Supervision meets Self-supervision: A Deep Multitask Network for Colorectal Cancer Histopathological Analysis

Aritra Marik, Soumitri Chattopadhyay, and Pawan Kumar Singh*

Department of Information Technology, Jadavpur University, Kolkata, INDIA
{aritramarik2002, soumitri.chattopadhyay, pawansingh.ju}@gmail.com

Abstract. Colorectal cancer is one of the most common types of cancer worldwide and the leading cause of death due to cancer. As such, an early detection and diagnosis is of paramount importance, which is however, limited due to insufficient medical practitioners available for large-scale histopathological screening. This demands for a reliable computer-aided framework that can automatically analyse histopathological slide images and assist pathologists in quick decision-making. To this end, we propose a novel deep learning framework that combines supervised learning with self-supervision for robust learning of histopathological features from colorectal tissue images. Specifically, our framework comprises a multitask training pipeline using deep metric learning that learns the embedding space using triplet loss, which is augmented using a self-supervised image reconstruction module that enhances learning of pixel-level texture features. The downstream classification is done by extracting features using the pre-trained encoder and feeding them into a support vector machine classifier. We perform qualitative and quantitative analysis on a publicly available colorectal cancer histopathology dataset, as well as compare the proposed framework against some state-of-the-art works, where the model is found to outperform several existing works in literature. The source codes of the proposed method can be found at: https://github.com/soumitri2001/DMTL-CRCH.

Keywords: Deep metric learning · Self-supervision · Image reconstruction · Colorectal cancer · Histopathology · CRCH dataset

1 Introduction

The uncontrolled cell division in certain epithelial tissues of the colon and rectum in the large intestine, due to mutation in certain genes, results in colorectal carcinoma or colorectal cancer (CRC) [27]. As of 2020, CRC is one of the leading causes of cancer related deaths in the world, accounting for around 11% of the cancer patients worldwide. The early detection of CRC can increase the survival rates by around 90% [28] which explains the need for its early detection.

*Corresponding Author

Traditionally, the standard diagnosis procedures for CRC, which are carried out by pathologists, include faecal occult blood test (FOBT) and faecal immunochemical test (FIT) for detection of hemoglobin in the blood, followed by colonoscopy for studying the cause behind it. However, the manual observation and diagnosis is susceptible to observer based variations. The necessity of higher efficiency in colorectal cancer histopathological (CRCH) analysis from the tissue images along with the extraction of underlying features from those images has paved the way for deep learning based methods in the literature.

Medical imaging-based histopathological analysis has been approached classically using traditional machine learning and handcrafted feature extraction [30,32], which, however, fail to capture complex underlying patterns within the image data. As such, deep learning methods, particularly CNNs [13], have gained popularity due to their capability in recognising salient and translationally invariant image features for robust classification and investigation. CNNs have been successfully applied to several facets of medical imaging including histopathology [31], chest X-rays [5] and CT scans [18].

In this paper, we propose a novel deep learning framework combining supervised and self-supervised techniques for feature learning for the purpose of CRCH analysis. The supervised learning framework is guided by deep metric learning using triplet loss for learning a discriminative embedding space along with self-supervised image reconstruction for learning pixel-level tissue image features. To the best of our knowledge, such a multitask pipeline has not been approached yet in regard to histopathological analysis. The downstream classification is done by feeding the extracted features into an SVM classifier [8].

The main contributions of this work may be summarized as follows:

1. A novel multitask training pipeline is proposed for learning robust representations of colorectal histopathological images.
2. Deep metric learning is used to learn a discriminative embedding space, augmented by a self-supervised image reconstruction module that enforces learning of pixel-level information for enhanced histopathological analysis.
3. Once trained, the encoder is used off-the-shelf to extract features which are used to train an SVM classifier [8] for the downstream classification. Upon comparison, the proposed framework outperforms several state-of-the-art works in literature on a publicly available CRCH dataset [16].

The rest of the paper is organised as follows. Section 2 discusses some recent works that are relevant to our proposed method. Section 3 describes the proposed multitask learning pipeline in detail. Section 4 outlines the experimental evaluation of the method on a public dataset. Finally, Section 5 concludes and also discusses future extensions of the present research.

2 Related Works

Colorectal Cancer Histopathology: The work in tissue-image classification in past few years has primarily been under two categories, texture feature-based methods and deep learning methods. The first study on multi texture-feature analysis for colorectal tissue classification was presented by Kather *et al.* [16], which obtained the highest accuracy of 87.4% in the multi-class classification task. The study also presented a new dataset of 5000 histological images of human CRC including 8 tissue classes (described in Table 1), the dataset we have used in our work. Among deep learning-based approaches, a bi-linear CNN model was proposed by [31] which extracted and fused features from stain decomposed histological images, achieving an accuracy of 92.6%. The CNN architecture proposed by [7] highlighted the importance of stain normalization in the literature, although achieving an accuracy of only 79.66%. [25] proposed a method in which the classifier performance was enhanced by a fine-tuned CNN model to an accuracy of 92.74% on the multi-class classification task. [21] first introduced the concept of fine-tuning a deep learning model which was pre-trained on ImageNet with respect to the current dataset, while coping with the limited data available with respect to biomedical imaging. They used 108 different combinations of feature extractors and classifiers, out of which the pre-trained denseNet-169 model and the SVM classifier obtained an accuracy if 92.08%. It is to be noted that in all of the aforementioned works, the CRCH dataset by [16] has been used. To this end, we bring to table a novel and efficient multitask network for robust CRCH classification.

Deep Metric Learning: Metric learning [17] is based on the principle of similarity between data samples. Specifically, deep metric learning utilizes deep architectures by learning embedded feature similarity through training on raw data. Metric learning has been extensively applied in three-dimensional modelling [9], medical imaging [2], facial recognition [19,26,15], and signature verification [4]. The supervised module of deep metric learning of proposed architecture has been inspired from the triplet network proposed by [14] for learning distance-based metric embeddings of a multi-class dataset.

Self-supervised Learning: Self-supervised learning [22] has gained tremendous popularity in recent years due to it capability of learning very good quality representations without the need of explicit supervision. Such techniques include contrastive learning [6,12], generation/reconstruction-based [3,10], clustering-based [1] and so on. We take inspiration from self-supervised literature to augment our metric learning model with a reconstruction network to learn minute pixel-level visual information.

3 Methodology

Overview: We intend to combine supervised learning with self-supervision as a pre-training paradigm for robust CRCH image analysis. Specifically, we design a metric learning framework that learns to maximise intra-class similarity and simultaneously distinguishing samples from a different class. Further, to enhance learning of image-level texture features we introduce an image reconstruction decoder that takes in the embedding of a corrupted version of the histopathological image and tries to output its original version. The corrupted version is produced by random spatial transformations such as blurring or dropping randomly distributed pixels. A reconstruction loss is used to train this branch. Note that the encoder weights are shared across the two modules and gradients flow throughout the architecture. Once the network is trained, the frozen encoder is used to extract image features, which are used to train a SVM classifier for the final classification. Figure 1 shows the overall architecture of the proposed pipeline.

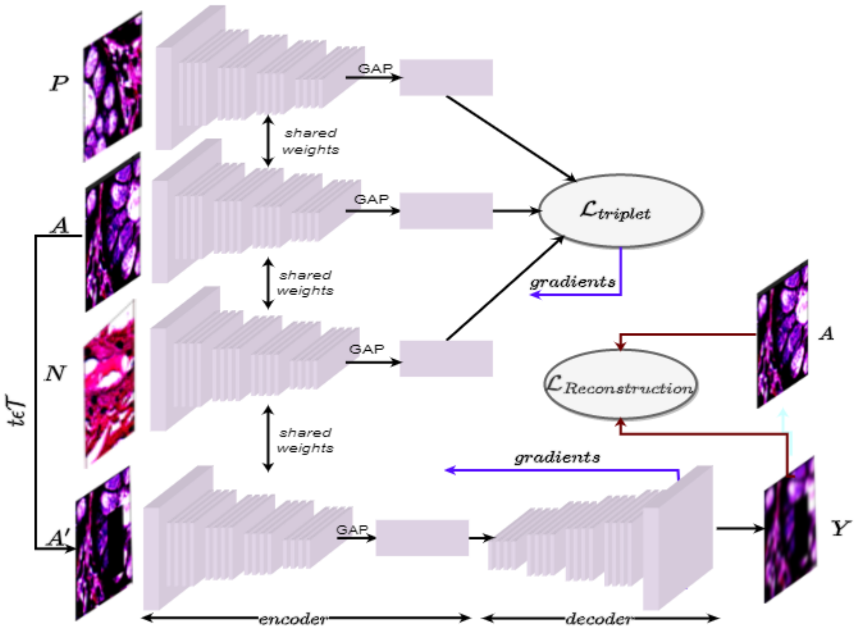


Fig. 1: Schematic representation of architecture proposed in this study.

3.1 Deep Metric Learning

The main purpose of metric learning is to learn a discriminative embedding space wherein samples belonging to the same class are close together while those from different classes are farther apart. Typically, metric learning losses aim at minimizing the distance metric between intra-class or "positive" pairs while maximizing the same between inter-class or "negative" pairs. In this work, we have used the state-of-the-art Siamese convolutional networks [14] that comprise three CNN branches with shared weights, each corresponding to the anchor, positive and negative images comprising the triplet. The network is trained using a simple triplet loss objective. To keep our framework simple and lightweight, we have used an ImageNet pre-trained ResNet-18 [13] network as the encoder backbone, with the output embedding dimension being set to 512.

$$\mathcal{L}_{triplet} = \sum_{i=1}^{K} max(0, \|\mathcal{F}(A^{(i)}) - \mathcal{F}(P^{(i)})\|^2 - \|\mathcal{F}(A^{(i)}) - \mathcal{F}(N^{(i)})\|^2 + \mu) \quad (1)$$

where, (A, P, N) constitute a triplet, $\mathcal{F}(\cdot)$ denotes the shared encoder and μ is the permissible margin value of the triplet loss, set to 0.2 experimentally.

3.2 Image Reconstruction Network

Image reconstruction aims at learning pixel-level information so as to restore its original version from a corrupted one. We achieve this by introducing a decoder module that starts from an embedding vector into a series of upsampling layers to finally yield the output image dimensions. We have used the state-of-the-art U-Net decoder [24] network, *excluding* the skip connections from the encoder. A corrupted view of the original "anchor" image is formed using random image transformations such as affine transformations, blurring or dropping pixels, and is passed through shared encoder network, the output of which is then passed through the decoder. A reconstruction loss is employed between the output image (Y) and the original image (A) before corruption. In our work, we have used the simple mean squared error (MSE) as the reconstruction loss to train our network.

$$\mathcal{L}_{reconstruction} = \frac{1}{M} \sum_{i=1}^{K} \sum_{j=1}^{M} \|A_j^{(i)} - Y_j^{(i)}\|_2^2 \quad (2)$$

Combining Equations 1 and 2, the overall training loss objective can be written as follows:

$$\mathcal{L}_{train} = \lambda \cdot \mathcal{L}_{triplet} + \mathcal{L}_{reconstruction} \quad (3)$$

Here, λ is a hyperparameter that is used for relative weighting of the losses. Experimentally, it has been set to 10.

3.3 Final Classification

Once the model is fully trained, we use the encoder while keeping its weights fixed and extract features from the images which are then fed into an SVM classifier [8] so as to train it for the final classification step. The SVM classifier is an supervised algorithm that aims at finding the optimal hyperplane(s) that separate the respective classes by mapping the samples onto a space such that the distance between class boundaries is maximized. The unseen sample features are likewise mapped on the sample space and thus, classes are predicted based on where they fall.

4 Results and Discussion

Dataset Description: We train and evaluate our proposed framework on a publicly available dataset by Kather *et al.* [16] comprising 5000 colorectal histopathological images uniformly distributed across 8 classes. Each image is of dimensions 150×150 px and belongs to exactly one category among those in Table 1. For our purpose, we split the dataset as 0.75/0.25 for train/test respectively.

Table 1: Image categories in the CRCH dataset [16] used in this research. Each class contains 625 images, which is split as 0.75/0.25 for train/test respectively.

Label	Category
0	Tumour Epithelium
1	Simple Stroma
2	Complex Stroma
3	Immune Cells
4	Debris
5	Normal Mucosa
6	Adipose Tissue
7	Background

Implementation Details: Our model has been implemented in PyTorch [23] on a 12GB K80 Nvidia GPU. The encoder-decoder framework was trained for 200 epochs using the Stochastic Gradient Descent (SGD) optimizer [29] with a learning rate of 0.01. To keep the pipeline fairly straightforward, we chose the backbone encoder as the ImageNet pre-trained ResNet-18 [13] that outputs an embedding of dimension 512. The decoder used is the state-of-the-art U-Net decoder [24] network *without* the skip connections from the encoder. Each image was resized to 256×256 *px* using bilinear interpolation before being passed through the encoder, the batch size being set to 32.

Evaluation Metrics: The four metrics used used for evaluating our proposed method on the CRCH dataset [16] are, namely, *Accuracy*, *Precision*, *Recall*, and *F1-Score*. The formulas for the metrics, derived from a confusion matrix, C, are given in the following Equations 4, 5, 6, 7.

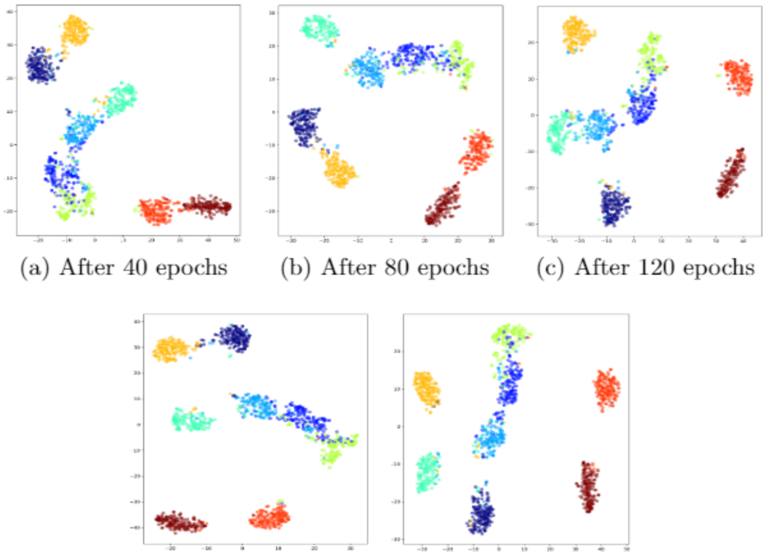
$$Accuracy = \frac{\sum_{i=1}^{N} C_{ii}}{\sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij}} \tag{4}$$

$$Precision_i = \frac{C_{ii}}{\sum_{j=1}^{N} C_{ji}} \tag{5}$$

$$Recall_i = \frac{C_lii}{\sum_{j=1}^{N} C_{ij}} \tag{6}$$

$$F1 - Score_i = \frac{2}{\frac{1}{Precision_i} + \frac{1}{Recall_i}} \tag{7}$$

Here, N signifies the number of classes in the respective dataset.



(a) After 40 epochs (b) After 80 epochs (c) After 120 epochs

(d) After 160 epochs (e) After 200 epochs

Fig. 2: t-SNE plots obtained by the encoder at different stages of the training process.

4.1 Qualitative Analysis

We analyse the discriminative embedding space learned by the joint encoder-decoder model by visualising the embedding space of the extracted features in two-dimensional plane using t-distributed stochastic neighbourhood embedding (t-SNE) [20] at intervals during the training process. The t-SNE plots have been put in Figure 2. It is evident from the plots that over the training epochs, the embedding space gets more and more discriminative and the classes get fairly well separated, thus qualitatively suggesting that a robust representation has been learned by the pre-training paradigm.

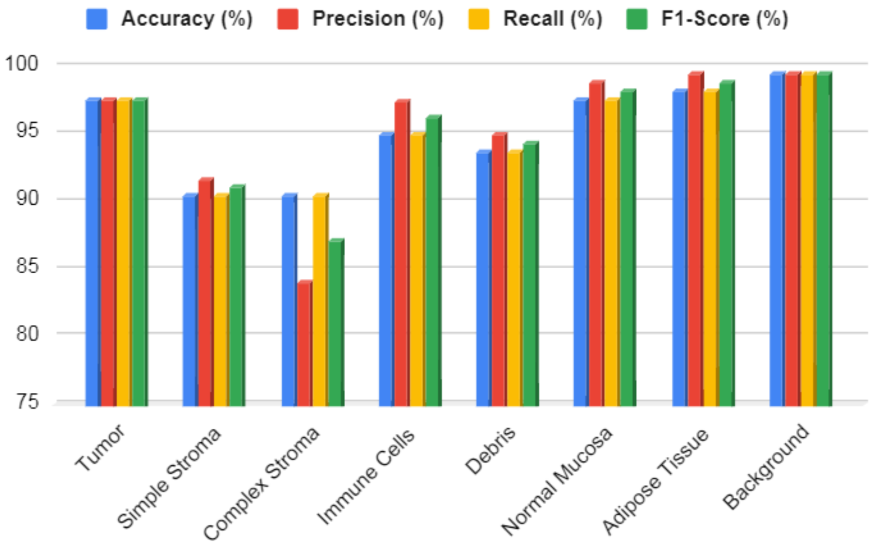


Fig. 3: Class-wise evaluation metrics obtained by the proposed method on the CRCH dataset.

4.2 Comparison with state-of-the-art

Performance comparison between our proposed method and other state-of-the-art methods for the CRCH dataset [16] has been provided in Table 2. It can be examined that our proposed framework outperforms all other existing state-of-the-art methods by a significant margin. It may be taken into consideration that some of the previously existing works in the literature reported accuracy as the only evaluation metric. It is insufficient and does not provide enough insights regarding false positives and true negatives. Since CRCH analysis a multi-class classification task, the absence of sufficient evaluation metrics makes the works unreliable. On the other hand, our proposed pipeline achieves commendable performance on all of the mentioned metrics considered.

4.3 Ablation Study

Since our proposed framework comprises two components i.e. a supervised metric learning model and a self-supervised image reconstruction model, we quantitatively determine the contribution of each of them by performing an ablation study. We define the baselines for the same as follows:

- **Triplet**: This denotes the metric learning model alone excluding the decoder part, such that the encoder is trained using triplet loss only. All other experimental parameters are kept identical. Please refer to Section 3.1 for full details.

Table 2: Comparison of the proposed framework with state-of-the-art methods on publicly available CRCH dataset.

Method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Ciompi *et al.* [7]	79.66	–	–	–
Kather *et al.* [16]	87.40	–	–	–
Ohata *et al.* [21]	92.08	–	–	92.12
Wang *et al.* [31]	92.60	–	92.80	–
Sabol *et al.* [25]	92.74	92.50	92.76	92.64
Ghosh *et al.* [11]	92.83	92.83	93.11	92.97
Proposed method	**95.22**	**95.34**	**95.22**	**95.26**

- **Reconstruction**: This denotes the encoder-decoder model which takes in a randomly transformed histopathological image and tries to reconstruct the original image from it, trained using reconstruction loss. All other experimental parameters are kept identical. Please refer to Section 3.2 for full details.

The overall performance results are shown in Table 3. From Table 3, it can be observed that the triplet model shows a better performance than the reconstruction baseline, thereby highlighting the importance of optimizing the embedding space for better discrimination. Furthermore, it can also be noticed that combining the two components together i.e., our proposed framework improves the classification performance of the individual stand-alone baselines, affirming the contributions of the respective modules.

Table 3: Ablation study on the proposed pipeline.

Method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Triplet	94.58	94.65	94.58	94.61
Reconstruction	93.39	93.50	93.39	93.41
Triplet+Reconstruction (Proposed)	**95.22**	**95.34**	**95.22**	**95.26**

5 Conclusion & Future Work

In this work, we present a novel training strategy that leverages supervised metric learning and self-supervised image reconstruction for robust representation learning of histopathological images. While metric learning optimizes the embedding space, the reconstruction module enforces pixel-level information learning, which aids the overall training pipeline, improving the downstream evaluation. We have analysed our framework qualitatively as well as compared with several existing state-of-the-art works in literature, along with a suitable ablation study to investigate the contributions of the respective modules. The results highlight the prowess of the proposed method for image classification on CRCH dataset.

However, our work does have certain limitations, such as the relatively poor performance for the class 'Complex Stroma' as shown in Figure 3. We intend to investigate this in future. Possible extensions of our work maybe on the lines of improving the respective modules using alternate metric learning losses, or using VAE/GAN based models for reconstruction, and so on. Further, this work provides a foundation for CRCH representation learning and paves the way towards contrastive learning based self-supervised approaches [6,12] as large-scale supervised learning gradually becomes infeasible. We intend to explore on these lines as well in our future works.

References

1. Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., Tran, D.: Self-supervised learning by cross-modal audio-video clustering. arXiv preprint arXiv:1911.12667 (2019)
2. Annarumma, M., Montana, G.: Deep metric learning for multi-labelled radiographs. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing (2018)
3. Atito, S., Awais, M., Kittler, J.: Sit: Self-supervised vision transformer. arXiv preprint arXiv: Arxiv-2104.03602 (2021)
4. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. IJPRAI, World Scientific (1993)
5. Chattopadhyay, S., Kundu, R., Singh, P.K., Mirjalili, S., Sarkar, R.: Pneumonia detection from lung x-ray images using local search aided sine cosine algorithm based deep feature selection method. International Journal of Intelligent Systems (2021)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
7. Ciompi, F., Geessink, O., Bejnordi, B.E., de Souza, G.S.e.a.: The importance of stain normalization in colorectal tissue classification with convolutional networks. In: IEEE ISBI (2017)
8. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning (1995)
9. Dai, G., Xie, J., Zhu, F., Fang, Y.: Deep correlated metric learning for sketch-based 3d shape retrieval. In: AAAI (2017)
10. Deepak, P., Philipp, K., Jeff, D., Trevor, D., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR (2016)
11. Ghosh, S., Bandyopadhyay, A., Sahay, S., Ghosh, R., Kundu, I., Santosh, K.: Colorectal histology tumor detection using ensemble deep neural network. EAAI, Elsevier (2021)
12. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR (2016)
14. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International workshop on similarity-based pattern recognition (2015)
15. Hu, J., Lu, J., Tan, Y.P.: Discriminative deep metric learning for face verification in the wild. In: IEEE CVPR (2014)

16. Kather, J.N., Weis, C.A., Bianconi, F., et al.: Multi-class texture analysis in colorectal cancer histology. Scientific Reports, Nature (2016)
17. Kaya, M., Bilge, H.Ş.: Deep metric learning: A survey. Symmetry (2019)
18. Kundu, R., Basak, H., Singh, P.K., Ahmadian, A., Ferrara, M., Sarkar, R.: Fuzzy rank-based fusion of cnn models using gompertz function for screening covid-19 ct-scans. Scientific Reports. Nature (2021)
19. Liu, J., Deng, Y., Bai, T., Wei, Z., Huang, C.: Targeting ultimate accuracy: Face recognition via deep embedding. arXiv preprint arXiv:1506.07310 (2015)
20. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research (2008)
21. Ohata, E.F., Chagas, J.V.S.d., Bezerra, G.M., Hassan, M.M., de Albuquerque, V.H.C., Filho, P.P.R.: A novel transfer learning approach for the classification of histological images of colorectal cancer. The Journal of Supercomputing, Springer (2021)
22. Ohri, K., Kumar, M.: Review on self-supervised image recognition using deep neural networks. Knowledge-Based Systems (2021)
23. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
25. Sabol, P., Sinčák, P., Hartono, P., Kočan, P., et al.: Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. JBI, Elsevier (2020)
26. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR (2015)
27. Society, A.C.: What is colorectal cancer ? American Cancer Society (2020), www.cancer.org/cancer/colon-rectal-cancer/about/what-is-colorectal-cancer.html
28. Society, A.C.: Survival rates for colorectal cancer. American Cancer Society (2021), www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html
29. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: ICML (2013)
30. Takamatsu, M., Yamamoto, N., Kawachi, H., Chino, A., Saito, S., Ueno, M., Ishikawa, Y., Takazawa, Y., Takeuchi, K.: Prediction of early colorectal cancer metastasis by machine learning using digital slide images. CMPB (2019)
31. Wang, C., Shi, J., Zhang, Q., Ying, S.: Histopathological image classification with bilinear convolutional neural networks. In: IEEE EMBC (2017)
32. Xu, Y., Ju, L., Tong, J., Zhou, C.M., Yang, J.J.: Machine learning algorithms for predicting the recurrence of stage iv colorectal cancer after tumor resection. Scientific Reports, Nature (2020)