Learning Interpretable Microscopic Features of Tumor by Multi-task Adversarial CNNs Improves Generalization

Mara Graziani (✉ mara.graziani@hevs.ch)

University of Geneva

Sebastian Otalora

University of Geneva

Stéphane Marchand-Maillet

University of Geneva

Henning Müller

HES-SO Valais-Wallis

Vincent Andrearczyk

HES-SO Valais-Wallis

Research Article

Keywords: Interpretable Deep Learning, Histopathology, Multi-task learning

Posted Date: July 14th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-744740/v3>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Learning Interpretable Microscopic Features of Tumor by Multi-task Adversarial CNNs Improves Generalization

Mara Graziani*

mara.graziani@hevs.ch

University of Applied Sciences Western Switzerland (HES-SO Valais), 3960, Sierre, Switzerland

University of Geneva (UNIGE), Department of Computer Science (CUI), 1227, Carouge, Switzerland

Sebastian Otálora

juan.otalamontenegro@hevs.ch

University of Applied Sciences Western Switzerland (HES-SO Valais), 3960, Sierre, Switzerland

University of Geneva (UNIGE), Department of Computer Science (CUI), 1227, Carouge, Switzerland

Stéphane Marchand-Maillet

University of Geneva (UNIGE), Department of Computer Science (CUI), 1227, Carouge, Switzerland

Henning Müller

henning.mueller@hevs.ch

University of Applied Sciences Western Switzerland (HES-SO Valais), 3960, Sierre, Switzerland

University of Geneva (UNIGE), Department of Radiology and Medical Informatics, 1211, Geneva, Switzerland

Vincent Andrearczyk

vincent.andrearczyk@hevs.ch

University of Applied Sciences Western Switzerland (HES-SO Valais), 3960, Sierre, Switzerland

Abstract

Adopting Convolutional Neural Networks (CNNs) in the daily routine of pathological diagnosis requires not only near-perfect precision, but also sufficient generalization to data shifts and transparency. Existing CNN models act as black boxes, not ensuring to the physicians that important diagnostic features are used by the model. Building on top of successfully existing techniques such as multi-task learning, domain adversarial training and concept-based interpretability, we address the challenge of introducing diagnostic factors in the training objectives. Our architecture, by learning end-to-end an uncertainty-based weighting combination of multi-task and adversarial losses, is encouraged to focus on pathology features such as density and pleomorphism of nuclei, e.g. variations in size and appearance, while discarding misleading features such as staining differences. Our results on breast lymph node tissue show significantly improved generalization in the detection of tumorous tissue, with best average AUC 0.89 ± 0.01 against the baseline AUC 0.86 ± 0.005 . By applying the interpretability technique of linearly probing intermediate representations, we also demonstrate that interpretable pathology features such as nuclei density are learned by the proposed CNN architecture, confirming the increased transparency of this model. This result is a starting point towards building transparent multi-task architectures that are robust to data heterogeneity. Our code is available at <https://bit.ly/356yQ2u>.

Keywords: Interpretable Deep Learning, Histopathology, Multi-task learning

1. Introduction

The analysis of microscopic tissue images by Convolutional Neural Networks (CNNs) is an important part of computer-aided systems for cancer detection, staging and grading (Litjens et al., 2017; Janowczyk and Madabhushi, 2016; Campanella et al., 2019; Ilse et al., 2020). The automated suggestion of Regions of Interest (RoIs) is one task, in particular, that

may help pathologists in increasing their performance and inter-rater agreement in the diagnosis (Wang et al., 2016). Hard annotations of tumor regions, however, are costly and rarely pixel-level precise. Moreover, the existing imaging datasets are highly heterogeneous, being subject to staining, fixation, slicing variability, multiple scanner resolutions, artifacts, and, at times, permanent ink annotations (Lafarge et al., 2017). While physicians can naturally adapt to the variability of the images in the datasets, deep features are sensitive to confounding factors and their reliability for clinical use is thus still questionable (Janowczyk and Madabhushi, 2016; Campanella et al., 2019). The multiple sources of variability and the limited availability of training data lead to deep features that often present unwanted biases and that do not mirror clinically relevant diagnostic features. For example, using ImageNet for model pretraining introduces an important bias towards texture features (Geirhos et al., 2018) and this impacts the classification of tumorous tissue by transfer learning (Graziani et al., 2018).

As largely argued in the literature (Caruana et al., 2015; Rudin, 2019; Tonekaboni et al., 2019), transparent and interpretable modelling should be prioritized to ensure model safety and reliability, but transparent models are still struggling to appear in the context of digital pathology. This friction is due to the complexity of microscopy images as opposed to working with tabular data with specific clinical descriptions for each variable. We thus propose a novel convolutional architecture that increases the transparency and control of the learning process. The main technical innovation here is the combination of two successful techniques, namely multi-task learning (Caruana, 1997) and adversarial training (Ganin et al., 2016), with the purpose of guiding model training to focus on relevant features, called *desired targets*, and to discard *undesired targets* such as confounding factors. Moreover, we bring new insights on balancing multiple tasks for digital pathology, a technique that only recently arose interest in the digital pathology landscape (Gamper et al., 2020b). The joint optimization of main, auxiliary and adversarial task losses is a novel exploration in the histopathology field.

We propose an application of this architecture on the histopathology task of breast cancer classification. In this particular context, CNNs should extract from the vast amount of information contained in Whole Slide Images (WSIs) fine grained features of the tissue structure. Our objective function is thus built in such a way to obtain features representative of highly informative clinical factors (e.g. nuclei neoplasticity) and invariant to confounding concepts (e.g. staining variability). Our results show that the learned features contain information about the desired targets of nuclei density, area and texture, which are chosen to match the diagnostic procedure of physicians. At the same time, we achieve invariance to domain shifts in the tissue staining, fixation, processing and digitalization. As our experiments show, the main challenge of the proposed architecture is the combination various multi-task and adversarial losses that may have differing error metrics. From our analysis of loss weighting strategies, it emerges that the uncertainty-based approach proposed by Kendall et al. (2018) best handles the convergence and stability of the joint optimization. Our results, besides, show a significant increase in the performance and generalization to unseen data. The results are reported in Section 4 and discussed in Section 5. Further details on experimental setup and methods are presented in Section 3

Figure 1: Intuitive illustration about multi-task learning in (a): given two related tasks M and A, the optimization process is driven to choose solutions that satisfy both tasks. In (b) no connection exists between the tasks, hence the multi-task approach may result in negative transfer, providing only sub-optimal models for all the tasks. In (c), an adversarial task is added and the optimization is pushed to representations that satisfy both main and auxiliary tasks, but that avoids the minimum of the adversarial task.

2. Related Works

2.1 Multi-task Learning

Similarly to how learning happens in humans, multi-task architectures aim at simultaneously learning multiple tasks that are related to each other (Ruder, 2017). In principle, learning two related tasks generates mutually beneficial signals that lead to more general and robust representations than traditional or multimodal learning. Figure 1 better illustrates the original concept proposed by Caruana (1997) and our the extension that we propose in this paper. Suppose that a complex model, e.g. a CNN, is trained on a main task M. In Figures 1 (a) and (c) the optimization objective of M has two local minima, represented as the set $\{a, b\}$. In Figure 1 (a), the auxiliary task A is related to the main task, with which it shares the local minimum in a . The joint optimization of M and A is thus likely to identify the shared local minima a as the optimal solution (Caruana, 1997). The search is biased by the extra information given by the task A towards representations that lay at the intersection of what could be learned individually for each task. In Figure 1 (b), the auxiliary task is totally unrelated to the main task. No local minima are shared in this case and a negative transfer may happen without positive improvements to the performance. Finally, in Figure 1 (c), we show our contribution in this paper, where we extend the multi-task framework to introduce an adversarial task C that is learned by adversarial training Ganin et al. (2016). In this case, the main task M has local minima in $\{a, b, c\}$, but the minimum in c is also a solution of the adversarial task C. By being adversarial to C, the optimization is likely to prefer solutions that satisfy M and A, while avoid solutions that satisfy the adversarial task C. Hence, the solution a should be favored by the concurrent action of both tasks A and C.

Multi-task architectures divide into two families depending on the hard or soft sharing of the parameters. In architectures with hard parameter sharing such as the one proposed in this paper, multiple supervised tasks share the same input and some intermediate representation (Caruana, 1997). The parameters learned up to this intermediate point are

called *generic parameters* since they are shared across all tasks. In soft parameter sharing, the weight updates are not shared among the tasks and the parameters are task-specific, introducing only a soft constraint on the training process (Duong et al., 2015). As explained by Caruana (1997), multi-task learning leads to various benefits if the tasks are linked by a valid relationship. For instance, the generalization error bounds are improved and the risk of overfitting is reduced (Baxter, 1995). The speed of convergence is also increased, since fewer training samples are required per task (Baxter, 2000).

Multi-task learning has been successful in various applications, such as natural language processing (Subramanian et al., 2018), computer vision (Kokkinos, 2017), autonomous driving (Leang et al., 2020), radiology (Andrearczyk et al., 2021) and histology (Gamper et al., 2020b). The preliminary work by Gamper et al. (2020b), in particular, shows a decrease in the loss variance as an effect of multi-task for oral cancer, suggesting that this work may have high potential for histology applications. Note that, depending on the applications and on the loss functions used to represent the multiple tasks, multiple strategies exist for weighting each task contribution in the objective function Gong et al. (2019). Alternatively to uniform weighting, for example, dynamical task re-weighting during training was proposed by Leang et al. (2020). Kendall et al. (2018), moreover, proposed the use of uncertainty estimates to weight each task.

2.2 Adversarial Learning

Proposed by Ganin et al. (2016), adversarial learning introduced a novel approach to solving the so-called problem of domain adaptation, namely the minimization of the domain shift in the distributions of the training (also called source distribution) and testing data (i.e. target). Typically treated as either an instance re-weighting operation (Gong et al., 2013) or as an alignment problem (Long et al., 2015), domain adaptation is handled by adversarial learning as the optimization of a domain confusion loss. A domain classifier discriminates between the source and the target domains during training and its parameters are optimized to minimize the error when discriminating the domain labels. This can be extended to more than two domains by a multi-class domain classifier. The adversarial learning of domain-related features is obtained by a gradient reversal operation on the branch learning to discriminate the domains. Because of this operation, the network parameters are optimized to maximize the loss of the domain classifier, thus making multiple domains impossible to distinguish one from another in the internal network representation. This causes a competition between the main task and the domain branch during training that is referred to as a min-max optimization framework. As a downside, the optimization of adversarial losses may be complicated, with the min-max operation affecting the stability of the training (Ganin et al., 2016). Convergence can be promoted, however, by activating and de-activating the gradient reversal branch according to a training schedule as in Lafarge et al. (2017).

2.3 Concept-based Interpretable Modelling

Interpretability of deep learning in digital pathology applications is becoming increasingly important H.M. et al. (2022). Most of the methods proposed so far, however, focus on achieving post-hoc explanations rather than direct interpretable modelling. Linear models,

in particular, were proposed as an inherently interpretable approach to probe the internal activations of the network after training. Regression Concept Vectors (RCVs) and Concept Activation Vectors (CAVs) were shown to generate insightful explanations in terms of diagnostic measures (Kim et al., 2018; Graziani et al., 2018, 2019a,b; Yeche et al., 2019). RCVs, particularly, solve a linear regression problem in the space of the internal activations of a CNN layer. They were used to evaluate the relevance of a *concept*, e.g. a clinical feature such as nuclei area or tumour extension, in the model output generation. The performance of the linear regressor indicated how well the concept was learned. Both CAVs and RCVs constitute a baseline of linear interpretability of CNNs, formalized for applications in the medical domain as concept attribution (Graziani et al., 2020). No possibility is given by these methods, however, to act on the training process and modify the learning of a concept. It is not possible, for example, to discourage the learning of a confounding concept, e.g. domain, staining, watermarks. Similarly, the learning of discriminant concepts cannot be further encouraged. With this paper we aim at addressing this gap.

3. Methods

3.1 Proposed Architecture

The architecture for guiding the training of CNNs is described in this section for a general application with pre-defined features. This general framework can be applied to multiple tasks. The diagnosis of cancerous tissue in breast microscopy images is proposed in this paper as an application for which the implementation details are described in Sections 3.5 and 3.6.

In the following, we clarify the notation used to describe the model. We assume that a set of N observations, i.e. the input images, is drawn from an unknown underlying distribution and split into a training subset $\{\mathbf{x}_i\}_{i=1}^n$ and a testing subset $\{\mathbf{x}_i\}_{i=n+1}^N$. The main task, namely the one for which we aim at improving the generalization, is the prediction of the image labels $\mathbf{y} = \{y_i\}_{i=1}^n$, for which ground truth annotations are available. A CNN of arbitrary structure is used as a feature encoder, of which the features are then passed through a stack of dense layers. The model parameters up to this point are defined as $\boldsymbol{\theta}_f$. The parameters of the label prediction output layers are identified by $\boldsymbol{\theta}_y$. The structure described up to this point replicates a standard CNN with a single main task branch that is addressing the classification. The remaining parameters of the architecture implement (i) the learning of auxiliary tasks by multi-task learning (Caruana, 1997) and (ii) the adversarial learning of detrimental features to induce invariance in the representations, as in the domain adversarial approach by Ganin et al. (2016). We combine these two approaches by introducing K extra targets representing desired and undesired tasks that must be introduced to the learning of the representations. The targets are modeled as the prediction of the feature values $\{c_{k,i}\}_{i=1}^N$, where $k \in 1, \dots, K$ is an index representing the extra task being considered. The feature values may be either continuous or categorical. Additional parameters $\boldsymbol{\theta}_k$ are trained in parallel to $\boldsymbol{\theta}_y$ for the K extra targets. We refer to all model outputs for all inputs $\mathbf{x}$ as $f(\mathbf{x}) \in \mathcal{R}^{K+1}$.

The architecture is illustrated in Figure 2 and consists of two blocks. The first block is used to extract features from the input images. A state of the art CNN of arbitrary choice without the decision layer is used as a feature encoder generating a set of feature maps.

Figure 2: Multi-task adversarial architecture

179 The feature maps are passed through a Global Average Pooling (GAP) operation that is
 180 performed to spatially aggregate the responses and connect them to a stack of dense layers.
 181 For this specific architecture, we use a stack of three dense layers of 1024, 512 and 256
 182 nodes respectively. The second block comprises one branch per task, taking as input the output
 183 of the first block. The main task branch consists of the prediction of the labels $\mathbf{y}$ and has
 184 as many dense nodes as there are of unique classes in $\mathbf{y}$. For binary classification tasks,
 185 e.g. discrimination of tumorous against non-tumorous inputs, the main task branch has a
 186 single node with a sigmoid activation function. K branches are added to model the extra
 187 targets. We refer to *extra* tasks for all the additional targets to the main task whether
 188 desired or undesired. *Auxiliary* tasks refer to the modeling of the desired targets, while
 189 *adversarial* tasks refer to that of undesired targets. The extra tasks are modeled by linear
 190 models as in Graziani et al. (2018). For continuous-valued targets, the extra branch consists
 191 of a single node with linear activation function. For categorical targets, the extra branch
 192 has multiple nodes followed by a softmax activation function. A gradient reversal operation
 193 (Ganin et al., 2016) is performed on the branches of the undesired targets to discourage the
 194 learning of these features.

195 3.2 Objective Function

196 The objective function of the proposed architecture balances the losses of the main task and
 197 the extra tasks for the desired and undesired targets. This is obtained by a combination of
 198 multi-task and adversarial learning. The main task loss is $\mathcal{L}_y^i(\theta_f, \theta_y) = \mathcal{L}_y(x_i, y_i; \theta_f, \theta_y)$,
 199 where θ_f are the parameters of the first block (namely of the CNN encoder and the dense
 200 layers) in Figure 2 and θ_y those of the main task branch in the second block of the same
 201 figure. The extra parameters θ_k ($k \in 1, \dots, K$) are trained for the branches of the desired
 202 and undesired target predictions, with the loss being $\mathcal{L}_k^i(\theta_f, \theta_k) = \mathcal{L}_k(x_i, c_{k,i}; \theta_f, \theta_k)$.

203 Training the model on n training and $(N - n)$ testing samples consists of optimizing the
 204 function:

$$E(\boldsymbol{\theta}_y, \boldsymbol{\theta}_f, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K) = \lambda_m \frac{1}{n} \sum_{i=1}^n \mathcal{L}_y^i(\boldsymbol{\theta}_f, \boldsymbol{\theta}_y) + \sum_{k=1}^K \lambda_k \frac{1}{N} \sum_{i=1}^N \mathcal{L}_k^i(\boldsymbol{\theta}_f, \boldsymbol{\theta}_k). \quad (1)$$

205 The gradient update is:

$$\boldsymbol{\theta}_f \leftarrow \boldsymbol{\theta}_f - \left(\lambda_m \frac{\partial \mathcal{L}_y^i}{\partial \boldsymbol{\theta}_f} + \sum_{k=1}^K \lambda_k \alpha_k \frac{\partial \mathcal{L}_k^i}{\partial \boldsymbol{\theta}_f} \right), \quad (2)$$

$$\boldsymbol{\theta}_y \leftarrow \boldsymbol{\theta}_y - \lambda_m \frac{\partial \mathcal{L}_y^i}{\partial \boldsymbol{\theta}_y}, \quad (3)$$

$$\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}_k - \lambda_k \frac{\partial \mathcal{L}_k^i}{\partial \boldsymbol{\theta}_k}, \quad (4)$$

208 where λ_m and λ_k are positive scalar hyperparameters to tune the trade-off between the
 209 losses. For each extra branch, the hyperparameter $\alpha_k \in \{-1, 1\}$ is used to specify whether
 210 the update is adversarial or not. A value of $\alpha_k = -1$ activates the gradient reversal operation
 211 and starts an adversarial competition between the feature extraction and the corresponding
 212 k^{th} extra branch. The main task is only trained on the training data, since $\mathcal{L}_y^i = 0$ for $i > n$
 213 in Eq. (2) and (3). Following the work in Ganin et al. (2016), the extra tasks are learned
 214 on both training and test data. The training on test data can also be removed, since it is
 215 not always possible to fully retrain a network for new data.

216 3.3 Loss weighting strategy

217 The proposed architecture requires the combination of multiple objectives in the same loss
 218 function. The vanilla formulation in Eq. 1 simply performs a weighted linear sum of the
 219 losses for each task. This is the predominant approach used in prior work with multi-
 220 objective losses (Gong et al., 2019) and adversarial updates (Ganin et al., 2016; Lafarge
 221 et al., 2017). The appropriate choice of weighting of the different task losses is a major
 222 challenge of this setting. The tuning of the hyperparameters may reveal tedious and non-
 223 trivial due to the combination of classification and regression tasks with different ranges of
 224 the loss function values (e.g. combining the bounded binary cross-entropy loss in $[0,1]$ with
 225 the unbounded mean squared error loss).

226 An optimal weighting approach may be learned simultaneously with the other tasks by
 227 adding network parameters for the loss weights λ_m and λ_k . The direct learning of λ_m and
 228 λ_k , however, would just result in weight values quickly converging to zero. Kendall et al.
 229 (2018) proposed a Bayesian approach that makes use of the homoscedastic uncertainty
 230 of each task to learn the optimal weighting combination. In loose words, homoscedastic
 231 uncertainty reflects a task-dependent confidence in the prediction. The main assumption to
 232 obtain an uncertainty-based loss weighting strategy is that the likelihood of the task output
 233 can be modeled as a Gaussian distribution with the mean given by the model output and
 234 a scalar observation noise σ :

$$p(\mathbf{y}|f(\mathbf{x})) = \mathcal{N}(f(\mathbf{x}), \sigma^2) \quad (5)$$

This assumption is also applied to the outputs of the extra tasks. The loss weights λ_m and λ_k are then learned by optimizing the minimization objective given by the negative log likelihood of the joint probability of the task outputs given the model predictions. To clarify this concept, let us focus on a simplified architecture with the main task being the logistic regression of binary labels (e.g. tumor v.s. non-tumor) with noise σ_1 and one auxiliary task consisting of the linear regression of feature values $\mathbf{c} = \{c_i\}_{i=1}^N$, with noise σ_2 . The minimization objective for this multi-task model is:

$$-\log p(\mathbf{y}, \mathbf{c} | \mathbf{f}(\mathbf{x})) \propto \frac{1}{2\sigma_1^2} \mathcal{L}_y(\boldsymbol{\theta}_f, \boldsymbol{\theta}_y) + \frac{1}{2\sigma_2^2} \mathcal{L}_k(\boldsymbol{\theta}_f, \boldsymbol{\theta}_k) + \log \sigma_1 + \log \sigma_2 \quad (6)$$

By minimizing Eq. 6 with respect to σ_1 and σ_2 , the optimal weighting combination is learned adaptively based on the data (Kendall et al., 2018). As σ_1 increases, the weight for its corresponding loss decreases, and vice-versa. The last term $\log \sigma_1 + \log \sigma_2$, besides, acts as a regularizer discouraging each noise to increase unreasonably. This construction can be extended easily to multiple regression outputs and the derivation for classification outputs is given in Kendall et al. (2018).

3.4 Dataset

The experiments are run on three publicly available datasets, namely Camelyon 16, Camelyon 17 (Litjens et al., 2018a) and the breast subset of the PanNuke dataset (Gamper et al., 2019, 2020b) (The data are available for download at the links <https://camelyon17.grand-challenge.org> and https://warwick.ac.uk/fac/cross_fac/tia/data/pannuke, last accessed in January 2022).. The Camelyon challenge collections of 2016 and 2017 contain respectively 270 and 899 WSIs. All training slides of both challenges contain annotations of metastasis type (i.e. negative, macro-metastases, micro-metastases, isolated tumor cells), and 320 of them contain manual segmentations of tumor regions. The analysis also includes the breast tissue scans in the PanNuke dataset, for which multiple nuclei types were annotated by the semi-automatic instance segmentation tool in Gamper et al. (2019). Labels of neoplastic, inflammatory, connective, epithelial, and dead nuclei are given together with the images by the dataset creators. We pre-process the data by extracting smaller patches of 224×224 pixels at the highest magnification level and reducing the staining variability by Reinhard normalization Reinhard et al. (2001). Training, validation and test splits are build as in Table 1. Oversampling is applied to the PanNuke images to balance their under-representation. For each input we extract smaller image patches located in the center, upper left, upper right, bottom left and bottom right corners of the image. The three pre-existing PanNuke folds were used to separate the patches in the splits by using two folds in the training set and the third fold in the internal testing set. No PanNuke images were used for the external validation since all the three folds contain images for the multiple centers.

3.5 Main task and architecture backbone

The main task that we address is the binary classification of input images that include tumor tissue from those without tumor. Inception V3 pretrained on the ImageNet (Szegedy et al., 2016) is used as the backbone CNN for feature encoding. The parameters up to the last

Table 1: Summary of the train, validation, internal and external test splits.

	Label	Cam16	Cam17 (5 Centers)					PanNuke (3 Folds)		
			C. 0	C. 1	C. 2	C. 3	C. 4	F. 1	F. 2	F. 3
Train	Neg.	12954	31108	25137	38962	25698	0	1425	1490	0
	Pos.	6036	8036	5998	2982	1496	0	2710	2255	0
Val.	Neg.	0	325	0	495	0	0	0	0	0
	Pos.	0	500	0	500	0	0	0	0	0
Int. Test	Neg.	0	0	274	483	458	0	0	0	1475
	Pos.	0	500	999	0	0	0	0	0	2400
Ext. Test	Neg.	0	0	0	0	0	500	0	0	0
	Pos.	0	0	0	0	0	500	0	0	0

convolutional layer are kept frozen to avoid overfitting to the pathology images. The output of the CNN is passed through the GAP and the three fully connected layers as illustrated in Figure 2. The fully connected layers have respectively 2048, 512 and 256 units. A dropout probability of 0.8 and L2 regularization are added to these three fully connected layers to avoid overfitting. The main task is the detection of patches containing tumor as a binary classification task. The branch consists of a single node with sigmoid activation function connected to the output of the third dense layer. The architecture as described up to here, hence without extra branches, is used as the baseline for the experiments. The extra tasks consist of either the linear regression or the linear classification of continuous or categorical labels respectively. For linear regression, the extra branch is a single node with linear activation function. The Mean Squared Error (MSE) between the predicted value and the label is added to the optimization function in Eq. 1. For the linear classification, the extra branch has a number of dense nodes equal to the number of classes to predict and a softmax activation function, also connected to the third dense layer. The Categorical Cross-Entropy (CCE) loss is added to the optimization in Eq. 1. Further details about the extra branches used for the experiments are given in Section 3.6.

The architecture is trained end-to-end with mini-batch Stochastic Gradient Descent (SGD) with standard parameters (learning rate of 10^{-4} and Nesterov momentum of 0.9). The main task loss function is the class-Weighted Binary Cross-Entropy (WBCE). The class weights are set to weight more heavily every instance of the positive class, for instance they are set to the ratio of negative samples $136774/29513 + 136774 = 0.82$ for the positive class and the ratio of positive samples 0.18 for the negative class.

We evaluate the convergence of the network by early stopping on the total validation loss with patience of 5 epochs. The Area Under the ROC Curve (AUC) is used to evaluate model performance. For each experiment, we perform five runs with multiple initialization seeds to evaluate the performance variation due to initialization. The splits are kept unchanged for the multiple seed variations. To evaluate the performance on multiple test splits, we perform bootstrapping of the test sets. A number of 50 test sets of 7589 images (the total number of test images in the two sets) are obtained by sampling with repetition from the total pool of testing images. This method evaluates the variance of the test set without prior assumption on the data distribution and it shows the performance difference due to variation of the sampling of the population.

Concept	Clinical reference	Description	Visual examples	Magnification	Source	Type	Task
Count of cavities	NGH tubular formation	tumour cells in gland structure	well formed poorly formed	low	annotation or automated	D	auxiliary regression
Nuclei area	NGH nuclear pleomorphism	abnormality in size	regular enlarged	high	annotation or automated	C	auxiliary regression
Nuclei Texture		vesicular appearance	uneven stain				
Mitotic count	NGH mitotic count	number of mitosis	mitosis	high	annotation or automated	D	auxiliary regression
Nuclei density	Proliferation index	Reproduction rate	regular overgrowth	any	annotation or automated	D	auxiliary regression
Staining	Staining procedure	dye applied on the tissues	different appearance	any	metadata	D	adversarial classification

Figure 3: Control targets for breast cancer. C and D stand for continuous and discrete respectively.

3.6 Configuration of the extra targets

The experiments focus on the integration of four desired and one undesired target with multiple combinations. The auxiliary targets relate to the main task, being important diagnostic features. We expect that learning of these desired features will improve the solution robustness and generalization of the model. Discarding the undesired targets may improve the invariance of the learned features to confounding factors. The Nottingham Histologic Grading (NHG) of breast tissue identifies the key diagnostic features for breast cancer (Bloom and Richardson, 1957). By analyzing this we derived the desired and undesired features that are illustrated in Figure 3. From this set, we retain cancer indicators at the nuclear level, since the input images are at the highest magnification. We model the variations of the nuclei size, appearance (e.g. irregular, heterogeneous texture) and density shown in Figure 3 as real-valued variables. Because of the heterogeneity of the data, we also guide the network training to discard information about staining and tissue representation differences in the images. The processing center of the slides is modeled as an undesired target, encouraging feature invariance to staining and acquisition differences.

Hand-crafted features representing the variations in the nuclei size and appearance are automatically extracted either from the images or from the nuclear contours. The nuclear contours are available in the form of manual annotations only for the PanNuke data. Automated contours of the nuclei in the Camelyon images are obtained by the multi-instance deep segmentation model in Otálora et al. (2020). This model is a Mask R-CNN model (He et al., 2017), fine-tuned from ImageNet weights on the Kumar dataset for the nuclei segmentation task (Kumar et al., 2017). The R-CNN identifies nuclei entities and then generates pixel-level masks by optimizing the Dice score. ResNet50 (He et al., 2017) is used for the convolutional backbone as in (Otálora et al., 2020). The network is optimized by SGD with standard parameters (learning rate of 0.001 and momentum of 0.9).

The number of pixels inside nuclear contours is averaged for each input image to represent variations of the nuclei area, referred to as *area* in the experiments. Nuclei *density* is

estimated by counting the nuclei in the image. Haralick descriptors of texture contrast and correlation (Haralick, 1979) are also extracted from the entire input images as in Graziani et al. (2018). Being continuous and unbounded measures, the values for these features are normalized to have zero mean and unitary standard deviation before training the model. In the paper, we refer to these features as *area*, *density*, *contrast* and *correlation*. The values of these features are used as prediction labels for the auxiliary target branches, that are also named as the feature that they should predict. These auxiliary branches perform a linear regression task, trying to minimize the Mean Squared Error between the predicted value of the feature and the extracted values used as labels.

Information about the center that performed the data acquisition is present in the dataset as metadata. We model it as a categorical variable that may take values from 0 to 7, namely one for each known center in the training data. Since there is no specific information on acquisition centers in Camelyon16 and PanNuke, these have been modeled as two distinct acquisition centers in addition to the five known centers of Camelyon17. This information is partly inaccurate, since we know that in both datasets more than a single acquisition center was involved (Litjens et al., 2018b; Gamper et al., 2020a). The noise introduced by this information may limit the benefits introduced by the adversarial branch but it should not affect negatively the performance. In the future, unsupervised domain alignment methods may also be explored. The prediction of this variable is added to the architecture as an undesired target branch, referred to as *center* in the experiments.

4. Results

4.1 Single-task baseline model

In the baseline model, only the main task branch is trained and no extra tasks are used. The results for the baseline are shown in the first row of Table 2, that is identified by a unique ID, namely model-ID 1. The experiments are performed on the dataset configuration in Table 1, where an internal and external test sets are used to evaluate the generalization. Two columns are used to report the results on the internal (int.) and external (ext.) test sets. Where not stated otherwise, the average AUC (avg. AUC) over ten repetitions with multiple initialization seeds is used for the evaluation.

4.2 Double-task combinations

An ablation study is performed by adding to the baseline model a single additional task at a time. This study aims at identifying the benefits of encouraging each task individually. The desired and undesired targets described in Section 3.6 are added as extra branches to the backbone architecture detailed in Section 3.5. The gradient reversal operation is only active for the *center* branch. The losses of each task are combined by two strategies, namely a vanilla and the uncertainty-based approach in Kendall et al. (2018). In the vanilla configuration, the loss weight values are set to 1 for all branches. The results for single task combinations are reported in Table 2 with unique IDs ranging from 2 to 5. Model-ID 2, for example, is given by the combination of the main task branch with the additional task *area*, namely of predicting the area of the nuclei in the images. A single auxiliary branch already outperforms the baseline (int. avg AUC 0.819 ± 0.001 , ext. avg. AUC 0.868 ± 0.005 ,

int. avg. F1 0.783 ± 0.002 , ext. avg. F1 0.315 ± 0.001), as for example in model-ID 3 by encouraging nuclei *count* (int. avg AUC 0.836 ± 0.005 , ext. avg. AUC 0.890 ± 0.009 , int. avg. F1 0.768 ± 0.001 , ext. avg. F1 0.660 ± 0.003) The avg. F1 score of model-ID3 on the two test sets is at 0.746 ± 0.009 , a significantly higher value than the 0.701 ± 0.001 of the baseline model. On the external test set, the best generalization is achieved by adding *count* as a desired target (ext. avg. AUC 0.890 ± 0.009).

4.3 Multi-task adversarial combinations

The most promising branches are then combined together to further improve performance. The following combinations of multiple auxiliary and adversarial tasks are tested in the experiments: *center* + *density*, *center* + *area*, *center* + *density* + *area*. The combination of all the branches in model-ID 8 leads to the best performance on the int. test (int. avg. AUC 0.874 ± 0.009), with an increase of 0.05 AUC points compared to the baseline. Model-ID 6 reports comparable performance on the external test set to model-ID 3. The addition of the *center* adversarial branch in model-ID 6 leads to the best model overall with overall avg. AUC (on both internal and external sets) at 0.824 ± 0.006 and avg. F1 score 0.755 ± 0.005 for the uncertainty trained model. This represents a significant improvement compared to the overall avg. AUC 0.79 ± 0.001 and avg. F1 score 0.701 ± 0.004 of the baseline model, with $p\text{-value} < 0.001$. The statistical significance of the results is evaluated by the non-parametric Wilcoxon test (two-sided) applied on the bootstrapping of the test set as described in Sec. 3.5.

Table 2: Average AUC on the main task and standard deviations from different starting points of the network parameter initialization. Results for the vanilla and uncertainty based weighting strategies. The adversarial task, i.e. *center*, is marked by an overline.

ID	main	area	count	contrast	<u>center</u>	int. test		ext. test	
1	✓					0.819±0.001		0.868±0.005	
						vanilla	unc.	vanilla	unc.
2	✓	✓				0.718±0.11	0.834±0.01	0.560±0.06	0.871±0.01
3	✓		✓			0.853±0.03	0.836±0.005	0.874±0.02	0.890±0.009
4	✓			✓		0.854±0.07	0.835±0.008	0.883±0.02	0.876±0.007
5	✓				✓	0.845±0.10	0.822±0.005	0.884±0.04	0.871±0.005
6	✓		✓		✓	0.863±0.06	0.841±0.004	0.623±0.10	0.890±0.01
7	✓	✓	✓		✓	0.838±0.05	0.848±0.003	0.490±0.03	0.864±0.01
8	✓	✓	✓	✓	✓	0.858±0.02	0.874± 0.009	0.686±0.20	0.825±0, 01

4.4 Sanity checks

To confirm the benefit of the added related tasks, we compare these results with those obtained with random noise as additional targets. This experiment is performed as a sanity check, where an auxiliary task is trained to predict random values. As expected, the overall, internal and external avg. AUCs are lower for this experiment and have larger standard deviations (overall avg. AUC 0.819 ± 0.04 , int. test AUC 0.834 ± 0.001 and ext. avg. AUC

0.879 $\pm$ 0.03). This shows that the selected tasks are more relevant to the main task than the regression of random values.

4.5 Interpretability and visualizations

At this point, one may ask whether it is possible to interpret the internal representation of the model to verify that the additional tasks were learned by the guided architectures. Table 3 evaluates how well the representations of nuclei area, count and contrast are learned by the baseline and the proposed CNNs. For model-ID3 (trained with the uncertainty-based weighting strategy), the prediction of the nuclei *count* values has average determination coefficient $R^2 = 0.81 \pm 0.05$, showing that the concept was learned during training, passing from an initial Mean Squared Error (MSE) of the prediction of 0.46 to 0.17 at the end of training. Similar results apply to the other model-IDs 2 to 4 when only a single branch is added. Table 3 compares the performance on the extra-tasks to learning the concepts directly on the baseline model activations, where the network parameters are not optimized to learn the extra tasks. The classification of the *center* in model-ID 5 reduces in accuracy as the gradient reversal is used during training. The centers of the validation sets are predicted with accuracy 0.29 ± 0.01 at the end of the training (starting from an initial accuracy of 0.53 ± 0.01). When more extra tasks are optimized together the performance on the side tasks is affected, with Model-IDs 6, 7 and 8 not reporting high R^2 values. The average R^2 of nuclei *count* for model-ID 6, for example, decreases from -2.25 ± 0.05 and plateaus at around -0.63 ± 0.05 .

Table 3: Performance on the extra-tasks for the baseline and guided models with the uncertainty-based strategy. The average and standard deviation of the determination coefficient are reported (the closer to 1 the better).

ID	area	count	contrast
baseline	0.66 ± 0.003	0.85 ± 0.007	0.56 ± 0.01
2	0.70 ± 0.005	-	-
3	-	0.88 ± 0.004	-
4	-	-	0.64 ± 0.003

Figure 4 shows the dimensionality reduction of the internal representations learned by the baseline and model-ID 3. The visualization is obtained by applying the Uniform Manifold Approximation and Projection (UMAP) method by McInnes et al. (2018) (the hyper-parameters for the visualization were kept to the default values of 15 neighbors, 0.1 minimum distance and local connectivity of 1). Note that the model-ID 3 selected for visualization was trained with the uncertainty-based weighting strategy. In the representation, the two classes are represented with different colors, whereas the size of the points in the plot is indicative of the values of nuclei counts in the images. The top row shows the projection of the internal representation of the last convolutional layer (known as mixed10 in the standard InceptionV3 implementation) of the two models. The bottom row shows the projection of the first fully connected layer after the GAP operation. Since the nuclei count values were normalized to zero mean and unit variance, these are represented in the plot as ranging

Figure 4: Uniform Manifold Approximation and Projection (UMAP) representation of the internal activations of the baseline and guided model-ID3 (obtained with the UMAP default hyperparameter set up). The top row shows the activations at the last convolutional layer of both models, known as mixed10 in the standard implementation of InceptionV3 (Szegedy et al. (2016)). The bottom row shows the activations of the first fully connected layer after the GAP operation.

between a minimum of -2 and a maximum of 2. For clarity of the representation, the image shows the UMAP of a random sampling of 4000 input images.

5. Discussion

The central question of this work is whether expert-knowledge can be used as a guidance to induce the learning of robust representations that generalize better to new data than the classic training of CNNs. The proposed experiments give multiple insights on this question that we discuss in this section.

The clinical features used for diagnosis can be modeled as auxiliary and adversarial tasks. Since the extra tasks are modeled as regression tasks, this approach favors model transparency, since it ensures that specific features of the data are learned during network

training. The features *area* and *contrast*, for example, were already modeled by Graziani et al. (2018) as linear regression tasks that were used to probe the internal activations of InceptionV3 fine-tuned on the Camelyon data. These features emerged as relevant concepts learned by the network to drive the classification. The architecture in this paper further guides the training towards learning a predictive relationship for these concepts. This is obtained by jointly optimizing the extra regression tasks together with the main task, encouraging even further the attention of the backbone CNN on these aspects through multi-task learning (Caruana, 1997). Our results in Table 2 show that the overall performance is significantly better than the baseline approach, even when a single extra task is added to the training as in model-ID3, for example. For this model, the representations of the positive class organize in a more compact cluster than in the baseline model, as shown by the UMAP visualization in Figure 4. The representations on the right side of the figure (for model-ID3) also appear more structured than those on the left, being organized as following a direction for increasing values of the nuclei count. With the feature values being extracted automatically, this modification does not require extra annotations, and only introduces a neglectable increase in complexity. One extra task, for instance, requires the training of only 2049 additional parameters, namely the 0.008% of Inception V3.

The auxiliary and adversarial tasks are balanced in the same end-to-end training without additional tuning of the loss weight nor of a specific training schedule that would help the convergence of the adversarial task. This novel approach exploits the benefits of another paper in the machine learning research field that uses task-dependent uncertainty to balance structurally different losses such as MSE and BCE (Kendall et al., 2018). The multi-task framework is sensitive to the choice of the auxiliary tasks. Tasks to include during training should be selected based on prior knowledge and existing observations on whether they may cooperate or compete towards learning the main task. The result of concept-based interpretability analyses in Graziani et al. (2020) was informative about for us to choose the tasks that were included in this study. The structured framework for identifying cooperating and competing tasks suggested in Standley et al. (2020) may be used when the prior knowledge about each tasks is not sufficient to establish whether a concept should be encouraged or discouraged. It is important to notice that appropriate weighting is also fundamental to avoid that the additional tasks overtake the main task. The vanilla weighting of the losses shows instability on unseen domains and poor performance on the external test set. The uncertainty-based approach, conversely, is robust to data variability and consistent over random seed initializations for all model-IDs. The stability to data variability is shown by the performance on the external test set and by the testing with bootstrapping. The consistency over seed reinitializations is shown by the small standard deviation of the AUC on both test sets. This gives insight on how to handle the multiple loss types for the multi-task modeling on histopathology tasks. With the uncertainty-based weighting strategy the architecture did not require any specific tuning of the loss weights, whereas a fine-tuning of the weighting parameters appears highly necessary in the vanilla approach, particularly for the combinations with more than one extra task (model-IDs 6, 7, 8). The manual fine-tuning of the loss weights in the vanilla approach may lead to the over-specification of the model to the specific requirements of the test data considered in this study. These results not only extend the preliminary work by Gamper et al. (2020b) to a different histology tissue and model architecture, but also give more insights on how

to handle multiple auxiliary losses and adversarial losses without requiring tedious tuning of hyper-parameters.

6. Conclusion

We show how expert-knowledge can be used pro-actively during the training of CNNs to drive the representation learning process. Clinically relevant and easy-to-interpret features describing the visual inputs are introduced as extra tasks for the learning objective, significantly improving the robustness and generalization performance of the model. From a design perspective, our framework aligns ethically with the intent of not replacing humans, but rather making them part of the development of deep learning algorithms. Our experiments focus on tumor detection, but other tasks may benefit from a similar architecture. Deepfake detection, for example, may be improved by the auxiliary task of classifying checkerboard artifacts (Wang et al., 2020). Human-computer interfaces, besides, may be designed to directly collect user-specific feedback. The extra tasks may be used as a weak-supervision to extend the training data with unlabeled datasets at a marginal cost of some extra automatic processing such as the extraction of nuclei contours or texture features. One may argue that additional annotations may be required for other clinical features. This represents, however, only a minor limitation of this method since a few annotated images may already suffice to train the extra tasks.

A few limitations of our method require further work and analyses. Our analysis is restricted to uncertainty-based weighting strategies, although several approaches were proposed in the literature (Leang et al., 2020). The results on *center* do not show a marked improvement by the adversarial branch. This could be due to the fact that the acquisition centers were not annotated for the PanNuke dataset. An unsupervised domain adaptation approach such as the domain alignment layers proposed by Carlucci et al. (2017) may be used to discover this latent information. Depending on the application, a different loss weighting approach may be used for the adversarial task and other undesired control targets can also be included, such as rotation, scale and image compression methods. In addition, our experiments show that the auxiliary tasks become harder to learn when they are scaled up in number, with model-ID 8 having a lower R^2 for the regression of the individual features than those reported for model-IDs 2 to 5 in Table 3. As explained also by Caruana (1997), the poor performance on the extra tasks is not necessarily an issue as long as these help with improving the model performance and generalization on unseen data. Further research is necessary to verify how this architecture may be improved to ensure high performance on all the extra tasks, while maintaining its transparency and complexity at similar levels. In future work we will also focus on extracting extra features solely from unlabeled data and on introducing them during training as weak supervision.

Data Availability

The Camelyon data that support the findings of this study are available at <https://camelyon17.grand-challenge.org/Data/> as accessed in August 2021, with the DOI identifier of the dataset paper doi.org/10.1109/TMI.2018.2867350. The PanNuke data are

527 available at https://warwick.ac.uk/fac/cross_fac/tia/data/pannuke (accessed in Au-
 528 gust 2021), with DOI identifier of the paper doi.org/10.1007/978-3-030-23937-4_2.

529 Code Availability

530 The code used for the experiments is available online for reproducibility on Github (https://github.com/maragraziani/multitask_adversarial) and Zenodo at <https://doi.org/10.5281/zenodo.5243433> (accessed in August, 2021). An executable version on Code
 532 Ocean is currently being developed.
 533

534 Acknowledgments

535 This work is supported by the European Union’s projects ExaMode (grant s825292) and
 536 AI4Media (grant 951911).

537 Ethical Standards

538 The work follows appropriate ethical standards in conducting research and writing the
 539 manuscript, following all applicable laws and regulations regarding treatment of animals or
 540 human subjects.

541 Conflicts of Interest

542 We declare we don’t have conflicts of interest.

543 References

- 544 Vincent Andrearczyk, Pierre Fontaine, Valentin Oreiller, and Adrien Depeursinge. Multi-
 545 task Deep Segmentation and Radiomics for Automatic Prognosis in Head and Neck Can-
 546 cer. In *under revision*, page 1, 2021.
- 547 Jonathan Baxter. Learning internal representations. In *Proceedings of the eighth annual*
 548 *conference on Computational learning theory*, pages 311–320, 1995.
- 549 Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence*
 550 *research*, 12:149–198, 2000.
- 551 HJG Bloom and WW Richardson. Histological grading and prognosis in breast cancer: a
 552 study of 1409 cases of which 359 have been followed for 15 years. *British Journal of*
 553 *Cancer*, 11(3):359, 1957.
- 554 Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Mirafior, Vitor W K Silva,
 555 Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs.
 556 Clinical-grade computational pathology using weakly supervised deep learning on whole
 557 slide images. *Nature medicine*, 25(8):1301–1309, 2019.

- 558 Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Buló.
 559 Just dial: Domain alignment layers for unsupervised domain adaptation. In *International*
 560 *Conference on Image Analysis and Processing*, pages 357–369. Springer, 2017.
- 561 Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
- 562 Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad.
 563 Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day read-
 564 mission. In *International Conference on Knowledge Discovery and Data Mining*, 2015.
- 565 Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing:
 566 Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd*
 567 *annual meeting of the Association for Computational Linguistics and the 7th international*
 568 *joint conference on natural language processing (volume 2: short papers)*, pages 845–850,
 569 2015.
- 570 Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Raj-
 571 poot. PanNuke: an open pan-cancer histology dataset for nuclei instance segmentation
 572 and classification. In *European Congress on Digital Pathology*, pages 11–19. Springer,
 573 2019.
- 574 Jevgenij Gamper, Navid Alemi Koohbanani, Simon Graham, Mostafa Jahanifar, Syed Ali
 575 Khuram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. Pannuke dataset exten-
 576 sion, insights and baselines. *arXiv preprint arXiv:2003.10778*, 2020a.
- 577 Jevgenij Gamper, Navid Alemi Koohbanani, and Nasir Rajpoot. Multi-task learning in
 578 histo-pathology for widely generalizable model. *arXiv preprint arXiv:2005.08645*, 2020b.
- 579 Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle,
 580 François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial train-
 581 ing of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030,
 582 2016.
- 583 Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann,
 584 and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape
 585 bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.
- 586 Boqing Gong, Kristen Grauman, and Fei Sha. Connecting the dots with landmarks: Dis-
 587 criminatively learning domain-invariant features for unsupervised domain adaptation. In
 588 *International Conference on Machine Learning*, pages 222–230. PMLR, 2013.
- 589 Ting Gong, Tyler Lee, Cory Stephenson, Venkata Renduchintala, Suchismita Padhy, An-
 590 thony Ndirango, Gokce Keskin, and Oguz H Elibol. A comparison of loss weighting
 591 strategies for multi task learning in deep neural networks. *IEEE Access*, 7:141627–141632,
 592 2019.
- 593 Mara Graziani, Vincent Andrearczyk, and Henning Müller. Regression concept vectors for
 594 bidirectional explanations in histopathology. In *Understanding and Interpreting Machine*
 595 *Learning in Medical Image Computing Applications*, pages 124–132. Springer, 2018.

- 596 Mara Graziani, James M Brown, Vincent Andrearczyk, Veysi Yildiz, J Peter Campbell,
597 Deniz Erdogan, Stratis Ioannidis, Michael F Chiang, Jayashree Kalpathy-Cramer, and
598 Henning Müller. Improved interpretability for computer-aided severity assessment of
599 retinopathy of prematurity. In *Medical Imaging 2019: Computer-Aided Diagnosis*, 2019a.
- 600 Mara Graziani, Henning Muller, and Vincent Andrearczyk. Interpreting intentionally flawed
601 models with linear probes. In *IEEE International Conference on Computer Vision Work-*
602 *shops*, 2019b.
- 603 Mara Graziani, Vincent Andrearczyk, Stephane Marchand-Maillet, and Henning Müller.
604 Concept attribution: Explaining CNN decisions to physicians. *Computers in Biol-*
605 *ogy and Medicine*, page 103865, 2020. ISSN 0010-4825. doi: [https://doi.org/10.1016/](https://doi.org/10.1016/j.compbimed.2020.103865)
606 [j.compbimed.2020.103865](https://doi.org/10.1016/j.compbimed.2020.103865). URL [http://www.sciencedirect.com/science/article/](http://www.sciencedirect.com/science/article/pii/S0010482520302225)
607 [pii/S0010482520302225](http://www.sciencedirect.com/science/article/pii/S0010482520302225).
- 608 Robert M Haralick. Statistical and structural approaches to texture. *Proceedings of the*
609 *IEEE*, 67(5):786–804, 1979.
- 610 Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings*
611 *of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- 612 Bas H.M., Hugo J. K., Kenneth G.A. G., and Max A. V. Explainable artificial intelligence
613 (XAI) in deep learning-based medical image analysis. *MIA*, 79, 2022. ISSN 1361-8415.
614 doi: <https://doi.org/10.1016/j.media.2022.102470>. URL [https://www.sciencedirect.](https://www.sciencedirect.com/science/article/pii/S1361841522001177)
615 [com/science/article/pii/S1361841522001177](https://www.sciencedirect.com/science/article/pii/S1361841522001177).
- 616 Maximilian Ilse, Jakub M Tomczak, and Max Welling. Deep multiple instance learning for
617 digital histopathology. In *Handbook of Medical Image Computing and Computer Assisted*
618 *Intervention*, pages 521–546. Elsevier, 2020.
- 619 Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image anal-
620 ysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*,
621 7, 2016.
- 622 Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to
623 weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on*
624 *computer vision and pattern recognition*, pages 7482–7491, 2018.
- 625 Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas,
626 et al. Interpretability beyond feature attribution: Quantitative testing with concept
627 activation vectors (TCAV). In *International Conference on Machine Learning*, pages
628 2673–2682, 2018.
- 629 Iasonas Kokkinos. Ubertnet: Training a universal convolutional neural network for low-,
630 mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings*
631 *of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6129–6138,
632 2017.

- 633 N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi. A dataset and a
 634 technique for generalized nuclear segmentation for computational pathology. *IEEE Trans-*
 635 *actions on Medical Imaging*, 36(7):1550–1560, 2017. doi: 10.1109/TMI.2017.2677499.
- 636 Maxime W Lafarge, Josien PW Pluim, Koen AJ Eppenhof, Pim Moeskops, and Mitko Veta.
 637 Domain-adversarial neural networks to address the appearance variability of histopathol-
 638 ogy images. In *Deep Learning in Medical Image Analysis and Multimodal Learning for*
 639 *Clinical Decision Support*, 2017.
- 640 Isabelle Leang, Ganesh Sistu, Fabian Bürger, Andrei Bursuc, and Senthil Yogamani. Dy-
 641 namic task weighting methods for multi-task networks in autonomous driving systems. In
 642 *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*,
 643 pages 1–8. IEEE, 2020.
- 644 Geert Litjens, Thijs Kooi, Babak E. Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco
 645 Ciompi, Mohsen Ghafoorian, Jeroen A. Van Der Laak, Bram Van Ginneken, and Clara I
 646 Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*,
 647 42:60–88, 2017.
- 648 Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balken-
 649 hol, Peter Bult, Altuna Halilovic, Meyke Hermesen, Rob van de Loo, Rob Vogels, et al.
 650 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAME-
 651 LYON dataset. *GigaScience*, 7(6):giy065, 2018a.
- 652 Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balken-
 653 hol, Peter Bult, Altuna Halilovic, Meyke Hermesen, Rob van de Loo, Rob Vogels, and et al.
 654 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAME-
 655 LYON dataset. *GigaScience*, 7(6), 2018b.
- 656 Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable
 657 features with deep adaptation networks. In *International conference on machine learning*,
 658 pages 97–105. PMLR, 2015.
- 659 Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation
 660 and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- 661 Sebastian Otálora, Manfredo Atzorib, Amjad Khanb, Oscar Jimenez-del Toroa, Vincent
 662 Andrearczykb, and Henning Müllera. Systematic comparison of deep learning strategies
 663 for weakly supervised gleason grading. *Medical Imaging 2020: Digital Pathology*, 2020.
- 664 Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between
 665 images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.
- 666 Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint*
 667 *arXiv:1706.05098*, 2017.
- 668 Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions
 669 and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

- 670 Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio
 671 Savarese. Which tasks should be learned together in multi-task learning? In *International*
 672 *Conference on Machine Learning*, pages 9120–9132. PMLR, 2020.
- 673 Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. Learning
 674 general purpose distributed sentence representations via large scale multi-task learning.
 675 In *International Conference on Learning Representations*, 2018.
- 676 Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Re-
 677 thinking the inception architecture for computer vision. In *IEEE Conference on Computer*
 678 *Vision and Pattern Recognition*, 2016.
- 679 Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What
 680 clinicians want: contextualizing explainable machine learning for clinical end use. In
 681 *Machine Learning for Healthcare Conference*, pages 359–380. PMLR, 2019.
- 682 Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck.
 683 Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*,
 684 2016.
- 685 Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-
 686 generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF*
 687 *Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020.
- 688 Hugo Yèche, Justin Harrison, and Tess Berthier. UBS: A Dimension-Agnostic Metric for
 689 Concept Vector Interpretability Applied to Radiomics. In *Interpretability of Machine*
 690 *Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision*
 691 *Support*, 2019.