

# Guiding CNNs towards Relevant Concepts by Multi-task and Adversarial Learning

Mara Graziani<sup>1,2</sup>, Sebastian Otálora<sup>1,2</sup>, Henning Müller<sup>1,2,3</sup>, and Vincent Andrearczyk<sup>1</sup>

<sup>1</sup> University of Applied Sciences and Arts Western Switzerland, Hes-so Valais, Sierre, Switzerland

<sup>2</sup> Centre Universitaire d'Informatique, University of Geneva, 1227 Carouge, Switzerland

<sup>3</sup> Medical Faculty, University of Geneva, 1211 Geneva, Switzerland  
`mara.graziani@hevs.ch`

**Abstract.** The opaqueness of deep learning limits its deployment in critical application scenarios such as cancer grading in medical images. In this paper, a framework for guiding CNN training is built on top of successful existing techniques of hard parameter sharing, with the main goal of explicitly introducing expert knowledge in the training objectives. The learning process is guided by identifying concepts that are relevant or misleading for the task. Relevant concepts are encouraged to appear in the representation through multi-task learning. Undesired and misleading concepts are discouraged by a gradient reversal operation. In this way, a shift in the deep representations can be corrected to match the clinicians' assumptions. The application on breast lymph nodes histopathology data from the Camelyon challenge shows a significant increase in the generalization performance on unseen patients (from 0.839 to 0.864 average AUC, p-value = 0,0002) when the internal representations are controlled by prior knowledge regarding the acquisition center and visual features of the tissue. The code will be shared for reproducibility on our GitHub repository.

**Keywords:** human-machine interaction · histopathology · multi-task learning · adversarial learning

## 1 Introduction

The analysis of tissue images by Convolutional Neural Networks (CNNs) is an important part of computer-aided systems for cancer detection, staging and grading [1,2,3,4]. The learning process is made very challenging by the costly and rarely pixel-level precise annotations and the heterogeneity of the data acquisition procedures [2,3]. Domain shifts in the data can cause the learning process to be biased towards specific acquisition procedures, an unwanted effect that can be explicitly corrected by adversarial learning [7,8]. Besides, additional objectives can improve the final performance if added as extra tasks in the end-to-end network framework [17]. The lack of model interpretability, however, makes it

difficult to interpret the internal mechanisms of CNN decisions, hindering the formulation of such additional objectives that could be used to guide the learning process towards solutions that are in line with the clinical guidelines and requirements. The recently-developed approach of concept attribution [10,11] generated new possibilities to explore the combination of additional tasks to guide the learning process. Explanations in terms of clinical factors can be obtained with high versatility [10,11,12,13,14,15], hence easing the interaction of domain-experts with deep learning models [16].

This paper uses the idea of concept learning in [14] to guide the learning of specific concepts (e.g. clinical factors) during training. The resulting CNN is guided to focus on the *desired* and *undesired control targets* specified by the users. *Desired control targets* specify the concepts (i.e. features in the data) that should be encouraged to appear in the internal representations. This is obtained by auxiliary tasks in a multi-task learning setting [17]. The *undesired control targets* represent features in the data that should be discouraged during the training process. A gradient reversal operation [18,19] is added in this case to obtain invariance to the undesired feature. Physicians, for example, can rank clinical factors according to their importance for the diagnosis and encourage their presence in the internal representations. A deep learning framework of this type can be used to ensure the users that their inductive bias is transferred to the network, i.e. a similar set of assumptions is used by the user and by the network to predict unseen inputs. While multi-task learning [17] and adversarial learning [18] are widely used techniques, fundamental in our contributions is their combination for steering the learning process. Moreover, the auxiliary tasks defined to learn the control targets are inherently interpretable, being modeled as linear regressors [20].

To demonstrate the benefits of this approach, experiments are proposed on a patch-based classifier of breast tumor tissue, using the data from the Camelyon16 and 17 challenges [21]. The detection of tumor areas from images at a high magnification level is the first step in the assessment of the TNM degree of lymph nodes spreading. Prognostic markers for breast cancer, such as those in the Nottingham grading system [22], are modeled with continuous or categorical values that can be extracted from the raw images, from segmentation (automatic or manual) of nuclei contours or the metadata. Introducing such interpretable features as control targets during network training leads to significantly higher performance.

## 2 Methods

### 2.1 Framework for Control Targeted Training

Our method combines and extends the techniques of adversarial [7,18,19] and multi-task learning [17]. In this section, we formulate the problem and describe a general framework that introduces the *desired* and *undesired* control targets as extra tasks. A set of  $N$  observations drawn from an unknown underlying distribution is split into  $\{\mathbf{x}_i\}_{i=1}^n$  for training and  $\{\mathbf{x}_i\}_{i=n+1}^N$  for testing. The main

task is defined by the prediction of output labels  $\{y_i\}_{i=1}^n$ , modeled by the loss  $\mathcal{L}_y^i(\theta_f, \theta_y) = \mathcal{L}_y(\mathbf{x}_i, y_i; \theta_f, \theta_y)$ , where  $\theta_f$  and  $\theta_y$  are respectively the parameters of a neural network feature extractor and of the label prediction output layers as illustrated in Figure 1a. The user’s control is introduced by  $K$  extra branches. An extra branch with parameters  $\theta_{c_k}$  ( $k \in 1, \dots, K$ ) is trained to predict the feature values  $\{c_{k,i}\}_{i=1}^N$ , with loss  $\mathcal{L}_{c_k}^i(\theta_f, \theta_{c_k}) = \mathcal{L}_{c_k}(\mathbf{x}_i, c_{k,i}; \theta_f, \theta_{c_k})$ . Training the model consists of optimizing the function:

$$E(\theta_y, \theta_f, \theta_{c_1}, \dots, \theta_{c_K}) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_y^i(\theta_f, \theta_y) + \lambda \sum_{k=1}^K \sum_{i=1}^N \mathcal{L}_{c_k}^i(\theta_f, \theta_{c_k}). \quad (1)$$

The gradient update is:

$$\theta_f \leftarrow \theta_f - \lambda \left( \frac{\partial \mathcal{L}_y^i}{\partial \theta_f} + \sum_{k=1}^K \alpha_k \frac{\partial \mathcal{L}_{c_k}^i}{\partial \theta_f} \right), \quad (2)$$

$$\theta_y \leftarrow \theta_y - \lambda \frac{\partial \mathcal{L}_y^i}{\partial \theta_y}, \quad (3)$$

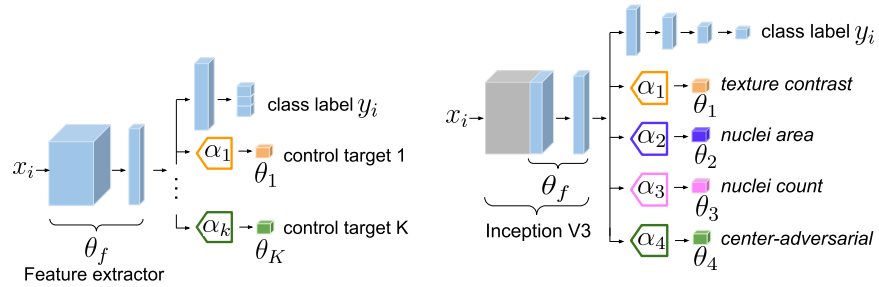
$$\theta_{c_k} \leftarrow \theta_{c_k} - \lambda \frac{\partial \mathcal{L}_{c_k}^i}{\partial \theta_{c_k}}, \quad (4)$$

where  $\lambda \in (0, 1]$  is the global learning rate,  $\alpha_k \in \mathbb{R}$  modulates the rate of update of the parameters  $\theta_f$  for the corresponding extra task. A positive  $\alpha_k$  encourages the feature extractor to include information about  $c_k$ , while a negative value triggers an adversarial competition between the feature extraction and the corresponding  $k^{\text{th}}$  extra branch. Setting  $\alpha_k = 0$  does not impact the main task since it annihilates the contribution of  $\mathcal{L}_{c_k}^i$  from the update of  $\theta_f$  in (2). Finally, the main task is only trained on the training data, since  $\mathcal{L}_y^i = 0$  for  $i > n$  in Eq. (2) and (3). The extra tasks, however, are trained on both training and test data<sup>4</sup>. The resulting general architecture is shown in Figure 1a.

## 2.2 From Expert Knowledge to Extra Tasks

Expert knowledge is a valuable source of information about prior beliefs on the underlying distribution describing the data. In some cases, experts may prefer encouraging high relevance for specific features to prevent the learning of misleading correlations in the training data. As an example, the model in [20] learned to assign a lower risk of death to cases of pneumonia with concurring asthma than to the general population. A correct diagnosis would have rather taken the opposite decision. It was the effective care given to these patients to actually cause the lower risk reported in the data. By introducing user control,

<sup>4</sup> The training of the extra tasks on testing data is optional as it is not always possible to fully retrain a network for new data.








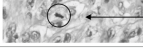



(a) General framework

(b) Histopathology application

**Fig. 1.** (a) The proposed framework for guiding CNN training. The hyper-parameters  $\alpha_1, \dots, \alpha_K$  are applied to the gradient back-propagation. The signs control the distinction between *desired* and *undesired* control targets. (b) Network architecture for the proposed application on breast histopathology. The frozen parameters are in gray.

the model decision could have been fixed to react appropriately to the concurring symptom.

In several applications, guidelines for manual decision-making have been perfected and validated over several years of study, e.g. the TNM and the Nottingham system in breast histopathology, the Gleason score in prostate cancer grading, etc. Even more, valuable work defined informative handcrafted visual features (e.g [23]) that were employed before the deep learning era and have recently been used to complement the deep features in fusion approaches [24].

Concept	Clinical reference	Description	Visual examples	Magnification	Source	Type	Task
Count of cavities	NGH tubular formation [20]	tumour cells in gland structure	 well formed  poorly formed	low	annotation or automated	D	auxiliary regression
Nuclei area	NGH nuclear pleomorphism [20]	abnormality in size	 regular  enlarged	high	annotation or automated	C	auxiliary regression
Nuclei Texture		vesicular appearance	 uneven stain				
Mitotic count	NGH mitotic count [20]	number of mitosis	 mitosis	high	annotation or automated	D	auxiliary regression
Nuclei density	Ki-67 protein expression	cell proliferation	 regular  overgrowth	any	annotation or automated	D	auxiliary regression
Staining	Staining procedure [5]	dye applied on the tissues	 different appearance	any	metadata	D	adversarial classification

**Fig. 2.** Possible control targets for breast cancer. D and C stand for continuous and discrete respectively.

Figure 2 shows examples of prior knowledge on the breast histopathology application with the respective clinical references. Continuous or discrete values representing a conceptual description of a phenomenon of interest are extracted from the images. For example, abnormality in nuclei size is modeled as variations in the *area* of the nuclei segmentation at a high magnification level. Nuclei segmentations are automatically extracted by a multi-instance deep segmentation model as in [25]. Similarly, nuclei density is estimated by counting the number of nuclei instances in the image (referred to as *nuclei count*). The Haralick descriptor of texture *contrast* [23] is also automatically extracted from the input images, being informative about the overall tissue texture and thus of eventual changes in the nuclei morphology. Being continuous, these measures are predicted as an extra-regression task in our framework. Hematoxylin and Eosin (H&E) stains often differ across different analysis centers, depending, among others, on the chemical mixture, temperature, etc. A *center-adversarial* approach can be implemented to obtain staining invariance as proposed in [7,8]. Information about the center from which images are acquired is often present in the metadata as in Camelyon17.

The illustration in Figure 1b shows the framework including the main task with four extra branches. Differently from feature-fusion approaches, additional

information is not passed to the model as input but as a target for the extra branches. This procedure is described in [17] as more robust than feature fusion since possible sources of noise in the features are better handled during training when added as outputs.

### 2.3 Datasets

The experiments are conducted on the Camelyon16 and 17 challenge data [21], which include Whole Slide Images (WSIs) of lymph node sections together with slide-level annotations and some local segmentations of tumor regions. The WSIs were prepared at different pathology centers, i.e. five for Camelyon17 (with available metadata) and two for Camelyon16. From the two datasets, 468 WSIs are preprocessed by local adaptive thresholding for background removal as in [26]. Patches are uniformly sampled at the highest magnification level from the annotated tumor regions for the tumor class. For the non-tumor class, patches are sampled not only from the annotated images but also from 297 non-tumor WSIs in Camelyon17. To evaluate the generalization on an unseen patient coming from an unseen center the last center of Camelyon17 (named LPON in [21] that is not present in Camelyon16) was kept out from the training and internal test sets and used for external testing. More information about the training, validation, internal and external test splits are given in Table 1.

**Table 1.** Data splits. Patients were randomly selected for validation and test.

	tumor patches	non-tumor patches	patients	WSIs
train	18,268	120,015	244	459
val	1,000	820	2	2
internal test	1,499	1,215	4	6
external test	500	500	1	1
total	21,267	122,550	251	468

### 2.4 Architectures and Training

An Inception V3 [27] model is used for the experiments as in [26,28]. Four dense layers<sup>5</sup> are placed on top of the Global Average Pooling layer (GAP) to solve the main task. Early layers are frozen to avoid overfitting, with only the top four convolutional layers and the four dense layers being updated during training. The binary cross-entropy loss is minimized for the main task. Class weights are used during batch-training to deal with the strong class imbalance in the training data. Four extra branches composed of a single node with linear activations are connected to the output of the GAP. The hyperparameters for each branch are

<sup>5</sup> dense 2048, ReLU, dropout 0.8, dense 512, ReLU, dropout 0.8, dense 256, ReLU, dropout 0.8, dense 1, sigmoid

defined as  $\alpha_k$  (see Eq. (2)). Hyperparameter tuning was performed with a non-exhaustive search on the validation set, although a finer tuning could have lead to different values of the  $\alpha_k$  and results. The extra tasks consist of either linear regression (of the continuous targets *texture contrast*, *nuclei area*, *nuclei count*) or classification (of the *center* labels). In the first case, the target measures are normalized to zero mean and unit variance and the Mean Squared Error (MSE) loss is minimized. The center adversarial branch is trained with categorical cross-entropy (CCE) loss. The network is trained end-to-end with early stopping on the validation Area Under the ROC Curve (AUC). Since the validation set is not sufficiently representative of the external test data, the network is left to train until a plateau is reached for all methods, i.e. for 35 epochs. The performance on the main task is evaluated by AUC. The performance on the additional tasks is monitored to ensure that the information in the internal network representation is modified (although not reported due to space limitations). The evaluation of the additional tasks is based on the  $R^2$  of the regression (in case of MSE loss) and the prediction accuracy (in case of CCE loss).

### 3 Experiments and Results

#### 3.1 Individual Encouraging of Control Targets

We evaluate the impact of activating each extra branch individually by setting a single non-zero  $\alpha_k$  at a time. The first experiment evaluates the benefit of including the regression of the *contrast* (Haralick feature) of the patches as an extra task with  $\alpha_1 = 10$  (all the other  $\alpha_k = 0$  for  $k \neq 1$ ). We then evaluate the performances for encouraging the *nuclei area* with  $\alpha_2 = 10$ , and *nuclei count* with  $\alpha_3 = 10$ . *Center-adversarial* is trained in a similar way to [7,8], with  $\alpha_4 = -10$  triggering the gradient reversal operation. This task, in particular, is modeled as a multi-class classification problem with five center labels, namely the four centers of Camelyon17 used for training, and an additional center grouping the data from Camelyon16 for which the centers are unknown<sup>6</sup>. For comparison, an experiment is run with an additional task of white *noise* regression solved by a single node with linear activations on top of the GAP and with  $\alpha_5 = 1$ . These five models are tested on both the internal and external test sets. A baseline model without any active branch is trained for comparison, together with a model with intensive data augmentation as in [28].

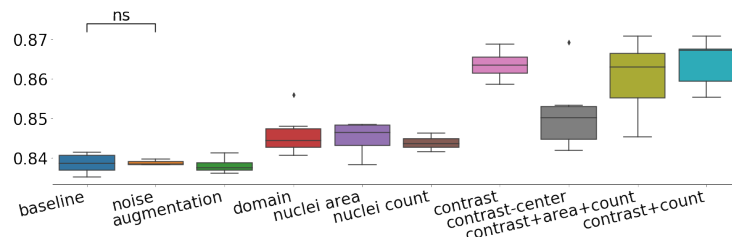
#### 3.2 Joint Multi-Task Adversarial

Experiments are also performed to observe the performance when multiple extra-tasks are optimized jointly with the main task. We encourage the learning of *contrast* while being *center-adversarial* by setting  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_4) = (10, 0, 0, -1)$ . For simplicity, we refer to this experiment as contrast-center. Combinations of desired control targets are then analyzed by encouraging with  $\boldsymbol{\alpha} = (10, 1, 0, 0)$

<sup>6</sup> Better performance would be expected with center information in Camelyon16.

*contrast* and *nuclei count* (referred to as *contrast+count*). The triplet *contrast*, *nuclei area* and *nuclei count* is then analyzed with  $\alpha = (10, 5, 1, 0)$ . This combination is called *contrast+area+count*.

### 3.3 Results



**Fig. 3.** Box plot of AUC on ten repetitions on the internal test set. Apart from *noise*, our proposed method is significantly higher than the baseline with p-values  $\leq 0.05$  (one-tailed heteroscedastic t-test).

Figure 3 shows the performance of the networks described in Sections 3.1 and 3.2. The AUC on the internal test set is used for comparison of the results over ten repetitions with different initializations of the weights. Of the models with only one extra task (described in Section 3.1), only the performance of the model with the additional branch encouraging the learning of *noise* are not significantly better than the baseline average. This shows that the control targets analyzed are relevant for the main task more than the introduction of just random values. The best result on the internal test set is obtained by combining two additional branches, namely the *contrast* and the *nuclei count*. With average AUC of 0.8640, this model significantly outperforms the baseline (as well as the baseline with data augmentation) average AUC at 0.8391 with p-value= 0,0002. On the external test set (not reported in a figure for brevity), the model with only the *center-adversarial* branch outperforms the baseline, with average AUC of 0.9228 compared to 0.9067 and p-value= 0.00052. Due to the domain shift of the external test set, this branch is necessary to obtain good performance on this set. The performance of the model encouraging *nuclei-area*, for example, is increased from 0.9076 to 0.9259 when combining with the *center-adversarial*.

## 4 Discussion and Conclusion

The central question of this work was whether expert-knowledge can be used in deep learning to learn more general representations. Prognostic factors can indeed be modeled as auxiliary and adversarial tasks to jointly train with the main task of recognition of breast tumor tissue. Each extra-task requires the



training of only 2,049 additional parameters, a very little increase from Inception V3, i.e. 0.008%. Results show that the performance of purely data-driven models is significantly improved when encouraging the learning of the diagnostic measures with either single or multiple branches. Generalization on unseen patients coming from unseen centers is best obtained when the additional tasks are learned with the domain invariance previously proposed in the literature [7,8]. Domain-adversarial training is unified by our framework as an undesired control target. Depending on the application, other undesired control targets could include rotation, scale, image compression methods and presence of watermarks or text.

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. This method could be used, for example, in human-computer interfaces to introduce user feedback during training. To achieve better transfer, an adversarial branch on over-specific representations learned from a different dataset could be removed, contrasting a phenomenon called overlearning. A limitation of this approach is, however, the need for labeled data to train the additional branches. While some information can be extracted automatically, e.g. the nuclei contours for nuclei count, this may not be possible for some types of control.

Finally, the control target measures automatically extracted from unlabeled data could be introduced during training as weak supervision. Data augmentation can also be used in combination with the extra branches to generate augmented images that correspond to the user's control targets. These last two will be the focus of our future work.

## Acknowledgements

This work was supported by the project PROCESS, part of the European Unions Horizon 2020 research and innovation program (grant agreement No 777533).

## References

1. Geert Litjens, Thijs Kooi, Babak E. Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
2. Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016.
3. Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Mirafior, Vitor W K Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.
4. Maximilian Ilse, Jakub M Tomczak, and Max Welling. Deep multiple instance learning for digital histopathology. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, pages 521–546. Elsevier, 2020.
5. Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. *Proceedings of Machine Learning Research, Machine Learning for Healthcare, 121, 2019*, 2019.
6. Babak E. Bejnordi, Geert Litjens, Nadya Timofeeva, Irene Otte-Höller, André Homeyer, Nico Karssemeijer, and Jeroen AWM van der Laak. Stain specific standardization of whole-slide histopathological images. *IEEE transactions on medical imaging*, 35(2):404–415, 2015.
7. Maxime W Lafarge, Josien PW Pluim, Koen AJ Eppenhof, Pim Moeskops, and Mitko Veta. Domain-adversarial neural networks to address the appearance variability of histopathology images. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2017.
8. Sebastian Otálora, Manfredo Atzori, Vincent Andrearczyk, Amjad Khan, and Henning Müller. Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology. *Frontiers in Bioengineering and Biotechnology*, 7:198, 2019.
9. Miriam Hägele, Philipp Seegerer, Sebastian Lapuschkin, Michael Bockmayr, Wojciech Samek, Frederick Klauschen, Klaus-Robert Müller, and Alexander Binder. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *arXiv preprint arXiv:1908.06943*, 2019.
10. Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, pages 2673–2682, 2018.
11. Mara Graziani, Vincent Andrearczyk, and Henning Müller. Regression concept vectors for bidirectional explanations in histopathology. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, 2018.
12. Mara Graziani, Henning Muller, and Vincent Andrearczyk. Interpreting intentionally flawed models with linear probes. In *IEEE International Conference on Computer Vision Workshops*, 2019.
13. Mara Graziani, James M Brown, Vincent Andrearczyk, Veysi Yildiz, J Peter Campbell, Deniz Erdogmus, Stratis Ioannidis, Michael F Chiang, Jayashree Kalpathy-Cramer, and Henning Müller. Improved interpretability for computer-aided severity assessment of retinopathy of prematurity. In *Medical Imaging 2019: Computer-Aided Diagnosis*, 2019.

14. Graziani M., Andrearczyk V., Marchand-Maillet S., and Mller H. Concept attribution: Explaining CNN decisions to physicians. *Computers in Biology and Medicine*, page 103865, 2020.
15. Hugo Yeche, Justin Harrison, and Tess Berthier. UBS: A Dimension-Agnostic Metric for Concept Vector Interpretability Applied to Radiomics. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, 2019.
16. Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Conference on Human Factors in Computing Systems*, 2019.
17. Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
18. Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.
19. Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems*, 2017.
20. Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *International Conference on Knowledge Discovery and Data Mining*, 2015.
21. Geert Litjens, Peter Bandi, Babak EhteshamiBejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob vandeLoo, Rob Vogels, and et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*, 7(6), 2018.
22. HJG Bloom and WW Richardson. Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years. *British Journal of Cancer*, 11(3):359, 1957.
23. Robert M Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.
24. Haibo Wang, Angel Cruz-Roa, Ajay Basavanhally, Hannah Gilmore, Natalie Shih, Mike Feldman, John Tomaszewski, Fabio Gonzalez, and Anant Madabhushi. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, 1(3), 2014.
25. Sebastian Otáloraa, Manfredo Atzorib, Amjad Khanb, Oscar Jimenez-del Toroa, Vincent Andrearczykb, and Henning Müllera. Systematic comparison of deep learning strategies for weakly supervised gleason grading. *Medical Imaging 2020: Digital Pathology*, 2020.
26. F. G. Zanjani, S. Zinger, and P. N. De. *Automated Detection and Classification of Cancer Metastases in Whole-slide Histopathology Images Using Deep Learning*, 2017.
27. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
28. Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*, 2017.