

Local Attention Graph-based Transformer for Multi-target Genetic Alteration Prediction

Daniel Reisenbüchler^{1,2}, Sophia J. Wagner^{1,2}, Melanie Boxberg⁴, and Tingying Peng²

¹ Technical University Munich, Munich, Germany

² Helmholtz AI, Neuherberg, Germany

³ Institute of Pathology Munich-North, Munich, Germany
reisenbuechler@helmholtz-muenchen.de

Abstract. Classical multiple instance learning (MIL) methods are often based on the identical and independent distributed assumption between instances, hence neglecting the potentially rich contextual information beyond individual entities. On the other hand, Transformers with global self-attention modules have been proposed to model the interdependencies among all instances. However, in this paper we question: Is global relation modeling using self-attention necessary, or can we appropriately restrict self-attention calculations to local regimes in large-scale whole slide images (WSIs)? We propose a general-purpose local attention graph-based Transformer for MIL (LA-MIL), introducing an inductive bias by explicitly contextualizing instances in adaptive local regimes of arbitrary size. Additionally, an efficiently adapted loss function enables our approach to learn expressive WSI embeddings for the joint analysis of multiple biomarkers. We demonstrate that LA-MIL achieves state-of-the-art results in mutation prediction for gastrointestinal cancer, outperforming existing models on important biomarkers such as microsatellite instability for colorectal cancer. Our findings suggest that local self-attention sufficiently models dependencies on par with global modules. Our LA-MIL implementation is available at https://github.com/agentdr1/LA_MIL.

Keywords: Multiple instance learning · Graph transformer · Local attention · Whole slide images · Mutation prediction

1 Introduction

Advances in slide-scanning microscopes and deep learning-based image analysis have significantly increased interest in computational pathology [2]. Whole slide images typically contain billions of pixels and reach up to several gigabytes in size. To mitigate the resulting computational burden, WSIs are commonly tessellated into smaller tiles [16]. However, patient diagnosis is typically only available as weakly-supervised slide-level annotation, e.g., cancer vs. non-cancer classification, cancer subtyping, or genomic analyses.

In histopathological image analysis, this task is formulated as multiple instance learning (MIL), where a WSI is considered as a bag, and tiles as contained

instances. Hence, efficiently learning representations and aggregating them from tiles to a bag label is crucial. One simple solution for this is to pass the bag label onto each tile, reducing MIL to supervised learning. This approach is particularly favored because of its ease of implementation, e.g., to predict microsatellite instability or tumor mutational burden [10,22]. The final bag-level prediction is obtained by aggregating all instance-level predictions with average pooling. These methods have two drawbacks: (i) a fraction of instance labels may differ from the bag label and therefore form label noise in supervised learning, and (ii) no morphological or spatial correlation between tiles is taken into account.

To remedy (i), MIL can learn from bag-level annotation without assuming the same label for each tile. In particular, Ilse et al. [8] propose an attention-based pooling layer aiming to weight each tile individually for its relevance within the bag prediction task. To tackle (ii), recently, convolutional neural networks (CNNs) were combined with self-attention-based Transformers [21]. Here, the tiles are condensed into feature vectors and subsequently the resulting sequence is fed into a Transformer, where the interdependence between tiles is incorporated by self-attention mechanisms. For instance, Li et al. [11] propose a deformable Transformer-based encoder-decoder structure and evaluate it across encoder only based Transformer. Shao et al. [20] uses the Neystrom method to approximate self-attention, aiming to decrease the computational complexity. Myronenko et al. [17] suggest incorporating feature vectors of different scales into an encoder-based Transformer.

However, general Transformer approaches suffer from quadratic complexity with respect to the sequence length. This complexity is a general problem across computer vision and neural language processing (NLP) domains. To alleviate this concern, Transformer using local attention in the NLP domain [15] showed that it is sufficient for a token to restrict the attention calculation to a local neighborhood inside the sequence, i.e., the surrounding words. On the other hand, in computer vision self-attention can be modified by introducing local windowed attention [12]. However, this approach for WSIs is not as conveniently applicable as for domain areas where the images have the same size, such as in well-curated datasets like ImageNet [4]. Whole slide images come with varying geometrical shapes and the representation obtained by using tiles while excluding some entities (e.g. due to artifacts, pen marks, etc) leads to holes within the visual representation. A handcrafted selection of which entities participate in the key-to-query product is not generally applicable for all WSIs, as this would not effectively adapt to the varying local neighborhoods.

Combining the local windowed attention idea with Graph Transformer [5], we propose a computationally light training pipeline consisting of a CNN and local attention-based Transformer with the following contributions:

- We present LA-MIL, a novel local attention graph-based Transformer that restricts self-attention calculations in Transformers by using k-nearest neighbor (kNN) graphs to model local regimes with respect to tiles inside the WSI.

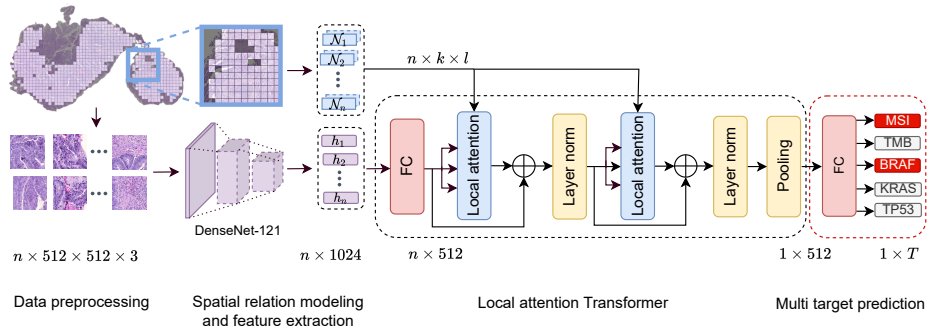


Fig. 1. LA-MIL overview: The pipeline consists of preprocessing, spatial relation modeling and feature extraction, and a local attention-based Transformer.

- To our knowledge, LA-MIL is the first pipeline to predict microsatellite instability and tumor mutational burden jointly with genetic alterations as well as the first transformer-based approach for mutation prediction.
- An efficiently adapted loss function enables our approach to learn meaningful bag representations for a joint analysis of multiple imbalanced biomarkers.
- We evaluate our approach extensively on two datasets for gastrointestinal cancer and demonstrate that our local attention mechanism sufficiently leverages information on par with global self-attention modules.
- LA-MIL shows great modeling interpretability by visualizing local attention scores, consisting of spatial and morphological dependencies.

Our experiments indicate that LA-MIL outperforms state-of-the-art approaches or is on par for mutation prediction tasks in gastrointestinal cancer.

2 Method

The pipeline of our LA-MIL approach is visualized in Fig. 1. In the following, we introduce the key components of our algorithm.

2.1 LA-MIL framework

Data preprocessing. First, a given gigapixel WSI is tessellated into N tiles, where each tile $t_i \in \mathbb{R}^{H \times W \times C}$. Further, we extract the coordinates $c_i \in \mathbb{R}^2$ with respect to the WSI for each tile. Tiles containing background, artifacts, and non tumor-tissue are excluded using Otsu’s method [18] and region of interest (RoI) annotations, reducing the number of tiles for downstream processing to $n < N$.

Per-tile feature extraction and spatial relation modeling. We compress the visual information contained in each tile t_i by extracting features using KimiaNet [19], a pretrained DenseNet-121 [7]. Thus, the WSI is represented

as a sequence of feature vectors $\{h_i\}_{i=1}^n \in \mathbb{R}^{n \times D}$, where the dimension D is the output size of the feature extraction CNN. For the spatial relation among the tiles of a WSI, we build l k-nearest neighbor (kNN) graphs \mathcal{G}_{kNN}^l using the Euclidean distance of the coordinates c_i . A kNN graph can be represented by a matrix $A \in \mathbb{R}^{n \times k}$, indicating the k neighbors for all n tiles.

Transformer architecture. Given a sequence of features $\{h_i\}_{i=1}^n$ and l graphs \mathcal{G}_{kNN}^l , we further downscale the feature vectors from D to d by using a fully-connected (FC) layer. Subsequently, a Transformer with l blocks of local attention layers is applied. These layers utilize the kNN graphs to update neighboring tiles, thereby modeling local morphological and spatial correlations. Note that a graph can also be shared between layers. Applying a residual connection and layer normalization [1] after each attention layer aims to improve the gradient flow and generalization performance. Finally, the sequence is aggregated into a bag-embedding vector $b \in \mathbb{R}^d$ by mean pooling as done in [21]. Another FC layer projects the bag-embedding vector into a target vector $t \in \mathbb{R}^T$, where T is the number of targets to predict. The sigmoid function is applied element-wise on the target vector t to obtain the scores for each target individually. In the context of mutation prediction, the thresholded scores indicate whether a particular gene occurs as wildtype or mutated, respectively.

Loss function. Mutation prediction is a challenging task since the targets often only occur in small frequencies (see Table 1). Hence, we use a loss which penalizes the model for wrong decisions about the prediction of underrepresented classes by weighting each binary cross-entropy (BCE) term individually. We take the mean of T BCE losses, thus treating each target equally:

$$L(\mathbf{x}, \mathbf{y}) = -\frac{1}{T} \sum_{t=1}^T \frac{n_t^{\text{neg}}}{n_t^{\text{pos}}} (y_t \log(\sigma(x_t)) + (1 - y_t) \log(1 - \sigma(x_t))).$$

As mentioned in the introduction, most mutation prediction studies use tile-level supervised learning rather than bag-level MIL. Hence, a common strategy to tackle highly imbalanced classification is to apply downsampling to reach an equilibrium of classes in the dataset splits [9,10]. However, this may not be possible in the multi-target and bag-operating setup, as downsampling may exclude nearly all samples, depending on the individual class distributions.

2.2 Local attention layer

Self-attention is a key component in Transformer architectures, where each token h_i is updated with global information of the complete input sequence $\{h_1, \dots, h_n\}$. In contrast, our local attention modules constrain the updates for each token h_i associated with node n_i to all tokens h_j with nodes $j \neq i$ that are connected with node n_i , as shown in Fig. 2. As input we consider a n -dimensional

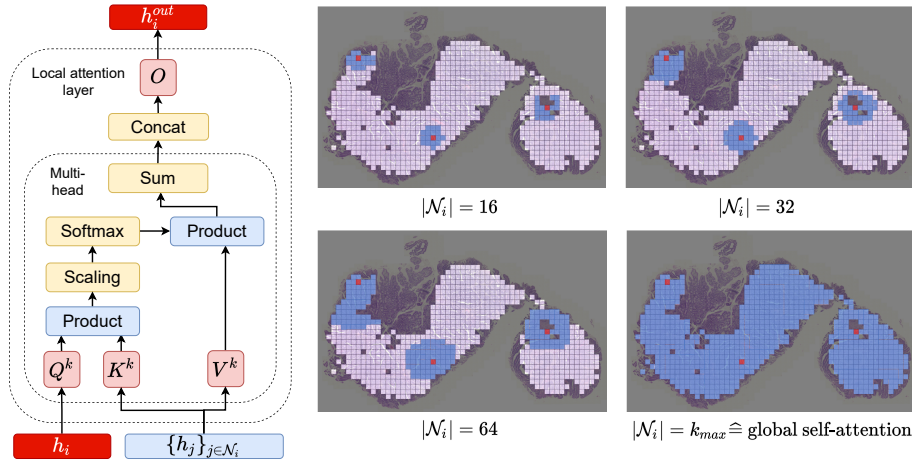


Fig. 2. Left: Computational block for a local attention layer. Right: Illustration of the spatial field of view for the attention calculations. Tiles in red visualize a query tile and blue colored tiles visualize the adaptive local neighborhoods of different sizes.

sequence of tokens h_i for $i = 1, \dots, n$ associated with n nodes of a graph \mathcal{G} where the nodes are connected by $n \times k$ edges. The update equation for token h_i is

$$h_i = O \cdot \text{Concat}_k \left(\sum_{j \in \mathcal{N}_j} w_{i,j}^k V^k h_j \right), \quad w_{i,j}^k = \text{softmax}_j \left(\frac{Q^k h_i \cdot K^k h_j}{\sqrt{d_k}} \right) \quad (1)$$

where $Q^k, K^k, V^k \in \mathbb{R}^{d_k \times d}$ and $O \in \mathbb{R}^{d \times d}$ with $k = 1, \dots, H$ denoting the number of the respective attention head. The notation $j \in \mathcal{N}_i$ refers to a set of indices j of nodes connected to the i th node by j edges. The cardinality of \mathcal{N}_i is equal to the number of neighbors for all tiles. To calculate local attention scores a_i for each tile from local attention layers, we first cache the intermediate outputs $w_{i,j}^k$ from Equation 1 and sum them across heads and local regimes, i.e.

$$a_i = \sum_{k=1}^H \sum_{j \in \mathcal{N}_j} w_{i,j}^k. \quad (2)$$

Subsequently, we normalize all n attention values a_i into the range $[0, 1]$ and denote the outcomes as local attention scores for each tile.

3 Experiments

We applied the proposed method on diagnostic formalin-fixed paraffin-embedded diagnostic slides for two cohorts in the The Cancer Genome Atlas (TCGA) dataset [24]. From tissue contained in WSIs, we followed recent works [6,9,3]

Table 1. Distribution of genetic alterations in TCGA-CRC and TCGA-STAD. We denote the number of positive samples for each target.

Cohort	n	MSI	TMB	BRAF	ALK	ERBB4	FBXW7	KRAS	PIK3CA	SMAD4	TP53
CRC	594	78	85	66	40	62	107	223	178	81	332
STAD	440	75	86	13	19	62	38	40	86	36	225

and include only tumor-occupied tissue regions. All images were downsampled to $20\times$ magnification, corresponding to a resolution of $0.5\frac{\mu m}{px}$. The task is to predict genetic alterations, the microsatellite status and the tumor mutational burden (TMB) as biomarkers [16] directly from WSIs.

TCGA colorectal and TCGA stomach datasets. Our dataset TCGA-CRC consists of tiled WSIs from tumor regions of colorectal tissue. We used the preprocessed tumor tissue tiles from kather.ai. As a second dataset, we tiled WSIs of stomach tissue from TCGA, downloaded at portal.gdc.cancer.gov. After noise removal, we excluded all tiles which were not contained in the tumor region by using manual tumor annotations available at: kather.ai. We retrieved the genetic annotations matching the WSIs from xenabrowser.net. Annotations for the microsatellite stability/instability (MSS/MSI) and TMB are available at cbiportal.org. Following Luchini et al. [14], we binarized MSS and MSI-Low as MSS and MSI-High as MSI.

Implementation. Each tile was embedded into a 1024-dimensional feature space by a DenseNet-121 model that was pretrained on histopathological data. By using the coordinates of each tile, we built two kNN-Graphs with $k = 16$ and $k = 64$ for subsequent attention restriction in the first and second local attention module, respectively. In the training phase, each feature vector associated with the tiles was further compressed from 1024 to 512 by a FC layer. After a stack of two local attention layers, we averaged the feature vectors across all tiles. The resulting bag embedding was passed through a classification head, consisting of another fully connected layer from 512 to 10, to compute the logits. We applied the sigmoid activation function element-wise to calculate the probabilities for each individual target. For optimization, we employed the Lookahead optimizer [25] together with AdamW [13], and used a learning rate of $2e-05$ and $2e-04$ (for TCGA-CRC and TCGA-STAD, respectively) for 10 epochs, weight decay of $2e-05$, and batch size 1. The LA-MIL model with 2.1M parameter was implemented in PyTorch and DGL [23], and trained on a single Tesla V100 GPU.

Evaluation. To evaluate the mutation prediction task on both datasets TCGA-CRC and TCGA-STAD, we compared the performance of LA-MIL with state-of-the-art methods. The fact that most of the recent advances predict T biomarkers individually results in training, validating, and hyperparameter tuning for T separate models, while we used a single model to predict all biomarkers. Moreover,

Table 2. Mean AUROC scores for mutation prediction on the datasets TCGA-CRC and TCGA-STAD. For the competitive methods, we report results from the original publications; for our methods, we report the mean over five folds (see supplementary material for results with standard deviation).

Dataset	Method	MSI	TMB	BRAF	ALK	ERBB4	FBXW7	KRAS	PIK3CA	SMAD4	TP53
TCGA-CRC	Kather et al. [10]	0.77	-	-	-	-	-	-	-	-	-
	Wang et al. [22]	-	0.82	-	-	-	-	-	-	-	-
	Kather et al. [9]	-	-	0.66	0.51	-	0.49	0.60	0.62	0.63	0.68
	Fu et al. [6]	-	-	0.57	-	-	0.66	0.55	0.59	0.58	0.68
	T-MIL (Ours)	0.85	0.82	0.73	0.61	0.57	0.64	0.61	0.60	0.60	0.64
	LA-MIL (Ours)	0.85	0.83	0.72	0.63	0.60	0.66	0.62	0.61	0.58	0.63
TCGA-STAD	Kather et al. [10]	0.81	-	-	-	-	-	-	-	-	-
	Wang et al. [22]	-	0.75	-	-	-	-	-	-	-	-
	Kather et al. [9]	-	-	0.37	0.45	-	0.74	0.64	0.67	0.61	0.60
	Fu et al. [6]	-	-	-	-	-	-	-	0.47	0.49	0.63
	T-MIL (Ours)	0.80	0.78	0.73	0.52	0.47	0.71	0.65	0.58	0.62	0.57
	LA-MIL (Ours)	0.78	0.77	0.67	0.52	0.47	0.72	0.70	0.61	0.64	0.58

we implemented a Transformer MIL approach where we exchanged all local attention blocks with global self-attention [21], denoted as T-MIL. To stick with common evaluation procedures for mutation prediction in recent works, we evaluated our pipeline with a 5-fold cross validation (CV). We split the datasets into folds such that individual class distributions in each fold were approximately the same and ensured that no patient appeared in the training and validation set at the same time. We measured the performance using the area under the receiver operating characteristic curve (AUROC) for each target individually.

4 Results

Current methods, such as Kather et al. [10] for MSI and Wang et al. [22] for TMB, predict the biomarkers instance-wise as single targets for each tile with the corresponding label inherited from its parent WSI. Similarly, Kather et al. [9] and Fu et al. [6] train one model for each target gene when predicting mutations, but evaluate their results on WSI-level by average pooling of tile-wise predictions. In contrast, we train and evaluate only one model on WSI-level to predict multiple biomarkers using a MIL transformer that aggregates features from all tiles with global or local self-attention layers.

Table 2 shows the AUROC scores of state-of-the-art instance-wise methods compared to our methods on the dataset TCGA-CRC. The results suggest that our models can leverage information from multiple targets to achieve better overall performance. Interestingly, this holds especially for the prediction of MSI, where our approach improves the score by 8% from 0.77 to 0.85. The prediction

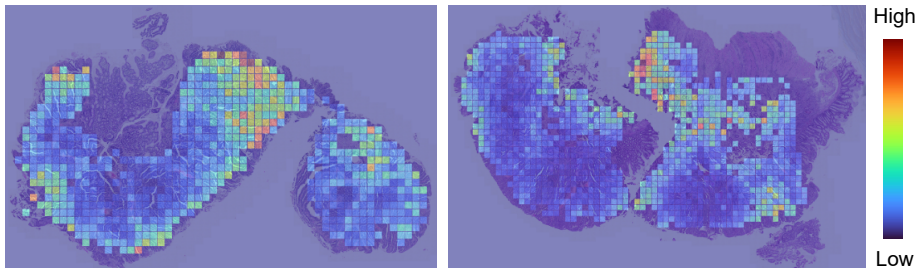


Fig. 3. Local attention scores visualization for the last local attention layer with restricted self-attention in a neighborhood of size 64.

of TMB and most of the gene mutations (except for SMAD4 and TP53) are on par or marginally better (+1%).

The results for the TCGA-STAD dataset in MSI prediction are marginally worse (-1%). This could arise due to the fact that the compared work uses a slightly different label strategy for MSI-Low cases, which also affects the evaluation. The results for TMB improves by up to 3%. Similar as on the TCGA-CRC dataset, we observe an improvement of results using our Transformer-based approaches for the remaining targets except for a few genes.

Local attention visualization. As described in Eq. 2, we can calculate the local attention scores from the query-to-key product. In Fig. 3 we colorized the tiles according to their corresponding local attention score. The high attentive regions include the nuclear chromatin that seems to be hyperchromatic, as well as crowded glands or solid areas. Yet, to the best of our knowledge, there are no distinguishing patterns from WSI for mutated genes, which makes it difficult for a quantitative evaluation. Nevertheless, LA-MIL provides an interpretation based on the contribution of each tile for the bag-level prediction task and thus paves a way for a deeper investigation of highly scored tiles.

5 Conclusion

In this work, we proposed a novel MIL framework with local attention for WSI analysis. Local attention is achieved through a graph-based transformer that models region-wise inter-dependencies. As the size of the region can be set arbitrarily, our approach bridges the gap between instance-wise and global relation approaches by modeling local relations of arbitrary size. Moreover, an effective adapted loss enables us to learn multiple biomarkers at once, for low computational cost compared to CNN-only based methods.

However, there is often more than one WSI for a patient available in the TCGA database. Future work will investigate the strategy of combining all WSIs for a patient, while suitably scaling the coordinates for tiles of different WSIs.

Thus, each tile will only be updated with local information from its direct parent WSI and transformed into a bag-embedding consisting of locally correlated tiles of all WSIs from a patient. We believe that our approach provides a base for further applications in other WSI analysis tasks where structured relation modeling is crucial.

Acknowledgements S.J.W. was supported by the Helmholtz Association under the joint research school "Munich School for Data Science - MUDS".

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization (2016). <https://doi.org/10.48550/ARXIV.1607.06450>
2. Cooper, L.A., Demicco, E.G., Saltz, J.H., Powell, R.T., Rao, A., Lazar, A.J.: Pan-Cancer insights from the cancer genome atlas: the pathologist's perspective. *The Journal of Pathology* **244**(5), 512–524 (Feb 2018)
3. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A.: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine* **24**(10), 1559–1567 (Sep 2018)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Dwivedi, V.P., Bresson, X.: A generalization of transformer networks to graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications* (2021)
6. Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L.R., Jimenez-Linan, M., Moore, L., Gerstung, M.: Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer* **1**(8), 800–810 (Jul 2020). <https://doi.org/10.1038/s43018-020-0085-8>
7. Huang, G., Liu, Z., Pleiss, G., Van Der Maaten, L., Weinberger, K.: Convolutional networks with dense connectivity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Dy, J., Krause, A. (eds.) *Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 80, pp. 2127–2136. PMLR (10–15 Jul 2018)
9. Kather, J.N., Heij, L.R., Grabsch, H.I., Loeffler, C., Echle, A., Muti, H.S., Krause, J., Niehues, J.M., Sommer, K.A.J., Bankhead, P., Kooreman, L.F.S., Schulte, J.J., Cipriani, N.A., Buelow, R.D., Boor, P., Ortiz-Brüchle, N., Hanby, A.M., Speirs, V., Kochanny, S., Patnaik, A., Srisuwananukorn, A., Brenner, H., Hoffmeister, M., van den Brandt, P.A., Jäger, D., Trautwein, C., Pearson, A.T., Luedde, T.: Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer* **1**(8), 789–799 (Jul 2020)
10. 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* **25**(7), 1054–1056 (Jun 2019)

11. Li, H., Yang, F., Zhao, Y., Xing, X., Zhang, J., Gao, M., Huang, J., Wang, L., Yao, J.: DT-MIL: Deformable transformer for multi-instance learning on histopathological image. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 206–216. Springer International Publishing (2021)
12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. *International Conference on Computer Vision (ICCV)* (2021)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net (2019)
14. Luchini, C., Bibeau, F., Ligtenberg, M., Singh, N., Nottegar, A., Bosse, T., Miller, R., Riaz, N., Douillard, J.Y., Andre, F., Scarpa, A.: ESMO recommendations on microsatellite instability testing for immunotherapy in cancer, and its relationship with PD-1/PD-L1 expression and tumour mutational burden: a systematic review-based approach. *Annals of Oncology* **30**(8), 1232–1243 (Aug 2019)
15. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pp. 1412–1421. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015). <https://doi.org/10.18653/v1/D15-1166>
16. Murchan, P., Ó'Brien, C., O'Connell, S., McNevin, C.S., Baird, A.M., Sheils, O., Broin, P.Ó., Finn, S.P.: Deep learning of histopathological features for the prediction of tumour molecular genetics. *Diagnostics* **11**(8), 1406 (Aug 2021)
17. Myronenko, A., Xu, Z., Yang, D., Roth, H.R., Xu, D.: Accounting for dependencies in deep learning based multiple instance learning for whole slide imaging. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 329–338. Springer International Publishing (2021)
18. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **9**(1), 62–66 (1979). <https://doi.org/10.1109/TSMC.1979.4310076>
19. Riasatian, A., Babaie, M., Maleki, D., Kalra, S., Valipour, M., Hemati, S., Zaveri, M., Safarpour, A., Shafiei, S., Afshari, M., Rasoolijaberi, M., Sikaroudi, M., Adnan, M., Shah, S., Choi, C., Damaskinos, S., Campbell, C.J., Diamandis, P., Pantanowitz, L., Kashani, H., Ghodsi, A., Tizhoosh, H.: Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Medical Image Analysis* **70**, 102032 (2021)
20. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification (2021)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 30. Curran Associates, Inc. (2017)
22. Wang, L., Jiao, Y., Qiao, Y., Zeng, N., Yu, R.: A novel approach combined transfer learning and deep learning to predict TMB from histology image. *Pattern Recognition Letters* **135**, 244–248 (Jul 2020)
23. Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., Zhou, J., Ma, C., Yu, L., Gai, Y., Xiao, T., He, T., Karypis, G., Li, J., Zhang, Z.: Deep graph library: A graph-centric, highly-performant package for graph neural networks (2020)
24. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The cancer genome atlas pan-cancer analysis project. *Nature Genetics* **45**(10), 1113–1120 (Sep 2013)

25. Zhang, M.R., Lucas, J., Hinton, G., Ba, J.: Lookahead optimizer: k steps forward, 1 step back (2019)