6.S897/HST.956 Machine Learning for Healthcare

Lecture 13: Machine Learning for Mammography

Instructors: David Sontag, Peter Szolovits

1 Breast Image Interpretation: Background and Challenges

Breast cancer affects over 2 million women of the 3.8 billion women globally, and is estimated to contribute to over 40,000 U.S. deaths and 600,000 worldwide deaths annually. The primary method for breast cancer diagnosis is the interpretation of mammography images, which are breast X-ray images. Mammograms are normally acquired from two different angles: craniocaudal (CC, which is a 2D projection of the axial view), and mediolateral oblique (MLO, which is a 2D projection of the frontal coronal view). Ideally, we would like to use these images for early detection of breast cancer, which would allow for more effective treatments and potential cures. It is particularly important to detect cancerous tissue before the non-metatstatic to metastatic transition, after which a cure is no longer feasible. The two primary challenges with mammographic analysis for this purpose are:

- Accurate risk assessment tools
- Effective screening tests

Specifically, the three main problems that need to be addressed to resolve these challenges are:

- No risk assessment models are able to predict individual risk accurately
- There is significant variability in human interpretation of mammograms
- Widespread standardization and application of mammography are limited by the dearth of specialialists.

1.1 Mammogram interpretation

The two questions that radiologists ask when analyzing a mammogram are:

- How dense is the breast tissue?
- Is it a normal or abnormal mammogram?

In a mammogram (an example of which is given in Figure 1, the white pixels represent the breast tissue and black pixels represent fatty tissue inside the breasts; thus, breasts with higher densities of white pixels have higher breast density. The current standard medical practice is for radiologists to bin the tissue density into four categories: fatty (low density), scattered, heterogeneously dense, and dense. However, since cancerous tumors are also small blobs of white pixels, dense breast tissue can sometimes obscure or make it more difficult to detect anomalous cancerous tissues. The difficulty of finding small cancerous tissue in mammograms is illustrated by the case studies in Figure 2.
Because of the 2D nature of traditional mammograms, patients with higher breast densities tend to receive more false negative mammograms, in which cancerous tissue is mistakenly diagnosed as healthy. This is shown in Figure 3, in which a 3D reconstruction of the breast tissue acquired using tomosynthesis (in which optical slices in the axial dimension are taken for 3D tomography) shows that cancerous tissue was missed in the 2D projection due to breast density. Although false negatives are the least frequent type of mammogram, it is particularly devastating because of the consequences it entails for the affected patient, who does not receive the appropriate treatment or care.

[Image removed due to copyright restrictions.]

Figure 1: Mediolateral oblique (MLO) view of a breast mammography with relevant components labeled. [And12]

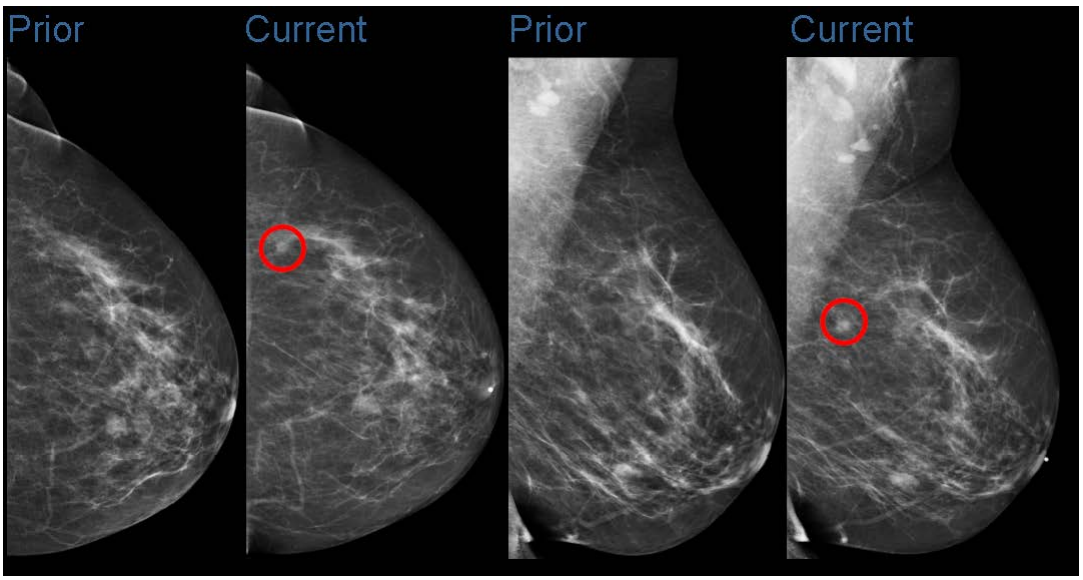

Figure 2: Two examples of time-evolving mammograms with developing cancerous tissue, indicated by the red circles. The cancerous tissue is small and blends in with the rest of the breast tissue.

1.2 Risk assessment

For early detection of breast cancer, it's important for accurate risk assessment tools to be available to interpret mammograms. With more accurate risk forecasting, more action can be taken prior to lymph node localization or metastasis. While false negatives (as described in Section 1.1) are particularly harmful to individuals, false positive mammograms can also have deleterious effects on patients. The side effects of chemotherapy or radiation therapy can be particularly damaging to a patient, especially if they are a false positive. Thus, the preferred method of investigating perceived high-risk patients is through the use of magnetic resonance imaging (MRI), which is an extremely useful but expensive non-invasive imaging tool.

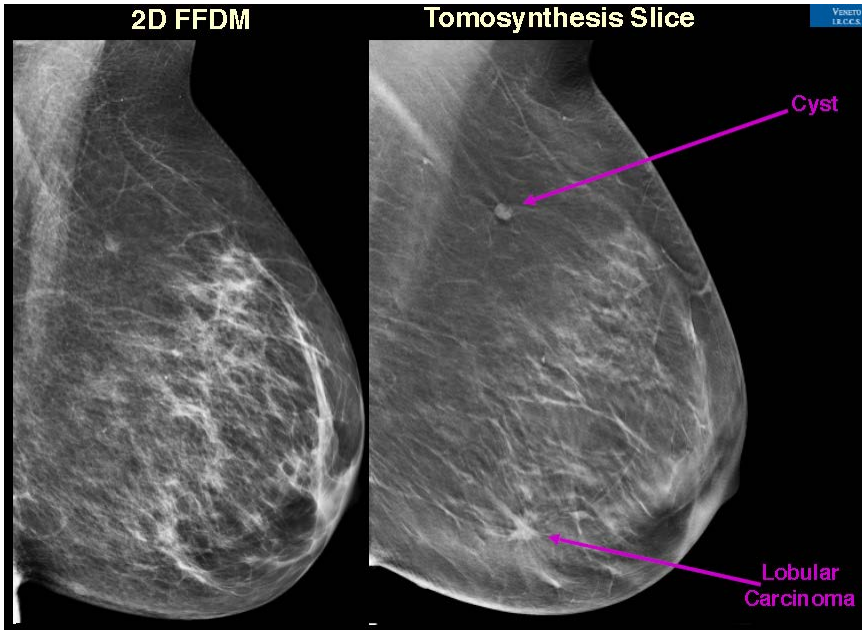

Figure 3: 2D vs. 3D Tomosynthesis mammograms. The 2D projection obscures the cancerous tissues that the tomosynthesis reveals. Images courtesy of Drs. Di Maggio G Gennaro, Istituto Oncologico Veneto I.R.C.C.S. (Padova, Italia)

However, classical risk models are quite limited, in that they use some basic patient features (e.g. age, race, breast density, family history) to generate risk assessment, and oftentimes are quite sensitive to certain variables. Additionally, the subjective categorization of tissue density (as described in Section 1.1) is subject to high variance between different radiologists, making that qualitative feature somewhat unreliable in classical risk models. We can see this exemplified in Figure 4.
The limitations of current clinical risk assessment procedures are elegantly captured by the fact that 75% of all early MRI screens are performed in women with less than 20% lifetime risk, while only 2% of the women with over 20% lifetime risk receive an early MRI. In this sense, the current risk assessment tools are unable to be effective because they are not correctly screening the at-risk population.

2 Deep learning models for mammogram interpretation

The usual process for triaging mammograms can be described by the following steps:

1. Routine screening (1000 patients)
2. Callback for additional imaging (100 patients)
3. Biopsy sample collection (20 patients)
4. Diagnosis / triage (6 patients)

To harness deep learning (DL) for mammogram triaging or risk assessment, we follow a standard procedure for clinical deployment and utilization:

1. Dataset collection
2. Modeling

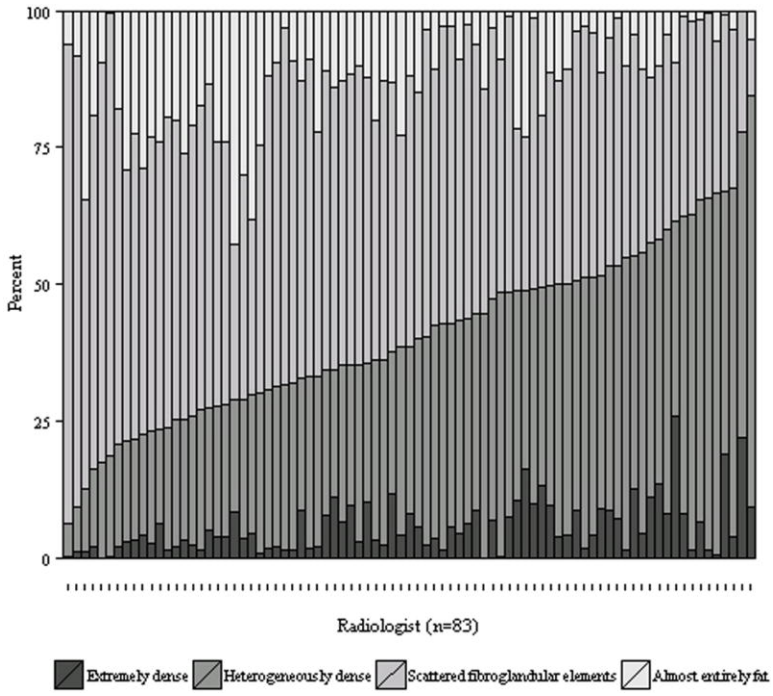

Figure 4: Variation in radiologist assessment of breast density. Radiologists vary in "dense" breast tissue classification rates from 6% to 85%. [SCO+16]

3. Clinical implementation

In this section, we introduce the development of a model for automated mammogram triaging and another model for risk assessment.

2.1 Mammogram triaging

The goal of the mammogram triaging model is to improve efficiency by reducing the false positives of cancer triaging (especially since >99% of patients are cancer-free). Given a cancer-free threshold that is set from training and development of the algorithm, the triaging process can be streamlined for radiologists by having them skip analysis of mammograms below the threshold.

2.1.1 Dataset collection

Data was collected from mammograms in 5 hospital registries, and patient outcomes from Radiology EHR and Partners. These data were pulled from all available screening mammograms between January 2009 and December 2016 (inclusive), with no exclusions made based on race, age, implants or other features of that nature; however, patients with other cancers in the breast were excluded as well as patients with negative exams that did not have a 1-year followup. After these exclusions, there remained 223,109 distinct mammograms from 66,661 unique patients. Since some patients had multiple data points, the training, developmental (dev), and test sets were split by patient to prevent model memorization of patient inputs. The training set contained 1,472 positive exams (1,453 unique patients) and 210,804 negative exams (56,790 unique patients); the development set contained 167 positive exams (163 unique patients) and 25,832 negative exams (7,019 unique patients); the test set contained 191 positive exams (187 unique patients) and 26,349 negative exams (7,170 unique patients).

2.1.2 Challenges of applying DL to mammogram triaging

While many aspects of the task of cancer detection based on mammograms are similar to that of natural image classification (ImageNet), there are some key differences that make this task tricky. First of all, the signal-to-noise-ratio (SNR) is much lower in mammograms compared to natural images. In other words, the cancerous tissue and region that is relevant for image classification is much smaller (tends to be around 1% of the image, exemplified in Figure 2) than the size of the relevant object(s) in natural images.

The images in mammogram datasets are also much larger than natural images. For instance, mammograms tend to be on the order of 3000 by 2600 pixels, whereas ImageNet images tend to be orders of magnitude smaller. This leads to issues with memory and parallelization – in particular, the batch size is limited, which makes the stochastic gradient updates much noisier, and can impede model learning. Large individual images also just presents memory and storage issues in general; for example, the data set for this particular project was over 12 TB. Image size and low SNR ratio both lead to a separate challenge: if patches are used, this can lead to removal of important context, which can make it difficult to differentiate between cancerous and non-cancerous tissues.

Lastly, if patch-based data is used, there is a heavy class imbalance, in that only around 0.7% of the inputs are positive, while the vast majority are negatives. While there are methods for resolving this being actively researched and implemented in the machine learning community, it is still a non-trivial challenge.

2.1.3 Building the model

In building the DL model, the first thign considered was initialization of the model. Adam Yala and colleagues decided to use ImageNet pre-trained CNN models (e.g. VGG, ResNet, Wide-ResNet, DenseNet) to initialize the network, and then to mitigate the small batch size issue (talked about in section 2.1.2) that is a result of the large size of the images, they performed several forward and backpropagation steps through the network before updating the weights via gradient descent. In essence, this allows them to use a larger batch size than GPU memory allows, but doing it in series rather than in parallel. For this, they utilized a batch size of 24 (by taking 2 full steps with batch size 3 distributed over 4 GPU's). They found that initializing with ImageNet allowed for learning to occur much quicker (demonstrated in Figure 5) than when initializing the model randomly. One possible explanation for this is that the batch normalization statistics are more stable to begin with when using pre-trained weights from ImageNet.

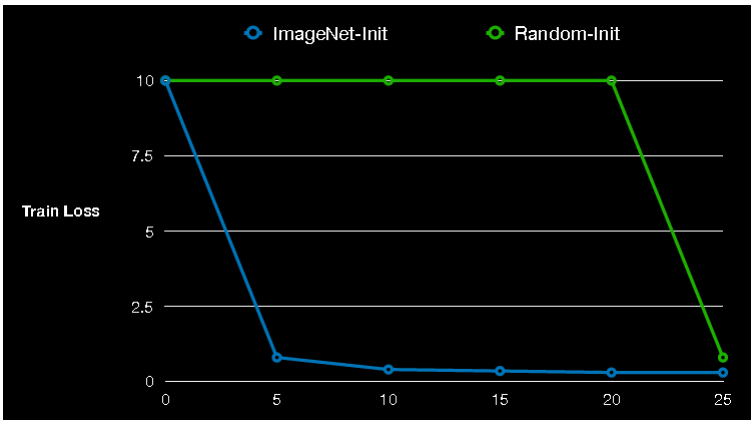

Figure 5: Loss vs. Epoch with ImageNet initialization and random initialization. ImageNet initialization allows for learning to occur quicker.

In selecting a base architecture of the model, they preferred models that were fully convolutional to allow for flexibility in feeding in inputs of varying resolutions; thus, the ResNet-18 model was chosen. This was

especially important in implementing a two-stage training method in which patches from the mammograms were sampled and fed to the network to pre-train the network, and then followed by fine tuning the model using the full-sized images. The researchers found that with the aforementioned serial batch size method, they could resolve the noisy stochastic gradient problem and directly train the model using the full sized images.

Finally, in order to calibrate the class balance to the real incidence of either class, Platt's method is used, in which a sigmoid function is learned to scale and shift the probabilities calculated from the model (based on training data) to the true incidence rate. In this manner, the incidence rate of classes in the training data does not skew the distribution of predictions during inference.

The model triaged by ranking radiologist true positives by the probability assigned by the model, and then returning the minimum probability of radiologist-identified true positive in the development set. To evaluate the best model, instead of using AUC, the triage ability of the model on the dev set was used.

2.1.4 Results

The objectives of the model during analysis were:

- Is the model discriminative across all populations?
- Does the model capture radiologist mistakes or do they overlap?
- Assess the triage on test set

The model achieved an AUC of 0.82 (confidence interval: [0.8, 0.85]) on the dev set, with no statistically significant difference in predictive value between varying ages, races, and breast densities. The main goal of the model was to not miss a single cancer the radiologist would have caught, in order to be used as an aid (and not replacement) for medical professionals. Thus, it is important to assess how the performance compared to radiologist assessments. It was determined that the model was able to triage significant portions of radiologist-assigned true positives, false positives, and true negatives that were below the model's triage threshold, both reducing radiologist labor (on the true positives and negatives) and limiting the false positives (by triaging those below the threshold as cancer-free, shown in Figure 6). These results are summarized quantitatively in Table 1, with the model increasing specificity by 0.7% and reducing the % of mammograms read by radiologist by 20%.

Setting	**Sensitivty (95% CI)**	**Specificity (95% CI)**	**Mammograms Read (95% CI)**
Original Interpreting Radiologist	90.6% (86.7, 94.8)	93.0% (92.7, 93.3)	100% (100, 100)
Original Interpreting Radiologist + Triage	90.1% (86.1, 94.5)	93.7% (93.0, 94.4)	80.7% (80.0, 81.5)

Table 1: Test set triage simulation results.

2.2 Mammogram risk assessment

2.2.1 Dataset collection and model construction

A similar method was followed to build a DL model for mammogram risk assessment. Here, the data set was constructed similarly by using mammogram images from a 5 Hospital Registry and patient outcomes from Radiology EHR and Partners from January 2009 to December 2012 with no exclusions based on race and other factors. Again, patients with other cancers in the breast were excluded, but (different from the

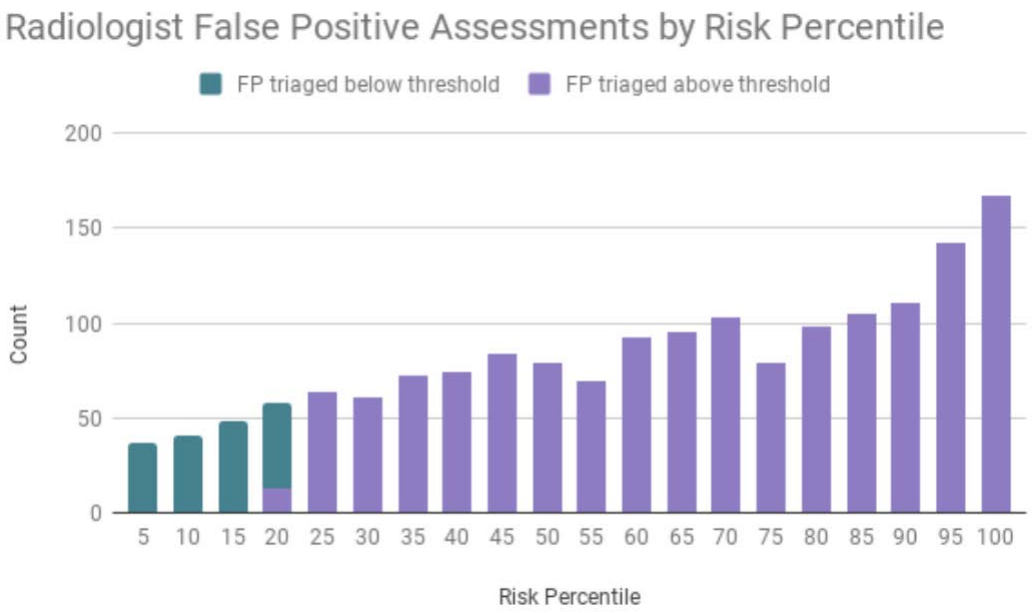

Figure 6: With the cancer-free triage threshold of just above 0.2 model-predicted risk percentile, the radiologist's false positives colored in blue can be eliminated by the automated triaging done by the model.

triage dataset) patients with negative exams which lacked a 5-year followup exam were excluded because the goal was to predict 5-year breast cancer risk. Once again, the train/dev/test sets were split by patient to prevent model memorization. In the end, the train set had 71,689 mammograms (31,806 unique patients), the validation set had 8,554 mammograms (3,804 unique patients), and the test set contained 8,751 mammograms (3,937 unique patients). The test set was constructed with an additional exclusion of excluding exams within one year of a cancer diagnosis. In this manner, the predictive value of the model could be limited to one temporal image per patient.

2.2.2 Results

The objectives of this study were to:

- Is the model discriminative across all populations?
- How does the DL model relate to classical models?

The same considerations for building the DL model from the previous section are used for this one, and the researchers decided to compare the risk assessment performance of (1) the DL model using image only, (2) DL using image and other classical features (e.g. age, race, breast density), and (3) Tyrer-Cuzick (a classical approach). Briefly, the Tyrer-Cuzick model estimates breast cancer risk from various family history, physical statistics, and demographic data. As shown in Figure 7, the hybrid method using DL image + features achieved an AUC of 0.7, whereas the DL image only model achieved an AUC of 0.68, and the traditional feature-based Tyrer-Cuzick achieved an AUC of 0.62.

It is particularly interesting to note that the Tyrer-Cuzick model, since it was trained on only white women, performed especially poorly on risk assessment of African-American women, achieving an AUC of 0.45 (the DL + features model achieved 0.71 AUC, as seen in Figure 8). This is worse than a model that is guessing randomly. Additionally, we see that the DL image + features model is discriminative across different populations by performing significantly better than the Tyrer-Cuzick model on various sub-populations, as

shown in Figure 9. Thus, we see that harnessing deep learning can have real potential to improve the risk assessment models that are used currently.

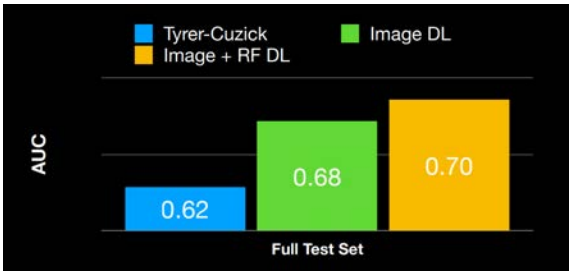

Figure 7: AUC on the test set using various risk models. The deep learning combined with classical features model performed the best.

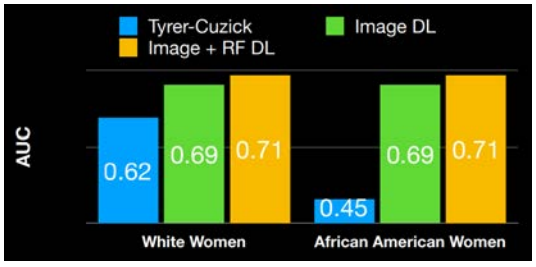

Figure 8: AUC comparison between models on white and African American women sub-populations

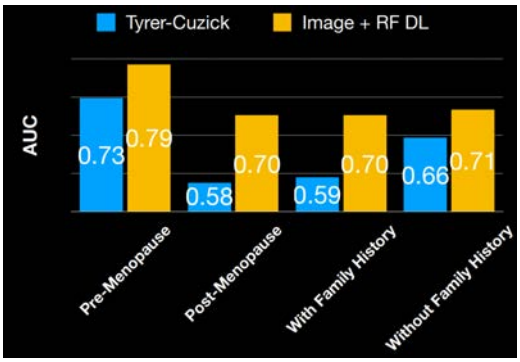

Figure 9: AUC comparison for various sub-populations, showing that the novel image + features DL model outperforms classical approaches.

2.2.3 Potential pitfalls

While deep learning is able to perform particularly well in mammogram interpretation, there are still challenges associated with it before it can achieve widespread clinical implementation. In particular, most data sets collected for mammogram interpretation are enriched, potentially not reflecting the true distribution of the data, and certainly contain biases (e.g. machine-specific collection at a certain hospital). For example, one dataset collected at a hospital in Shanghai had collected all of the cancer-positive data first, before collecting the negatives consecutively. This resulted in a skew dataset when researchers tried to use an

incomplete version of the data set. Another potential pitfall to automated mammogram interpretation is to interpret performance on reader studies as if they were clinical implementations. Reader studies should not be interpreted this way not only because of the dataset pitfalls mentioned previously, but because oftentimes inconvenient cases are excluded from the data set. For instance, a research study might filter out a dataset by breast size, or race, due to insufficient data points, or some other reason of analysis inconvenience. This will certainly further bias the dataset towards certain demographics and types of mammograms, thus making it non-generalizable and unsuitable for clinical implementation.

3 Future outlook

In the past, traditional computer-aided diagnosis (CAD) did not perform well in mammogram interpretation, and was primarily used as a money-driven endeavor. With the development of new imaging modalities such as tomosynthesis (which takes 2D optical sections of the breast to provide a 3D reconstruction), AI technology and hospitals have struggled to keep up. However, combined with ever-increasing amounts of data, recent applications of deep learning towards mammogram triaging and risk assessment have greatly improved upon classical methods, providing a promising outlook of a more efficient and life-saving clinical pipeline.

References

[And12] Savvas Andronikou. *The Breast*, pages 359–375. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[SCO+16] B.L. Sprague, E.F. Conant, T. Onega, E.F. Beaber, S.D. Herschorn, C.D. Lehman, A.N. Tosteson, R. Lacson, M.D. Schnall, D. Kontos, J.S. Haas, D.L. Weaver, W.E. Barlow, and PROSPR Consortium. Variation in mammographic breast density assessments among radiologists in clinical practice: A multicenter observational study. *Ann Intern Med*, 165(7):457–464, 2016.

6.S897 / HST.956 Machine Learning for Healthcare

Spring 2019