Prototypical Models for Classifying High-Risk Atypical Breast Lesions

Akash Parvatikar[1], Om Choudhary[1], Arvind Ramanathan[2], Rebekah Jenkins[1], Olga Navolotskaia[3], Gloria Carter[3], Akif Burak Tosun[4], Jeffrey L. Fine[3], and S. Chakra Chennubhotla[1,4(✉)]

[1] Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, USA
{akp47,opc3,rcj17,chakracs}@pitt.edu
[2] Data Science and Learning, Argonne National Laboratory, Lemont, IL, USA
ramanathana@anl.gov
[3] Department of Pathology, UPMC Magee-Womens Hospital, Pittsburgh, USA
{navolotskaiao,finejl}@upmc.edu, cartgj@mail.magee.edu
[4] SpIntellx Inc., Pittsburgh, USA
{burak,chakra}@spintellx.com

Abstract. High-risk atypical breast lesions are a notoriously difficult dilemma for pathologists who diagnose breast biopsies in breast cancer screening programs. We reframe the computational diagnosis of atypical breast lesions as a problem of prototype recognition on the basis that pathologists mentally relate current histological patterns to previously encountered patterns during their routine diagnostic work. In an unsupervised manner, we investigate the relative importance of ductal (global) and intraductal patterns (local) in a set of pre-selected prototypical ducts in classifying atypical breast lesions. We conducted experiments to test this strategy on subgroups of breast lesions that are a major source of inter-observer variability; these are benign, columnar cell changes, epithelial atypia, and atypical ductal hyperplasia in order of increasing cancer risk. Our model is capable of providing clinically relevant explanations to its recommendations, thus it is intrinsically explainable, which is a major contribution of this work. Our experiments also show state-of-the-art performance in recall compared to the latest deep-learning based graph neural networks (GNNs).

Keywords: Atypical breast lesions · Prototype-based recognition · Diagnostic explanations · Digital and computational pathology

1 Introduction

Breast cancer screening and early detection can help reduce the incidence and mortality rates [19]. Although effective, screening relies on accurate pathological diagnoses of breast biopsies for more than one million women per year in the

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-87237-3_14) contains supplementary material, which is available to authorized users.

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12908, pp. 143–152, 2021.
https://doi.org/10.1007/978-3-030-87237-3_14

US [4,18]. Most benign and malignant biopsy diagnoses are straightforward, but a subset are a significant source of disagreement between pathologists and are particularly troublesome for clinicians. Pathologists are expected to triage their patients' biopsies rapidly and accurately, and they have routines for difficult or ambiguous cases (e.g., second-opinion consults, additional stains). Still, disagreement remains an issue; while the literature suggests that diagnosis should be straightforward if diagnostic rules are followed [17], concordance remains elusive in real world diagnosis, reported in one study as low as 48% [4].

Our Approach: In this study, we focus on modeling and differentiating difficult breast lesion subtypes: atypical ductal hyperplasia (ADH), flat epithelial atypia (FEA), columnar cell changes (CCC), and Normal (including usual ductal hyperplasia (UDH) and very simple non-columnar ducts). Our approach originates from the method that pathologists practice, which is to carefully assess alterations in breast ducts before making diagnostic decisions [8,10,19]. Pathologists continually observe tissue patterns and make decisions supported by the morphology. In doing so, they look at an entire duct (*global*) and patterns within portions of the duct (*local*) striving to generate mental associations with prototypical ducts and/or their parts they previously encountered in training or clinical practice. We propose an end-to-end computational pathology model that can imitate this diagnostic process and provide explanations for inferred labels.

We hypothesize that ductal regions-of-interest (ROIs) having similar global and local features will have similar diagnostic labels and some features are more important than others when making diagnostic decisions. Our approach is related to other prototypes-driven image recognition systems that favor visual interpretability [3,6,16].

Contributions: To the best of our knowledge, our work is the first one to: (1) use a diverse set of concordant *prototype* images (diagnostic class agreed by all 3 pathologists) for learning, (2) characterize clinically relevant global and local properties in breast histopathology images, and (3) provide explanations by measuring the relative importance of prototype features, global and local, for the differential diagnosis of breast lesions. We also show that our approach facilitates diagnostic explanations with accuracies comparable to the state-of-the-art methods.

2 Related Work

Although there have been numerous efforts in using prototypes for scene recognition [3,6,16], to date, this idea has not been explored to classify breast lesions. One of the first studies to detect high-risk breast lesions was proposed in [20] which was based on encoding cytological and architectural properties of cells within the ducts. The work in [12] used structural alterations of the ducts as features to classify breast lesions into benign, atypia, ductal carcinoma in-situ (DCIS), and invasive. A different approach was proposed in [13], where the authors used analytical models to find clusters within ROIs with strong histologically relevant structures. However, their approach lacked a good learning strategy to infer the diagnostic label from these clusters. Further, two recent

studies approached this problem using attention-based networks to generate global representation of breast biopsy images [11] and biological entity-based graph neural networks (GNNs) [14] (also tested as a baseline method in Table 2). Both methods were tested on an unbalanced dataset like ours and both reported low performance measures in detecting high-risk lesions.

3 Methodology

3.1 Machine Learning Framework

In this paper, we propose an end-to-end computational pathology system that models the entire duct (global) and the patterns occurring within selective portions of the duct (local) with the goal of generating associations with similar ducts and/or parts (prototypical). *We hypothesize that images with one or more ducts having similar global and local features will have similar diagnostic labels and some features are more important than others when making diagnostic decisions.* We will first introduce a composite mapping function to learn the relative importance of global and local features in a prototype set $\mathcal{P}$ for differential diagnoses:

$$h(x;\mathcal{P}) = \sum_{k=1}^{p} \beta_k \left[\exp^{-\lambda_k^G c_k(x)} \times \prod_{j=1}^{m_k} \exp^{-\lambda_{kj}^L f_{kj}(x)} \right]. \tag{1}$$

Here $h(x;\mathcal{P})$ captures the association of a previously unseen image x with a set of prototype images in $\mathcal{P}$. The index k varies over the images in the prototype set $\mathcal{P}$ (size $= p$), while j indexes over a local feature set (size $= m_k$) in a given prototype image indexed by k. β_k determines if the resemblance of a previously unseen image x to the prototype k has a positive (β_+) or negative influence (β_-). λ_k^G and λ_{kj}^L indicate the relative importance of global (ductal) and local (intra-ductal) features in the prototype k respectively. The relative importance can be imagined as a distance measure, so we enforce non-negativity constraints on λ_k^G and λ_{kj}^L values. The functions $c_k(x)$ and $f_{kj}(x)$ compute the global and local differences respectively between x and the prototype set $\mathcal{P}$ (more details below). Finally, in formulating $h(x;\mathcal{P})$ we assume that the prototype images are independent and that the global and local information in each prototype can be functionally disentangled into a product form.

Since our goal is to learn the relative importance of global and local features in a prototype set, we solve the following optimization problem:

$$\arg\min_{\beta,\lambda} \mathcal{L}(\beta,\lambda) = \arg\min \sum_{i=1}^{n} \text{CrsEnt}(\sigma(h(x_i)), y_i) + C_\beta ||\beta||^2 + C_\lambda |\lambda| \tag{2}$$

using gradient descent. We use cross-entropy loss function (CrsEnt) to penalize misclassifications on the training set $\mathcal{X} = \{x_i\}$ and to obtain $\beta_{\text{optimal}} = \{\beta_k\}$ and $\lambda_{\text{optimal}} = \{\lambda_k^G, \lambda_{kj}^L\}$. We use a $tanh(\sigma)$ activation function on $h(x)$ from Eq. 1. To avoid overfitting, we invoke ℓ_2^2 and ℓ_1 regularization with coefficients C_β and C_λ respectively. Following the intuition that a pathologist might pay no attention to some features, e.g., small-round nuclei do not feature typically in the diagnosis of ADH, we choose ℓ_1 regularization for λ to sparsify the weights.

3.2 Encoding Global and Local Descriptions of a Duct

The functions $c_k(x)$ and $f_{kj}(x)$ in Eq. 1 compute the global and local differences between input image x and prototype set $\mathcal{P}$, as outlined in the steps below.

Step 1: For a proof-of-concept, we adopt the approach from [13] to build analytical models of 16 diagnostically relevant histological patterns following the guidelines presented in the WHO classification of tumors of the breast [8].

Analytical model of a cribriform pattern: Fig. 1 illustrates how to model a histological pattern, *cribriform*, that is critical to diagnosing ADH. By considering a spatial neighborhood of 100 µm around each cell (Fig. 1A) in ground-truth annotations of cribriform patterns in ROIs, the model incorporates three different components (Fig. 1B): (1) polarization of epithelial cells around lumen inside the ROI; (2) distance of any given nucleus in the ROI to two nearest lumen; and (3) circularity of lumen structure adjacent to a nucleus inside the ROI. For the ROI in Fig. 1A, the analytical models driving these three components are: (1) mixture of Gaussians (MoG) ($\mu_1 = 0.87, \mu_2 = 0.94, \mu_3 = 0.72$, $\sigma_1 = 0.002, \sigma_2 = 0.002, \sigma_3 = 0.003$, $\pi_1 = 0.44, \pi_2 = 0.35, \pi_3 = 0.21$) for modeling the distribution of clustering coefficients [21]; (2) Gamma distribution ($\alpha = 3.11$, $\beta = 34.37$) for modeling distance values to lumen and (3) a uniform distribution ($a = 0.2$, $b = 0.92$) to model the circularity values of nuclei inside the ROI. We further combine these three components with a mixture model, performing grid-search to optimize the mixing coefficients (Fig. 1B), to form the histological pattern of cribriform (P_{gt}^{crib}).

We pursue a similar approach to modeling other histological patterns using ground-truth ROI annotations: 1. *small*, 2. *large*, 3. *round*, 4. *crowded*, and 5. *spaced*, each modeled as a Gamma distribution; 6. *elliptical*, 7. *large-round*, 8. *small-elliptical*, 9. *spaced-large*, 10. *crowded-small*, 11. *spaced-small*, 12. *crowded-elliptical*, and 13. *spaced-round* each modeled as two-component MoG; and more complex patterns 14. *large-round-spaced*, 15. *picket-fence*, and 16. *cribriform* using a combination of Gamma, MoG, and Uniform distributions. Details on parameter estimation are discussed in [13].

Generating Likelihood Scores: Next, to compare ground-truth model of any histological pattern P_{gt} with a new model generated from the reference nucleus of an input image (P_{new}), we use two distance measures, 2-sample Kolmogorov-Smirnov

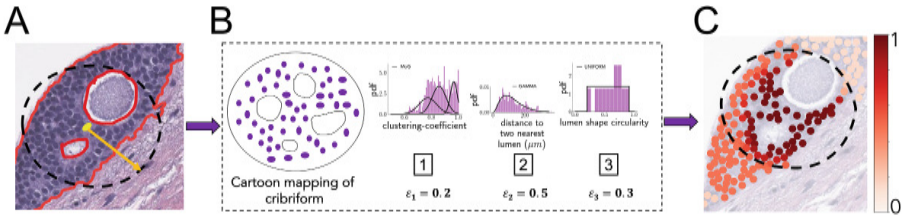

Fig. 1. Modeling cribriform pattern in a sample ROI (A) using parametric models for three component patterns in (B) and generating cell-level likelihood scores (C). Ductal region and intra-ductal lumen are outlined in red in (A). (Color figure online)

test and Kullback-Leibler divergence to compare Gamma and MoG distributions respectively. To map smaller distances that indicate stronger presence of the feature, we compute likelihood scores by applying an inverted S-function on the distances. In Fig. 1C the final likelihood score from evaluating the cribriform pattern is a weighted sum of the likelihood scores of the component patterns. A similar operation is carried out for generating cell-level likelihood scores for the remaining 15 features. The principal advantage of these analytical models is in their ability to handle heterogeneity that emerges from running imprecise low-level image processing routines, such as methods for segmenting nuclei or identifying boundaries of ductal ROIs. The heatmap visualization in Fig. 1C is a mechanism for explaining the model to pathologists, informing where these patterns are and how strongly they influence the overall diagnosis of a ROI.

Step 2: To encode the global description of a duct, we will represent it by a matrix of size $n \times l$ populated with likelihood scores, where n and l refer to the total number of cells and the number of histomorphological patterns respectively ($l = 16$). Additionally, we include the size of the largest duct if the ROI has a cluster of ducts. However, considering only the global information may lead to diagnostic inconsistencies. For example, a duct resembling FEA is better diagnosed as ADH if it contains a local cribriform pattern or as a CCC duct if it contains some hyperplasia (further meriting a comparison of local hyperplastic area with models of FEA/ADH).

Step 3: To encode the local description of a duct, we adopt a strategy followed by most expert pathologists. To this extent, for every histomorphological pattern, we identify islands within the duct where that particular feature is dominant and consider the largest island for further analysis. We detect feature islands by performing non-maxima suppression on cell-level likelihood scores using a threshold (=0.8) based on cross-validation.

Step 4: Finally, we have the machinery to compute the functions $c_k(x)$ and $f_{kj}(x)$ from Eq. 1. We define $c_k(x) = \|d(p_k, x)\|$, where a small value of $c_k(x)$ implies high similarity of image x to prototype p_k. We combine two measures to generate d: Kolmogorov-Smirnov test comparing 16-dim probability distributions of cell-level likelihood scores individually between x and p_k and an inverted S-function on the ratio of the duct sizes between x and p_k. This leads to a 17-dim vector d, which is further compressed by its ℓ_2 norm to obtain a single scalar value $c_k(x)$ for every pair of x and p_k. We further simplify the computation of $f_{kj}(x)$ by applying an inverted S-function on the ratio of the largest feature island sizes from the same histological feature between x and p_k, suitably modified to account for islands that are missing in either x or p_k.

4 Results and Discussion

4.1 Dataset

We collected a cohort of 93 WSIs which were labeled by an expert pathologist on the team to contain at least one ADH ROI. The breast biopsy slides were scanned

Table 1. Statistics of the atypical breast lesion ROI dataset

Prototype Set	PS-1	PS-2	PS-3
No. of ROIs	20	20	30
No. of feature islands	84	86	145

Class	NORMAL	CCC	FEA	ADH	Total
Train	420	99	116	119	754
Test	371	105	33	32	541

at 0.5 μm/pixel resolution at 20× magnification using the Aperio ScanScope XT (Leica Biosystems) microscope from which 1295 ductal ROI images of size $\approx 1K \times 1K$ pixels were extracted using a duct segmentation algorithm described in [13]. Briefly, the algorithm first breaks down the image into non-overlapping superpixels and then evaluates each superpixel's stain level together with its neighboring superpixels and assigns probabilities of them belonging to a duct. These guesses are then used to perform Chan-Vese region-based active contour segmentation algorithm [2] that separates the foreground (i.e., ducts) from the background.

We collected ground truth annotations of extracted ROIs from 3 breast pathology sub-specialists (P1, P2, and P3), who labeled the ROIs with one of the four diagnostic categories: Normal, CCC, FEA, and ADH. The diagnostic concordance for the four categories among P1, P2, and P3 were moderate with a Fleiss' kappa score of ≈0.55 [20]. The entire dataset was split into two sets.

i. Prototype set: We formed three prototype sets (PS-1, PS-2, and PS-3) containing ROIs with consensus diagnostic labels from the 3 pathologists having a balanced distribution over the four diagnostic categories. The final set of prototype ROIs were verified by P1 to confirm adequate variability is obtained. The number of aforementioned *islands* are also listed in Table 1.

ii. Train and test set: The training set consists of 754 ROIs labeled by P1 and the test set contains 541 ROIs consensus labeled by P1-P3. The training and test set were separated at WSI level to avoid over-fitting, since ROIs belonging to the same WSI can be correlated histologically. Due to limited number of ROIs belonging to the non-Normal category as seen in Table 1, the ROIs which do not participate in the prototype set were also included in the dataset.

4.2 Model Training and Evaluation

Our ML model (Eq. 1) is trained to minimize the objective function (Eq. 2) using gradient descent (learning rate $= 1 \times 10^{-4}$ and convergence tolerance $= 1 \times 10^{-3}$). Regularization coefficients C_β and C_λ were initialized to 2. To speed up convergence, we shuffle the training data after each iteration so that successive training examples rarely belong to the same class. Prior to training, the model parameters β and λ were initialized with weights randomly drawn from *LeCun normal* [9]. After each iteration, the parametric values of the objective function ($\mathcal{L}$), error-rate (ϵ), β, and λ are stored. After model convergence, we use β_{optimal} and λ_{optimal} parameters in the mapping function (1) to obtain h_{test}. We generate prediction probabilities p by first applying a *tanh* (σ) activation to h_{test} and then projecting it to the positive octant. If $p \geq 0.5$, the diagnostic label is 1 and 0 otherwise.

4.3 Baseline Models (B1-B3)

Following the method laid out in [14], we define two baseline models, B1 and B2, by re-implementing their cell-graph GNNs. We chose GNNs, a recently emerged state-of-the-art technique for encoding spatial organizations, over pixel-based convolutional neural networks (CNNs) as our experiments with CNNs showed poor performances in capturing the spatial context [13]. B1 is obtained by generating a cell-graph topology and cells within each graph are embedded with cytological features as in [14]. To assess the effect of histological patterns in cell embeddings, we generate B2 by replacing the duct-level cytological features with likelihood scores generated by our method. Finally, B3 is obtained by implementing a Logistic Regression classifier using the duct-level likelihood scores, following a similar strategy as in [13].

Table 2. Diagnostic results from the binary classification task expressed in %

		Baseline			PS-1			PS-2			PS-3		
	Model	B1	B2	B3	G1	L1	GL1	G2	L2	GL2	G3	L3	GL3
HR	**R**	56±6	68±6	62±3	66±4	71±1	73±4	68±4	72±2	68±3	66±7	74±2	69±3
	wF	77±2	82±3	76±1	65±2	61±1	65±1	67±4	61±1	63±2	63±1	64±1	64±1
ADH	**R**	38±8	45±7	56±3	70±7	61±8	78±8	59±13	80±4	71±4	72±6	70±11	68±5
	wF	78±4	86±2	79±1	70±3	64±2	67±5	64±3	62±1	60±6	64±2	67±1	64±1
FEA	**R**	48±12	40±6	35±4	54±6	64±5	68±7	58±6	60±3	63±5	63±6	67±2	62±5
	wF	81±5	82±3	78±1	71±2	65±2	69±3	66±4	66±3	69±3	66±2	69±2	66±3

4.4 Classification Results

For the sake of differential diagnosis of atypical breast lesions, we implemented several models using global (G), local (L), and both global and local information (GL) from three prototype sets (PS1-PS3) and compared it with the baseline models (B1-B3) (see Table 2). During the training step of each model, we created a balanced training set by randomly subsampling ROIs from each category so that we have equal number of ROIs for each classification category. To check for statistical significance, for each classification task, we run our ML algorithm on 10 training sets wherein the images are randomly selected and we report the classification scores as the mean and standard deviation over 10 runs (Table 2). The top panel of Table 2 (HR row) compares the classification performance of low-risk (Normal+CCC, −ve class) vs high-risk (FEA+ADH, +ve class) cases. For each diagnostic category (+ve class), we further implemented a different binary classifier for each modeling strategy proposed. The bottom panel of Table 2 (ADH and FEA row) shows the comparative performances of ADH- and FEA-vs-rest diagnostic classification. We highlight results from high-risk category because ADH lesion presents both - a risk of currently existing cancer (about 4%) and there is a high absolute future risk of about 1% per year, up to 30% lifetime risk [5]. FEA lesion combines the nuclear atypia seen in ADH, but lacks hyperplasia

and has simpler architecture [7]. Some patients with FEA will be offered surgery and they would also be treated as high-risk in the future.

Performance Metrics: For each classification scenario, we use *recall* (R) as the performance metric to focus on the correct detection of positive class, since there is a significant class imbalance (see Table 1) and the consequence of misdiagnosis (false negative) implies increased chance of developing cancer with lack of providing early treatment. We include *weighted F-measure* (wF) as an additional metric which gives importance to the correct detection of both positive and negative classes [15]. The class specific weights in wF are proportional to the number of positive and negative examples present in the test set.

Classification Performance: We highlight the best *recall* performances in Table 2, that are achieved using state-of-the-art baseline models against our method in black and gray boxes, respectively. Our method shows significant improvement ($p < 0.01$) in detecting diagnostically critical high-risk ADH and FEA ROIs compared to the baseline methods (the best average recall achieved is 80% for ADH classifier and 68% for FEA). We also observe that baseline models are performing better on detecting Normal ROIs (see Supp. Table 1 for comparative results of CCC-, and Normal-vs-rest classification). This behaviour explains higher weighted F-measure of baseline models in low- vs. high-risk classification, since in the testing set low-risk ROIs are 7-fold more than high-risk ROIs (i.e., baseline models are biased to detect low-risk lesions even when the training set was balanced). It is critical to note that real-life clinical observance of high-risk lesions is also around 15% [8], which is naturally reflected in our testing set, and it is crucial to catch these less-seen high-risk lesions for pre-cancer interventions while being able to provide diagnostic explanations to given recommendations.

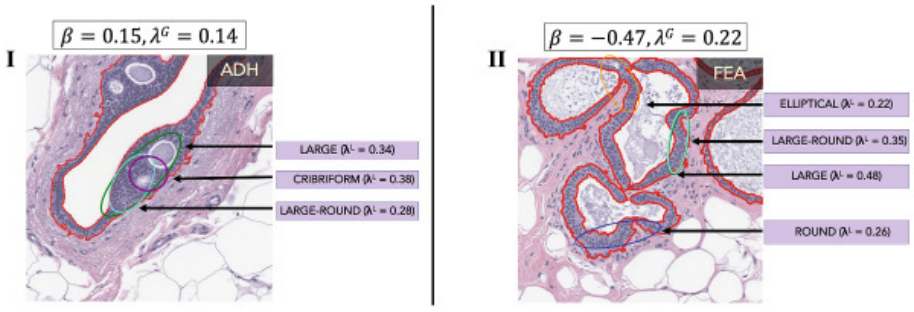

Fig. 2. Highlighting the relative importance of the global and local features from different prototypes (I and II) in ADH-vs-rest classifier.

4.5 Discussion

The explainability of our model is depicted in Fig. 2, which shows that our model leverages both global (λ_G) and local (λ_L) information of the ductal ROIs of two prototypical images, I and II, in detecting ADH from one of the experiments using GL3 classifier built using prototype set PS3. The values of model parameters: absolute change in the objective function ($\Delta\mathcal{L}$), training error-rate (ϵ),

β, and λ after each iteration are shown in Supp. Figure 1 and more examples of explainability are shown in Supp. Figure 2. Figure 2-I positively guides in detecting ADH category ($\beta = 0.15$) whereas Fig. 2-II is counterintuitive in detecting ADH lesions ($\beta = -0.47$). Although two of the histological feature islands, large and large-round present within these ROIs overlap, we assert that the absence of complex architectural pattern such as cribriform within Fig. 2-II might have led to a negative influence of this prototype's influence to detect ADH. Although it is possible that an FEA type lesion could be upgraded to ADH pathologically without cribriform architecture, this would require thickening of the duct lining to more than 5 cell layers which is uncommon in clinical practice.

Computational Cost: The entire pipeline is implemented in native Python 3.8. Total time required to obtain a diagnostic label with computation of all features for a previously unseen ROI is less than $30s$ on a 64-bit single $3.4GHz$ Intel Xeon processor.

Limitations: (1) Features like *bulbous micropapillae* and *rigid cellular bars* which are diagnostically relevant to high-risk lesions are missing; (2) Selection of prototypes was made on the basis of expert visual inspection. There is a need for more sophisticated statistical approaches [1] for prototype selection and (3) for a more detailed ablation study to test the robustness and reliability of our ML framework; (4) To offset the issue of unbalanced datasets, we are collecting expert annotations on additional high-risk lesion images.

Future Work: Our intent is to create an approach that generalizes, not only to other, more straightforward breast diagnoses but also to tissue histologies from other organs. Explainable machine learning approaches like ours will support pathologists during their transition to digital and computational pathology.

Acknowledgments. The grant NIH-NCI U01CA204826 to SCC supported this work. The work of AP and OC was partially supported by the sub-contracts 9F-60178 and 9F-60287 from Argonne National Laboratory (ANL) to the University of Pittsburgh from the parent grant DE-AC02-06CH1135 titled, Co-Design of Advanced Artificial Intelligence Systems for Predicting Behavior of Complex Systems Using Multimodal Datasets, from the Department of Energy to ANL.

References

1. Bien, J., Tibshirani, R.: Prototype selection for interpretable classification. Ann. Appl. Statist. **5**, 2403–2424 (2011)
2. Chan, T.F., et al.: Active contours without edges. IEEE Trans. Image Process. **10**(2), 266–277 (2001)
3. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. In: Advances in Neural Information Processing Systems, pp. 8930–8941 (2019)
4. Elmore, J.G., et al.: Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA **313**(11), 1122–1132 (2015)

5. Hartmann, L.C., Degnim, A.C., Santen, R.J., Dupont, W.D., Ghosh, K.: Atypical hyperplasia of the breast–risk assessment and management options. New England J. Med. **372**(1), 78–89 (2015)
6. Hase, P., Chen, C., Li, O., Rudin, C.: Interpretable image recognition with hierarchical prototypes. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 7, pp. 32–40 (2019)
7. Hugar, S.B., Bhargava, R., Dabbs, D.J., Davis, K.M., Zuley, M., Clark, B.Z.: Isolated flat epithelial atypia on core biopsy specimens is associated with a low risk of upgrade at excision. Am. J. Clin. Pathol. **151**(5), 511–515 (2019)
8. Lakhani, S.R.: WHO Classification of Tumours of the Breast. International Agency for Research on Cancer (2012)
9. LeCun, Y.A., Bottou, L., Orr, G.B., Müller, K.-R.: Efficient BackProp. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) Neural Networks: Tricks of the Trade. LNCS, vol. 7700, pp. 9–48. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35289-8_3
10. Li, B., et al.: Classifying breast histopathology images with a ductal instance-oriented pipeline
11. Mehta, S., Lu, X., Weaver, D., Elmore, J.G., Hajishirzi, H., Shapiro, L.: Hatnet: an end-to-end holistic attention network for diagnosis of breast biopsy images. arXiv preprint arXiv:2007.13007 (2020)
12. Mercan, E., Mehta, S., Bartlett, J., Shapiro, L.G., Weaver, D.L., Elmore, J.G.: Assessment of machine learning of breast pathology structures for automated differentiation of breast cancer and high-risk proliferative lesions. JAMA Netw. Open **2**(8), e198777–e198777 (2019)
13. Parvatikar, A., et al.: Modeling histological patterns for differential diagnosis of atypical breast lesions. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12265, pp. 550–560. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_53
14. Pati, P., et al.: HACT-Net: a hierarchical cell-to-tissue graph neural network for histopathological image classification. In: Sudre, C.H., et al. (eds.) UNSURE/GRAIL -2020. LNCS, vol. 12443, pp. 208–219. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60365-6_20
15. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
16. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 413–420. IEEE (2009)
17. Schnitt, S.J., Connolly, J.L.: Processing and evaluation of breast excision specimens: a clinically oriented approach. Am. J. Clin. Pathol. **98**(1), 125–137 (1992)
18. Silverstein, M.: Where's the outrage? J. Am. College Surgeons **208**(1), 78–79 (2009)
19. American Cancer Society: Breast cancer facts & figures 2019–2020. Am. Cancer Soc. 1–44 (2019)
20. Tosun, A.B., et al.: Histological detection of high-risk benign breast lesions from whole slide images. In: Descoteaux, M., et al. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 144–152. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_17
21. Zhou, N., Fedorov, A., Fennessy, F., Kikinis, R., Gao, Y.: Large scale digital prostate pathology image analysis combining feature extraction and deep neural network. arXiv preprint arXiv:1705.02678 (2017)