Divide-and-Rule: Self-Supervised Learning for Survival Analysis in Colorectal Cancer

Christian Abbet[1]([✉]), Inti Zlobec[4], Behzad Bozorgtabar[1,2,3],
and Jean-Philippe Thiran[1,2,3]

[1] Signal Processing Laboratory 5, EPFL, Lausanne, Switzerland
{christian.abbet,behzad.bozorgtabar,jeanphilippe.thiran}@epfl.ch
[2] Department of Radiology, Lausanne University Hospital, Lausanne, Switzerland
[3] Center of Biomedical Imaging, Lausanne, Switzerland
[4] TRU – Translational Research Unit, Bern, Switzerland

Abstract. With the long-term rapid increase in incidences of colorectal cancer (CRC), there is an urgent clinical need to improve risk stratification. The conventional pathology report is usually limited to only a few histopathological features. However, most of the tumor microenvironments used to describe patterns of aggressive tumor behavior are ignored. In this work, we aim to learn histopathological patterns within cancerous tissue regions that can be used to improve prognostic stratification for colorectal cancer. To do so, we propose a self-supervised learning method that jointly learns a representation of tissue regions as well as a metric of the clustering to obtain their underlying patterns. These histopathological patterns are then used to represent the interaction between complex tissues and predict clinical outcomes directly. We furthermore show that the proposed approach can benefit from linear predictors to avoid overfitting in patient outcomes predictions. To this end, we introduce a new well-characterized clinicopathological dataset, including a retrospective collective of 374 patients, with their survival time and treatment information. Histomorphological clusters obtained by our method are evaluated by training survival models. The experimental results demonstrate statistically significant patient stratification, and our approach outperformed the state-of-the-art deep clustering methods.

Keywords: Self-supervised learning · Histology · Survival analysis · Colorectal cancer

1 Introduction

Colorectal cancer is the third leading cause of cancer-related mortality worldwide. Five-year survival rates are low, at 60%. Although standard histopathological of cancer reporting based on features such as staging and grading identifies

patients with a potentially worse outcome to therapy, there is still an urgent need to improve risk stratification. Pathologists typically limit their reporting of colorectal cancers to approximately ten features, which they describe as single elements in their report (e.g., depth of invasion, pT; lymph node metastasis, etc.). However, the histopathological (H&E) slide is a "snapshot" of all occurring tumor-related processes, and their interactions may hold a wealth of information that can be extracted to help refine prognostication. These slides can then be digitized and used as input for computational algorithms to help support pathologists in their decision-making. The distribution of tissue types within the slide, the proximity of cell types or tissue components, and their spatial arrangement throughout the tissue can identify new patterns not previously detectable to the human eye alone.

Few studies have performed unsupervised clustering of whole slide images (WSIs) based on patch descriptors. They have been used to address the problem of image segmentation [16] or latent space clustering [4,6]. Among DL-based survival models, a recent study [13] used a supervised CNN for end-to-end classification of tissues to predict the survival of patients with colorectal cancer. Similar to our approach, several recent works have proposed unsupervised methods [14,17,22] for slide-level survival analysis. In [22], one of the first unsupervised approaches, DeepConvSurv has been proposed for survival prediction based on WSIs. More recently, DeepGraphSurv [14] has been presented to learn global topological representations of WSI via graphs. However, they heavily relied on noisy compressed features from a pre-trained VGG network. Recently, self-supervised representation learning methods [2,8,23] have been proposed to utilize the pretext task for extracting generalizable features from the unlabeled data itself. Therefore, the dataset does not need to be manually labeled by qualified experts to solve the pretext task.

Contributions. In this work, we propose a new approach to learn histopathological patterns through self-supervised learning within each WSI. Besides, we present a novel way to model the interaction between tumor-related image regions for survival analysis and tackle the inherent overfitting problem on tiny patient sets. To this end, we take advantage of a well-characterized, retrospective collective of 374 patients with clinicopathological data, including survival time and treatment information. H&E slides were reviewed, and at least one tumor slide per patient was digitized. To accelerate research we have made our code and trained models publicly available on GitHub[1].

2   Method

We first introduce our self-supervised image representation (Sect. 2.2) for the cancerous tissue area identified by our region of interest (RoI) detection scheme (Sect. 2.1). Then, we propose our deep clustering scheme and baseline algorithms

---

[1] https://github.com/christianabbet/DnR.

in Sect. 2.3 and Sect. 2.4, respectively. The clustering approach's usefulness is assessed by conducting survival analysis (Sect. 2.5) to measure if the learned clusters can contribute to disease prognostication. Finally, we discuss our implementation setup and experimental results in Sect. 3.

2.1   RoI Detection

Our objective is to learn discriminative patterns of unhealthy tissues of patients. However, WSI does not include information about the cancerous regions or the location of the tumor itself. Therefore, we seek a transfer learning approach for the classification of histologic components of WSIs. To do so, we choose to use the dataset presented in [12] to train a classifier to discriminate relevant areas. The dataset is composed of 100 K examples of tissue from CRC separated into nine different classes. For our task, we choose to retain three classes: lymphocytes (LYM), cancer-associated stroma (STR), and colorectal adenocarcinoma epithelium (TUM) that show the discriminative evidence for the class-of-interest and have been approved by the pathologist. Note that the presence of a large number of lymphocytes around the tumor is an indication of the immune reaction and, therefore, possibly linked to a higher survival score. We first train our classifier with the ResNet-18 backbone [9]. Then we use the stain normalization approach proposed in [15] to match the color space of the target domain and prevent the degradation of the classifier on transferred images. An example of RoI estimation is presented in Fig. 1. Such a technique allows us to discard a large part of the healthy tissue regions.

2.2   Self-Supervised Representation Learning

In this paper, we propose a self-supervised transfer colorization scheme to learn a more meaningful feature representation of the tissues and reduce the requirement for intensive tissue labeling. Unsupervised learning methods such as autoencoder trained by minimizing reconstruction error tend to ignore the underlying structure of the image as the model usually learns the distribution of the color space. To avoid this issue, we use colorization learning as a proxy task. As the input image, we convert the original unlabeled image through mapping function $\zeta(x)$ to a two-channel image (hematoxylin and eosin) that describes the nuclei and amount of extracellular material, respectively. To sidestep the memory bottleneck, we represent the WSI as a set of adjacent/overlapping tiles (image patches) $\{x_i \in \mathcal{X}\}_{i=1}^{N}$.

We define a function $\zeta : \mathcal{X} \rightarrow \mathcal{X}^{HE}$ that converts the input images to their HE equivalent [15,18]. Then, we train a convolutional autoencoder (CAE) to measure the per-pixel difference between transformed image(s) and input image(s) using MSE loss:

$$\min_{\phi,\psi} \mathcal{L}_{\mathrm{MSE}} = \min_{\phi,\psi} \|x - \psi \circ \phi \circ \zeta(x)\|_2^2. \qquad (1)$$

The encoder $\phi : \mathcal{X}^{HE} \rightarrow \mathcal{Z}$ is a convolutional neural network that maps an input image to its latent representation $\mathcal{Z}$. The decoder $\psi : \mathcal{Z} \rightarrow \mathcal{X}$ is an upsampling convolutional neural network that reconstructs the input image given
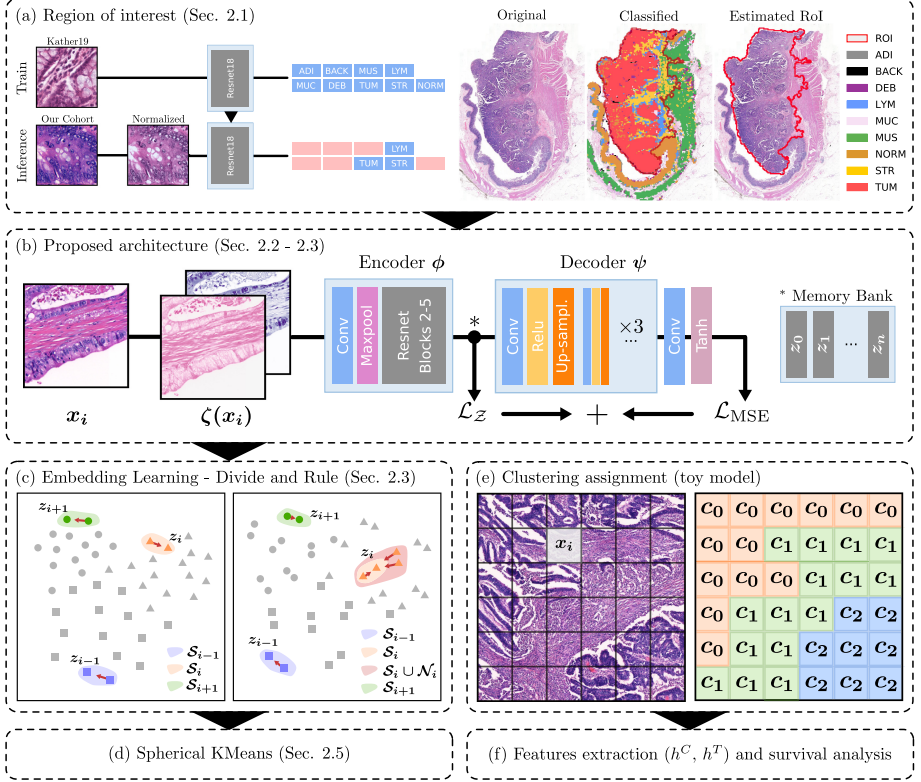
Fig. 1. The pipeline of the proposed approach. Estimation of the region of interest (a), learning of the embedding space (b–c), fitting of the cluster, assignment of all patient patches, and survival analysis (d–f).

a latent space representation. As a result, we use a single input branch to take into account the tissue's structural aspect.

2.3   Proposed Divide-and-Rule Approach

The principle behind our self-supervised learning approach is to represent image patches based on their spatial proximity in the feature space, meaning any two adjacent image patches (positive pairs) are more likely to be close to each other in the feature space $\mathcal{Z}$ than two distant patches (negative pairs). Such characteristics are met for overlapping patches as they share similar histomorphological patterns. We let $\mathcal{S}_i$ denote the set of patches that overlap with patch i spatially. Besides, we can assume that image patches in which their relative distances are smaller than a proximity threshold in the feature space should share common patterns. We define $\mathcal{N}_i$ as the set of top-k patches that achieve the lowest cosine distance to the embedding z_i of the image patch i.

Firstly, we initialize the network parameters using the self-supervised reconstruction loss in Eq. 1. Then, for each patch embedding i, we label its overlapping set of patches $\mathcal{S}_i$ as similar patches (positive pairs). Otherwise, we consider any distant patches as a negative pair, whose embeddings should be scattered. Motivated by [19], we use a variant of the cross-entropy to compute the instance loss (Eq. 2):

$$\mathcal{L}_{\text{Divide}} = - \sum_{i \in \mathcal{B}_{\text{inst}}} \log \left(\sum_{j \in S_i} p(j \mid i) \right), \quad p(j \mid i) = \frac{\exp(z_j^\top z_i / \tau)}{\sum_{k=1}^{N} \exp(z_k^\top z_i / \tau)}. \tag{2}$$

where $\tau \in \,]0,1]$ is the temperature parameter and $\mathcal{B}_{\text{inst}}$ denotes the set of samples in the mini-batch.

Secondly, we jointly optimize the training of network with reconstruction loss and a Rule loss $\mathcal{L}_{\text{Rule}}$ that takes into account the similarity of different images in the feature space (Eq. 3). We gradually expand the vicinity of each sample to select its neighbor samples. If samples have high relative entropy, they are dissimilar and should be considered as individual classes, $z \in \mathcal{B}_{\text{inst}}$. On the contrary, if samples have low relative entropy with their neighbors, they should be tied together, $z \in \mathcal{Z} \backslash \mathcal{B}_{\text{inst}}$. In practice, the entropy acts as a threshold to decide a boundary between close and distant samples and is gradually increased during training such that we go from easy samples (low entropy) to hard ones (high entropy). Finally, the proposed training loss, $\mathcal{L}_{\text{DnR}}$, joins the above losses with a weighting term λ (see Eq. 4):

$$\mathcal{L}_{\text{Rule}} = - \sum_{i \in \mathcal{Z} \backslash \mathcal{B}_{\text{inst}}} \log \left(\sum_{j \in \mathcal{S}_i \cup \mathcal{N}_i} p(j \mid i) \right). \tag{3}$$

$$\min_{\phi, \psi} \mathcal{L}_{\text{DnR}} = \min_{\phi, \psi} \mathcal{L}_{\text{MSE}} + \lambda \min_{\phi} [\mathcal{L}_{\text{Divide}} + \mathcal{L}_{\text{Rule}}]. \tag{4}$$

Dictionary Learning. Measuring similarities between samples requires the computation of features in the entire dataset for each iteration. The complexity grows as a function of the number of samples in the dataset. To avoid this, we use a memory bank, where we keep track and update the dictionary elements as in [19,23].

2.4   Algorithm Baselines

Deep Clustering Based on Spatial Continuity (DCS). As our first baseline, we leverage an inherent spatial continuity of WSIs. Spatially adjacent image patches (tiles) are typically more similar to each other than distant image patches in the slide and therefore should have similar feature representation $\mathcal{Z}$. Hence, we force the model to adopt such behavior by minimizing the distance between feature representations of a specific tile z_i and its overlapping tiles $\mathcal{S}_i$.

Deep Cluster Assignment (DCA). The downside of the first baseline is that in some cases, two distant image patches may be visually similar, or there may exist some spatially close patches that are visually different. This introduces noise in the optimization process. To tackle this issue, we can impose cluster membership as in [17].

Deep Embedded Clustering (DEC). Unlike the second baseline, the objective of our last baseline is not only to determine the clusters but also to learn a meaningful representation of the tiles. Therefore, we consider to jointly learn deep feature representation (ϕ, ψ) and image clusters U. The optimization is performed through the joint minimization of reconstruction loss and the KL divergence to gradually anneal cluster centers by fitting the model to an auxiliary distribution (see [20] for details).

2.5 Survival Analysis

Clustering and Assignment. The learned embedding space is assumed to be composed of a limited number of homogeneous clusters. We fit spherical KMeans clustering (SPKM) [21] to the learned latent space with K clusters. As a result, every patch within a patient slide will be assigned to a cluster, $c_k = \arg\min_{k \in \{0...K-1\}} \text{SPKM}(x_i, \mu_k)$.

Our objective is to model the interaction between tumor-related image regions (neighbor patches and clusters). To do so, we define a patient descriptor $h = [h^C, h^T] \in \mathbb{R}^{N \times (K+K^2)}$ as:

$$h_k^C = p(s = k) \quad \text{and} \quad h_{j \to k}^T = p(s = k \mid N(s) = j), \tag{5}$$

where s is a patch, h_k^C denotes the probability that a patch belongs to cluster k and h_k^T is the probability transition between a patch and its neighbors $N(s)$ (i.e. local interactions between clusters within the slide).

Survival. Survival analysis is prone to overfitting as we usually rely on a small patient set and a large number of features. To counter this issue, we first apply forward variable selection [10] using log partial likelihood function with tied times [5], $\mathcal{L}_{ll}$, and likelihood-ratio (LR) test to identify the subset of relevant covariates:

$$\text{LR} = -2[\mathcal{L}_{ll}(\beta^{\text{new}} \mid h^{\text{new}}) - \mathcal{L}_{ll}(\beta^{\text{prev}} \mid h^{\text{prev}})]. \tag{6}$$

Here $(h, \beta)^{\text{prev}}$ and $(h, \beta)^{\text{new}}$ are the previous and new estimated set of covariates, respectively. To validate that the selected covariates do not overfit the patient data, we use leave-one-out cross-validation (LOOCV) on the dataset and predict linear estimators [3] as $\hat{\eta}_i = h_i \cdot \beta^{-i}$ and $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2, \ldots \hat{\eta}_N)$ to compute C-Index [7]. Here, β^{-i} is estimated on the whole patient set minus patient i.

(a) Clusters samples

(b) Survival (Kaplan-Meier estimator)
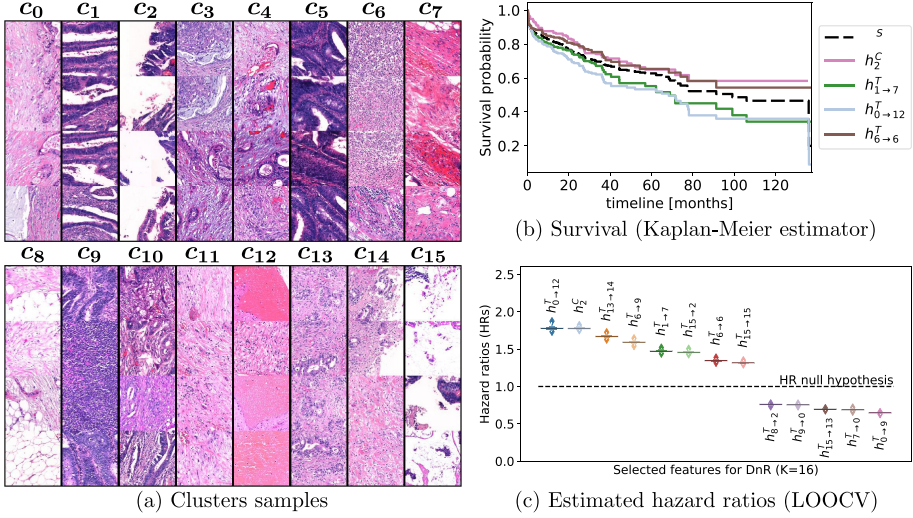
(c) Estimated hazard ratios (LOOCV)

Fig. 2. Comparison of estimated clusters representation. (a) Survival results and estimated hazard ratios over LOOCV (b–c). For Kaplan-Meier estimators, we choose a subset of curves that do not overlap too much for better visualization.

3   Experimental Results

Dataset. We use a set of 660 in-house unlabeled WSIs of CRC stained with hematoxylin and eosin (H&E). The slides are linked to a total of 374 unique patients diagnosed with adenocarcinoma. The dataset was filtered such that we exclude cases of mucinous adenocarcinoma in which their features are considered independent with respect to standard adenocarcinoma. A set of histopathological features (HFs) is associated with each patient entry (i.e. depth of invasion, pT, etc.). The survival time is defined as the period between resection of the tissue (operation) and the event occurrence (death of the patient). We denote $\mathcal{D}^S$ as the dataset that contains slides images and $\mathcal{D}^{S \cap HF}$ as the dataset that contains both information of the HFs and slides for each patient. Note that $|\mathcal{D}^{S \cap HF}| < |\mathcal{D}^S|$ as some patients have missing HFs and were excluded.

Experimental Settings. We use ResNet-18 for the encoder where the input layer is updated to support 2 input channels. The latent space has dimensions $d = 512$. The decoder is a succession of convolutional layers, ReLUs, and up-samplings (bicubic). The model was trained with the reconstruction loss $\mathcal{L}_{\mathrm{MSE}}$ for 20 epochs with early stopping. We use Adam optimizer $\beta = (0.9, 0.999)$ and learning rate, $lr = 1\mathrm{e}{-3}$. Then, we add $\mathcal{L}_{\mathrm{Divide}}$ for an additional 20 epochs with $\lambda = 1\mathrm{e}{-3}$ and $\tau = 0.5$. Finally, we go through 3 additional rounds using $\mathcal{L}_{\mathrm{Rule}}$ while raising the entropy threshold between each round.

Table 1. Multivariate survival analysis for the proposed approach and baselines. K and N_{feat} denote the number of clusters and the number of features that achieve statistical relevance when performing forward selection ($p < 0.05$). n denotes the number of patient in each set. Brier and Concordance Index are indicators of the performance.

Method	K	N_{feat}	$\mathcal{D}^{S \cap HF}$ ($n = 253$)		$\mathcal{D}^{S}$ ($n = 374$)	
			Brier [1]	C-Index [7]	Brier	C-Index
Histo. features (HFs)		8	0.2896	0.6076***	–	–
DCS	8	3	0.2840	0.5398+	0.2848	0.5562**
DCA† [17]	8	2	0.2887	0.5452**	0.2850	0.5555***
DEC† [20]	8	4	0.2884	0.6089**	0.2830	0.5765**
DnR w/o $\mathcal{L}_{\text{Divide}}$, $\mathcal{L}_{\text{Rule}}$	8	3	0.2870	0.6070*	0.2824	0.6040***
DnR w/o $\mathcal{L}_{\text{Rule}}$	8	3	0.2828	0.5951**	0.2840	0.5919***
DnR (ours)	8	4	0.2854	0.6107*	0.2832	0.6243***
DCS	16	9	0.2934	0.6073	0.2879	0.6464***
DCA† [17]	16	7	0.2827	0.6246+	0.2852	0.6322**
DEC† [20]	16	7	**0.2758**	0.6410**	0.2763	0.6426***
DnR w/o $\mathcal{L}_{\text{Divide}}$, $\mathcal{L}_{\text{Rule}}$	16	5	0.2819	0.6364*	0.2795	0.6324***
DnR w/o $\mathcal{L}_{\text{Rule}}$	16	10	0.3006	0.6207+	0.2934	0.6468***
DnR (ours)	16	13	0.2849	**0.6736****	**0.2725**	**0.6943****

† Autoencoder is replaced with the self-supervised objective function.
+ $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (log-rank test).

Clustered Embedding Space. We fit SPKM with $K = 8$ and $K = 16$. The sampled tiles for each cluster are presented in Fig. 2. Clusters demonstrate different tumor and stroma interactions (c_0, c_1, c_5, c_9), inflammatory tissues (c_6), muscles and large vessels (c_7), collagen and small vessels (c_8), blood and veins (c_{11}) or connective tissues (c_{12}). Some clusters do not directly represent the type of tissue but rather the positioning information such as c_2, which describe the edge of the WSI.

Ablation Study and Survival Analysis Results. We build our survival features (Eq. 5) on top of the predicted clusters, and their contribution is evaluated using Eq. 6. In Table 1, we observe that our model outperforms previous approaches by a safe 5% margin on C-Index [7]. The second step of the learning (DnR w/o $\mathcal{L}_{\text{Rule}}$) tends to decrease the prediction score. Such behavior is to be expected as the additional term ($\mathcal{L}_{\text{Divide}}$) will scatter the data and focus on self instance representation. When $\mathcal{L}_{\text{Rule}}$ is then introduced, the model can restructure the embedding by linking similar instances. Also, we observe an augmentation in features, N_{feat}, that achieve statistical relevance for prognosis as we go through our learning procedure (for $K = 16$), which proves that our proposed framework can model more subtle patches interactions. We show in Fig. 2

the distribution of hazard ratios for all models (from LOOCV) and the Kaplan-Meier estimator [11] for a subset of the selected covariates. In the best case, we identify 13 features that contribute to the survival outcome of the patients. For example, the interaction between blood vessels and tumor stroma $(h_{1\to7}^T)$ is linked to a lower survival outcome. A similar trend observed in the relation between tumor stroma and connective tissues $(h_{0\to12}^T)$.

4 Conclusion

We have proposed a self-supervised learning method that offers a new approach to learn histopathological patterns within cancerous tissue regions. Our model presents a novel way to model the interactions between tumor-related image regions and tackles the inherent overfitting problem to predict patient outcome. Our method surpasses all previous baseline methods and histopathological features and achieves state-of-the-art results, i.e., in C-Index without any data-specific annotation. Ablation studies also show the importance of different components of our method and the relevance of combining them. We envision the broad application of our approach for clinical prognostic stratification improvement.

References

1. Brier, G.W.: Verification of forecasts expressed in terms of probability. Mon. Weather Rev. **78**(1), 1–3 (1950)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
3. Dai, B., Breheny, P.: Cross validation approaches for penalized Cox regression. arXiv preprint arXiv:1905.10432 (2019)
4. Dercksen, K., Bulten, W., Litjens, G.: Dealing with label scarcity in computational pathology: a use case in prostate cancer classification. arXiv preprint arXiv:1905.06820 (2019)
5. Efron, B.: The efficiency of Cox's likelihood function for censored data. J. Am. Stat. Assoc. **72**(359), 557–565 (1977)
6. Fouad, S., Randell, D., Galton, A., Mehanna, H., Landini, G.: Unsupervised morphological segmentation of tissue compartments in histopathological images. PLoS ONE **12**(11), e0188717 (2017)
7. Harrell Jr., F.E., Lee, K.L., Mark, D.B.: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat. Med. **15**(4), 361–387 (1996)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Hosmer Jr., D.W., Lemeshow, S., May, S.: Applied Survival Analysis: Regression Modeling of Time-to-Event Data, vol. 618. Wiley, Hoboken (2011)

11. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. **53**(282), 457–481 (1958)
12. Kather, J.N., Halama, N., Marx, A.: 100,000 histological images of human colorectal cancer and healthy tissue, April 2018. https://doi.org/10.5281/zenodo.1214456
13. Kather, J.N., et al.: Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLoS Med. **16**(1), e1002730 (2019)
14. Li, R., Yao, J., Zhu, X., Li, Y., Huang, J.: Graph CNN for survival analysis on whole slide pathological images. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 174–182. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_20
15. Macenko, M., et al.: A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1107–1110, June 2009
16. Moriya, T., et al.: Unsupervised pathology image segmentation using representation learning with spherical k-means. In: Medical Imaging 2018: Digital Pathology, vol. 10581, p. 1058111. International Society for Optics and Photonics (2018)
17. Muhammad, H., et al.: Unsupervised subtyping of cholangiocarcinoma using a deep clustering convolutional autoencoder. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 604–612. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_67
18. Vahadane, A., et al.: Structure-preserving color normalization and sparse stain separation for histological images. IEEE Trans. Med. Imaging **35**(8), 1962–1971 (2016)
19. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via nonparametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)
20. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: International Conference on Machine Learning, pp. 478–487 (2016)
21. Zhong, S.: Efficient online spherical k-means clustering. In: Proceedings of 2005 IEEE International Joint Conference on Neural Networks, vol. 5, pp. 3180–3185. IEEE (2005)
22. Zhu, X., Yao, J., Zhu, F., Huang, J.: WSISA: making survival prediction from whole slide histopathological images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7234–7242 (2017)
23. Zhuang, C., Zhai, A.L., Yamins, D.: Local aggregation for unsupervised learning of visual embeddings. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6002–6012 (2019)