Space-filling Curves for Modeling Spatial Context in Transformer-based Whole Slide Image Classification

Cihan Erkan and Selim Aksoy

Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey

ABSTRACT

The common method for histopathology image classification is to sample small patches from large whole slide images and make predictions based on aggregations of patch representations. Transformer models provide a promising alternative with their ability to capture long-range dependencies of patches and their potential to detect representative regions, thanks to their novel self-attention strategy. However, as a sequence-based architecture, transformers are unable to directly capture the two-dimensional nature of images. While it is possible to get around this problem by converting an image into a sequence of patches in raster scan order, the basic transformer architecture is still insensitive to the locations of the patches in the image. The aim of this work is to make the model be aware of the spatial context of the patches as neighboring patches are likely to be part of the same diagnostically relevant structure. We propose a transformer-based whole slide image classification framework that uses space-filling curves to generate patch sequences that are adaptive to the variations in the shapes of the tissue structures. The goal is to preserve the locality of the patches so that neighboring patches in the one-dimensional sequence are closer to each other in the two-dimensional slide. We use positional encodings to capture the spatial arrangements of the patches in these sequences. Experiments using a lung cancer dataset obtained from The Cancer Genome Atlas show that the proposed sequence generation approach that best preserves the locality of the patches achieves 87.6% accuracy, which is higher than baseline models that use raster scan ordering (86.7% accuracy), no ordering (86.3% accuracy), and a model that uses convolutions to relate the neighboring patches (81.7% accuracy).

Keywords: Digital pathology, space-filling curves, vision transformer, whole slide image classification

1. INTRODUCTION

Digitized histopathology slides, called whole slide images (WSIs), have enabled pathologists to use computer aided diagnosis tools for detecting and grading tumors. Conventional grading methods are usually tedious and the grades assigned by the pathologists often depend on the experience of the particular experts that are carrying out the grading. Utilization of image analysis tools to classify WSIs can significantly reduce both the variability among the pathologists and their workload.

Automated WSI classification methods typically employ a patch-based strategy. First, the WSI is divided into smaller patches and a convolutional neural network (CNN) is used to extract feature vectors from those patches. Then, it becomes possible to make a classification by treating the WSI as a bag of feature vectors. This final classification can be done in multiple ways. The methods that are based on feature aggregation use the patch features to generate a feature representation for the bag and use this global feature vector to make a prediction.[1,2] An alternative is to use multiple instance learning (MIL) to classify the bag that is composed of patches as the instances.[3–5]

Another classification approach that has gained popularity in recent years is the *vision transformer*.[6] The core idea of the transformer-based models, the self-attention mechanism, provides an effective way of capturing relationships between different patches of an image. As opposed to CNNs, transformers can relate distant parts of an image to obtain a global representation, which makes them particularly useful for WSI classification as the images are usually very large and long-range relationships are common. Moreover, the attention-based approach of the transformers can identify the diagnostically relevant parts of the WSIs and discard the numerous irrelevant

Send correspondence to S.A.: E-mail: saksoy@cs.bilkent.edu.tr, Telephone: +90 (312) 2903405

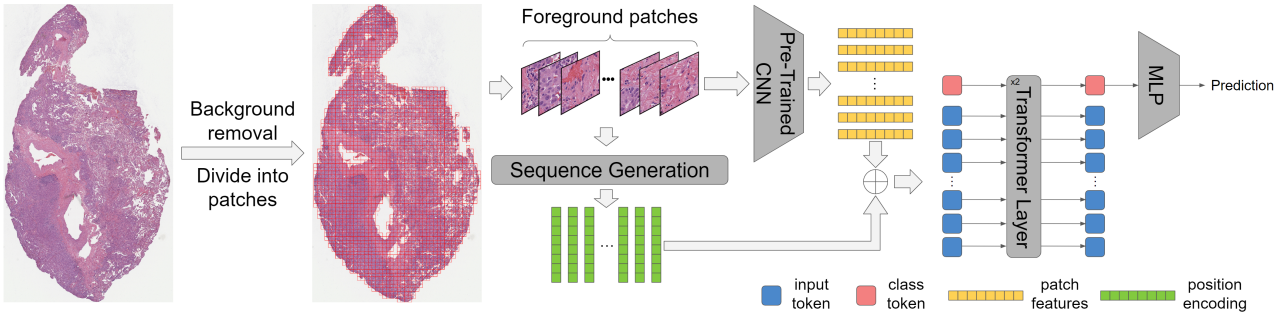


Figure 1. An overview of the proposed approach.

sections by shifting its attention. Recently, Shao et al.[7] and Mehta et al.[8] proposed transformer-based models, named TransMIL and HATNet, respectively, for histopathological image classification. The former utilizes two transformer layers with a convolution-based layer in between that aims to incorporate position information to classify a WSI. The latter approach employs multiple transformers to model the relationships between patches and bags of patches to classify manually identified regions of interest.

Unlike CNNs, transformers were originally designed to work on sequences. This poses an important problem regarding the transformer-based models as, in their original form, they are insensitive to the locations of the patches of an image. While it is possible to insert position information into the model, the common methods expect images to have a fixed size, which makes them unfit for WSIs with variable shapes and sizes. This problem can be observed in the aforementioned works as well. For example, TransMIL's convolution layer reorganizes the linear sequence of patches into a square form that ultimately leads to loss of information, particularly along the vertical dimension. On the other hand, HATNet does not incorporate any explicit position information and the resulting model is unaware of the spatial locations of the patches.

For WSI classification, it is important for the model to be attentive to the spatial context of the patches as neighboring patches are likely to be part of the same diagnostically relevant structure. If a model treats every patch as a standalone instance, it might miss important patterns. Even though it may be impossible to perfectly map a WSI into a one-dimensional sequence, the amount of information captured by the model can be enhanced by employing better sequence generation methods.

In this paper, we propose a transformer-based WSI classification framework that utilizes space-filling curves to generate patch sequences that are adaptive to the variations in the shapes and sizes of WSIs while encapsulating valuable information regarding the spatial arrangements of the patches. In particular, the goal is to preserve the locality of the patches so that neighboring patches in the one-dimensional sequence are closer to each other in the two-dimensional slide. The resulting position information in the produced sequence is incorporated into the patch representations by using positional encodings. Our experiments show that the proposed strategy that uses space-filling curves performs better than the commonly used sequence generation method of raster scan ordering, a model that uses convolutions to relate the neighboring patches, and a model that does not use any spatial encoding. In the rest of the paper, we describe the details of the methodology and present the experimental results.

2. METHODOLOGY

The proposed methodology follows a patch-based strategy to tackle the WSI classification task. Our method consists of three main steps: feature extraction, sequence generation, and classification. In the first step, we extract patch features from the WSI. Then, we generate a patch sequence to capture the spatial context and encode the position information. Finally, we predict the label of the WSI by using a classifier that consists of a transformer encoder and a multilayer perceptron (MLP) head. Figure 1 provides an overview of the proposed approach. In the following, we first summarize the feature extraction step, then describe the transformer-based classifier, and finally provide the details of the sequence generation methods used with this classifier.

2.1 Feature extraction

The feature extraction process generates a patch-based representation for the slides. First, we divide the WSI into non-overlapping 256×256 patches. Then, among the resulting patches, we remove the ones identified as background patches. Background removal is done by using a threshold on the average saturation values of the patches, as background patches usually have lower saturation values. After removing the background patches, we use an ImageNet pre-trained ResNet-based feature extractor to obtain a 1024-dimensional feature vector for each patch.

2.2 Classification

Classification step of the proposed method utilizes a transformer and an MLP head. The transformer architecture, originally proposed as a sequence-to-sequence model, has an encoder-decoder structure.[9] The transformer encoder can be used to generate a global representation for a sequence of tokens, which in turn can be used for sequence classification. The encoder employs two layers, a multi-head self-attention mechanism and a feed-forward network. Self-attention is given by the function

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V are learnable linear projections of the tokens and are named as query, key, and value, respectively, while d_k is the length of the Q and K vectors. Applying self-attention multiple times with different projections and concatenating them finalizes the multi-head self-attention mechanism. This novel approach to attention gives transformers the capacity to model long-range dependencies and the ability to generate powerful global representations.

We follow the common way of employing the transformer model as a classifier by using a special token called the class token. Here, the class token is a learnable parameter of the model and is attached to the beginning of every input sequence. The assumption is that the output of the model corresponding to the class token is the global representation of the sequence. Our overall classifier pipeline consists of a fully connected layer to reduce the size of the patch feature vectors to 512, two transformer layers, layer normalization, and an MLP layer connected to the class token to make the final prediction.

An important drawback of the transformer-based models is that they require a significant amount of memory because of their quadratic space complexity. As the sizes of the WSIs increase, it becomes increasingly infeasible to use the transformers in the default setting. Thus, instead of using the original implementation of the transformer model, we use an approximation that applies the Nyström method[10] to reduce the complexity of the self-attention mechanism. This allows us to train our model without discarding any important patches.

2.3 Sequence generation

WSIs are composed of biological structures with different shapes and sizes, and some of those structures may be diagnostically significant. The feature extraction network can encode those structures if they do not span more than one patch. To understand the larger structures, the classifier needs to know the larger context a patch resides. However, transformers have no inherent way of understanding the spatial context of the patches. To pass this information to the classifier, positional information should be injected into the input tokens of the transformer.

The common way of passing this information involves adding either constant or learnable positional encodings to the tokens in a raster scan order. Here, learnable positional encodings are defined as a set of model parameters, and the model learns the most appropriate values for them during the training process. On the other hand, constant positional encodings are fixed before the training starts, but they vary in a predictable manner throughout the sequence. In both cases, the model learns to pinpoint the exact locations of the patches as it associates each encoding with a fixed spot in the patch sequence. When the shape and size of the image is fixed, such an approach transfers tokens from the two-dimensional space to a single dimension without losing any information because the exact spatial position of the token in the two-dimensional image can easily be derived from its location in the sequence. However, this method provides suboptimal results for WSIs as the shapes and

sizes of the tissue structures are variable. When the raster scan order is used, the location information along the horizontal axis is somewhat preserved, but this method completely discards the vertical axis information. Moreover, large portions of the WSIs are usually composed of background regions that should be discarded, and this makes the horizontal flattening approach even less reliable. Because of these reasons, the model might not recognize diagnostically relevant structures as the patches that form these structures could be scattered around the sequence.

It is impossible to convert two-dimensional information to a single dimension without any loss of information when the shape of the WSI is variable. Nevertheless, considering that the diagnostic label of the WSI is invariant to rotation and translation operations, it can be assumed that the model does not require the global contextual information. We hypothesize that the more important component of the spatial context is the part that concerns the local neighborhoods of the diagnostically relevant structures. To this end, our novel sequence generation approach aims to preserve the spatial integrity of the regions while ordering the tokens. This allows the transformers to *see* the structures that cover multiple patches.

We achieve such an ordering by utilizing space-filling curves. Space-filling curves provide a mapping from $\mathbb{R}^n$ to $\mathbb{R}$. Such a mapping can be trivially translated to a mapping from $\mathbb{Z} \times \mathbb{Z}$ to $\mathbb{Z}$ by discretization, which in turn can be used to generate an ordered sequence based on the locations of the patches that constitute the WSI. We particularly focus on the space-filling curves' capability of preserving locality, i.e., two patches located in a close proximity in the one-dimensional sequence are expected to be close to each other in the two-dimensional slide space. Accordingly, consecutive patches are more likely to be part of the same larger diagnostic structure. This enables the transformer model to understand which patches are coming from the same structure.

In this paper, we evaluate four different approaches to generate the input sequence: Hilbert, Morton, Spiral, and Scan. Figure 2 presents a visualization of the sequences generated by the following ordering strategies:

- **Hilbert and Morton:** These methods generate the sequence by following the paths modeled by the discrete approximations of the Hilbert[11] and Morton[12] space-filling curves, respectively.
- **Spiral:** This method follows the path described by an outward-traveling spiral. The spiral originates approximately in the middle of the foreground region of the WSI.
- **Scan:** This method creates the sequence by following the commonly used raster scan order, traveling from left to right and top to bottom.

After generating a sequence, the last step is injecting the spatial information into the model. Our method handles this by utilizing sine-cosine based constant positional encodings.[9] We use this form of sinusoidal encodings as they are more suitable for variable sequence lengths. Here, the positional encodings are generated with the same size as the token feature descriptors. The final representation for each patch is obtained by an element-wise addition of its positional encoding vector and its token descriptor vector.

3. EXPERIMENTS

3.1 Dataset and experimental setting

Dataset: To train and test our model, we use the lung cancer WSIs released by The Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga) under projects TCGA-LUAD and TCGA-LUSC. Our dataset contains 532 LUAD (Lung Adenocarcinoma) slides and 508 LUSC (Lung Squamous Cell Carcinoma) slides. We split the dataset into five folds so that each fold contains roughly the same number of samples. The slides obtained from the same patient remain in the same fold.

Implementation details: During background removal of the feature extraction step, we set the saturation threshold as 15. We use a ResNet-50 model pre-trained on the ImageNet dataset to encode each patch with a 1024-dimensional feature vector. We train our models with the cross entropy loss and optimize the parameters using the AdamW optimizer with a weight decay of 0.1 and a learning rate of 2×10^{-5}.

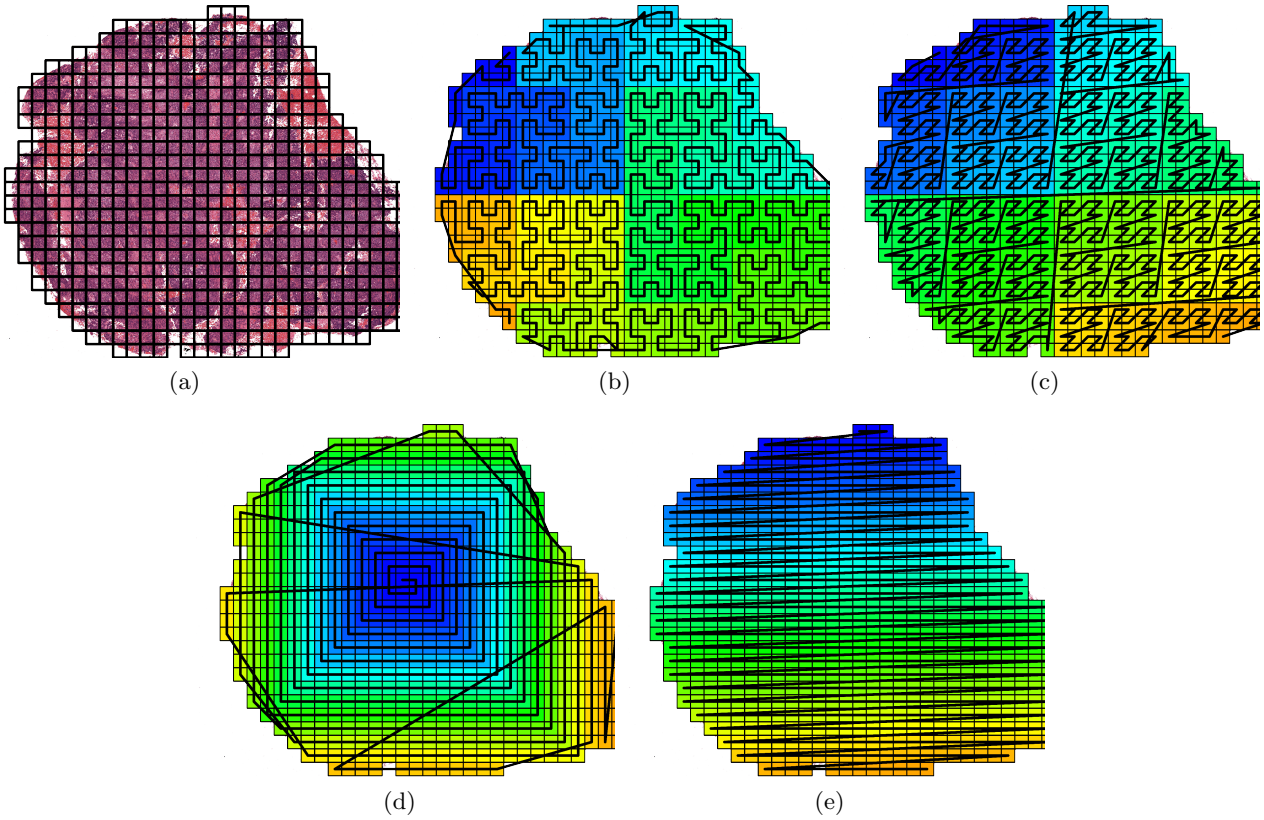


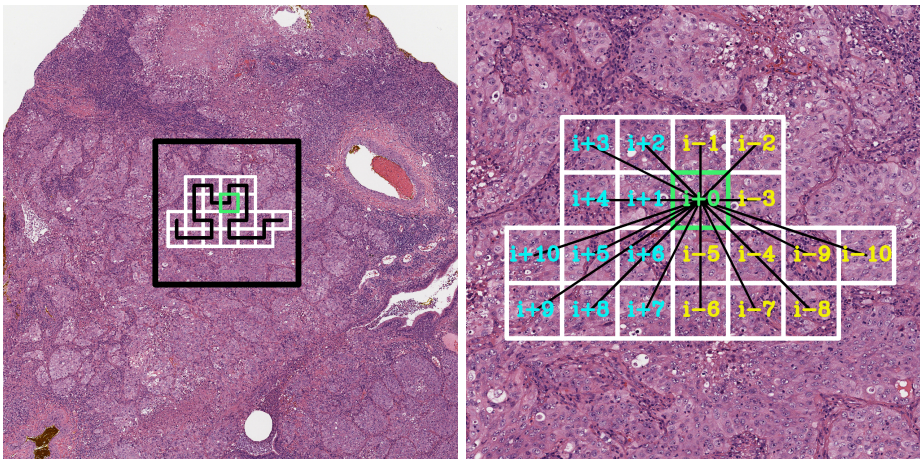
Figure 2. Illustration of sequence generation methods. (a) An example WSI. The black squares denote the foreground patches extracted from the WSI. They have no particular order when they are treated as a bag of patches. (b)-(e) present the patch sequences generated by the Hilbert, Morton, spiral, and scan strategies, respectively. Each patch is drawn with a different color (by using a colormap from blue to orange) where neighboring patches have similar colors. The black lines also connect the patches in the order they appear in the sequence.

Experimental setting: We train and test our models in two steps. In the first step, we use three folds for training and one fold for validation. This step is used to tune the hyperparameters of the model. After obtaining the best hyperparameters in this step, we fix them throughout the experiments and continue with the second step. The second step includes cross-validation experiments. We use three folds for training, one fold for validation, and one fold for testing. The validation fold is used to pick the best performing model during the iterations. We cycle the training, validation, and test folds five times so that each sample in the dataset is used for testing once. The results below use the statistics obtained from the cross-validation experiments.

3.2 Locality preservation quality

As mentioned previously, we hypothesize that an effective sequence should maintain the spatial integrity of the regions by preserving the locality of the patches. Moreover, it should be robust to the variations in the shapes and sizes of the WSIs while handling the possible discontinuities in the foreground region. In order to quantify the locality preserving quality of different sequence generation methods, we define a measure named Mean Distance to N Nearest Neighbors (MDNNN), which measures a patch's average distance to its one-dimensional sequence neighbors in the two-dimensional slide space. Figure 3 illustrates MDNNN calculation for a single patch. We calculate the MDNNN score for a single slide by averaging the MDNNN scores of all of its patches. Then, to obtain an overall MDNNN score for the whole dataset, we compute the average of the MDNNN scores for all slides.

Figure 4 presents MDNNN scores for different N values for different sequence generation methods. Results show that the Hilbert curve approach performs significantly better than other sequence generation methods for



(a) (b)

Figure 3. Illustration of locality preservation score (MDNNN) computation. (a) An example patch (green) and its neighbors (white) in the Hilbert sequence. The black lines that indicate the sequence progress from right to left in this example. (b) Illustration of MDNNN calculation for the i-th patch in the sequence for $N = 2 \times 10$. In this case, MDNNN is calculated as the average of the i-th patch's distance in the two-dimensional slide space to each of the 10 predecessor and 10 successor patches in the one-dimensional sequence.

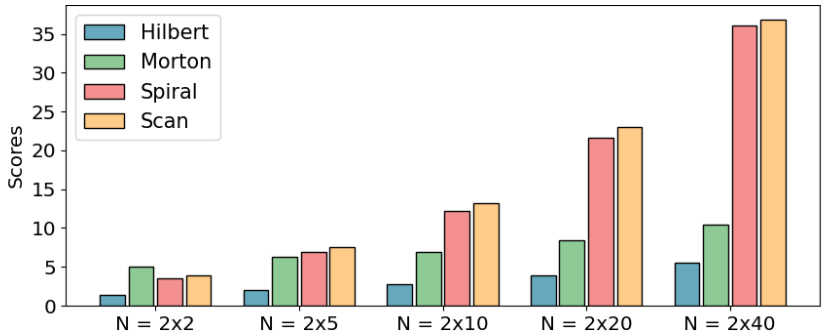


Figure 4. MDNNN scores for different sequence generation methods for different N values. Scores are calculated as patch-wise distance; thus, distance between two neighboring patches in the image equals 1. Smaller scores indicate that neighboring patches in the one-dimensional sequence are closer to each other in the two-dimensional slide.

all N values, indicating that it preserves the locality better than other methods. Performance of the Morton curve remains consistent as N increases compared to spiral and scan. Morton, spiral, and scan perform similarly for smaller N values. However, spiral and scan differ from the others significantly when N increases.

3.3 Classification experiments

To evaluate whether the proposed sequence generation method successfully transfers the position information to the transformer-based classifier, we conduct six classification experiments. The first four experiments incorporate the position information by utilizing different sequence generation methods: Hilbert, Morton, spiral, and scan. The fifth experiment tests the case where no position information is used and the patch encodings only include the outputs of the feature extractor. Finally, we use the TransMIL[7] method that includes a CNN-based module for capturing the spatial context of the WSI. We compare the methods according to their mean accuracy on five different folds when used as the test set. Table 1 presents the detailed experimental results in terms of classification accuracy for every fold, with mean accuracy and standard deviation. Figure 5 shows the mean accuracy of the models and their 95% confidence intervals.

The method that utilizes Hilbert space-filling curves to generate the sequence performs the best in terms of both mean accuracy and standard deviation by achieving 87.6% accuracy on average with a 1.59 standard

Table 1. Mean classification accuracies and standard deviations for different methods in five-fold cross-validation experiments.

	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Mean ± std.dev.**
Hilbert	86.67%	87.62%	86.89%	86.47%	90.34%	87.6 ± 1.59
Morton	85.71%	88.57%	88.35%	81.64%	86.47%	86.1 ± 2.80
Spiral	87.62%	88.10%	84.95%	80.19%	88.41%	85.9 ± 3.45
Scan	85.71%	87.62%	86.89%	83.57%	89.86%	86.7 ± 2.33
No Order	83.81%	89.05%	86.89%	83.09%	88.89%	86.3 ± 2.79
TransMIL	80.95%	78.10%	83.01%	82.13%	84.06%	81.7 ± 2.29

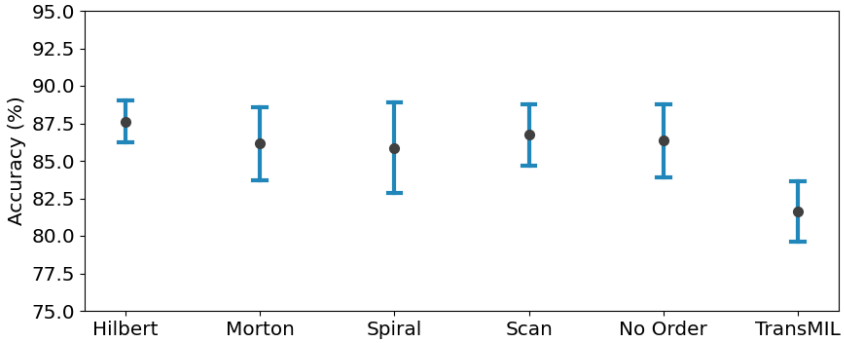


Figure 5. Mean classification accuracies and confidence intervals for different methods in five-fold cross-validation experiments.

deviation. Considering that the Hilbert curve also obtains the best (lowest) MDNNN score, this result supports our hypothesis that the performances of transformer-based models can be boosted by using sequences that are generated in a way that preserves the locality of the patches.

Other methods of sequence generation —Morton, spiral, and scan— do not offer a significant improvement over the scenario where the model receives no positional information. This implies that generating sequences with high MDNNN scores do not necessarily model the spatial context in a way that is helpful for transformer-based classifiers. This results also suggest that MDNNN scores for lower N values ($N \leq 2 \times 5$) are stronger indicators for the appropriateness of the sequences as Morton performs similar to spiral and scan only when N is low.

The TransMIL method only achieves 81.7% mean accuracy in our experiments. This is much lower than the accuracy reported by the authors of this method on the same dataset. This might show that their method of incorporating positional information by applying two-dimensional convolutions over the tokens does not generalize well over the whole dataset, as they only test their model by using a portion of the dataset.

In conclusion, the experimental results show that ordering the input tokens in a way that preserves the locality of the patches and utilizing this ordering via positional encodings boost the performance of the transformer-based classifier in the WSI classification task.

4. CONCLUSION

WSI classification, which usually employs a CNN-based approach, can also be performed by utilizing transformer-based architectures. However, transferring two-dimensional positional information of a WSI into a transformer-based classifier is a challenging task as transformers accept one-dimensional sequences as input. Even though it is more straightforward to model images with a fixed size as a sequence of patches by using positional encodings that precisely specify the location information, this becomes a significant problem when the image size is not fixed, as in the case of WSIs. Common methods of generating sequences of patches, such as the raster scan order, have the risk of scattering diagnostically relevant structures around the sequence and consequently hinder the model's ability to effectively diagnose the WSI. To overcome this issue, we proposed a novel sequence generation method that maintains the spatial integrity of such structures by employing space-filling curves that preserve the locality of the patches. Experimental results showed that the proposed method of sequence generation boosts the performance of the transformer-based WSI classifiers.

Acknowledgment

This work was supported in part by the GEBIP Award from the Turkish Academy of Sciences.

REFERENCES

[1] Akbarnejad, A., Ray, N., and Bigras, G., "Deep Fisher vector coding for whole slide image classification," in [*2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*], 243–246 (2021).

[2] Sun, C. et al., "Deep learning-based classification of liver cancer histopathology images using only global labels," *IEEE Journal of Biomedical and Health Informatics* **24**(6), 1643–1651 (2020).

[3] Campanella, G. et al., "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine* **25**, 1301–1309 (Aug 2019).

[4] Ilse, M., Tomczak, J. M., and Welling, M., "Attention-based deep multiple instance learning," (2018). arXiv:1802.04712.

[5] Lu, M. Y. et al., "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering* **5**(6), 555–570 (2021).

[6] Dosovitskiy, A. et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in [*International Conference on Learning Representations*], (2021).

[7] Shao, Z. et al., "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in [*Advances in Neural Information Processing Systems*], **34**, 2136–2147 (2021).

[8] Mehta, S., Lu, X., Wu, W., Weaver, D., Hajishirzi, H., Elmore, J. G., and Shapiro, L. G., "End-to-end diagnosis of breast biopsy images with transformers," *Medical Image Analysis* **79**, 102466 (2022).

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., "Attention is all you need," (2017). arXiv:1706.03762.

[10] Xiong, Y. et al., "Nyströmformer: A Nyström-based algorithm for approximating self-attention," in [*Proceedings of the AAAI Conference on Artificial Intelligence*], (2021).

[11] Hilbert, D., "Ueber die stetige abbildung einer linie auf ein flächenstück," *Mathematische Annalen* **38**, 459–460 (1891).

[12] Morton, G. M., "A computer oriented geodetic data base; and a new technique in file sequencing," tech. rep., IBM (1966).