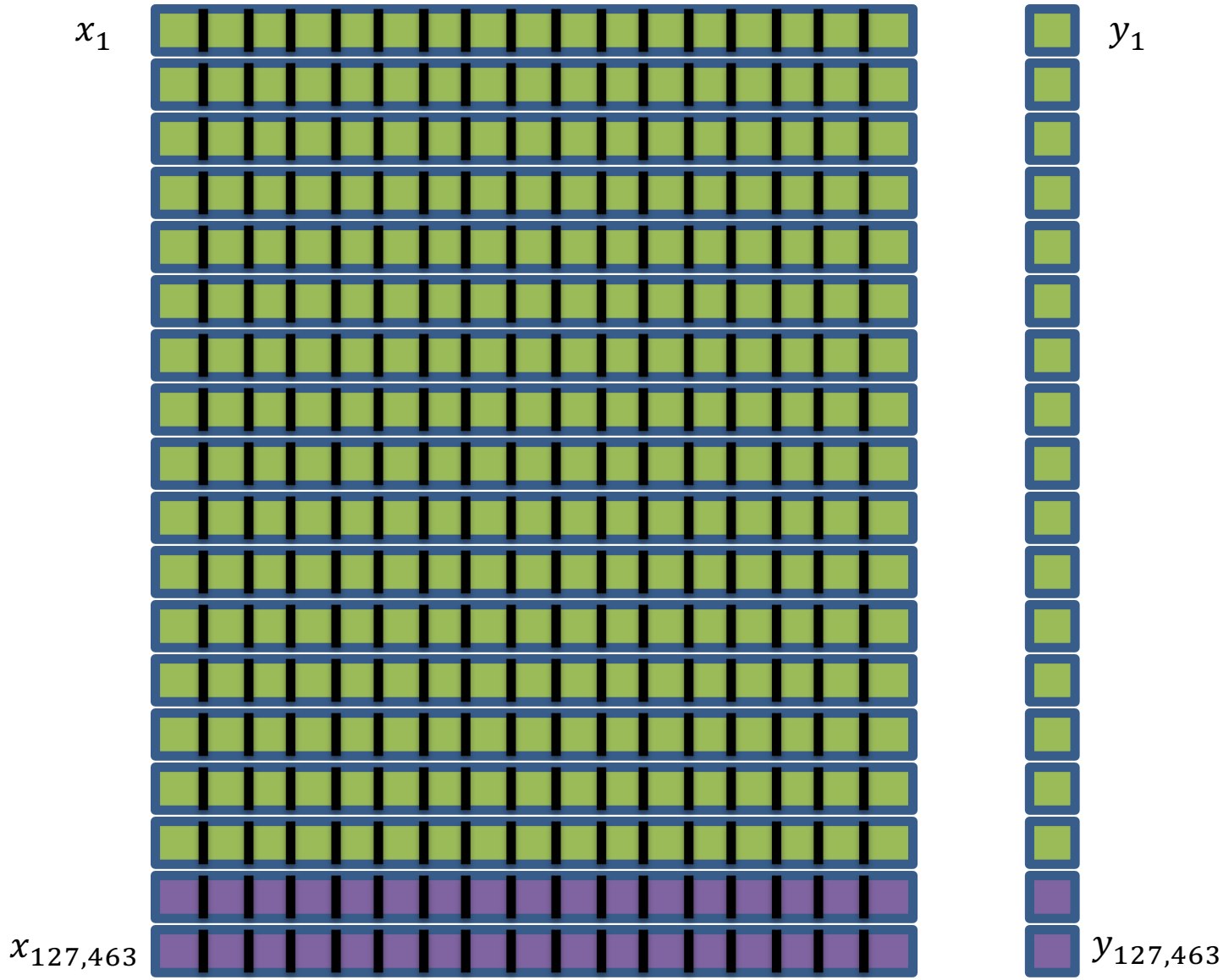
Multiple Classification Tasks

1. Match dermatologists' labeling:
   - Three-class disease partition
   - Nine-class disease partition

2. Biopsy-proven benign versus malignant lesions:
   - Keratinocyte carcinoma vs benign seborrheic keratosis
   - Malignant melanoma vs benign nevus
     - Standard images
     - Dermoscopy

Evaluation, Part 1:

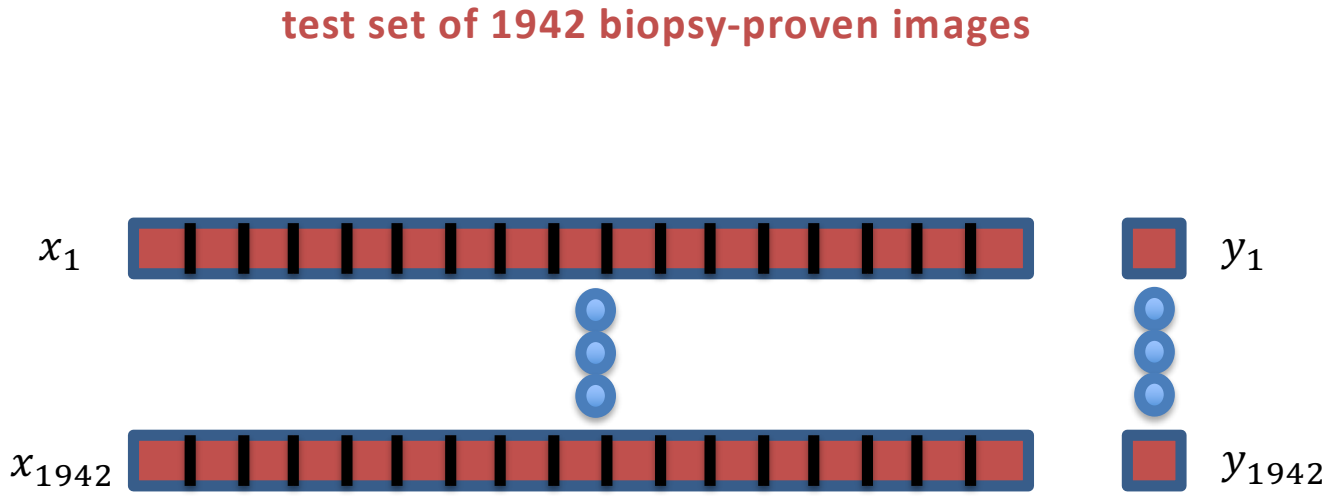
- 9-fold cross-validation
  - 757 training classes derived from dermatologists' annotations
  - 3 and 9-class validation partitions
  - two dermatologists
- Not clear what data were used to decide when to stop training



training set **validation set**

Evaluation, Part 2:

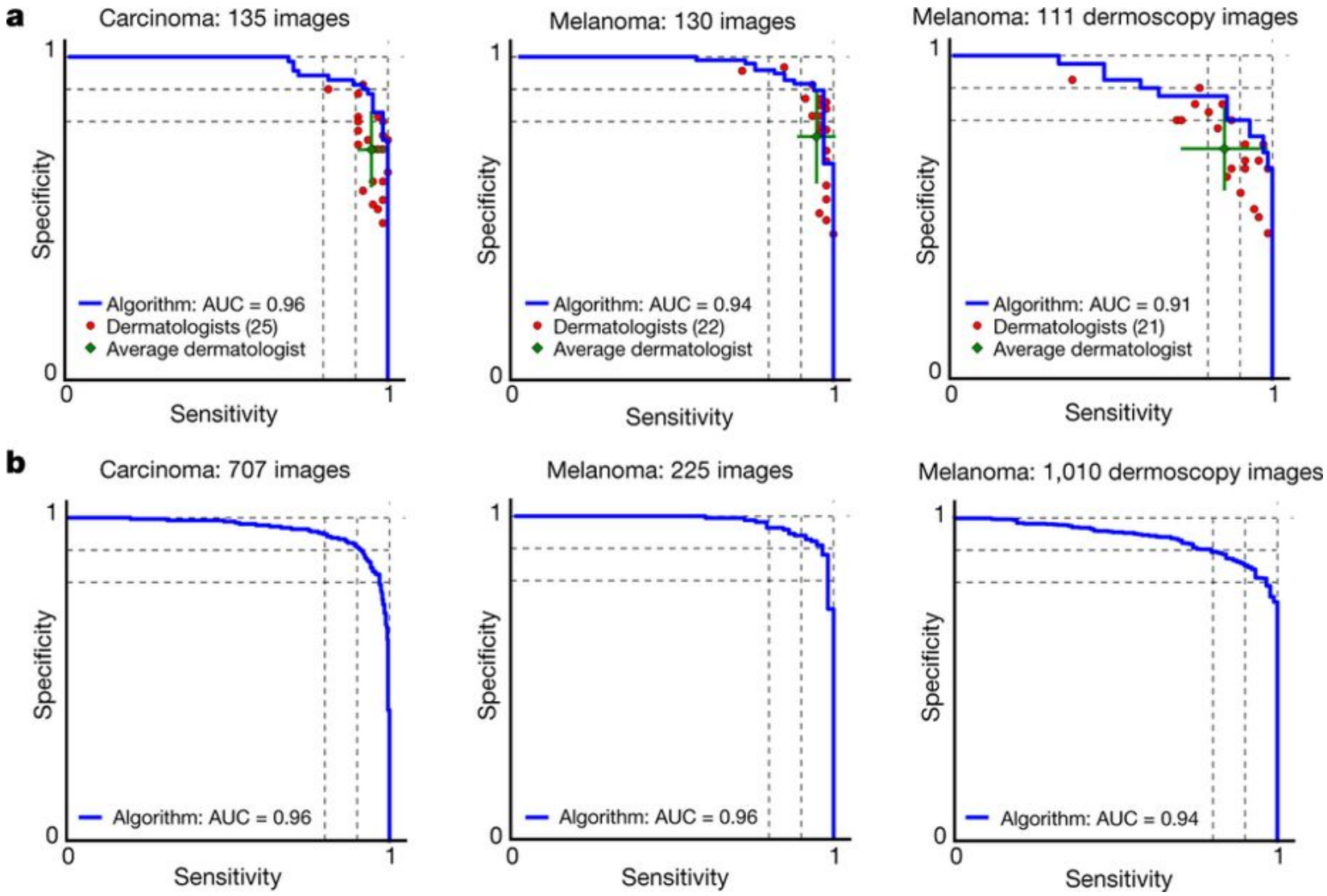
Model trained during Part 1 is compared to 21 dermatologists on a test set of biopsy-proven images



Q: In the “data preparation” section, authors emphasize that images of the same lesion (from the same person) were not split between the training and validation sets. Why is this important?

“Our dataset contains sets of images corresponding to the same lesion but from multiple viewpoints, or multiple images of similar lesions on the same person. While this is useful training data, extensive care was taken to ensure that these sets were not split between the training and validation sets.”

Results: CNN Performance vs Dermatologists

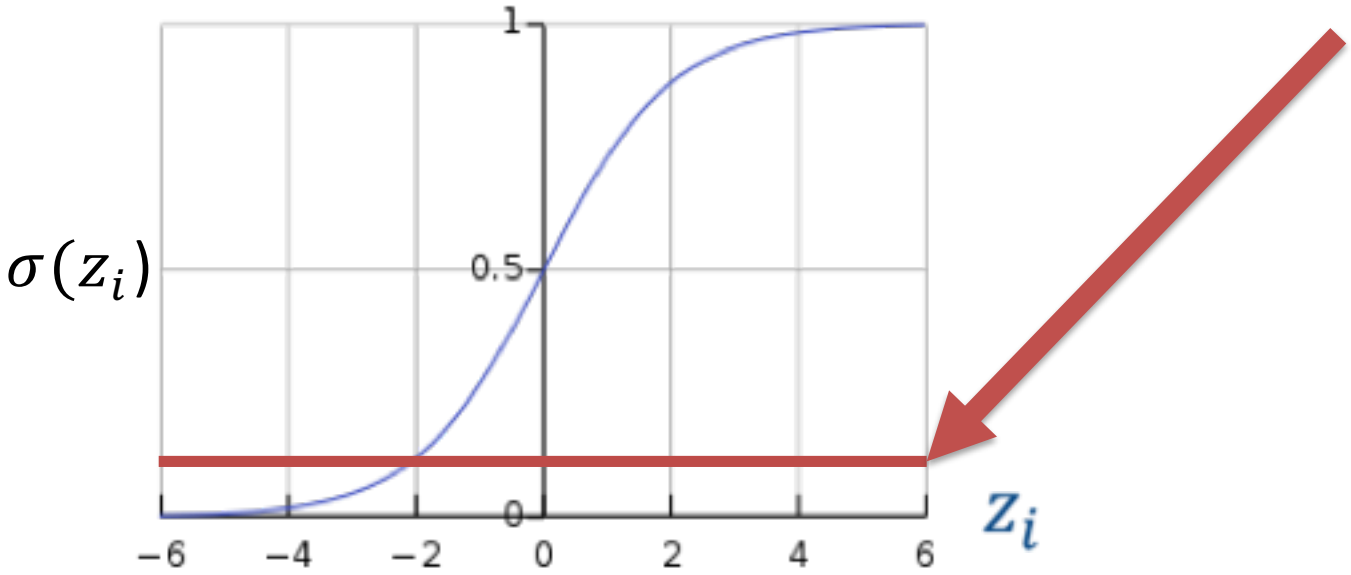


Q: In the previous figure, each dermatologist's performance is represented as a single point (i.e. sensitivity/specificity pair), whereas the CNN's performance is represented as a curve. Why is this?

Probabilistic Classification

With a probabilistic classifier, we can explicitly balance sensitivity vs specificity

$$p(y_i = 1|x_i) = \sigma(z_i)$$



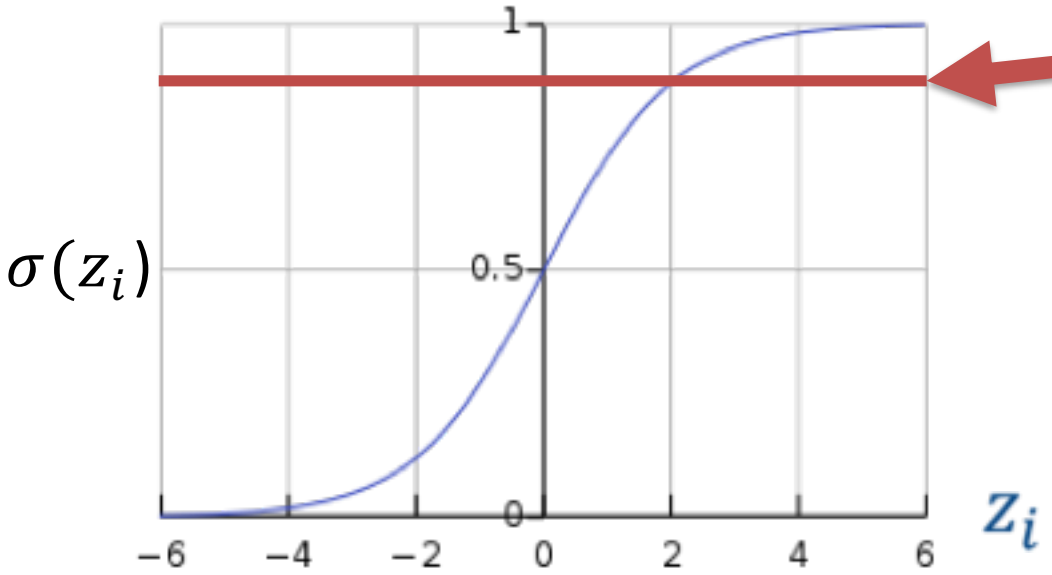
classification threshold

A low threshold favors sensitivity, because more points are classified as condition positive (i.e. 1s)

Probabilistic Classification

With a probabilistic classifier, we can explicitly balance sensitivity vs specificity

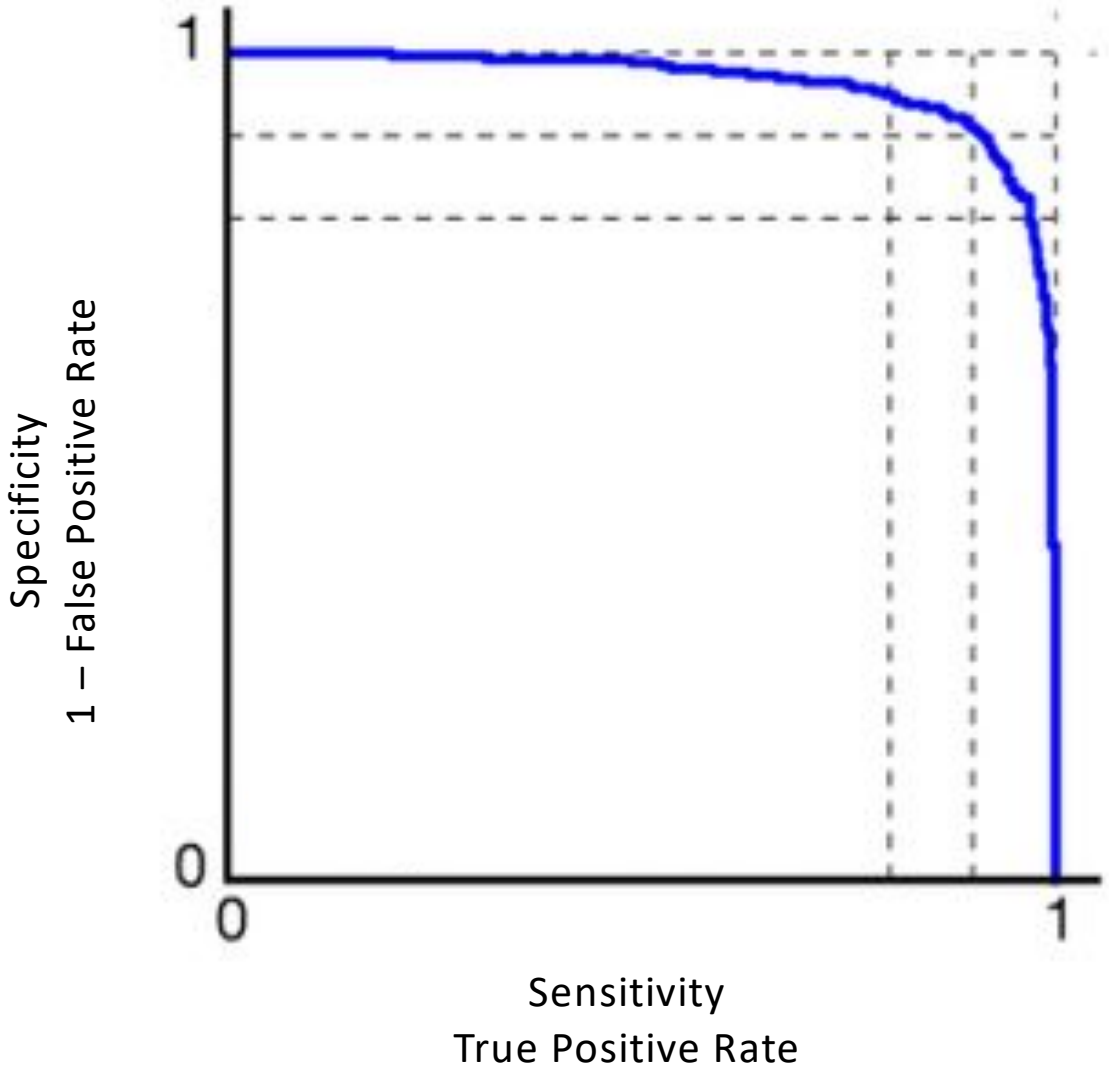
$$p(y_i = 1|x_i) = \sigma(z_i)$$



classification threshold

A high threshold favors specificity, because more points are classified as condition negative (i.e. 0s)

Sensitivity/Specificity Curve



Sensitivity, or True Positive Rate:

$$\frac{\text{true positives}}{\text{all real positives}}$$

Specificity, or (1 – False Positive Rate):

$$\frac{\text{true negatives}}{\text{all real negatives}}$$

Accuracy:

$$\frac{\text{true positives} + \text{true negatives}}{\text{total cases}}$$

Results: Cross-Validation Accuracy

Extended Data Table 2 | General validation results

a.

Classifier	Three-way accuracy
Dermatologist 1	65.6%
Dermatologist 2	66.0%
CNN	69.4 ± 0.8%
CNN - PA	**72.1 ± 0.9%**

b.

Classifier	Nine-way accuracy
Dermatologist 1	53.3%
Dermatologist 2	55.0%
CNN	48.9 ± 1.9%
CNN - PA	**55.4 ± 1.7%**

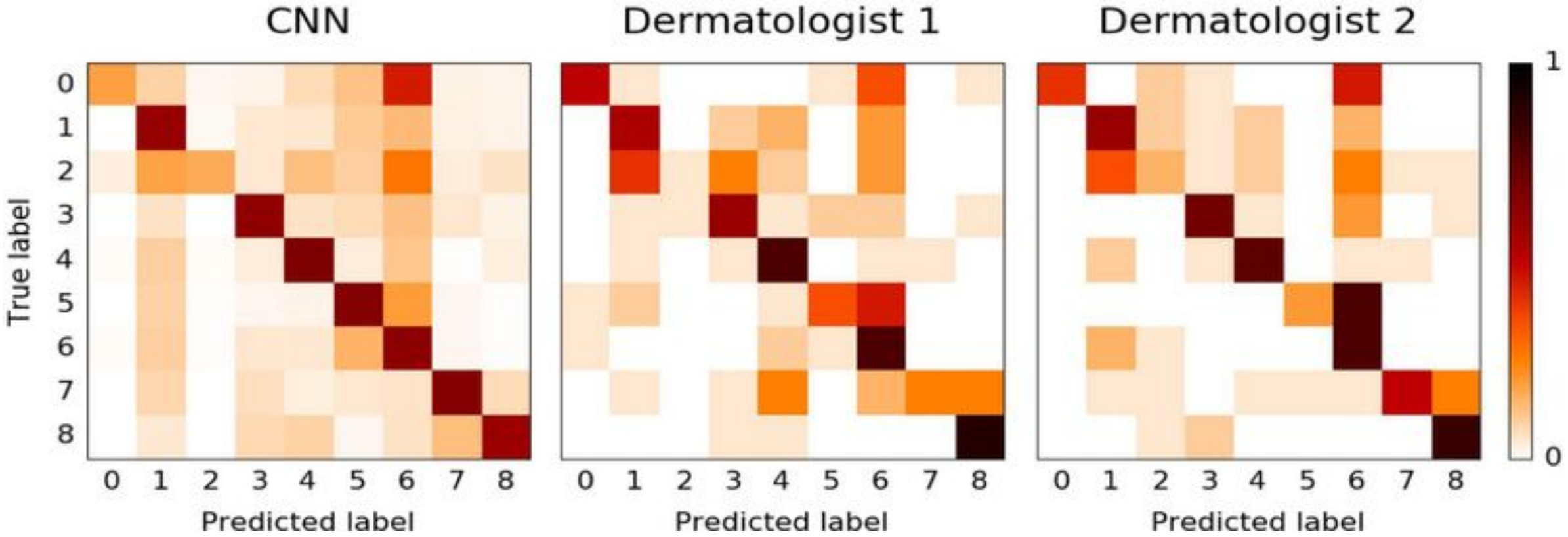
c. Disease classes: three-way classification

0. Benign single lesions
1. Malignant single lesions
2. Non-neoplastic lesions

d. Disease classes: nine-way classification

0. Cutaneous lymphoma and lymphoid infiltrates
1. Benign dermal tumors, cysts, sinuses
2. Malignant dermal tumor
3. Benign epidermal tumors, hamartomas, milia, and growths
4. Malignant and premalignant epidermal tumors
5. Genodermatoses and supernumerary growths
6. Inflammatory conditions
7. Benign melanocytic lesions
8. Malignant Melanoma

Results: Confusion Matrix



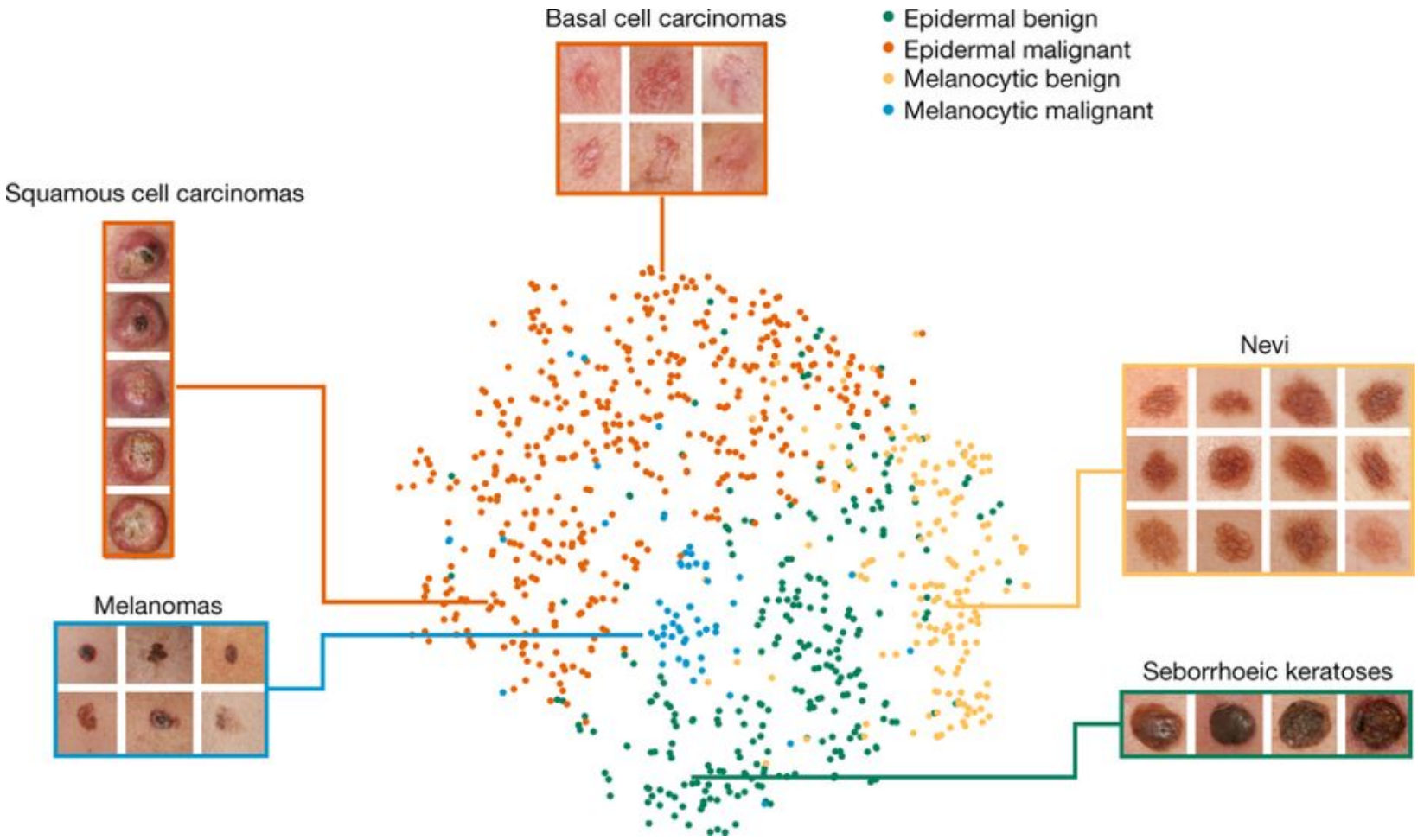
Q: Can't we vary the classification thresholds for multi-class problems as well?

A: Yes, but the common approach is to reduce to two-class problems by combining classes (e.g. class A vs not class A). Multi-class extensions of the ROC curve and AUC exist, but are not commonly used.

How do the authors attempt to look inside the “black box”?

MODEL INTERPRETATION

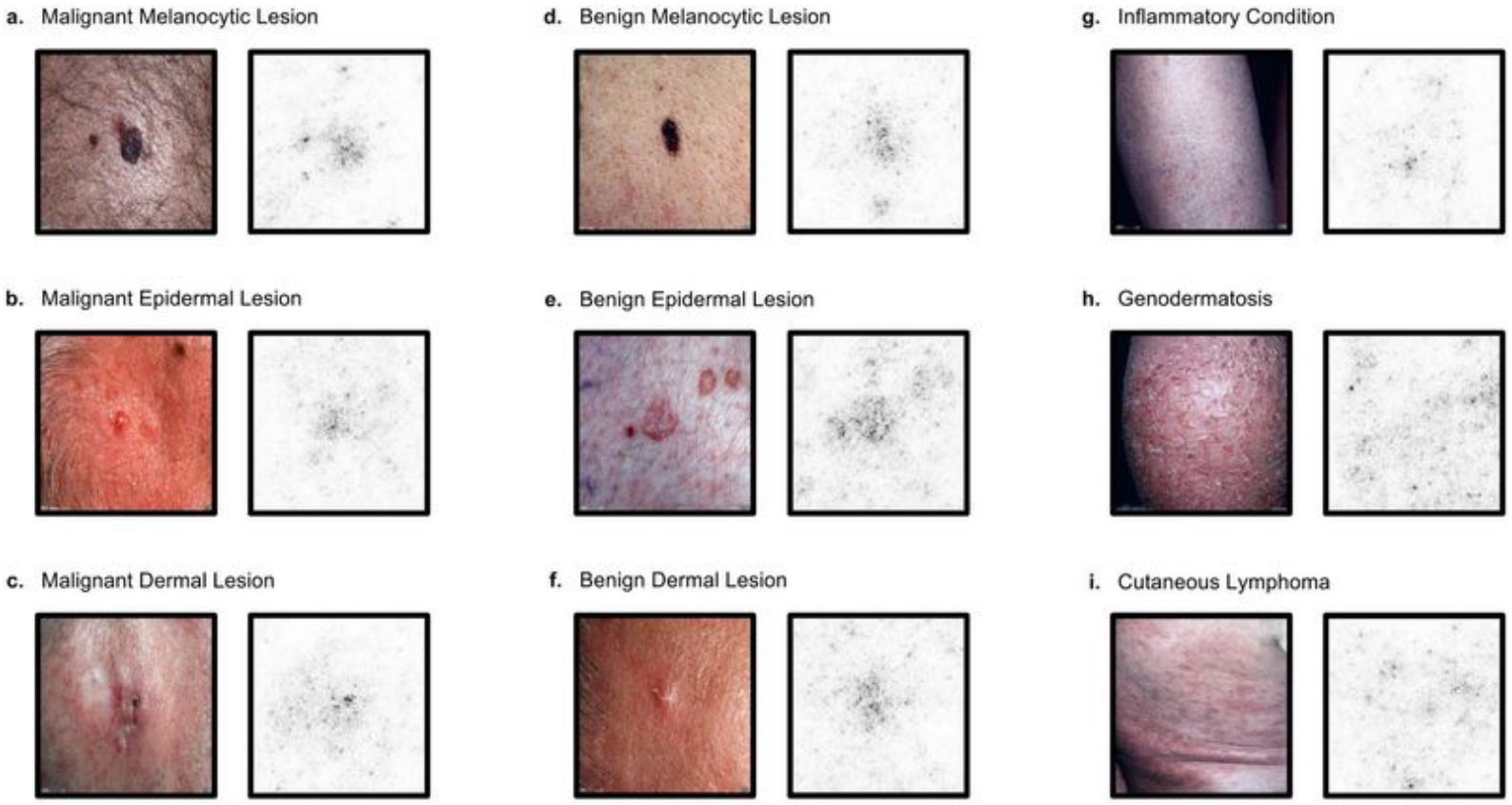
t-SNE visualization of last hidden layer



t-distributed stochastic neighbor embedding (t-SNE) maps high-dimensional points to two-dimensional points such that similarity between pairs of points is (approximately) preserved

Q: How much does this visualization help us understand the model?

Saliency maps for example images



Saliency maps show gradients for each pixel with respect to the CNN's loss function. Darker pixels represent those with more influence.

Q: How much does this visualization help us understand the model?

DISCUSSION: SIGNIFICANCE OF RESULTS

Q1: Is the comparison to dermatologists a fair and/or reasonable one?

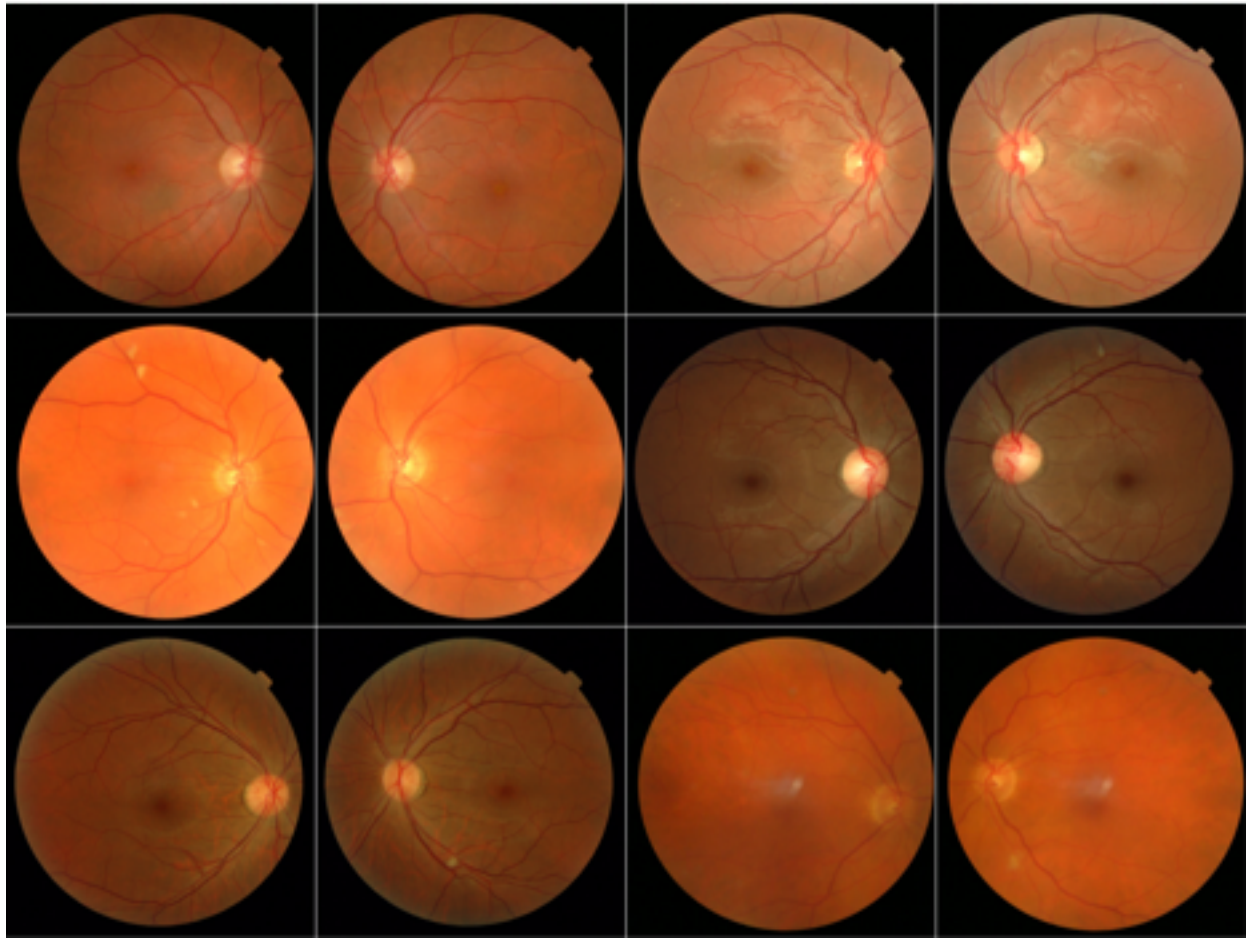
Q2: Are there important methodological limitations?

Q3: What are the barriers to adoption in clinical practice?

Convolutional Neural Networks for Images

SURVEY OF MEDICAL APPLICATIONS

Survey of Applications: Classification



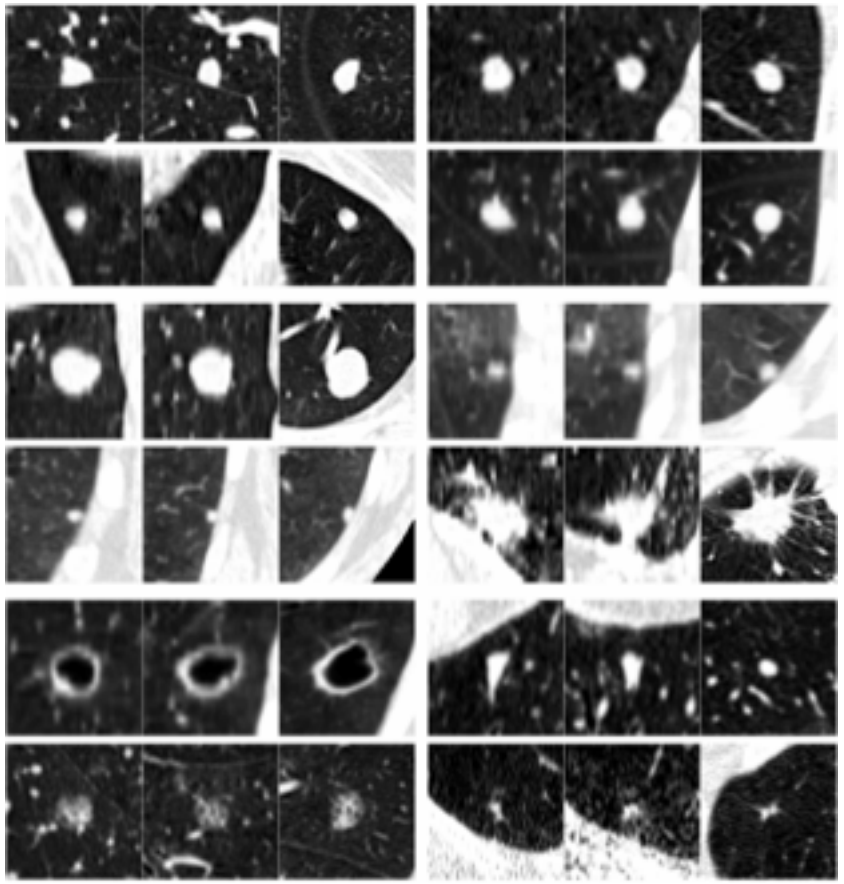
Right and left eye images of the 6/874 subjects diagnosed as having rDR, per the adjudicated consensus of the retinal experts, that were false-negatives for the rDR output (i.e., were missed by the device). All six subjects had consensus grading of moderate DR without ME.

From: Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning

Invest. Ophthalmol. Vis. Sci.. 2016;57(13):5200-5206. doi:10.1167/iovs.16-19964

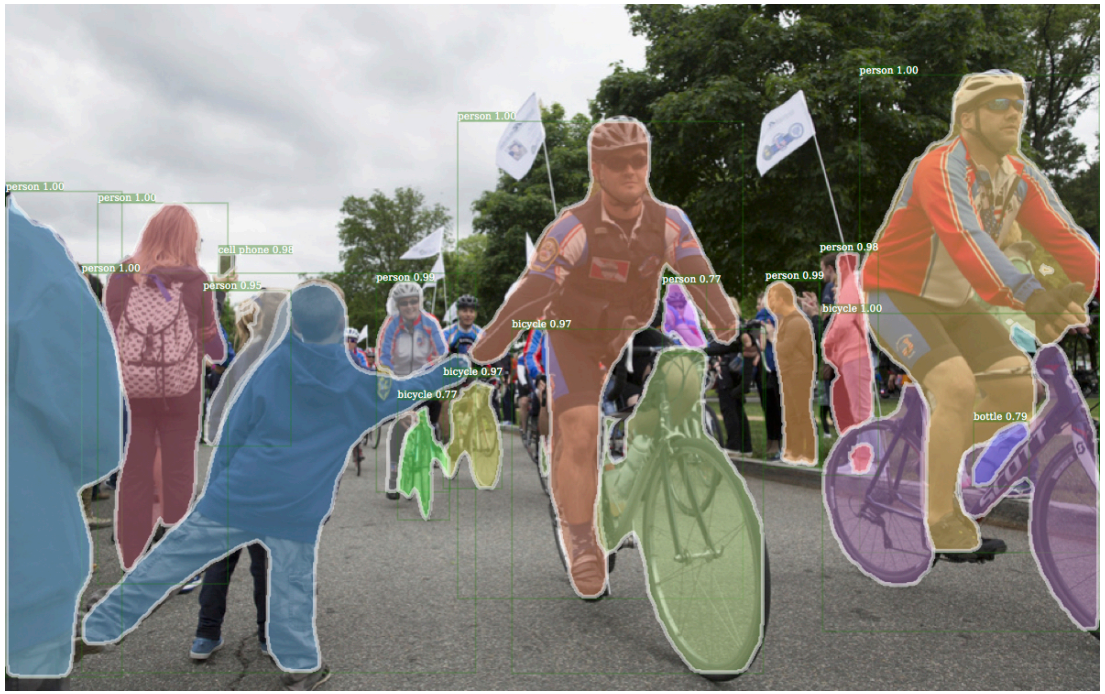
Survey of Applications: Detection

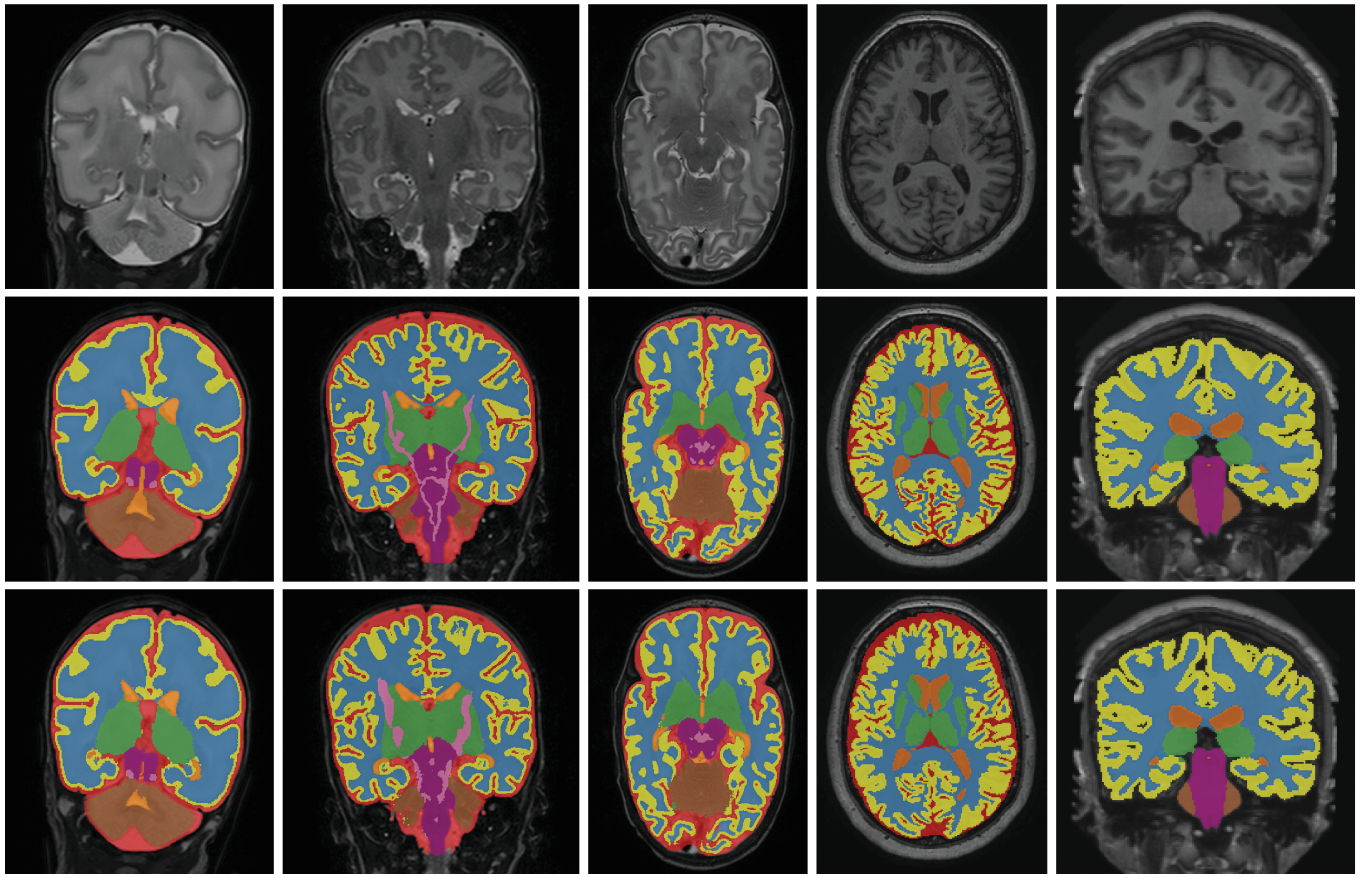
https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/img/kites_detections_output.jpg



van Ginneken B, Setio AA, Jacobs C, Ciompi F. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. InBiomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on 2015 Apr 16 (pp. 286-289). IEEE.

Survey of Applications: Segmentation

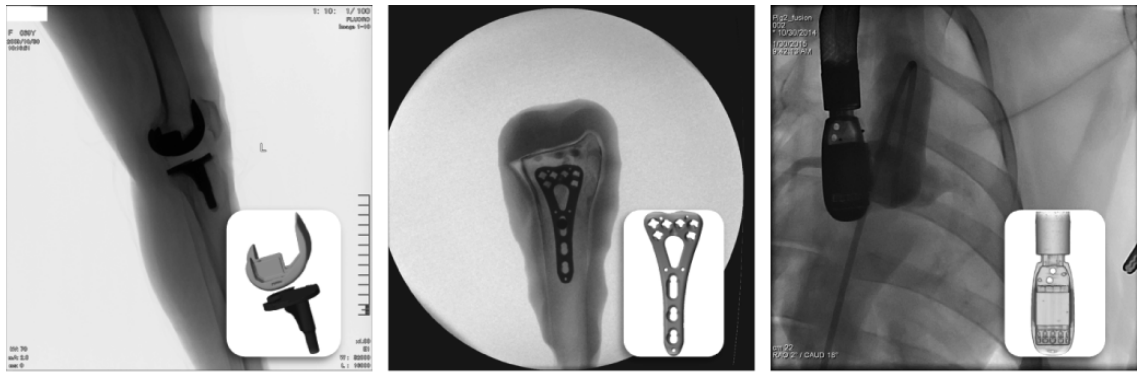




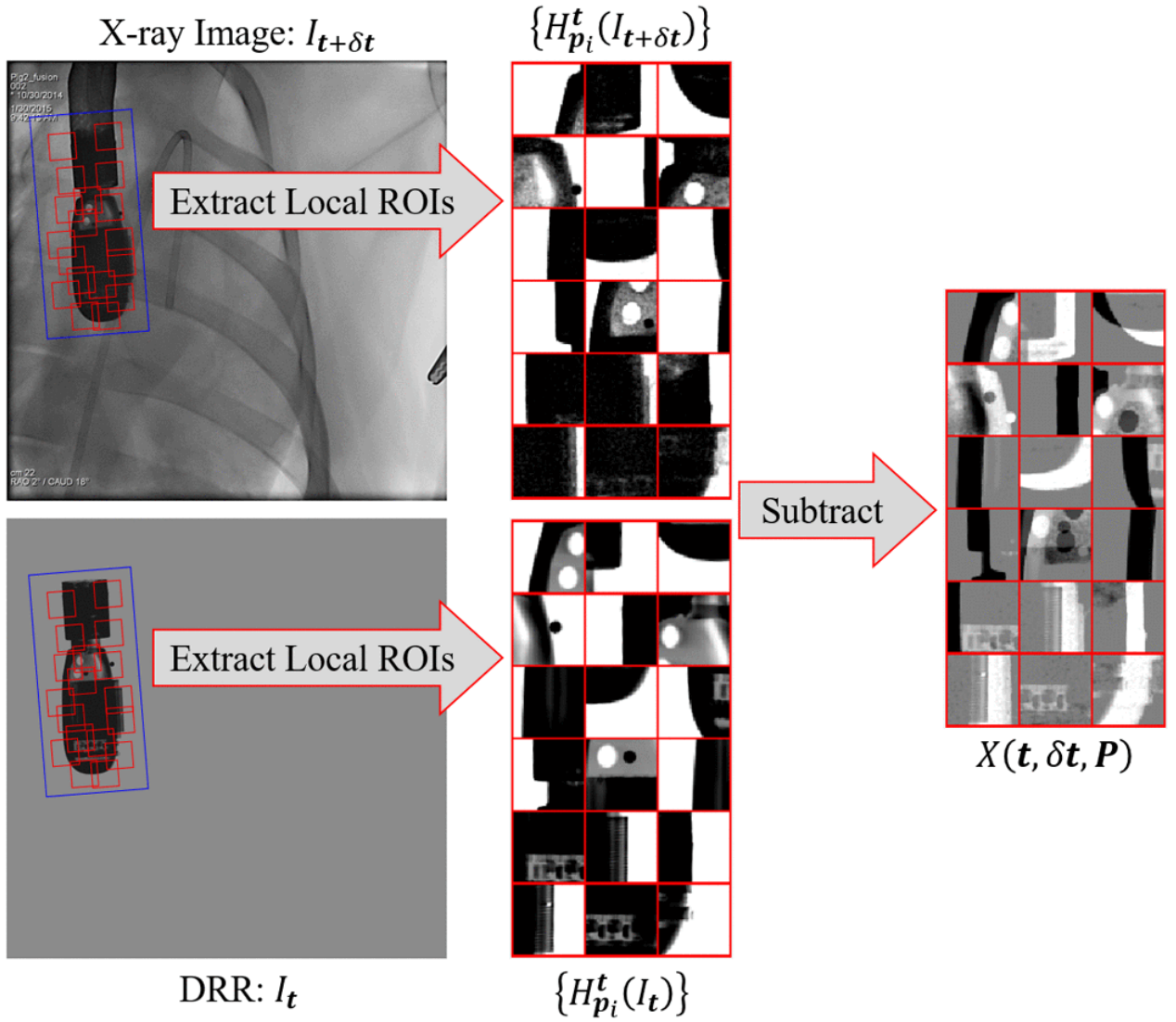
Moeskops P, Viergever MA, Mendrik AM, de Vries LS, Benders MJ, Išgum I. Automatic segmentation of MR brain images with a convolutional neural network. IEEE transactions on medical imaging. 2016 May;35(5):1252-61.

https://github.com/facebookresearch/Detectron

Survey of Applications: Registration



(a) (b) (c)



Miao S, Wang ZJ, Liao R. A CNN regression approach for real-time 2D/3D registration. IEEE transactions on medical imaging. 2016 May;35(5):1352-63.

THANK YOU!

Questions or ideas? Please contact me at m.engelhard@duke.edu