HHS Public Access

Author manuscript

*Med Image Anal*. Author manuscript; available in PMC 2022 January 01.

Published in final edited form as:

*Med Image Anal*. 2021 January ; 67: 101814. doi:10.1016/j.media.2020.101814.

# Weakly Supervised Instance Learning for Thyroid Malignancy Prediction from Whole Slide Cytopathology Images

**David Dov**[a,*], **Shahar Z. Kovalsky**[b], **Serge Assaad**[a], **Jonathan Cohen**[c], **Danielle Elliott Range**[d], **Avani A. Pendse**[d], **Ricardo Henao**[a], **Lawrence Carin**[a]

[a]Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA

[b]Department of Mathematics, Duke University, Durham, NC 27708, USA

[c]Department of Surgery, Duke University Medical Center, Durham, NC 27710, USA

[d]Department of Pathology, Duke University Medical Center, Durham, NC 27710, USA

## Abstract

We consider machine-learning-based thyroid-malignancy prediction from cytopathology whole-slide images (WSI). Multiple instance learning (MIL) approaches, typically used for the analysis of WSIs, divide the image (bag) into patches (instances), which are used to predict a single bag-level label. These approaches perform poorly in cytopathology slides due to a unique bag structure: sparsely located informative instances with varying characteristics of abnormality. We address these challenges by considering multiple types of labels: bag-level malignancy and ordered diagnostic scores, as well as instance-level informativeness and abnormality labels. We study their contribution beyond the MIL setting by proposing a maximum likelihood estimation (MLE) framework, from which we derive a two-stage deep-learning-based algorithm. The algorithm identifies informative instances and assigns them local malignancy scores that are incorporated into a global malignancy prediction. We derive a lower bound of the MLE, leading to an improved training strategy based on weak supervision, that we motivate through statistical analysis. The lower bound further allows us to extend the proposed algorithm to simultaneously predict multiple bag and instance-level labels from a single output of a neural network. Experimental results demonstrate that the proposed algorithm provides competitive performance

[*]Corresponding author: david.dov@duke.edu (David Dov).

Credit Author Statement

David Dov: Conceptualization, Investigation, Methodology, Software, Writing
Shahar Z. Kovalsky: Conceptualization, Visualization, Data Curation, Writing
Serge Assaad: Methodology, Writing
Jonathan Cohen: Conceptualization, Data Curation, Project administration
Danielle Elliott Range: Conceptualization, Data Curation
Avani A. Pendse: Data Curation
Ricardo Henao: Conceptualization, Resources
Lawrence Carin: Supervision, Funding acquisition, Resources

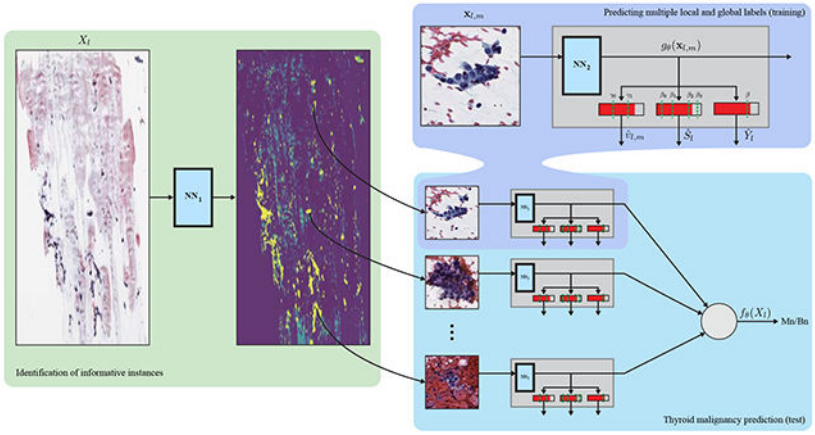


Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

compared to several competing methods, achieves (expert) human-level performance, and allows augmentation of human decisions.

## Graphical Abstract




### Keywords

medical image analysis; multiple instance learning; AI; deep learning; healthcare; pathology; human level; Thyroid

## 1. Introduction

The prevalence of thyroid cancer is increasing worldwide (Aschebrook-Kilfoy et al., 2013). The most important test in the *preoperative* diagnosis of thyroid malignancy is the analysis of a fine needle aspiration biopsy (FNAB), which is stained and smeared onto a glass slide. The FNAB sample is examined under an optical microscope by a cytopathologist, who determines the risk of malignancy according to various features of follicular (thyroid) cells, such as their size, color and the architecture of cell groups. The diagnosis of FNAB, however, involves substantial clinical uncertainty and often results in unnecessary surgery.

We consider the prediction of thyroid malignancy from FNAB, for which we have established in Dov et al. (2019); Elliott Range et al. (2020) a dataset of 908 samples. Each sample comprises a whole slide image (WSI) scanned at a typical resolution of ~ 40,000×25,000 pixels, as well as the postoperative histopathology diagnosis, that is considered the ground truth in this study. The goal in this paper is to predict the ground truth malignancy label from the WSIs. Each sample also includes the diagnostic score assigned to the slide by a cytopathologist according to the Bethesda System (TBS) (Cibas and Ali, 2009), which is the universally accepted reporting system for thyroid FNAB (there are six TBS categories). TBS 2 indicates a benign slide, TBS 3, 4 and 5 reflect inconclusive findings with an increased risk of malignancy, and TBS 6 indicates malignancy. TBS 1 is assigned to inadequately prepared slides and is out of the scope of this work. Further, we consider a set of 4494 manually annotated *local* labels of informative image regions

containing follicular groups. The local labels indicate three categories of abnormality: “0” - normal, “1” - atypical, and “2” malignant.

Machine learning, and in particular deep neural networks, have become prevalent in the analysis of WSIs (Ozolek et al., 2014; Litjens et al., 2016; Kraus et al., 2016; Sirinukunwattana et al., 2016; Djuric et al., 2017; Ilse et al., 2018; Zhang et al., 2019; Campanella et al., 2019; Glass et al., 2020a,b). Due to the large resolution of WSIs, gigabytes in size, each image is typically split into a set (bag) of small regions (instances) that are processed individually into local estimates, then aggregated into a global image-level prediction. This approach, often referred to as multiple instance learning (MIL) (Quellec et al., 2017), addresses memory-capacity limitations of existing graphical processor unit (GPU) computing platforms. Widely used MIL approaches include Zhang et al. (2006) and Kraus et al. (2016), which propose to aggregate local predictions via *noisy-or* or *noisy-and* pooling functions, respectively. In Ilse et al. (2018) a weighted combination of local decisions is proposed, incorporating an attention mechanism to form a global decision.

The vast majority of previous studies consider the analysis of *histopathology* biopsies, which comprise whole tissues covering large regions of the WSI. In contrast, FNABs *(cytopathology* biopsies), as we consider in this paper, contain separate, sparsely located groups of follicular cells, which are informative for diagnosis. The diagnosis of the FNABs, performed by a trained (cyto-)pathologist, includes the identification of follicular groups followed by evaluation of their characteristics. A WSI containing even as few as six follicular groups with a size of tens of pixels, which corresponds to less than 0.01% of the area of the slide, is considered sufficient for diagnosis. FNABs are considered significantly more challenging for diagnosis by pathologists due to their sparsity, and since in many cases, the characteristics of individual follicular groups are subject to subjective interpretation. An example of a large image region of 10000×5000 pixels containing merely a single follicular group, as well as examples of follicular groups with different abnormality levels, are presented in Fig. 1. Due to these challenges, the automated analysis of FNAB is addressed in the literature in a limited scale and scope. Specifically for thyroid FNAB, Daskalakis et al. (2008); Varlatzidou et al. (2011); Gopinath and Shanthi (2013); Kim et al. (2016); Gilshtein et al. (2017); Savala et al. (2018); Sanyal et al. (2018) consider manually selected individual follicular cells in extreme magnification or a small number of “zoomed-in” regions. However, these studies do not address the problem of intervention-free malignancy prediction from cytopathology WSIs.

The paper Cheplygina et al. (2019) surveyed MIL, semi- and weakly-supervised learning approaches. These scenarios consider classification tasks with different assumptions on the availability of training labels: in MIL, only global labels are available at the bag (WSI) level, while in semi/weakly supervised setting local labels at the instance (image region) are only partially available or are noisy(Zhou, 2018). Cheplygina et al. (2019) pointed out three gaps in the existing literature of medical image analysis associated with these scenarios. In the following, we address these gaps in the context of thyroid malignancy prediction. First, Cheplygina et al. (2019) claim that MIL, semi- and weakly-supervised learning are typically studied as separate problems, despite the close relation between them. Here, we investigate how only a few local, instance-level, labels can improve prediction beyond the classical MIL

setting, where only a global label at the WSI/bag level is available. This is important in medical applications, where the collection of local labels requires significant manual effort, raising the question of what kind of labels to collect and what is the expertise required for their collection. For example, a non-expert could identify informative instances containing groups of follicular cells, while only a cytopathologist expert can determine the level of their abnormality (normal/atypical/malignant). In this context, we note the closely related task of region-of-interest detection, studied extensively for object detection (Uijlings et al., 2013; Girshick et al., 2014; Girshick, 2015; Ren et al., 2017). However, here we are not strictly concerned with the accurate estimation of bounding boxes of individual instances, a difficult challenge in the case of cytopathology, as our goal is to predict the global per-slide label.

The second gap is related to the structure of the bag in MIL in terms of the *prevalence of positive instances* (PPI) in a bag, which is typically not taken into account. The classical definition of MIL assumes at least one positive instance in a positive bag, while Kraus et al. (2016), for example, assume a certain number of positive instances triggering a global positive label. In our context, PPI measures the fraction of the positive instances (in a positive WSI), *i.e.,* those containing follicular groups with clear characteristic of malignancy. In contrast, a positive bag also contains non-malignant follicular groups, as well as uninformative instances. The uninformative instances constitute the vast majority of the scan, mainly containing red blood cells, considered in our case as background. This forms a unique bag structure of low PPI. On the other hand, once background instances are filtered out, as we propose in our approach, the bags composed of only informative instances have a high PPI structure; namely, the follicular groups are *consistent* in their indication of malignancy to a certain level, which we explore in this paper.

The third gap is the question of how to use multiple labels for improving classification. To this end, we consider the joint prediction of the malignancy labels, the TBS categories, and the local abnormality labels. Since both TBS categories and the local labels correspond to the increasing probability of malignancy, we consider their joint prediction using ordinal regression (Gutierrez et al., 2016; McCullagh, 1980; Agresti, 2003; Dorado-Moreno et al., 2012). The joint prediction is motivated by the observation that the local labels, as well as TBS categories, are a consistent proxy for the probability of malignancy (Jing et al., 2012; Pathak et al., 2014), and so their joint prediction induces cross-regularization.

This paper extends a previous conference publication Dov et al. (2019), where we presented an algorithm that provides predictions of thyroid malignancy comparable to those of cytopathology experts (we compared to three such experts). In Dov et al. (2019), we focused on a more thorough description of the clinical problem we address and provided complete details on the dataset and its acquisition. This paper focuses on the detailed derivation and the analysis of the proposed algorithm. Novel contributions, which go beyond Dov et al. (2019), include: We propose a maximum likelihood estimation (MLE) framework for classification in the mixed setting, where multiple global and local labels are available for training. While in classical MIL, informative instances are implicitly identified, the MLE framework allows explicit identification of them using the local labels, which we show to be especially useful in the low-PPI setting. We further derive a lower bound of the MLE, which corresponds to a weakly supervised training strategy, in which the global labels are

propagated to the instance level and used as noisy local labels. Statistical analysis and experiments on synthetic data show that this training strategy is particularly useful for high-PPI bags obtained by filtering out the background instances. From the lower bound of the MLE, we derive the algorithm for malignancy prediction, that is based on deep-learning and comprises two stages. The algorithm identifies instances containing groups of follicular cells and incorporates local decisions based on the informative regions into the global slide-level prediction. The lower bound of the MLE further allows us to investigate the simultaneous prediction of the global malignancy and the TBS category scores, as well as the local abnormality scores. Specifically, using ordinal regression, we extend our framework to jointly predict these labels from a single output of a neural network. Extensive cross-validation experiments comparing the proposed approach to competing methods, as well as ablation experiments, demonstrate the competitive performance of the proposed algorithm. We further show that the proposed ordinal regression approach allows application of the proposed algorithm to augment cytopathologist decisions.

## 2. Problem formulation

Let $\mathbb{X} = \{X_l\}$ be a set of WSIs, where $X_l = \{\mathbf{x}_{l,m}\}$ is the set of $M_l$ instances in the $l$th WSI. The $m$th instance $\mathbf{x}_{l,m} \in \mathbb{R}^{w \times h \times 3}$ is a patch from an RGB digital scan, whose width and height are $w$ and $h$, respectively. Let $\mathbb{Y} = \{Y_l\}$ be the corresponding set of malignancy labels: $Y_l \in \{0,1\}$, where 0 and 1 correspond to benign and malignant cases, respectively. The goal is to predict thyroid malignancy $\hat{Y}_l$. Similar to $\mathbb{Y}$, consider the set $\mathbb{S} = \{S_l\}$, where $S_l \in \{2,3,4,5,6\}$ is the TBS category assigned to a WSI by a pathologist.

We consider an additional set of *local* labels $\mathbb{U} = \{U_l\}$, where $U_l = \{\mathbf{u}_{l,m}\}$ and $u_{l,m} \in \{0,1\}$. $u_{l,m} = 1$ if instance $\mathbf{x}_{l,m}$ contains a group of follicular cells, and $u_{l,m} = 0$ otherwise. Our dataset includes 4494 such informative instances, manually selected (by a trained pathologist) from 142 WSIs. These local labels are exploited in the proposed framework for the improved identification of the informative instances. The instances containing follicular groups are further labeled according to their abnormality, forming the set $\mathbb{V} = \{v_{l,m}\}$, $v_{l,m} \in \{0,1,2\}$ (normal, atypical and malignant). While in the classical MIL setting, only the set of binary malignancy labels $\mathbb{Y}$ is available, we explore in this paper the contribution of the additional label sets $\mathbb{S}$, $\mathbb{U}$ and $\mathbb{V}$ for the improved prediction of thyroid malignancy.

## 3. Proposed framework for thyroid malignancy prediction

### 3.1. MLE formulation

Let $\mathscr{L}$ be the likelihood over the dataset given by:

$$\mathscr{L} \triangleq P(\mathbb{X}, \mathbb{Y}, \mathbb{U}) = \prod_l P(Y_l|X_l, U_l)P(U_l|X_l)P(X_l), \tag{1}$$

where for simplicity we only consider at this point the sets of labels $\mathbb{Y}$, $\mathbb{U}$. We drop the right most term by assuming a uniform distribution over the WSIs, and further assume the following conditional distribution on the label $Y_l$:

$$Y_l|X_l, U_l \sim \text{Bernoulli}\left(\frac{1}{\widetilde{M}}\sum_m \sigma(g_\theta(\mathbf{x}_{l,m}))u_{l,m}\right), \tag{2}$$

where $g_\theta(\mathbf{x}_{l,m}) \in \mathbb{R}$ is the output of a neural network with parameters $\theta$, $\sigma(\cdot)$ is the sigmoid function, and $\widetilde{M} \triangleq \sum_m u_{l,m}$ (note $\widetilde{M} \ll M_l$). This statistical model suggests the estimation of $Y_l$ from an average of local, instance-level estimates $g_\theta(\mathbf{x}_{l,m})$, weighted by $u_{lm}$ according to the level of their informativeness. We note that the true labels $u_{l,m}$ are available only for a small subset of instances. Therefore, $u_{l,m}$ in (2) and throughout the paper, refers to estimates of these labels unless otherwise noted. In addition, we consider $u_{l,m}$ as binary variables for simplicity, and note that our framework can be extended to continuous variables as well. We further analyze (2) in Subsection 3.2. Substituting (2) into (1) leads to the following log likelihood expression, the derivation of which is presented in Appendix 1:

$$\begin{aligned}\log \mathscr{L} = &\sum_l Y_l \log\left[\frac{1}{\widetilde{M}}\sum_m \sigma(g_\theta(\mathbf{x}_{l,m}))u_{l,m}\right] \\ &+ (1-Y_l)\log\left[1-\frac{1}{\widetilde{M}}\sum_m \sigma(g_\theta(\mathbf{x}_{l,m}))u_{l,m}\right] \\ &+ \sum_m \log P(u_{l,m}|\mathbf{x}_{l,m}).\end{aligned} \tag{3}$$

Maximizing the first two terms on the right hand side of (3) is equivalent to minimizing the binary cross entropy (BCE) loss in the MIL setting. For example, the average pooling method is obtained by setting $u_{l,m}$ = const, and the noisy-or algorithm (Zhang et al., 2006) is obtained by setting $u_{l,m}$ = 0 for all instances except the one providing the highest prediction value.

In fact, one can obtain a more general form of MIL classifier by considering a more general form of (2): $Y_l|X_l, U_l \sim \text{Bernoulli}(h(\sum_m g(\mathbf{x}_{l,m})u_{l,m}))$, where $h(\cdot) \in \mathbb{R}$ and $g(\cdot) \in \mathbb{R}^D$. This follows from Zaheer et al. (2017); Ilse et al. (2018), who showed that any function invariant to the order of the instances, *i.e.,* the MIL classifier in our case, can be decomposed into the form $h(\sum_m g(\mathbf{x}_{l,m})u_{l,m})$ with a particular selection of $h$, $g$. The attention mechanism of Ilse et al. (2018), for example, explicitly identifies informative instances, $u_{l,m}$, in a data-driven manner. Hou et al. (2016) proposed an EM-based iterative algorithm for MIL by heuristically estimating $u_{l,m}$ in the last term of (3) from instance level malignancy predictions. We show in Section 4 that classical MIL algorithms, in which selection of informative instances is implicit, completely fail to predict malignancy due to the low PPI of FNABs, which mostly comprise irrelevant background instances.

Equation (3) is more general than the classical MIL setting, as it also allows use of the local labels to estimate the informativeness of the instances. To that end, we propose to greedily maximize (3) in two steps: we use another neural network, denoted by $r_\phi(\cdot)$, trained using the last term of (3) and the local labels to estimate the informativeness of instances $u_{l,m}$ (see details in Subsection 3.4), and predict slide-level malignancy from the informative instances. Once trained, the network for the identification of informative instances $r_\phi(\cdot)$ is applied to the WSIs, and the estimated weights $u_{l,m}$ are set to 1 for the $\widetilde{M}$ most informative instances, and zero otherwise; hence the definition $\widetilde{M} \triangleq \sum_m u_{l,m}$ in (2) holds. We fix $\widetilde{M} = 1000$

instances, a value that balances the tradeoff between having a sufficient amount of training data to predict malignancy and using instances that with high probability are informative.

Once the informative instances are identified, we turn to the prediction of malignancy from the first two terms in (3). Since $\Sigma_m u_{l,m} \Big/ \widetilde{M} = 1$, we can write:

$$\begin{aligned}\log \mathscr{L} = & \sum_l Y_l \log \left[ \sum_m \sigma(g_\theta(\mathbf{x}_{l,m})) \frac{u_{l,m}}{\widetilde{M}} \right] \\ & + (1 - Y_l) \log \left[ \sum_m \frac{u_{l,m}}{\widetilde{M}} (1 - \sigma(g_\theta(\mathbf{x}_{l,m}))) \right] \\ & + \sum_m \log P(u_{l,m} | \mathbf{x}_{l,m}) .\end{aligned} \tag{4}$$

Using Jensen's inequality, we get the lower bound:

$$\begin{aligned}\log \mathscr{L} \geq & \sum_{l,m} \frac{u_{l,m}}{\widetilde{M}} [Y_l \log (\sigma(g_\theta(\mathbf{x}_{l,m}))) \\ & + (1 - Y_l) \log (1 - \sigma(g_\theta(\mathbf{x}_{l,m})))] \\ & + \log P(u_{l,m} | \mathbf{x}_{l,m}) \\ \triangleq & \log \mathscr{L}^{\mathbb{Y}} + \sum_{l,m} \log P(u_{l,m} | \mathbf{x}_{l,m}) .\end{aligned} \tag{5}$$

Recalling that $u_{l,m}$ are binary, the term $-\log \mathscr{L}^{\mathbb{Y}}$ is the BCE loss calculated using only the informative instances. The lower bound implies the global labels $\{Y_l\}$ are assumed to hold locally, *i.e.,* separately for each instance. We propose to train the neural network $g_\theta(\cdot)$ according to (5), and consider $\{g_{\boldsymbol{\theta}}(\mathbf{x}_{l,m})\}$ as local, instance-level, predictions of thyroid malignancy, which are averaged into a global slide-level prediction:

$$f_\theta(X_l) = \frac{1}{\widetilde{M}} \sum_m g_\theta(\mathbf{x}_{l,m}) u_{l,m}, \tag{6}$$

where high values of $f_{\boldsymbol{\theta}}(X_l)$ correspond to high probability of malignancy. Accordingly, the predicted slide-level thyroid malignancy $\hat{Y}_l$ is given by:

$$\hat{Y}_l = \begin{Bmatrix} 1; & \text{if } f_\theta(X_l) > \beta \\ 0; & \text{else} \end{Bmatrix}, \tag{7}$$

where $\beta$ is a threshold value.

### 3.2. Analysis of the lower bound in the high-PPI setting

The extent to which the assumption that the global label holds locally and separately for each instance, which stems from (5), is directly related to the bag structure. This assumption holds perfectly in the extreme case of PPI = 1, *i.e.,* that all instances are malignant in a malignant WSI and all of them are benign in a benign WSI. Yet, PPI smaller than 1 corresponds to a weakly supervised setting where instances are paired with noisy labels. Experimental studies, such as the those presented in Alpaydin et al. (2015); Rolnick et al. (2017), previously reported on the robustness of neural networks to such label noise. In this subsection, we analyze the utility of the lower bound in (5), (6) and (7) for MIL in the high

PPI setting. We note that the PPI of the bags is indeed high once the uninformative labels were filtered out, as we show by the analysis of the abnormality labels $v_{l,m}\}$ in Section 4.

A common practice in binary classification is to predict the conditional class probability $P(Y_l = 1|X_l)$. Specifically, in standard (single-instance) classification $P(Y_l = 1|X_l)$ is predicted, for example, from an output of a neural network trained with BCE loss. For simplicity, we analyze logit $(Y_l = 1|X_l)$, where $\text{logit}(\cdot) \triangleq \log\left(\frac{P(\cdot)}{1-P(\cdot)}\right)$, rather than $P(Y_l = 1|X_l)$. The following proposition shows that $f_{\boldsymbol{\theta}}(X_l)$ in (6) is related directly to logit $(Y_l = 1|\mathbf{x}_{l,m})$.

**Proposition 1.** *The estimate of logit*$(Y_l = 1|X_l)$ is given by a linear function of $f_{\boldsymbol{\theta}}(X_l)$*:*

$$logit(Y_l = 1|X_l) = \widetilde{M} f_\theta(X_l) + C, \tag{8}$$

where C is a constant and $\widetilde{M}$ is the number of the informative instances.

Proposition 1 implies that making a prediction according to (7) by comparing $f_{\boldsymbol{\theta}}(X_l)$ to a threshold value $\beta$ is equivalent to comparing the estimated logit function to the threshold $\gamma \triangleq \widetilde{M}\beta + C$. The proof is provided in Appendix 2. We further note that the logit function is directly related to the likelihood ratio test. Using Bayes rule: logit$(Y_l = 1|X_l)$ = log $\Lambda$ + $P(Y = 0)/P(Y = 1)$, where $\Lambda$ is the likelihood ratio defined as $\Lambda \triangleq P(X_l|Y_l = 1)/P(X_l|Y_l = 0)$. This implies that thresholding $f_{\boldsymbol{\theta}}(X_l)$ is equivalent to applying the likelihood ratio test, widely used for hypothesis testing (Casella and Berger, 2002).

Proposition 1 provides further insight into the training strategy suggested in (5). An implicit assumption made in the proof is that $\sigma\,(g_{\boldsymbol{\theta}}(\mathbf{x}_{l,m}))$ estimates the probability $P(Y_l = 1|\mathbf{x}_{l,m})$ of the slide being malignant given a single instance $\mathbf{x}_m$; a similar assumption is made in the derivation of the noisy-and MIL in Kraus et al. (2016). Proposition 1 therefore implies that $f_{\boldsymbol{\theta}}(X_l)$ predicts well the likelihood ratio provided that $g_{\boldsymbol{\theta}}(\mathbf{x}_{l,m})$ is a good estimate of $P(Y_l = 1|\mathbf{x}_{l,m})$. Equation (5) indeed suggests to directly predict the global label from each instance separately. The higher the PPI is, the lower is the noise level in the the labels used to predict $P(Y_l = 1|\mathbf{x}_{l,m})$ and, according to the proposition, the better is the global prediction of $P(Y_l = 1|X_l)$. This comes in contrast to (3) and, specifically, to classical MIL approaches, wherein the network is optimized to predict the global label from the *multiple* instances, and there is no guarantee on the quality of predictions of individual instances.

### 3.3. Simultaneous prediction of multiple global and local label

We now consider prediction of the TBS categories $\mathbb{S}$ and the local abnormality scores $\mathbb{V}$ using the likelihood over the full dataset $P(\mathbb{X}, \mathbb{Y}, \mathbb{U}, \mathbb{S}, \mathbb{V})$. To make the computation of the likelihood tractable, we assume that $P(\mathbb{Y} \mid \mathbb{X}, \mathbb{U})$, $P(\mathbb{S} \mid \mathbb{X}, \mathbb{U})$ and $P(\mathbb{V} \mid \mathbb{X}, \mathbb{U})$ are independent. The straightforward approach under this assumption is to extend (3) and (5) by adding two cross entropy loss terms to predict the labels $\mathbb{S}$ and $\mathbb{V}$, which leads to a standard multi-label scenario. However, this does not encode the strong relation between $\mathbb{Y}$, $\mathbb{S}$ and $\mathbb{V}$, in the sense that all indicate various abnormality (malignancy) levels. We therefore propose to encode these relations into the architecture of the neural network $g_\theta(\cdot)$. Specifically, we take

advantage of the ordinal nature of $\mathbb{S}$ and $\mathbb{V}$, where higher values of the labels indicate a higher probability of malignancy, and propose an ordinal regression framework to predict all three types of labels from a *single* output of the network. In what follows, we consider for simplicity only the prediction of the global TBS category $\mathbb{S}$. Extending the framework to predict the local labels $\mathbb{V}$ is straightforward, as our lower bound formulation in (5) treats local and global labels in the same manner.

Similar to (7), we propose to predict the TBS category by comparing the output of the network $f_{\boldsymbol{\theta}}(X_l)$ to threshold values $\beta_0 < \beta_1 < \beta_2 < \beta_3 \in \mathbb{R}$. Recall that the TBS category takes an integer value between 2 and 6, yielding:

$$\hat{S}_l = \begin{cases} 2; & \text{if } f_{\boldsymbol{\theta}}(X_l) < \beta_0 \\ n+2; & \text{if } \beta_{n-1} < f_{\boldsymbol{\theta}}(X_l) < \beta_n, n \in [1,2,3] \\ 6; & \text{if } f_{\boldsymbol{\theta}}(X_l) > \beta_3 \end{cases}. \tag{9}$$

The proposed framework for ordinal regression is inspired by the proportional odds model, also termed the cumulative link model (McCullagh, 1980; Dorado-Moreno et al., 2012). The original model suggests a relationship between $f_{\boldsymbol{\theta}}(X_l)$, the threshold $\beta_n$ and the cumulative probability $P(S_l - 2 \leq n)$, *i.e.,*

$$\text{logit}\,(S_l - 2 \leq n) = \beta_n - f_{\boldsymbol{\theta}}(X_l)\,. \tag{10}$$

The proportional odds model imposes order between different TBS by linking them to $f_{\boldsymbol{\theta}}(X_l)$ so that higher values of $f_{\boldsymbol{\theta}}(X_l)$ correspond to higher TBS categories. Recalling that the logit function is a monotone mapping of a probability function into the real line, values of $f_{\boldsymbol{\theta}}(X_l)$ that are significantly smaller than $\beta_n$ correspond to high probability that the TBS category is smaller than $n + 2$.

We deviate from McCullagh (1980); Dorado-Moreno et al. (2012) by estimating $P(S_l - 2 > n)$ rather than $P(S_l - 2 \leq n)$, which gives (derivation presented in Appendix 3):

$$P(S_l - 2 > n) = \frac{1}{1 + \exp[-(f_{\boldsymbol{\theta}}(X_l) - \beta_n)]}\,. \tag{11}$$

We note that this deviation is not necessary for the prediction of TBS, yet it allows combining the predictions of the thyroid malignancy and the TBS category in an elegant and interpretable manner. We observe that the right term in the last equation is the sigmoid function $\sigma(f_{\boldsymbol{\theta}}(X_l) - \beta_n)$. Accordingly, we can train the network to predict $P(S_l - 2 > l)$ according to:

$$\begin{aligned} \log \mathscr{L}^{\mathbb{S}} \triangleq \sum_{l,m} u_{l,m} \sum_{n=0}^{3} S_l^n \log\,[\sigma(g_{\boldsymbol{\theta}}(\mathbf{x}_{l,m}) - \beta_n)] \\ + (1 - S_l^n) \log\,[1 - \sigma(g_{\boldsymbol{\theta}}(\mathbf{x}_{l,m}) - \beta_n)], \end{aligned} \tag{12}$$

where $S_l^n = \mathbb{1}(S_l - 2 > n)$ and $\mathbb{1}(\cdot)$ is the indicator function. Specifically, maximizing $\log \mathscr{L}^{\mathbb{S}}$ is equivalent to minimizing 4 BCE loss terms with the labels $S_l^n$, $n \in (0,1,2,3)$, whose

explicit relation to TBS is presented in Table 5 in the Appendix. The use of $g_{\boldsymbol{\theta}}(\mathbf{x}_{l,m})$ in (12), instead of the more natural choice of $f_{\boldsymbol{\theta}}(X_l)$, is enabled by the lower bound in (5). The lower bound also allows us to extend this framework to predict the local abnormality score, which we denote by $\log \mathscr{L}^{\mathbb{V}}$, similar to (12) by considering two additional thresholds, $\gamma_0$, $\gamma_1$ and two corresponding BCE loss terms.

For the simultaneous prediction of thyroid malignancy, TBS category and the local labels, the total loss function is given by the sum of (5), (12) and $\log \mathscr{L}^{\mathbb{V}}$. We note the similarity between $\log \mathscr{L}^{\mathbb{S}}$ in (12) and $\log \mathscr{L}^{\mathbb{Y}}$ in (5), a result of our choice to estimate $P(S_l - 2 > n)$ rather than $P(S_l - 2 \leq n)$ and has the following interpretation: $\log \mathscr{L}^{\mathbb{Y}}$ can be considered a special case of ordinal regression with a single fixed threshold value of 0. The total loss function simultaneously optimizes the parameters $\boldsymbol{\theta}$ of the network $g_{\boldsymbol{\theta}}(\cdot)$ according to 7 classification tasks, corresponding to threshold values $\{0, \beta_0, \beta_1, \beta_2, \beta_3, \gamma_0, \gamma_1\}$.

The threshold values $\{\beta_0, \beta_1, \beta_2, \beta_3\}$ (and $\{\gamma_0, \gamma_1\}$) are learned along with the parameters of the networks, via stochastic gradient descent. While the training procedure does not guarantee the correct order of $\beta_0 < \beta_1 < \beta_2 < \beta_3$ (Dorado-Moreno et al., 2012), we have found in our experiments that this order is indeed preserved.

We note that, in some cases, the term of the loss function corresponding to the prediction of malignancy may conflict with that of the TBS category or the local label. For example, consider a malignant case ($Y_l = 1$) with TBS category 3 assigned by a pathologist. The term of the loss, in this case, which corresponds to TBS penalize high values of $f_{\boldsymbol{\theta}}(X_l)$ whereas the term corresponding to malignancy encourages them. We therefore interpret the joint estimation of TBS category, the local labels, and malignancy as a cross-regularization scheme. Given two scans with the same TBS but different final pathology, the network is trained to provide higher prediction values for the malignant case. Likewise, in the case of two scans with the same pathology but different local labels, the prediction value of the scan with the higher abnormality score is expected to be higher. Thus, the network adopts properties of the Bethesda system and the abnormality scores, such that the higher the prediction value $f_{\boldsymbol{\theta}}(X_l)$ the higher is the probability of malignancy. Yet the network is not strictly restricted to the Bethesda system and the local labels, so it can learn to provide better predictions.

### 3.4. Identification of the informative instances

We predict the informativeness of the instances using a second neural network $r_{\boldsymbol{\phi}}(\mathbf{x}_{l,m})$, optimized according to:

$$\begin{aligned} \log \mathscr{L}^{\mathbb{U}} \triangleq \sum_{l,m} \big[ & u_{l,m} \log(\sigma(r_{\boldsymbol{\phi}}(\mathbf{x}_{l,m}))) \\ & + (1 - u_{l,m}) \log\left(\sigma(r_{\boldsymbol{\phi}}(\mathbf{x}_{l,m}))\right)\big], \end{aligned} \tag{13}$$

where here $\{u_{l,m}\}$ is the set of the local labels. The term $-\log \mathscr{L}^{\mathbb{U}}$ is the standard BCE loss obtained from the last term in (5), assuming $u_{l,m} \mid \mathbf{x}_{l,m} \sim \text{Bernoulli}\left(\sigma(r_{\boldsymbol{\phi}}(\mathbf{x}_{l,m}))\right)$. Training the network requires sufficiently many labeled examples, the collection of which was done manually by an expert pathologist through an exhaustive examination of the slides. To make

the labeling effort efficient, the cytopathologist only marked positive examples of instances containing follicular groups ($u_{l,m} = 1$). We further observed in our experiments that instances sampled randomly from the whole slide mostly contain background. Therefore, to train the network $r_{\phi}(\mathbf{x}_{l,m})$, we assume that $u_{l,m} = 0$ for all instances in the last equation except those manually identified as informative. More specifically, we propose the following design of training batches. We use batches comprising an equal number of positive and negative examples to overcome the class imbalance. As positive examples, we take follicular groups sampled uniformly at random from the set of the labeled instances, *i.e.,* for which $u_{l,m} = 1$. Negative examples are obtained by sampling uniformly at random instances from the whole slide. Since in some cases informative instances can be randomly sampled and wrongly considered uninformative, the proposed training strategy can be considered weakly supervised with noisy negative labels.

To summarize, our complete log likelihood function is:

$$\log \mathscr{L}_{\text{total}} \triangleq \log \mathscr{L}^{\mathbb{Y}} + \log \mathscr{L}^{\mathbb{S}} + \log \mathscr{L}^{\mathbb{V}} + \log \mathscr{L}^{\mathbb{U}}, \tag{14}$$

where $\log \mathscr{L}_{\text{total}}$ is the lower bound on the full log-likelihood of the probabilistic model we assume for $\mathbb{X}$, $\mathbb{Y}$, $\mathbb{U}$, $\mathbb{S}$, $\mathbb{V}$. Note that one can further weight the different likelihood components if desired. This however is not considered in the experiments, for simplicity.

## 4. Experiments

### 4.1. PPI analysis on synthetic data

In Subsection 4.2, we evaluate the performance of the proposed algorithm of predicting thyroid malignancy compared to baseline MIL algorithms, considering the two settings of low PPI, when a bag comprises all instances in the WSI, and in the high PPI, after background instances were filtered out as a preprocessing step. To better understand the effect of the PPI on the performance of the different methods, we experimented with the CIFAR10 dataset Krizhevsky et al. (2009), designing a MIL setting where we can synthetically control the PPI. In this experiment, we consider each image as an instance and group them into bags. A bag is assigned with a positive label if at least one instance is positive and the PPI is controlled by setting the proper number of positive and negative instances in the bag. In this manner we construct multiple MIL datasets with different PPI values and evaluate the performance of the methods for each one of them. Specifically, CIFAR10 comprises natural images from 10 classes; we assign a positive label to an instance (image) if it belongs to one of 5 arbitrarily chosen classes and a negative label if it belongs to the other 5 classes. Each dataset comprises 1000 bags, with 100 instances each. The instances are assumed independent and are sampled uniformly at random from the original dataset, with equal probability to positive and negative bags. We note that we assume independence between the instances to facilitate the simulation, an assumption which may not hold in practice as instances from the same slide may be correlated. Given an average PPI value of a dataset, we allow slight variation of the PPI in each bag by sampling, uniformly at random, bag-level PPI values in the range of 0.8 – 1.2 of the average PPI. For each MIL dataset, we train each algorithm for 30 epochs and repeat the experiment 10 times.

We compare the proposed weakly supervised training strategy derived from the lower bound in (5) to the following MIL algorithms: noisy-or MIL, where the global prediction value is the highest local prediction, noisy-and MIL Kraus et al. (2016), the attention-based MIL algorithm presented in Ilse et al. (2018), and average-pooling MIL obtained by maximizing the first two terms of (3) rather than their lower bound. The methods are denoted “NoisyOr,” “NoisyAnd,” “AttentionMIL” and “AveragePooling,” respectively. The performance of the different algorithms is presented in Fig. 2.

As expected, the performance of the methods is improved with the increase of the PPI since there are more positive instances indicating that a bag is positive. Noisy-or MIL provides inferior performance compared to the other methods for most PPI values, and only for low PPIs it performs comparably. This is because the global decision is based only on a single instance, so this approach does not benefit from the multiple positive instances present in the slides when the PPI is high. This method was excluded from the following experiments due to poor performance.

Noisy-and MIL performs on par with average-pooling, where in both methods equal weights are assigned to the different instances. The improved performance obtained by the proposed training strategy compared to average-pooling MIL supports the use of the lower bound in 5, and the analysis in Section 3.2 implying that a better global prediction is obtained by training the network to directly predict the global label from each instance separately. For low PPI, the attention-based MIL provides the best performance indicating the advantage of using the attention mechanism to properly weight the instances. The proposed training strategy performs well for high PPI values, and provides the best performance even for PPI values as low as 0.18. This highlights an important advantage of the proposed training strategy, that allows prediction of the global label separately from each instance, even in the presence of a large amount of label noise.

### 4.2. Thyroid malignancy prediction

**Experimental Setting.—**To evaluate the proposed algorithm, we performed a 5-fold cross-validation procedure, splitting the 908 scans by 60%, 20%, 20% for training, validation, and testing, respectively, such that a test scan is never seen during training. The algorithm is trained using a Tesla P100-PCIE GPU with 16 Gb of memory. We use instances of size 128 × 128 pixels. This size is large enough to capture large groups of follicular cells while allowing the use of sufficiently many instances in each minibatch. Both the network for the identification of the informative instances $r_\phi(\cdot)$ and the network for the prediction of malignancy $g_\theta(\cdot)$ are based on the small and the fast converging VGG11 architecture Simonyan and Zisserman (2014), details of which are summarized in Table 4. We observed in our experiments that the training procedure of $r_\phi(\cdot)$ converges after a few epochs, so we set a stopping criterion to avoid over-fitting. Specifically, we use the average of predictions of positive examples, a criterion we find more reasonable than, *e.g.,* the area under the (ROC) curve (AUC). The latter takes into account negative examples, the accuracy of which we are uncertain since negative examples are randomly sampled from the WSI. We stop the training process if this measure does not increase between epochs, which typically occurs after 1 to 5 epochs. We use 10 instances per minibatch, a value set arbitrarily and that has a

small effect on the performance. The malignancy prediction network $g_\theta(\cdot)$ is trained for 100 epochs with a minibatch size of 288 instances, which corresponds to the maximum memory capacity of the GPU.

**Identification of instances containing follicular groups.—**We evaluated the performance of the network for the identification of informative instances $r_\phi(\cdot)$ using the annotated 142 WSIs, obtaining a test AUC of 0.985. A limitation of this analysis is that negative labels were sampled uniformly at random (as in the training procedure). We also calculated the average prediction value over the true informative instances, *i.e.*, those annotated by the pathologist, and received a test average prediction value of 0.97. Lastly, as we show below, the proposed approach leads to a significant improvement in the prediction of thyroid malignancy, in turn implying that the informative instances are indeed identified properly. In Fig. 3, we present examples of detected informative instances containing follicular groups.

We further illustrate in Figs. 4 and 5 a heat map of prediction values and a corresponding histogram of informativeness predictions in an example scan. With low prediction values, the majority of the instances contain background, as is seen in both figures. Specifically, the follicular groups (Fig. 4 top) are highlighted with bright colors in the heat map (Fig. 4 middle). In Fig. 5, the majority of instances contain background with low prediction values, however, the histogram is bimodal, with a second peak in the range of 0.95 to 1. These high predictions indeed correspond to instances containing follicular groups, which we select for thyroid malignancy prediction. This illustrates the extremely low PPI of FNAB WSIs, where only ~ 2% are informative and can be used to determine malignancy. We note that this example scan contains a relatively large amount of informative regions, selected for ease of presentation. In practice, the amount of informative regions can be as small as 0.01% as already stated.

In this context, we note that the number, size and complexity of follicular groups are features that may indicate malignancy. Follicular group count alone is not a reliable proxy for malignancy. For example, TBS 6 slides tend to have increased numbers of large follicular groups. However, malignant slides in lower TBS categories typically have lower counts of follicular groups. Moreover, there exists benign cases *(e.g.,* cases known as ‘Follicular Adenomas’) which exhibit similar characteristics in which the WSI is typically covered with a large number of follicular groups. For that reason, we avoid counting follicular groups.

We further note that we do not consider in this paper the accurate detection of bounding boxes Liu et al. (2019) nor pixel-level segmentation of follicular groups, rather just classifying instances of constant size as informative or not. The prediction of bounding boxes and segmentation could allow for the explicit estimation of the size and the shape of follicular groups and have the potential to improve classification performance. Yet, these are much more challenging tasks that require a significant amount of annotation effort both for training and evaluation data. Specifically, our data set does not include accurate boxes around the bounds of the follicular groups, and in many cases, only a part of the group is annotated.

**PPI analysis.—**While the large number of background instances pose low PPI, filtering them out as a preprocessing step significantly changes the PPI in the bag. To shed light on the structure of the bag, restricted to the subset of the informative instances, we present in Table 1 the distribution of the manually annotated local abnormality scores across the binary labels of malignancy and TBS categories. We note that the local abnormality labels were collected from an arbitrarily selected WSIs, and the cytopathologist, who was blinded to the malignancy and TBS categories of the WSIs, labeled each follicular group independently of other groups. In Table 1 top, 80% of the instances in malignant WSIs ($Y_l = 1$) are labeled malignant, and most of them originated in TBS 6 slides. This demonstrates the consistency between the local and the global labels, *i.e.,* high PPI. Yet the PPI is lower than one: for example, the cytopathologist assigned an atypical category to 17% of instances in malignant slides implying that they do not contain clear characteristics of malignancy. This demonstrates the label noise induced by the use of the lower bound in (5), according to which the global labels are propagated to the instance level. Interestingly, as seen in Table 1 bottom, benign slides include some instances marked malignant ($v_{l,m} = 2$) by the pathologist. This contradicts the classical MIL assumption that in a negative bag all bags are negative illustrating the uncertainty in the diagnosis of cytopathology FNABs.

**Prediction of thyroid malignancy.—**To evaluate the proposed algorithm and the contribution of the different label sets in its design, we first consider the prediction of thyroid malignancy from the whole slide using only the global labels, and use the baseline approaches "NoisyAND ($\mathbb{Y}, \mathbb{S}$)," "AttentionMIL ($\mathbb{Y}, \mathbb{S}$)" and a standard CNN ("CNN ($\mathbb{Y}, \mathbb{S}$)"), all of which are trained to simultaneously predict malignancy and the TBS category (notations indicate the labels used for training). These baselines are designed originally to process whole images, which is not possible in our case due to memory limitations. Therefore, we use crops of size 448 × 448 pixels, which allows 10 crops per minibatch, subject to memory limitations. These values were selected to optimize performance over the validation set. We compare these methods, to a version of the proposed algorithm trained according to (5) without the use of the local abnormalities, denoted by "Proposed ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$)." This comparison highlights the contribution of the local label set $\mathbb{U}$ for better identification of informative instances in the low PPI setting.

Once we applied the network to filter out the uninformative instances, each slide is represented by a set of the informative instances only, leading to a high PPI regime. We evaluate competing MIL approached in this case also, denoting them "NoisyAND ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$)," "AttentionMIL ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$)" and "AveragePooling ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$). These MIL methods are trained to predict the global labels ($\mathbb{Y}, \mathbb{S}$) from the set of the the informative instances representing each slide. We note that there is no straightforward way to incorporate the local abnormality labels $\mathbb{V}$ into the competing MIL approaches, since they are designed to use only global, slide level labels. We compare these methods to a variant of the proposed method "Proposed ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$)", which uses the same labels.

In addition, we consider multiple variants of the proposed algorithm, where each uses a different combination of the local and the global labels $\mathbb{Y}$, $\mathbb{S}$ and $\mathbb{V}$, respectively. Lastly, to better understand the advantage of using a single output of $g_\theta(\cdot)$ for the joint prediction of

the labels, we consider a version termed "Proposed2Heads ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$)," in which the network has two outputs, one for the prediction malignancy and the other for the prediction of the TBS category.

Table 2 summarizes the performance of the algorithms in the form of the area under receiver operating characteristic curve (AUC) and the average precision (AP). As can be seen in the table, "CNN ($\mathbb{Y}, \mathbb{S}$)," "NoisyAND ($\mathbb{Y}, \mathbb{S}$)" and "AttentionMIL ($\mathbb{Y}, \mathbb{S}$)" achieve markedly inferior performance compared to other methods. This is because their decisions are largely dominated by irrelevant background data. Specifically, the attention mechanism in "AttentionMIL ($\mathbb{Y}, \mathbb{S}$)" does not properly identify the informative instances due to low PPI. The method "Proposed ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$)" performs significantly better reflecting the large importance of separately identifying the informative instances according to the last term in (5) using the local labels.

In the high-PPI MIL setting, where each bag comprises only informative instances, "Proposed ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$)" marginally outperforms the methods "NoisyAND ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$)," "AttentionMIL ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$)" and "AveragePooling ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$)," and has among the lowest standard deviation. In particular, the higher AUC and AP values of the proposed algorithm, trained using the lower bound of the MLE in (5), compared to "AveragePooling ($\mathbb{Y}, \mathbb{U}, \mathbb{S}$)" devised from (3) are consistent with the experiment on synthetic data, as well as our analysis in Subsection 3.2, which suggest that better local predictions lead to improved global decisions. Moreover, as the analysis suggests, in the high-PPI setting, there is no advantage to the sophisticated aggregation of decisions from multiple instances presented in Ilse et al. (2018), relative to the simple averaging in (6).

The method "Proposed ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$)" marginally outperforms all other variants of the proposed method including both "Proposed ($\mathbb{U}, \mathbb{Y}$)" and "Proposed2Heads ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$)". This demonstrates the advantage of the proposed framework in the joint prediction of TBS categories, along with the binary malignancy labels from a single output of a neural network presented in Subsection 3.3. Interestingly, "Proposed ($\mathbb{U}, \mathbb{Y}, \mathbb{S}, \mathbb{V}$)" provides inferior performance compared to "Proposed ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$)". We trained the method "Proposed ($\mathbb{U}, \mathbb{Y}, \mathbb{S}, \mathbb{V}$)" using both the 4,494 manually annotated instances, as well as ~ 545,000 instances ($\widetilde{M} = 1000$ instances per WSI), for which we considered the global labels as noisy local labels. To balance the large difference in the size of these sets and better understand the contribution of the local labels, we considered minibatches comprising of 20% instances with local annotations. While it is possible to further tune the proportion of the instances with the local labels in the minibatches, this experiment suggests that there is no significant advantage to further incorporating local abnormalities scores in the proposed framework. This further demonstrates how well the network is trained using weak supervision by the global labels. This result further provides insight on the role the local labels and on the potential effect of inter-reviewer variability in their collection, which in our case was performed by a single pathologist. Specifically, we expect a small inter-reviewer variability in the identification of the follicular groups, which does not require a special expertise. On the other hand, assigning abnormality scores to the follicular groups can be done only by expert pathologists, and we do expect variability between reviewers. In the setting of our

experiments, and under the constraints we had on collecting expert annotations, the small number of 4,494 abnormality labels did not improve the results.

**Comparison to human-level performance.—**For the comparison of the algorithm to human-level performance, we use a subset of 109 slides which were reviewed by 3 expert cytopathologists (Experts 1 to 3), who assigned TBS categories, in addition to the TBS available in the original medical record (MR TBS). The performance of the proposed algorithm ("Proposed ($\mathbb{U}, \mathbb{Y}, \mathbb{S}$)") is compared to those of human in Fig. 6, using receiver operating characteristic (ROC) and precision-recall (PR) curves. Curves representing the performance of the human experts are obtained by considering the TBS categories as "human predictions of malignancy" such that TBS categories 2 to 6 correspond to increasing probability of thyroid malignancy. The AUC score obtained by the proposed algorithm is comparable to those of humans, and the algorithm provides an improved AP score compared to the human experts.

Figure 7 further presents a comparison of TBS scores assigned by the algorithm and the human experts. High values are obtained at the top-left and right-bottom of the matrix, while off-diagonal values decay. This block diagonal structure is exactly what is expected from the algorithm rather than, *e.g.,* a diagonal structure. For the indeterminate cases, assigned TBS 3 to 5 by the experts, the term of the loss function corresponding to final pathology $\mathscr{L}^{\mathbb{Y}}$ encourages the algorithm to deviate from the original TBS, and provide either lower values in the benign cases or higher values in the malignant ones. On the other hand, cases assigned with TBS 2 and 6 by cytopathologists are benign and malignant, respectively, in more than 97% of the cases. This high confidence in TBS 2 and 6 cases is similarly encoded in the algorithm, as we note that *all* the cases for which the algorithm predicts TBS 2 or 6 are indeed benign or malignant, respectively.

This implies the potential to apply the algorithm for augmenting cytopathologists' decisions. By the joint prediction of TBS and malignancy from a single output of the network, the proposed framework allows the grouping of predictions according to increasing probabilities of malignancy, using the thresholds $\{\beta_0,\beta_1,\beta_2,\beta_3\}$ in (12). This allows one to naturally combine the human and algorithm decisions according to the following rule: use human or the algorithm's decision if either of them assign TBS 2 or 6. In the case both of them assign an indeterminate score of TBS 3 to 5, we consider two variants: 1) use human decision, and 2) use the algorithm's decision. Table 3 shows that the combined rule where indeterminate decisions are held by the algorithm indeed improves the decisions of all three experts, further implying that the algorithm performs beyond human-level in indeterminate cases.

## 5. Conclusions

We have considered machine-learning-based prediction of thyroid malignancy from cytopathology WSIs, in the setting where multiple local and global labels are available for training. An MLE formulation has been presented, that extends MIL to this setting, and, using a lower bound of the MLE, devised a two-stage algorithm. Inspired by the work of a cytopathologist, the algorithm identifies informative instance containing follicular cells, and then assigns a reliable slide-level malignancy score, similar to the Bethesda system, where

higher values correspond to higher probabilities of malignancy. We showed that the MLE framework facilitates the use of local labels for the improved identification of informative instances in the low-PPI regime, where most instances are uninformative. In the high-PPI setting, after the uninformative instances have been excluded, statistical analysis and experiments on both synthetic and cytopathology WSIs data showed the advantage of the weakly supervised training strategy induced by the lower bound of the MLE. Experimental results further showed that the proposed framework for simultaneous prediction of binary malignancy labels and TBS categories does not benefit from the use of the manually collected abnormality scores. While a non-expert can manually identify informative instances, assigning abnormality scores requires the expertise of an expert cytopathologist and is costly and time-consuming. We showed that the proposed algorithm, without using these labels, achieves performance comparable to three cytopathologists, and demonstrated the application of the algorithm to improve human decisions. The proposed MLE framework and the lower bound have two important properties that are general rather than specific to thyroid data: the framework decouples the identification and classification of instances, and it naturally associates between local and global labels. Our future plans are to apply the framework to the diagnosis of prostate cancer, where these properties may be particularly useful. First, prostate diagnosis is determined by the classification of prostate glands, so that it may be useful first to separate them from the irrelevant background. Moreover, in prostate slides there is an explicit relation between the local and global labels: the global diagnostic score, termed the Gleason score, is given by the frequency and the severity of the local labels.

## Appendix 1

By substituting (2) into (1) we get:

$$\begin{aligned}\mathscr{L} = &\prod_l \left(\frac{1}{\widetilde{M}}\sum_m (\sigma(g_\theta(\mathbf{x}_{l,m}))u_{l,m})\right)^{Y_l} \\ &\cdot \left(1-\frac{1}{\widetilde{M}}\sum_m (\sigma(g_\theta(\mathbf{x}_{l,m}))u_{l,m}\right)^{1-Y_l} \\ &\cdot P(U_l|X_l)\,.\end{aligned} \tag{15}$$

We take a log from both sides of the equation:

$$\begin{aligned}\log \mathscr{L} = &\sum_l Y_l \log\left(\frac{1}{\widetilde{M}}\sum_m \sigma(g_\theta(\mathbf{x}_{l,m}))u_{l,m}\right) \\ &+ (1-Y_l)\log\left(1-\frac{1}{\widetilde{M}}\sum_m \sigma(g_\theta(\mathbf{x}_{l,m}))u_{l,m}\right) \\ &+ \log P(U_l|X_l),\end{aligned} \tag{16}$$

and we get (3) by rewriting the right term by assuming that being an instance informative is independent of other instances.

## Appendix 2

*Proposition 1* The estimate of logit($Y_l = 1|X_l$) is given by a linear function of $f_{\boldsymbol{\theta}}(X_l)$:

$$\text{logit}(Y_l = 1|X_l) = \widetilde{M} f_\theta(X_l) + C, \tag{17}$$

where $C$ is a constant and $\widetilde{M}$ is the number of the informative instances.

The proof is based on the assumption that the instances $\mathbf{x}_{l,m}$ are independent random variables. We note that this assumption is used to facilitate the derivation and it might not hold in practice for instances taken from the same scan. Yet, we motivate this assumption by the large variability between the follicular groups in their size, architecture and the number of cells as demonstrated in Fig. 3.

Proof. From the definition of logit( $Y_l = 1|X_l$), and using the Bayes rule we get:

$$\begin{aligned}\text{logit}(Y_l = 1|X_l) &= \log\left(\frac{P(Y_l = 1|X_l)}{P(Y_l = 0|X_l)}\right) \\ &= \log\left(\frac{P(X_l|Y_l = 1)}{P(X_l|Y_l = 0)}\right) + \log\left(\frac{P(Y_l = 0)}{P(Y_l = 1)}\right).\end{aligned} \tag{18}$$

By further using the independence assumption, we have:

$$\text{logit}(Y_l = 1|X_l) = \sum_m \log\left(\frac{P(\mathbf{x}_{l,m}|Y_l = 1)}{P(\mathbf{x}_{l,m}|Y_l = 0)}\right) + \log\left(\frac{P(Y_l = 0)}{P(Y_l = 1)}\right). \tag{19}$$

Since for the uninformative instances $P(\mathbf{x}_{l,m}|Y_l = 1) = P(\mathbf{x}_{l,m}|Y_l = 0)$, the sum in last equation is in fact over the $\widetilde{M}$ informative instances rather than over the whole set of size $M$. Another application of the the Bayes on the first right term leads to:

$$\begin{aligned}\text{logit}(Y_l = 1|X_l) &= \sum_m \log\left(\frac{P(Y_l = 1|\mathbf{x}_{l,m})P(Y_l = 0)}{P(Y_l = 0|\mathbf{x}_{l,m})P(Y_l = 1)}\right) \\ &+ \log\left(\frac{P(Y_l = 0)}{P(Y_l = 1)}\right) \\ &= \sum_m \log\left(\frac{P(Y_l = 1|\mathbf{x}_{l,m})}{P(Y_l = 0|\mathbf{x}_{l,m})}\right) + C \\ &= \sum_m \text{logit}(Y_l = 1|\mathbf{x}_{l,m}) + C,\end{aligned} \tag{20}$$

**Table 4.**

VGG11 based architecture used for both the first and the second neural networks in the proposed algorithm. Each conv2d layer comprises 2D convolutions with the parameters kernel_size = 3 and padding = 1. Parameters of the Max-pooling layer: kernel_size = 2, stride = 2. The conv2d and the linear layers (except the last one) are followed by batch normalization and ReLU. The network is trained using the binary cross entropy (BCE) loss via stochastic gradient descent with learning rate 0.001, momentum 0.99 and weight decay with decay parameter $10^{-7}$.

| **Feature extraction layers** | |
|---|---|
| **Layer** | **Number of filters** |
| conv2d | 64 |
| Max-pooling(M-P) | |
| conv2d | 128 |
| M-P | |
| conv2d | 256 |
| conv2d | 256 |
| M-P | |
| conv2d | 512 |
| conv2d | 512 |
| M-P | |
| **Classification layers** | |
| **Layer** | **Output size** |
| Linear | 4096 |
| Linear | 4096 |
| Linear | 1 |

where: $C \triangleq (M+1)\log(\frac{P(Y_l=0)}{P(Y_l=1)})$. According to (5), the last equation is estimated by:

$$\text{logit}(Y_l = 1|X_l) = \sum_m u_{l,m} g_\theta(\mathbf{x}_{l,m}) + C\,. \tag{21}$$

Finally, (17) is given by assigning (6) into (21).

## Appendix 3

By definition of the logit function and since $P(S_l - 2 \le n) = 1 - P(S_l - 2 > \text{n})$ we have:

$$\text{logit}\,(S_l - 2 > n) = -\text{logit}\,(S_l - 2 \le n)\,. \tag{22}$$

Further substituting the last equation into (10), gives:

$$\text{logit}\,(S_l - 2 > n) = -(f_\theta(X_l) - \beta_n)\,. \tag{23}$$

Last, we rewrite (23) as

$$P(S_l - 2 > n) = \frac{1}{1 + \exp\left[-(f_\theta(X_l) - \beta_n)\right]}. \quad (24)$$

**Table 5.**

Binary labels used in the proposed ordinal regression framework to predict the Bethesda score.

| | | $S_l^0$ | $S_l^1$ | $S_l^2$ | $S_l^3$ |
|---|---|---|---|---|---|
| Bethesda score | $S_l = 2$ | 0 | 0 | 0 | 0 |
| | $S_l = 3$ | 1 | 0 | 0 | 0 |
| | $S_l = 4$ | 1 | 1 | 0 | 0 |
| | $S_l = 5$ | 1 | 1 | 1 | 0 |
| | $S_l = 6$ | 1 | 1 | 1 | 1 |

## References


Agresti A, 2003 Categorical data analysis. volume 482 John Wiley & Sons.

Alpaydin E, Cheplygina V, Loog M, Tax DM, 2015 Single-vs. multiple-instance classification. Pattern recognition 48, 2831–2838.

Aschebrook-Kilfoy B, Schechter RB, Shih YCT, Kaplan EL, Chiu BCH, Angelos P, Grogan RH, 2013 The clinical and economic burden of a sustained increase in thyroid cancer incidence. Cancer Epidemiology and Prevention Biomarkers.

Campanella G, Hanna MG, Geneslaw L, Miraflor A, Silva VWK, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ, 2019 Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Medicine 25, 1301–1309.

Casella G, Berger RL, 2002 Statistical Inference. volume 2 Duxbury Pacific Grove, CA.

Cheplygina V, de Bruijne M, Pluim JP, 2019 Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Analysis 54, 280–296.

Cibas ES, Ali SZ, 2009 The bethesda system for reporting thyroid cytopathology. American J. of Clinical Pathology 132, 658–665.

Daskalakis A, Kostopoulos S, Spyridonos P, Glotsos D, Ravazoula P, Kardari M, Kalatzis I, Cavouras D, Nikiforidis G, 2008 Design of a multi-classifier system for discriminating benign from malignant thyroid nodules using routinely h&e-stained cytological images. Computers in Biology and Medicine 38, 196–203. [PubMed: 17996861]

Djuric U, Zadeh G, Aldape K, Diamandis P, 2017 Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. NPJ Precis. Oncology 1, 22.

Dorado-Moreno M, Gutierrez PA, Hervas-Martinez C, 2012 Ordinal classification using hybrid artificial neural networks with projection and kernel basis functions, in: Proc. International Conference on Hybrid Artificial Intelligence Systems, Springer pp. 319–330.

Dov D, Kovalsky SZ, Cohen J, Range DE, Henao R, Carin L, 2019 Thyroid cancer malignancy prediction from whole slide cytopathology images, in: Machine Learning for Healthcare Conference, pp. 553–570.

Elliott Range DD, Dov D, Kovalsky SZ, Henao R, Carin L, Cohen J, 2020 Application of a machine learning algorithm to predict malignancy in thyroid cytopathology. Cancer Cytopathology 128, 287–295. [PubMed: 32012493]

Gilshtein H, Mekel M, Malkin L, Ben-Izhak O, Sabo E, 2017 Computerized cytometry and wavelet analysis of follicular lesions for detecting malignancy: A pilot study in thyroid cytology. Surgery 161, 212–219. [PubMed: 27839932]

Girshick R, 2015 Fast r-cnn, in: Proc. of the IEEE International Conference on Computer Vision, pp. 1440–1448.

Girshick R, Donahue J, Darrell T, Malik J, 2014 Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580–587.

Glass C, Davis R, Xiong B, Dov D, Glass M, 2020a The use of artificial intelligence (ai) machine learning to determine myocyte damage in cardiac transplant acute cellular rejection. The Journal of Heart and Lung Transplantation 39, S59.

Glass M, Davis R, Dov D, Glass C, 2020b The use of artificial intelligence in diagnosing acute cellular rejection in cardiac transplant patients, in: LABORATORY INVESTIGATION, NATURE PUBLISHING GROUP 75 VARICK ST, 9TH FLR, NEW YORK, NY 10013-1917 USA pp. 301–301.

Gopinath B, Shanthi N, 2013 Computer-aided diagnosis system for classifying benign and malignant thyroid nodules in multi-stained fnab cytological images. Australasian Physical & Engineering Sciences in Medicine 36, 219–230. [PubMed: 23690210]

Gutierrez PA, Perez-Ortiz M, Sanchez-Monedero J, Fernandez-Navarro F, Hervas-Martinez C, 2016 Ordinal regression methods: survey and experimental study. IEEE Transactions on Knowl. and Data Engineering 28, 127–146.

Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH, 2016 Patch-based convolutional neural network for whole slide tissue image classification, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2424–2433.

Ilse M, Tomczak JM, Welling M, 2018 Attention-based deep multiple instance learning. arXiv preprint arXiv:1802.04712 (ICML18).

Jing X, Knoepp SM, Roh MH, Hookim K, Placido J, Davenport R, Rasche R, Michael CW, 2012 Group consensus review minimizes the diagnosis of follicular lesion of undetermined significance and improves cytohistologic concordance. Diagnostic Cytopathology 40, 1037–1042. [PubMed: 21538963]

Kim E, Corte-Real M, Baloch Z, 2016 A deep semantic mobile application for thyroid cytopathology, in: Proc. Medical Imaging 2016: PACS and Imaging Informatics: Next Generation and Innovations, International Society for Optics and Photonics p. 97890A.

Kraus OZ, Ba JL, Frey BJ, 2016 Classifying and segmenting microscopy images with deep multiple instance learning. Bioinform. 32, i52–i59.

Krizhevsky A, Hinton G, et al., 2009 Learning multiple layers of features from tiny images. Technical Report Citeseer.

Litjens G, Sanchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, Hulsbergen-Van De Kaa C, Bult P, Van Ginneken B, Van Der Laak J, 2016 Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific reports 6, 26286. [PubMed: 27212078]

Liu T, Guo Q, Lian C, Ren X, Liang S, Yu J, Niu L, Sun W, Shen D, 2019 Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks. Med. Image Analysis, 101555.

McCullagh P, 1980 Regression models for ordinal data. J. of the R. Statistical Society. Series B (Methodological), 109–142.

Ozolek JA, Tosun AB, Wang W, Chen C, Kolouri S, Basu S, Huang H, Rohde GK, 2014 Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning. Med. Image Analysis 18, 772–780.

Pathak P, Srivastava R, Singh N, Arora VK, Bhatia A, 2014 Implementation of the bethesda system for reporting thyroid cytopathology: interobserver concordance and reclassification of previously inconclusive aspirates. Diagnostic Cytopathology 42, 944–949. [PubMed: 24692395]

Quellec G, Cazuguel G, Cochener B, Lamard M, 2017 Multiple-instance learning for medical image and video analysis. IEEE Reviews in Biomedical Engineering 10, 213–234. [PubMed: 28092576]

Ren S, He K, Girshick R, Sun J, 2017 Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence , 1137–1149. [PubMed: 27295650]

Rolnick D, Veit A, Belongie S, Shavit N, 2017 Deep learning is robust to massive label noise. arXiv preprint arXiv:1705.10694.

Sanyal P, Mukherjee T, Barui S, Das A, Gangopadhyay P, 2018 Artificial intelligence in cytopathology: A neural network to identify papillary carcinoma on thyroid fine-needle aspiration cytology smears. J. of Pathology Informatics 9.

Savala R, Dey P, Gupta N, 2018 Artificial neural network model to distinguish follicular adenoma from follicular carcinoma on fine needle aspiration of thyroid. Diagnostic Cytopathology 46, 244–249. [PubMed: 29266871]

Simonyan K, Zisserman A, 2014 Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sirinukunwattana K, Raza SEA, Tsang Y, Snead DR, Cree IA, Rajpoot NM, 2016 Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Transactions on Medical Imaging 35, 1196–1206. [PubMed: 26863654]

Uijlings JR, Van De Sande KE, Gevers T, Smeulders AW, 2013 Selective search for object recognition. International J. of Computer Vis 104, 154–171.

Varlatzidou A, Pouliakis A, Stamataki M, Meristoudis C, Margari N, Peros G, Panayiotides JG, Karakitsos P, 2011 Cascaded learning vector quantizer neural networks for the discrimination of thyroid lesions. Anal. Quant. Cytol. Histol 33, 323–334. [PubMed: 22590810]

Zaheer M, Kottur S, Ravanbakhsh S, Poczos B, Salakhutdinov RR, Smola AJ, 2017 Deep sets, in: Proc. Advances in Neural Information Processing Systems (NIPS), pp. 3391–3401.

Zhang C, Platt JC, Viola P, 2006 Multiple instance boosting for object detection, in: Proc. Advances in Neural Information Processing Systems (NIPS), pp. 1417–1424.

Zhang Z, Chen P, McGough M, Xing F, Wang C, Bui M, Xie Y, Sapkota M, Cui L, Dhillon J, et al., 2019 Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat. Machine Intelligence 1, 236.

Zhou Z, 2018 A brief introduction to weakly supervised learning. National Science Review 5, 44–53.

## Highlights

- Machine-learning-based thyroid-malignancy prediction from cytopathology whole slides
- Beyond multiple instance learning: incorporating multiple global and local labels
- Weakly supervised method derived from a lower bound of a maximum likelihood estimator
- Ordinal regression framework for multi-label predictions augments human decisions

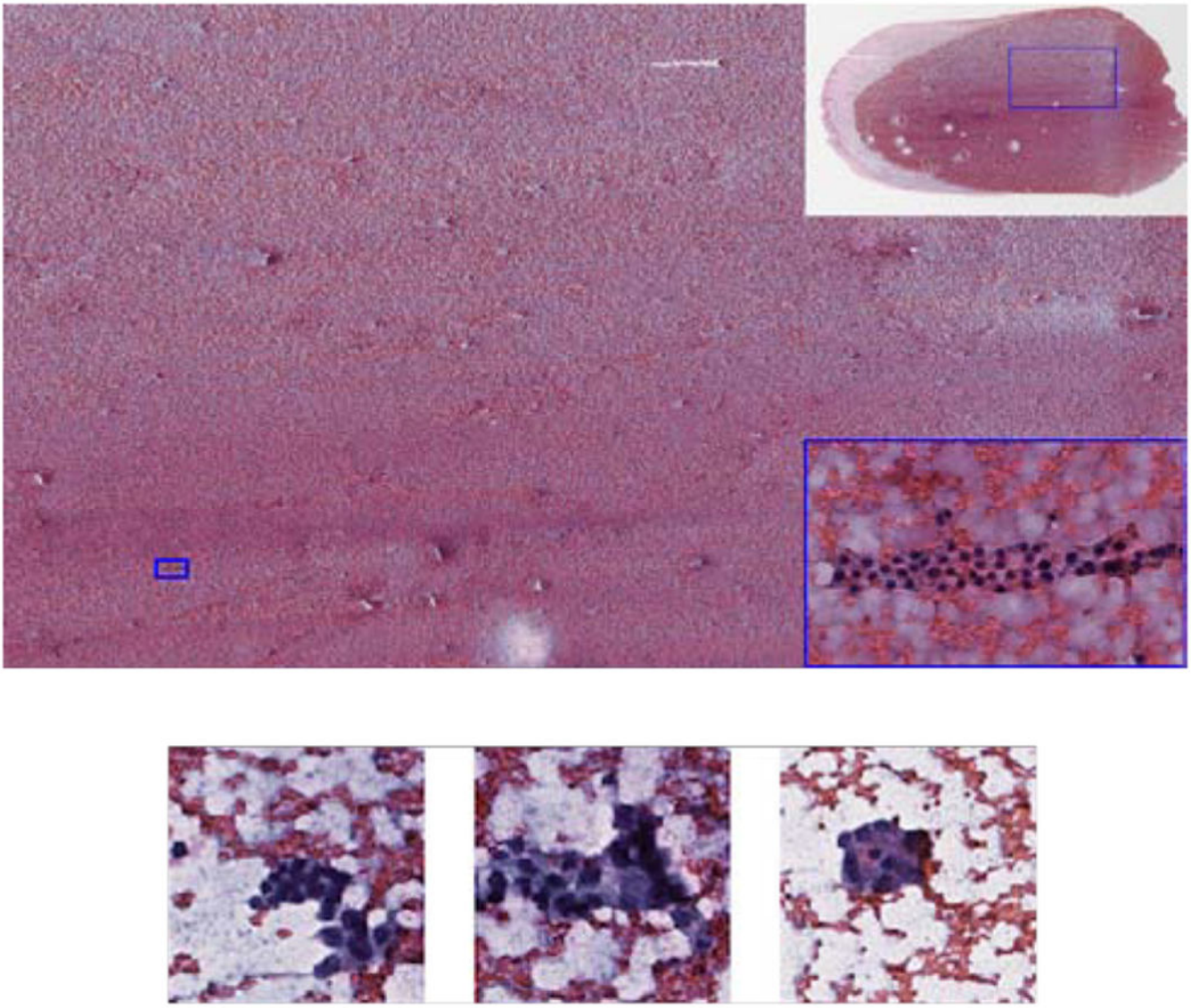


**Fig. 1.**

(Top) Example of a large image region of 10000 × 5000 pixels containing only a single group of follicular cells marked by the small rectangle. Top right corner: WSI with a rectangle indicating the location of the large image region. Bottom right corner: x10 zoomed in image of the informative follicular group. (Bottom) Examples of follicular groups with different abnormality levels. From left to right: benign, atypical and malignant.

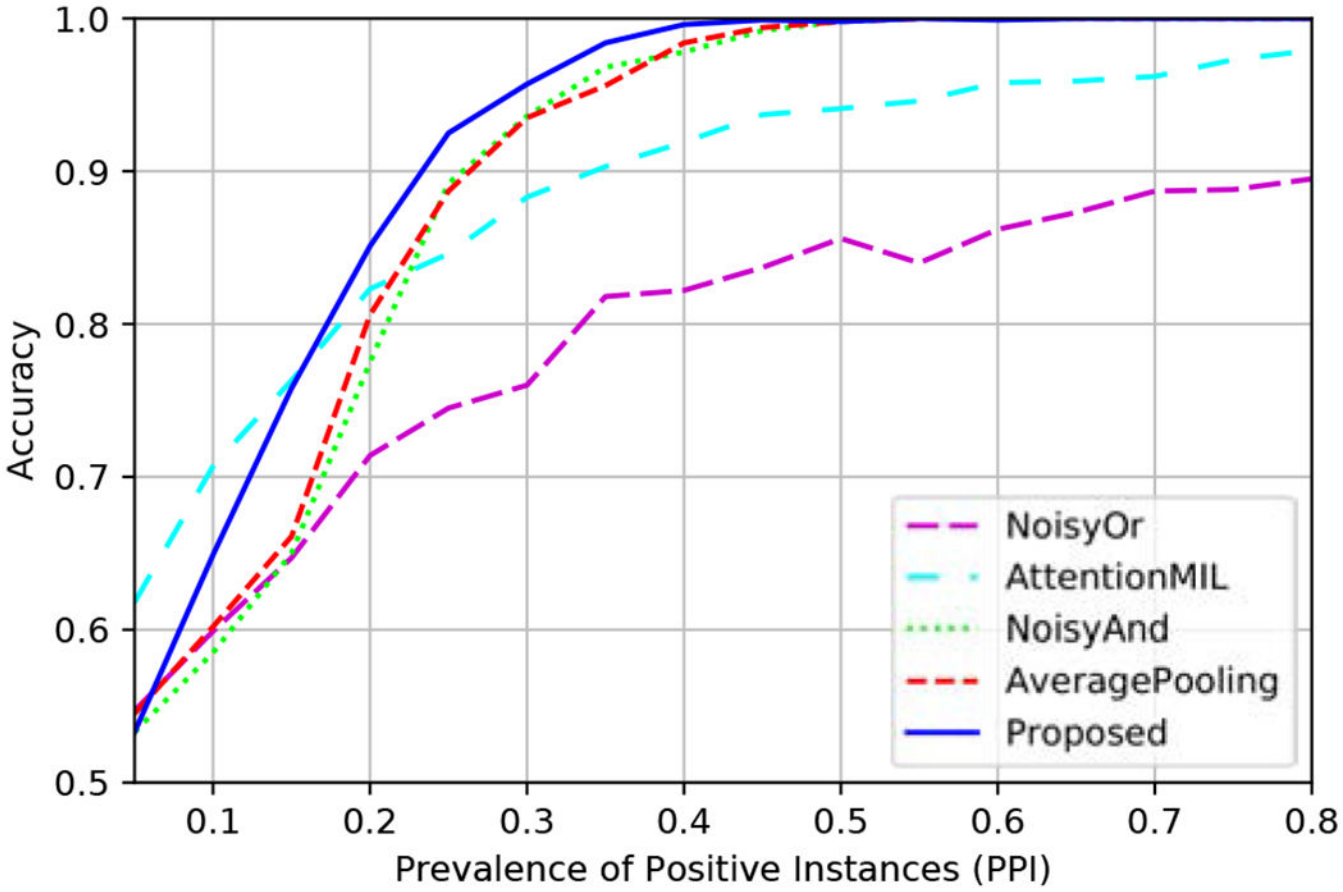



**Fig. 2.**
Accuracy *vs.* PPI on CIFAR10 data.

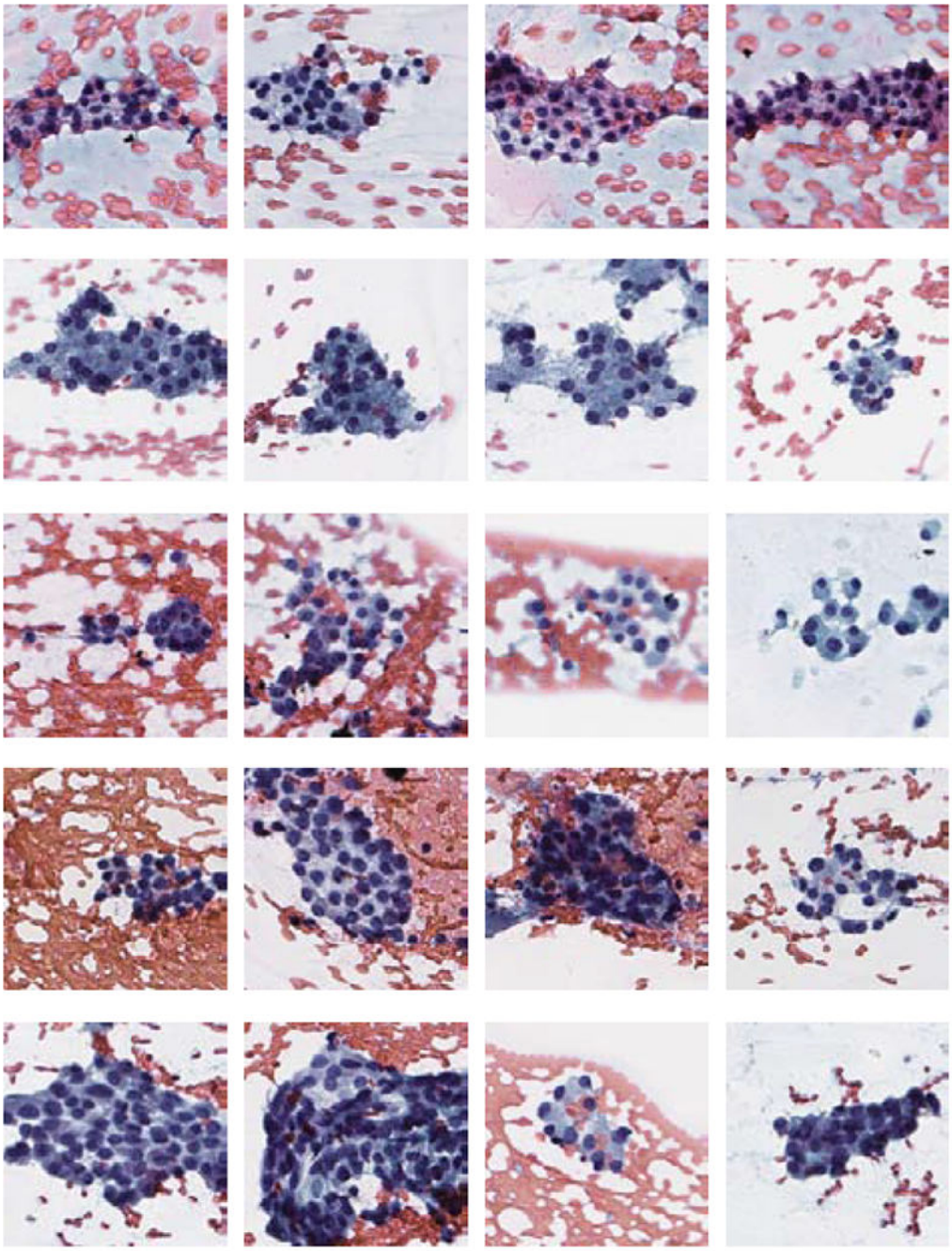


**Fig. 3.**

Instances containing follicular groups. The rows, from top to bottom, correspond to TBS 2 – 6 categories.

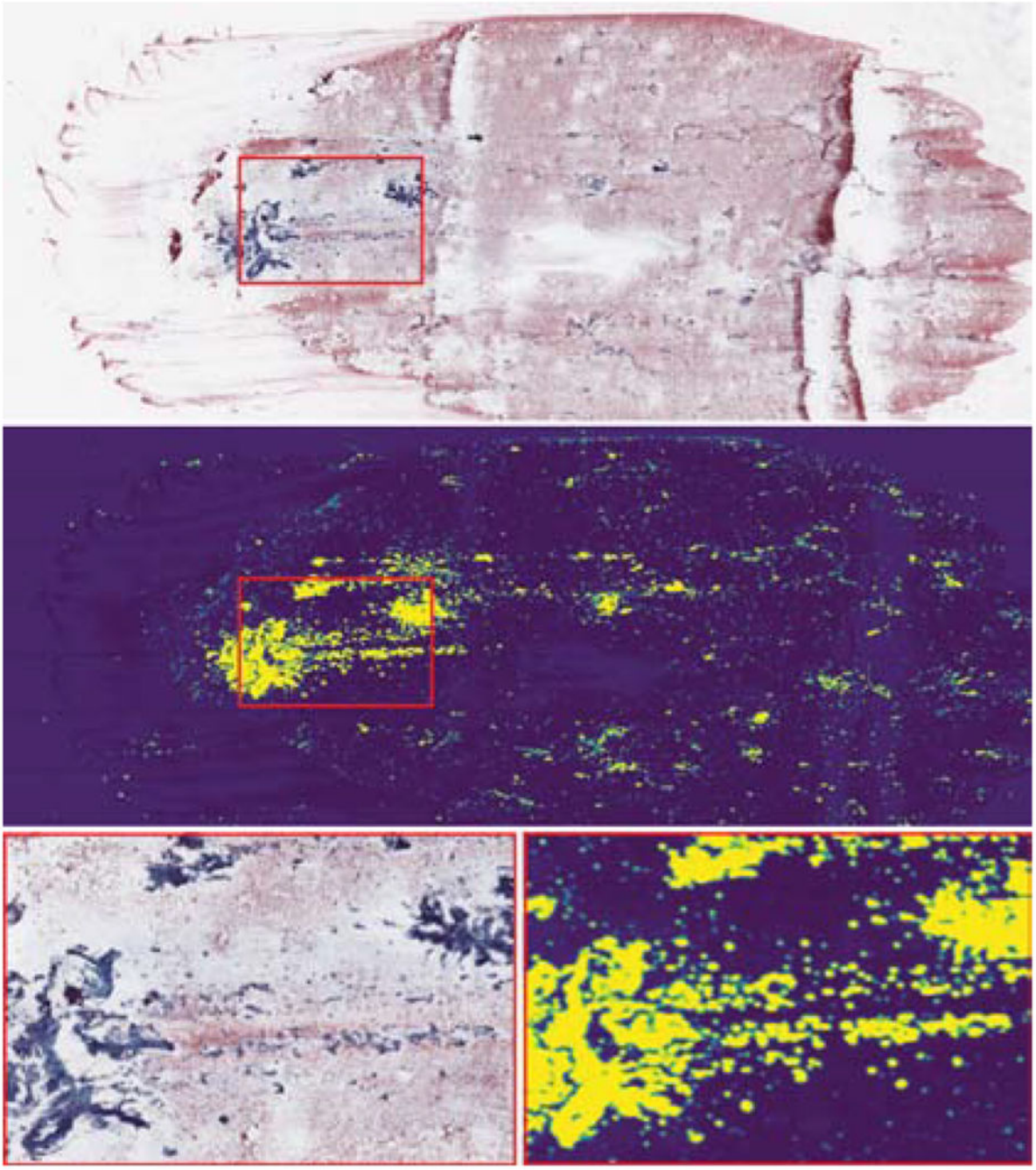


**Fig. 4.**

(Top) Whole-slide cytopathology scan. (Bottom left) Detail of the area marked by the red rectangle. (Middle) Heat map of prediction values of the first neural network. Instances predicted to contain follicular groups correspond to bright regions. (Bottom right) Detail of the are marked by the red rectangle.

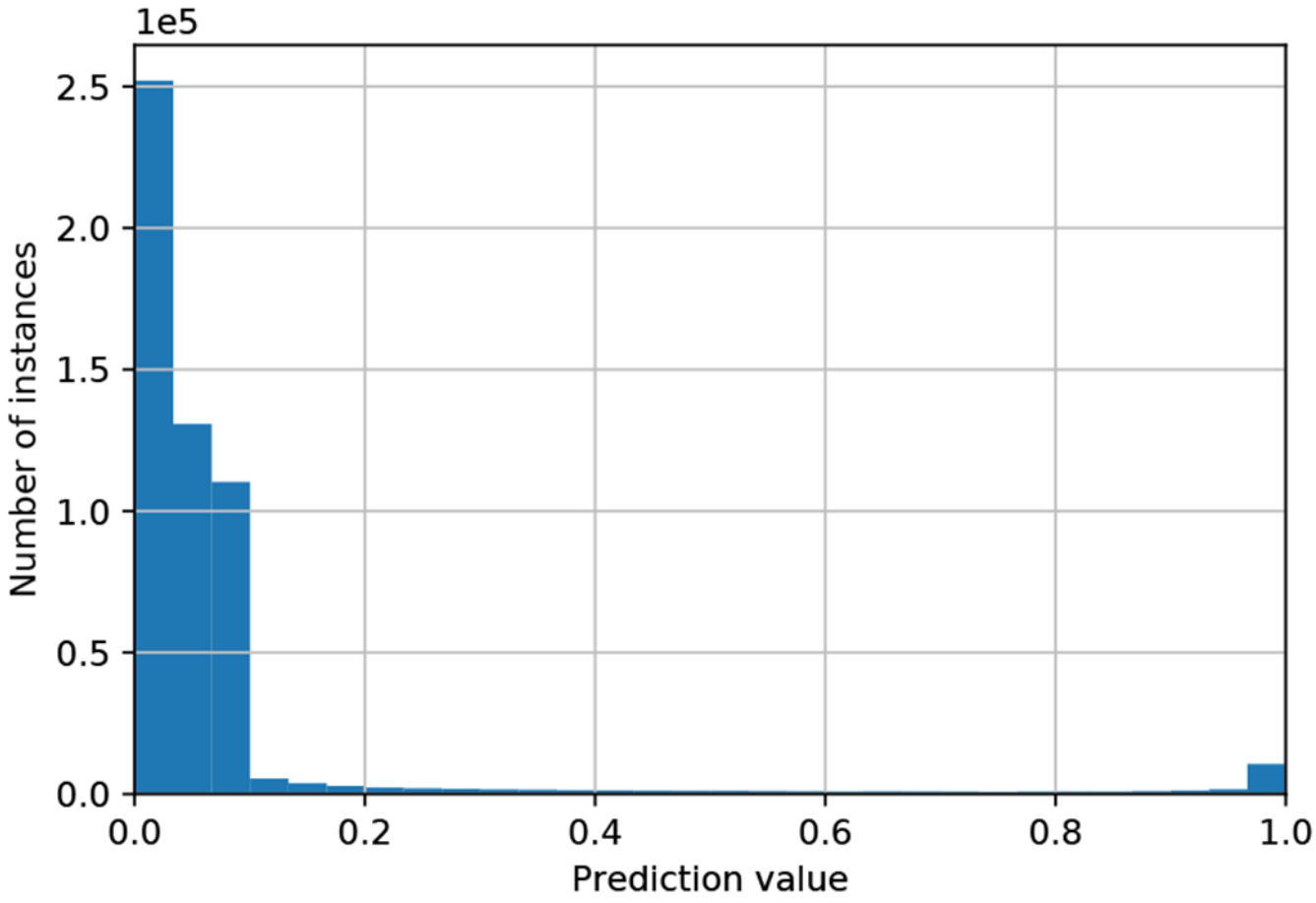



**Fig. 5.**

Histogram of predictions for instances taken from a single slide. High prediction values correspond to high probabilities that an instance contain follicular groups.

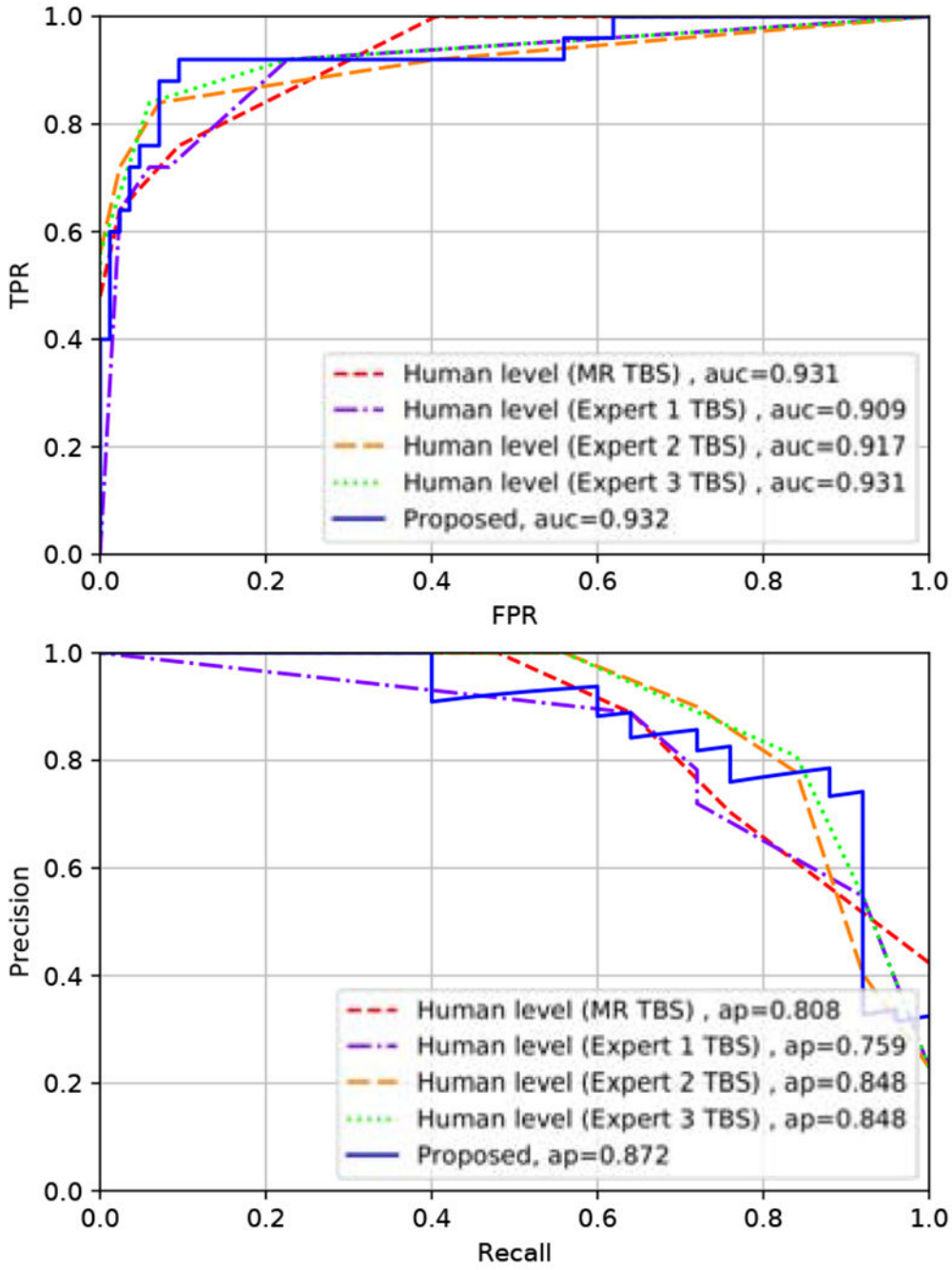



**Fig. 6.**

ROC (Left) and PR (Right) curves comparing the performance of the proposed algorithm and human experts in predicting thyroid malignancy. Blue curve - the proposed algorithm. Red curve - pathologist from the medical record. Purple, orange and green curves - expert cytopathologists 1, 2, and 3, respectively (these three individuals analyzed the same digital image considered by the algorithm, and these experts were not the same as the clinicians from the medical record).

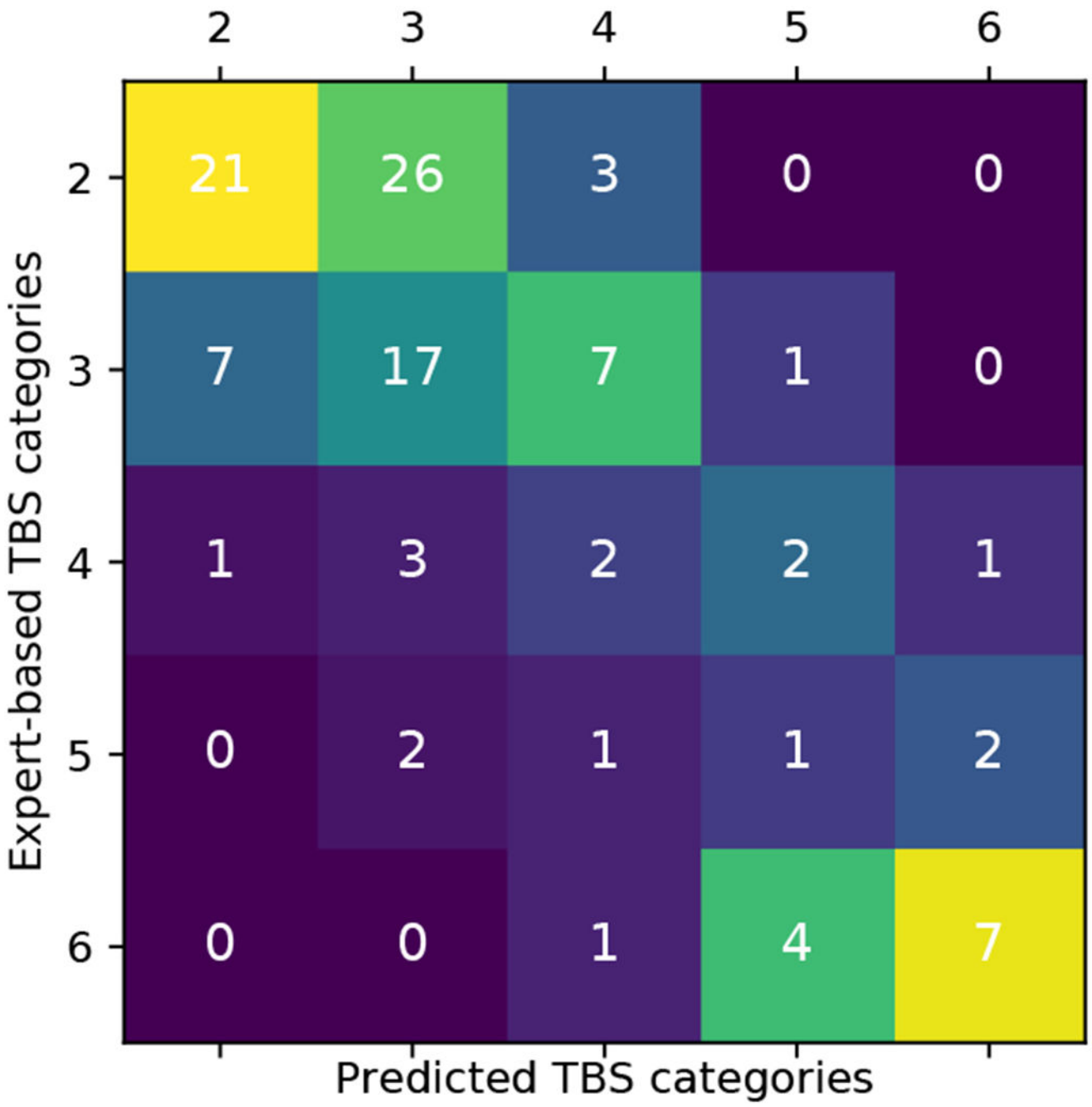



**Fig. 7.**

Confusion matrix of TBS categories assigned by the proposed algorithm *vs.* human experts. The colors in the plot correspond to a column normalized version of the confusion matrix.

**Table 1.**

Distribution of local abnormality labels. (Top) Distribution in malignant slides ( $Y_I = 1$). (Bottom) Distribution in benign slides ( $Y_I = 0$).

| | **Local abnormality labels ($v_{l,m}$)** | | |
|---|---|---|---|
| **Global label** | **Benign (0)** | **Atypical (1)** | **Malignant (2)** |
| $S_I = 2$ | 15 (33%) | 30 (67%) | 0 (0%) |
| $S_I = 3$ | 20 (38%) | 32 (62%) | 0 (0%) |
| $S_I = 4$ | 0 (0%) | 51 34%) | 100 (66%) |
| $S_I = 5$ | 4 (6%) | 28 (47%) | 28 (47%) |
| $S_I = 6$ | 4 (0%) | 110 (10%) | 1027 (90%) |
| Total | 43 (3%) | 251 (17%) | 1155 (80%) |
| | **Local abnormality labels ($v_{l,m}$)** | | |
| **Global label** | **Benign (0)** | **Atypical (1)** | **Malignant (2)** |
| $S_I = 2$ | 1234 (80%) | 328 (20%) | 0 (0%) |
| $S_I = 3$ | 281 (40%) | 373 (53%) | 46 (7%) |
| $S_I = 4$ | 41 (6%) | 619 (89%) | 38 (5%) |
| $S_I = 5$ | 4 (5%) | 74 (95%) | 0 (0%) |
| $S_I = 6$ | 0 | 0 | 0 |
| Total | 1565 (51%) | 1396 (46%) | 84 (3%) |

**Table 2.**

Comparison of the performance of the competing algorithms in the form of AUC and AP scores.

| Method | AUC | AP |
|---|---|---|
| CNN (𝕐, 𝕊) | 0.748 ± 0.035 | 0.498 ± 0.037 |
| NoisyAND (𝕐, 𝕊) | 0.761 ± 0.027 | 0.538 ± 0.037 |
| AttentionMIL (𝕐, 𝕊) | 0.743 ± 0.055 | 0.486 ± 0.095 |
| NoisyAND (𝕌, 𝕐, 𝕊) | 0.845 ± 0.016 | 0.708 ± 0.041 |
| AttentionMIL (𝕌, 𝕐, 𝕊) | 0.823 ± 0.021 | 0.643 ± 0.048 |
| AveragePooling (𝕌, 𝕐, 𝕊) | 0.850 ± 0.025 | 0.693 ± 0.037 |
| Proposed (𝕌, 𝕐) | 0.858 ± 0.017 | 0.719 ± 0.029 |
| Proposed (𝕌, 𝕊) | 0.852 ± 0.024 | 0.713 ± 0.040 |
| Proposed (𝕌, 𝕍) | 0.835 ± 0.024 | 0.693 ± 0.049 |
| Proposed (𝕌, 𝕐, 𝕍) | 0.858 ± 0.014 | 0.721 ± 0.035 |
| Proposed (𝕌, 𝕊, 𝕍) | 0.857 ± 0.018 | 0.733 ± 0.048 |
| Proposed2Heads (𝕌, 𝕐, 𝕊) | 0.860 ± 0.019 | 0.711 ± 0.046 |
| Proposed (𝕌, 𝕐, 𝕊) | **0.870 ± 0.017** | **0.743 ± 0.037** |
| Proposed (𝕌, 𝕐, 𝕊, 𝕍) | 0.860 ± 0.024 | 0.730 ± 0.047 |

**Table 3.**

Combined human and algorithm decisions. Decision rule: use human or algorithm decisions if either of them assign TBS 2 or 6. “Combined345Human”: the decision in TBS 3,4,5 cases is made by human. “Combined345Alg: the decision in TBS 3,4,5 cases is by the algorithm.

| | **Human** | **Combined345Human** | **Combined345Alg** |
|---|---|---|---|
| Expert 1 | 0.909 | 0.918 | **0.925** |
| Expert 2 | 0.917 | 0.925 | **0.929** |
| Expert 3 | 0.931 | 0.934 | **0.937** |
| | **Human** | **Combined345Human** | **Combined345Alg** |
| Expert 1 | 0.759 | 0.784 | **0.812** |
| Expert 2 | 0.848 | 0.867 | **0.886** |
| Expert 3 | 0.848 | 0.864 | **0.888** |