Deep Learning With Conformal Prediction for Hierarchical Analysis of Large-Scale Whole-Slide Tissue Images

Håkan Wieslander, Philip J. Harrison, Gabriel Skogberg, Sonya Jackson, Markus Fridén, Johan Karlsson, Ola Spjuth, and Carolina Wählby

***Abstract*—With the increasing amount of image data collected from biomedical experiments there is an urgent need for smarter and more effective analysis methods. Many scientific questions require analysis of image sub-regions related to some specific biology. Finding such regions of interest (ROIs) at low resolution and limiting the data subjected to final quantification at full resolution can reduce computational requirements and save time. In this paper we propose a three-step pipeline: First, bounding boxes for ROIs are located at low resolution. Next, ROIs are subjected to semantic segmentation into sub-regions at mid-resolution. We also estimate the confidence of the segmented sub-regions. Finally, quantitative measurements are extracted at full resolution. We use deep learning for the first two steps in the pipeline and conformal prediction for confidence assessment. We show that limiting final quantitative analysis to sub-regions with full confidence reduces noise and increases separability of observed biological effects.**

***Index Terms*—Conformal prediction, deep learning, digital pathology, hierarchical analysis.**

Manuscript received October 31, 2019; revised February 19, 2020 and March 27, 2020; accepted May 12, 2020. Date of publication May 28, 2020; date of current version February 4, 2021. This work was supported by the Swedish Foundation for Strategic Research under Grant BD150008 and in part by the European Research Council under Grant ERC2015CoG 682810. *(Corresponding author: Håkan Wieslander.)*

Håkan Wieslander and Carolina Wählby are with the Department of Information Technology and SciLifeLab, Uppsala University, 751 05 Uppsala, Sweden (e-mail: h.wieslander@it.uu.se; carolina.wahlby@it.uu.se).

Philip J. Harrison is with the Department of Pharmaceutical Biosciences, Uppsala University, 752 37 Uppsala, Sweden (e-mail: philip.harrison@farmbio.uu.se).

Gabriel Skogberg and Ola Spjuth are with the Department COPD and IPF, Respiratory, Inflammation and Autoimmunity, R&D, AstraZeneca, 431 50 Gothenburg, Sweden (e-mail: gabriel.skogberg@astrazeneca.com; ola.spjuth@farmbio.uu.se).

Sonya Jackson is with the Department of Translational Science and Experimental Medicine, Respiratory, Inflammation and Autoimmunity, R&D, AstraZeneca, 431 50 Gothenburg, Sweden (e-mail: sonya.jackson@astrazeneca.com).

Markus Fridén is with the Department of Drug Metabolism and Pharmacokinetics, Respiratory, Inflammation and Autoimmunity, R&D, AstraZeneca, 431 50 Gothenburg, Sweden, and also with the Translational PKPD Group, Department of Pharmaceutical Biosciences, Uppsala University, 752 37 Uppsala, Sweden (e-mail: markus.friden@astrazeneca.com).

Johan Karlsson is with Data Sciences & Quantitative Biology, Discovery Sciences, R&D, Astra Zeneca, 431 50 Gothenburg, Sweden (e-mail: johan.karlsson1@astrazeneca.com).

Digital Object Identifier 10.1109/JBHI.2020.2996300

I. Introduction

MICROSCOPY imaging is one of the most powerful tools used to investigate complex biomedical processes, and automated analysis methods are capable of measuring a large number of parameters from a broad range of samples in parallel. Rapidly developing high-throughput techniques are generating data at an unprecedented rate thus placing the biomedical sciences on the verge of a digital explosion. Transformative approaches to analyze this massive spatial and temporal multichannel image data are urgently needed, or there is a real risk that the promised, rapid advancement in knowledge will not materialize. In digital pathology, images generated by whole-slide scanning often reach a couple of gigabytes in size and can span ten to hundreds of thousands of pixels in both the x- and y-direction. This full resolution is often needed to acquire accurate stain quantification or patient diagnosis. Furthermore, scientific and/or diagnostic information is often sparse, confined to small regions of interest (ROIs). Specifically, analysis of lung tissue remains a challenging task due to the heterogenous nature of the lung with its myriad of sub-compartments, all with distinct physiological functions and roles in pathophysiology of respiratory disease. Although there have been important advancements made during the past decades in the management of lung disease there is still a large unmet need in major lung diseases such as chronic obstructive pulmonary disease (COPD) and idiopathic pulmonary fibrosis (IPF).

Globally, COPD is the third leading cause of death [1] and a major challenge with COPD is to subtype it into endotypes that share underlying pathological mechanisms. However, two typical manifestations of COPD within different sub-compartments are bronchitis and emphysema which affect the airways and alveolar bed respectively. IPF is another deliberating lung disease with a 5-year survival rate of 20–40% [2]. For IPF, the major affected lung sub-compartment is the alveolar bed.

As of today, image analysis of histological lung samples typically is conducted on either large scanned images with no segmentation into different sub-compartments or directly on

small selected or randomized areas for all of which the analysis fail to be either specific, unbiased or based on the wealth of the data in a whole slide scan.

We make two fundamental observations: (1) Not all data contain valuable information. With datasets outgrowing resources we cannot afford to analyze data that lack scientifically relevant information. (2) Given limited resources, or if real-time decisions are needed, we have to be smart about which subsets of the data we use for detailed, costly analyses. We should therefore focus on processing only the data that is most likely to answer the scientific question under study.

Deep learning has become one of the most competitive and successful machine learning approaches for exploring microscopy images of cells and tissue samples [3]–[5]. Deep learning approaches used for detection of ROIs and semantic segmentation often output scores between 0 and 1 defined via a softmax function [6]. The highest scoring class defines the predicted class. If these scores could be thought of as estimates of prediction confidence, further analysis could be confined to high-confidence regions. However, the softmax function has been shown to produce overconfident predictions with uncalibrated outputs [7], and softmax values should not be thought of as confidence measures. Guo and Pleiss *et al.* [7] proposed a temperature scaling where a constant temperature (T), optimized with respect to a validation set, dampens the softmax output to obtain calibrated confidence. This can however fail to produce class-wise calibration which is important when the predictions are to be used as a confidence measure [8]. An alternative approach guaranteed to obtain valid measures of confidence is to use the statistical learning theory of conformal prediction (CP) proposed by Vovk *et al.* [9], where predictions are hedged: They incorporate a valid indication of their own accuracy and reliability. CP works atop any machine learning algorithm and can be readily applied to deep learning applications at almost no additional cost [10].

In this paper we present an approach for the analysis of large-scale whole-slide lung tissue images aiming to limit costly computations to the parts of the data most likely to answer a given scientific question. Our main contribution is to combine deep learning with CP for tissue sub-region prediction with confidence. Using hierarchical identification of tissue regions and a measure of confidence in sub-region detection, we quantify region-specific drug response. A key objective is to start from low resolution to predict and rank regions that should be investigated at higher resolution, motivated by the fact that analysis at lower resolution involves fewer pixels and is therefore cheaper. Costly quantitative analysis of fluorescent markers is thus limited to very few and small specific regions at the highest resolution.

The work we present here, adds a valuable methodology to address the understanding of both the drug target localization and drug target engagement in specific lung compartments of interest, based on whole lung slide data thus enabling direct insight into the spatial effect of therapeutic compounds and the relationship of this to histological and pathological structures.

II. Problem Description

We present a generalizable method for analyzing large-scale image data applicable to a wide range of problems where precision of region detection is prioritised over recall. This is often the case when evaluating tissue specific drug response, but may not be suitable for eg. detection of malignancies. We evaluate the method by quantifying the distribution of fluorescent signals in well-defined sub-regions of lung tissue.

More specifically, we first use low-resolution images to detect bounding box ROIs of certain parts of the lung tissue (blood vessels and airways). This is done using deep learning, but without CP to maximize detections. Once ROIs are defined, we move to medium resolution to do semantic segmentation of ROIs consisting of background and four other sub-regions; 1. blood vessels 2. alveolar bed 3. epithelial of conducting airways and 4. sub-epithelial layer of conducting airways. All pixels classified into these sub-regions are also given a confidence score which tells us how likely it is that a pixel belongs to the given class compared to the other classes. In the sub-regions a fluorescent signal is quantified at full resolution by summing the intensity values and dividing by the area of the region. Finally, we investigate if more statistically significant results can be obtained by focusing the analysis to well-defined ROI sub-regions defined with high confidence. In other words, we hypothesize that we can provide a better answer to our scientific question if we focus our analysis to the part of the data that is most likely to provide relevant information.

Assumption 1: There is a difference in drug response between cell layers (epithelium, sub-epithelium and alveoli) around airways and blood vessels [11].

Assumption 2: These differences can be measured by defining these cell layers and quantifying fluorescent signals from gene expression in response to drug uptake in fluorescence microscopy images.

Hypothesis: The significance in the quantified difference will be higher if we are more confident in the definition of the cell layers.

III. Related Work

The most common strategy when applying deep learning based methods to large whole-slide images is to divide the image into smaller patches and to make class predictions on these individual patches. Hou *et al.* [12], for example, proposed a patch-based approach where histograms of the patch predictions are used as feature vectors for image-level predictions. In [13] Graham *et al.* proposed a multi-step approach where patch-wise prediction maps are extracted and their subsequent feature vectors are fed through a random forest classifier for lung cancer grading. Although promising, patch-based approaches can become quite computational expensive and time-consuming when applied to images of large tissue slides. Instead, by utilizing the characteristics of convolutions, analysis can be done in one pass of the network: All dense layers are replaced by convolutional layers making the network fully convolutional, referred to as Fully Convolutional Networks (FCNs) [14]. Thus,

the network is no longer restricted by a specific input image size. As the size of the input image is increased the network generates a coarse prediction map as output. One of their major strengths is that these networks can be trained on patch-based annotations and subsequently utilized for semantic segmentation. This was shown for example in [15] for quantifying biomarkers and in [16] for cancer region segmentation. A more efficient approach utilizing FCNs to detect lymph node metastasis in whole slide breast images was proposed by Lin *et al.* [17]. This idea was later extended showing that FCNs remove a lot of the redundant convolutional operations in patch based methods [18]. They introduce anchor layers, which in contrast to standard convolutional layers, can jump and move to a different position when making dense predictions, thus increasing the speed even further. Other learning-based methods for region localization use ground truth annotations consisting of ROIs defined by bounding boxes. With the annotated data limited to patch-based annotations, architectures like Faster RCNN [19] and Yolo [20] are unfortunately unfeasible.

In many cases, definition of ROIs is not sufficient. Within defined ROIs, pixels have to be assigned to object-specific classes or sub-regions by semantic segmentation. The problem of segmentation of large whole-slide tissue images (and many other image analysis applications), is that it requires making accurate local predictions whilst accounting for global context. One of the first applications of deep learning for segmentation in medical images used CNNs with patch-based sliding windows to classify pixels [21]. As deep learning requires larger training datasets than are generally available for biomedical applications, this patch-based representation of the input data can be beneficial. However, there is a trade-off between local and global information when using such patches, whereby small patches sacrifice contextual information over location accuracy and vice versa. Furthermore, the use of a sliding window results in a large amount of parameter redundancy when computing the feature maps of neighbouring pixels. U-Net [22], a popular modified version of the FCN architecture combines global and local information into one network. It uses contracting convolving encoder layers (learning global context) skip-connected to expanding "up-convolving" decoder layers (learning high-resolution location). The U-Net architecture has for instance been utilized in epithelium segmentation in prostate cancer whole slide images [23].

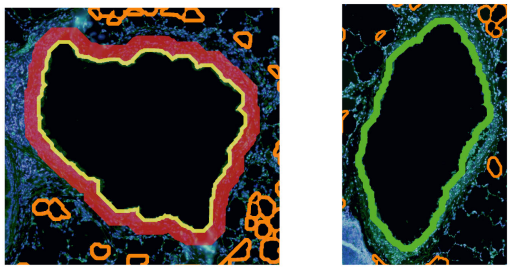
A major bottleneck when applying deep learning methods to cell and tissue images is the lack of labeled data; manual data annotation is time-consuming and requires a high level of expertise, and few alternative approaches exist [24]. However, studies have shown that reusing models pre-trained on different tasks provides a good network initialization [25], [26]. This is known as transfer learning, whereby the transferred parameter values provide good initial values for gradient descent prior to fine-tuning to fit the target data. This idea has been successfully applied using large annotated datasets like ImageNet [27]. Others have shown that pre-training on data from a similar task can be even more beneficial [28].

Confidence estimates in deep learning have mainly focused on Bayesian based methods. Gal and Ghahramani proposed using Monte Carlo dropout [29] for model (epistemic) uncertainty by using active dropout layers at test time to obtain a variance around the predictions. This results in an approximate Bayesian posterior distribution for the predicted probabilities. However, these approximate Bayesian methods [30], being based on rather limited distributional assumptions, are liable to underestimate uncertainty. The Bayesian hypernetworks of Krueger *et al.* [31], which combine Bayesian methods, deep learning and generative modelling, provide one means of overcoming this problem of uncertainty underestimation. Other methods propose letting the model predict the variance [32] as a form of aleatoric (data driven) uncertatinty. Test time augmentations together with Monte Carlo dropout was investigated in [33] for fetal brain segmentation in MRI slices and brain tumor segmentation in MRI volumes.

As an alternative to Bayesian approaches there is a method known as Conformal Prediction (CP). CP was initially devised to work in an on-line transductive setting, such that learning and prediction occur simultaneously. In this sense confidence in a prediction is tailored both to the previously seen objects (whose features and labels are known) and to the features of the new object, whose label is to be predicted [34]. The fully on-line mode of CP can be computationally demanding (with the learning algorithm updated for each new data point). The theory however extends easily to the off-line inductive mode (which we use in this paper). CP has been used in moderately sized problems, e.g. to predict quantitative structure-activity relationships of molecules [35], to assess complication risks following coronary procedures [36] and to detect anomalies in fishing vessel trajectories [37]. It has also been shown to scale up well on a distributed computing implementation to very large datasets, such as the Higgs boson dataset [38], the largest binary classification dataset in the UCI machine learning repository [39].

IV. Data

The dataset used in this work was created to investigate drug distribution in lungs of rats as previously described in [11]. Briefly, rats were treated with different doses of the drug (fluticasone propionate) and with two methods of administration; either inhaled or intravenously. As response to drug uptake, cells produce mRNA from glucocorticoid receptor response genes. Rats were sacrificed followed by tissue fixation and sectioning. Cell nuclei were stained with DAPI and two different fluorescent markers were applied to detect mRNA for the glucocorticoid receptor response genes. Resulting tissue slides were imaged in four fluorescent channels, one for each mRNA detector, one for nuclei (DAPI) and one for auto fluorescence, using a slide scanner. All image analysis, apart from the final quantitative evaluation of drug response, was performed on the image channels showing nuclei and auto fluorescence, in other words representing general tissue morphology. Full resolution images (each around 23000 px $\times$ 35000 px) were sub-sampled to medium resolution (16% of the original size $\approx$ 9200 px $\times$ 14000 px) and low resolution (4% of the original size $\approx$ 4600 px $\times$ 7000 px) for further processing as described below.



(a) (b)

Fig. 1. Example annotations of lung tissue divided into (a) Airway: epithelium (yellow), sub-epithelium (red) and alveoli (orange) and (b) blood vessels (green) and alveoli (orange).

TABLE I
NUMBER OF IMAGES USED FOR TRAINING IN THE DIFFERENT STEPS OF THE PIPELINE

Data set	Number of images		
	Airways	Blood vessels	Other
ROI localization	352	202	495
sub-region segmentation	28	28	-

For model training ROIs were manually extracted and labeled as 'Airway,' 'Blood vessel' or 'Other'. The 'Other' class contains large holes and broken tissue not belonging to either airways or blood vessels. In total, 1159 ROIs from 58 tissue sections (58 different animals) were labeled.

Further annotations were needed for creating a deep learning model for semantic segmentation of sub-regions within the ROIs. These annotations were generated by a semi-automatic method where images of ROIs were first binarized to capture large hole structures. Around each hole (representing an airway or a blood vessel), a sub-region was defined by dilation, and the width of the region was limited by the intensity values in the nuclei and auto-fluorescence channels. This method gave a rough label mask for both the epithelial and sub-epithelial layers around airways and also for the blood vessels. Since we were only interested in finding representative areas of the different regions, we selected, for annotation only one airway and one blood vessel from each tissue section, thus reducing the amount of annotation work required. The generated annotations were subsequently visually inspected and filtered to remove images where the method produced sub-optimal results, resulting in 40 annotated examples per class. Cell layers representing alveoli were annotated by identifying smaller hole regions and dilating a small region around those (Fig. 1).

A test set was defined and removed from the dataset. This test set was selected to include four images from different animals, all treated with the same drug dose. The resulting number of ROIs and annotated sub-regions used for training the different models can be seen in Table I. For ROI localization, 10 images per class were removed for validation and the rest were used for training. For region segmentation, 10 images were removed for validation/calibration.

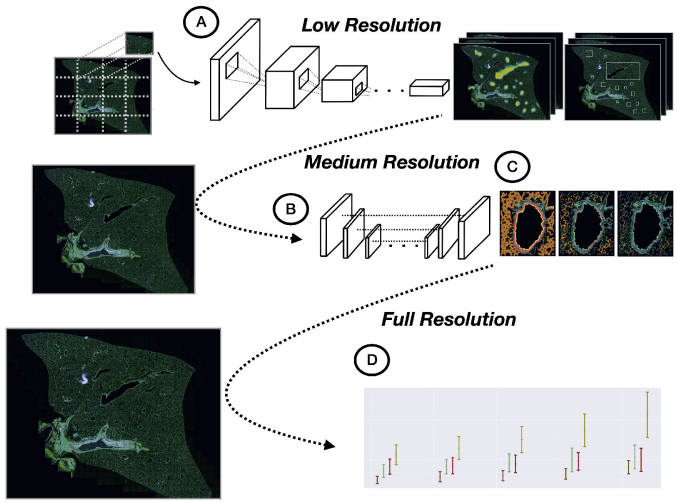


Fig. 2. Illustration of proposed workflow. Regions of interest are located, through Fully Convolutional Networks, in low resolution (a). The located regions are mapped to a higher resolution where region segmentation is performed with a U-net architecture (b). Conformal predictions are applied to the output of U-Net to obtain confidence predictions (c). Lastly, the segmented regions are mapped to the full resolution where drug response quantification is done (d).

V. METHODS

For the proposed pipeline our focus lies on identifying the most informative regions to answer the scientific question at hand, namely if there are differences in drug response in different cell layers. ROIs are located in low resolution, where finer details are missing, but larger structures are still visible. A finer-scale semantic segmentation of ROIs is then produced at a medium resolution, and combined with a confidence measure based on conformal predictions. In the final full resolution step quantification of stains is performed (Fig. 2).

A. ROI Localization at Low Resolution

To make the analysis pipeline as general as possible we aimed to make the number of tunable hyper-parameters small. The network had to be able to learn from a small amount of training data thus making large model architectures unfeasible. We therefore decided to use ResNet 18 [40], which has, as its name suggests, only 18 layers. We initialized the network with pre-trained weights from ImageNet [41] and augmented the training data by mirroring and 90 degrees rotations to create eight times as many images. We used the Adam [42] optimizer with the default learning rate of 0.001 and early stopping (with a maximum of 100 epochs) based on the loss on a validation set. We thus used the validation set only for determining how long the network should be trained. After training, the weights of the last fully connected layer were transferred and replaced by a 1×1 convolutional layer making it fully convolutional [14] (Fig. 3).

At inference the large tissue image is unfortunately too large to fit on a GPU and to be processed by the network. Thus, smaller tiles of the image needed to be extracted. Since the original network was trained with a predefined input size the

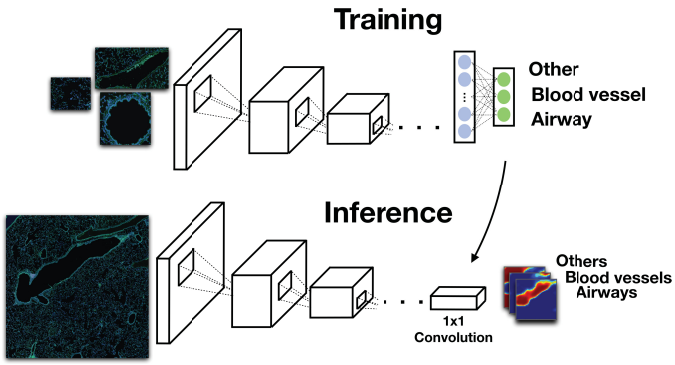


Fig. 3. Illustration of the network for ROI localization at training and inference time. The arrow from Training to Inference illustrates how the weights of the last fully connected layer are transferred and replaced by a 1 × 1 convolutional layer making the model fully convolutional.

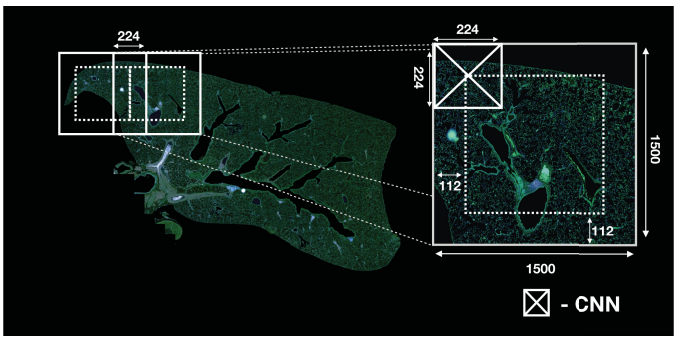


Fig. 4. Overlapping cutouts were applied to account for loss in spatial output size as the full tissue sample can not fit on a GPU during inference.

network will output point predictions for this same fixed size. Most pretrained ImageNet models have a spatial size of 224 px × 224 px, which results in, at inference, a 224 px × 224 px window being convolved over the larger input (Fig. 4). To account for this, the tiles had to overlap. With a 224 px × 224 px model this overlap had to be 224 px. To produce a result for the entire image, also the edges should be padded with 112 px. In this work, this step was not needed since there is a large empty space between the actual tissue and the edges of the image, and no information was lost. Similarly, stitching and rebuilding the full image required re-sizing the output to a width and height 224 px smaller than the input and then padding 112 pixels at each side.

Due to the varying size of the training ROIs, the training images had to be sub-sampled to 224 px × 224 px to fit the size of the network. An average sub-sampling factor $(\Delta x_{od},\ \Delta y_{od})$ was calculated so that it could be used for sub-sampling the test set. Inference was done by first down-sampling the test images to a size $ImageSize_x \cdot \Delta x_{od},\ ImageSize_y \cdot \Delta y_{od}$. The images were then tiled into 1500 px × 1500 px patches (size chosen based on hardware limitations) and passed through the network. The outputs were finally re-sized and stitched together.

With the obtained ROI prediction maps a couple of post processing steps were applied. First the prediction maps were thresholded by 0.5 and ROI candidates were proposed via connected components. Two filtering methods were applied to the proposed ROIs:

1) Remove ROIs with an area smaller than 1000 pixels
2) Remove ROIs less than 100 pixels from the edge of the tissue sample

The edge of the tissue was found through active contours based on the implementation from [43] available at [44].

Lastly, bounding boxes were defined for the final ROIs and boxes with intersection over union larger than 0.5 were removed to avoid multiple copies of the same object being forwarded to the next step.

B. Semantic Sub-Region Segmentation of ROIs at Medium Resolution

Depending on the question at hand, ROIs may have to be further divided into sub-regions at a higher resolution. This is especially the case when structures of interest are comparably small, such as cell clusters or epithelial layers. In the presented case, we had to move to medium resolution to do semantic segmentation of ROIs into sub-regions consisting of background and four types of cell layers. To simplify the task, we used knowledge of ROI class (airway or blood vessel), obtained from the previous step to train two separate networks for sub-region segmentation. Some initial experiments showed that a single network with more classes made more errors than two separate networks focusing only on a specific object.

The ROI segmentation was done with a U-Net architecture with an encoder initialized with weights pre-trained on ImageNet [27]. The training was based on weighted cross entropy loss with the weight of the background class set to a third of the other classes. This weight was set due to the fact that the background class was not fully annotated and includes regions from all different classes. A lower weight on the background class encourages the network to get the actual annotated labels correct.

The training data was sub sampled to twice the size of the test images for the ROI localization step to simulate a medium resolution $(ImageSize_x \cdot 2 \cdot \Delta x_{od},\ ImageSize_y \cdot 2 \cdot \Delta y_{od})$. The U-Net architecture is fully convolutional and thus not restricted by any specific input size. However, to enable batch training, the data has to have the same size. We therefore ensured all training ROIs were larger or equal to 512 px × 512 px (based on the maximum size not exceeding hardware limitation with batch training). Images smaller than this size were padded with zeros. The data was then loaded via random crops of size 512 × 512 to make batches of similar sizes. Training set augmentations were performed with shifting, scaling, rotating and applying random re-sized crops. The scaling and shifting factors was limited to 10% of the image size [45]. The two different networks (one for airways and one for blood vessels) were trained for a maximum of 400 epochs (again early stopping was performed using the validation set).

At inference the images were passed through the network one by one and thus do not need to meet the same size criteria. Instead, since the network has five pooling layers (each halving

the size of its input) and skip connections from the encoder to the decoder, the input sizes are required to be divisible by $2^5 = 32$. This was ensured by padding the test and calibration images to the necessary size. As output, all pixels are given a Softmax value, and the pixels are assigned to the class for which it has the highest Softmax value.

C. Conformal Prediction for Pixel Classification With Confidence

As described in the introduction, softmax values do not provide confidence values for a pixel belonging to the given class compared to the other classes. Therefore, we apply CP to achieve a confidence measure that can guarantee the prediction sets to contain the true label of the object with a probability equal to a user-defined significance level ϵ, under the weak assumption of data exchangeability. We achieve this by comparing new objects to previous examples of known outcome through a nonconformity function, indicating the "strangeness" of the new object. We expand the binary classification setting, where classes 0 and 1 translate into four possible prediction sets: {0}, {1}, {0,1} and Ø (the empty set) to multi class classification with three and four classes resulting in additional prediction sets. For instance three classes has eight prediction sets {0}, {1}, {2}, {0,1}, {0,2}, {1,2},{0,1,2} and Ø (the empty set). With guarantees of validity, the efficiency of the predictor remains to be evaluated [46]. We used the ratio of single-label prediction sets as our efficiency metric, yielding a value between 0 and 1 whereby lower values are preferable (more efficient). The efficiency is thus calculated as the number of predictions with a multi-labeled prediction set over the total number of observations.

We used the following nonconformity measure (here described for a single prediction)

$$\alpha = A(\sigma_t) = \frac{max_{j=0\ldots c, j\neq t}(\sigma_j)}{\sigma_t} \tag{1}$$

where σ is the output from softmax, t the true class and c the number of classes. In other words the maximum output of all but the true class divided by the output for the true class, where $\alpha \in [0, \infty]$ [10].

Conformal prediction was originally developed in an online transductive setting for which the model needs to be updated or re-trained for each new example which is computationally demanding. Inductive conformal prediction (ICP) operates under the same assumptions and provides the same guarantees as CP but with reduced computational load [10], [47]. In ICP, the training dataset is first split into a calibration set C and a proper training set P. The underlying machine learning model (such as SVM, Random Forest, or DNN) is then trained only once on P, yielding a single model that can be used to calculate the nonconformity scores. The model is then applied on C to calculate nonconformity scores for all calibration instances. In this work we use Mondrian conformal predictors, where each label category is treated individually with respect to the comparison of nonconformity scores, yielding one set of nonconformity scores (α_l) per label l. The Mondrian approach has attractive properties

Algorithm 1: Calculate Alphas for Calibration Set.

Initialize:
$\alpha_{cal} \in \mathbb{R}^{c \times m}$
$c =$ number of classes
$m =$ number of sampled pixels per class
for Images **in** Calibration set **do**
  random sample m pixels per class t
  for pixel x **in** sampled pixels **do**
    Calculate nonconformity measure:
    $A(\sigma_t^x) = \frac{max_{j=0\ldots c, j\neq t}(\sigma_j^x)}{\sigma_t^x}$
    $\alpha_{cal}(t, x)$ append $A(\sigma_t^x)$
  end for
end for

Algorithm 2: Calculate p-values for Test Image.

for Pixel x **in** test image **do**
  for each possible classification u, $u = 1..c$ **do**
    $\alpha_{x,u} = A(\sigma_u^x) = \frac{max_{j\ldots c, j\neq u}(\sigma_j^x)}{\sigma_u^x}$
    $p_{x,u} = \frac{\#\{n=1\ldots m : \alpha_{cal}(u,n) \geq \alpha_{x,u}\}}{m}$
  end for
end for

when e.g. data is unbalanced [48]. When making predictions on new objects using ICP, the objects nonconformity score is first calculated using the trained underlying model. P-values $p(y_j)$ are then calculated for each class j as the fraction of α - values in the calibration set larger than the newly calculated value divided by the total number of α - values in the calibration set (m):

$$p(y_j) = \frac{\#\{n = 1\ldots m : \alpha_j \geq \alpha_n\}}{m} \tag{2}$$

For more details on conformal prediction, see [34].

To examine the calibration of the output from CP a separate test was set up. All the data (except the test data) was randomly shuffled and divided into three parts: calibration/validation; proper training; and calibration testing. To examine the optimal size for the calibration set, different sizes were explored. For all experiments the calibration test size was kept fixed at 15 images. The evaluation was done using five fold cross validation and the final results were obtained by averaging across the folds. The final model was then trained on all data with the calibration set size chosen as the lowest amount showing sufficient resolution (evaluated by comparing the calibration plots) in the calibration to ensure valid results. With the softmax output for the images in the calibration set, alpha values were calculated in a Mondrian way as follows:

The images in the test set were first passed through the sub-region segmentation network to obtain the softmax output. The p-values were then calculated for each pixel position as follows:

The predicted output being the class with the largest p-value. The confidence in the prediction is defined as the largest p-value minus the second largest p-value. Confidence is a measure of

TABLE II
QUANTITATIVE RESULTS FOR THE RESNET 18 MODEL TRAINED FOR CLASSIFYING AIRWAYS, BLOOD VESSELS AND OTHER, EVALUATED ON THE VALIDATION SET

	Precision	Recall	F1-Score
Validation set:	0.89	0.87	0.87

how likely the prediction of the current class is compared to the other classes. A large measure of confidence means that the predicted class conforms well to the calibration set, whilst the other classes conform poorly. Lower confidence is obtained when multiple classes conform similarly well to the calibration set or if all classes conform poorly.

D. Quantification of Drug Response at Full Resolution

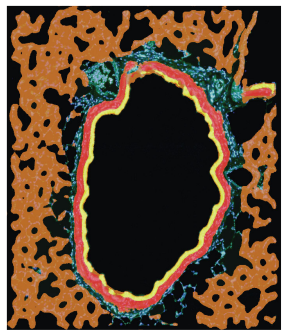
High image resolution is often required to get accurate quantification of weak or small fluorescent signals only visible at full resolution. We therefore mapped the sub-regions defined in the previous step to the highest image resolution available. Different binary mappings were created for a set of pre-defined confidence levels. Per-region drug response was thereafter approximated as the sum of pixel intensities belonging to a given region divided by the area of the region. The measurements were made in the image channels describing mRNA content.

VI. RESULTS

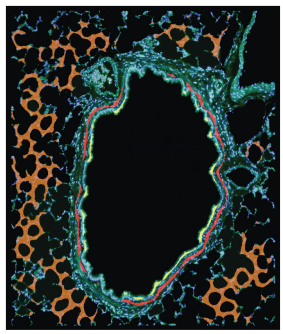
When training the network for ROI localization, it was observed that the network converged quickly, hence limiting the training to 100 epochs was sufficient. The average down sampling factor of the training data was found to be approximately 20% of the original size in each direction ($\Delta x_{od} \approx 0.197, \Delta y_{od} \approx 0.206$). In other words, the number of pixels we process during ROI localization is only 4% of the original amount. Quantitative results from the ROI localization step can be seen in Table II.

In the semantic sub-region segmentation step, the networks needed longer training to converge, hence a limit of 400 epochs was set to reach a good convergence. An example of the output from sub-region prediction can be seen in Fig. 5(a).

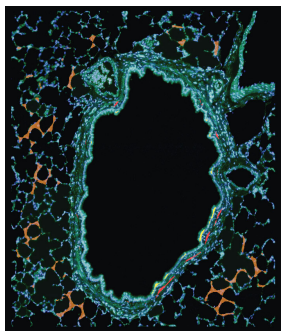
In CP confidence level estimation, varying the calibration set size was shown to give similar results with the only difference being that using more than five images showed slightly more stable results per fold, but similar results on average. We therefore chose to use ten images in the calibration set to not exclude too much data but to still reach stability in our evaluation. The resulting calibration plot comparing the softmax output, temperature scaling, and the CP output can be seen in Fig. 6. This figure shows that using the softmax would result in substantial underestimates of the actual error, whereas CP produces valid results. The actual error is measured as the ratio of incorrect predictions (at a given threshold) when measured against the labeled examples. For each image 1000 pixels were sampled for inclusion in the calibration set. Taking more or fewer pixels changes the resolution in the calibration set but has little effect on the results. Evaluation of the non-conformity



(a) Output prediction with no threshold for confidence, predicted is the largest p-value ($\epsilon = 0$).



(b) Output prediction for confidence level 0.8 ($\epsilon = 0.2$).



(c) Output prediction for confidence level 0.95 ($\epsilon = 0.05$).

Fig. 5. Semantic segmentation results of tissue surrounding an airway, with epithelium (yellow), sub-epithelium (red), alveoli (orange) and background (no color overlay) shown on top of the raw image. The regions shrink (from a to c) as the threshold for the confidence level in pixel classification is increased (i.e decreasing ϵ).

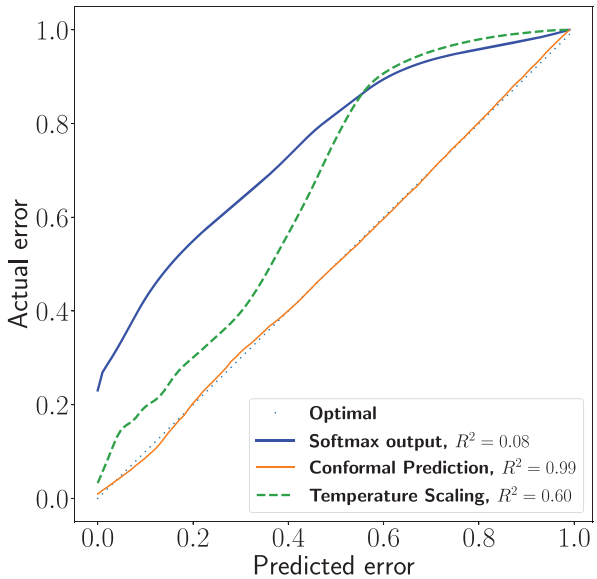


Fig. 6. Calibration plot for predicted and observed error comparing direct output from the softmax, temperature scaling and the CP output. The data was split so that 13 images were used for training, ten for calibration and 15 for calibration testing. The method is evaluated with five-fold cross validation and presented is the average of all folds.

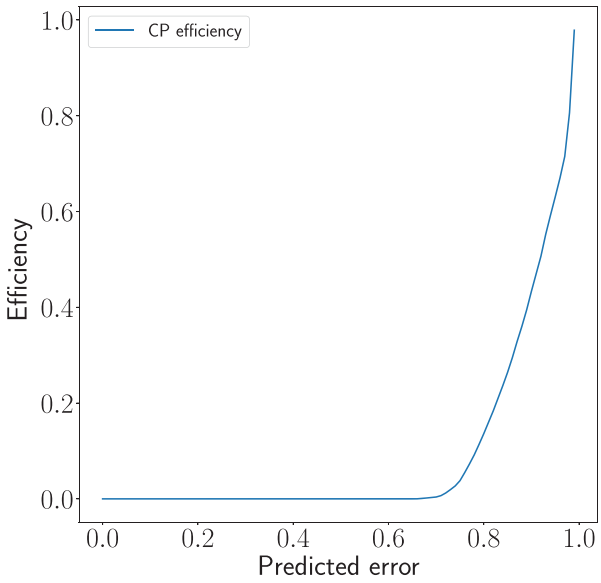


Fig. 7. Efficiency measured through single valued predictions (lower is better) at different ϵ. The efficiency indicates how well the non-conformity measure works for capturing the difference between new examples and seen examples (i.e calibration set).

measure at different ϵ is presented in Fig. 7. A lower value means a more efficient measure. This is achieved when the non-conformity measure successfully captures the difference between new examples and known examples, i.e, making single valued predictions.

Using the final pixel confidence level output from CP, low-confidence pixels can be excluded from each sub-region class based on thresholding at the pixel class level. An example of the resulting reduced regions at two different confidence thresholds can be seen in Fig. 5(b and c).

For final quantification of drug response we show the effect of focusing measurements to regions with low confidence $(0.1 - 0.2)$ and gradually increasing the confidence interval up to $(0.8 - 1)$. For each confidence interval, drug response was quantified in the corresponding tissues sub regions, and the results from quantification of drug response per image sub-region class are presented in Fig. 8a. The image data in the presented result was not included during testing or validation of ROI detection or sub-region segmentation, and represents tissue slides from four animals exposed to the same drug concentration by inhalation. Values shown are means and standard deviations across all regions within each sub-region class. The plot indicates that the drug response is higher in the epithelium of the airways as compared to the airway sub-epithelium and blood vessel. The lowest level is observed in the alveoli. As the analysis is confined to regions with higher confidence, larger differences, with higher statistical significance are observed. To evaluate the pixel classification performance of the models, 4 airways and 4 blood vessels where selected for more accurate manual annotations (focusing on the epithelium, sub-epithelium and blood vessel areas). The resulting precision and recall are presented in Fig. 8b.

VII. Discussion

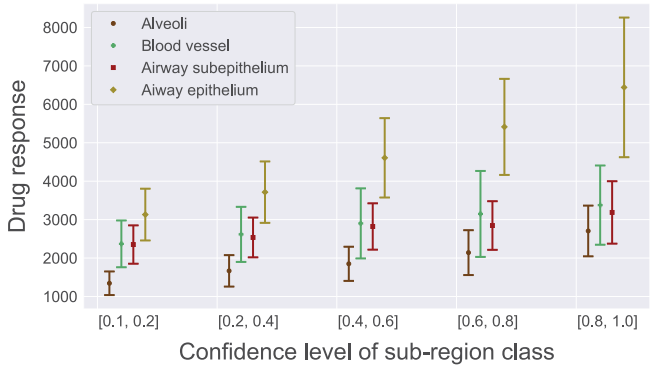
The pipeline we present here is created in such a way that for each part, the data could potentially be replaced with a different dataset to explore alternative problems. With generalizability, however, one sacrifices specificity, hence depending on the problem, more tailored solution may be constructed to outperform our proposed pipeline. We show that FCNs can detect ROIs in low resolution reducing the amount of data that needs to be analyzed (as compared to analyzing all data at full resolution). An interesting approach would be to apply the hierarchical approach with ROI detection already during data acquisition at the microscope, and thereby limiting the collection of irrelevant data. Furthermore, we show that U-Nets can separate finer regions even with approximate ground truth labels during training.

With the effectiveness of transfer learning our models were able to learn from a rather limited training set (Table II, Fig. 8b). The results (Fig 8b) show an increase in precision at a higher confidence threshold, but with the sacrifice of recall. Thus, a high confidence threshold results in a final readout with higher detection precision, but lower recall, focusing the final measurements on regions with more accurate predictions. The models were pre-trained on ImageNet which consists of natural images. Nevertheless, our baseline approach was able to distinguish between the different tissue regions with a relatively low amount of training data. Improved results could also be obtained with networks pretrained on a (large) more domain specific dataset.

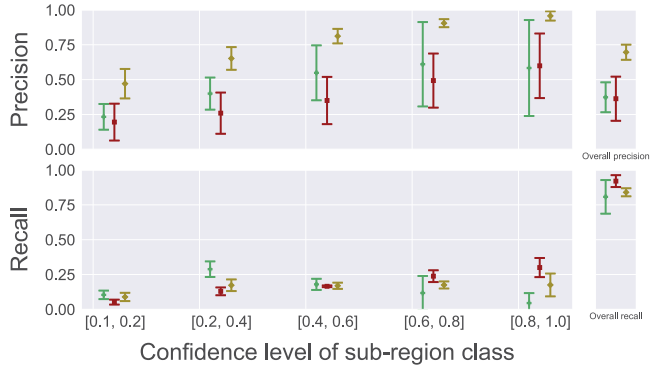
In deep learning, and machine learning in general, there are often considerable uncertainties in many of the predictions. As mentioned earlier the predicted probabilities from the softmax output are also not well calibrated. This was shown clearly in Fig. 6, where temperature scaling gave better calibrations, but still far from the nearly perfectly calibrated results given by CP. Furthermore, these probabilities are simply point estimates without any information on their variability. In general, for medical image data there also often exists a high level of uncertainty in the annotated labels [49]. Indeed in our data the annotations were done quite roughly and in some cases the deep learning segmentation results were more accurate than the annotations. Awareness of these various forms of uncertainty is invaluable for accurate conclusions, and deep learning methods that assign confidence to predictions may also be better received by clinicians.

Our results show that the higher confidence regions mainly enclose the most central part of the segmented regions, whereas the transition between labels in the output often results in the lowest confidence. With the rough annotations, for training, used in this work, these pixels were the hardest to learn confidently. Confidence based prediction methods like CP are therefore a useful tool in applications where the user can adjust the amount of errors he/she is willing to make. Allowing more errors equates to including more uncertain regions and in the case of sparse labels can also result in larger regions. Allowing less error results in smaller regions but with more precision.

Furthermore, our results show a larger separation of drug response mainly between the sub-epithelial and epithelial regions



(a) Drug response, quantified as per sub-region average intensity of mRNA stain, measured within sub-regions at five different pixel classification confidence intervals. The measurements represent mean and standard deviation across all regions from four different animals, all treated with the same drug dose administered by inhalation. For airways N = 47 and blood vessels N = 21.



(b) Precision and recall measured over different pixel classification confidence intervals. The measurements represent mean and standard deviation across all regions from four different animals. The overall score is measured with the prediction being the class with the largest p-value. Note that the overall score leads to significantly higher recall (while reduced precision) as all pixels are included. N = 4 for both airways and blood vessels (same legend as in a).

Fig. 8. Measurements of drug response, precision and recall for different pixel classification confidence intervals.

in airways when measured at a higher confidence level. Since the epithelial region of the airway is narrower, more precision is required in the segmentation results for accurate predictions. For this type of region, high confidence predictions help exclude noise. In contrast, the sub-epithelial cell layer is generally wider and requires less precision for high confident sub-region classification. This can be seen in Fig. 8a where the measured drug response for sub-epithelial regions does not change as much for the higher confidence regions. For the epithelial region the concentration steadily increases with higher confidence, as initially hypothesized.

Perhaps the optimal means of accounting for uncertainty will come with the fusion of deep learning and Bayesian modelling, permitting the inclusion of parameter, model and observational uncertainty in a natural probabilistic manner. However, due to their high computational cost and the need to specify prior distributions on all the network weights, such fully Bayesian approaches are currently unfeasible and alternative solutions are required. One alternative and less computationally costly approach is to use CP, as we have done here. CP does not make any distributional assumptions, circumvents the need to specify priors on the parameters, and interestingly can provide stronger guarantees of validity than Bayesian methods, even when based on the true probability distribution of the data [34]. Due to the limited amount of images available for training in the presented experiment, we derive our calibration set from a number of randomly sampled points in each image. This strategy may have an effect on the exchangability assumption of the data points. Further studies of the lowest number of images and the highest number of pixels that is required (with unlimited data one could sample a single point per class and image) are needed. Since temperature scaling only dampens the softmax output it is difficult to find an optimal temperature that gives calibrated outputs for all classes. This is, however, a simpler method and might work better for simpler problems, or when the ground truth annotations are more complete. CP is a more powerful (but more demanding) method where the prediction distribution is utilized in a more efficient way. We here conclude that our empirical studies show that for our methodology the predictions are valid (Fig. 6) with an efficient non-conformity measure (Fig. 7) and that using CP helps to calibrate the softmax output, giving valid prediction regions.

References

[1] S. A. Quaderi and J. R. Hurst, "The unmet global burden of COPD," *Global Health, Epidemiol. Genomics*, vol. 3, p. e4, 2018.
[2] D. S. Kim, H. R. Collard, and T. E. King Jr., "Classification and natural history of the idiopathic interstitial pneumonias," *Proc. Amer. Thoracic Soc.*, vol. 3, no. 4, pp. 285–292, 2006.
[3] A. Gupta *et al.*, "Deep learning in image cytometry: A review," *Cytometry Part A*, vol. 95, no. 4, pp. 366–380, 2019.
[4] F. Xing, Y. Xie, H. Su, F. Liu, and L. Yang, "Deep learning in microscopy image analysis: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4550–4568, Oct. 2018.
[5] E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, and D. Van Valen, "Deep learning for cellular image analysis," *Nature Methods*, vol. 16, pp. 1233–1246, 2019.
[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, pp. 184–187.
[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, vol. 70, pp. 1321–1330.
[8] M. Kull, M. P. Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach, "Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 12 295–12 305.
[9] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, 1st ed. Berlin, Germany: Springer, 2010.
[10] H. Papadopoulos, "Inductive conformal prediction: Theory and application to neural networks," in *Tools in Artificial Intelligence*. Rijeka, Croatia: IntechOpen, 2008.
[11] M. Friden *et al.*, "Understanding and quantifying the spatial distribution of inhaled drugs and their effects," in *Proc. Respiratory Drug Del.*, 2018, vol. 1, pp. 45–50.
[12] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2016, pp. 2424–2433.
[13] S. Graham, M. Shaban, T. Qaiser, N. Alemi Koohbanani, S. A. Khurram, and N. Rajpoot, "Classification of lung cancer histology images using patch-level summary statistics," 2018. [Online]. Available: https://doi.org/10.1117/12.2293855

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3431–3440.
[15] F. Sheikhzadeh, R. K. Ward, D. van Niekerk, and M. Guillaud, "Automatic labeling of molecular biomarkers of immunohistochemistry images using fully convolutional networks," *PLOS ONE*, vol. 13, no. 1, pp. 1–18, Jan. 2018.
[16] B. Peng, L. Chen, M. Shang, and J. Xu, "Fully convolutional neural networks for tissue histopathology image classification and segmentation," in *Proc. 25th IEEE Int. Conf. Image Process.*, Oct. 2018, pp. 1403–1407.
[17] H. Lin, H. Chen, Q. Dou, L. Wang, J. Qin, and P. Heng, "Scannet: A fast and dense scanning framework for metastastic breast cancer detection from whole-slide image," in *Proc. IEEE Winter Conf. Appl. Comput. Vision*, Mar. 2018, pp. 539–546.
[18] H. Lin, H. Chen, S. Graham, Q. Dou, N. Rajpoot, and P. Heng, "Fast scannet: Fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis detection," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1948–1958, Aug. 2019.
[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2016, pp. 779–788.
[21] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2012, pp. 2843–2851.
[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.
[23] W. Bulten *et al.*, "Epithelium segmentation using deep learning in h&e-stained prostate specimens with immunohistochemistry as reference standard," *Scientific Rep.*, vol. 9, no. 1, 2019, Art. no. 864.
[24] S. K. Sadanandan, P. Ranefall, S. Le Guyader, and C. Wählby, "Automated training of deep convolutional neural networks for cell segmentation," *Scientific Rep.*, vol. 7, no. 1, 2017, Art. no. 7860.
[25] A. Kensert, P. J. Harrison, and O. Spjuth, "Transfer learning with deep convolutional neural networks for classifying cellular morphological changes," *SLAS DISCOVERY: Advancing Life Sci. R&D*, vol. 24, no. 4, pp. 466–475, 2019,
[26] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, Jun. 2014, pp. 512–519.
[27] V. Iglovikov and A. Shvets, "Ternausnet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation," 2018, *arXiv:1801.05746*.
[28] R. K. Samala, H. Chan, L. Hadjiiski, M. A. Helvie, C. D. Richter, and K. H. Cha, "Breast cancer diagnosis in digital breast tomosynthesis: Effects of training sample size on multi-stage transfer learning using deep neural nets," *IEEE Trans. Med. Imag.*, vol. 38, no. 3, pp. 686–696, Mar. 2019.
[29] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.
[30] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Statistical Assoc.*, vol. 112, no. 518, pp. 859–877, 2017.
[31] D. Krueger, C.-W. Huang, R. Islam, R. Turner, A. Lacoste, and A. Courville, "Bayesian Hypernetworks," Oct. 2017, *arXiv:1710.04759*.
[32] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Advances Neural Inf. Process. Syst. 30*, I. Guyon *et al.*, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2017, pp. 5574–5584.
[33] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, 2019.
[34] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Berlin, Germany: Springer-Verlag, 2005.
[35] U. Norinder, L. Carlsson, S. Boyer, and M. Eklund, "Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination," *J. Chem. Inf. Model.*, vol. 54, no. 6, pp. 1596–1603, 2014.
[36] V. Balasubramanian, R. Gouripeddi, S. Panchanathan, J. Vermillion, A. Bhaskaran, and R. Siegel, "Support vector machine based conformal predictors for risk of complications following a coronary drug eluting stent procedure," in *Proc. 36th Annu. Comput. Cardiol. Conf.*, Sep. 2009, pp. 5–8.
[37] J. Smith, I. Nouretdinov, R. Craddock, C. Offer, and A. Gammerman, "Conformal anomaly detection of trajectories with a multi-class hierarchy," in *Proc. Int. Symp. Statistical Learn. Data Sci.*, 2015, pp. 281–290.
[38] M. Capuccini, L. Carlsson, U. Norinder, and O. Spjuth, "Conformal prediction in spark: Large-scale machine learning with confidence," in *Proc. IEEE/ACM 2nd Int. Symp. Big Data Comput.*, Dec. 2015, pp. 61–67.
[39] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.
[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[43] P. Mrquez-Neila, L. Baumela, and L. Alvarez, "A morphological approach to curvature-based evolution of curves and surfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 2–17, Jan. 2014.
[44] 2018. [Online]. Available: https://github.com/pmneila/morphsnakes
[45] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125
[46] V. Vovk, V. Fedorova, I. Nouretdinov, and A. Gammerman, "Criteria of efficiency for conformal prediction," in *Proc. Symp. Conf. Probabilistic Prediction Appl.*, 2016, pp. 23–39.
[47] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman, "Inductive confidence machines for regression," in *Proc. Eur. Conf. Mach. Learn.*, 2002, pp. 345–356.
[48] U. Norinder and S. Boyer, "Binary classification of imbalanced datasets using conformal prediction," *J. Mol. Graph. Model.*, vol. 72, pp. 256–265, 2017.
[49] S. G. Armato *et al.*, "Assessment of radiologist performance in the detection of lung nodules," *Academic Radiol.*, vol. 16, no. 1, pp. 28–38, 2009.