Classifying Breast Histopathology Images with a Ductal Instance-Oriented Pipeline

Beibin Li*‡, Ezgi Mercan‡, Sachin Mehta*,
Stevan Knezevich‖, Corey W. Arnold†, Donald L. Weaver§, Joann G. Elmore†, Linda G. Shapiro*

*University of Washington, Seattle, WA †University of California, Los Angles, CA
‡Seattle Children's Hospital, Seattle, WA §University of Vermont, Burlington, VT
‖Pathology Associates, Clovis, CA

***Abstract*—In this study, we propose the Ductal Instance-Oriented Pipeline (DIOP) that contains a duct-level instance segmentation model, a tissue-level semantic segmentation model, and three-levels of features for diagnostic classification. Based on recent advancements in instance segmentation and the Mask R-CNN model, our duct-level segmenter tries to identify each ductal individual inside a microscopic image; then, it extracts tissue-level information from the identified ductal instances. Leveraging three levels of information obtained from these ductal instances and also the histopathology image, the proposed DIOP outperforms previous approaches (both feature-based and CNN-based) in all diagnostic tasks; for the four-way classification task, the DIOP achieves comparable performance to general pathologists in this unique dataset. The proposed DIOP only takes a few seconds to run in the inference time, which could be used interactively on most modern computers. More clinical explorations are needed to study the robustness and generalizability of this system in the future.**

***Index Terms*—biomedical imaging, deep learning, cancer diagnosis, biopsy, histopathology, machine learning, whole slide images**

I. Introduction

Breast cancer is one of the most common cancers for females: about 13% of women will develop breast cancer over their lifetimes, and 2.6% of women will die from breast cancer in the United States [1]. The recent development of Artificial Intelligence (AI) screening tools for mammography [2] could reduce the second reader's workload, but diagnosing breast cancer is still a time-consuming and challenging task. Physicians usually recommend breast biopsies for diagnosis and for the development of treatment plans after finding suspicious areas in a mammogram, ultrasound, or magnetic resonance imaging (MRI).

When analyzing breast biopsies, pathologists usually focus on ducts, because most breast cancers begin in the terminal ducts or lobules of the breast [1]. In ductal carcinoma in situ (DCIS), cells inside ducts undergo malignant transformation to cancer cells; in invasive breast cancer, these abnormal cells have escaped from the duct and are growing in the surrounding tissue [1]. Hence, tissues surrounding duct borders are the most relevant regions for pathologists and also machine learning models to focus on.

Based on the importance of ductal regions, Mercan et al. designed structure features [3] to summarize the architectural characteristics in duct-based structures. Their method can emulate pathologists' behaviors to interpret diagnostic decisions and has been shown to outperform pathologists on the difficult task of categorically differentiating DCIS from Atypia. They identified duct instances (i.e. all pixels that are part of an individual breast duct or lobule) by applying a union-find algorithm to split semantic segmentation predictions of specific tissue classes into smaller duct regions. However, many ducts are entangled in breast biopsies, and this approach cannot distinguish a duct instance that is adjacent to other ducts, as shown in Figure 1. Moreover, extracting their structure features from biopsies can take hours on a computer, because this algorithm is not suitable for parallel processing on multiple computer cores. The accuracy and computational requirements need improvement if the goal is to create an interactive and real-time tool for clinical use.

On the other hand, many research groups have designed end-to-end deep learning systems to classify breast histopathology images, including fully-convolutional networks (FCNs) [4], patch- to ROI-level feature representation [5], and graph convolutional networks [6]. Mehta et al. also designed Y-Net [7], which can perform image segmentation and diagnostic classification at the same time. These systems are usually accurate and fast because of the recent advancement of Graphical Processing Units (GPUs) and parallelism for matrix manipulation. Nevertheless, these methods are somewhat blind to the underlying pathological and structural features that led to the clinical diagnosis. They only provide heat maps, attention maps, or patches for visualization; they do not guarantee a focus on ductal regions and cannot offer a decisive interpretation as [3] does for pathologists. Knowing why an AI algorithm makes a certain diagnostic decision is crucial in clinical practice.

Researchers often have to balance the trade-offs between speed, accuracy, and interpretability within these computer-aided diagnosis (CAD) tools. In this study, we created a Ductal Instance-Oriented Pipeline (DIOP) that can identify individual duct structures in breast biopsies, extract features, and classify breast cancer diagnoses. The proposed pipeline improves upon

Research reported in this article was supported by grants R01 CA172343, R01 CA140560, U01CA231782, and R01 CA200690 from the National Cancer Institute of the National Institutes of Health.

previous approaches on all diagnosis tasks, outperforms human general pathologists in two out of three binary classification tasks, and achieves comparable performance to general pathologists in a four-way diagnostic classification task, distinguishing among Benign, Atypia, Ductal Carcinoma in Situ, and Invasive Cancer category examples.

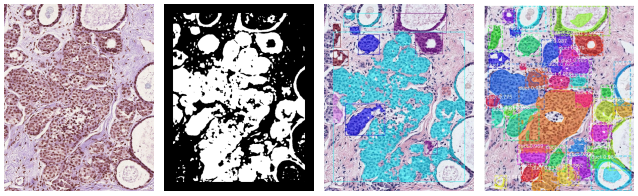
II. Background

Recent developments in breast cancer assessment, semantic segmentation, instance segmentation, and weakly supervised learning for medical imaging provide the groundwork for our study. Semantic segmentation is a common task in medical imaging; it partitions an image into multiple tissues, grades, or classes by classifying each pixel inside the image. For example, LSBB [8], Y-Net [7], and ESPNet [9] were designed for breast biopsy semantic segmentation; multi-scale U-Net [10], vanilla FCNs [11], EM-based models [12], and attention models [13] were created for prostate cancer; specialized auto-encoder [14], FCNs [15], and U-Net [16] were used for melanoma segmentation tasks. Unfortunately, these methods could not identify duct instances inside a region of interest (ROI), because semantic segmentation could not differentiate these instances from the semantic labels. Instead, instance segmentation labels, which contain all pixels that are part of each breast duct or lobule, are needed.

While semantic segmentation has been widely applied to cancer diagnosis, instance segmentation is rarely used. Li et al. designed Path R-CNN (regions with convolutional neural network features) [17] based on Mask R-CNN [18] to classify glands and grade prostate cancer, which pioneered instance segmentation for medical imaging. In the recent two years, instance segmentation for nuclei [19], cluttered cells [20], polyps [21], and other tissues have been developed. In breast biopsies, each individual duct might contain different structural information about tissues, which could not be provided by tissue-level semantic segmentation. Hence, performing instance segmentation for ducts could be valuable for classifying histopathology images.

Acquiring instance segmentation labels for ducts is a tedious and time-consuming process, and no public datasets are available for instance segmentation on ducts to the best of our knowledge. For images with simple structures, semantic segmentation labels can be easily converted to instance segmentation labels by using union-find, connected components, or other rule-based algorithms. Unfortunately, ducts inside of breast biopsies are too complex for these rule-based label conversion algorithms, and instance annotations are needed in order to train an instance segmenter.

Geometrically, the shape of a duct is similar to the shape of a "doughnut": it is usually ring-shaped with a circular thick border and central empty space, but often it does not have any holes because microscopic histopathology images are two-dimensional cross sections through three-dimensional structures. Breast ducts are analogous to pipes or tubes, and breast lobules are analogous to a hollow ball such as a tennis ball. Pre-cancerous conditions can also cause the lining of the ducts and lobules to proliferate and fill the holes. All these features complicate rule-based algorithms. The difficulties compound when many ducts are adjacent to each other so that the borders of these ducts are difficult to distinguish. In DCIS and invasive cases, cancerous cells first begin to distort then escape from ducts, respectively, and thus duct cross sections develop into various and complex shapes.



(a) ROI image (b) Binary (c) Previous (d) Ours

Fig. 1: **Duct instances:** (a) the input ROI image in RGB color space; (b) the binary image inferred from tissue-level semantic segmentation, where the white pixels are ducts; (c) duct instances found by mathematical morphology and connected component algorithm; (d) the ducts inferred from our system. In (c) and (d), each color represents one duct instance. The connected component method (c) could not distinguish duct instance from the conglomerated region, even if it has been used to solve similar problems (e.g. in [3], [17], etc.).

Recent studies show that weak annotation [22], imperfect annotation [23], active learning, and human-in-the-loop methods [24] can be used to alleviate this problem. These methods encouraged us to design an efficient annotation plan to find ducts inside breast biopsies.

III. System for Breast Cancer Diagnosis

Several clinical studies have shown that stromal tissues and ducts are important biomarkers for diagnosing breast cancer [1], [25]. Motivated by these studies, we introduce a machine learning-based framework that accounts for these important bio-markers in cancer diagnosis. Our system consists of three components: (1) a duct-level instance segmentation model (Section III-A), (2) a tissue-level semantic segmentation model (Section III-B), and (3) a classifier with three-levels of extracted features (Section III-C). The input ROI is fed to both duct-level and tissue-level segmentation modules simultaneously to produce instances of ducts and tissue-level segmentation masks. Histogram frequency and co-occurrence features are extracted at three different levels, from the duct, from the bounding boxes, and from the entire region of interest to predict the diagnosis. Our experimental results show that the proposed framework outperforms the previous state-of-the-art methods and matches the performance of pathologists.

A. Duct-level instance segmentation

The instance segmentation network adopts the same structure as Mask R-CNN [18]. The network consists of two stages. The first stage takes an ROI as an input and produces duct candidates. In the second stage, these candidates are then classified as duct or not. In addition to this classification, the

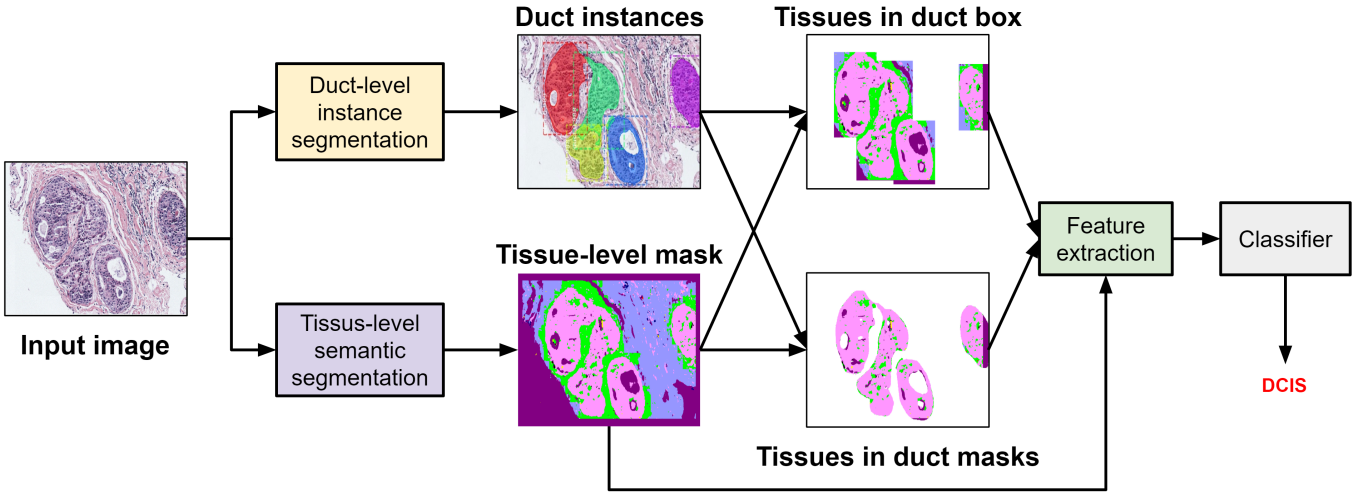


Fig. 2: Ductal Instance-oriented Pipeline: this pipeline leverages existing tissue-level semantic segmentation, Mask R-CNN for duct-level instance segmentation, extracting features, and a classifier for diagnosis. The feature extractors calculate the tissue histogram frequency and co-occurrence matrix for ROI-, box-, and mask-levels.

second stage also produces bounding box coordinates as well as a pixel-wise mask of the duct.

As in any other medical imaging task, collecting data is difficult because it requires experts to annotate data. To generate duct-level instance annotations, a weakly-supervised annotation framework is introduced that combines an annotator's bounding box annotations with tissue-level semantic labels to create duct-level instance masks (details in Section IV-A).

B. Semantic segmentation

An off-the-shelf segmentation network [8] is applied to generate tissue-level semantic segmentation. The off-the-shelf network splits the input image into non-overlapping patches and predicts a segmentation mask for each patch using a multi-resolution encoder-decoder structure.

C. Feature Extraction and Classification

Several methods (e.g., histogram features, co-occurrence features, and structural features) have been proposed to extract features from tissue-level segmentation masks. Deriving from superpixels (regions of similar color with an area), these features allow the encoding of information about tissues (e.g., stromal tissue) and structures (e.g., ducts) present in biopsy images and help improve the diagnostic classification compared to multi-instance learning-based methods [5]. Histogram frequency features can convey the distribution of tissues in an image, and the co-occurrence features can encode simple spatial relationships. The structure features, extracting frequencies from five layers inside and five layers outside of a duct instance, can capture the changes in the shape of the epithelial structures.

Though the structure features used in [3] would allow capturing architectural information around the ducts, they are computationally expensive as compared to the histogram and co-occurrence features. While computing structure features takes over an hour averagely for each image, computing the proposed features only takes about one second. The proposed DIOP leverages the duct-, box-, and tissue-level information based on tissue histogram frequency and co-occurrence frequency. This design allows us to replace these computationally expensive features with simple features at three levels. These features are then fed to a classification network (e.g., a random forest or a multi-layer perceptron) to predict the diagnostic category. The ability of our system to aggregate clinically relevant information at different levels allows our framework to outperform existing methods by a significant margin (Section IV-D).

IV. Experimental Results

A. Dataset

Diagnostic labels: The Ductal Instance-Oriented Pipeline was developed by using digital whole slide images created from residual breast biopsy material [26]–[28]. The dataset consists of a total of 428 ROI images, which were extracted from 240 breast biopsies and categorized by three expert pathologists, who selected the ROIs for each whole slide image and agreed on a consensus diagnosis for each slide. Similar to many previous studies on this dataset [3], [7], [29], we performed diagnosis classification for each of the 428 ROIs into 4 classes: Benign, Atypia, Ductal Carcinoma in Situ, or Invasive Cancer. The dataset is unique and is enriched with additional cases in the challenging Atypia and DCIS categories to assist in establishing statistical confidence in accuracy predictions for categorical classification. The enriched image dataset provides additional ROI input data on lower prevalence disease categories for developing the AI pipeline. The comparison of participants' diagnoses performance with expert pathologists was previously reported [30].

Duct-level instance annotations: Ductal regions are important bio-markers in diagnosing breast cancer. However, collecting duct-level instance segmentation masks is difficult because

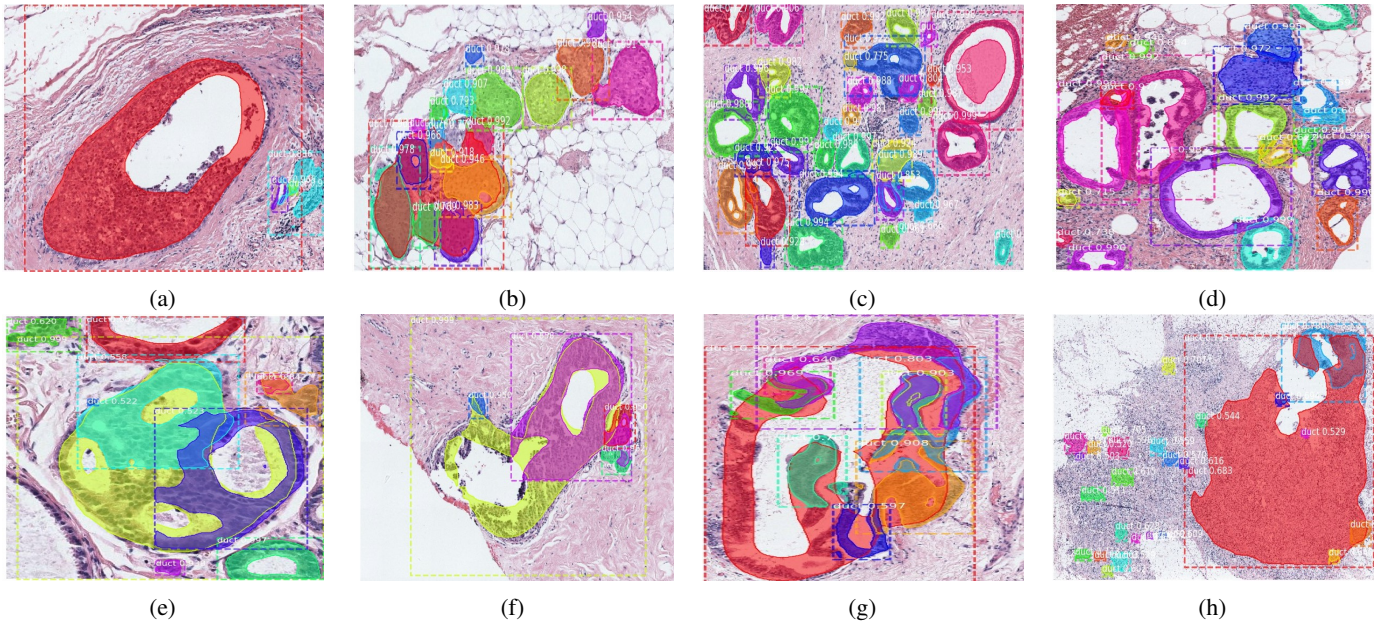


Fig. 3: **Ductal segmentation testing result:** each color represents one instance of a duct in the biopsy. The top row shows four examples with satisfiable duct identification results, and the bottom row shows four imperfect examples. Examples (e), (f), and (g) have taken a single irregular and expanded duct and split it into multiple duct structures. Cancerous cells have escaped from ducts in (h), and our system mistakenly marks a big region (in red) as one duct.

pathologists are required to annotate the instances. We created a weakly supervised annotation tool to collect duct-level instance segmentation masks. Our annotation tool is shown in Figure 4. We first applied off-the-shelf tissue-level semantic segmentation network to ROIs. Benign epithelium (BE), malignant epithelium (ME), secretion (SC), and necrosis (NC) are tissues that surround ducts. Therefore, we created binary masks by assigning pixels in these tissues as foreground with the remaining pixels as background, as shown in Figure 1b. These masks were then combined with the bounding box annotations[1] to produce duct-level instance segmentations. Overall, 4,347 duct instances were marked in 100 ROIs. Note that one bounding box might contain pixels from several ducts, and this annotation strategy can still mistakenly mark pixels from other ducts to the main duct in the bounding box. So, our annotations are only “silver standard" rather than “ground truth," because they are inexact.

Tissue-level semantic annotations: The dataset also provides pixel-wise tissue-level labels for 58 ROIs. An expert pathologist annotated these ROIs into eight tissue classes: background (BG), benign epithelium (BE), malignant epithelium (ME), normal stroma (NS), desmoplastic stroma (DS), secretion (SC), blood (BL), and necrosis (NC). These ROIs were used to train a tissue-level segmentation model. See [8] for more details.

B. Implementation details

Duct-level instance segmentation: We fine-tuned Mask R-CNN[2] [32] (pretrained on the MS-COCO dataset) with ResNet-50 as a backbone network for 30 epochs using SGD with an initial learning rate of 0.01 and a momentum of 0.9. We used duct-level instance segmentation masks produced using our weakly supervised annotation tool for fine-tuning Mask R-CNN. Compared to cellular entities, ductal regions are larger in size and can be easily detected with lower resolution images. Therefore, we resized all ROI mages to a fixed spatial dimension of 512×512. The dataset was split into an 80:20 ratio: 80 ROIs for training and 20 ROIs for validation. On the validation set, Mask R-CNN achieved a mean intersection over union (mIOU) of 72% and mean average precision (mAP) of 32%. Figure 3f visualizes duct-level instance segmentation masks produced by Mask R-CNN.

Tissue-level semantic segmentation: Ductal regions can be identified at lower image resolutions because the shape and texture can help recognition. However, tissue-level segmentation methods do not perform well at lower resolutions, because low-resolution images may lose information about cellular entities, which help differentiate between different tissues. Similar to [3], the off-the-shelf semantic segmentation method [8] was applied to produce tissue-level semantic masks at x40 magnification.

Diagnostic classification: we tried a random forest model, a 3-degree polynomial support vector machine (SVM), an SVM with radial basis function (RBF) kernel, and a multi-layer perceptron (MLP) with four hidden layers (256, 128, 64, and 32 neurons for each layer, similar to [7]) for comparison.

C. Classification methods

Diagnostic classification tasks and baseline methods are introduced below. We run each experiment 100 times and

[1]The bounding boxes around ducts were marked by an engineering student (Beibin Li) under the supervision of an expert pathologist (Stevan Knezevich).

[2]https://github.com/matterport/Mask_RCNN

Task	Features	Sensitivity	Specificity	Accuracy	F_1
Invasive vs Non-invasive	*Pathologists* [30]	*0.84*	*0.99*	*0.98*	*0.86*
	Superpixel Features [3]	**0.70**	0.95	0.94	0.62
	Structure Features [3]	0.49	0.96	0.91	0.51
	Duct-RCNN (Ours)	0.62	**0.98**	**0.95**	**0.73**
Atypia and DCIS vs Benign	*Pathologists*	*0.72*	*0.62*	*0.81*	*0.51*
	Superpixel Features	0.79	0.41	0.70	0.46
	Structure Features	**0.85**	0.45	0.70	0.50
	Duct-RCNN (Ours)	**0.85**	**0.63**	**0.79**	**0.59**
DCIS vs Atypia	*Pathologists*	*0.70*	*0.82*	*0.80*	*0.76*
	Superpixel Features	0.88	0.78	0.83	0.86
	Structure Features	0.89	0.80	0.85	0.87
	Duct-RCNN (Ours)	**0.91**	**0.89**	**0.90**	**0.92**

TABLE I: **Diagnosis results for binary classification:** we show sensitivity, specificity, accuracy, and F_1 score for all models. We highlight the best machine performances in this table, and pathologists' performances are provided for comparison. The results for superpixel features and structure features are from [3], [31], where the standard deviations (STD) are not reported.

report the mean performance.

Binary classification: Emulating the successive decisions made by pathologists, [3] performed three binary classification tasks (i.e. invasive v.s. non-invasive, atypia & DCIS v.s. benign, and DCIS v.s. atypia) in their studies. We performed similar experiments, using leave-one-out cross-validation to evaluate each binary classification model. If the number of features was more than the number of ROIs in a classification task, we performed principal components analysis (PCA) to reduce the number of features to 20 dimensions. We applied a weighted random sampling approach to sample balanced positive and negative samples before training these binary classifiers.

Multi-class classification: We also conducted a 4-way classification experiment to compare our results with previous studies [5], [7]. We used the same training/validation/testing split as Y-Net [7], an extension of U-net with a separate branch for diagnostic classification; their discriminative masks improved classification accuracy by 7% over previous feature-engineering methods. On the other hand, a multi-instance learning based method [5], analyzing extracted features from a CNN instead of tissue-level semantic information, outperformed previous methods. We will compare the proposed DIOP with these baseline methods.

D. Main Results

Binary classification: Table I compares the performance of our method with the superpixel and structural features [3] in terms of sensitivity, specificity, accuracy, and F_1 score. Overall, our method outperforms both methods in all binary classification tasks. We observe that the super-pixel-feature-based method delivers the best performance for the invasive vs. non-invasive task. This is because cancer cells spread out from the ducts in invasive cancer (as shown in Figure 3h). This limits both our method and the structure-feature-based method to aggregate information around ducts, resulting in lower performance. In contrast, the super-pixel-feature-based method only accounts for pixel-level information and not structure-level information. Therefore, such methods are resilient to structural changes.

Method	Accuracy
Pathologists [30]	0.70
MIL with max-pooling [33]	0.55
MIL with learned fusion [5]	0.67
Semantic Learning [8]	0.55
Y-Net [7]	0.63
DIOP (Ours)	**0.70** $\pm$ 0.02

TABLE II: Multi-class classification results on the breast biopsy dataset. Our model outperforms existing methods by a significant margin and also, matches the performance of pathologists.

Multi-class classification: Table II compares 4-way classification performance of our method with state-of-the-art methods. Compared to these methods, our method delivers significantly better performance. For example, our method is about 7% and 3% more accurate than Y-Net and the multiple instance learning (MIL)-based method. Importantly, our method matches the performance of pathologists on this dataset.

E. Ablations

To understand the components of our system in detail, we perform the following experiments:

Impact of duct-level instance segmentation and tissue-level semantic segmentation: Following [31], we extracted L*a*b, hematoxylin and eosin (H&E), and local binary pattern (LBP) histogram features for the duct-only method. For the other two methods (tissue-only and tissue+duct), we extracted histogram and co-occurrence tissue-level features (similar to [7], [8]). Table III shows that the method that uses both duct- and tissue-level information delivers the best performance.

Impact of different features: To aggregate the information about different structures present in the breast biopsy images, previous methods have proposed different features, namely structure features, histogram features, and co-occurrence features. Table IV compares the performance of our method with these different features. Using both histogram and co-occurrence features delivers the best performance. We note that

our method delivers a good performance with co-occurrence features alone. This is because the co-occurrence matrix encodes strong relationships between different tissues.

Impact of extracting features from different levels: Our framework in Section III encodes information from three different levels: (1) tissue-level segmentation mask for the whole image, (2) tissue-level mask for duct bounding boxes and (3) tissue-level mask for duct instances. Table V shows that extracted features from all levels help improve performance.

Impact of classifiers: We study the impact of different classifiers in Table VI. Compared to widely used MLP and SVM, the random forest delivered the best performance. This is likely because random forests reduce high-variance by ensembling many trees into one model. This reduces over-fitting and improves performance, especially on small datasets (like ours).

V. Discussion

Interpretation: While parameters inside a random forest can be hard to understand, we adapted SHAP [34], a game theory-based approach, for interpretation. After training our diagnosis model, we applied SHAP to interpret the diagnostic decision for each ROI and search for the most important features among all ROIs. Table VII compares the 10 most important features from the DIOP and from the tissue-level machine learning model (with 0.67 accuracy) in the ablation experiments.

The BD (boundary of ducts) values in co-occurrence features occur when a pixel is adjacent to the border of a mask or bounding box, which matches the boundary of duct tissues and thus does not have a second pixel to co-occur with. DIOP identifies two co-occurrence features related to BD as important features, which is consistent with the intuition of structure features. Even if the tissue-level model can identify similar co-occurrence features, it is unable to use information inside ductal regions; on the other hand, DIOP mostly focuses on duct masks and also bounding boxes, and it only ranks one ROI-level feature to the top-10 most important features. More clinical studies are needed to verify the consistency of our interpretations with pathologists in the future.

Diagnosis: The proposed DIOP outperforms existing end-to-end and feature engineering approaches. For the four-way classification task, the DIOP achieves comparable performance to general pathologists. In Section IV-D, the general pathologists' diagnostic accuracy is 70% for this unique dataset, which over-sampled DCIS and Atypia cases. In the real-world setting, pathologists diagnosing accuracy is over 92% [35].

Ducts	Tissue	Accuracy
✓		0.57
	✓	0.67
✓	✓	**0.70**

TABLE III: Impact of duct-level instance segmentation and tissue-level semantic segmentation.

Histogram	Co-occurrence	Accuracy
✓		0.66
	✓	0.69
✓	✓	**0.70**

TABLE IV: This table studies the impact of different features extracted from duct-level and tissue-level masks (Section III-C). We did not use superpixel and structural features [31] because (1) they are computationally expensive and (2) our method delivers better performance with these simple features (**0.70** vs. 0.66).

Method	Accuracy
Tissue in ROI	0.67
Tissue in Duct box	0.66
Tissue in Duct mask	0.69
Tissue in Duct mask + ROI	0.69
Tissue in Duct box + ROI	0.67
Tissue in Duct box + mask	0.69
Tissue (All)	**0.70**

TABLE V: Impact of extracting features from different levels (segmentation ROI mask, duct mask, and duct boxes).

Method	Accuracy
SVM (polynomial)	0.62
SVM (RBF-kernel)	0.65
MLP	0.66
Random Forest (DIOP)	**0.70**

TABLE VI: Impact of different classification algorithms.

Rank	DIOP (ours)	Tissue-level model
1	BD & BE in duct mask	ME & NC in ROI
2	ME & NC in duct mask	BG & NC in ROI
3	BD & NC in duct mask	SC freq in ROI
4	BE & NS in bounding box	BE freq in ROI
5	BG & NC in duct mask	BE & SC in ROI
6	BE & SC in ROI	ME & NS in ROI
7	ME & SC in bounding box	BE & NS in ROI
8	NC freq in bounding box	NC freq in ROI
9	BE & SC in bounding box	NS & NC in ROI
10	DS freq in duct mask	SC & NC in ROI

TABLE VII: The top-10 important features from SHAP interpretation. **Left:** results from the Ductal Instance-oriented Pipeline (DIOP). **Right:** results from the classifier for tissue-level semantic features (aka ROI-level features).

Limitation: Note that our dataset was only obtained from 240 biopsies, and additional studies are needed to fully examine our method. With future improvement in semantic segmentation and instance segmentation approaches, our system has the potential to achieve higher accuracy. In this study, we explored clinical relevant features for cancer diagnosis, but we used two separate networks for duct-level and tissue-level segmentation; these could be combined in the future.

Data Annotation: In this study, AI-generated semantic segmentation masks helped the annotator to understand breast biopsies and to perform some easy annotation tasks. These an-

notations were then used to train instance segmentation models for diagnosis purposes. More comprehensive studies, such as controlled and counterbalanced human factor experiments, are needed to investigate the effectiveness of this human-in-the-loop design and other types of human-AI interaction, such as educating pathologist trainees.

Future directions: Many questions need to be answered before deploying these CAD systems, for examples: the generalizability to different datasets, the interpretability of models, the robustness of application to diverse images sets, and the vulnerability to noise and adversarial attacks. Abnormal breast histopathology and breast cancer are complex and heterogeneous disease processes that still require human experts to supervise the important diagnostic decision process.

VI. Conclusion

In this study, we proposed the Ductal Instance-Oriented Pipeline for breast pathology and cancer diagnosis, which contains a duct-level instance segmenter, a tissue-level semantic segmenter, three-levels of pixel-wise features, and a diagnostic classifier. To train the special instance-level segmentation model, we adapted weak annotation and human-in-the-loop design to acquire training data. The proposed method outperforms previous computer-aided approaches in all diagnostic tasks. It also outperforms general pathologists in 2 out of 3 binary classification tasks and almost matches overall performance for general pathologists on our unique dataset. More comprehensive studies are needed to validate our approach in the future.

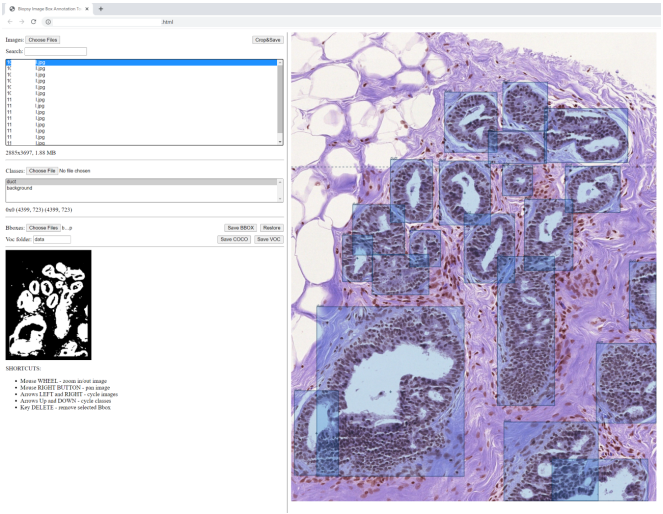


Fig. 4: **The graphical user interface (GUI) to annotate duct instances:** the main panel (right-hand side) shows the input image and allows users to create bounding box annotations. The bottom left side shows the binary tissue mask to guide the annotator. The top left section allows the annotator to load, select, and save images and annotations. Filenames are removed in this visualization for privacy concerns. This GUI is developed based on HTML, Javascript, and YOLO BBox Annotation Tool[3].

References

[1] American Cancer Society. Breast cancer facts & figures 2019-2020. *Atlanta: American Cancer Society*, 2019.

[2] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C Corrado, Ara Darzi, et al. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[3] Ezgi Mercan, Sachin Mehta, Jamen Bartlett, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. Assessment of machine learning of breast pathology structures for automated differentiation of breast cancer and high-risk proliferative lesions. *JAMA network open*, 2(8):e198777–e198777, 2019.

[4] Baris Gecer, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. *Pattern recognition*, 84:345–356, 2018.

[5] Caner Mercan, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. From patch-level to ROI-level deep feature representations for breast histopathology classification. In *Medical Imaging 2019: Digital Pathology*, volume 10956, page 109560H. International Society for Optics and Photonics, 2019.

[6] Bulut Aygüneş, Selim Aksoy, Ramazan Gökberk Cinbiş, Kemal Kösemehmetoğlu, Sevgen Önder, and Ayşegül Üner. Graph convolutional networks for region of interest classification in breast histopathology. In *Medical Imaging 2020: Digital Pathology*, volume 11320, page 113200K. International Society for Optics and Photonics, 2020.

[7] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann G Elmore, and Linda Shapiro. Y-net: joint segmentation and classification for diagnosis of breast biopsy images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 893–901. Springer, 2018.

[8] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann Elmore, and Linda Shapiro. Learning to segment breast biopsy whole slide images. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 663–672. IEEE, 2018.

[9] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 552–568, 2018.

[10] Jiayun Li, Karthik V Sarma, King Chung Ho, Arkadiusz Gertych, Beatrice S Knudsen, and Corey W Arnold. A multi-scale u-net for semantic segmentation of histological images from radical prostatectomies. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1140. American Medical Informatics Association, 2017.

[11] Nathan Ing, Zhaoxuan Ma, Jiayun Li, Hootan Salemi, Corey Arnold, Beatrice S Knudsen, and Arkadiusz Gertych. Semantic segmentation for prostate cancer grading by convolutional neural networks. In *Medical Imaging 2018: Digital Pathology*, volume 10581, page 105811B. International Society for Optics and Photonics, 2018.

[12] Jiayun Li, William Speier, King Chung Ho, Karthik V Sarma, Arkadiusz Gertych, Beatrice S Knudsen, and Corey W Arnold. An em-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies. *Computerized Medical Imaging and Graphics*, 69:125–133, 2018.

[13] Jiayun Li, Wenyuan Li, Arkadiusz Gertych, Beatrice S Knudsen, William Speier, and Corey W Arnold. An attention-based multi-resolution model for prostate whole slide imageclassification and localization. *arXiv preprint arXiv:1905.13208*, 2019.

[14] Mohamed Attia, Mohamed Hossny, Saeid Nahavandi, and Anousha Yazdabadi. Skin melanoma segmentation using recurrent and convolutional neural networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 292–296. IEEE, 2017.

[15] Manu Goyal and Moi Hoon Yap. Multi-class semantic segmentation of skin lesions via fully convolutional networks. *arXiv preprint arXiv:1711.10449*, 2017.

[16] Mike van Zon, Nikolas Stathonikos, Willeke AM Blokx, Selim Komina, Sybren LN Maas, Josien PW Pluim, Paul J van Diest, and Mitko Veta. Segmentation and classification of melanoma and nevus in whole slide images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 263–266. IEEE, 2020.

[3]https://github.com/drainingsun/ybat

[17] Wenyuan Li, Jiayun Li, Karthik V Sarma, King Chung Ho, Shiwen Shen, Beatrice S Knudsen, Arkadiusz Gertych, and Corey W Arnold. Path R-CNN for prostate cancer diagnosis and gleason grading of histological images. *IEEE transactions on medical imaging*, 38(4):945–954, 2018.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[19] Donghao Zhang, Yang Song, Siqi Liu, Dagan Feng, Yue Wang, and Weidong Cai. Nuclei instance segmentation with dual contour-enhanced adversarial network. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 409–412. IEEE, 2018.

[20] Fidel A Guerrero-Pena, Pedro D Marrero Fernandez, Tsang Ing Ren, Mary Yui, Ellen Rothenberg, and Alexandre Cunha. Multiclass weighted loss for instance segmentation of cluttered cells. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2451–2455. IEEE, 2018.

[21] Jaeyong Kang and Jeonghwan Gwak. Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. *IEEE Access*, 7:26440–26447, 2019.

[22] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.

[23] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, page 101693, 2020.

[24] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *arXiv preprint arXiv:1910.02923*, 2019.

[25] Therese B Bevers, Benjamin O Anderson, Ermelinda Bonaccio, Sandra Buys, Mary B Daly, Peter J Dempsey, William B Farrar, Irving Fleming, Judy E Garber, Randall E Harris, et al. Breast cancer screening and diagnosis. *Journal of the National Comprehensive Cancer Network*, 7(10):1060–1096, 2009.

[26] Natalia V Oster, Patricia A Carney, Kimberly H Allison, Donald L Weaver, Lisa M Reisch, Gary Longton, Tracy Onega, Margaret Pepe, Berta M Geller, Heidi D Nelson, et al. Development of a diagnostic test set to assess agreement in breast pathology: practical application of the guidelines for reporting reliability and agreement studies (grras). *BMC Women's Health*, 13(1):3, 2013.

[27] Donald L Weaver, Pamela M Vacek, Joan M Skelly, and Berta M Geller. Predicting biopsy outcome after mammography: what is the likelihood the patient has invasive or in situ breast cancer? *Annals of surgical oncology*, 12(8):660–673, 2005.

[28] Patricia A Carney, Steven P Poplack, Wendy A Wells, and Benjamin Littenberg. The new hampshire mammography network: the development and design of a population-based registry. *AJR. American journal of roentgenology*, 167(2):367–372, 1996.

[29] Wenjun Wu, Beibin Li, Ezgi Mercan, Sachin Mehta, Jamen Bartlett, Donald L Weaver, Joann G Elmore, and Linda G Shapiro. MLCD: A unified software package for cancer diagnosis. *JCO Clinical Cancer Informatics*, 4, 2020.

[30] Joann G Elmore, Gary M Longton, Patricia A Carney, Berta M Geller, Tracy Onega, Anna NA Tosteson, Heidi D Nelson, Margaret S Pepe, Kimberly H Allison, Stuart J Schnitt, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama*, 313(11):1122–1132, 2015.

[31] Ezgi Mercan. *Digital Pathology: Diagnostic Errors, Viewing Behavior and Image Characteristics*. PhD thesis, 2017.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Caner Mercan, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. *IEEE transactions on medical imaging*, 37(1):316–325, 2017.

[34] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[35] Joann G Elmore, Heidi D Nelson, Margaret S Pepe, Gary M Longton, Anna NA Tosteson, Berta Geller, Tracy Onega, Patricia A Carney, Sara L Jackson, Kimberly H Allison, et al. Variability in pathologists' interpretations of individual breast biopsy slides: a population perspective. *Annals of internal medicine*, 164(10):649–655, 2016.