Automatic Quantification of Tumour Hypoxia from Multi-modal Microscopy Images using Weakly-Supervised Learning Methods

Gustavo Carneiro, Tingying Peng, Christine Bayer, Nassir Navab

***Abstract*—In recently published clinical trial results, hypoxia-modified therapies have shown to provide more positive outcomes to cancer patients, compared with standard cancer treatments. The development and validation of these hypoxia-modified therapies depend on an effective way of measuring tumour hypoxia, but a standardised measurement is currently unavailable in clinical practice. Different types of manual measurements have been proposed in clinical research, but in this paper we focus on a recently published approach that quantifies the number and proportion of hypoxic regions using high resolution (immuno-)fluorescence (IF) and hematoxylin and eosin (HE) stained images of a histological specimen of a tumour. We introduce new machine learning-based methodologies to automate this measurement, where the main challenge is the fact that the clinical annotations available for training the proposed methodologies consist of the total number of normoxic, chronically hypoxic and acutely hypoxic regions without any indication of their location in the image. Therefore, this represents a weakly-supervised structured output classification problem, where training is based on a high-order loss function formed by the norm of the difference between the manual and estimated annotations mentioned above. We propose four methodologies to solve this problem: 1) a naive method that uses a majority classifier applied on the nodes of a fixed grid placed over the input images; 2) a baseline method based on a structured output learning formulation that relies on a fixed grid placed over the input images; 3) an extension to this baseline based on a latent structured output learning formulation that uses a graph that is flexible in terms of the amount and positions of nodes; and 4) a pixel-wise labelling based on a fully convolutional neural network. Using a dataset of 89 weakly annotated pairs of IF and HE images from eight tumours, we show that the quantitative results of methods (3) and (4) above are equally competitive and superior to the naive (1) and baseline (2) methods. All proposed methodologies show high correlation values with respect to the clinical annotations.**

***Index Terms*—Microscopy, Structured output learning, Deep learning, Weakly-supervised training, High-order loss functions**

I. Introduction

Tumour hypoxia is characterised by a poor tissue oxygenation that is negatively associated with the effectiveness of standard cancer therapies [46]. There is also evidence that fluctuating tumour hypoxia levels with time indicates the development of aggressive survival strategies, such as local invasion, metastasis, and acquired treatment resistance [2]. Tumour hypoxia can be classified into chronic or acute, depending on its causes, duration and consequences, where chronic hypoxia results in a limitation of tumour growth, while acute hypoxia promotes tumour aggressiveness [24]. Recently published clinical trial studies show that hypoxia-modified therapies appear to be beneficial for cancer patients that are classified with respect to their hypoxic tumour status, compared with standard cancer therapies [46]. Nevertheless,

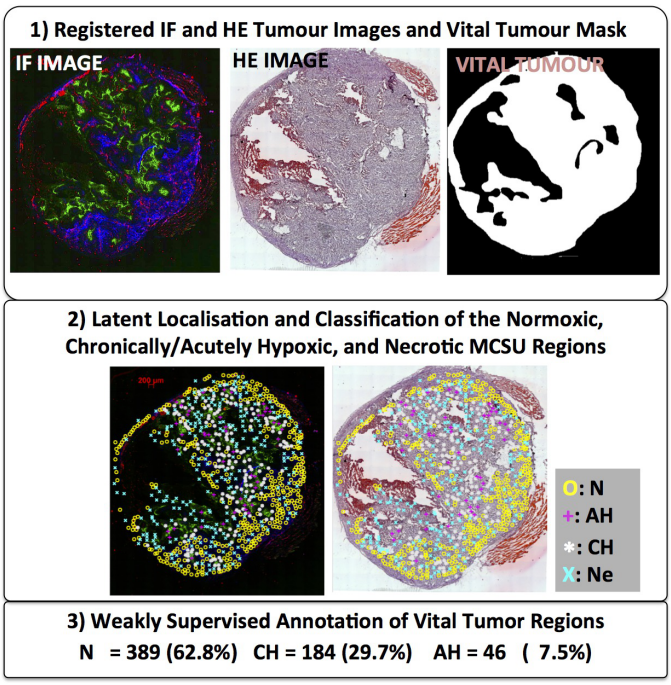


Fig. 1. From the IF and HE images and the vital tumour mask (box 1) obtained from a tumour tissue, the proposed methodologies must produce a high-level annotation (box 3) consisting of the number and proportion of normoxic (N), chronically hypoxic (CH), and acutely hypoxic (AH) regions. Box 2 shows the latent localisation and classification of MCSUs in the input images, where it is also necessary to detect the necrotic regions because of the inaccurate segmentation provided by the vital tumour mask. Note that in the HE image, pink regions denote vital tumour regions, red regions represent necrotic tissue, and white regions indicate missing tissue caused by the imaging process, as explained below in Sec. III; while in the IF image, red denotes micro-vessel, green represents hypoxia and blue means perfusion. This figure is better visualised electronically - please zoom in the IF/HE images in the middle box to notice the region annotations.

G. Carneiro is with the Australian Centre for Visual Technologies, University of Adelaide; T. Peng is with the Computer Aided Medical Procedures, Technische Universität München; C. Bayer is with the Department of Radiation Oncology, Technische Universität München; and N. Navab is with the Computer Aided Medical Procedures, Technische Universität München and Johns Hopkins University.

G. Carneiro thanks the Alexander von Humboldt Foundation for the Fellowship for Experienced Researchers and the Australian Research Councils Discovery Projects funding scheme (project DP140102794). T. Peng thanks the Alexander von Humboldt Foundation for the Fellowship for Postdoctoral Researchers.

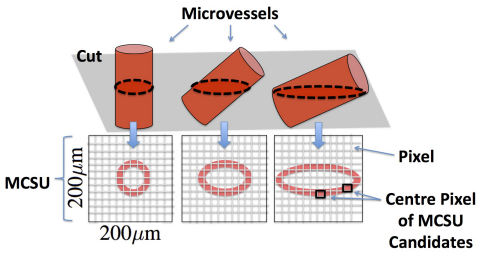


Fig. 2. The imaging process of an MCSU. Notice that a micro-vessel can be cut in different ways, generating visually different MCSUs. Also notice that each pixel with a strong red component in the IF stained image represents the centre pixel of an MCSU candidate, which must be clustered with other neighbouring detected MCSU candidates in order to form an MCSU.

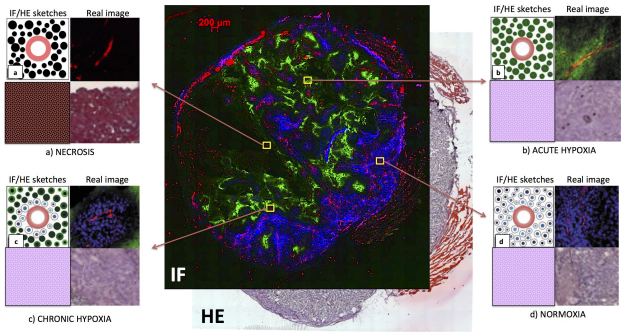


Fig. 3. Sketch of the appearance of MCSU classes [23]. Necrotic regions (a) have a red region at the centre of the IF image, followed by black pixels around it (indicating necrotic tissue); acute hypoxia (b) also has a red centre, but immediately followed by green regions (indicating hypoxia); chronic hypoxia (c) is denoted by a red region at the centre with a blue region immediately around it (indicating perfusion), followed by a green region towards the border; and normoxic MCSUs (d) again have a red region at the centre, and a blue region around it. Moreover, normoxic, chronic and acute hypoxic MCSUs have a smooth pink appearance in the HE image (indicating vital tumour tissue), while necrotic regions have a broken red appearance.

in spite of the apparent success of such therapies, there is no standard way of measuring tumour hypoxia in clinical practice [46]. Therefore, one of the critical points for the validation and further development of hypoxia-modified therapies is the implementation of an effective and practical way for measuring tumour hypoxia [46].

Several methods for manually measuring tumour hypoxia have been proposed [46], but a central problem present in most of the proposed measurements is a lack of sufficient spatial resolution to quantify the heterogeneity of the oxygenation supply of the tumour (e.g., pO2 probe do not have any spatial information and PET measurement has a poor resolution). Maftei et al. [24] addressed this issue with the use of high resolution (immuno-)fluorescence (IF) and hematoxylin and eosin (HE) stained images of a histological specimen of a tumour to highlight hypoxic regions. Essentially, the measurement proposed by Maftei et al. [24] (see Fig. 1) consists of estimating the number and percentage of normoxic (N: regions with adequate oxygen supply), chronically hypoxic (CH), and acutely hypoxic (AH) micro-circulatory supply units (MCSU), where an MCSU is an area of the tissue supplied by a micro-vessel (see Fig. 2). In particular, this measurement comprises the following steps: 1) registration of the IF and HE stained images [33] that not only allows for the simultaneous local feature extraction from both image modalities, but also for the transferring of the coarse vital tumour mask manually defined on the HE image to the corresponding IF image, hence removing the majority of necrotic tissue; 2) localisation of MCSUs [23]; 3) classification of each MCSU into N, CH, AH and necrosis (Ne - residual necrotic tissue as the manually defined vital tumour mask is quite coarse); and 4) calculation of the number and proportion of MCSUs classified as N, CH and AH. Currently, this quantification depends on a laborious and subjective manual annotation that requires an expertise which is not generally available in clinical practice. Therefore, the acceptance of this way of measuring tumour hypoxia would be promoted by the availability of an automated tool that can robustly emulate this manual annotation. In turn, the availability of such tool would facilitate the design and implementation of large-scale clinical tests to validate future hypoxia-modified therapies.

In this paper, we propose the use of machine learning-based models to automate the annotation proposed by Maftei et al. [24]. In particular, we explore the use of structured output models to automatically quantify the number and proportion of MCSUs classified as N, CH and AH using the IF and HE stained images of a histological specimen of a tumour [24]. A major challenge in the development of these models is that the training set, annotated by an clinical expert, contains the IF and HE images with the respective number and proportion of MCSUs classified as N, CH and AH, without indication of MCSU locations, sizes and classes [24] . This challenge is alleviated by the following general description of an MCSU and respective classification [2], [24]: 1) an MCSU is represented by a square region of size $200\times200\mu$m characterised by a cluster of red pixels in the IF image denoting a transversal cut of a microvessel at the centre of the region (Fig. 2), and 2) the appearance of the N, CH, AH and Ne classes are as defined in Fig. 3. This description allows the production of a rough manual annotation by a non-expert to train local detectors that can find **pixels** representing MCSU candidates and local classifiers that can discriminate $200 \times 200\mu$m regions around each detected pixel with respect to the N, CH, AH and Ne classes. These clinical expert and non-expert annotations are then used for training the following structured output learning models: 1) a naive method that uses a majority classifier applied on the nodes of a fixed grid placed over the input images; 2) a baseline method based on a structured output learning formulation that relies on a fixed grid placed over the input images; 3) an extension to this baseline based on a latent structured output learning formulation that uses a graph that is flexible in terms of the amount and positions of nodes [48], [34]; and 4) a pixel-wise labelling method based on fully convolutional neural network [18], [22]. These models are trained with a high-order loss function that minimises the norm of the difference between the clinically annotated and automatically estimated number of N, CH and AH MCSUs present in the IF and HE images of a histological specimen of a tumour. Hence, given that the clinical annotation does not contain the location and classification of MCSUs and the output of the method depends on such latent localisation and classification, this is a weakly supervised learning problem.

Moreover, the fact that we need to estimate the number and proportion of three separate MCSU classes indicates that this is a structured output learning problem.

In the experiments, we use a dataset of 89 weakly annotated pairs of IF and HE images from eight tumours, where 16 pairs of images from two tumours are explored for training an MCSU candidate detector and classifier using the annotations provided by a non-expert based on the description of the N, CH and AH classes above. The remaining 73 pairs of images from six tumours are used for training and testing the proposed structured output learning models using the clinical annotations based on the number and proportion of N, CH and AH MCSUs. Based on a leave-one-tumour-out cross validation experiment, we show a high correlation between the manual and automated annotations in terms of the number and proportion of MCSU classes for the four methodologies, but the quantitative results of methods (3) and (4) above are equally competitive and superior to the naive (1) and baseline (2).

1) Contributions: This paper extends the following papers by the same authors: 1) improvement of the generalisation ability of the convolutional neural network model (CNN) for the classification of MCSU candidates from [6]; 2) statistically significant improvement of the structured support vector machine model based on a latent flexible graphical model [7], where the unary potential function has been updated with the new CNN model from point (1) above; 3) statistically significant improvement of the pixel-wise labelling method based on fully convolutional neural network [8], where the input has been updated with the new CNN from point (1) above; 4) implementation of the naive method based on a majority classifier applied on the nodes of a fixed grid placed over the input images; and 5) implementation of the baseline method based on a structured output model that uses a fixed underlying graphical model. The naive and baseline methods (items 4 and 5 above) have been implemented in order to provide a benchmark in the assessment of the proposed methodologies (items 2 and 3 above) based on flexible and latent structured learning [7] and deep learning [8]. The weakly annotated dataset used in this paper can be downloaded from `http://cs.adelaide.edu.au/~carneiro/humboldt/`.

II. Literature Review

This paper is focused on the training of structured output models using probabilistic graphical models [42] and deep learning models [20]. Such models have been successfully explored in several computer vision problems, such as semantic segmentation [27], [45], instance segmentation [37], human pose estimation [21], [41], [3], depth and normal estimation [10], multiple organ detection and segmentation from medical images [30], [43], [32], [50], [9], and text recognition [16].

The weakly annotated training set to fit probabilistic graphical models implies that the loss function must minimise an error computed from the weak annotation, which in this paper is the number of MCSUs classified as N, CH and AH. The minimisation of this error involves the use of a high-order loss function that in general does not decompose well over the model variables [40], [34], [36]. Recent attempts to solve similar problems rely on low-order loss functions that are easier to optimise, but may not represent well the high-level error to minimise [13], [25], [35]. Another natural consequence of such weak annotation in probabilistic graphical models is the need for an underlying latent graph [48], which has been explored in 3-D human pose estimation [15], [47] and weakly supervised semantic segmentation [25], [35], [47], [13]. Current structured output learning methods based on probabilistic graphical models that combine flexible latent graphs, weakly supervised training and high-order loss functions, like the approaches being proposed in this paper, have not been proposed, to the best of our knowledge.

Structured output models based on deep learning methods have been thoroughly explored recently [11], [22], [20], [9]. Similarly, weakly-supervised learning to fit such deep learning models has also been investigated in the field [28], [29], but the use of high-order loss has just recently been studied by Pathak et al. [31] in a work that was developed in parallel to our own work presented in this paper. Compared to these approaches, our proposal is novel in terms of the loss function that combines a semantic labelling low-order loss with a high-order loss that minimises the error computed from the weak annotation mentioned above. We show that this new loss function is critical for the effectiveness of our approach.

III. Dataset

The dataset used in this paper contains the images prepared by Maftei et al. [23], that were acquired from xenografted human squamous cell carcinoma lines of the head and neck (FaDu), transplanted subcutaneously into the right hind leg of mice. Tumor excision followed immediately upon animal sacrifice. Cryosections of each tumour were scanned and photographed using AxioVision 4.7 and the multidimensional and mosaix modules. The IF images have been formed using three stainings: Pimonidazole for hypoxia stain (green regions), CD31 for vessel stain (red regions), and Hoechst 33342 for perfusion stain (blue). The cover slip was then removed to stain the same slice with HE to enable the manual labelling of vital tumour regions, where the process of removing this cover slip can cause severe tearing and folding in HE images, as shown by the white regions inside the tumour in the HE image of Fig. 1. Maftei et al. [23] have produced 89 pairs of IF and HE images from eight tumours using the imaging process described above. Furthermore, Maftei et al. [23] have annotated each of the 89 images in terms of: 1) the number of normoxic, chronically hypoxic and acutely hypoxic MCSUs; and 2) the vital tumour region. It is worth noting that the location and individual classification of MCSUs are not available in the manual annotation by Maftei et al. [23], and the vital tumour region generally contains necrotic regions given that it is inaccurately annotated.

The IF and HE images have been registered [33] and down-sampled such that the largest size (between vertical and horizontal) is 1024 pixels, which means that the resolution is approximately 10μm per pixel. It is important to mention that the deformable registration proposed by Peng et al. [33] has been tested in a subset comprising 26 images from the dataset described below in Sec. V, by computing the alignment error between ten pairs of anatomical landmarks. The mean error

achieved was 2.5 pixels (median of 2.1 pixels) - this error was found to be significantly smaller than the ones produced by competing methods (please see [33] for more details).

IV. Methodology

In this section, we first provide a formal definition of the dataset used, followed by the detection and classification of MCSU candidates into normoxia (N), chronic hypoxia (CH), acute hypoxia (AH) and Necrosis (Ne) using the HE and IF images acquired from a histological specimen of a tumour. Then, we explain the proposed naive method (referred to as NAIVE) that is based on a majority classifier applied on the nodes of a fixed grid placed over the input images. We then introduce the baseline method based on rigid structured output model trained with structured support vector machine (SSVM) [42] that is named RSSVM, followed by an introduction to the flexible latent structured output model, also trained with SSVM, and referred to as FLSSVM. Finally, we explain the structured output deep learning model, which is named DCNN.

A. Formal Dataset Definition

The dataset introduced in Sec. III is formally defined by $\mathcal{D} = \{\mathbf{x}_n, \mathbf{v}_n, \mathbf{y}_n\}_{n=1}^{N}$, with $\mathbf{x} = \{\mathbf{x}^{(\text{IF})}, \mathbf{x}^{(\text{HE})}\}$ denoting the input IF and HE images, where $\mathbf{x}^{(\text{IF})}, \mathbf{x}^{(\text{HE})} : \Omega \rightarrow [0,1]^3$ ($\Omega \in \mathbb{R}^2$ denotes the image lattice), $\mathbf{v} : \Omega \rightarrow \{0,1\}$ representing a segmentation map of the image regions that contain vital tumour tissue, and $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{N}^3$ denoting the annotation comprising the number of N, CH and AH MCSUs. It is worth noting that the vital tumour segmentation map $\mathbf{v}$ provides an inaccurate labelling of vital region of the tumour, so it is likely that the necrotic (Ne) tumour region will have to be analysed. For this reason, we include the class Ne as a possible class for a detected MCSU candidate, even though Ne is not part of the manual annotation $\mathbf{y}$.

B. MCSU Candidate Detection and Classification

The goal of the proposed system is to detect and classify MCSUs. An MCSU is defined by a box of size $200 \times 200\mu m$ centred at a micro-vessel [23]. As shown in Fig. 2, a micro-vessel is composed of a cluster of micro-vessel pixels that can be easily detected by thresholding the red channel of the IF image. Hereafter, we refer to such micro-vessel pixels as MCSU candidates, which are detected with:

$$\mathbf{t}(i) = \begin{cases} 1 & \text{, if } \mathbf{x}^{(\text{IF,RED})}(i) > \gamma, \\ 0 & \text{, otherwise.} \end{cases} , \tag{1}$$

where $\mathbf{t} : \Omega \rightarrow \{0,1\}$ represents the MCSU candidate map, and $\mathbf{x}^{(\text{IF,RED})}$ denotes the red channel of the IF image (the yellow dots of the first image of the block in Fig. 5-(b) represent the MCSU candidates).

Each region of size $200 \times 200\mu m$ centred at MCSU candidates defined in (1) can be classified into one of the four classes defined in Fig. 3. Given that such annotation is unavailable, we use the sketch representation in Fig. 3 to annotate MCSU candidates into one of the four classes: N, CH, AH, Ne. The annotation process consists of randomly selecting MCSU candidates, cropping a region of size $200 \times 200\mu m$ around them, and request the annotation by an non-expert. This annotation results in the dataset $\mathcal{D}_c = \{\mathbf{x}_n(i), c_n(i)\}_{n\in\{1,\dots,N\}, i\in\mathcal{I}_n}$, where n indexes one of the N images, $\mathcal{I}_n = \{i \in \Omega | \mathbf{t}_n(i) = 1\}$ represents a set of MCSU candidates in image n, $\mathbf{x}(i) = \{\mathbf{x}^{(\text{IF})}(i), \mathbf{x}^{(\text{HE})}(i)\}$ denotes the IF and HE image regions of size $200 \times 200\mu m$ centred at i, and $c_n(i) \in \{1,2,3,4\}$ representing the classes N, CH, AH, and Ne, respectively. Using this annotation, we train the following multi-class classifiers: 1) Adaboost [51], 2) linear SVM (LSVM) [42], 3) random forest (RF) [5], and 4) convolutional neural networks (CNN) [18]. These four classifiers are chosen based on their superior performances presented in a recent study [12]. Moreover, the use of several classifiers has the potential to increase the classification robustness [6], particularly given the unreliability of the annotations in $\mathcal{D}_c$. Figure 4 shows the Adaboost, LSVM and RF input features, which are represented by a set of three radial histograms from the RGB channels of the IF and HE images, and the CNN input features that are simply the RGB values from $\mathbf{x}_n(i)$. This training process generates the following $K = 4$ classifiers:

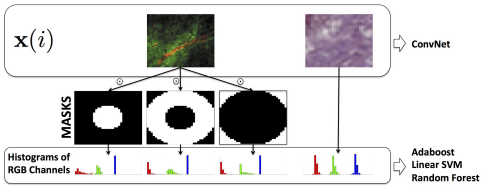


Fig. 4. Features used for the MCSU candidate classification.

$$\{P^{(k)}(c|\mathbf{x}(i), \theta^{(1,k)})\}_{k=1}^{K}, \tag{2}$$

where $k = 1$ denotes the Adaboost classifier, $k = 2$ represents LSVM, $k = 3$ means RF, and $k = 4$ is the CNN, $P^{(k)}(c|\mathbf{x}(i), \theta^{(1,k)}) \in [0,1]$ and $\sum_c P^{(k)}(c|\mathbf{x}(i), \theta^{(1,k)}) = 1$. The classification of $\mathbf{x}$ produces $\mathbf{r} : \Omega \rightarrow \mathbb{R}^{4\times K}$, defined by $\mathbf{r}(i) = \left[\mathbf{t}(i) \times P^{(k)}(c|\mathbf{x}(i), \theta^{(1,k)})\right]_{c\in\{1,2,3,4\}, k\in\{1,2,3,4\}}$,representing the probability of each of the $K = 4$ classifiers with respect to each of the four classes (N, CH, AH, and Ne) at position $i \in \Omega$.

C. Naive Baseline (NAIVE)

We first formulate a naive baseline method that relies solely on the non-expert annotation to train local detectors, defined in (2), and an underlying rigid grid of points represented by $\mathbf{g} : \Omega \rightarrow \{0,1\}$, where $\mathbf{g}(i) = 1$ if $i \in \Omega$ is located at places at the intersection of multiples of 200μm in the horizontal and vertical directions and $\mathbf{g}(i) = 0$, otherwise. The set of grid points forming MCSUs is represented by $\mathcal{V} = \{i \in \Omega | \mathbf{g}(i) = 1 \text{ AND } \mathbf{t}(i) = 1\}$, where $\mathbf{t}(.)$ is defined in (1). The naive MCSU classification at $v \in \mathcal{V}$ is then defined by:

$$n_v = \text{mode}\left(\{\arg\max_{c\in\{1,2,3,4\}} P^{(k)}(c|\mathbf{x}(i), \theta^{(1,k)})\}_{k=1}^{K}\right), \tag{3}$$

where mode(.) denotes the mode operator, and $P^{(k)}(c|\mathbf{x}(i), \theta^{(1,k)})$ is one of the $K = 4$ classifiers,

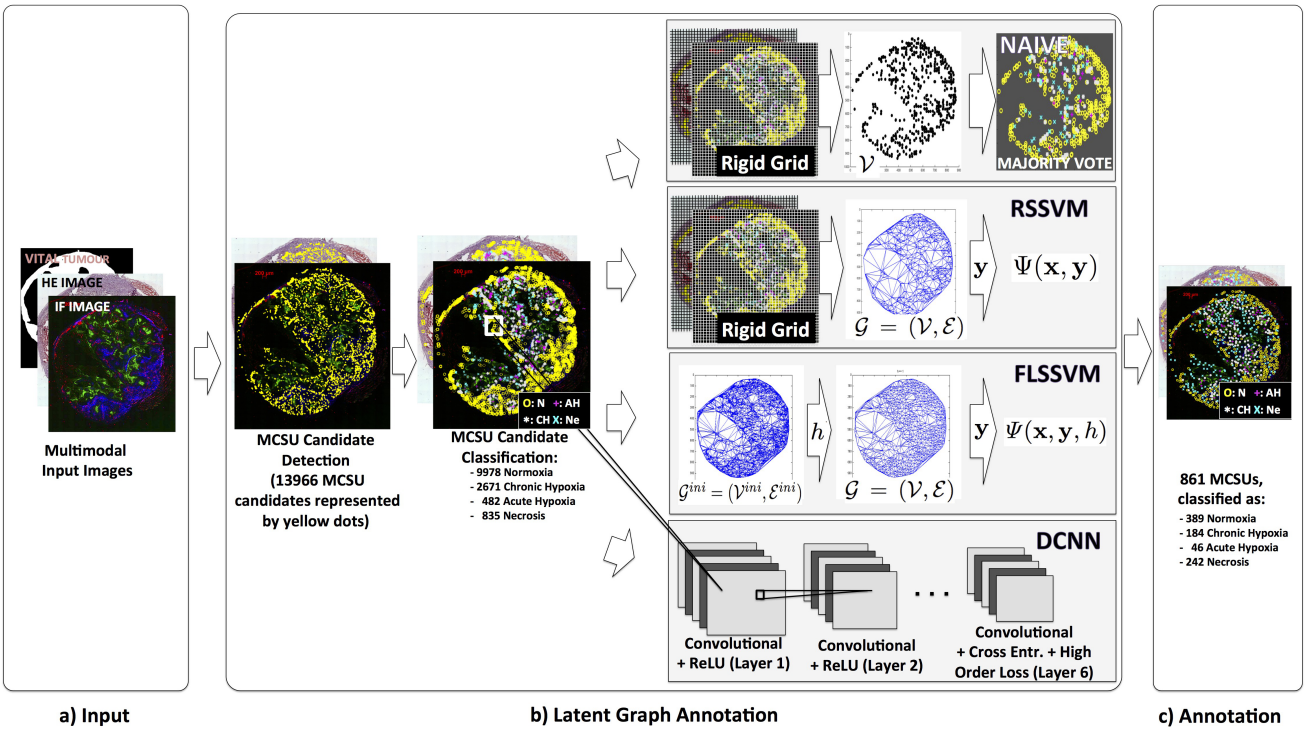


Fig. 5. The methodologies proposed in this paper receive as input the IF and HE images (a), from which micro-vessel pixels are detected and classified (first two frames in (b)). Then for the FLSSVM, the graph $\mathcal{G}$ is built and labelled using the initial graph $\mathcal{G}^{ini}$ in order to represent the MCSUs and form $\Psi(.)$ for (7) and (18). For DCNN, a series of convolutional layers applied to the micro-vessel pixel classification images produce a final map containing the MCSUs and their classes. From the outputs of FLSSVM and DCNN, it is trivial to obtain the final annotation in (c). The naive baseline based on the majority vote of the rigid grid nodes of the set $\mathcal{V}$, defined in Sec. IV-C, is represented in the box NAIVE, and the more sophisticated baseline based on probabilistic graphical model, but using a rigid grid, is depicted in the bos RSSVM. This figure is better visualised with an electronic reader reader - please zoom in the IF/HE images to notice the MCSU annotations.

defined in (2). The annotation for the image is then obtained by summing the results produced in (3), as follows:

$$\mathbf{y}^* = \left[\sum_{v \in \mathcal{V}} \delta(n_v - 1), \sum_{v \in \mathcal{V}} \delta(n_v - 2), \sum_{v \in \mathcal{V}} \delta(n_v - 3) \right], \quad (4)$$

where $\delta(.)$ represents the Dirac delta function. Note that this naive baseline method emulates the manual process of annotating these microscopic images that consists of taking the IF and HE images and scanning them in steps of 200μm in horizontal and vertical directions in order to detect and classify MCSUs - (see NAIVE box of Fig. 5-(b)).

D. Rigid Structure Support Vector Machine (RSSVM)

A more sophisticated baseline can be formed using a probabilistic graphical model with the goal of building a latent graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that represents the spatial distribution and classification of MCSUs, and use this graph in a structured SVM model that is learned using a high-order loss function. The graph $\mathcal{G}$ is formed by the nodes of the set $\mathcal{V}$, defined in Sec. IV-C, and the graph edges $\mathcal{E}$ are obtained with Delaunay triangulation (see middle of the RSSVM box of Fig. 5-(b)).

The learning and inference for RSSVM uses the feature vector $\Psi(\mathbf{x}, \mathbf{y})$ (right of the RSSVM box of Fig. 5-(b)) that is formed by labelling of the graph $\mathcal{G}$ using the annotation $\mathbf{y}$ and input image $\mathbf{x}$, as follows:

$$\begin{aligned} \underset{\mathbf{M}}{\text{minimise}} \quad & -\|\mathbf{M} \odot \mathbf{P}\|_F^2 + \sum_{c=1}^{3} \left(\mathbf{y}(c) - \|\mathbf{M} \odot \mathbf{E}_c\|_F^2\right)^2 \\ \text{subject to} \quad & \mathbf{1}_4^\top \mathbf{M} = \mathbf{1}_{|\mathcal{V}|}^\top, \ \mathbf{M} \in \{0,1\}^{4 \times |\mathcal{V}|}, \end{aligned} \quad (5)$$

where $\mathbf{P} \in \mathbb{R}^{4 \times |\mathcal{V}|}$, with

$$\mathbf{P}(c, v) = \prod_{k=1}^{K} \prod_{i \in \mathcal{V}} P^{(k)}(c | \mathbf{x}(i), \theta^{(1,k)}) \quad (6)$$

for $c \in \{1, 2, 3, 4\}$, $\mathbf{E}_1 = [\mathbf{1}_{|\mathcal{V}|}, \mathbf{0}_{|\mathcal{V}|}, \mathbf{0}_{|\mathcal{V}|}, \mathbf{0}_{|\mathcal{V}|}]^\top \in \{0,1\}^{4 \times |\mathcal{V}|}$ denotes a matrix with ones in first row and zeros elsewhere (similarly for $c = 2, 3$ with ones in rows 2 and 3), $\mathbf{1}_N$ and $\mathbf{0}_N$ represent column vector of ones or zeros of size N , $\|.\|_F$ denotes the Frobenius norm, $\odot$ represents the Hadamard product, and the summation varies from 1 to 3 given that $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{N}^3$ has the annotation for three classes: N, CH and AH. The optimisation in (5) maximises the probability of label assignment and minimises the difference between the number of each of the MCSU classes in $\mathbf{M}$ and in $\mathbf{y}$. The original integer programming in (5) is relaxed with $\mathbf{M} \in [0, 1]$ in order to make the optimisation feasible. The output $\mathbf{M}$ in (5) is then used for labelling the graph with $m_v(\mathbf{y}) = \arg\max_{c \in \{1,\ldots,4\}} \mathbf{M}(c, v)$ for each node $v \in \mathcal{V}$.

The RSSVM model is based on a linear structured support vector machine introduced by Szummer et al. [39], where inference is denoted by:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}\in\mathcal{Y}} \mathbf{w}^\top \Psi(\mathbf{x},\mathbf{y}), \tag{7}$$

with $\Psi(\mathbf{x},\mathbf{y})$ representing the unary and binary potential functions defined based on the graph labels $m_v(.)$ from (5) and the K classifiers $\{P^{(k)}(c|\mathbf{x}(i),\theta^{(1,k)})\}k = 1^K$ from (2). The details of the inference in (7) and learning of $\mathbf{w}$ are defined in Appendix A.

E. Flexible Latent Structure Support Vector Machine (FLSSVM)

The FLSSVM formulation extends the baseline method RSSVM (Sec. IV-D) by integrating a variable that can change the structure of graph $\mathbf{G}$ in terms of the number and location of nodes and set of edges, which increases the flexibility of the model. This means that graph structure becomes a hidden variable in a latent structured SVM model that is learned using the same high-order loss function defined in Sec. IV-D. The estimation of $\mathcal{G}$ uses the detected MCSUs provided by the map $\mathbf{t}$ from (1), which form the initial graph $\mathcal{G}^{ini} = (\mathcal{V}^{ini},\mathcal{E}^{ini})$, with nodes $v \in \mathcal{V}^{ini}$ labelled with position $i_v \in \mathbb{R}^2$ (where $\mathbf{t}(i_v) = 1$), and classification result $\mathbf{r}_v = [P^{(k)}(c_v|\mathbf{x},\theta^{(1,k)})]_{c_v\in\{1,...,4\},k\in\{1,...,K\}} \in \mathbb{R}^{4\times K}$, and the edges $\mathcal{E}^{ini}$ defined by Delaunay triangulation (leftmost image in FLSSVM box from Fig. 5-(b)). The structure of $\mathcal{G}$ is estimated using the minimum spanning tree (MST) clustering [14] over $\mathcal{G}^{ini}$, with the edge weight between nodes v and t (where $v,t \in \mathcal{V}^{ini}$) defined by $\|i_v - i_t\| \times \|\mathbf{r}_v - \mathbf{r}_t\|$. The idea of MST clustering is to merge MCSU candidates into clusters $\{\mathcal{C}_1,...,\mathcal{C}_{|\mathcal{V}|}\}$, where $\mathcal{C}_v \subset \mathcal{V}^{ini}$ represents an MCSU. The main parameter controlling the result of the MST clustering is h that constrains the size of each MCSU denoted by $\mathcal{C}_i$. Specifically, the size of an MCSU is constrained by $h \times 200\mu\text{m} \geq \max_{v,t\in\mathcal{C}_i} \|i_v - i_t\|$, where $h \in [0.5, 2]$ (note that h around 1 is related to the definition that an MCSU has a diameter of around $200\mu m$). Finally, the graph $\mathcal{G}$ has nodes $v \in \mathcal{V}$ formed by the clusters $\{\mathcal{C}_v\}_{v=1}^{|\mathcal{V}|}$, with the position of each node v computed from the centroid of the nodes $t \in \mathcal{C}_v$, and edges in $\mathcal{E}$ estimated with Delaunay triangulation (middle of the FLSSVM box of Fig. 5-(b)).

The inference used in the FLSSVM model is similar to the one in RSSVM, but with the introduction of the latent variable h, as follows:

$$(\mathbf{y}^*, h^*) = \arg\max_{\mathbf{y}\in\mathcal{Y},h\in\mathcal{H}} \mathbf{w}^\top \Psi(\mathbf{x},\mathbf{y},h), \tag{8}$$

where $\Psi(\mathbf{x},\mathbf{y},h)$ has the same definition as $\Psi(\mathbf{x},\mathbf{y})$ in (15). The learning of $\mathbf{w}$ in (8) follows the latent structured support vector machine introduced by Kumar [19], and is defined in detail in Appendix B.

F. Deep Convolutional Neural Network (DCNN)

The DCNN model consists of a fully convolutional neural network [22] that uses as input the map $\mathbf{r}$ defined in (2) that includes the results of $K = 4$ classifiers for each of the four classes, plus an additional background class, with probability map defined by $1 - \mathbf{t}$, where $\mathbf{t}$ is the MCSU candidate map defined in (1). This background class is needed because we minimise an objective function (described below) that uses a cross-entropy loss function that needs the four original classes (N, CH, AH, Ne) and the background class for regions without MCSUs. More specifically, the input is represented by $K\times 5 = 20$ channels, defined by,

$$\mathbf{p}_c^{(k)}(i) = \begin{cases} P^{(k)}(c|\mathbf{x}(i),\theta^{(1,k)}) & \text{if } \mathbf{t}(i) = 1 \\ 0 & \text{if } \mathbf{t}(i) = 0 \end{cases} \tag{9}$$

for $c \in \{1,2,3,4\}$, where $k \in \{1,2,3,4\}$ represents the classifier index, and $\mathbf{t}(i) = 1$ indicates an MCSU detection at location $i \in \Omega$. The background class is defined for $c = 0$ in (9) as $\mathbf{p}_0^{(k)}(i) = 1 - \mathbf{t}(i)$. The output of the DCNN consists of five binary maps $\mathbf{o}_c : \Omega \to \{0,1\}$, where $c \in \{1,...,4\}$ represents locations $i \in \Omega$ containing an MCSU classified as N, CH, AH or Ne, and $c = 0$ denotes regions without MCSUs (i.e., background). From these binary maps, it is possible to compute the number of MCSUs classified as N, CH and AH. Given that the location and classification of MCSUs are not available from the training set, we use (5) to produce a proxy annotation $\mathbf{M}$ for the DCNN training, where the annotation at $i \in \Omega$ is defined by:

$$m(i) = \begin{cases} \arg\max_{c\in\{1,...,4\}} \mathbf{M}(c,v) & \text{, if } \exists v \in \mathcal{V} \text{ s.t. } i_v = i \\ 0 & \text{, otherwise} \end{cases}, \tag{10}$$

where the set of nodes $\mathcal{V}$ of graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ is estimated as explained in Sec. IV-E. An example using the IF and HE images in Fig. 5-(a) is displayed in Fig. 6, with inputs $\mathbf{p}_c^{(k)}$ (only one of the classifiers $k \in \{1,...,4\}$ i shown) and outputs for the DCNN, represented by five binary maps $\mathbf{m}_c : \Omega \to \{0,1\}$, where $\mathbf{m}_c(i) = 1$ if $m(i) = c$, and zero otherwise. The training of this DCNN model relies on the minimisation of the following loss function

$$\ell = \left(-\sum_{i\in\Omega} \left(\sum_{c=0}^{C} \delta(m(i)-c) \log \frac{\exp(\mathbf{W}_c^\top \mathbf{x}(i))}{\sum_{l=0}^{C} \exp(\mathbf{W}_l^\top \mathbf{x}(i))} \right) \right) + \left(\sum_{c=1}^{3} \left(\sum_{i\in\Omega} \delta(m(i)-c) - \sum_{i\in\Omega} \delta(\widehat{m}(i)-c) \right)^2 \right), \tag{11}$$

where $\mathbf{x}(i)$ represents the input from the second to last DCNN layer, the first term is the cross-entropy loss that uses the proxy annotation defined in (10), and the second term is the high-order error based on the squared difference between the number of MCSUs annotated and classified as N, CH and AH, which is based on the DCNN classification at image location $i \in \Omega$ represented by

$$\widehat{m}(i) = \arg\max_{c\in\{0,...,4\}} \frac{\exp(\mathbf{W}_c^\top \mathbf{x}(i))}{\sum_{l=0}^{C} \exp(\mathbf{W}_l \mathbf{x}(i))}. \tag{12}$$

The optimisation process to minimise the loss in (11) is based on stochastic gradient descent, which is problematic given the difficulty in computing the derivative of $\delta(\widehat{m}(i)-c)$. However, such derivative can be computed with an approximation based on a softmax function with a temperature parameter τ, defined by

$$\widetilde{\delta}(\widehat{m}(i)-c) = \frac{\exp\left(\frac{\mathbf{W}_c^\top \mathbf{x}(i)}{\tau}\right)}{\sum_{l=0}^{C} \exp\left(\frac{\mathbf{W}_l^\top \mathbf{x}(i)}{\tau}\right)}, \tag{13}$$

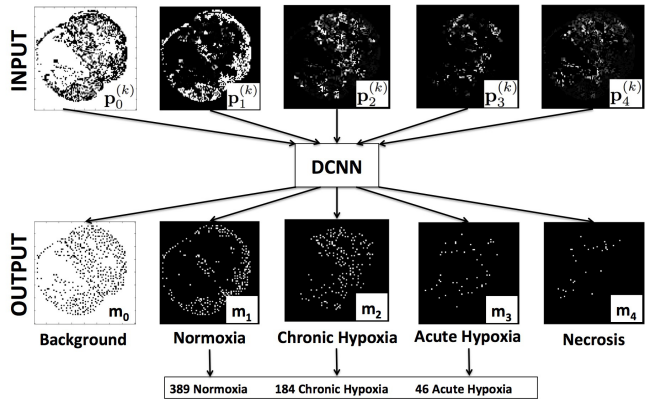


Fig. 6. Inputs and outputs for the DCNN model.

with $0 < \tau << 1$. With such approximation, we can compute the following derivative for the loss in (11):

$$\frac{\partial \ell}{\partial \mathbf{W}_j} = -\sum_{i\in\Omega} \mathbf{x}(i)\left(\delta(m(i)-j) - \frac{\exp(\mathbf{W}_j^\top \mathbf{x}(i))}{\sum_l \exp(\mathbf{W}_l^\top \mathbf{x}(i))}\right) + \sum_{i\in\Omega} 2\mathbf{x}(i)\left(\sum_{c=1}^{3}\left(\delta(m(i)-c) - \widetilde{\delta}(\widehat{m}(i)-c)\right)\times \left(\frac{\exp\left(\frac{\mathbf{W}_c^\top \mathbf{x}(i)}{\tau}\right)}{\sum_l \exp\left(\frac{\mathbf{W}_l^\top \mathbf{x}(i)}{\tau}\right)} - \delta(c-j)\right) \times \frac{\exp\left(\frac{\mathbf{W}_j^\top \mathbf{x}(i)}{\tau}\right)}{\sum_l \exp\left(\frac{\mathbf{W}_l^\top \mathbf{x}(i)}{\tau}\right)}\right). \quad (14)$$

V. Experimental Setup

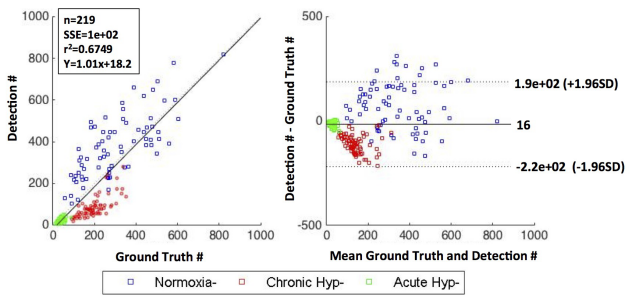
For the experiments, we use the 89 pairs of IF and HE images from eight tumours introduced in Sec. III. For the MCSU candidate detection we use a threshold value $\gamma = 0.1$ in (1), defined based on the assumption that micro-vessel pixels must have a red component of at least 0.1. Empirically, we have observed that the number of micro-vessel pixels selected with $\gamma = 0.1$ is roughly 10 times the number of actual MCSUs present in the same image, which reduces considerably the chances of missing true MCSUs.

For the MCSU candidate classification, we use 16 pairs of IF and HE images from two tumours, and the non-expert annotation was performed with an active learning scheme, where one image (out of 16) was randomly chosen, from which 500 MCSU candidates using the detection defined in (1) were manually annotated. Using these initial annotations, the four classifiers defined in (2) are initially trained and applied to a new set of 500 MCSU candidates, and new user annotation is requested for the MCSU candidates that presented a disagreement amongst the classifiers in terms of the classification result. This annotation process is repeated for the remaining 15 images, forming a dataset of 1000 annotated MCSU candidates for each of the 16 images. The accuracy of these classifiers are tested using a 2-fold cross validation experiment, where the dataset is divided into eight images for training, four for validation and four for testing (this division is done randomly), where the validation set is used to estimate the hyper-parameters of each classifier (i.e., the number of weak classifiers of Adaboost, C-value for LSVM, number and depth of trees in RF and number of layers, number of layers and filters per layer, and size of filters in CNN). The performance is measured by computing the error on the testing set (that contains 4000 samples from the four testing images and 1000 MCSU candidates per image): $\frac{1}{4000}\sum_{i=1}^{4000} 1-\delta(c_i - c_i^*)$, where i indexes the annotated MCSU candidates from the testing images, c_i is the manually annotated class of the i^{th} MCSU candidate, and $c_i^* = \arg\max_c P^{(k)}(c|\mathbf{x}(i), \theta^{(1,k)})$ from (2).

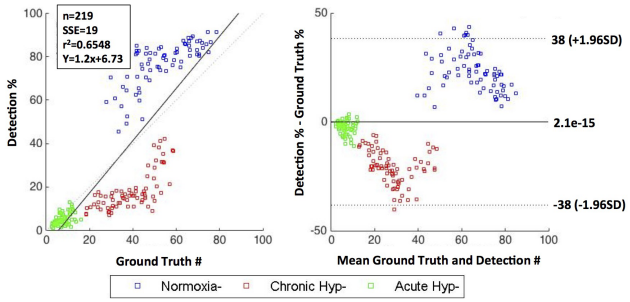
The quantitative assessment of NAIVE, RSSVM, FLSSVM and DCNN models is based on a six-fold cross validation experiment, where the IF and HE image pairs of five tumours are used for training the models and the images from the hold-out tumour for testing. For these experiments, the MCSU candidate detection uses $\gamma = 0.1$, and the MCSU candidate classifier is the one from the first fold of the cross validation experiment above. The NAIVE model requires no training (except for the original the $\{P^{(k)}(c|\mathbf{x}(i), \theta^{(1,k)})\}_{k=1}^{K}$ from (2)), and inference follows the majority classification of (4). For RSSVM, the inference to estimate $\mathbf{y}^*$ in (7) and the loss augmented inference in (18) to estimate $\widehat{\mathbf{y}}_n$ are based on graph cuts (alpha-expansion) [4]. Likewise for FLSSVM, the inference for $\mathbf{y}^*$ and h^* in (8) and the loss augmented inference in (20) for $\widehat{\mathbf{y}}_n$ and $\widehat{h}_n$ are also based on graph cuts (alpha-expansion) [4] with $h \in \mathcal{H} = \{0.5, 1, 1.5, 2\}$. For RSSVM and FLSSVM inferences, graph cuts produces a graph labelling, and we only take the number of normoxic, chronically hypoxic and acutely hypoxic MCSUs to build the vector $\widehat{\mathbf{y}} \in \mathbb{N}^3$ that is then used to build $\Psi(\mathbf{x}, \widehat{\mathbf{y}}, h)$ from the optimisation in (5). The training of the DCNN [44] uses the temperature parameter $\tau = 0.01$ in (14) and it runs for 100 epochs using mini-batches of size 10, learning rate 0.001, and momentum 0.9. The DCNN model used in this work has 6 convolutional layers with activation functions based on the rectified linear unit (ReLU) [26], except for the last layer, which uses the loss defined in (11), as depicted in Fig. 5. The input image has $4 \times 5 = 20$ channels with the five classes estimated by four classifiers, and is scaled to 100×100 pixels and normalised by mean subtraction (see Fig. 6). Stages 1-6 use: 1) 10 (5×5) filters, 2) 10 (5×5) filters, 3) 50 (5×5) filters, 4) 100 (5×5) filters, 5) 100 (5×5) filters, and 6) 5 (5×5) filters. The output has five channels (representing classes $\{0, ..., 4\}$) of size 80×80 (see Fig. 6). This quantitative assessment measures the correlation between the manual and estimated number and proportion of MCSU classes (N, CH and AH) using the six test sets (for the six-fold cross validation) with the Bland Altman plots [1] that display the number of samples, sum of squared error (SSE), Pearson r-value squared (r^2), and linear regression. We also report the inference running time using an un-optimised Matlab code running on a 2.3 GHz Intel Core i7 with 8GB of RAM and Nvidia GeForce 650M.

VI. Results

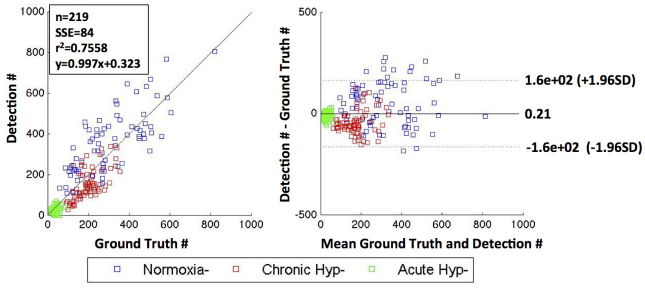
Table I shows the mean and standard deviation of the training and testing errors in the 2-fold cross validation test using the MCSU candidate Adaboost, RF, LSVM and CNN classifiers (2). We show the training and testing results of each classifier in order to assess their generalisation abilities,



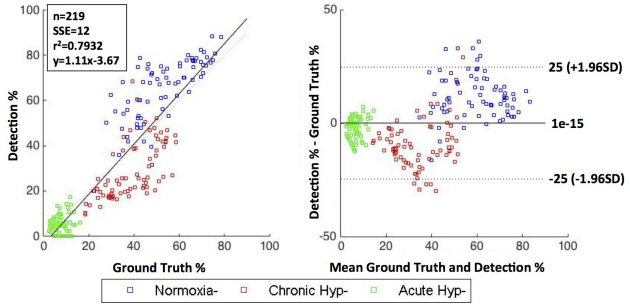
a) NAIVE (number of MCSU classes)



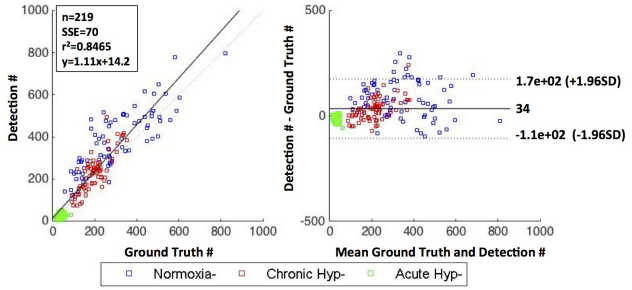
b) NAIVE (percentage of MCSU classes)



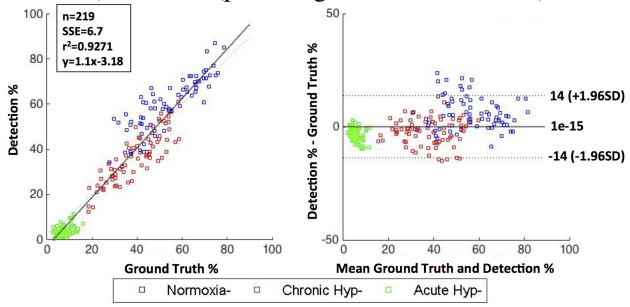
c) RSSVM (number of MCSU classes)



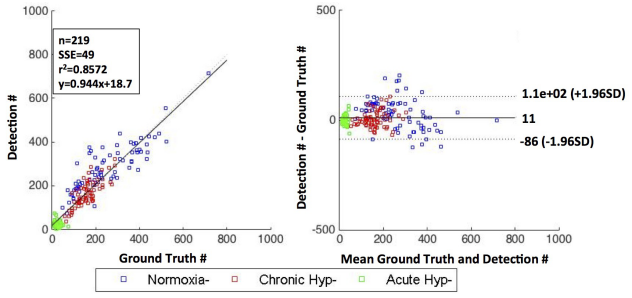
d) RSSVM (percentage of MCSU classes)



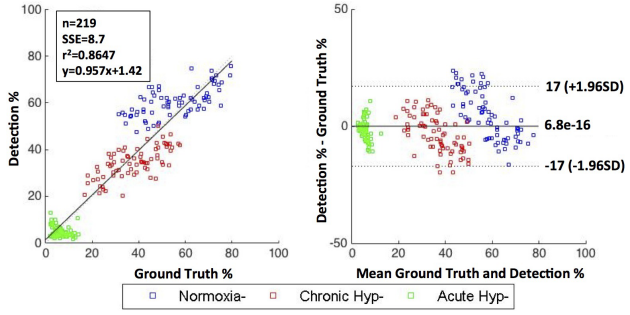
e) FLSSVM (number of MCSU classes)



f) FLSSVM (percentage of MCSU classes)



g) DCNN (number of MCSU classes)



h) DCNN (percentage of MCSU classes)

Fig. 7. Bland Altman graphs of MCSU classification in terms of the numbers (left) and proportion (right) of MCSU classes for the NAIVE (1,b), RSSVM (c,d), FLSSVM (e,f) and DCNN (g,h) models.

and we also show the relevant hyper-parameters (described in Sec. V) estimated for each classifier for the first fold of the cross validation process (the second fold shows similar results). The quantitative assessment of the proposed models NAIVE, RSSVM, FLSSVM and DCNN is shown in Figure 7 that displays the Bland Altman graphs of the number and proportion of MCSU classes. We have also run an additional experiment to assess the importance of the high-order loss function in the DCNN loss (11). Basically, we removed the high-order loss from (11), which means that the DCNN loss in this experiment consists only of the cross-entropy loss, and the results show that all MCSUs (in all test images) are classified as background (i.e., $c = 0$ in Sec. IV-F). This is a reasonable result because background is the most dominant label in the DCNN training.

Figure 8 shows the manual and estimated annotations of several test images produced by the proposed methods, allowing a qualitative visual comparison between them with respect to the number and proportion of MCSU classes and the visual distribution of MCSU classes in the image.

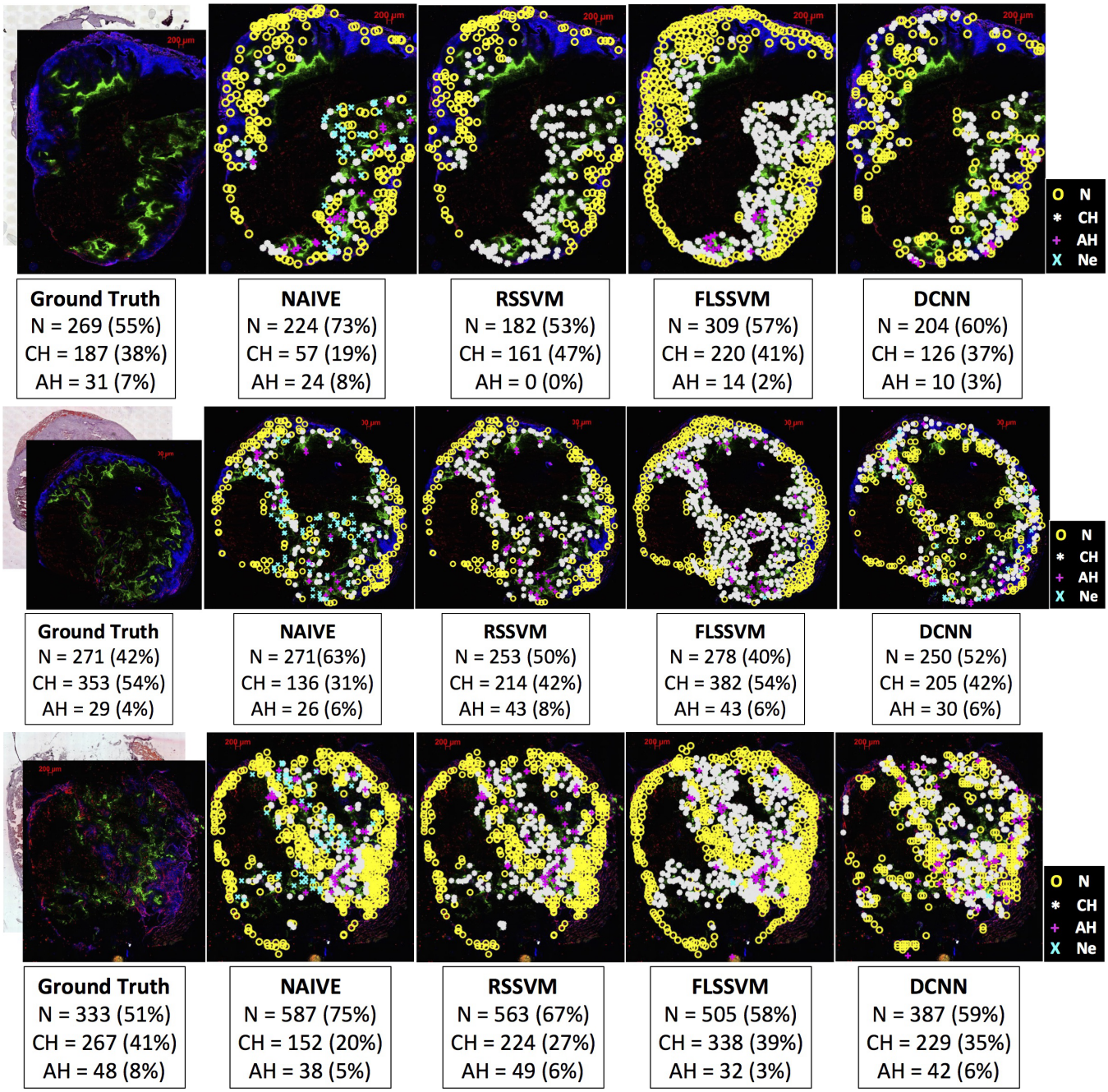


Fig. 8. Results of three different test images that show a qualitative comparison of NAIVE, RSSVM, FLSSVM, and DCNN in terms of the number and percentage of MCSU classes estimated by each model compared to the manual annotation (left image in each case). This figure is better visualised electronically - please zoom in the IF images to notice the MCSU annotations (note that we only show the results on the IF image, but the input consists of the IF and HE images, as shown on the leftmost column).

Finally, the inference running time of each stage of both methods are as follows (mean average from all test images): MCSU candidate detection (0.03s), MCSU candidate classification (157s), NAIVE - building set $\mathcal{V}$ (0.05s), RSSVM - from MCSU candidates to $\Psi(\mathbf{x}, \mathbf{y})$ (3s), FLSSVM - from MCSU candidates to $\Psi(\mathbf{x}, \mathbf{y}, h)$ (26s), NAIVE inference in (4) (0.005s) RSSVM inference in (7) (1.5s), FLSSVM inference in (8) (3.5s), and DCNN inference (0.35s). Thus, the running time for NAIVE is around 150.1s, RSSVM is 161.53s, for FLSSVM is 186.53s, and for DCNN is 157.38s.

VII. Discussion and Conclusion

The results presented in Fig. 7 show that all proposed methods produce accurate classification results for the automatic quantification of the number and proportion of MCSU classes from the HE and IF images, particularly considering the large correlation coefficients (r^2) and small errors (SSE) with respect to the manual annotations. Nevertheless, compared to the NAIVE baseline ($SSE = 100$ and $r^2 = 0.6749$ for the number of MCSU classes, and $SSE = 19$ and $r^2 = 0.6548$ for the percentage of MCSU classes) and the RSSVM baseline

TABLE I
MEAN AND STANDARD DEVIATION OF THE ERRORS PRODUCED BY THE MCSU CANDIDATE CLASSIFIERS IN THE 2-FOLD CROSS VALIDATION TEST (THIS RESULT UPDATES THE RESULTS FROM [6]). IN THE LAST COLUMN, WE SHOW THE HYPER-PARAMETERS ESTIMATED FOR EACH CLASSIFIER FOR THE FIRST FOLD OF THE CROSS VALIDATION PROCESS.

Method	Training	Testing	Hyper-parameters
Adaboost	0.1321 ± 0.0047	0.1510 ± 0.005	1000 WEAK CLASSIFIERS
Rand. Forest	0.0800 ± 0.0031	0.1298 ± 0.0013	100 TREES WITH DEPTH = 10
lin. SVM	0.1861 ± 0.0140	0.2178 ± 0.0284	C-VALUE = 0.1
CNN	0.0980 ± 0.0071	0.1714 ± 0.0049	1 LAYER W/ 200 1×1 FILTERS

($SSE = 84$ and $r^2 = 0.7558$ for the number of MCSU classes, and $SSE = 12$ and $r^2 = 0.7929$ for the percentage of MCSU classes), the results from FLSSVM ($SSE = 70$ and $r^2 = 0.8465$ for the number of MCSU classes, and $SSE = 6.7$ and $r^2 = 0.9271$ for the percentage of MCSU classes) and DCNN ($SSE = 49$ and $r^2 = 0.8572$ for the number of MCSU classes, and $SSE = 8.8$ and $r^2 = 0.8647$ for the percentage of MCSU classes) are superior, with FLSSVM presenting the best result in terms of the percentage of MCSU classes and DCNN with the best result for the number of MCSU classes. These results are better than our previously published results [8] because of the improved CNN MCSU candidate classifier, where we improved its generalisation ability with the use of dropout [38] in the training of the CNN model (see Tab. I). In particular, we measure the statistical significance of the new results of this paper with respect to the prior results in [7], [8] by comparing each of the N, CH and AH number and percentage estimates using the Wilcoxon signed-rank test. For the FLSSVM, out of the six results (three measurements for the number and for the percentage estimates), five are statistically significant, and for the DCNN, four are significant (assuming 5% significance level).

An interesting conclusion from the results above is that the increased sophistication of the models FLSSVM and DCNN, compared with NAIVE and RSSVM, provides considerable accuracy improvements, as demonstrated by the decreasing SSE value and increasing r^2 value. Another important observation is that the NAIVE model provides a quantitative assessment of the non-expert annotation, where it is worth noticing in Fig. 7 that NAIVE appears quite biased towards a larger number of normoxic and a smaller number of chronic hypoxic MCSUs. The use of the clinical annotation by RSSVM, FLSSVM and DCNN has fixed this bias present in the non-expert annotation.

Moreover, from the visual results in Fig. 8, we can speculate that RSSVM and FLSSVM (and NAIVE to a certain extent) produce results that are more likely to be visually correct. Although we do not have the manual annotation for the location and classification of MCSUs to quantitatively validate such a statement, we show some exemplary results in Fig. 8 for a qualitative demonstration. In particular, in all IF images of Fig. 8, large regions stained in red/blue are expected to be annotated with normoxia, which can be easily seen in the results from NAIVE, RSSVM and FLSSVM, but not from DCNN. Regions of IF images that contain transitions between blue to green should show chronically hypoxic MCSUs, which is the general result displayed by NAIVE, RSSVM and FLSSVM, but not by DCNN. Similarly, green regions in IF images must contain a large proportion of acutely hypoxic MCSUs, which is shown for NAIVE, RSSVM and FLSSVM, but not for DCNN. Finally, necrotic regions appear mostly in the boundaries of the vital tumour mask (generally represented by regions without any MCSU candidates in the images), which are correctly classified by NAIVE, RSSVM and FLSSVM, but not by DCNN. We believe that one of the issues causing the worse qualitative results by DCNN lies in the lack of a spatial prior for the MCSUs, such as the one used by the other models. Regardless of the qualitative analysis of the visual results presented above, it is important to note that the main clinical interest in this method is the final counting of normoxic, chronically hypoxic and acutely hypoxic MCSUs, which means that DCNN can be considered one of the two best methods (along with FLSSVM). It also is important to notice that when the DCNN is trained with cross-entropy loss only (i.e., without the high-order loss), then it is observed that all MCSUs are classified as background - this shows the importance of integrating the high-order loss into the DCNN training.

This paper addresses a problem that we believe is crucial for future medical image analysis applications, which is the implementation of systems with the exclusive use of the high-level annotations that are currently present in clinical datasets. Currently, the vast majority of systems in the field require the use of artificial and low-level annotations that are not present in clinical datasets. For instance, when classifying tumours, medical image analysis systems will generally produce a segmentation of the tumour that is used in the classification process. However, in order to design segmentation systems, we need artificial segmentation annotations that are not currently present in clinical datasets. This issue generates several problems: 1) small amount of annotations available for training the system, 2) unreliable systems due the usually large inter- and intra-user variability of such annotations, and 3) expensive annotation process. If we start to directly process medical data and their accompanying high-level annotations already present in clinical datasets, then we will be able to produce more robust systems given the large amount of data available in hospitals and clinics in a hopefully less expensive manner given that we will completely eliminate the artificial annotation process. This paper partially achieves such a goal given that we still use the artificial non-expert annotation for training the MCSU classifier, but the fact that we did not require clinical annotations in that stage helped reduce the annotation process cost. Nevertheless, a future plan for this work is the development of a methodology that no longer needs such non-expert annotation. We expect that with the availability of this dataset, other researchers in the field will try different methodologies for solving the challenging problem presented in this paper.

Finally, a potential criticism faced by this paper is the fact that the clinical usefulness of the method has not been tested given that we did not assess the accuracy of the automated MCSU classification in a clinical setting. It is worth noting that the accuracy of the automated MCSU classification has in fact been tested against the clinical annotation provided by Maftei et al. [23], consisting of the number of normoxic, chronically hypoxic and acutely hypoxic MCSUs. One can argue that the accuracy of this automated MCSU classification already shows the clinical relevance of our paper. It is also

important to emphasise that data on the aggressiveness of the individual tumours are not available for the dataset used in this work [23], so it is not possible to assess clinical outcome with this dataset. However, large-scale clinical studies of hypoxia-modified therapies (that can produce such annotations on treatment prognosis) are planned to be conducted in the future, but one of the major impediments for the validation of such studies is the availability of a standardised way of measuring tumour hypoxia that is practical and effective. We believe that the tool developed in this paper has the potential to address this issue and become crucial in validating future hypoxia-modified cancer therapies.

REFERENCES

[1] D. G. Altman and J. M. Bland. Measurement in medicine: the analysis of method comparison studies. *The statistician*, pages 307–317, 1983.

[2] C. Bayer, K. Shi, S. T. Astner, C.-A. Maftei, and P. Vaupel. Acute versus chronic hypoxia: why a simplified classification is simply not enough. *International Journal of Radiation Oncology* Biology* Physics*, 80(4):965–968, 2011.

[3] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures for multiple human pose estimation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1669–1676. IEEE, 2014.

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(11):1222–1239, 2001.

[5] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] G. Carneiro, T. Peng, C. Bayer, and N. Navab. Automatic detection of necrosis, normoxia and hypoxia in tumors from multimodal cytological images. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 2429–2433. IEEE, 2015.

[7] G. Carneiro, T. Peng, C. Bayer, and N. Navab. Flexible and latent structured output learning. In *International Workshop on Machine Learning in Medical Imaging*, pages 220–228. Springer, 2015.

[8] G. Carneiro, T. Peng, C. Bayer, and N. Navab. Weakly-supervised structured output learning with flexible and latent graphs using high-order loss functions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 648–656, 2015.

[9] N. Dhungel, G. Carneiro, and A. P. Bradley. Deep learning and structured prediction for the segmentation of mass in mammograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–612. Springer, 2015.

[10] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *arXiv preprint arXiv:1411.4734*, 2014.

[11] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2013.

[12] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

[13] L. Fiaschi, F. Diego, K. Gregor, M. Schiegg, U. Koethe, M. Zlatic, and F. A. Hamprecht. Tracking indistinguishable translucent objects over time using weakly supervised structured learning. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2736–2743. IEEE, 2014.

[14] O. Grygorash, Y. Zhou, and Z. Jorgensen. Minimum spanning tree based clustering algorithms. In *Tools with Artificial Intelligence, 2006. ICTAI'06. 18th IEEE International Conference on*, pages 73–81. IEEE, 2006.

[15] C. Ionescu, F. Li, and C. Sminchisescu. Latent structured models for human pose estimation. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2220–2227. IEEE, 2011.

[16] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Deep structured output learning for unconstrained text recognition. *arXiv preprint arXiv:1412.5903*, 2014.

[17] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[19] M. P. Kumar. *Weakly Supervised Learning for Structured Output Prediction*. PhD thesis, Ecole Normale Supérieure de Cachan, 2014.

[20] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[21] S. Li, W. Zhang, and A. B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2848–2856, 2015.

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[23] C.-A. Maftei, C. Bayer, K. Shi, S. T. Astner, and P. Vaupel. Changes in the fraction of total hypoxia and hypoxia subtypes in human squamous cell carcinomas upon fractionated irradiation: evaluation using pattern recognition in microcirculatory supply units. *Radiotherapy and Oncology*, 101(1):209–216, 2011.

[24] C.-A. Maftei, C. Bayer, K. Shi, S. T. Astner, and P. Vaupel. Quantitative assessment of hypoxia subtypes in microcirculatory supply units of malignant tumors using (immuno-) fluorescence techniques. *Strahlentherapie und Onkologie*, 187(4):260–266, 2011.

[25] D. Mahapatra, A. Vezhnevets, P. J. Schuffler, J. A. Tielbeek, F. M. Vos, and J. M. Buhmann. Weakly supervised semantic segmentation of crohn's disease tissues from abdominal mri. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 844–847. IEEE, 2013.

[26] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.

[27] S. Nowozin, P. V. Gehler, and C. H. Lampert. On parameter learning in crf-based approaches to object class image segmentation. In *Computer Vision–ECCV 2010*, pages 98–111. Springer, 2010.

[28] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015.

[29] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. *arXiv preprint arXiv:1502.02734*, 2015.

[30] B. Patenaude, S. M. Smith, D. N. Kennedy, and M. Jenkinson. A bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*, 56(3):907–922, 2011.

[31] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1796–1804, 2015.

[32] O. Pauly, B. Glocker, A. Criminisi, D. Mateus, A. M. Möller, S. Nekolla, and N. Navab. Fast multiple organ detection and localization in whole-body mr dixon sequences. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*, pages 239–247. Springer, 2011.

[33] T. Peng, M. Yigitsoy, A. Eslami, C. Bayer, and N. Navab. Deformable registration of multi-modal microscopic images using a pyramidal interactive registration-learning methodology. In *Biomedical Image Registration*, pages 144–153. Springer, 2014.

[34] P. Pletscher and P. Kohli. Learning low-order models for enforcing high-order statistics. In *International Conference on Artificial Intelligence and Statistics*, pages 886–894, 2012.

[35] G. Quellec, M. Laniard, G. Cazuguel, M. D. Abràmoff, B. Cochener, and C. Roux. Weakly supervised classification of medical images. In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, pages 110–113. IEEE, 2012.

[36] M. Ranjbar, A. Vahdat, and G. Mori. Complex loss optimization via dual decomposition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2304–2311. IEEE, 2012.

[37] N. Silberman, D. Sontag, and R. Fergus. Instance segmentation of indoor scenes using a coverage loss. In *Computer Vision–ECCV 2014*, pages 616–631. Springer, 2014.

[38] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[39] M. Szummer, P. Kohli, and D. Hoiem. Learning crfs using graph cuts. In *Computer Vision–ECCV 2008*, pages 582–595. Springer, 2008.

[40] D. Tarlow and R. S. Zemel. Structured output learning with high order loss functions. In *International Conference on Artificial Intelligence and Statistics*, pages 1212–1220, 2012.

[41] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*, pages 1799–1807, 2014.

[42] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, page 104. ACM, 2004.

[43] Z. Tu, K. L. Narr, P. Dollár, I. Dinov, P. M. Thompson, and A. W. Toga. Brain anatomical structure segmentation by hybrid discrimi-

native/generative models. *Medical Imaging, IEEE Transactions on*, 27(4):495–508, 2008.
[44] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. *CoRR*, 2014.
[45] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 845–852. IEEE, 2012.
[46] J. C. Walsh, A. Lebedev, E. Aten, K. Madsen, L. Marciano, and H. C. Kolb. The clinical importance of assessing tumor hypoxia: relationship of tumor hypoxia to prognosis and therapeutic opportunities. *Antioxidants & redox signaling*, 21(10):1516–1554, 2014.
[47] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE, 2011.
[48] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1169–1176. ACM, 2009.
[49] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural computation*, 15(4):915–936, 2003.
[50] S. K. Zhou. Discriminative anatomy detection: Classification vs regression. *Pattern Recognition Letters*, 43:25–38, 2014.
[51] J. Zhu, H. Zou, S. Rosset, and T. Hastie. Multi-class adaboost. *Statistics and Its*, 2009.

APPENDIX A
INFERENCE AND LEARNING OF RSSVM

The inference of the RSSVM model introduced in (7) [39] depends on $\Phi(.)$, defined as follows:

$$\Psi(\mathbf{x},\mathbf{y}) = [f_1^{(1,1)},...,f_4^{(1,1)},...,f_1^{(1,K)},...,f_4^{(1,K)},f^{(2,1)},...,f^{(2,L)}]. \tag{15}$$

The unary features in (15) are defined as

$$f_c^{(1,k)} = \sum_{v\in\mathcal{V}} \delta(m_v(\mathbf{y})-c)\phi^{(1,k)}(c,\mathbf{x};\theta^{(1,k)}), \tag{16}$$

where $m_v(\mathbf{y}) \in \{1,2,3,4\}$ denotes the label of node $v \in \mathcal{V}$ from (5), and $k \in \{1,...,K\}$ with $\phi^{(1,k)}(c,\mathbf{x};\theta^{(1,k)}) = -\log P^{(k)}(c|\mathbf{x}_v,\theta^{(1,k)})$ representing the k^{th} unary potential function in (2) that computes the negative log probability of assigning class c to node v. The binary features in (15) are defined as

$$f^{(2,l)} = \sum_{(v,t)\in\mathcal{E}} \phi^{(2,l)}(c_v,c_t,\mathbf{x};\theta^{(2,l)}), \tag{17}$$

where $l \in \{1,...,L\}$, $\phi^{(2,1)}(c_v,c_t,\mathbf{x};\theta^{(2,l)}) = (1-\delta(c_v-c_t))g(c_v,c_t,\mathbf{x};\theta^{(2,l)})$ represents the binary potential function that estimates the compatibility between nodes v and t if their labels are different. In particular, the binary potential functions used are the following: 1) $g(c_v,c_t,\mathbf{x};\theta^{(2,1)}) = 1/\|i_v - i_t\|$ (with $i_v \in \Omega$ denoting the position of node v in the image), 2) $g(c_v,c_t,\mathbf{x};\theta^{(2,2)}) = 1/\|\mathbf{r}_v - \mathbf{r}_t\|$ (with $\mathbf{r}_v$ defined in Sec. IV-B as $[P^{(k)}(c_v|\mathbf{x},\theta^{(1,k)})]_{c_v\in\{1,...,4\},k\in\{1,...,K\}} \in \mathbb{R}^{4\times K}$ representing a vector of classifier responses for node v); and 3) $g(c_v,c_t,\mathbf{x};\theta^{(2,3)}) = 1/(\|i_v - i_t\| \times \|\mathbf{r}_v - \mathbf{r}_t\|)$.

The learning of model RSSVM is defined by [39]:

$$\begin{aligned} \underset{\mathbf{w},\{\xi_n\}_{n=1}^N}{\text{minimise}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{n=1}^N \xi_n \\ \text{subject to} \quad & \Big(\mathbf{w}^\top \Psi(\mathbf{x}_n,\mathbf{y}_n)\Big) - \\ & \Big(\mathbf{w}^\top \Psi(\mathbf{x}_n,\widehat{\mathbf{y}}_n)\Big) \geq \Delta(\mathbf{y}_n,\widehat{\mathbf{y}}_n) - \xi_n \\ & \xi_n \geq 0, \forall \widehat{\mathbf{y}}_n \in \mathcal{Y}, n = 1,...,N, \end{aligned} \tag{18}$$

where $\{\xi_n\}_{n=1}^N$ denotes the slack variables and $\Delta(\mathbf{y}_n,\widehat{\mathbf{y}}_n) = \sum_{c=1}^3 |\mathbf{y}_n(c) - \widehat{\mathbf{y}}_n(c)|$ computes the high-order loss between the manual and estimate annotations $\mathbf{y}_n$ and $\widehat{\mathbf{y}}_n$, respectively.

The estimation of $\mathbf{w}$ in (18) is performed using the cutting plane algorithm [17] that iteratively solves a loss augmented inference problem by inserting a new constraint in the set of most violated constraints with $\widehat{\mathbf{y}}_n = \arg\max_{\mathbf{y}\in\mathcal{Y}} \Delta(\mathbf{y}_n,\mathbf{y}) + \mathbf{w}^\top\Psi(\mathbf{x},\mathbf{y})$. The loss augmented inference and the inference in (7) are based on graph cut (using alpha expansion) [4], where the high-order loss function, defined by $\Delta(\mathbf{y}_n,\widehat{\mathbf{y}}_n)$ is integrated into graph cuts using the decomposition proposed by Pletscher and Kohli [34]. This decomposition consists of using of a lower envelope of the high-order loss function with the integration of auxiliary variables $\{z_c\}_{c=1}^3$, which forms the following loss augmented inference:

$$\begin{aligned} \widehat{\mathbf{y}}_n = \arg \min_{\mathbf{y}\in\mathcal{Y},z_1,z_2,z_3\in\{0,1\}} & -\mathbf{w}^\top\Psi(\mathbf{x},\mathbf{y}) + \\ \sum_{c=1}^3 2z_c & \left(\mathbf{y}_n(c) - \sum_{v\in\mathcal{V}} \delta(m_v(\mathbf{y})-c)\right) + \\ & \left(\sum_{v\in\mathcal{V}} \delta(m_v(\mathbf{y})-c) - \mathbf{y}_n(c)\right), \end{aligned} \tag{19}$$

which is decomposable and can be solved by graph cut [4].

APPENDIX B
LEARNING OF FLSSVM

The learning process for FLSSVM is formulated as [19]:

$$\begin{aligned} \underset{\mathbf{w},\{\xi_n\}_{n=1}^N}{\text{minimise}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{n=1}^N \xi_n \\ \text{subject to} \quad & \left(\max_{h_n\in\mathcal{H}} \mathbf{w}^\top\Psi(\mathbf{x}_n,\mathbf{y}_n,h_n)\right) - \\ & \left(\mathbf{w}^\top\Psi(\mathbf{x}_n,\widehat{\mathbf{y}}_n,\widehat{h}_n)\right) \geq \Delta(\mathbf{y}_n,\widehat{\mathbf{y}}_n) - \xi_n \\ & \xi_n \geq 0, \forall \widehat{\mathbf{y}}_n \in \mathcal{Y}, \forall \widehat{h}_n \in \mathcal{H}, n = 1,...,N, \end{aligned} \tag{20}$$

where $\{\xi_n\}_{n=1}^N$ and $\Delta(\mathbf{y}_n,\widehat{\mathbf{y}}_n)$ are defined above in (20). The learning algorithm to solve (20) is the concave-convex procedure [49], consisting of the following stages: 1) update the latent variable h_n for n^{th} training sample using the latest estimate for $\mathbf{w}$, with $\max_{h_n\in\mathcal{H}} \mathbf{w}^\top\Psi(\mathbf{x}_n,\mathbf{y}_n,h_n)$; and 2) update $\mathbf{w}$ with (18) with $\{h_n\}_{n=1}^N$ from step 1 using the cutting plane algorithm [17], similarly to (18). Also similarly to (7), the loss augmented inference and the inference are based on graph cuts (alpha expansion) [4], where the high-order loss is integrated into graph cuts using the same decomposition shown in (19).