MASTER THESIS

RADBOUD UNIVERSITY NIJMEGEN

Algorithmic differentiation of STIC lesions from normal epithelium in the fallopian tube on H&E stained slides

Author:
Niels van den Hork

Student number:
s4572602

Supervisor Radboud University:
Elena Marchiori

Supervisors Radboud UMC:
Japser Linmans
Jeroen van der laak
Joep Bogaerts

Second reader:
Francesco Ciompi

March 10, 2021

ALGORITHMIC DIFFERENTIATION OF STIC LESIONS FROM NORMAL EPITHELIUM IN THE FALLOPIAN TUBE ON H&E STAINED SLIDES

N. van den Hork, J. Linmans, J. Bogaerts, J. van der Laak

Radboud University & Radboud UMC
Nijmegen, The Netherlands

ABSTRACT

Serous tubal intraepithelial carcinoma (STIC) is a precursor lesion to ovarian carcinoma. We developed an automated deep learning model that can detect and segment Serous tubal intraepithelial carcinoma (STIC) lesions in Whole Slide Images (WSIs) of the Fallopian tube by exclusively making use of the morphological properties of STIC found in H&E stained WSIs. We trained a fully convolutional densenet and compared different class groupings and ratios used during training. We also compared a one-step and two-step approach on this task and lastly evaluated whether hard negative mining improved model performance. We evaluated our methods using the ROC and FROC Curves in addition to pixel-level class accuracy.

1. INTRODUCTION

High grade serous ovarian carcinoma (HGSC) is the leading cause of death amongst gynecologic carcinomas. While the incidence rate is comparatively low, the mortality rate is high at 65% [1]. Since there are no specific symptoms for ovarian carcinoma, it is often only detected at an advanced stage. Early detection may be crucial for improving patient outcomes [2]. As such, extensive research into possible biomarkers [3][4][5][6][7], precursors [8] and other methods [9] that lead to early detection has been done, but has not yet proven successful enough to be used in a clinical setting. Often due to low accuracy and the inability to detect small pre-invasive tumors [10][11][12][13][14].

1.1. STIC

Serous tubal intraepithelial carcinoma (STIC) is considered to be a precursor lesion to HSGC and is found in the epithelial tissue of the fallopian tube [2]. It is mainly found in women with a BRCA1/2 mutation and in those who have a family history of ovarian cancer [8][15].

Apart from STIC, other premalignant Fallopian tube lesions are known such as serous tubal intraepithelial lesion (STIL) and p53 lesions. Especially the STIL lesions share morphological similarities with STIC such as nuclear and architectural atypia albeit at a lower rate [16]. Both lesions only occur in the epithelium of the Fallopian tube.

It is hypothesized that STIC lesions can acquire the ability of invasive growth and thus progress to Fallopian tube cancer [17]. Similarly, via shedding of STIC cells to the ovary or the peritoneal cavity, it can progress to ovarian carcinoma (OC) or primary peritoneal carcinoma (PPC) respectively. In practice Fallopian tube carcinoma, ovarian carcinoma and primary peritoneal carcinoma are all considered OC and are treated similarly.

BRCA 1 and 2 are genes that produce proteins which help repair damaged DNA. When either of these genes are mutated or when a harmful mutation is inherited from ones parents, this repairing function may be lost. This in turn increases the lifetime risk of breast cancer from 12% to 72% or 69% for BRCA 1 and 2 respectively, The life time risk of developing ovarian carcinoma is increased from 1.3% to 44% or 17% respectively [18][19]. Although the risk of ovarian cancer is lower than other carcinomas, the mortality rate is much higher[20] due to the absence of effective screening methods. BRCA1/2 carriers are offered enhanced screening or preemptive risk reducing surgery whereby both the ovaries and the fallopian tubes are removed, known as risk reducing salpingo oophorectomy (RRSO). This removal is often performed between 35 and 45 year of age when the desire to have children has been fulfilled, meaning that STIC has had ample time and opportunity to grow and spread. Whilst removing the ovaries reduces the risk of OC, it has many negative side-effects related to the induced early menopause.

With the growing evidence that ovarian carcinoma originates in the Fallopian tubes, alternative surgical methods are offered in an experimental setting, whereby only the Fallopian tubes are removed and the ovaries stay in situ, thus not inducing early menopause. The removed Fallopian tubes are extensively examined histologically. When the fallopian tube contains no (pre)cancerous cells, such as STIC, no additional treatment is considered necessary. However when a STIC lesion is found, the possibility that such a lesion has already shed aberrant cells to the ovaries is considered a reason to remove the ovaries. As such it is vital to successfully identify STIC lesions in order to be able to offer risk reducing salgingectomy with delayed oopharctomy as a safe alternative treatment option to RRSO.

1.2. Detection of STIC

Currently STIC is detected based on histopathological examination of fallopian tube tissue. H&E staining is used to look for aberrant epithelium [17] after which MIB (Ki-67) and p53 stainings [21][22] are performed to confirm the diagnosis. The p53 staining highlights the TP53 tumor suppression gene [23]. Whereas the Ki-67 staining is a marker for cell growth or division [24]. Excessive cell growth is a possible sign of cancerous growth. Detection of STIC without both stainings is considerably more difficult [25]. Unfortunately the marker specific Ki-67 and p53 stainings are relatively expensive to make. Whilst the cost is relevant, it is not a hard limiting factor. A pathologist initially examines an H&E slide and only requests additional stainings if they suspect STIC is present in the slide. If the pathologist does not suspect STIC to be present in the tissue, the additional stainings are not requested. This can lead to the false negatives as it is more difficult to detect STIC in H&E stainings than in the marker specific stainings.

The pathologist often bases their initial decision on the cheaper H&E staining. In which we can only make use of the morphological properties of STIC lesions to differentiate it from other epithelial cells. These properties include:

- Polymorpism [2]
- Larger cells [2]
- The absence of cilia [2]
- Nuclear polychromasia [21]
- Nuclear stratification [21]
- Epithelial stratification [25]
- Mitotic figure [25]

To an untrained eye these cells generally look more chaotic and irregular than regular epithelium. It is difficult, even for a pathologist, to detect STIC in tissue solely using H&E stainings and there is a large interobserver variability [25]. All in all, a lot of effort is required by the pathologist, as they are expected to examine histopathological slides on high magnification for potential STIC cells, which makes diagnosing STIC a highly time consuming task.

Recently, certain computer algorithms have shown promising results on visual tasks [26][27][28][29][30][31], sometimes outperforming humans. These algorithms often make use of trained deep convolutional neural networks that are well suited for tasks such as object detection, segmentation, classification and more. We may be able to utilise these neural networks to assist the pathologist in their workflow to either improve their accuracy and robustness, or to decrease the time required to diagnose a patient.

We hypothesise that a neural network will be able to differentiate STIC from normal tissue even when based solely on the morphological properties that can be observed using H&E stainings. Which means that Ki-67 and p53 stainings might no longer be necessary in detecting STIC. Alternatively, the neural network could be applied to minimize the number of false positives of the pathologist.

Using a neural network will allow pathologist to spend their time more efficiently by highlighting suspicious areas and will help detect STIC. Additionally, our work will enable further research into STIC Lesions. At this moment, quantitized research into STIC is difficult due to the limited known cases and thus limited amount of available data. One of our goals is for the neural network to be able to search through archived fallopian tube tissue samples and return only tissues where the model suspects the presence of STIC. The model will then generate annotations for all these digitized fallopian tubes. This will enable further research into STIC on a scale that is magnitudes greater than was previously possible.

2. RELATED WORK

2.1. The rise of digital pathology

To this day, pathologists look through microscopes to diagnose the tissue of a patient. This is done for each patient individually and the pathologist has to search the entire tissue at high resolution for potential signals of abnormalities. The pathologist has a high work-load which makes it not surprising that some small areas of a tissue may be overlooked or misinterpreted during examination [32][33][34][35]. The pathologist could in theory delegate some of the easier cases to assistants, but even they have to be highly educated in order to do this work. Assisting the pathologist using custom made computer algorithms can make this whole process more scalable, robust and cost-effective.

Since the rise of computers, much effort has been made to help automate various processes in medical sector. In some cases there was early success [36] but the effectiveness was mainly limited to simple tasks. These algorithms were rule based and were heavily reliant on domain knowledge [37]. One example of rule-based algorithms can be found in radiology with the use of the left-right symmetry assumption of the breast to help highlight areas that are more likely to contain an abnormality [38]. These tools were used to assist the radiologist and help them diagnose cases more quickly.

2.2. The use of AI within digital pathology

Deep learning and in particular convolutional neural networks(CNNs) have significantly increased the potential for digital pathology as these networks are able to learn to diagnose tissue while using minimal domain knowledge in the form of labelled training data [39]. CNNs are able to detect visual patterns in images. Conversely, a pathologist follows a much more complex decision process, where they consider data other than the tissue. Pathologists may for example consider the patients medical history. These methods involing deep neural networks could not be used to their full extend until recently due to low computing power and expensive storage. Additionally, *deep* neural networks were not yet feasible due to the vanishing gradient problem[40]. Now that computing power and storage has become significantly cheaper and the vanishing gradient problem has been solved [41], digital pathology is becoming a part of many diagnosis processes[42] [43].

Currently, much research is being done to help pathologists make diagnoses by utilising artificial intelligence. The most obvious way of assisting the pathologist is by automatically highlighting specific objects such as carcinomas or by filtering out cases that are exceedingly normal. The technology is not at the stage where it can be used without the supervision and interpretation of a pathologist. The short term goal is that the pathologist can make use of the predictions in their own diagnosis. This may save them a considerable amount of time and would allow them to serve more patients while still making confident diagnoses with the same level of accuracy. This is especially important with the growing number of patients that have to be served by one pathologist [42][44][45][37]. Example applications of AI within Digital pathology include nuclei-segmentation [46], lymph node metastate detection [47], prostate cancer [48] and Gleason grading [49] in which AI was able to outperform the average pathologist. This work also shows that pathologists assisted by AI were performing better than pathologists without AI assistance.

2.3. Deep Learning in detection of Ovarian Carcinomas

As far as we are aware, this is the first time that algorithms have been applied for the purpose of detecting and segmenting STIC in Whole Slide Images. There has been more research into detecting ovarian carcinomas in general. These efforts are based on varying input data such as biomarkers, histopathological & cytological images, CT scans or multiphoton microscopy images.

A neural network trained on RNA sequencing has produced a miRNA algorithm for diagnosis of epithelial ovarian cancer (EOC) [50]. Zhang et al. trained an artificial neural network that used four tumor marker values (CA-125 II, CA 72–74, CA 15–13 and M-CSF) as input to detect early-stage ovarian cancer [51]. Similarily Donach et al. used an artificial neural network with four marker values (OVX1, M-CSF, CA19–19 and CA 72–74) as input [52]. Lu et al. predict the origin of Cancers of Unknown Primary (CUP) using a convolutional neural network [53]. Wu et al. apply deep convolutional neural networks to cytological images to classify ovarian cancer types [54]. Wang et al. Predict the risk of recurrence using a densenet in high-grade serous ovarian cancer based on CT scans [55]. Huttunen et al. classifies multiphoton microscopy images of ovarian tissue using deep learning into the classes healthy or high-grade serous carcinoma [56].

2.4. Whole Slide Images

In digital pathology, representation of a scanned tissue are stored digitally as Whole slide images(WSIs). These slides regularly exceed the size of 100GB and are stored in a so called multi resolutional image format (mrxs), which stores representations of the tissue in 'tiles' at different levels of resolution. This allows the pathologist to digitally observe, diagnose and annotate the tissue at one of several levels of zoom or resolution. This is similar to how a pathologist would observe tissue through a microscope. Since the filesizes of WSIs are much larger than what a normal computer is able to load into memory, the mrxs format must allow us to partially load the WSI.

For the purpose of diagnosing a patient, a pathologist will come to a slide-level or patient-level conclusion by inspection slides in a digital slide viewer. In contrast, in order to create a dataset usable for the training of neural networks they make pixel-level annotations by encircling regions of the tissue data. ASAP (Automated Slide Analysis Platform)[1] is used at the Radboud UMC as the digital slide viewing and annotating software. In our case, the annotations are stored in xml format which will later be converted to labels per pixel that can be used in the training and evaluation of neural networks.

2.5. Patching

Similarly to the pathologist, a neural network is able to make either pixel-level or slide-level predictions. A pixel-level prediction in the form of a heatmap is much more interpretable than a slide-level prediction. Due to the large size of the highest resolution layer of WSIs, it becomes more complex to create a slide-level predictor, as we are not able to load the a whole WSI into (GPU) memory at once. These two reasons motivate us to create a patch-based pixel-level predictor. A patch is a small part of a larger image, which is often used when the original image cannot be loaded into memory. The input to the neural network will be a patch, the output can be either an output patch or a center pixel prediction. The former is often accomplished with a U-net [30] style architecture, while the latter can be achieved with several different architectures such as a densenet [57], in this case the training data will be pairs of patches and labels of the center pixel. The densenet will use the whole patch to compute a prediction for the center pixel.

[1] https://computationalpathologygroup.github.io/ASAP/

2.6. Heatmaps

To apply a trained network to a WSI is what we refer to as inference. We will use the network to predict a value for each pixel in the WSI, all these predictions together are what forms a heatmap. A heatmap can be interpreted as an image that can be overlayed on the original WSI, where it can highlight specific tissue. For example, our heatmap can be used to visually highlight STIC. We can use this heatmap further to determine on a slide level whether or not STIC is present. If, for example, the heatmap shows that a large region of the WSI is predicted as STIC, we conclude that STIC is present in that WSI.

2.7. Fully convolutional Densenet

In the current digital pathology the main form of AI is the convolutional neural network. This type of network excels at learning complex patterns from labelled image data. For our research we will use a fully convolutional Densenet [57] which has so far shown good results in digital pathology. A densenet is an neural network architecture which makes extensive use of residual connections that are also seen in Resnets[41]. A densenet is made up of blocks, each several layers deep. Each block is connected to the other blocks by means of feed forward residual connections. This means that feature maps of previous layers can be used in later layers without redundancy. A fully convolutional densenet is a modification of a densenet in which only convolutional layers are used. As a result, this network can accept inputs of any size. This architecture has already been used to detect malignant regions in the prostate [58], to classify interstitial lung disease [59], the automatic scoring of nuclear pleomorphism spectrum in breast cancer [60] and more.

3. DATA

Our dataset consists of 60 Whole Slide Images (WSIs) of the Fallopian tube of individual patients each with a BRCA1/2 mutation. This data was collected from several hospitals within the Netherlands and the tissues have been excised and archived between 2001 and 2019. Each of these images has been annotated by a pathologist, who marked areas within the image as either normal epithelium, STIC or STIL. Annotations of non epithelial regions were provided by pathologist in training. These classes include bloodvessels, cysts, cancer, fat and inflammation. The STIC and STIL classes only occur in epithelial tissue and are annotated exhaustively in our dataset. Each other class has been sparsly annotated, which means that no effort has been made to annotate, for example, each bloodvessel. In practice, the majority of these tissues are left unnanotated, as a sufficient amount of annotations had already been collected. A WSI is not guaranteed to include each class, except for one of STIC or STIL. We will refer to all aforementioned classes as *annotation classes*. Later we will group these into *model classes*, on which we will train and evaluate our model.

We also have access to 65 healthy WSIs containing tissue of the Fallopian tube. These contain no malignant tissue and have not been annotated any further. We will make use of these slides when we evaluate our model using the ROC and FROC curves. The absence of annotations is a non-issue for this purpose. Since our model should distinguish only between STIC and non-STIC, we assume that any prediction of STIC in these slides is a false positive. For the purpose of testing we also do not require any non-epithelial annotations, as we are only interested in the STIC/ non-STIC performance.

We will split our dataset into 50 WSIs for training, 5 for validation and 5 for testing. Each split was made to contain annotations from all classes. We have created 5 crossvalidated permutations of this dataset, in which we ensured that each validation and test slide was not included in more than one split. These five splits allows us to evaluate on 25 validation WSIs and 25 test WSIs. There is no leakage since all WSIs are from unique patients. For 250 epochs, we will sample 400 batches of 32 patches of 279x279 pixels from the WSI, where the center pixel of the patch is within the annotations. We sample our patches equally from all model classes unless stated otherwise, this means we will train on the same amount of STIC, healthy epithelium and non epithelium. The patches are extracted from the second highest resolution level available in our WSIs: 0.5 micrometer per pixel. Afterwards they are normalized and randomly augmented using flipping, rotating, scaling, by changing the color or contrast, by blurring the image, by adding noise or by using a filter which changes the staining.

After our initial training, we observed different annotation habits between the two annotators, where one included lumen in the epithelium annotations whereas the other did not. This caused a bias towards one of the classes and caused the model to mispredict lumen and perform poorly on STIC. We changed these annotations such that they would include as little lumen as possible which lessened the observed effect.

4. METHODS

Here we will introduce the experiments that we performed and we will provide motivation for each. These experiments all aim to improve the performance at which the model can differentiate STIC from normal epithelium tissue.

4.1. Class grouping & ratios

There are several choices that can be made in how we use the dataset most effectively. One example is how we process the annotation classes into model classes that can be supplied to the network: what is the best way to group them into model classes? Another example is at which ratios we should sample the classes, is it better to use balanced training data, or should we match the distribution found in real data? In the latter case the model would mainly train on non-epithelium tissue.

In addition to STIC and normal epithelium annotations, we have made annotations of other tissue: Non epithelium, blood vessels, cancer, cysts, fat and inflammation. An important question is how to use this information to improve the performance on the differentiation between STIC and normal epithelium. For example: Cancer and STIC look similar to each other, it may be beneficial to group Cancer with STIC during training. Grouping Cancer with STIC means that the model does not need to learn to differentiate STIC from Cancer, which is much more difficult to do compared to distinguishing STIC from epithelium for example. The model would use a large part of its capacity to make the decision between STIC and Cancer, while this capacity can better be spend on learning the difference between STIC and epithelium. Additionally adding Cancer to STIC can be seen as increasing the amount of data for STIC classes. The model can then make use of patterns in both STIC and Cancer to be more robust and overfit less. Lastly, it is not an issue if our final model predicts Cancer as STIC, both STIC and Cancer are (pre)malignant and the pathologist can easily distinguish between the two. Conversely, we might decide that Cancer should be grouped with non-epithelium, as the cancerous cells are present in the stroma and not in the epithelium. While cancerous cells originate from epithelial cells, they have undergone significant changes that make them easily separable from epithelial cells. There is also the possibility that not including cancer at all, or at a lower sampling rate, is the best option. There are several tissue types that raise similar questions which we answer later on in this thesis.

4.2. Slide level metrics

As mentioned in section 1.2, One of our goals is to be able to filter archived Fallopian tube tissue samples and return only the slides where the model predicts the presence of STIC. As such, it is important to develop a slide-level metric that can be used to find only the cases that are most likely to contain STIC. The resulting availability of Fallopian tube tissue annotated on slide-level with STIC will allow for quantitative research into STIC. As it allows researchers to use larger datasets of tissue samples to build their work on.

To be able to calculate a correct slide level metric, our model should predict a correct pixel level prediction for every pixel in a WSI. We collect for each pixel in the WSI the corresponding model prediction, we call this a pixel-level heatmap. From this heatmap we should be able to compute the slide-level prediction: Is there STIC present in this WSI or not? Since our model is applied to all pixels in a WSI, it should produce a reasonable prediction for each of those pixels. That means that our model should perform well on any type of tissue that can be found in the Fallopian tube. In order to do this, our model should be able to classify STIC, (healthy) epithelium and all other tissue (non-epithelium). The non epithelium model class should encompass each type of tissue that is not either STIC, STIL or healthy epithelium, among others it will include stroma, blood vessels and lumen.

4.3. Two-step or one-step

We established that we need to differentiate not only between STIC and normal epithelium, but also between epithelium and non-epithelium. We can accomplish this in two different ways. The first is to train one model to distinguish these three classes. Another solution is to build a 2-step approach. This entails that we first differentiate only between epithelium and non-epithelium, which gives us a mask of all epithelium tissue. In the second step we will then, from this mask, differentiate between STIC and normal epithelium. This method may have higher performance as both tasks are easier in definition. It also allows for a post-processing step to combine both masks. The availability of two masks instead of one, may give us more insight into the performance of the individual models.

4.4. Hard negative mining

Lastly, we will perform hard negative mining [61][62][63] and evaluate whether or not this improves the slide and object level metrics. Hard negative mining is the act of training on data that the neural network has miss-predicted on. The assumption is that training more frequently on relatively difficult data will improve model performance, as the new additional labels correct the model on its miss-predictions. Hard negative mining aims to reduce the amount of false positives. To perform hard negative mining, we first generate a prediction mask using a trained model, then we compare this mask against the annotations provided by the pathologist. We identify the areas where the model predicted the presence of STIC, but where the pathologist did not. We assume that the pathologist is correct in all cases. We save all these locations to a separate mask where they are marked as healthy epithelium. This mask will be used to fine-tune our trained model for 10 epochs at a low learning rate. In our case we apply hard negative mining only to the STIC class. This is because the STIC class is the only annotation class that has the guarantee of being exhaustively annotated. If we were to, for example, perform hard negative mining on our epithelium annotations, we would find that the majority of the 'hard negatives' found are actually just unannotated epithelium, fine-tuning on these hard negatives would be catastrophic. This is the reason why we can only perform hard negative mining for the STIC class. Since we are only able to compute hard-negatives over the STIC annotations, we can only fine-tune the second model(STIC vs non-STIC) in the two-step approach. We cannot fine-tune the one-step model as we cannot automatically give a correct true label to this hard negative. We cannot apply this to the first two-step model (Epithelium

	STIC	Healthy Epithelium	Non-epithlium	Not included
Cancer	x		x	x
Cyst		x	x	x
STIL	x			x

Table 1. Class grouping experiment one-step. The rows indicate the annotations class. The columns indicate the model class, or the exclusion from the dataset. The crosses indicate that we evaluate a grouping strategy where the respective model class is assigned to the respective model class.

vs non-epithelium) as this model does not distinguish STIC. We can only fine-tune the second model of the two-step approach.

5. EXPERIMENTS

In this section we will state the design and implementation details of each experiment.

5.1. Class grouping & ratios

In section 4 we mentioned how we could group the annotation classes into model classes, to find the optimal configuration. For some annotation classes, such as cancer, we can use our domain knowledge to build hypotheses, which we can then test empirically. In the example of cancer, we can place it in the non epithelium class grouping, since cancer is not epithelium. But we could also place it in the STIC class, since it is visually very similar to STIC. Table 1 shows the configurations that we will evaluate. We will furthermore test the setup where we classify each of the 10 classes individually. We expect that this method will have inferior performance compared to the other setups, but we are interested in the resulting confusion matrix.

Lastly, we will try out different sampling ratios of the model class. We aim to better match the distribution of the class imbalance that can be found in real world data. We test three configurations in total, where we sample healthy epithelium classes at a rate of 1, 3 or 9 respectively.

5.2. Two-step or one-step

In the two step approach we must decide again which annotation labels constitute to epithelium or non epithelium and also which labels belong to STIC or non-STIC. Table 2 shows what types of annotation classes the model classes contain. We chose these classes based on the results from the class grouping experiment.

We combine the masks from the two networks to create a continuous heatmap that only activates on STIC. First we threshold the epithelium output masks, which results in a binary heatmap where a pixel is only either Epithelium or not epithelium, represented by ones and zeroes respectively. We then remove connected components of under 50 pixels in this mask and multiply it with the STIC mask to find the intersect of the two masks. This mask contains zeroes at all pixels where the prediction is non epithelium, and holds a continuous prediction for STIC on epithelial pixels.

Model class	**Annotation classes**
Epithelium	STIC, STIC near cancer, Epithelium, Cystic Epithelium
Non-epithelium	Non Epithelium, Bloodvessel, Fat, Inflammation
STIC	STIC, STIC near cancer
Non-STIC	Normal Epithelium, cystic Epithelium

Table 2. Class grouping two-step model

5.3. Hard negative mining

Machine learning models will almost always make some incorrect predictions. In the case of our two-step model we face a binary classification task, this means that there are two types of mistakes the model could make: false positives, where healthy tissue is predicted as STIC, and false negatives, where STIC is predicted as healthy. In our case we are provided with exhaustively annotated STIC labels. Conversely, healthy tissue is sparsely annotated. From this we can infer that unnanotated regions are healthy tissue, any STIC prediction in these regions will be a false positive. Our model will, by default, not be able to learn from this, as the data is not labelled and as such it will never be sampled at training. We can however still leverage this data, by identifying which areas of the WSI the network predicts incorrectly as a false positive, save these locations and retrain or fine-tune our model using these extra annotated datapoints. On additionally benefit of hard negative mining is that we only train on the patches which the model had difficulty with before. This means that extra time can be spend during training to perfect the model, instead of wasting resources going through mostly easy patches.

We performed hard negative mining on the two-step approach. We obtained hard negative masks by computing and thresholding the post processed output of the two step approach and subtracting from this the true positives as found in our annotation data. We have dilated these true positives with a uniform 10x10 kernel as to not include this area in our hard negative mask. This method will find hard negatives not just in already annotated areas, but anywhere in the complete WSI.

We then supplemented the previous dataset with these hard negative masks and fine-tuned the densenet on this new dataset for 10 epochs at a low learning rate. Since this method is expected to lead to less false positives, we do not expect to see much improvement within the annotations themselves as they are either true positives or false negatives to begin with.

We do expect to see an improvement on the ROC and FROC of the slide level STIC risk predictor.

As mentioned, we are only able to compute hard-negatives over the STIC annotations and can only fine-tune the second model(STIC vs non-STIC) in the two-step approach.

6. METRICS & EVALUATION

6.1. Class accuracy

We use class accuracy as the pixel level metric to compare how our models perform within the (non exhaustive) annotations and to observe the performance on a class to class basis.

6.2. STIC risk score

Since a STIC detection model would likely be used to identify cases of STIC in archived data, it is important to develop a slide-level metric that can be used to find only the cases that are most likely to contain STIC. We achieve this by computing the STIC-risk score which is the ratio between STIC and Epithelium in a given Whole Slide Image, which is calculated from the prediction heatmap:

$$\text{STIC-risk} = \frac{\#\text{STIC}}{\#\text{Epithelium}}$$

6.3. ROC

We evaluate this slide level risk score on both recall and precision using the ROC Curve, where we will prioritise precision to eliminate false positives. The false positive rate should be low as we can assume that the large majority of archived fallopian tube tissue will not contain STIC. We will also calculate the corresponding area under the ROC Curve (AUC). This measure will give insight into how well the slide-level STIC risk score can be used to filter archived data. For this use case we prioritise reducing the amount of false positives.

6.4. FROC

We will also compute the object-based free-response receiver operating characteristic (FROC)[64] and the area under the FROC curve (AUFC) in the same manner as is done in CAMELYON16 [65]. A FROC Curve is a plot of the sensitivity (or true positive rate) against the average number of false-positives per whole slide image [65].

When considering how the model will be used by a pathologist, we must evaluate how much time they will have to spend per slide in order to come to a conclusion. Without the model the pathologist has to verify for each epithelial cell that it is not STIC, this is naturally very time consuming. Our model could significantly speed up this process. The rate at which this happens correlates with the amount of false positives per image, as the pathologist has to consider each of these areas until they identify STIC. As such we use a method which compares the amount of false positives per slide against the sensitivity.

We will compute the FROC and ROC for each cross validation. Each of these have 5 test WSIs as positives, which are guaranteed to contain some amount of STIC, and all cross validation splits use the same 65 Healthy WSIs as negatives, which contain no STIC at all. We will compute seperate ROC and FROC curves per split. In total we will use 25 positive test slides. We will report the individual and averaged AUC/AUFC.

7. RESULTS

Here we will show the results for each experiment and highlight key results. The discussion and drawing of conclusions from these findings will be shared in the sections hereafter.

7.1. Class grouping and ratios

Table 3 shows the pixel level class accuracies for different class groupings. We have included Table 7 in the appendix, this table shows per grouping strategy exactly which annotation classes belong to which model class, we strongly recommend to use this table as a reference point.

Grouping strategy	STIC	HEALHTY epithelium	NON epithelium
CancerNON	.75	.69	**.96**
CancerSTIC *	.78	.73	.93
CystHEALTHY	.74	.71	.95
CystNON	.74	.68	**.96**
NOcancercyst	**.82**	.66	.95
NOcancercyststil *	.72	**.79**	.94
10class	.66	.58	n/a

Table 3. Class accuracies for different class grouping strategies. The strategies marked with an asterisk hold the best performance when considering the sum of STIC and Healthy class accuracies

Here follows the compacted overview our base grouping strategy entails: we include STIC and STIC near cancer in the STIC model class, normal epithelium as the healthy epithelium model class and Non-epithelium, Bloodvessels, fat and inflammation in the Non-epithelium model class. This base grouping strategy does not change. What does change, is where we assign Cancer, Cyst and STIL. For each strategy we will distribute these three annotation classes to different model classes. Each grouping type is evaluated on the same data using the class grouping of the base strategy to allow for a fair comparison.

Table 7 also shows exactly how each strategy differs from the base strategy. Again, we provide the compacted

Ratio strategy	STIC	HEALHTY	NON
		epithelium	epithelium
HEALTHYx1 *	**.72**	.79	.94
HEALTHYx3	.53	.86	**.95**
HEALTHYx9	.0	**.98**	.68

Table 4. Class accuracies for different class ratio strategies. The strategies marked with an asterisk hold the best performance when considering the sum of STIC and Healthy class accuracies

overview here: cancerNON means that the cancer annotations were interpreted as if they were non epithelium. CancerSTIC, conversely, would mean that we group cancer with the STIC model class. NOcancercyst means that we did not include either of cancer and cyst annotations in our training. HEALTHYx3 means that we oversampeled Healthy epithelium by a factor of 3, whereas we normally train on balanced data. This was an attempt to more closely match the actual distribution found in whole slide images. The 10class grouping was trained to observe the confusion matrix, which is included in the appendix as Figure 3

From Table 3 we can see that not including cancer and cysts (NOcancercyst) in our training is beneficial for STIC performance, the performance on healthy epithelium however drops when compared to other groupings. When considering the performance on healthy epithelium, we see that not including cancer, cysts or STIL at all (NOcancercyststil) is best. The non epithelium class accuracy does not depend heavily on the chosen grouping strategy. In practise we will use the strategy that leads to the maximal performance in both STIC and Healthy Epithelium accuracies. We see that, when considering the sum of these two, excluding cancer, cyst and STIL leads to the best class accuracies together with the strategy of assigning cancer to STIC. Table 4 shows that increasing the sampling ratio of healthy epithelium does not improve overall performance when compared to the base ratio (HEALTHYx1). This means that we should not increase the sampling rate of healthy epithelium.

7.2. Two-step or one-step

Within our annotations the two-step approach improves slightly upon the one-step approach by achieving a class accuracy on STIC of 84%. We also evaluate the choice between the one-step and two-step approach by using the area under the ROC and FROC curves to gain insight into the slide and region level performance. These metrics were higher in the two-step approach when compared to the one-step, this can be seen in Table 5. Especially the AUFC score is improved in the two-step approach. As mentioned before, we use a 5-fold cross validated dataset for to train and evaluate the two-step approach. The area under the ROC and FROC curves for each individual split can be observed in Table 6.

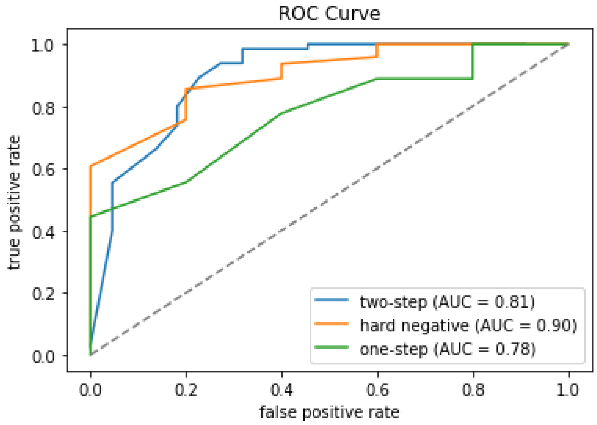


Fig. 1. ROC Curves for one-step, two-step and hard negative mining

Experiment type	AUC	AUFC
One-step	0.78	0.64
Two-step	0.81	0.74
Two-step + Hard negative mining	0.90	0.86

Table 5. AUC and AUFC values for one-step,two-step and hard negative mining

Figure 1 shows corresponding ROC curves of both approaches. We have included the ROC curves of the individual splits in the appendix as Figure 2

7.3. Hard negative mining

Hard negative mining significantly improves slide level performance as can be seen by the area under the ROC and FROC Curves in Tables 5 and 6 and Figure 1. This is as expected as the hard negative mining technique aims to remove false positives.

The original false positive predictions that were used in this hard negative mining were sometimes characterized by artefacts in the scanning process, but most often the false positives were regions of tissue that were visually similar to STIC or they were some other tissue that the network had not encountered before during training. These false positives could

cross validation split	AUC	AUFC
Two-step cv1	1.00	0.80
Two-step cv2	0.78	0.52
Two-step cv3	0.75	0.68
Two-step cv4	0.64	0.89
Two-step cv5	0.97	0.81

Table 6. AUC and AUFC values for the individual splits in the two-step approach. See Figures 2, 4 and 5l in the appendix for the ROC and FROC Curves per split

be significantly reduced through the process of hard negative mining. Reducing the amount of false positives is crucial to determining a slide level metric which considers the whole WSI.

8. DISCUSSION

In finding the optimal class grouping strategy we experimented mainly with where to assign cancer, Cysts and STIL. For each of these annotation classes, we provided arguments that it is not trivial to find the one correct model class, as each annotation class could potentially be assigned to another model class as well. We found through our experiments that the best class grouping strategies were: 1) not including cancer, cysts and STIL (NOcancercyststil), or 2) assigning cancer to STIC and assigning STIL and Cysts according to the base strategy (cancerSTIC). To support the first strategy of NOcancercyststil, we can make the argument that the cancer, cyst, STIL annotation classes do not add enough beneficial information for training the network and mostly distract the network. These annotation classes are, as mentioned, not obviously assignable to one specific model class. This added complexity does not outweigh the potential increase of relevant training data that these annotations could bring.

Including cancer in the STIC class achieves the same summed performance as NOcancercyststil. Cancer is visually very similar to STIC tissue, the largest difference is that Cancer is not epithelium and as such often present in the stroma instead of near the lumen. The network is able to make use of this extra training data of cancer to improve its classification on STIC. This also implies that there are not enough STIC annotation in the dataset and annotating and training on more STIC annotations will be beneficial for the model performance.

Since both strategies were virtually equal in performance, we can look at the practical application and identify the strategy that may be the most convenient to develop further. One aspect of future development will be collection a larger data set. This includes making annotation of the new tissue. As the Cyst and STIL annotations do not significantly contribute to the performance of the model, it may be optimal to focus on the annotations of tissue such as healthy epithelium in addition to the aforementioned STIC.

Our two step approach performs better than the one step. This is likely because the first step model segments the epithelium, which is relatively easy. Our second model then only has learn to be able to differentiate STIC versus Healthy epithelium. This split allows both model to specialise in these simpler tasks and effectively use twice the model capacity of a densenet. There are two added benefits to using the two step approach. The first is that since we have the intermediate epithelium prediction, the system is not as much of a black box, we have access to an intermediate epithelium mask. The second is that it is modular, it is simple to replace the epithelium detection network for a different model for example.

Hard negative mining improved the slide level performance of the two-step model as can be seen from the increased area under the ROC and FROC curves. The amount of false positives after the hard negative mining is reduced significantly. To reduce these further, effort should be made to add more STIC and epithelial annotations to the training data. The detection of artefacts would also improve the model performance.

9. CONCLUSION

We presented a deep convolutional neural network that can segment STIC lesions in Whole slide images and we found the optimal configuration to be a two-step approach with hard negative mining. The model can be used by a pathologist to assist them in the detection of STIC in a H&E stained slide. Furthermore, we defined a STIC-risk score that can be used as a slide level predictor to filter archived data for the purpose of future research. Our method achieved an AUC of 0.90 and an AUFC of 0.86. Additional performance and robustness can still be achieved through the expansion of the training set, mainly by increasing the amount of epithelium and STIC annotations.

10. FUTURE WORK

10.1. Different architectures

In this work we used DenseNet and obtained good results. Since our Densenet is a fully convolutional model that predicts the center pixel of a patch, experimenting with a pure segmentation network such as a U-net could be worthwhile. Alternatively, using a context aware architecture such as HookNet[66] could improve performance by combining context and detail.

10.2. Hyper parameter optimization

We have not performed extensive hyper parameter optimization on our models. We expect that experimenting with different patch sizes, resolution levels, model configurations or architectures will lead to a better performing model.

10.3. Exhaustively annotate data

We mentioned that our data is not exhaustively annotated, except for the STIC class. This means that on a pixel level we can only evaluate and make claims about sensitivity (recall), but not precision. Other measures such as a DICE score also become more difficult to implement and extract meaning from as a result of non-exhaustively annotated data. In order to more extensively evaluate a model, it is beneficial to create a limited number of test slides or test regions that are exhaustively annotated.

10.4. Annotate more data

Throughout our experiments we have noticed that there are likely not enough STIC and epithelium annotation present in the data for optimal performance. We expect that collecting and annotating more data can significantly increase the model performance on both the pixel level class accuracies and the slide level metrics.

11. REFERENCES

[1] Rebecca Siegel, Jiemin Ma, Zhaohui Zou, and Ahmedin Jemal, "Cancer statistics, 2014," *CA: a cancer journal for clinicians*, vol. 64, no. 1, pp. 9–29, 2014.

[2] Alison Karst and Ronny Drapkin, "Ovarian cancer pathogenesis: A model in evolution," *Journal of oncology*, vol. 2010, pp. 932371, 01 2010.

[3] Long Nguyen, Segundo Joel Cardenas-Goicoechea, Pierre Gordon, Christina Curtin, Mazdak Momeni, Linus Chuang, and David Fishman, "Biomarkers for early detection of ovarian cancer," *Women's health*, vol. 9, no. 2, pp. 171–187, 2013.

[4] Dmitriy W Gutkin, Michael R Shurin, El Azher, Mounia Alaoui, Galina V Shurin, Liudmila Velikokhatnaya, Denise Prosser, Namhee Shin, Francesmary Modugno, Paul Stemmer, et al., "Novel protein and immune response markers of human serous tubal intraepithelial carcinoma of the ovary," *Cancer Biomarkers*, vol. 26, no. 4, pp. 471–479, 2019.

[5] Qing Zhang, Guohong Hu, Qifeng Yang, Ruifen Dong, Xing Xie, Ding Ma, Keng Shen, and Beihua Kong, "A multiplex methylation-specific pcr assay for the detection of early-stage ovarian cancer using cell-free serum dna," *Gynecologic oncology*, vol. 130, no. 1, pp. 132–139, 2013.

[6] Martin Widschwendter, Michal Zikan, Benjamin Wahl, Harri Lempiäinen, Tobias Paprotka, Iona Evans, Allison Jones, Shohreh Ghazali, Daniel Reisel, Johannes Eichner, et al., "The potential of circulating tumor dna methylation analysis for the early detection and management of ovarian cancer," *Genome medicine*, vol. 9, no. 1, pp. 116, 2017.

[7] Ian Jacobs, Jane Bridges, Colin Reynolds, Isabel Stabile, Philippa Kemsley, Jurgis Grudzinskas, and David Oram, "Multimodal approach to screening for ovarian cancer," *The Lancet*, vol. 331, no. 8580, pp. 268–271, 1988.

[8] Vit Weinberger, Marketa Bednarikova, David Cibula, and Michal Zikan, "Serous tubal intraepithelial carcinoma (stic) – clinical impact and management," *Expert Review of Anticancer Therapy*, vol. 16, no. 12, pp. 1311–1321, 2016.

[9] JP Lerner, IE Timor-Tritsch, A Federman, and G Abramovich, "Transvaginal ultrasonographic characterization of ovarian masses with an improved, weighted scoring system," *American journal of obstetrics and gynecology*, vol. 170, no. 1, pp. 81–85, 1994.

[10] John R van Nagell Jr, Paul D DePriest, Frederick R Ueland, Christopher P DeSimone, Amy L Cooper, J Matt McDonald, Edward J Pavlik, and Richard J Kryscio, "Ovarian cancer screening with annual transvaginal sonography: findings of 25,000 women screened," *Cancer*, vol. 109, no. 9, pp. 1887–1896, 2007.

[11] David A Fishman, Leeber Cohen, Stephanie V Blank, Lee Shulman, Diljeet Singh, Kenny Bozorgi, Ralph Tamura, Ilan Timor-Tritsch, and Peter E Schwartz, "The role of ultrasound evaluation in the detection of early-stage epithelial ovarian cancer," *American journal of obstetrics and gynecology*, vol. 192, no. 4, pp. 1214–1221, 2005.

[12] Shigemi Sato, Yoshihito Yokoyama, Tomomi Sakamoto, Masayuki Futagami, and Yoshiharu Saito, "Usefulness of mass screening for ovarian carcinoma using transvaginal ultrasonography," *Cancer*, vol. 89, no. 3, pp. 582–588, 2000.

[13] JR van Nagell Jr, PD DePriest, MB Reedy, HH Gallion, FR Ueland, EJ Pavlik, and RJ Kryscio, "The efficacy of transvaginal sonographic screening in asymptomatic women at risk for ovarian cancer," *Gynecologic oncology*, vol. 77, no. 3, pp. 350–356, 2000.

[14] Thomas H Bourne, Stuart Campbell, Karina M Reynolds, Malcolm I Whitehead, Jayne Hampson, Patrick Royston, TJ Crayford, and William P Collins, "Screening for early familial ovarian cancer with transvaginal ultrasonography and colour blood flow imaging.," *British Medical Journal*, vol. 306, no. 6884, pp. 1025–1029, 1993.

[15] Emily E K Meserve, Jan Brouwer, and Christopher P Crum, "Serous tubal intraepithelial neoplasia: the concept and its application," *Modern Pathology*, 2017.

[16] Amy Gross, Robert Kurman, Russell Vang, Ie-Ming Shih, and Kala Visvanathan, "Gross al, kurman rj, vang r, shih iem, visvanathan kprecursor lesions of high-grade serous ovarian carcinoma: morphological and molecular characteristics. j oncol 2010: 126295," *Journal of oncology*, vol. 2010, pp. 126295, 04 2010.

[17] David W. Kindelberger, Yonghee Lee, Alexander Miron, Michelle S. Hirsch, Colleen Feltmate, Fabiola Medeiros, Michael J. Callahan, Elizabeth Garner,

Robert W. Gordon, Chandler Birch, Ross S. Berkowitz, Michael G. Muto, and Christopher P. Crum, "Intraepithelial carcinoma of the fimbria and pelvic serous carcinoma: Evidence for a causal relationship," *The American Journal of Surgical Pathology*, vol. 31, pp. 161–169, 2007.

[18] Krapcho M et al. Howlader N, Noone AM, "Seer cancer statistics review, 1975-2014," *National Cancer Institute. Bethesda, MD,*, 2018.

[19] Barnes DR et al. Kuchenbaecker KB, Hopper JL, "Risks of breast, ovarian, and contralateral breast cancer for brca1 and brca2 mutation carriers.," *JAMA. 2017*, vol. 317, no. 23, pp. 2402–2416, 2017.

[20] Krapcho M Miller D Brest A Yu M Ruhl J Tatalovich Z Mariotto A Lewis DR Chen HS Feuer EJ Cronin KA Noone AM, Howlader N, "Seer cancer statistics review, 1975-2015," *National Cancer Institute. Bethesda, MD,*, 2018.

[21] Elke Jarboe, Ann Folkins, Marisa R Nucci, David Kindelberger, Ronny Drapkin, Alexander Miron, Yonghee Lee, and Christopher P Crum, "Serous carcinogenesis in the fallopian tube: a descriptive classification," *International Journal of Gynecological Pathology*, vol. 27, no. 1, pp. 1–9, 2008.

[22] Y Lee, A Miron, R Drapkin, MR Nucci, F Medeiros, A Saleemuddin, J Garber, C Birch, H Mou, RW Gordon, DW Cramer, FD McKeon, and CP Crum, "A candidate precursor to serous carcinoma that originates in the distal fallopian tube," *The Journal of Pathology*, vol. 211, no. 1, pp. 26–35, 2007.

[23] Anna Yemelyanova, Russell Vang, Malti Kshirsagar, Dan Lu, Morgan A Marks, Ie Ming Shih, and Robert J Kurman, "Immunohistochemical staining patterns of p53 can serve as a surrogate marker for tp53 mutations in ovarian carcinoma: an immunohistochemical and nucleotide sequencing analysis," *Modern pathology*, vol. 24, no. 9, pp. 1248–1253, 2011.

[24] Thomas Scholzen and Johannes Gerdes, "The ki-67 protein: From the known and the unknown," *Journal of Cellular Physiology*, vol. 182, no. 3, pp. 311–322, 2000.

[25] Kala Visvanathan, Russell Vang, Patricia Shaw, Amy Gross, Robert Soslow, Vinita Parkash, Ie-Ming Shih, and Robert J Kurman, "Diagnosis of serous tubal intraepithelial carcinoma based on morphologic and immunohistochemical features: a reproducibility study," *The American journal of surgical pathology*, vol. 35, no. 12, pp. 1766, 2011.

[26] M. Z. Hossain, Ferdous Sohel, M. F. Shiratuddin, and Hamid Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys (CSUR)*, vol. 51, pp. 1 – 36, 2018.

[27] L. Jiao, Fan Zhang, F. Liu, Shuyuan Yang, L. Li, Zhixi Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019.

[28] Ross Girshick, "Fast r-cnn," *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun 2017.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, p. 234–241, 2015.

[31] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C. Berg, "Dssd : Deconvolutional single shot detector," 2017.

[32] Elizabeth A Krupinski, "Visual scanning patterns of radiologists searching mammograms," *Academic radiology*, vol. 3, no. 2, pp. 137–144, 1996.

[33] Richard E Bird, Terry W Wallace, and Bonnie C Yankaskas, "Analysis of cancers missed at screening mammography.," *Radiology*, vol. 184, no. 3, pp. 613–617, 1992.

[34] Joann G Elmore, Carolyn K Wells, Carol H Lee, Debra H Howard, and Alvan R Feinstein, "Variability in radiologists' interpretations of mammograms," *New England Journal of Medicine*, vol. 331, no. 22, pp. 1493–1499, 1994.

[35] EJ Potchen, "Variation in diagnostic accuracy: Potential role of computer-aided diagnosis," in *Computer-Aided Diagnosis in Medical Imaging*, p. 527. Elsevier Amsterdam, 1999.

[36] U. Bottigli, P. Cerello, P. Delogu, M. E. Fantacci, F. Fauci, G. Forni, B. Golosio, P. L. Indovina, A. Lauria, E. Lopez Torres, R. Magro, G. L. Masala, P. Oliva, R. Palmiero, G. Raso, A. Retico, A. Stefanini, S. Stumbo, and S. Tangaro, "A computer aided detection system for mammographic images implemented on a grid infrastructure," 2003.

[37] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec 2017.

[38] V. Raman, P. Then, and P. Sumari, "Digital mammogram tumor preprocessing segmentation feature extraction and classification," in *2009 WRI World Congress on Computer Science and Information Engineering*, 2009, vol. 2, pp. 507–511.

[39] Carl J Vyborny and Maryellen L Giger, "Computer vision and artificial intelligence in mammography.," *AJR. American journal of roentgenology*, vol. 162, no. 3, pp. 699–708, 1994.

[40] Sepp Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 06, no. 02, pp. 107–116, 1998.

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," 2015.

[42] Hamid Reza Tizhoosh and Liron Pantanowitz, "Artificial intelligence and digital pathology: challenges and opportunities," *Journal of pathology informatics*, vol. 9, 2018.

[43] Navid Farahani, Anil V Parwani, and Liron Pantanowitz, "Whole slide imaging in pathology: advantages, limitations, and emerging perspectives," *Pathol Lab Med Int*, vol. 7, no. 23-33, pp. 4321, 2015.

[44] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan, "Digital pathology and artificial intelligence," *The lancet oncology*, vol. 20, no. 5, pp. e253–e261, 2019.

[45] Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi, "Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology," *Nature reviews Clinical oncology*, vol. 16, no. 11, pp. 703–715, 2019.

[46] Fuyong Xing and Lin Yang, "Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review," *IEEE reviews in biomedical engineering*, vol. 9, pp. 234–263, 2016.

[47] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[48] Hans Pinckaers, Wouter Bulten, Jeroen van der Laak, and Geert Litjens, "Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels," 2020.

[49] Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens, "Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study," *The Lancet Oncology*, vol. 21, no. 2, pp. 233 – 241, 2020.

[50] Kevin M Elias, Wojciech Fendler, Konrad Stawiski, Stephen J Fiascone, Allison F Vitonis, Ross S Berkowitz, Gyorgy Frendl, Panagiotis Konstantinopoulos, Christopher P Crum, Magdalena Kedzierska, et al., "Diagnostic potential for a serum mirna neural network for detection of ovarian cancer," *Elife*, vol. 6, pp. e28932, 2017.

[51] Zhen Zhang, Yinhua Yu, Fengji Xu, Andrew Berchuck, Carolien van Haaften-Day, Laura J Havrilesky, Henk WA de Bruijn, Ate GJ van der Zee, Robert P Woolas, Ian J Jacobs, et al., "Combining multiple serum tumor markers improves detection of stage i epithelial ovarian cancer," *Gynecologic oncology*, vol. 107, no. 3, pp. 526–531, 2007.

[52] Martin Donach, Yinhua Yu, Grazia Artioli, Giuseppe Banna, Weiwei Feng, Robert C Bast, Zhen Zhang, and Maria O Nicoletto, "Combined use of biomarkers for detection of ovarian cancer in high-risk women," *Tumor Biology*, vol. 31, no. 3, pp. 209–215, 2010.

[53] Ming Y. Lu, Melissa Zhao, Maha Shady, Jana Lipkova, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood, "Deep learning-based computational pathology predicts origins for cancers of unknown primary," 2020.

[54] Miao Wu, Chuanbo Yan, Huiqiang Liu, and Qian Liu, "Automatic classification of ovarian cancer types from cytological images using deep convolutional neural networks," *Bioscience reports*, vol. 38, no. 3, 2018.

[55] Shuo Wang, Zhenyu Liu, Yu Rong, Bin Zhou, Yan Bai, Wei Wei, Meiyun Wang, Yingkun Guo, and Jie Tian, "Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer," *Radiotherapy and Oncology*, vol. 132, pp. 171–177, 2019.

[56] Mikko J Huttunen, Abdurahman Hassan, Curtis W McCloskey, Sijyl Fasih, Jeremy Upham, Barbara C Vanderhyden, Robert W Boyd, and Sangeeta Murugkar, "Automated classification of multiphoton microscopy images of ovarian tissue using deep learning," *Journal of biomedical optics*, vol. 23, no. 6, pp. 066002, 2018.

[57] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[58] Chiao-Min Chen, Yao-Sian Huang, Pei-Wei Fang, Cher-Wei Liang, and Ruey-Feng Chang, "A computer-aided diagnosis system for differentiation and delineation of malignant regions on whole-slide prostate histopathology image using spatial statistics and multidimensional densenet," *Medical Physics*, vol. 47, no. 3, pp. 1021–1033, 2020.

[59] Wenping Guo, Zhuoming Xu, and Haibo Zhang, "Interstitial lung disease classification using improved densenet," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 30615–30626, 2019.

[60] Caner Mercan, Maschenka Balkenhol, Roberto Salgado, Mark Sherman, Philippe Vielh, Willem Vreuls, Antonio Polonia, Hugo M. Horlings, Wilko Weichert, Jodi M. Carter, Peter Bult, Matthias Christgen, Carsten Denkert, Koen van de Vijver, Jeroen van der Laak, and Francesco Ciompi, "Automated scoring of nuclear pleomorphism spectrum with pathologist-level performance in breast cancer," 2020.

[61] K.-K. Sung and T. Poggio, "Learning and example selection for object and pattern detection.," *InMIT A.I.Memo*, , no. 1521, pp. 85, 1994.

[62] Meng Li, Lin Wu, Arnold Wiliem, Kun Zhao, Teng Zhang, and Brian C. Lovell, "Deep instance-level hard negative mining model for histopathology images," 2019.

[63] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[64] Chakraborty DP, "Recent developments in imaging system assessment methodology, froc analysis and the search model," *Nucl Instrum Methods Phys Res A. 2011 Aug 21;648 Supplement 1:S297-S301.*, 2011.

[65] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, and Jeroen A. W. M. van der Laak, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA. 2017;318(22):2199–2210.*, 2017.

[66] Mart van Rijthoven, Maschenka Balkenhol, Karina Siliņa, Jeroen van der Laak, and Francesco Ciompi, "Hooknet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images," *Medical Image Analysis*, vol. 68, pp. 101890, Feb 2021.

12. APPENDIX

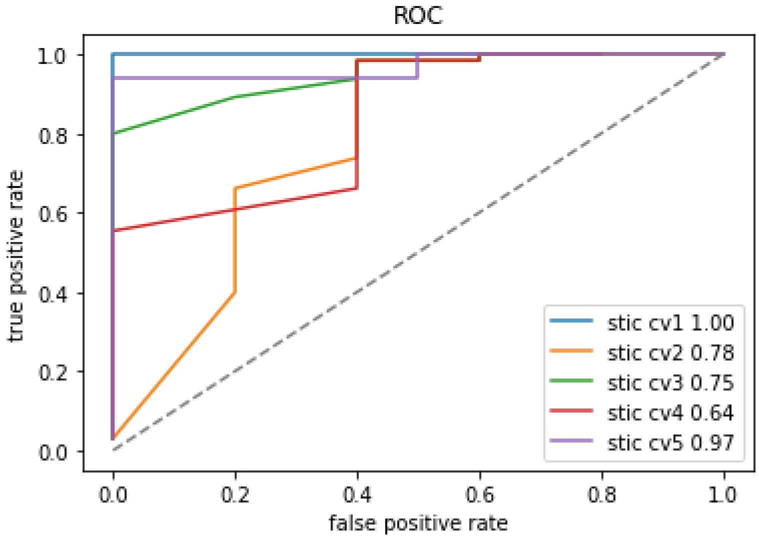
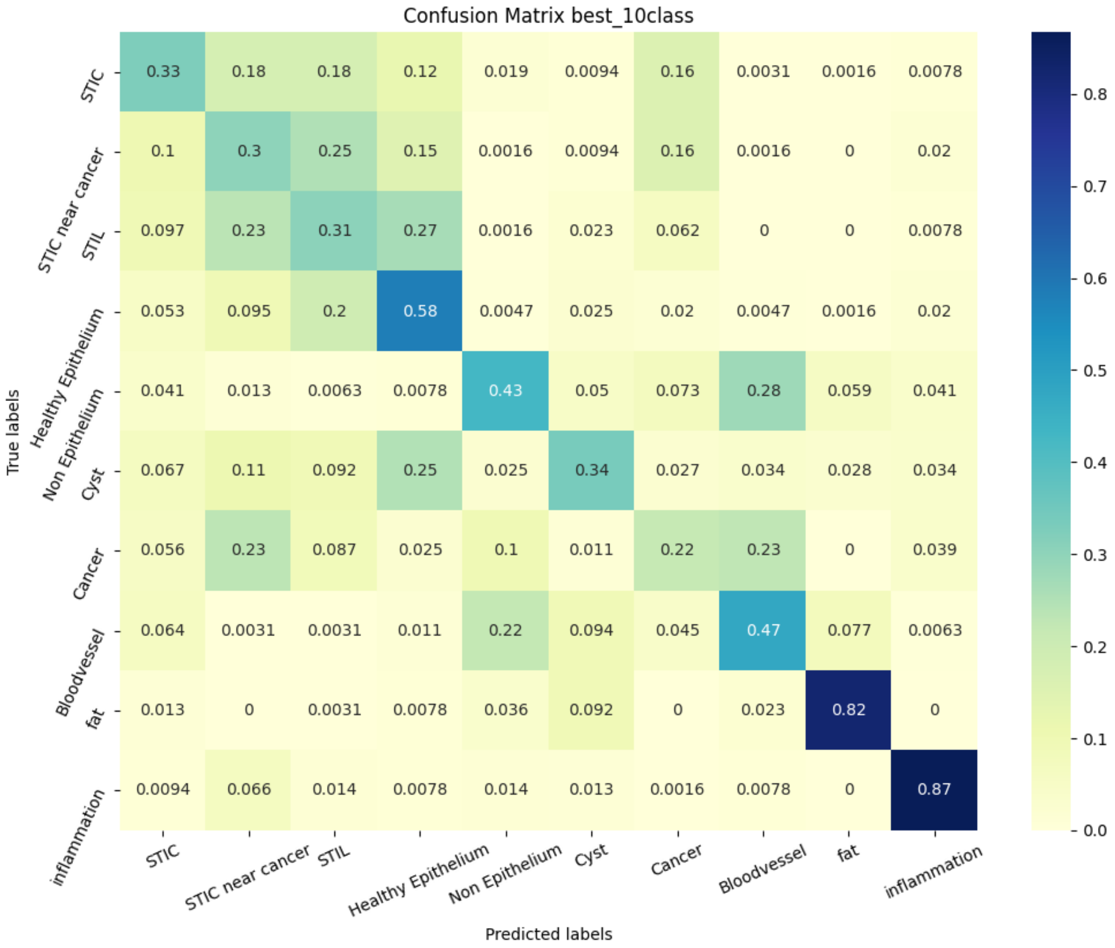


Fig. 2. ROC two-step. For each individual cross validation

	STIC	Healthy Epithelium	Non-epithelium	Not included
Base	STIC, STIC near Cancer	Normal epithelium	Non-epithelium, Blood vessel, Fat, Inflammation	
CancerNON	STIL	Cyst		Cancer
CancerSTIC	Cancer,STIL	Cyst		
CystHealthy	STIL	Cyst	Cancer	
CystNON	STIL		Cancer	Cyst
NOcancercyst	STIL			Cancer,Cyst
NOcancercyststil				Cancer,Cyst,STIL

Table 7. The exact class grouping strategies. The row 'Base' contains all annotation classes that remain static throughout the experiments, the following rows show how the configuration is enhanced compared to the 'Base' row



	STIC	STIC near cancer	STIL	Healthy Epithelium	Non Epithelium	Cyst	Cancer	Bloodvessel	fat	inflammation
STIC	0.33	0.18	0.18	0.12	0.019	0.0094	0.16	0.0031	0.0016	0.0078
STIC near cancer	0.1	0.3	0.25	0.15	0.0016	0.0094	0.16	0.0016	0	0.02
STIL	0.097	0.23	0.31	0.27	0.0016	0.023	0.062	0	0	0.0078
Healthy Epithelium	0.053	0.095	0.2	0.58	0.0047	0.025	0.02	0.0047	0.0016	0.02
Non Epithelium	0.041	0.013	0.0063	0.0078	0.43	0.05	0.073	0.28	0.059	0.041
Cyst	0.067	0.11	0.092	0.25	0.025	0.34	0.027	0.034	0.028	0.034
Cancer	0.056	0.23	0.087	0.025	0.1	0.011	0.22	0.23	0	0.039
Bloodvessel	0.064	0.0031	0.0031	0.011	0.22	0.094	0.045	0.47	0.077	0.0063
fat	0.013	0	0.0031	0.0078	0.036	0.092	0	0.023	0.82	0
inflammation	0.0094	0.066	0.014	0.0078	0.014	0.013	0.0016	0.0078	0	0.87

Fig. 3. Confusion matrix of the 10 class model

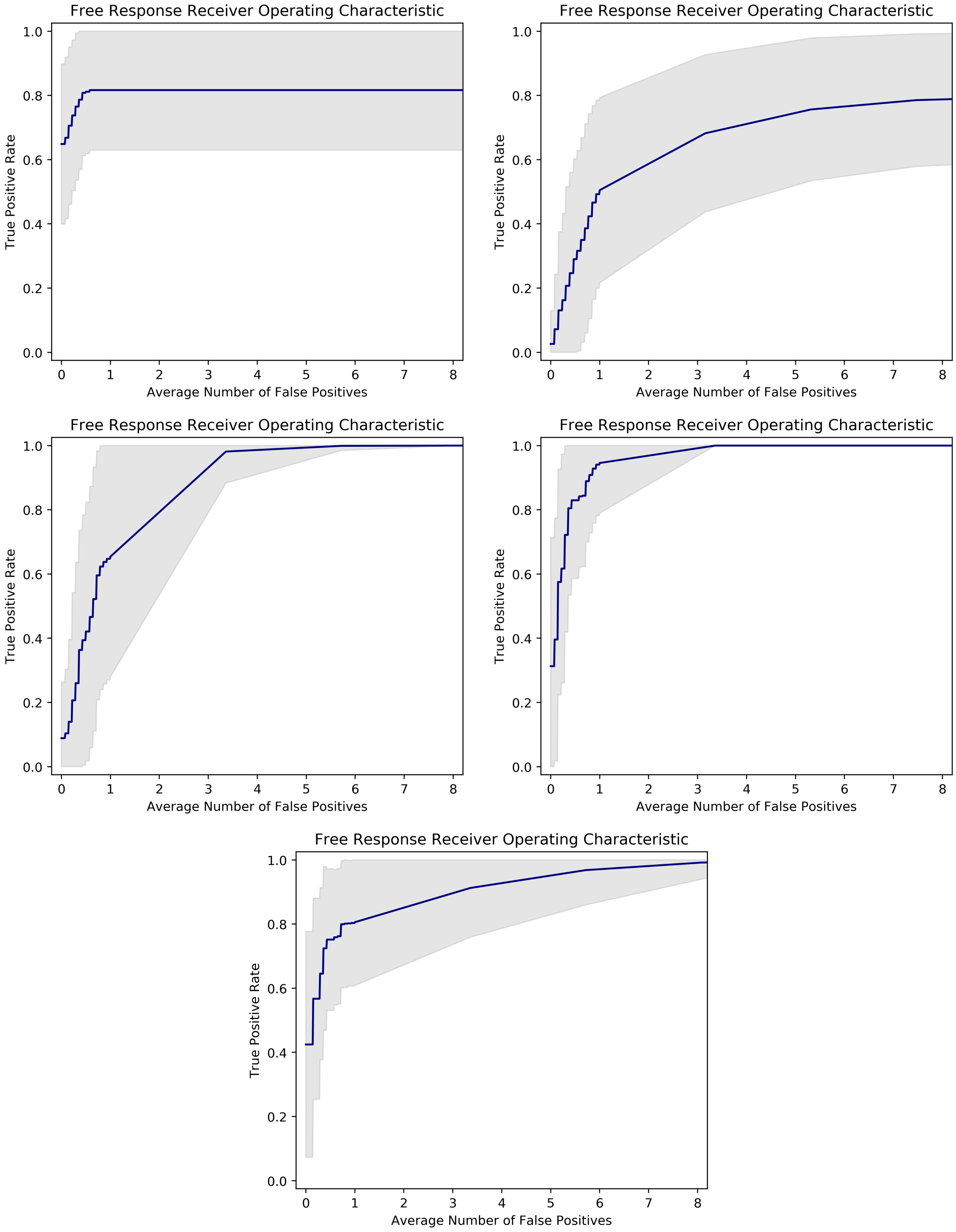


Fig. 4. Froc curves for each split of the two-step model

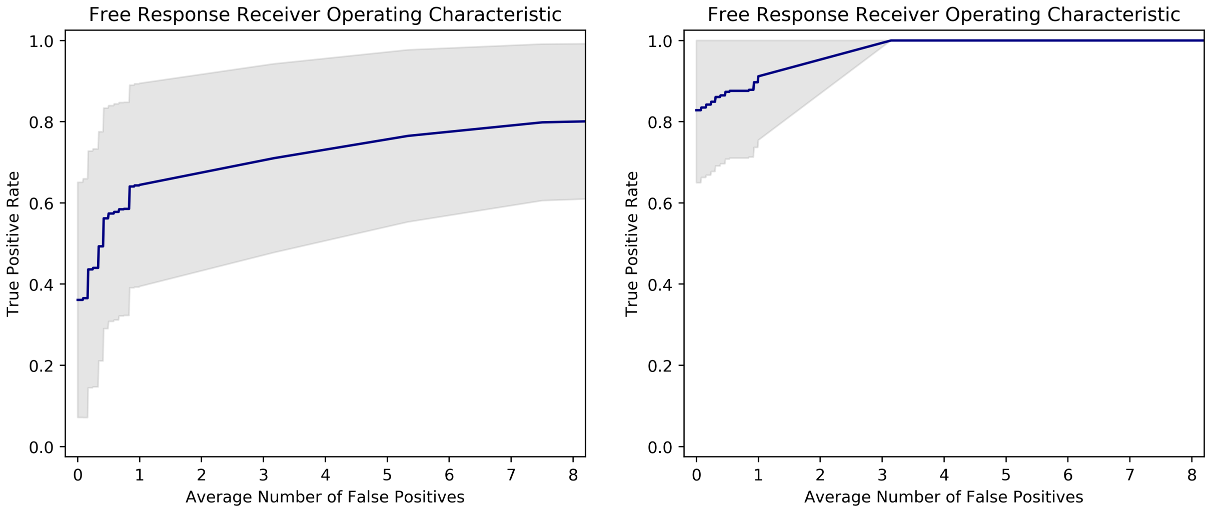


Fig. 5. FROC curves for the onestep (left) and hardnegative (right) model

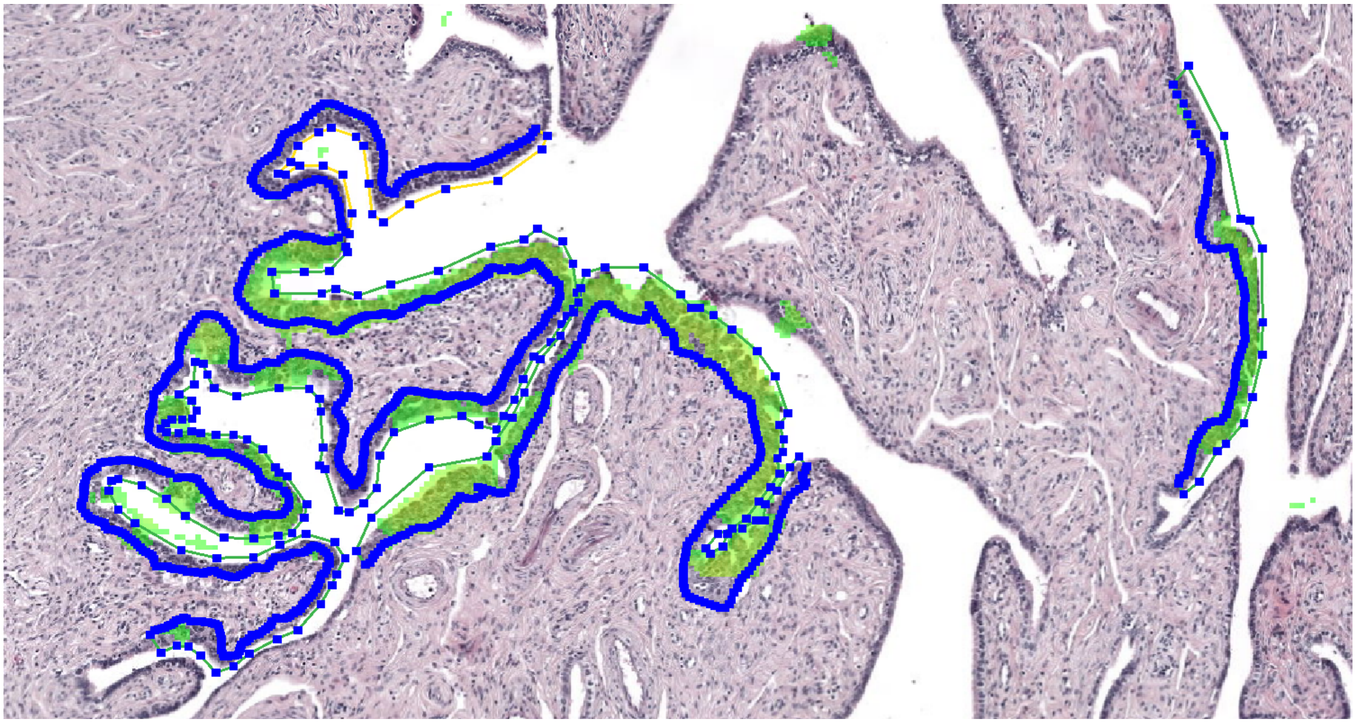


Fig. 6. Example model output on H&E stained tissue from the Fallopian tube. The blue dots each encircle a STIC (or STIL) region. Tissue highlighted in green is predicted as STIC by the model. This example was cherry picked to show the model working as expected

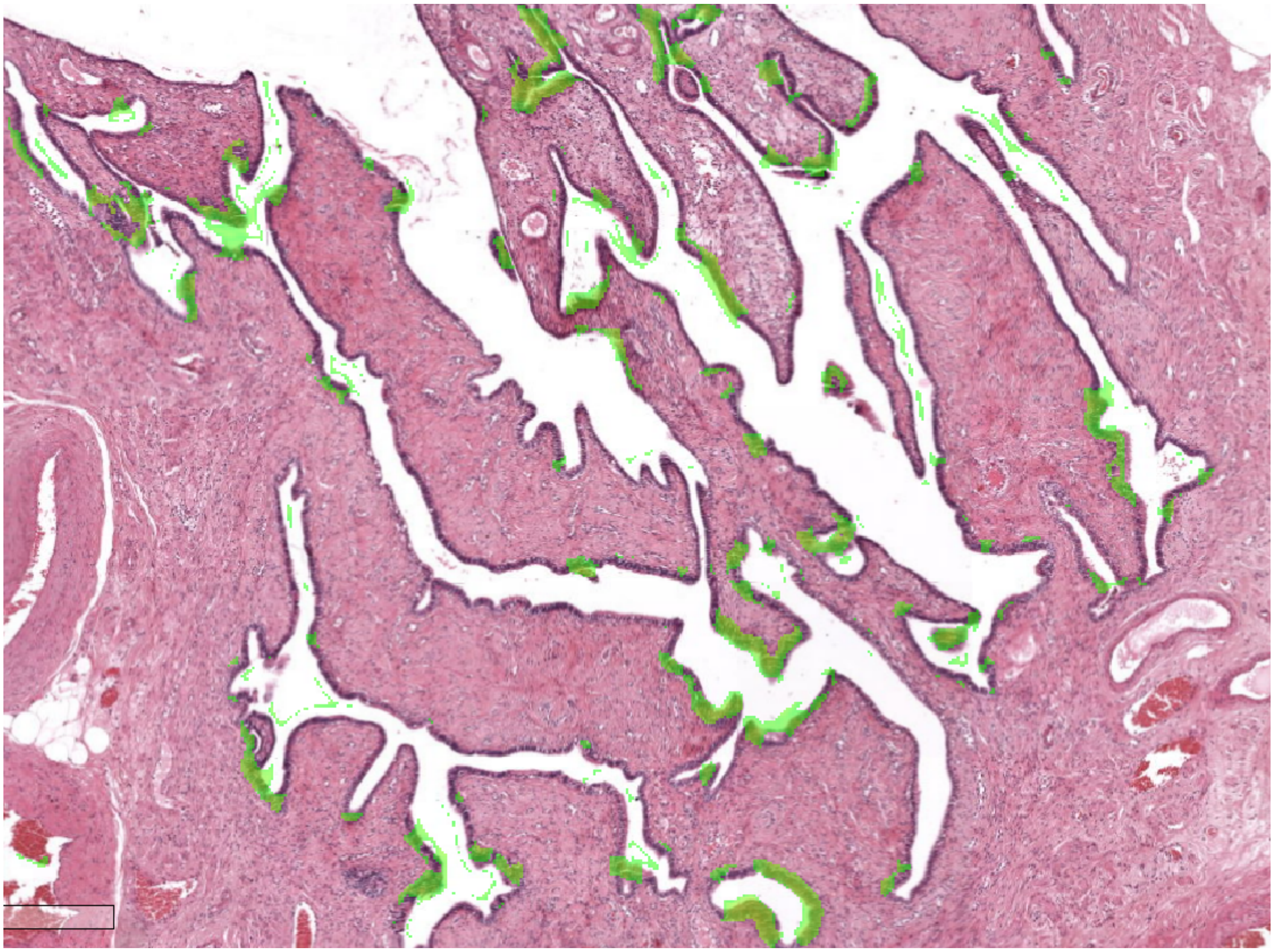


Fig. 7. This example was cherry picked to show the model performing poorly. The model generates false positives on healthy epithelium and sometimes on lumen.

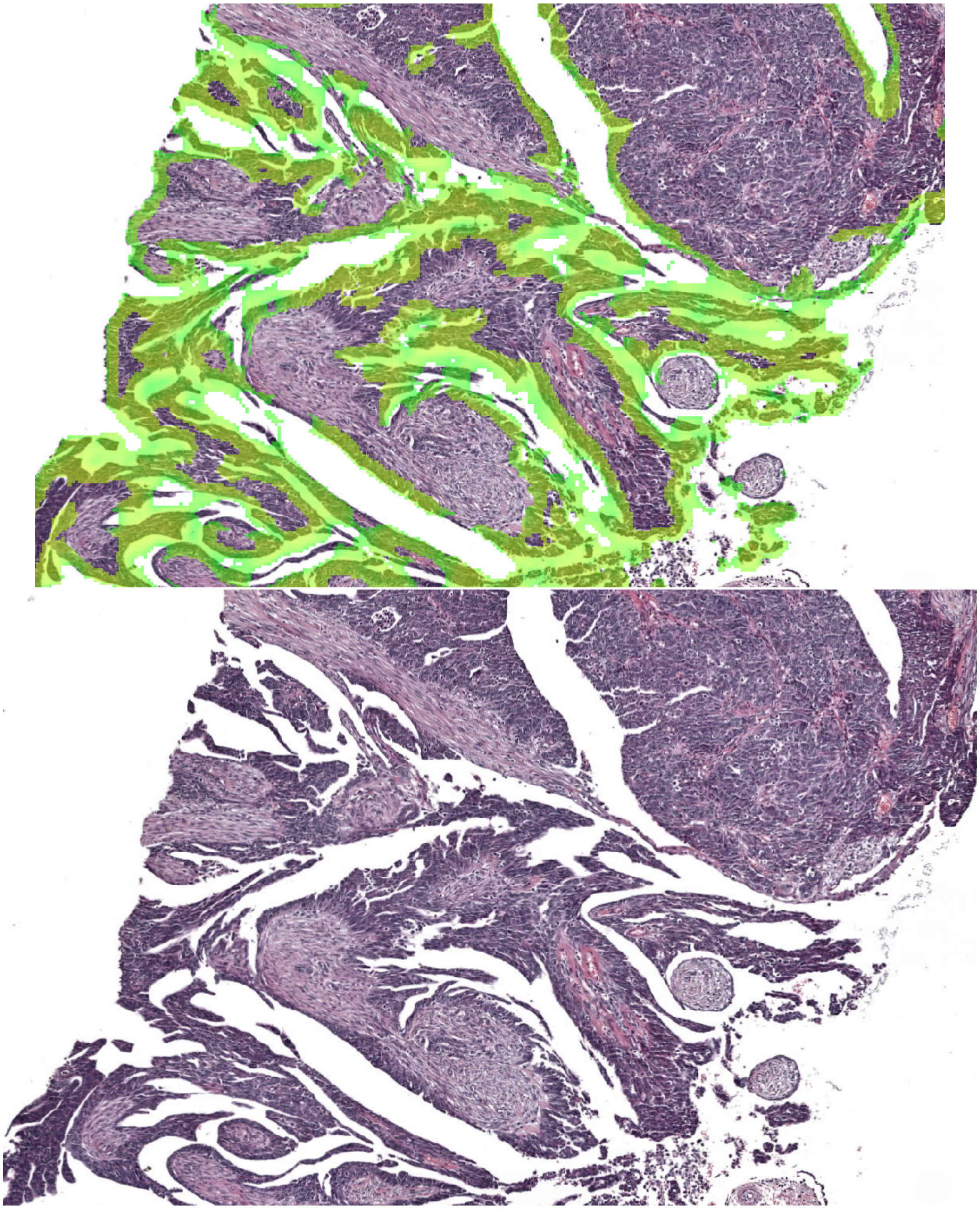


Fig. 8. This examples show the model correctly predicting unnanotated carcinoma as STIC. The pathologist confirmed that this region is malignant carcinoma. This figure shows the same tissue twice, one with the prediction heatmap overlayed and one without

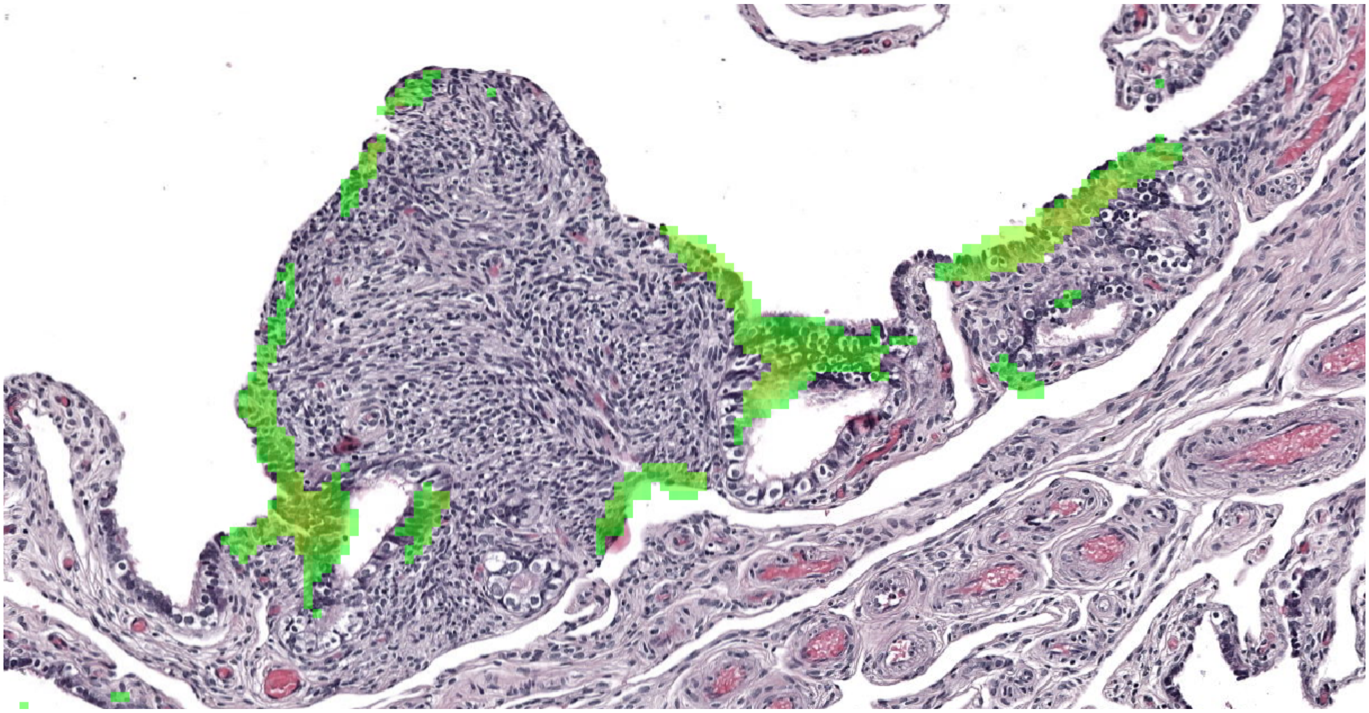


Fig. 9. This example show the model predicting unusual epithelium tissue as STIC. If the pathologist were to diagnose this patient, it would be good to take a closer look at this region.

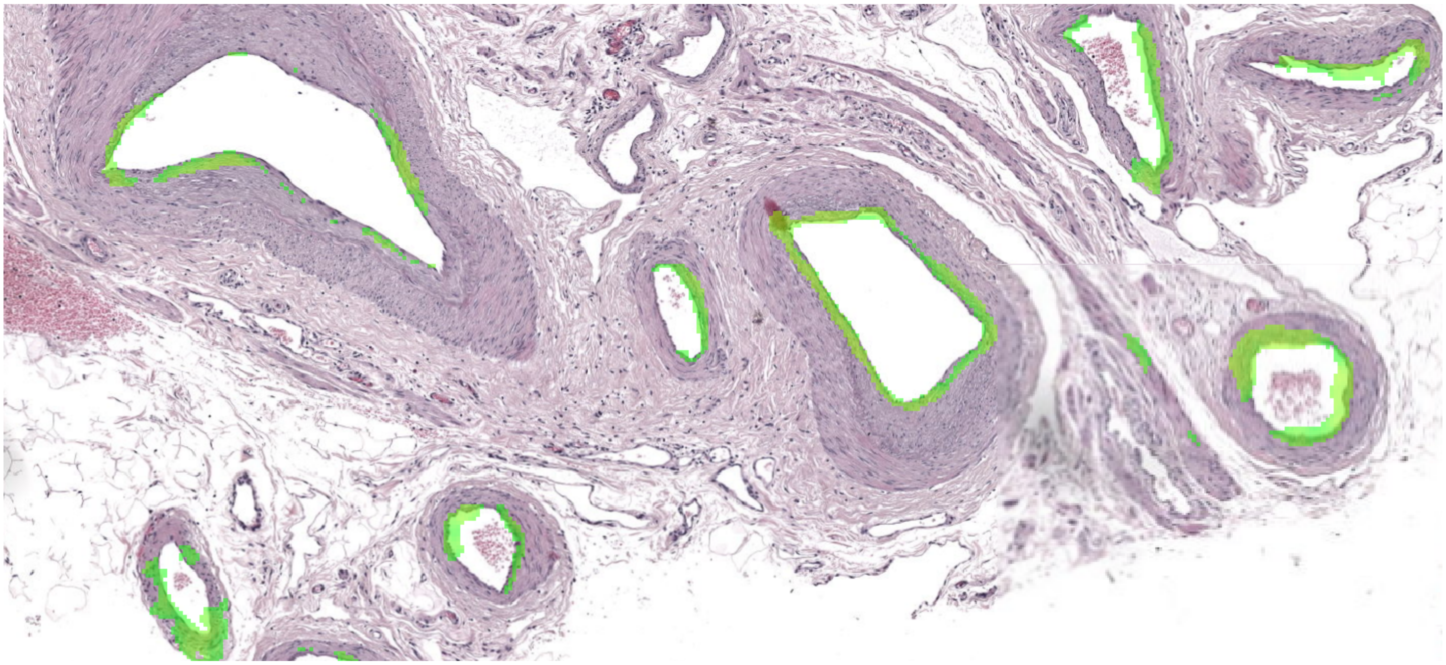


Fig. 10. This example show the model predicting blood vessel epithelium as STIC. This is a recurring false positive that can be easily dismissed by the pathologist during diagnosis.

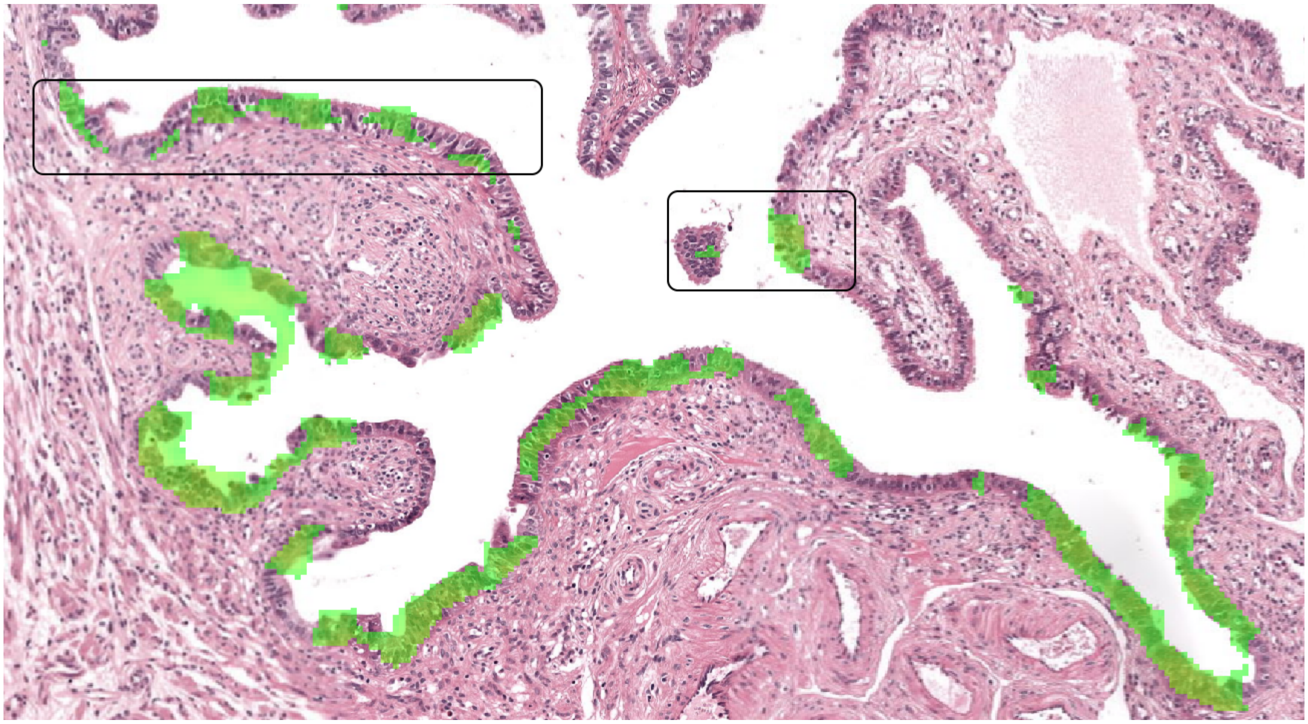


Fig. 11. This example show an average model prediction. Note that the pathologists annotations have been disabled for the purpose of this figure. The two areas surrounded by rounded squares are false positives. The rest of the model predictions are correct.

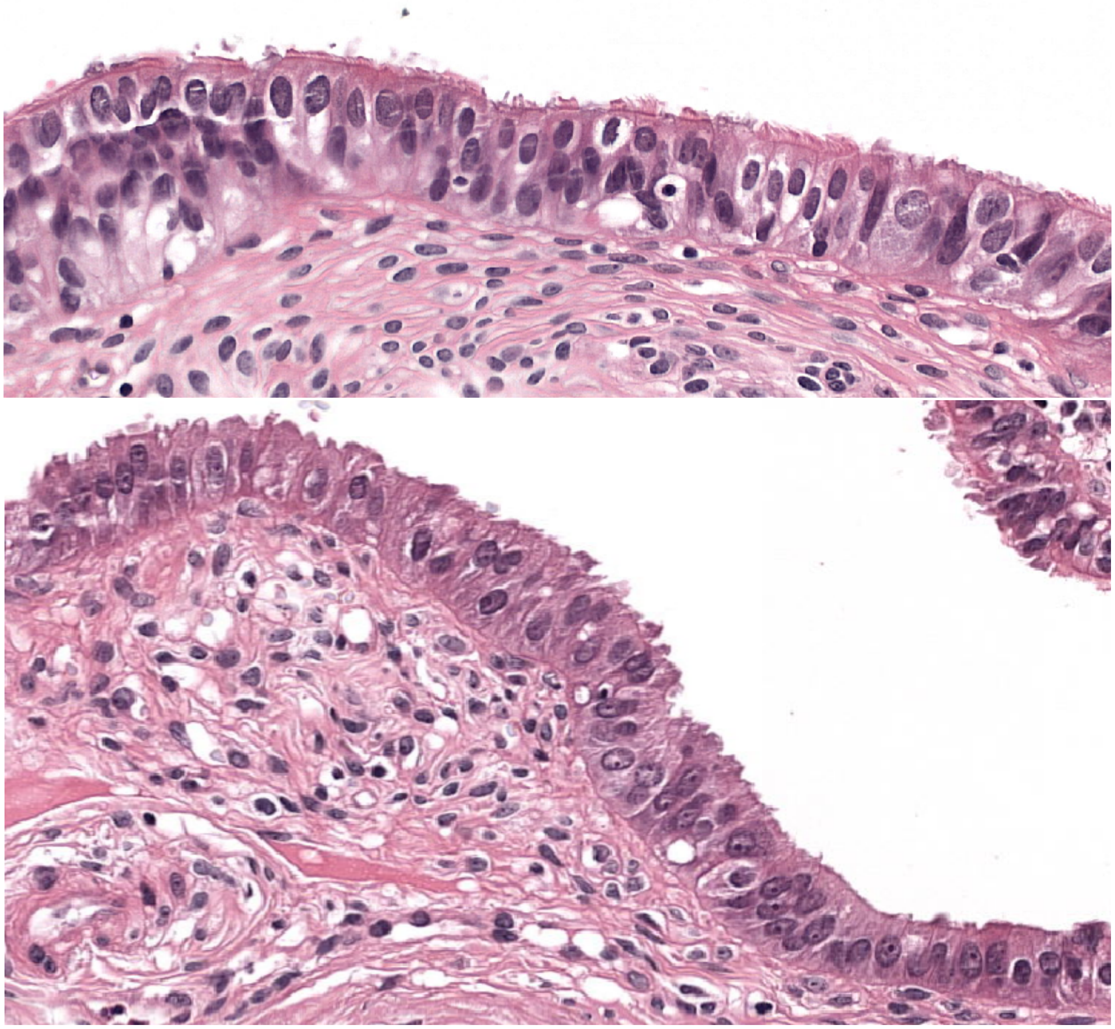


Fig. 12. The first image shows healthy tissue, the second image shows STIC tissue.