2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871553

Attention-based multiple instance learning with self-supervision to predict microsatellite instability in colorectal cancer from histology whole-slide images

Jacob S. Leiby, Jie Hao, Gyeong Hoon Kang, Ji Won Park*, and Dokyoon Kim*

Abstract— **Microsatellite instability (MSI) is a clinically important characteristic of colorectal cancer. Standard diagnosis of MSI is performed via genetic analyses, however these tests are not always included in routine care. Histopathology whole-slide images (WSIs) are the gold-standard for colorectal cancer diagnosis and are routinely collected. This study develops a model to predict MSI directly from WSIs. Making use of both weakly- and self-supervised deep learning techniques, the proposed model shows improved performance over conventional deep learning models. Additionally, the proposed framework allows for visual interpretation of model decisions. These results are validated in internal and external testing datasets.**

I. Introduction

Microsatellite instability (MSI) is a state of genetic hypermutation caused by defects in the mismatch repair system. It occurs in roughly 15% of all colorectal cancers [1]. MSI is clinically relevant in colorectal cancer as tumors expressing this pattern have shown the highest response rates to immunotherapies as well as improved overall survival [2]. MSI can be determined using genetic analyses [3]. However, these analyses are often limited to larger tertiary care centers and may add additional time and cost during diagnosis [4, 5].

Manual inspection of hematoxylin and eosin-stained (H&E) tissue slides remains the gold-standard in solid-tumor cancer diagnosis. With the adoption and validation of digitizing these slides as whole-slide images (WSIs) for manual examination [6], rich data are available for computational analyses beyond simple diagnosis. In recent years, deep learning models have been developed and applied to WSI analysis for a variety of tasks including mitosis detection, survival prediction, and predicting MSI [4, 7-10]. These studies have shown promising results in using deep learning models to objectively determine cancer characteristics from WSIs.

Histology WSIs are large (upwards of 100,000 × 100,000 pixels), so it is necessary to tile the WSI into smaller images. Thus, for each patient, a collection of tile instances represents the full WSI. MSI is a patient-level attribute, and therefore the label is associated with the entire WSI. Conventionally, each tile will inherit the slide-level label. However, not all regions of the WSI are informative of MSI status. Pathologic indicators associated with MSI include presence of mucin, poor or undifferentiated histology, and presence of tumor infiltrating lymphocytes (TIL) [11]. Due to intra-tumor heterogeneity, it is unknown which tiles will contain these informative features without prior annotation. Therefore, assigning the slide-level label to each tile is imprecise and can introduce error into training. One of the solutions to this is weakly-supervised learning, where the assumption of noisy or imprecise labels is incorporated into model training. Previous studies have used weakly-supervised learning for prediction of gestational age from placental biopsies and overall survival prediction from tumor histology slides [12, 13].

Due to the difficulty of curating annotations for histology datasets and the limited availability of labeled data, it is also difficult to train a model to extract meaningful representations of WSIs. One solution to this is to use self-supervised learning (SSL), where the data itself provides the supervision. In SSL, a model is trained to generate representative embeddings of the input data without the use of label information. Several studies have shown the benefit of using SSL to train models for downstream tasks [14, 15].

To overcome the challenge of noisy labels and the challenge of representation learning, we propose a model that integrates SSL into a weakly-supervised learning framework. Specifically, our model uses attention-based multiple instance learning with an auxiliary contrastive representation loss to predict MSI in colorectal cancer. Code is available at www.github.com/leibyj/WSI_attention_learning.

II. Materials and Methods

A. Deep Learning for Histopathological Analysis

Due to the large size of histology WSIs and the restrictions of input dimension for deep learning models, it is necessary to tile the WSI into smaller images to be used as model input. In this study, the tumor regions of WSIs were segmented into non-overlapping tiles of size 224 × 224 pixels.

This work was supported in part by the National Institutes of Health [U01 AG068057 and R01 NL012535]

J. S. Leiby is with the Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA.

J. Hao was with University of Pennsylvania, Philadelphia, PA, 19104, USA. She is now with the Institute of Medical Information at the Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China.

G. H. Kang is with the Department of Pathology, Seoul National University Hospital, Seoul National University College of Medicine, and Cancer Research Institute, Seoul National University, Seoul, South Korea.

* Co-corresponding authors.

J. W. Park is with the Department of Surgery, Seoul National University Hospital, Seoul National University College of Medicine, and Cancer Research Institute, Seoul National University, Seoul, South Korea. (sowisdom@gmail.com)

D. Kim is with Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, and Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, 19104, USA. (Dokyoon.kim@pennmedicine.upenn.edu)

This study was conducted in accordance with the Declaration of Helsinki. Publicly available data do not require IRB review.

B. Multiple Instance Learning

Multiple instance learning (MIL) is a weakly-supervised learning framework where instead of assigning a label to each instance, the instances are grouped into collections, *bags,* and each bag is assigned a label [16]. Thus, the individual label of each instance within a bag is unknown. This framework is well-suited for WSI analysis, as often the label is attributed to the full slide and needs to be inherited by the tiles, leading to potentially noisy labels. By forming bags of instances that inherit the label, the noise is reduced under the assumption that at least one instance in the bag is informative of the true slide-level label.

A recently proposed method to aggregate instance-level data into a bag-level representation is attention-based pooling [17]. Under this framework, a feature embedding and a learnable attention weight are generated for each instance in a bag. The attention-weighted summation of all instance embeddings then forms the aggregated bag representation. The bag representation is finally input into a classifier.

C. Contrastive Representation Learning

Contrastive learning is a form of SSL in which a model is trained to map data to informative embeddings by maximizing the similarity between certain embeddings and minimizing the similarity between others [18]. A recent framework was proposed that defines a simple model for contrastive learning of image representations [19]. It relies on extensive data augmentation to generate two correlated views (a positive pair) of a given datapoint. For each positive pair of samples (i, j) and their vector embeddings (z_i, z_j), the loss function is defined as:

$$\ell_{\mathrm{i,j}} = -\log \frac{\exp\left(\mathrm{sim}(z_i,z_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{I}_{[k\neq i]} \exp\left(\mathrm{sim}(z_i,z_k)/\tau\right)}, \qquad (1)$$

where $\mathbb{I}[k \neq i] \in \{0,1\}$ is an indicator function that evaluates to 1 iff $k \neq i$, and τ is a temperature parameter. $sim(\cdot, \cdot)$ represents a similarity function, i.e. cosine similarity.

D. Proposed Model

Our model employed both MIL and contrastive representation learning. The model input consists of a bag of tiles, $\boldsymbol{t} = (t_i, \ldots, t_b)$, for a bag size b. The proposed architecture consists of a feature extraction network, an attention sub-network, a weighted aggregation function, and a classifier (Fig. 1). Th e feature extraction network is the VGG19 architecture with parameters pretrained on ImageNet [20, 21]. The fully connected layers were removed and replaced with a single layer that encodes the feature map, f_i. An intermediate convolutional output is additionally used as input into the attention sub-network [12, 17]. This network consists of an average pooling layer, followed by two fully connected layers with rectified linear unit (ReLU) activation functions and outputs a single linear node, the attention score a_i. All tiles in a bag generate feature maps $\boldsymbol{f} = (f_i, \ldots, f_b)$ and attentions $\boldsymbol{a} = (a_i, \ldots, a_b)$. To generate the bag-level representation, the feature maps and attentions are pooled into an aggregate feature map, $\bar{f} = \frac{a^T f}{\sum a}$. The aggregate feature map is finally input into a classifier network consisting of two fully connected layers with ReLU activations. The output is a single node using the sigmoid activation function representing the probability, $\hat{y}$, of the bag being labeled as MSI. The bag-level loss function is the binary cross-entropy loss:

$$\mathcal{L}_{bce} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\hat{y} + (1-y_i)\log(1-\hat{y}). \qquad (2)$$

A recently proposed method adapts the contrastive representation learning loss function (1) to an MIL framework [14, 19]. Instead of performing data augmentation to generate positives pair embeddings of data, $(f_i, \bar{f}_j)$ is treated as positive pair where f_i is an instance-level feature map, and $\bar{f}_j$ is the associated bag-level aggregate feature map. Here, the instance feature maps of tiles within a bag and the associated aggregate feature map act as a pseudo-augmentation. N is the total number of tile instances and B is the number of bags in a minibatch. The resulting contrastive loss function is:

$$\mathcal{L}_{\mathrm{con}} = -\frac{1}{N}\sum_{i,j=1}^{N} \log \frac{\exp\left(\mathrm{sim}(f_i,\bar{f}_j)/\tau\right)}{\sum_{k=1}^{B} \mathbb{I}_{[k\neq j]} \exp\left(\mathrm{sim}(f_i,\bar{f}_k)/\tau\right)}. \qquad (3)$$

Our model training combined the cross-entropy (2) and the MIL contrastive loss (3), with τ = 0.5, into a single loss function. The model was trained end-to-end via stochastic gradient descent, with a bag size of eight and a batch size of twelve.

E. Datasets and Experimental Design

This study used datasets from The Cancer Genome Atlas Program (TCGA) and Pathology AI Platform (PAIP) [22, 23]. The WSIs from the TCGA colorectal cancer cohort were previously split into training and testing datasets (70% and 30%), and tiles were extracted and preprocessed to size 224×224 from the tumor region of each WSI [4]. The TCGA training dataset consists of 39 and 221 MSI and microsatellite stable (MSS) patients, respectively. The TCGA testing dataset consists of 26 MSI and 74 MSS patients. The PAIP dataset consists of 12 MSI and 35 MSI patients, and the tumor region of each WSI was also segmented into non-overlapping tiles of size 224×224. There were 700 tiles extracted from each WSI on average.

To evaluate the performance of our proposed model, we performed five-fold cross-validation within the TCGA training dataset. The hold-out fold was used to determine early stopping. Each model was then used to predict MSI status in the TCGA testing dataset. Patient level probabilities were calculated as the median probability of all bags created from the patient WSI.

Additionally, we retrained the model using the entire TCGA training dataset for the optimal number of epochs as determined by the cross-validation experiment. This model was then used to predict MSI status in the TCGA testing dataset as well as the external PAIP validation dataset.

III. RESULTS

To evaluate our proposed model, we compared it to two baseline models—VGG19 and ResNet18 [24], both with parameters pretrained on ImageNet and finetuned during training. These models were trained under a fully supervised learning framework. Additionally, we compared our model with and without the use of the contrastive loss function.

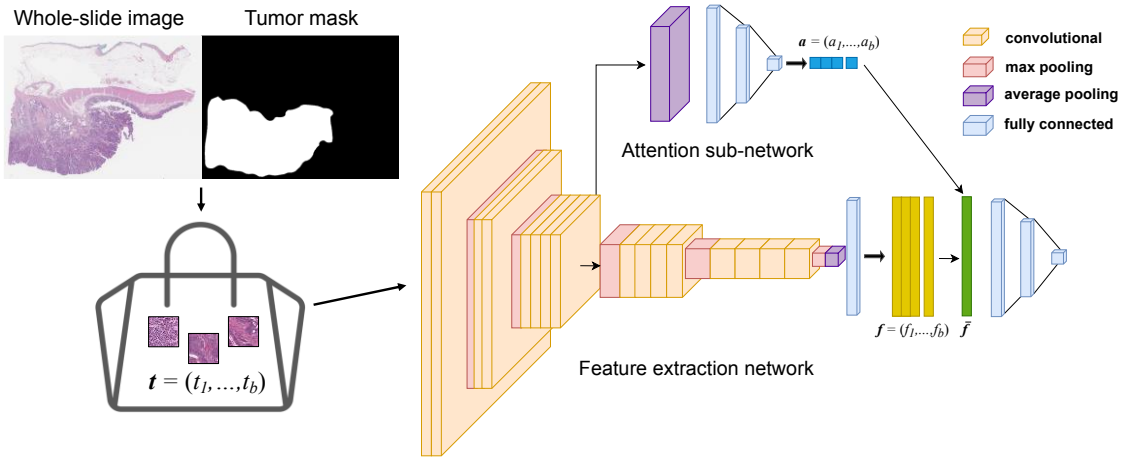

Figure 1. Proposed model overview. The tumor regions of WSIs are segmented into tiles and placed into bags as model input. The model consists of a feature extraction network, an attention sub-network, attention-weighted aggregation $\bar{f}$, and a classifier network.

A. *Model Performance*

Five-fold cross-validation was performed in the TCGA training dataset, and the resulting models were used to predict MSI in the TCGA testing dataset, shown in Table I. The proposed model outperforms both baseline models in terms of area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (AUPRC). We found that the addition of the contrastive loss function further improved AUC.

The final model was trained using the entire TCGA training dataset for the optimal number of epochs determined in cross-validation. The performance was evaluated in the internal TCGA testing set and the external PAIP dataset shown in Table II. We also see improvements in both performance metrics with the addition of attention-based aggregation. In the external dataset, we see that the inclusion of the contrastive loss leads to the top performing model.

B. *Interpretability*

With our proposed model we can visually interpret the informative tumor regions associated with outcome by analyzing the learned attention scores for each tile. For a well-predicted MSI patient, Fig. 2 shows an overlay of the tile attention score heatmap with the original WSI. The tiles with the highest attention scores from well-predicted MSI patients show previously known histologic features associated with MSI, including presence of mucin and stromal TIL (Fig. 2).

IV. Discussion

We proposed a method integrating self-supervised learning into a multiple instance learning framework for prediction of microsatellite instability from histology WSIs.

TABLE I. FIVE-FOLD CROSS-VALIDATION PERFORMANCE

Model	AUC	AUPRC
VGG19 Baseline	0.822 (0.02)	0.624 (0.05)
ResNet18 Baseline	0.828 (0.01)	0.689 (0.01)
VGG19 + Attention	0.861 (0.01)	**0.729 (0.02)**
VGG19 + Attention + Contrastive	**0.864 (0.01)**	0.690 (0.04)

TABLE II. INTERNAL AND EXTERNAL DATASET PERFORMANCE

Data	Model	AUC	AUPRC
TCGA	VGG19 Baseline	0.824	0.638
	ResNet18 Baseline	0.825	0.684
	VGG19 + Attention	0.821	**0.707**
	VGG19 + Attention + Contrastive	**0.876**	0.671
PAIP	VGG19 Baseline	0.770	0.512
	ResNet18 Baseline	0.686	0.434
	VGG19 + Attention	0.795	0.629
	VGG19 + Attention + Contrastive	**0.876**	**0.793**

We trained and evaluated our model using WSIs from the TCGA and PAIP colorectal cancer datasets. We compared our model to conventional fully supervised learning models.

We found that MIL outperformed fully supervised learning in both the internal and external datasets. By using an attention-weighted aggregation scheme, our model overcame intra-tumor heterogeneity by inherently learning which regions of the tumor were most informative of the outcome. Additionally, through learning attention scores, we can visually interpret model decisions. We found that the top-ranking tiles for MSI patients consisted of mucin and stromal TIL, which have previously been shown to be associated with MSI [11]. Thus, our model was able to effectively focus attention on important regions without the use of prior annotation. This is important in developing models for histopathological analysis as curating fine-grained annotations for WSIs is labor-intensive.

To learn meaningful embeddings of the data, we included a contrastive representation loss function in model training. SSL is especially crucial in biomedical image analysis, as it does not rely on labelled or annotated data, which is often limited. Additionally, many image models are pretrained using ImageNet, which does not include data resembling biomedical images [21]. Therefore, this type of self-supervised technique allowed for our model to learn improved representations of the data, and ultimately improve overall performance. We combined the contrastive representation and the weakly-supervised loss functions to jointly learn

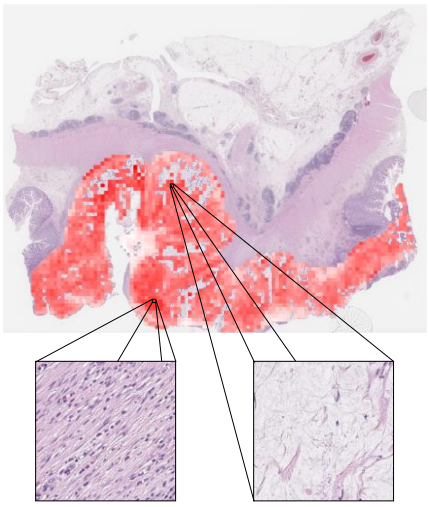

Figure 2. Heatmap showing tile attention scores for a well-predicted MSI patient, darker colors show higher attentions assigned to those regions. Example top ranked tiles showing features associated with MSI.

informative embeddings and predict outcomes in an end-to-end manner.

Deep learning models for histopathological analysis show great promise for improving the objectivity of information we can gather from WSIs. These types of models can serve as useful tools in clinical settings by assisting pathologists in histopathological assessments. Robustness and interpretability will be critical in future research to develop translational models. Thus, further studies are warranted to validate our model in additional independent cohorts.

REFERENCES

[1] Boland, C. R. & Goel, A. Microsatellite Instability in Colorectal Cancer. *Gastroenterology, Elsevier BV*, **2010**, 138, 2073-2087.e3

[2] Le, D. T.; Uram, J. N.; Wang, H.; Bartlett, B. R.; Kemberling, H.; Eyring, A. D.; Skora, A. D.; Luber, B. S.; Azad, N. S.; Laheru, D.; Biedrzycki, B.; Donehower, R. C.; Zaheer, A.; Fisher, G. A.; Crocenzi, T. S.; Lee, J. J.; Duffy, S. M.; Goldberg, R. M.; de la Chapelle, A.; Koshiji, M.; Bhaijee, F.; Huebner, T.; Hruban, R. H.; Wood, L. D.; Cuka, N.; Pardoll, D. M.; Papadopoulos, N.; Kinzler, K. W.; Zhou, S.; Cornish, T. C.; Taube, J. M.; Anders, R. A.; Eshleman, J. R.; Vogelstein, B. & Diaz, L. A. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine, Massachusetts Medical Society,* **2015**, *372*, 2509-2520

[3] Kather, J. N.; Halama, N. & Jaeger, D. Genomics and emerging biomarkers for immunotherapy of colorectal cancer. *Seminars in Cancer Biology, Elsevier BV,* **2018**, *52*, 189-197

[4] Kather, J. N.; Pearson, A. T.; Halama, N.; Jäger, D.; Krause, J.; Loosen, S. H.; Marx, A.; Boor, P.; Tacke, F.; Neumann, U. P.; Grabsch, H. I.; Yoshikawa, T.; Brenner, H.; Chang-Claude, J.; Hoffmeister, M.; Trautwein, C. & Luedde, T. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine, Springer Science and Business Media LLC,* **2019**, *25*, 1054-1056

[5] Singh, M. P.; Rai, S.; Pandey, A.; Singh, N. K. & Srivastava, S. Molecular subtypes of colorectal cancer: An emerging therapeutic opportunity for personalized medicine. *Genes & Diseases, Elsevier BV,* **2021**, *8*, 133-145

[6] Snead, D. R. J.; Tsang, Y.-W.; Meskiri, A.; Kimani, P. K.; Crossman, R.; Rajpoot, N. M.; Blessing, E.; Chen, K.; Gopalakrishnan, K.; Matthews, P.; Momtahan, N.; Read-Jones, S.; Sah, S.; Simmons, E.; Sinha, B.; Suortamo, S.; Yeo, Y.; Daly, H. E. & Cree, I. A. Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology, Wiley,* **2015**, *68*, 1063-1072

[7] Cireşan, D. C.; Giusti, A.; Gambardella, L. M. & Schmidhuber, J. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, Springer Berlin Heidelberg,* **2013**, 411-418

[8] Wulczyn, E.; Steiner, D. F.; Xu, Z.; Sadhwani, A.; Wang, H.; Flament-Auvigne, I.; Mermel, C. H.; Chen, P.-H. C.; Liu, Y. & Stumpe, M. C.
Hsieh, J. C.-H. *(Ed.)*. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLOS ONE, Public Library of Science (PLoS),* **2020**, *15*, e0233678

[9] Skrede, O.-J.; Raedt, S. D.; Kleppe, A.; Hveem, T. S.; Liestøl, K.; Maddison, J.; Askautrud, H. A.; Pradhan, M.; Nesheim, J. A.; Albregtsen, F.; Farstad, I. N.; Domingo, E.; Church, D. N.; Nesbakken, A.; Shepherd, N. A.; Tomlinson, I.; Kerr, R.; Novelli, M.; Kerr, D. J. & Danielsen, H. E. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet, Elsevier BV,* **2020**, *395*, 350-360

[10] Echle, A.; Grabsch, H. I.; Quirke, P.; van den Brandt, P. A.; West, N. P.; Hutchins, G. G.; Heij, L. R.; Tan, X.; Richman, S. D.; Krause, J.; Alwers, E.; Jenniskens, J.; Offermans, K.; Gray, R.; Brenner, H.; Chang-Claude, J.; Trautwein, C.; Pearson, A. T.; Boor, P.; Luedde, T.; Gaisa, N. T.; Hoffmeister, M. & Kather, J. N. Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. *Gastroenterology, Elsevier BV,* **2020**, *159*, 1406-1416.e11

[11] Greenson, J. K.; Huang, S.-C.; Herron, C.; Moreno, V.; Bonner, J. D.; Tomsho, L. P.; Ben-Izhak, O.; Cohen, H. I.; Trougouboff, P.; Bejhar, J.; Sova, Y.; Pinchev, M.; Rennert, G. & Gruber, S. B. Pathologic Predictors of Microsatellite Instability in Colorectal Cancer. *Ovid Technologies (Wolters Kluwer Health),* **2009**, *33*, 126-133

[12] Mobadersany, P.; Cooper, L. A. D. & Goldstein, J. A. GestAltNet: aggregation and attention to improve deep learning of gestational age from placental whole-slide images. *Laboratory Investigation, Springer Science and Business Media LLC,* **2021**

[13] Wulczyn, E.; Steiner, D. F.; Moran, M.; Plass, M.; Reihs, R.; Tan, F.; Flament-Auvigne, I.; Brown, T.; Regitnig, P.; Chen, P.-H. C.; Hegde, N.; Sadhwani, A.; MacDonald, R.; Ayalew, B.; Corrado, G. S.; Peng, L. H.; Tse, D.; Müller, H.; Xu, Z.; Liu, Y.; Stumpe, M. C.; Zatloukal, K. & Mermel, C. H. Interpretable survival prediction for colorectal cancer using deep learning. *npj Digital Medicine, Springer Science and Business Media LLC,* **2021**, *4*

[14] Chikontwe, P.; Luna, M.; Kang, M.; Hong, K. S.; Ahn, J. H. & Park, S. H. Dual attention multiple instance learning with unsupervised complementary loss for COVID-19 screening. *Medical Image Analysis, Elsevier BV,* **2021**, *72*, 102105

[15] Abbet, C.; Zlobec, I.; Bozorgtabar, B. & Thiran, J.-P. Divide-and-Rule: Self-Supervised Learning for Survival Analysis in Colorectal Cancer. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Springer International Publishing,* **2020**, 480-489

[16] Carbonneau, M.-A.; Cheplygina, V.; Granger, E. & Gagnon, G. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition, Elsevier BV,* **2018**, *77*, 329-353

[17] Ilse, M.; Tomczak, J. M. & Welling, M. Attention-based Deep Multiple Instance Learning. *arXiv,* **2018**

[18] Le-Khac, P. H.; Healy, G. & Smeaton, A. F. Contrastive Representation Learning: A Framework and Review. *IEEE Access 8*, **2020**, 193907–193934

[19] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E., 2020. A simple framework for con- trastive learning of visual representations. *In: Proceedings of the 37th Inter- national Conference on Machine Learning, ICML 2020*, 13-18 July 2020, Virtual Event. PMLR, pp. 1597–1607

[20] Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*, **2014**

[21] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K. & Fei-Fei, L. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE,* **2009**

[22] www.cancer.gov/tcga

[23] www.wisepaip.org/paip

[24] He, K.; Zhang, X.; Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *arXiv*, **2015**