BMI 826 Final Project Report: Detection of Ductile Carcinoma in Breast Tissue using Out-of-distribution Detection Frameworks

Demeng Feng Huan Liang Juan Rios
University of Wisconsin-Madison
dfeng24@wisc.edu hliang74@wisc.edu jsrios@wisc.edu

1. Introduction

The application of Deep Learning models such as Convolutional Neural Networks (CNNs) have obtained impressive results on biological image classification. For example, an application of ResNet-50 by Xie *et al.* boasts a true positive rate of 0.92 and a false positive rate of 0.06 on histopathology images of melanoma [29]. Hosny et al. demonstrated using AlexNet a true positive rate of 0.983 and a false positive rate of 0.014 on skin cancer images [26]. These and other methods such as [27], and [4] involve training CNNs using data carefully curated with patch-level annotations from medical doctors. One issue is the difficulty in obtaining physician-annotated images in both quality and quantity. For example [17] discusses the time-consuming and error-prone task for pathologists to manually label whole slides. We constrain this problem setting to detect cancerous tissue in histopathology images without access to patch-level annotations.

2. Related Works

The application of machine learning in whole-slide-imaging (WSI) can be categorized into 2 branches. The earlier work used traditional machine learning or statistics methods (for example, kernalized SVM [24], Random Forest [25], and Adaboost [14] algorithms) with hand-crafted features (for example, color [14] or texture descriptor [2]). However, classical methods did not perform well because they only take limited features into consideration [15].

Recently, deep-learning (DL) based methods are widely used for whose slide imaging (WSI) tasks [15]. DL methods still suffer from the fact that most regions in the (whole) tissue slice image contain less or nearly no information. Therefore, [15] has proposed an attention-based method to let the network learn which regions in the whole-slice-image contains key information. Furthermore, [7] proposed generative-adversarial-network(GAN)-based method, where GAN is trained to produced healthy tissue images, and by comparing those generated images and a new image, people can determine whether the new image contains cancer.

3. Approach

We propose to train a CNN to detect abnormal tissue such as tumors using whole slide annotations by only training to classify different types of normal tissue. Patch-level annotations will be used to evaluate the results. The proposed approach uses Out-of-Distribution (OOD) detection. OOD detection is identifying if an input x does not share the distribution of the training data. Many OOD detection methods have been recently proposed and is an active research area [9], [19], [21], [23], [31], [28], [5]. A number of these methods compute some score $S(x)$. If $S(x)$ is above (or below) some threshold τ, then x is labeled OOD. Ideally, by training to classify only normal types of tissue, the cancerous tissue is absent or rare during training.

3.1. Evaluation Criteria

We measure the performance of OOD detection frameworks by measuring the FPR95 on a test set of images containing benign tumor tissue, ductal carcinoma in situ, and invasive ductal carcinoma. The FPR95 is measured by selecting a threshold score τ where 95% of in-distribution are correctly labeled as such, and determining how many OOD images are incorrectly labeled as in-distribution as done in [9], [19]. The lower the FPR95 the better the performance.

We also qualitatively evaluate the performance on a region of interest (ROI) within a whole slide image via the following approach: Slide the CNN detector across an ROI within the whole slide with overlapping input. The pixels in the ROI will accumulate a score $S(x, y)$ where x is the column and y is the row of a pixel in the ROI . The intent is that pixels in OOD tissue will have on average, a measurably different normalized score $S(x, y)_\mu$ than in-distribution pixels. These normalized scores can be used to build a heat-map of the ROI. In addition, if the score

$S(x,y)_\mu > \tau$, then we can create a binary mask of the ROI signaling in- and out-of distribution pixels. The threshold τ is the same threshold used to measure the FPR95.

3.2. Datasets

Candidate image datasets include (1) lymph nodes with normal and metastatic tissue, [3][18] (2) breast with normal and carcinoma tissue [1], [20], and (3) patches of 19 tissue types such as breast, thyroid, stomach, liver, etc. [6]

The first dataset is crafted into patches from two healthy tissue classes: lymph node and breast tissue. The lymph node patches are obtained from [3], which contains 130 whole slide images of HE stained axillary lymph node specimens. These whole slides are labeled whether metastatic tissue is present. The patches are collected as 224 x 224 images from the non-metastatic whole slides. Patches with more than 50% background are discarded.

The breast tissue patches are obtained from [1] which contain 10 whole slides of breast tissue with benign and ductal carcinoma annotations. Ideally we would extract patches from whole slides simply labeled as not containing any cancerous tissue. However, it was difficult to find such dataset unlike the lymph node whole slides. The annotations provide a mask of the areas labeled by pathologists as: benign, in situ, or invasive ductal carcinoma. Un-annotated regions can be taken to be healthy regions. The breast tissue patches are taken from regions for which any mask does not overlap more than 10% of the area, and background is less than 50%. The breast tissue is also HE stained and at the same magnification level as the lymph node images. The microns per pixel differentiate by a negligible amount. this dataset consist of 32,340 training patches evenly split between breast and lymph tissue.

A second dataset is built from the patches with more than 50% mask area of any mask type and are saved and labeled according to the mask with the largest area. For example, if a patch contains 30% benign and 21% invasive then the patch is labeled as 'benign'. This dataset is primarily used to compute the FPR95 of the out-of-distribution detection frameworks.

A third dataset is crafted from 19 tissue types obtained from [6]. This dataset contains $4,118$ patches which all contain cancerous tissue. the breast patches are replaced by randomly-sampled breast patches from the 2-tissue dataset. Both the 2-tissue and 19-tissue datasets use a $80/20$ train/validation split.

A fourth dataset is obtained from [20], which contains pathologist-annotated whole slide images of breast tissue signaling invasive, in situ ductal carcinoma , and benign tissue along with tumor cellularity estimates. This dataset is primarily used to evaluate the binary masks and heat maps built by the models.

3.3. Implementation

Throughout the paper we use ResNet-18 [8] as well as WideResNet-50-2 [30] using PyTorch and other standard python libraries.

4. Out-of-Distribution Detection

This section summarizes the three out-of-distribution detection techniques used: Energy-Based [19], Outlier Exposure [10], and ODIN [16].

4.1. Energy-Based

Energy-Based detection is grounded on energy based models [13]. the score $S(x)$ is referred to as the energy score, where the lower the energy, the lower the probability a sample x is out-of-distribution and a higher energy the higher the probability. computing $S(x)_e$ is as follows:

$$S(x,f)_e = -T \cdot log \sum_i^k e^{f_i(x)/T} \tag{1}$$

where k is the number of classes, $f_i(x)$ is the pre-softmax output of the i^{th} neuron in the output layer refered to as a logit, and T is a temperature parameters held constant to be 1. One can use equation (1) using a trained model. This modality is referred to as *at-inference* detection. A second modality termed *energy-tuning* allows the model to be trained with a loss function that helps shape the energy-surface of the model. This loss function is: $L_{\text{total}} = L_{\text{ce}} + \lambda \cdot L_{\text{energy}}$, where L_{ce} is the standard cross entropy loss, and L_{energy} is the energy loss defined as:

$$\begin{aligned} L_{energy} = \max(0, E(x_{in}) - m_{in})^2 + \\ \max(0, m_{out} - E(x_{out}))^2 \end{aligned} \tag{2}$$

where λ is a regularization constant taken to be 0.1, and $E(x)$ is the energy-scored from equation (1), x_{in} is the training sample, x_{out} is a randomly selected unlabeled image from an auxiliary out-of-distribution dataset. In this paper we use Imagenet [22]. Finally, m_{in} and m_{out} are the mean energy scores for the training and auxiliary dataset respectively, computed using a normally-trained model and the at-inference modality. Equation (2) is a double-hinge loss function than punishes the model when x_{in} has too high energy, and/or when x_{out} has too low energy.

4.2. ODIN

ODIN, or Out-of-Distribution Image Detection, utilizes two components to differentiate in-distribution samples from the out-of distribution samples. The first is temperature scaling, where the softmax is computed with a temperature parameter, shown below:

$$S_i(x,T) = \frac{exp(f_i(x)/T)}{\sum_{j=1}^{N} exp(f_j(x)/T)} \quad (3)$$

Using the temperature adjusted outputs for each class, the softmax score is then defined as maximum softmax output across all classes, such that $S(x,T) = max_i S_i(x,T)$. These scores aggregated for all images then create the distribution themselves.

The other component that helps detect out-of-distribution images is by perturbing the input images slightly, given by:

$$\tilde{x} = x - \epsilon sign(-\nabla_x log S_{\hat{y}}(x,T)) \quad (4)$$

The aim of the perturbation is to increase the softmax score for both in-distribution and out-of-distribution images, where the perturbation increases the in-distribution scores more than the out-of-distribution scores. This separates the distribution even more.

4.3. Outlier-exposure

Outlier-exposure approach modifies the regular loss function of a neural network, such that the neural network can predict in-distribution data with high confidence while doing random guess for out-of-distribution data [11]. For maximum softmax probability output, the loss function is:

$$L = \mathbb{E}_{(x,y)\sim D_{in}}[-\log f_y(x)] + \lambda \mathbb{E}_{(x,y)\sim D_{out}}[H(\mathcal{U}; f(x))] \quad (5)$$

where D_{in} and D_{out} are sets of in-distribution and out-of-distribution data, respectively. $f_y(x)$ is the output of label y for sample x, $\mathcal{H}$ is the cross-entropy loss, $\mathcal{U}$ is the distribution of out-of distribution data, which is assumed to be the uniform distribution, and $f(x)$ is the vector of softmax output for data x. Here, $\lambda = 0.5$. For out-of-distribution, the outlier-exposure score is $-\max_c f_c(x)$. It is expected that for in-distribution data, the score is close to -1, and for out-of-distribution data, the score is higher.

5. Results

This section summarizes results for each out-of-distribution method by showing the FPR95 for 3 categories of OOD data: benign, in situ, and invasive.

5.1. Energy-Based

We vary 4 binary configurations to give a total of 16 combinations: (1) using the 2-class versus 19-class dataset, (2) using at-inference versus the energy-tuning modality, (3) using pre-trained versus randomly initialized weights, and (4) using ResNet-18 versus WideResNet-50-2. In this method, the best (lowest) FPR95 scores come from training with the 2-tissue set, and the at-inference modality. Pre-trained starting weights provides better benign tissue FPR95, while random starting weights provide better in-situ FPR95. The 4 best outcomes are shown in Table 1.

Table 1. The FPR95 for ductile carcinoma in situ, ductile carcinoma invasive, and benign tumors using energy-based out-of-distribution detection on breast tissue

Energy-Based FPR95				
Model	Weights	Benign	In Situ	Invasive
ResNet-18	pre-trained	0.32	0.52	0.86
	random	0.54	0.48	0.79
WideResNet-50-2	pre-trained	0.21	0.38	0.62
	random	0.36	0.39	0.72

Table 2. The FPR95 for ductile carcinoma in situ, ductile carcinoma invasive, and benign tumors using ODIN detection on breast tissue

ODIN FPR95				
Model	Weights	Benign	In Situ	Invasive
ResNet-18	pre-trained	0.88	0.92	0.95
	random	0.93	0.93	0.94
WideResNet-50-2	pre-trained	0.90	0.93	0.95
	random	0.92	0.91	0.89

The best performance comes from WideResNet50-2 starting with pre-trained weights on the 2-class dataset, and using at-inference detection. Figure 1 visualizes the distribution of energy-scores for the healthy tissue along with invasive ductile carcinoma when FPR95 is at 0.62.

5.2. ODIN

In this method, the best (lowest) FPR95 scores also come from training with the 2-tissue set, both using ResNet-18 and WideResNet-50-2, although the scores for the 2-tissue model are rather poor compared to the other methods. Similar to Energy-Based, pre-trained starting weights provides better benign tissue FPR95. The 4 best outcomes using combinations of ResNet vs WideResNet and pre-trained vs random weights are shown in Table. 2.

The best performance comes from WideResNet50-2 with random weights and ResNet-18 with pre-trained weights. However, the FPR-95 scores from each combination shown are so similar to each other that random error may have caused the slight differences of scores. A reason for the poor performance may be due to the lack of disparity between the in-distribution and out-of-distribution data sets (cancerous and healthy tissue).

5.3. Outlier-Exposure

For this method, the best (lowest) FPR95 score comes from training with 2-tissue set, both using ResNet-18 and WideResNet-50-2. The 4 best outcomes using combinations of ResNet vs WideResNet and pre-trained vs random

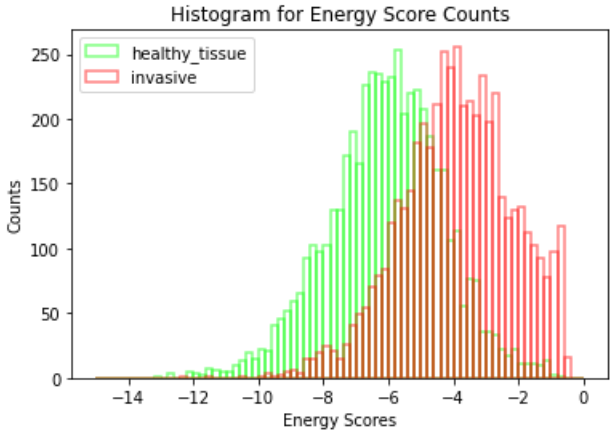

Figure 1. Distribution in energy scores for in-distribution (healthy breast, lymph nodes) and out-of-distribution (invasive ductal carcinoma) image-patches calculated using an energy-based framework

Table 3. The FPR95 for ductile carcinoma in situ, ductile carcinoma invasive, and benign tumors using outlier-exposure image detection on breast tissue

Outlier-Exposure FPR95				
Model	Weights	Benign	In Situ	Invasive
ResNet-18	pre-trained	0.71	0.46	0.88
	random	0.65	0.36	0.66
WideResNet-50-2	pre-trained	0.32	0.49	0.85
	random	0.46	0.31	0.32

weights are shown in Table. 3. The best performance comes from WideResNet50-2 starting with random weights on the 2-class dataset.

5.4. Binary masks and Heat Maps

Because of the poor results of ODIN, we only consider energy-based and Outlier-Exposure frameworks when building the binary mask over an ROI. Figure 2 in the appendix shows the binary mask applied over a $2,000$ x $4,000$ pixel ROI. This ROI contains in situ ductal carcinoma along with normal fat cells. The light-shaded regions are pixels labeled as OOD by either energy-based or outlier-exposure. The darker-shaded regions are labeled OOD by both frameworks. Although the in situ carcinoma is largely inside the shaded region, so are the fat cells, which are normal. Additionally, the connective tissue throughout the ROI is unlabeled and may potentially be invasive carcinoma. For reference, figure 3 in the appendix shows the mask over an ROI labeled as healthy. This masking does well in not masking the healthy lobules, but continues to mask fat cells as OOD.

The appendix also shows heat maps generated by the energy-based framework. Energy-based heat map is shown because the distribution of scores is more Gaussian-like and easier to show, while in outlier exposure, the in-distribution scores are narrow compared to the long-tailed OOD scores.

6. Challenges

One challenge is that these OOD frameworks are designed to detect OOD data whose semantics are substantially far from the in-distribution. For example energy-based distribution will consider CIFAR-10 (frog, cats, airplanes) [12] as in-distribution, while OOD datasets include SHVN (house numbers from street view), and Places365 (scenic images). Detecting cancer tissue resembling normal tissue is a much harder task because cancer tissue is embedded in normal tissue. This problem setting can be identified as near-OOD detection [5]. The authors of [5] provides some possible future research opportunities to explore as discussed in the 'Future work' section.

Another challenge is the collection of a representative dataset. We saw that fat cells are masked as OOD tissue while in reality these cells are normally found in breast tissue. One solution is to build a dataset that has equal representation of normal structures such as fat cells, which are not nearly as prevalent as the pink connective tissue making up most of the breast images.

7. Future work

The authors of [5] show promising results with the use of vision transformers rather than CNNs in a near-OOD detection setting. The main reason stated is that vision transformers can better distinguish class embeddings derived from the penultimate (pre-logit) layer of a model. [5] uses CIFAR-10 as in-distribution and CIFAR-100 as out-of-distribution and achieve a AUROC of 99.6% using a pre-trained visual transformer with 1-shot Outlier Exposure. Using vision transformers in cancer detection settings lay the foundation for continuing research. Combining vision transformers with a more carefully curated dataset may give better in terms of lower FPR95s along with not mislabeling fat cells as OOD.

8. Conclusion

We train a model to differentiate between healthy breast and healthy lymph node tissue. We employ 3 out-of-distribution detection frameworks (energy-based, outlier-exposure, and ODIN) to measure the FPR95 over a set of OOD data (benign tissue, invasive, and in situ ductal carcinoma). We measure the effectiveness of these frameworks using FPR95 score, and we build both binary masks shading pixels considered out-of-distribution along with a heat map of energy scores.

All authors contributed equal to this final project. Huan was in charge of implementing ODIN, Demeng was in charge of implementing outlier exposure, and Juan was in charge of implementing energy-based OOD detection.

References

[1] Guilherme Aresta et al. “BACH: Grand challenge on breast cancer histology images”. In: *Medical Image Analysis* 56 (2019), pp. 122–139. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2019.05.010.

[2] Sevcan Aytaç Korkmaz and Hamidullah Binol. “Classification of molecular structure images by using ANN, RF, LBP, HOG, and size reduction methods for early stomach cancer detection”. In: *Journal of Molecular Structure* 1156 (2018), pp. 255–263. ISSN: 0022-2860. DOI: https://doi.org/10.1016/j.molstruc.2017.11.093. URL: https://www.sciencedirect.com/science/article/pii/S0022286017315818.

[3] Hanna Campanella, M.G. Andrea Brogi, and T.J. Fuchs. “Breast Metastases to Axillary Lymph Nodes [Data set]”. In: (2019). DOI: https://doi.org/10.7937/tcia.2019.3xbn2jcc.

[4] Andre Esteva et al. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *Nature* (2017). doi:10.1038/nature21056.

[5] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. *Exploring the Limits of Out-of-Distribution Detection*. 2021. arXiv: 2106.03004 [cs.LG].

[6] Jevgenij Gamper et al. “PanNuke Dataset Extension, Insights and Baselines”. In: *arXiv preprint arXiv:2003.10778* (2020).

[7] Changhee Han et al. “MADGAN: unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction”. In: *BMC Bioinformatics* 22.2 (Apr. 2021), p. 31. ISSN: 1471-2105. DOI: 10.1186/s12859-020-03936-1. URL: https://doi.org/10.1186/s12859-020-03936-1.

[8] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[9] Dan Hendrycks and Kevin Gimpel. *A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks*. 2018. arXiv: 1610.02136 [cs.NE].

[10] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. *Deep Anomaly Detection with Outlier Exposure*. 2019. arXiv: 1812.04606 [cs.LG].

[11] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. *Deep Anomaly Detection with Outlier Exposure*. 2019. arXiv: 1812.04606 [cs.LG].

[12] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. “CIFAR-10 (Canadian Institute for Advanced Research)”. In: (). URL: http://www.cs.toronto.edu/~kriz/cifar.html.

[13] Yann LeCun et al. “A tutorial on energy-based learning”. In: *PREDICTING STRUCTURED DATA*. MIT Press, 2006.

[14] Jiayun Li et al. *A Multi-resolution Model for Histopathology Image Classification and Localization with Multiple Instance Learning*. 2020. arXiv: 2011.02679 [eess.IV].

[15] Yixin Li et al. *A Hierarchical Conditional Random Field-based Attention Mechanism Approach for Gastric Histopathology Image Classification*. 2021. arXiv: 2102.10499 [cs.CV].

[16] Shiyu Liang, Yixuan Li, and R. Srikant. *Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks*. 2020. arXiv: 1706.02690 [cs.LG].

[17] Min Ling et al. “Fast Whole Slide Image Analysis Of Cervical Cancer Using Weak Annotation”. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. 2021, pp. 1037–1041. DOI: 10.1109/ISBI48211.2021.9433964.

[18] Geert Litjens et al. “1399 H amp;E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset”. In: *GigaScience* 7.6 (May 2018). giy065. ISSN: 2047-217X. DOI: 10.1093/gigascience/giy065. eprint: https://academic.oup.com/gigascience/article-pdf/7/6/giy065/25045131/giy065.pdf. URL: https://doi.org/10.1093/gigascience/giy065.

[19] Weitang Liu et al. *Energy-based Out-of-distribution Detection*. 2021. arXiv: 2010.03759 [cs.LG].

[20] Mohammad Peikari et al. “Automatic cellularity assessment from post-treated breast surgical specimens”. In: *Cytometry Part A* 91.11 (2017), pp. 1078–1087. DOI: https://doi.org/10.7937/TCIA.2019.4YIBTJNO.

[21] Jie Ren et al. *Likelihood Ratios for Out-of-Distribution Detection*. 2019. arXiv: 1906.02845 [stat.ML].

[22] Olga Russakovsky et al. “ImageNet Large Scale Visual Recognition Challenge”. In: *CoRR* abs/1409.0575 (2014). arXiv: 1409.0575. URL: http://arxiv.org/abs/1409.0575.

[23] Chandramouli Shama Sastry and Sageev Oore. *Detecting Out-of-Distribution Examples with In-distribution Examples and Gram Matrices*. 2020. arXiv: 1912.12510 [cs.LG].

[24] Harshita Sharma et al. "Appearance-based necrosis detection using textural features and SVM with discriminative thresholding in histopathological whole slide images". In: *2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE)*. 2015, pp. 1–6. DOI: 10.1109/BIBE.2015.7367702.

[25] Harshita Sharma et al. "Cell nuclei attributed relational graphs for efficient representation and classification of gastric cancer in digital histopathology". In: *SPIE Medical Imaging*. 2016.

[26] *Skin Cancer Classification using Deep Learning and Transfer Learning — IEEE Conference Publication — IEEE Xplore*. URL: https://ieeexplore.ieee.org/document/8641762 (visited on 10/11/2021).

[27] Bastiaan S Veeling et al. "Rotation Equivariant CNNs for Digital Pathology". In: (June 2018). arXiv: 1806.03962 [cs.CV].

[28] Jim Winkens et al. *Contrastive Training for Improved Out-of-Distribution Detection*. 2020. arXiv: 2007.05566 [cs.LG].

[29] Peizhen Xie et al. "Interpretable Classification from Skin Cancer Histology Slides Using Deep Learning: A Retrospective Multicenter Study". In: (2019). arXiv: 1904.06156 [q-bio.TO].

[30] Sergey Zagoruyko and Nikos Komodakis. *Wide Residual Networks*. 2017. arXiv: 1605.07146 [cs.CV].

[31] Hongjie Zhang et al. *Hybrid Models for Open Set Recognition*. 2020. arXiv: 2003.12506 [cs.CV].

9. Appendix

a sample of images with an applied binary mask or heat mask.

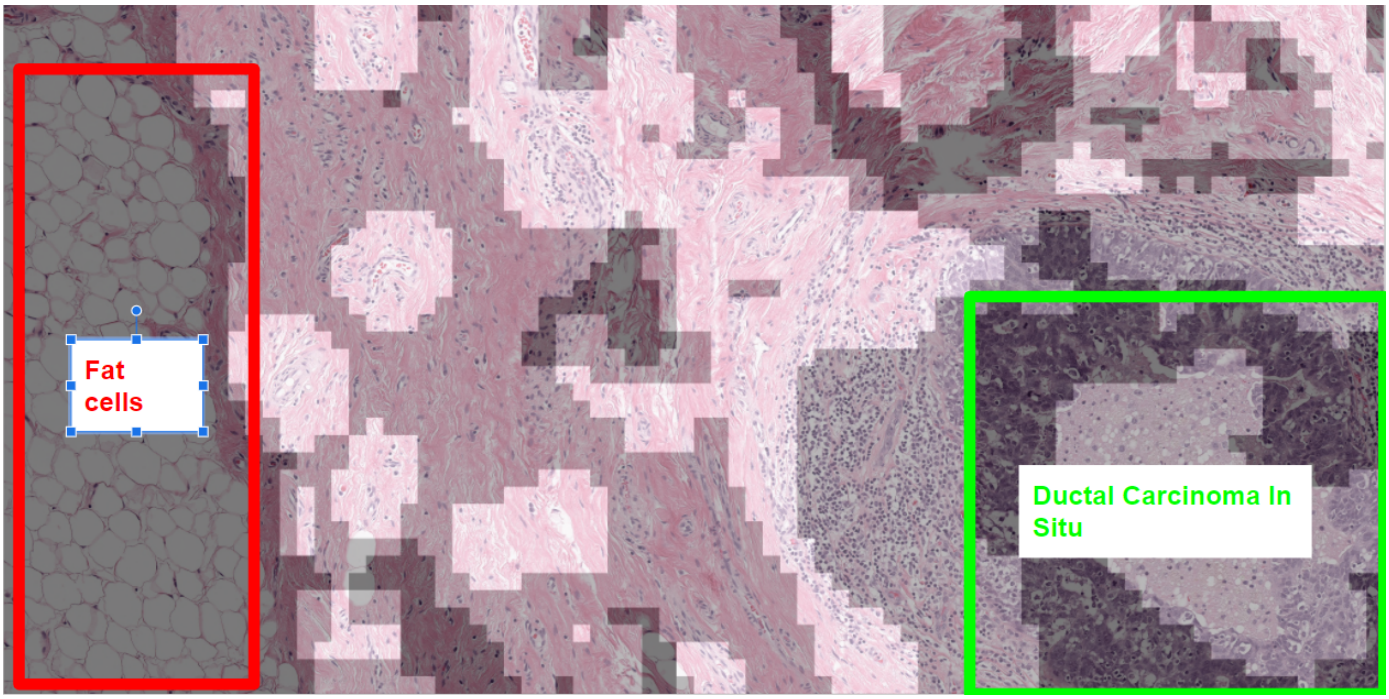

Figure 2. Masking of OOD pixels, where light-shaded regions were labeled OOD by either energy-based or outlier-exposure, and dark-shaded regions are agreed to be OOD by both frameworks. The majority of the tissue outside the bounding boxes are not explicitely labeled and may contain either healthy tissue or some degree of invasive carcinoma.

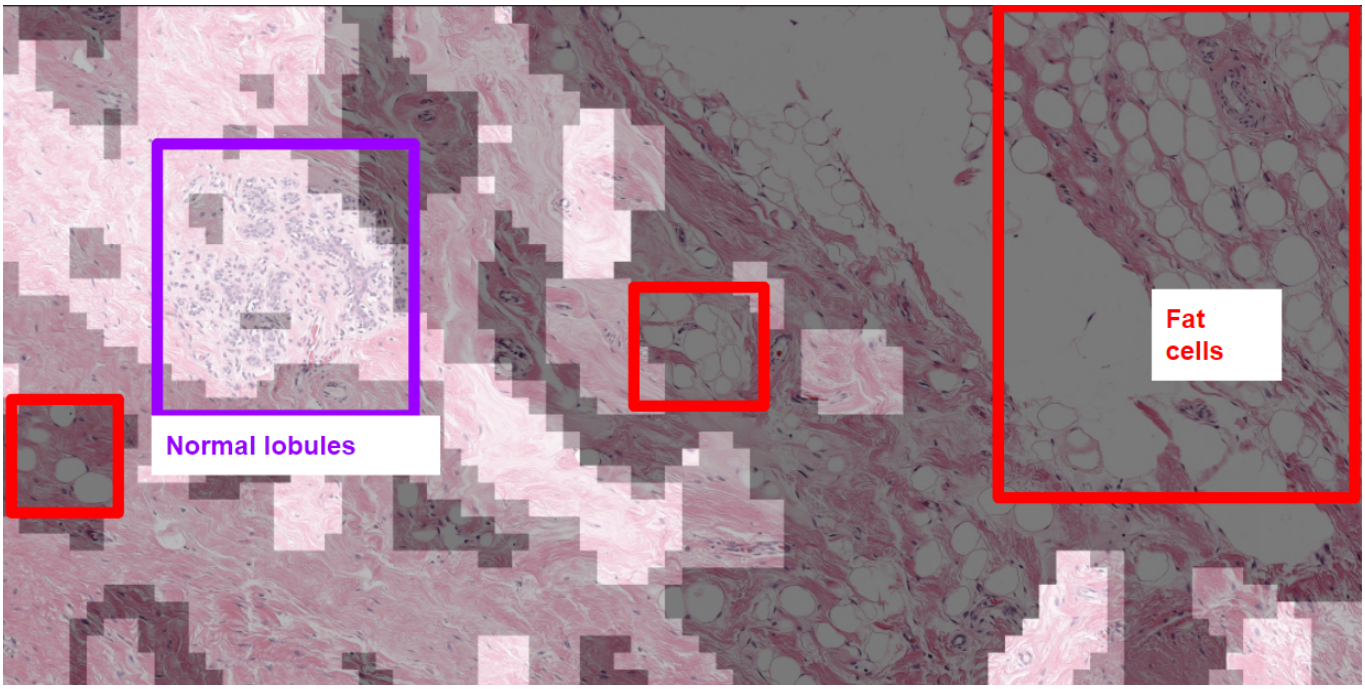

Figure 3. Masking of OOD pixels, where light-shaded regions were labeled OOD by either energy-based or outlier-exposure, and dark-shaded regions are agreed to be OOD by both frameworks. This is labeled normal tissue. Normal fat cells are mostly labeled OOD. Additionally, some of the connective tissue is also lightly-shaded. Normal lobules are largely not shaded thus labeled as in-distribution

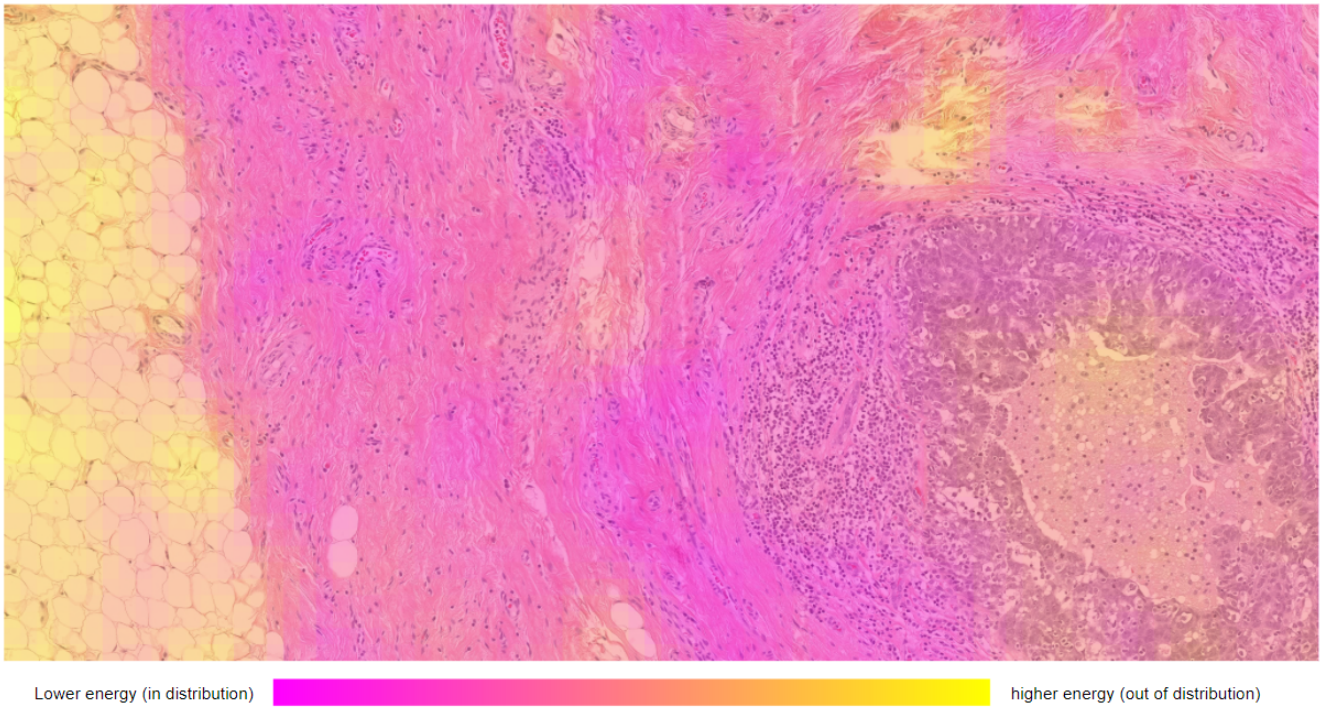

Figure 4. Heat map of averaged energy scores over an ROI. Both normal fat cells and the in situ tissue contain high energies (out of distribution). The energy scores range from -10 to 0

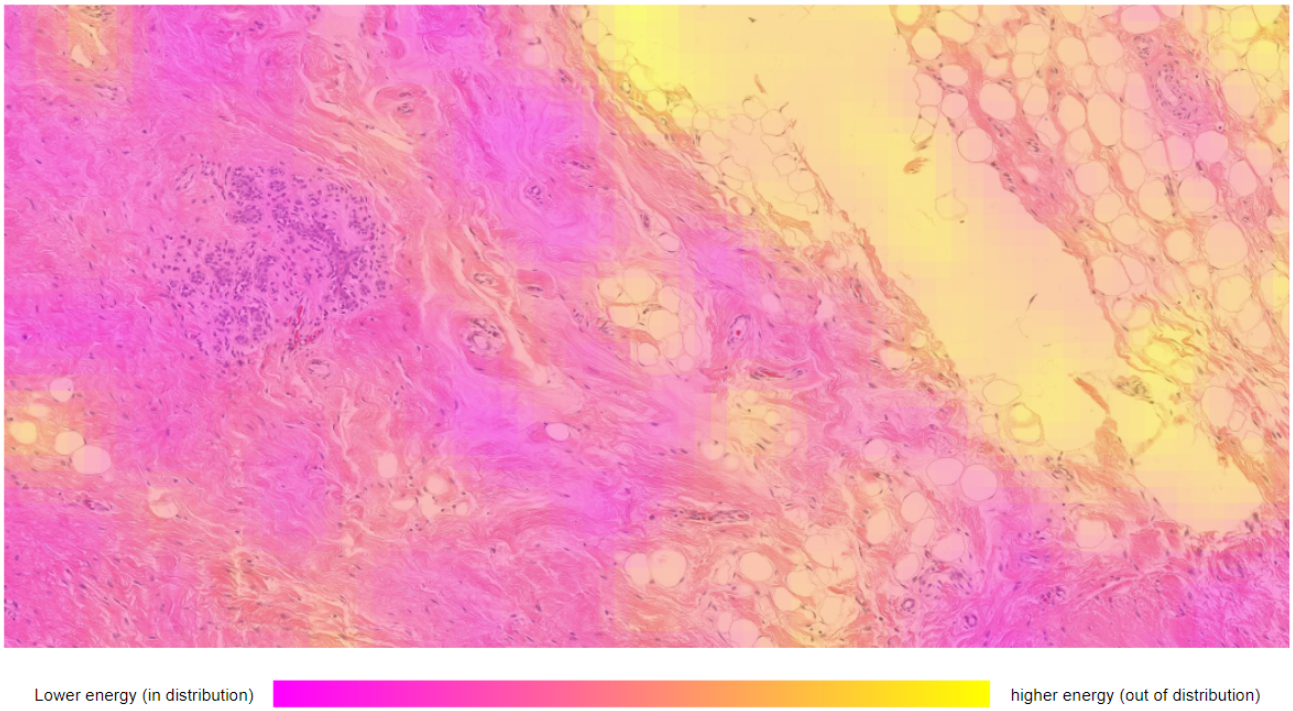

Figure 5. Heat map of averaged energy scores over an ROI. Normal fat cells have incorrectly assigned high energies, while connective tissue and normal lobules have correctly assigned low energies. Ideally this whole image should be low energy. The energy scores range from -10 to 0.