SparseXMIL: Leveraging spatial context for classifying whole slide images in digital pathology

Loïc Le Bescond, Marvin Lerousseau, Fabrice Andre, Hugues Talbot

SparseXMIL: Leveraging spatial context for classifying whole slide images in digital pathology

Loïc Le Bescond, Marvin Lerousseau, Fabrice André, and Hugues Talbot, *Member*

***Abstract*— The computer analysis of Whole Slide Images (WSI) is becoming increasingly prevalent in pathology-based diagnosis, although it presents considerable challenges due to the voluminous nature of the data. To address this issue, Multiple Instance Learning (MIL) has emerged as a viable approach that involves partitioning WSI into tiles for processing. Nevertheless, conventional MIL methodologies inadequately capture the essential spatial context between tiles, which is imperative for accurate diagnosis across various diseases. In this paper, we present a novel framework, SparseXceptionMIL (SparseXMIL), aiming to enhance the modeling of spatial interactions within WSI data by introducing a multi-dimensional sparse image representation and a novel pooling operator. This operator, integrating sparse convolutions within the Xception architecture, enables effective spatial information processing across multiple scales. Empirical evaluations conducted on various classification tasks, encompassing subtyping for breast and lung carcinomas and predicting abnormalities in the DNA damage response in breast cancer WSI, consistently demonstrate the superiority of our approach over benchmark methods. These results underscore the potential of sparse convolutional architectures to improve WSI classification. The source code for our experiments is made available at `https://github.com/loic-lb/SparseXMIL`.**

***Index Terms*— Multiple Instance Learning, Sparse Convolutions, Classification, Multi-scale analysis**

I. Introduction

Pathology delves into the study of disease processes and their manifestations within tissues and organs. Pathologists play a significant role in diagnosing diseases through the microscopic examination of tissue samples, unraveling the underlying pathological mechanisms crucial for diagnosis and clinical decision-making. Through the analysis of intricate histological patterns and molecular signatures, pathology provides invaluable insights into disease etiology, progression, and treatment response, thereby shaping the landscape of modern healthcare. Digital pathology builds upon traditional histopathology by harnessing digital imaging technology to digitize and automatically analyze tissue samples. Whole slide images (WSI) emerge as a pivotal modality in this context. WSI are high-resolution scans of biological samples stained with specific chemicals to highlight key features, such as nuclei, tissue architecture, and protein distribution. This promising field opens new avenues for advanced computational techniques to improve diagnostic accuracy for various pathologies, ranging from cancer to Parkinson's disease [1], [2].

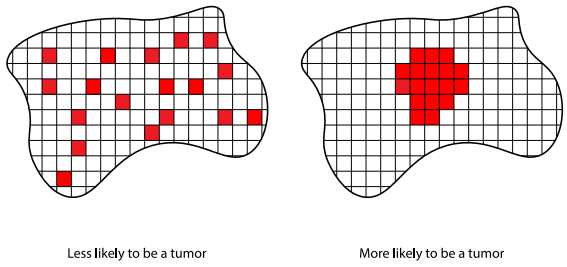

Fig. 1. The importance of leveraging the spatial context in WSI analysis. The following example considers a model classifying tiles as tumor (red squares) or non-tumor (white squares). Two different results for the model are displayed at the slide level, with the same number of tiles classified as positive (red) and negative (white). As tumor cells are known to be close to each other, it could be naturally concluded that the prediction on the right is more likely to be close to the ground truth than the one on the left. However, for most MIL approaches, both are indistinguishable since they do not consider the spatial context.

Computational pathology has the potential to yield significant contributions to the medical domain. It can first automate the analysis of vast amounts of complex histological data, reducing the time required for diagnosis and enhancing workflow efficiency. Moreover, this analysis relies on extracting nuanced, exhaustive, and deterministic information from tissue images, leading to more comprehensive and reproducible diagnoses. From another perspective, developing predictive models and biomarkers for disease prognosis and treatment response paves the way for personalized medicine approaches.

This work was supported in part by the PRISM project funded by France 2030 and grant number ANR-18-IBHU-0002.

Loïc Le Bescond is with the Centre for Visual Computing, CentraleSupélec, Gif-sur-Yvette, 91192 France, and UMR981, Gustave Roussy, Villeuif, 94805 France (e-mail: loic.le-bescond@centralesupelec.fr).

Marvin Lerousseau is with the Centre for Computational Biology, Mines Paris, Paris, 75006 France, and Institut Curie, PSL University, Paris, 75005 France, and UMR900, INSERM, Paris, 75005 France (e-mail:marvin.lerousseau@curie.fr).

Fabrice André is with UMR981, Gustave Roussy, Villejuif, 94805 France (e-mail:fabrice.andre@gustaveroussy.fr).

Hugues Talbot is with the Center for Visual Computing, CentraleSupélec, Gif-sur-Yvette, 91192 France (e-mail:hugues.talbot@centralesupelec.fr).

Indeed, computational pathology facilitates large-scale data aggregation and analysis, empowering researchers to uncover novel insights into disease mechanisms and therapeutic targets. Overall, the integration of computational methods into pathology holds immense promise for advancing both clinical practice and scientific discovery in the field [3]–[5].

The sheer size of WSI in computational pathology presents a formidable challenge, often requiring scalable and efficient analysis methods. Indeed, uncompressed WSI usually weigh dozens of gigabytes and contain billions of pixels. Because most of the information is at the cell level, it is not viable to downsample WSI for many tasks, including tumor grading, tumor subtyping, or survival prediction. To circumvent this issue, the Multiple Instance Learning (MIL) framework [6] has been widely adopted. MIL partitions WSI into smaller tiles or instances, allowing for more manageable processing and analysis. However, a significant challenge in MIL-based WSI analysis lies in effectively integrating spatial context information between these tiles. Histopathological images contain crucial spatial relationships between tissue structures and cellular features, essential for accurate diagnosis. Conventional MIL methodologies often struggle to capture and utilize this spatial context efficiently, which could lead to sub-optimal performance in tasks requiring precise localization or characterization of pathological findings. Overcoming these challenges necessitates the development of novel MIL frameworks that can effectively leverage spatial context information at multiple scales, thus improving the accuracy and reliability of WSI analysis in computational pathology.

A few recent computational pathology approaches have made strides toward integrating the spatial context between the instances but still present limitations. Transfomer-based MIL approaches are computationally expensive and require enough tissue at each magnification to integrate multi-scale information [7], [8]. Streaming convolutional neural networks (CNN) do not scale efficiently with image size, leading to prolonged training and inference times unsuitable for clinical applications [9]. Graph-based approaches are challenging to design due to the expertise required for graph construction, and many graph structures furthermore lack the equivariance and invariance properties of images, impeding the design of efficient training strategies [10].

Building upon the groundwork of Lerousseau *et al.* [11], our approach capitalizes on the spatial relationships between tiles in Whole Slide Image (WSI) classification by leveraging CNN architectures, a widely utilized spatially-aware model in image analysis that remains underutilized in MIL approaches. To this end, we represent WSI as multi-dimensional sparse images by randomly sampling tiles from the original images, thus avoiding the need to downsample the original WSI while preserving crucial spatial interactions. Additionally, we put forward the use of sparse convolutions to process such representations. These convolutional operations are specifically designed to model spatial relationships while being memory efficient in a sparse setting. By leveraging these properties, we harness the properties of conventional CNN architectures for WSI data analysis, effectively overcoming the computational challenges typically associated with processing such large-scale images.

More specifically, we introduce in SparseXMIL, a new MIL model that integrates sparse convolutions with the Xception architecture [12] to capture spatial interdependencies within WSI across multiple scales. We showcase its efficiency across five classification tasks spanning three datasets, where SparseXMIL achieves competitive or state-of-the-art results while ensuring interpretability. Furthermore, we assess the influence of tissue spatial relationships through sensitivity analyses, providing valuable insights into the integration of spatiality within MIL approaches and its impact on performance.

SparseXMIL is built on genericity, ensuring its usability for many clinically impactful tasks such as diagnostics, treatment response, or patient stratification. SparseXMIL could have significant implications for the fields of pathology and precision medicine, offering a more nuanced and comprehensive approach to histological slide analysis.

II. Background

A. Multiple Instance Learning

We begin by outlining the formalism of MIL applied to WSI data. In a conventional supervised classification scenario, the objective is to train a model $\mathcal{M}_\theta$ to predict a label $Y \in \{0, 1\}$ from a WSI $X \in \mathbb{R}^{d \times W_X \times H_X}$, where d represents the number of channels (typically 3), W_X denotes the slide's width, and H_X its height. Given that typical WSI consists of billions of pixels and can exceed 30 gigabytes when uncompressed, they cannot be entirely accommodated in GPU memory, let alone processed by deep learning models. To address this constraint, each WSI X is conventionally partitioned into a set of N_X images, commonly referred to as patches or tiles, resulting in an initial slide representation $X = \{x_1, ..., x_{N_X}\}$. For simplicity, we assume that all tiles have the same spatial dimensions (w, h), although they may be of different sizes.

A MIL model $\mathcal{M}_\theta$ typically comprises three sequential components:

- An instance embedder $f : \mathbb{R}^{(d,w,h)} \mapsto \mathbb{R}^l$ extracting from each tile x_i a fixed-dimensional (tile) representation vector v_i of size l,
- A pooling operator $\Theta : \mathbb{R}^l \mapsto \mathbb{R}^m$ mapping all the representations v_i into a single (slide) representation vector v of size m,
- A classifier $g : \mathbb{R}^m \mapsto \{0, 1\}$ generating the final prediction $\hat{Y}$ for the WSI,

In most cases, the instance embedder f remains frozen, meaning it is not trained during MIL training, potentially limiting performance by lacking tile-level features specific to histological slides. Accordingly, in this paper, we allow f to be trained alongside other components of the SparseXMIL architecture.

MIL methodologies for WSI classification can be partitioned into two groups depending on their pooling operators Θ: those that are permutation invariant, and those designed to exploit the spatial relationships among instances.

1) Permutation-invariant pooling operators: This first category comprises methods that process instances separately and so, independently of their spatial arrangement on X. Among these approaches, the pooling operator is invariant to

instance permutations, such that, for any random permutation σ: $\Theta(f(\sigma(X))) = \Theta(f(X))$. This category includes instance-based methods that identify specific instances to form the prediction, employing simple pooling operators like max or its variant, top-k max [13]. It also encompasses representation-based methods that merge the representations of all instances into a single WSI representation, using operators such as mean, log-sum-exp, and noisy logical operators [14], [15]. To enhance interpretability, Ilse *et al.* [16] introduced Attention-MIL, which incorporates attention and gated attention pooling operators to compute the bag representation by weighted sums of each instance's representation, with weights learned from the instance representations via a simple deep-learning model. Subsequent approaches include incorporating clustering constraints [17], exploiting WSI's hierarchical structure [18], and employing pre-training strategies with self-supervised learning [19]. However, these methodologies do not inherently consider spatial correlations between instances, potentially limiting their performance for tasks requiring spatial awareness.

2) MIL with spatial-aware pooling: The second category of MIL methods explicitly considers the relationships between instances. Zhou *et al.* [20] were among the first to emphasize the importance of treating instances as interconnected entities in image analysis, highlighting that $\Theta(f(\sigma(X))) \neq \Theta(f(X))$, where σ represents a random perturbation of X. To model these dependencies, they introduced a graph representation based on Euclidean distances between instances. Subsequent approaches have employed various graph-based strategies to capture instance dependencies using algorithms such as ϵ-graphs, k-nearest neighbors, or Delaunay triangulation, applied at different patch or cellular levels [21]. Graph convolutional neural networks (GNN) were then applied to these structures to pool instance representations while integrating neighborhood context [22]–[24]. Despite their effectiveness in modeling local relationships, graphs pose challenges in designing efficient training strategies when applied to WSI analysis. In particular, the performance of GNN-based methods relies on the initial graph structure. Still, no consensus is yet available on the proper structure to select for WSI analysis, and learning the graph's topology from the node is computationally expensive [21]. Moreover, traditional image augmentations, such as rotations or flipping, are not directly applicable to the non-Euclidean graph domain depending on the chosen graph structure, potentially compromising model generalization.

Transformer architectures have been explored in conjunction with graph-based methods to better capture spatial interdependencies. For example, the TransMIL model incorporates spatial dependency through its PPEG module, which assumes an ordering among the instances to create a projection into a 2-D image space. Position encoding is then computed by summing depth-wise convolutions of various kernel sizes applied to this image. This concept has been further expanded with diverse positional encoding strategies [25], multi-scale methodologies [26], [27], and the integration of self-supervised learning [8]. However, TransMIL's PPEG module assumes that tiles can be projected into square images, which does not hold in practice as tiles are extracted from the tissue with different shapes, and the multi-scale processing integration proposed by HIPT cannot be applied in slides with insufficient tissue at lower magnifications.

Alternative approaches have attempted to process entire slides in a single input, like streaming CNNs, but these methods' computation time scales linearly with the size of the images, which could impeach their direct application to WSI data [9].

In this paper, we propose to leverage augmentations naturally present in images (rotations, flipping, ...) through the use of CNN architectures tailored to exploit these characteristics. Our new sparse image representation, based on tile coordinates, overcomes the constraints from other spatially-aware pooling methods, such as the requirement for tiles to be projected onto a square grid structure and the difficulties associated with multi-scale processing in slides with limited tissue at lower magnifications. By integrating sparse convolutions, we achieve scalability of CNN architectures to the WSI level, facilitating comprehensive multi-scale analysis.

B. Sparse Convolutions

Sampling instances leads to a sparser representation of the WSI, rendering traditional convolutional approaches less effective [28]. We propose using sparse convolutions to model spatial dependencies efficiently between the instances in such a sparsified setting. Developed initially for point cloud analysis [29], sparse convolutions have demonstrated promising results when applied to WSI data, as demonstrated by Lerousseau *et al* [11]. However, their proposed network inadequately addresses the multi-scale nature of WSI. Comprising only a few convolutional layers with fixed kernel sizes, it fails to consider the spatial dependencies at the multiple scales present in large images. Our proposed methodology employs a sparse convolutional neural network architecture specifically designed to model these interactions across multiple magnifications.

Our approach mainly focuses on the utilization of the Minkowski convolutional neural network framework [30]. Unlike standard dense convolutions, this framework takes a sparse tensor $\mathcal{T}$ as input, which consists of two matrices: a coordinate matrix $\mathbf{C}$ representing the sampled locations, or active sites, and a feature matrix $\mathbf{F}$ composed of the associated features. In other words, if we consider a batch of images, a sparse tensor $\mathcal{T}$ can be represented as follows:

$$\mathcal{T} = \{\mathbf{C}, \mathbf{F}\}$$
$$\text{with } \mathbf{C} = \begin{bmatrix} b_1 & c_1^1 & c_1^2 & \dots & c_1^D \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_n & c_n^1 & c_n^2 & \dots & c_n^D \end{bmatrix}, \ \mathbf{F} = \begin{bmatrix} f_1^T \\ \vdots \\ f_n^T \end{bmatrix} \tag{1}$$

where $\forall i \ b_i \in \mathbb{Z}^+$ denotes the batch index, $\forall i \ c_i \in \mathbb{Z}^D$ represents the D-dimensional coordinates of each active site, and $\forall i \ f_i$ corresponds to the features associated with these specific locations in the images of the batch (such as RGB values).

In addition to this sparse tensor representation, Choy *et al.* [30] introduced a novel generalized convolution operation

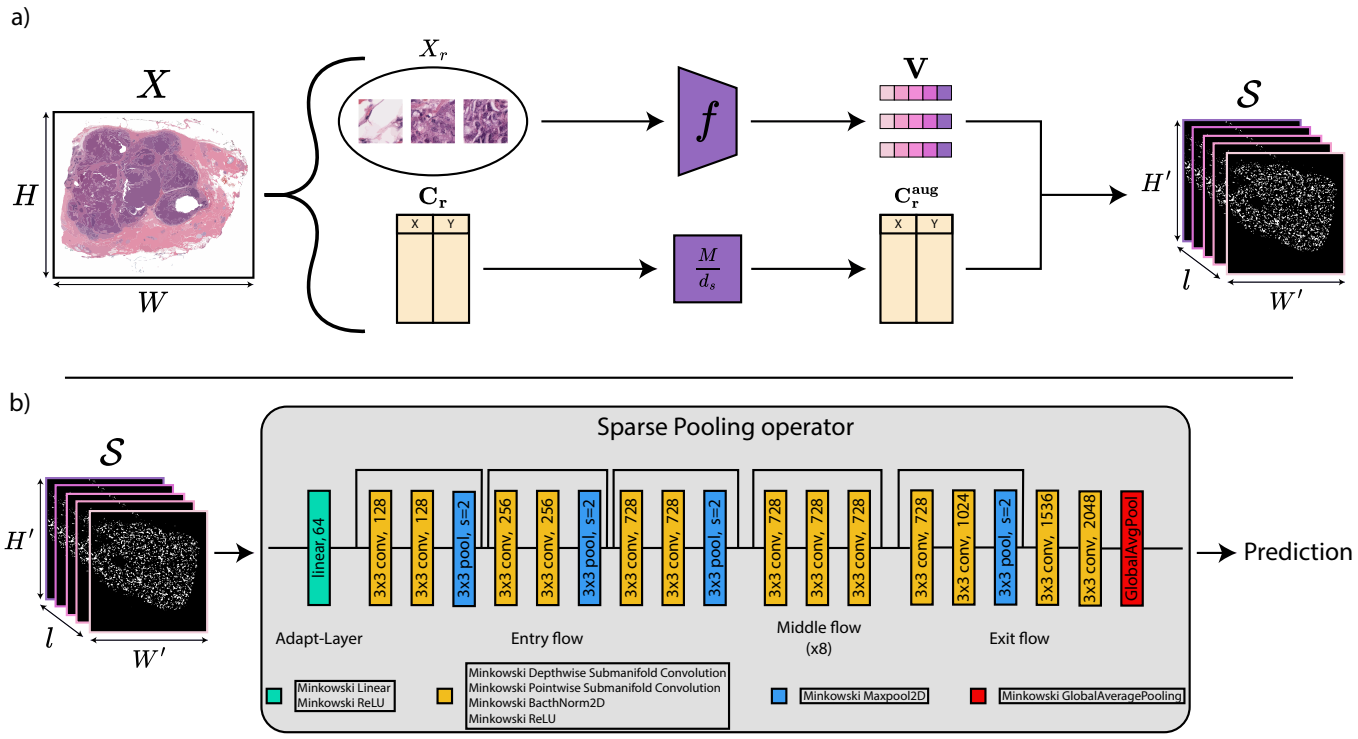

Fig. 2. a) Sparse Image Representation. From a given input WSI $\boldsymbol{X}$ of size $(\boldsymbol{W}, \boldsymbol{H})$, we sample a set of $\boldsymbol{r}$ instances $\boldsymbol{X_r}$ of coordinates $\mathbf{C_r}$. The set of instances is forwarded to the tile embedder $\boldsymbol{f}$ that extracts a fixed-sized representation from each instance, resulting in the matrix $\mathbf{V} \in \mathbb{R}^{\boldsymbol{r},\boldsymbol{l}}$. Meanwhile, the coordinates are downsampled by a factor $\boldsymbol{d_s} = (\boldsymbol{w}, \boldsymbol{h})$ and augmented through a transformation matrix $\boldsymbol{M}$, resulting in $\mathbf{C_r^{aug}}$. Finally, the embeddings from $\mathbf{V}$ are aligned with their corresponding coordinates in in $\mathbf{C_r^{aug}}$, leading to the sparse image $\boldsymbol{\mathcal{S}} = \{\mathbf{C_r^{aug}}, \mathbf{V}\}$, of dimension $\boldsymbol{W'} = \boldsymbol{W}/\boldsymbol{w}$ and $\boldsymbol{H'} = \boldsymbol{H}/\boldsymbol{h}$, with $\boldsymbol{l}$ channels. b) Sparse pooling operator architecture. The sparse pooling operator architecture takes the sparse image representation as input to produce a final prediction for the WSI. It is first composed of the Adapt-Layer that maps the $\boldsymbol{l}$ channels to a smaller dimension. Then, a sparse model built upon the Xception architecture and Minkowski submanifold convolutions construct the final prediction from the output of the Adapt-Layer.

that extends the concept of traditional dense convolutions to sparse tensors. Considering an input feature vector $v_u^{\text{in}} \in \mathbb{R}^{d^{\text{in}}}$ at coordinate $u \in \mathbb{R}^D$, and a convolutional kernel weights $W \in \mathbb{R}^{K^D \times d^{\text{out}} \times d^{\text{in}}}$, we have:

$$v_u^{\text{out}} = \sum_{i \in \mathcal{N}^D(u,K,\mathcal{C}^{\text{in}})} W_i v_{u+i}^{\text{in}} \text{ for } u \in \mathcal{C}^{\text{out}} \tag{2}$$

where K represents the kernel size, $W_i \in \mathbb{R}^{d^{\text{out}} \times d^{\text{in}}}$ corresponds to one of the K^D spatial weight matrices, $\mathcal{C}^{\text{in}}$ and $\mathcal{C}^{\text{out}}$ denote the lists of input and output active sites, respectively. $\mathcal{N}^D(u, K, \mathcal{C}^{\text{in}})$ represents the set of offsets i such that $u + i \in \mathcal{C}^{\text{in}}$ and $i \in \mathcal{N}^D(K)$, with:

$$\mathcal{N}^D(K) = \begin{cases} [0, K)^D \bigcap \mathbb{Z}_+^D \text{ for } K \text{ even} \\ [-\frac{K-1}{2}, \frac{K-1}{2}]^D \bigcap \mathbb{Z}_+^D \text{ otherwise} \end{cases} \tag{3}$$

This framework aligns well with our perspective of WSI as sparse data. In addition, the use of Minkowski convolutions in our model addresses a fundamental challenge in processing WSI: the balance between capturing detailed information at the tile level and understanding the broader spatial context of the slide. Thus, Minkowski convolutions represent a powerful tool for extracting meaningful patterns and relationships from the WSI data, which are critical for accurate classification and analysis.

III. Methods

This section introduces SparseXceptionMIL (SparseXMIL), extending the traditional Multiple Instance Learning framework to integrate spatial dependencies into the WSI analysis. The first part consists in constructing a sparse image representation from the WSI. To this end, we rely on tile embeddings and coordinates to project the information of each tile onto a multi-dimensional sparse image. Then, acknowledging the limitations of standard convolutions in processing large-scale sparse data, we propose a new sparse pooling operator built upon Minkowski convolutions and the Xception architecture, suitable to process large-scale data. The global design of our method is illustrated in Fig. 2.

A. Sparse image representation

Our approach diverges from previous MIL methodologies by retaining the original coordinates at which the tiles were extracted on the input slide in a matrix denoted as $\mathbf{C_r} = [c_1, ..., c_r]^T$, where $c_i \in [\![1, W]\!] \times [\![1, H]\!]$. By indexing each instance with its extraction coordinates, we are able to map each tile onto a sparse image $\mathcal{S}$ reflecting the actual spatial distribution of the tiles on the slide. For the sample size r, we set a sample size $r = p \times N$ proportional to the size of the input slide. This way, we can guarantee that the density within

the initial convolutional layers' receptive field $\mathcal{N}^D(u, K, \mathcal{C}^{in})$ remains consistent across varying input dimensions.

In practice, we first employ a tile embedder f to extract a fixed-size feature vector $v_i \in \mathbb{R}^l$ from each tile within $X_r = \{x_1, ..., x_r\}$. We then concatenate the extracted v_i into a feature matrix $\mathbf{V} = [v_1, ..., v_r]^T$, and create a sparse tensor $\mathcal{T} = \{\mathbf{C_r}, \mathbf{V}\}$. However, considering WSI typically measures tens of thousands of pixels, using the raw coordinates would result in huge gaps between adjacent tiles. To address this problem, we apply a downsampling factor d_s to the coordinates in $\mathbf{C_r}$, ensuring minimal spacing between consecutive tiles. This factor is set to $d_s = (w, h)$, facilitating a one-pixel difference between two adjacent tiles. Without a proper downsampling factor, the architecture's initial convolutional layers would essentially function as fully connected layers, which is undesirable. Conversely, a higher downsampling factor could lead to information loss, akin to reducing the resolution of the original slide. Therefore, $d_s = (w, h)$ represents an optimal balance, preserving information while making the spatial context exploitable.

To better exploit the spatial intricacies present in WSI data, we additionally apply geometric transformations on the coordinate matrix $\mathbf{C_r}$. These transformations, which encompass rotations, translations, flips, and resizing operations, are applied to enhance the model's perception of spatial relationships by introducing variability in the representation of tile arrangements. The composition of these transformations yields a transformation matrix M that varies with each data iteration. Our sparse image representation is thus defined as the sparse tensor $\mathcal{S} = \{\mathbf{C_r^{aug}}, \mathbf{V}\}$, where $\mathbf{C_r^{aug}} = \mathbf{C_r}\frac{M}{d_s}$ represents the adapted set of coordinates. This sparse image representation functions as a multi-dimensional sparse image $\mathcal{S} \in \mathbb{R}^{l \times W' \times H'}$, with $W' = W/w$ and $H' = H/h$.

Our method allows for the efficient processing of large-scale WSI through this new sparse image representation, balancing the inclusion of detailed tile features with the broader spatial context. By considering each tile's features and direct spatial locations, our model aims to provide a more comprehensive analysis of WSI data.

B. Sparse pooling operator

The first central component of the Sparse pooling operator is the Adapt-Layer, which serves as the initial stage of our sparse convolutional architecture. This layer refines the embeddings generated by the tile embedder, ensuring that they are optimal for subsequent processing. The Adapt-Layer comprises a straightforward yet effective arrangement of a single linear layer followed by a ReLU (Rectified Linear Unit) activation function. Its primary function is to downsample the embeddings from the feature matrix $\mathbf{V} \in \mathbb{R}^{r,l}$ into a reduced size $\mathbf{V'} \in \mathbb{R}^{r,l'}$, where $l' < l$. This downsampling operation is crucial to minimize the memory demands of the subsequent convolutional layers, which are often the most resource-intensive components of the network. Considering the typical dimensions of feature representations in the initial layers of widely-used CNN architectures like ResNet and Inception, we set $l' = 64$.

The second principal component of the Sparse pooling operator is a sparse convolutional architecture tailored to process the feature-downsampled sparse image $\mathcal{S}' = \{\mathbf{C_r^{aug}}, \mathbf{V'}\}$ as input to generate the final prediction $\hat{Y}$. We specifically employed depthwise separable convolutions, which separate spatial and channel processing into two stages: a depthwise convolution followed by a pointwise convolution. The advantage of depthwise separable convolutions lies in their efficiency while requiring fewer parameters than regular convolutions, which is particularly appealing in the context of large input data such as WSI.

In constructing our sparse model, we focus particularly on the concept of submanifold convolutions as they retain sparsity throughout the depth of the network. As highlighted by Graham *et al.* [31], regular sparse convolutions tend to reduce the sparsity in the image across layers by dilating the number of active sites. Indeed, for each sparse convolution, the set of new output coordinates C^{out} is defined by:

$$C^{\text{out}} = \{u + i; u \in \mathcal{C}^{\text{in}}, i \in \mathcal{N}^D(u, K, \mathcal{C}^{\text{in}})\} \tag{4}$$

To solve this problem, submanifold sparse convolutions suggest keeping the same set of input coordinates as output coordinates. By ensuring $\mathcal{C}^{\text{in}} = \mathcal{C}^{\text{out}}$, the architecture maintains the integrity of the sparse structure. However, solely using submanifold convolutions may restrict the ability of the model to capture global information. Hence, incorporating pooling operators or strided convolutions become essential to process the WSI data globally.

To this end, our sparse pooling operator builds upon the Xception model proposed by Chollet [12]. This architecture has been exploited in previous research in histopathology [32] and is known to handle large-scale data well. Our primary modification to the original architecture excludes the first two convolutional blocks. This adjustment accounts for the tile embedder and Adapt-Layer operations that already incorporate local context through tile embeddings. We then replicate the design of the original architecture, substituting dense convolutions with submanifold convolutions throughout the model. The inclusion of pooling operators and strided convolutions allows the model to capture broader information while maintaining sparsity.

To generate the final predictions $\hat{Y}$, we employ global average pooling on the features of the resulting active sites. This pooling operation aggregates the spatial information into a single feature vector. Subsequently, this singular feature vector is passed to a single linear layer, resulting in the prediction $\hat{Y}$.

IV. Experiments

We benchmarked the proposed approach against the traditional MIL approaches for WSI classification across three different datasets.

A. Datasets

The benchmark included three well-known subtyping tasks involving diagnosis slides obtained from the TCGA database:

- The first task involves the classification of Invasive Ductal Carcinoma (831 patients) and Invasive Lobular Carcinoma (210 patients) within the context of Invasive Breast Carcinoma (BRCA).
- The second task focuses on subtyping Non-Small Cell Lung Carcinoma, distinguishing between Lung Adenocarcinoma (528 patients) and Lung Squamous Cell Carcinoma (505 patients).
- The third task involves subtyping Clear Cell, Papillary, and Chromophobe Renal Cell Carcinoma (517, 294, and 120 patients, respectively) within Renal Cell Carcinoma (RCC).

Additionally, we explore the application of SparseXMIL in predicting more challenging outcomes that could benefit from spatial-context-aware methods. Specifically, we investigate the prediction of DNA Damage Response genomic alterations in Breast Cancer. Our objective is to predict the status of Homologous Recombination Deficiency (HRD) vs Homologous Recombination Proficiency (HRP) of cancer cells. This status, measured by the HRD score, plays a key role in the design of therapeutic strategies. For this purpose, we employ two binarization strategies -mHRD (447 HRD vs 465 HRP) and tHRD (318 HRD vs 316 HRP)- computed on BRCA patients data, as proposed in [33]. The labels were extracted from the TCGA database for each task, except for HRD prediction, where we collected the labels from [33].

To prepare the slides for analysis, we first applied several pre-processing steps for tissue extraction. They consist in applying a threshold on texture features extracted from the tissue at low magnification and discarding white and black patches based on RGB values. We then extracted 256×256 tiles at the highest magnification.

B. Baselines

We conducted a comprehensive comparative analysis to benchmark our method against the most prevalent and high-performing approaches. To this end, we implemented five distinct models under two distinct training protocols. The models include Average MIL and Attention MIL [16], which are permutation-invariant. Additionally, we evaluated context-aware MIL models, namely SparseConvMIL [11], TransMIL [7], and GCN-MIL [22].

C. Experimental Settings

For the experimentation, we maintained consistent settings across all implemented models to ensure fair comparisons. We explored two settings, one with instance sampling and one without, to evaluate the impact of sampling on the performance of the different MIL methodologies.

In our first experimental setting, we uniformly sampled **20%** of the tiles from each slide during training but also during both validation and testing, employing this approach as a test-time augmentation technique.

Our SparseXMIL method is designed to enable end-to-end training, spanning from the tile to the slide level. However, we identified a trade-off between the number of tiles sampled and the resolution at which they are extracted. It is crucial to sample a proportional number of tiles to capture spatial interactions effectively. Yet, achieving this at the highest resolution necessitates processing a large number of instances ($\approx 10^4$ tiles per slide), which was not tractable within our hardware constraints. To address this issue, we opted to freeze the tile embedder, which conserves memory space and reduces computation time. Following the pipeline introduced by Lu *et al.* [17], we adopted, for all benchmarked methods, a Resnet50 model truncated after the third residual block pre-trained on ImageNet. We thus extract feature vectors of size $l = 1024$ for each tile.

Each model was trained using a batch size of 16 WSI, Adam optimizer with a learning rate of 2e-4, and a weight decay of 1e-7. Training was carried out for up to 100 epochs (extended to 150 epochs for AverageMIL due to its longer convergence time) on an Nvidia A6000 GPU. During training, we additionally introduced an exponential moving average on the weights of the inference model with a decay factor of 0.99. Regarding SparseXMIL and SparseConvMIL, we applied a downsampling factor $d_s = (256, 256)$ and introduced various data augmentation strategies on our sparse image representations, including random rotation, flipping, resizing, and translation operations.

During inference, we implemented a test-time augmentation scheme for both the validation and testing stages for all methods. We performed 10 runs per WSI while preserving instance sampling (and sparse image augmentations for SparseConvMIL and SparseXMIL), and subsequently averaged the output probabilities to produce the final predictions.

In the second setting, we trained Attention MIL, TransMIL, and GCN-MIL using all the instances. Under this condition, we set the batch size to 1 and removed test-time augmentation. All the other parameters remained consistent with the first setting, including using a frozen tile embedder pre-trained with ImageNet weights and the use of exponential moving average.

We assessed model performances using the macro-averaged area under the curve (AUC) metric across a 10-fold cross-validation framework. We aligned our evaluation with publicly available splits to contextualize our results with the state-of-the-art performance on these tasks. For the subtyping tasks, we use the splits provided in [8](train/val/test decomposition), including those slides omitted in their study due to insufficient tissue at high magnification. As our method can handle varying tissue densities and resolutions, such slides remain valuable and analyzable within our framework. Regarding HRD prediction, we adhered to the 10-fold cross-validation splits detailed [33].

V. Results

A. Classification tasks

The results of the subtyping tasks are outlined in Table. I. Comparing SparseXMIL against other benchmarked models across various settings, our model achieves superior performance in the breast and lung subtyping tasks, with notable gains observed in the BRCA dataset. An interesting observation is the consistently high performance of context-aware methods, including SparseXMIL, Transmil, and GCN-MIL. This underscores the importance of incorporating spatial

TABLE I

RESULTS ON SUBTYPING TASKS. MEAN AND CONFIDENCE INTERVAL VALUES OF THE AVERAGED AUC COMPUTED ON A 10-FOLD CROSS-VALIDATION (THE MACRO-AVERAGED AUC FOR RCC SUBTYPING). THE TWO BEST VALUES FOR EACH METRIC IS HIGHLIGHTED IN **BOLD**. ✓* DENOTES METHODS WITH INSTANCE SAMPLING AND SPARSE IMAGE AUGMENTATIONS.

Method	Instance Sampling	BRCA	NSCLC	RCC
Average MIL	✗	0.883 ± 0.043	0.918 ± 0.021	0.985 ± 0.008
GCN-MIL	✗	0.866 ± 0.048	0.924 ± 0.017	0.985 ± 0.007
Attention MIL	✗	0.856 ± 0.047	0.931 ± 0.026	0.978 ± 0.007
TransMIL	✗	0.682 ± 0.143	0.902 ± 0.056	0.976 ± 0.012
Average MIL	✓	0.840 ± 0.047	0.882 ± 0.021	0.981 ± 0.008
Attention MIL	✓	0.877 ± 0.028	$0,948 \pm 0,024$	0.987 ± 0.005
SparseConvMIL	✓*	0.861 ± 0.043	0.902 ± 0.026	0.983 ± 0.008
GCN-MIL	✓	0.873 ± 0.050	0.932 ± 0.021	$\mathbf{0.989 \pm 0.005}$
TransMIL	✓	$\mathbf{0.891 \pm 0.032}$	$\mathbf{0,959 \pm 0,019}$	0.987 ± 0.006
Proposed	✓*	$\mathbf{0.910 \pm 0.035}$	$\mathbf{0,960 \pm 0,021}$	$\mathbf{0.988 \pm 0.005}$

TABLE II

RESULTS ON HRD PREDICTION. MEAN AND CONFIDENCE INTERVAL VALUES OF THE AVERAGED AUC COMPUTED ON A 10-FOLD CROSS-VALIDATION. THE TWO BEST VALUES FOR EACH METRIC IS HIGHLIGHTED IN **BOLD**. ✓* DENOTES METHODS WITH INSTANCE SAMPLING AND SPARSE IMAGE AUGMENTATIONS.

Method	Instance Sampling	BRCA mHRD	BRCA tHRD
Average MIL	✓	0.675 ± 0.071	0.784 ± 0.029
SparseConvMIL	✓*	0.686 ± 0.063	0.796 ± 0.033
Attention MIL	✓	$\mathbf{0.717 \pm 0.060}$	$\mathbf{0.823 \pm 0.037}$
GCN-MIL	✓	0.697 ± 0.061	0.818 ± 0.036
TransMIL	✓	0.714 ± 0.060	0.809 ± 0.056
Proposed	✓*	$\mathbf{0.723 \pm 0.055}$	$\mathbf{0.822 \pm 0.057}$

dependencies into WSI analysis. The distinct advantage of our approach in two of the three tasks highlights the effectiveness of our new sparse image representation and our proposed architecture in exploiting spatial relationships between tiles. Notably, the breast subtyping task, where our method excelled, is particularly sensitive to spatial context [34], further emphasizing the significance of spatially aware approaches in WSI classification.

Across all three datasets, instance sampling during both the training and testing phases consistently improved the performance of nearly all methods compared the results reported in the HIPT paper and our own findings without instance sampling. TransMIL, in particular, demonstrated notable performance improvements in breast and lung subtyping tasks, indicating that instance sampling is particularly advantageous for this model. Additionally, instance sampling also contributed to noticeable improvements in the performance of AttentionMIL and GCN-MIL, and reduced performances for Average MIL.

These observations underscore the potential of instance sampling for data augmentation and regularization, which is particularly effective in mitigating overfitting issues, as observed with TransMIL. This correlates well with research conducted by Tarkhan *et al.* [35], suggesting that instance sampling in WSI can save computing time and resources, but also improve the classifier's performance. Such evidence supports the potential effectiveness of instance sampling as a form of data augmentation, which could be beneficial during both the training and inference stages for specific MIL models. Based on these insights, we retained instance sampling for subsequent experiments.

The results for HRD prediction are presented in Table II. Here, our method SparseXMIL once again demonstrates superiority over other methods. However, the other high-performing method in these tasks is Attention MIL, which does not inherently account for spatial context. This indicates that these tasks are significantly more challenging and that the best-performing methods may be the ones least prone to overfitting. Nevertheless, our approach still outperforms the other models in mHRD prediction, confirming the robustness of our framework.

In summary, these results validate the interest of our proposed strategy in capturing the spatial context present in WSI data, and enhancing the classification performance across various tasks.

B. Sensitivity analysis

One main component of our method lies in its ability to integrate and leverage the spatial relationships between tiles. To evaluate the impact of spatial relationships on the performance of our method and other spatial-aware MIL models, we conducted a sensitivity analysis, assessing how two types of perturbations during the inference process affect the prediction.

1) Perturbation 1: Shuffling Tile Localizations: The first perturbation involved shuffling the localization of the tiles. This was achieved by applying a random permutation to the coordinate vectors of the tiles for each slide. Formally, this can be expressed as:

$$\mathbf{C}' = [c_j; j = \sigma(i), i \in \{1..N\}, c_j \in \mathbf{C}]^T \tag{5}$$

where N represents the number of tiles for a given slide, $\mathbf{C}$ denotes the coordinate vector of the slide, and σ represents the random permutation function.

We also applied this perturbation to SparseConvMIL and DCGN-MIL, and randomly permuted the order of the feature vectors for TransMIL to induce the same effect.

TABLE III

RESULTS OF THE SENTIVITY ANALYSIS. MEAN AND CONFIDENCE INTERVAL VALUES OF THE AVERAGED AUC COMPUTED ON A 10-FOLD CROSS-VALIDATION (THE MACRO-AVERAGED AUC FOR RCC SUBTYPING) WITH DIFFERENT PERTURBATIONS AT INFERENCE. ✓* DENOTES METHODS WITH INSTANCE SAMPLING AND SPARSE IMAGE AUGMENTATIONS.

Dataset	Method	Instance Sampling	$\emptyset$	Shuffling	Random
BRCA	GCN-MIL	✗	0.866 ± 0.048	0.813 ± 0.089	0.812 ± 0.088
	TransMIL	✗	0.682 ± 0.143	0.632 ± 0.121	-
	GCN-MIL	✓	0.873 ± 0.050	0.816 ± 0.049	0.809 ± 0.050
	TransMIL	✓	0.891 ± 0.032	0.889 ± 0.030	-
	SparseConvMIL	✓*	0.861 ± 0.043	0.852 ± 0.047	0.848 ± 0.053
	Proposed	✓*	0.910 ± 0.035	0.893 ± 0.032	0.878 ± 0.035
NSCLC	GCN-MIL	✗	0.924 ± 0.017	0.734 ± 0.094	0.730 ± 0.091
	TransMIL	✗	0.902 ± 0.056	0.882 ± 0.084	-
	GCN-MIL	✓	0.932 ± 0.021	0.800 ± 0.050	0.791 ± 0.051
	TransMIL	✓	0.959 ± 0.019	0.962 ± 0.015	-
	SparseConvMIL	✓*	0.902 ± 0.026	0.879 ± 0.035	0.875 ± 0.033
	Proposed	✓*	0.960 ± 0.021	0.954 ± 0.023	0.954 ± 0.020
RCC	GCN-MIL	✗	0.985 ± 0.007	0.929 ± 0.030	0.930 ± 0.029
	TransMIL	✗	0.976 ± 0.012	0.954 ± 0.018	-
	GCN-MIL	✓	0.989 ± 0.005	0.961 ± 0.019	0.959 ± 0.020
	TransMIL	✓	0.987 ± 0.006	0.987 ± 0.005	-
	SparseConvMIL	✓*	0.983 ± 0.008	0.981 ± 0.009	0.982 ± 0.008
	Proposed	✓*	0.988 ± 0.005	0.984 ± 0.007	0.983 ± 0.006

2) Perturbation 2: Randomizing Coordinate Values: The second perturbation involves assigning random coordinates for each tile within each slide's maximum and minimum coordinate values. This perturbation not only disrupts the spatial context of the tiles but also alters the overall shape of the tissue. This modification is particularly relevant for evaluating multi-scale approaches like our proposed method, as it introduces an additional layer of complexity by affecting both the local spatial relationships and the global tissue structure. Mathematically, this can be expressed as:

$$\mathbf{C}' = [(x,y)_i; i \in \{1..N\}, x \sim \mathcal{U}_{[x_{\min}, x_{\max}]}, \\ y \sim \mathcal{U}_{[y_{\min}, y_{\max}]}] \tag{6}$$

where $x_{\min}, x_{\max}$ and $y_{\min}, y_{\max}$ represent the minimum and maximum values along the axes for a given slide. It is worth noting that this second perturbation only applies to SparseConvMIL, DGCN-MIL, and our method, as the other methods do not directly incorporate tile coordinates.

3) Results and Observations: The results from our sensitivity analysis, detailed in Table III, reveal different degrees of sensitivity to spatial perturbations among the evaluated methods. Notably, GCN-MIL shows the most pronounced decline in performance after spatial perturbations across the three subtyping tasks. GCN-MIL's great sensitivity to spatial context perturbation within the instance sampling setting may lie in the KNN-algorithm used to construct the graph, which is not sensible to the number of sampled instances.

Similarly, SparseConvMIL and SparseXMIL present reduced performance when subjected to both perturbations, especially in breast cancer subtyping. Despite this performance drop, it's noteworthy that SparseXMIL continues to rank among the top-performing methods, maintaining robustness against spatial disruptions. Notably, our method SparseXMIL exhibits a more pronounced decrease when applying the second perturbation than SparseConvMIL. This result is likely due to the sensitivity of SparseXMIL to the global tissue structure, as our method operates on different scales with its sparse model architecture.

Finally, TransMIL is less affected by both perturbations. This is particularly interesting as TransMIL, when trained without instance sampling, shows a higher sensitivity to the shuffling perturbation. These findings would suggest that the TransMIL spatial information's processing is deeply impacted by the number of tiles processed, contrary to GCN-MIL.

These different responses to spatial perturbations highlight the varying degrees of dependency on spatial context among the methods, offering valuable insights into how each method relies on spatial information to enhance its performance.

C. Interpretation

TABLE IV

RESULTS OF THE INTERPRETATION ANALYSIS. AVERAGED AUC, F1 SCORE, AND PRECISION COMPUTED OVER TUMOR ANNOTATIONS FROM THE BRCA DATASET, WITH MODELS TRAINED USING INSTANCE SAMPLING ON SUBTYPING PREDICTION. THE BEST VALUE FOR EACH METRIC IS HIGHLIGHTED IN **BOLD**.

Method	AUC	F1 score	Precision
Attention MIL	0.832	0.668	0.590
GCN-MIL	0.801	0.634	0.549
Proposed (exit flow)	0.825	0.683	0.606
Proposed (middle flow)	**0.863**	**0.714**	**0.663**

Interpretability plays an important role in medical image analysis, particularly to ensure the reliability and trustworthiness of model predictions. Commonly, this is achieved through attention weights, which assign scores to each tile, indicating their contribution to the overall prediction.

To evaluate how the different benchmarked methods identify relevant information for prediction, we conducted an analysis using attention-weighted heatmaps as a proxy for tumor segmentation in the breast subtyping task. We used 489 external annotations of tumor segmentation performed on BRCA WSI provided by Gao *et al.* [36]. Our focus was to evaluate the AUC and the precision of these heatmaps compared to the provided annotations. We aimed to quantify the extent to

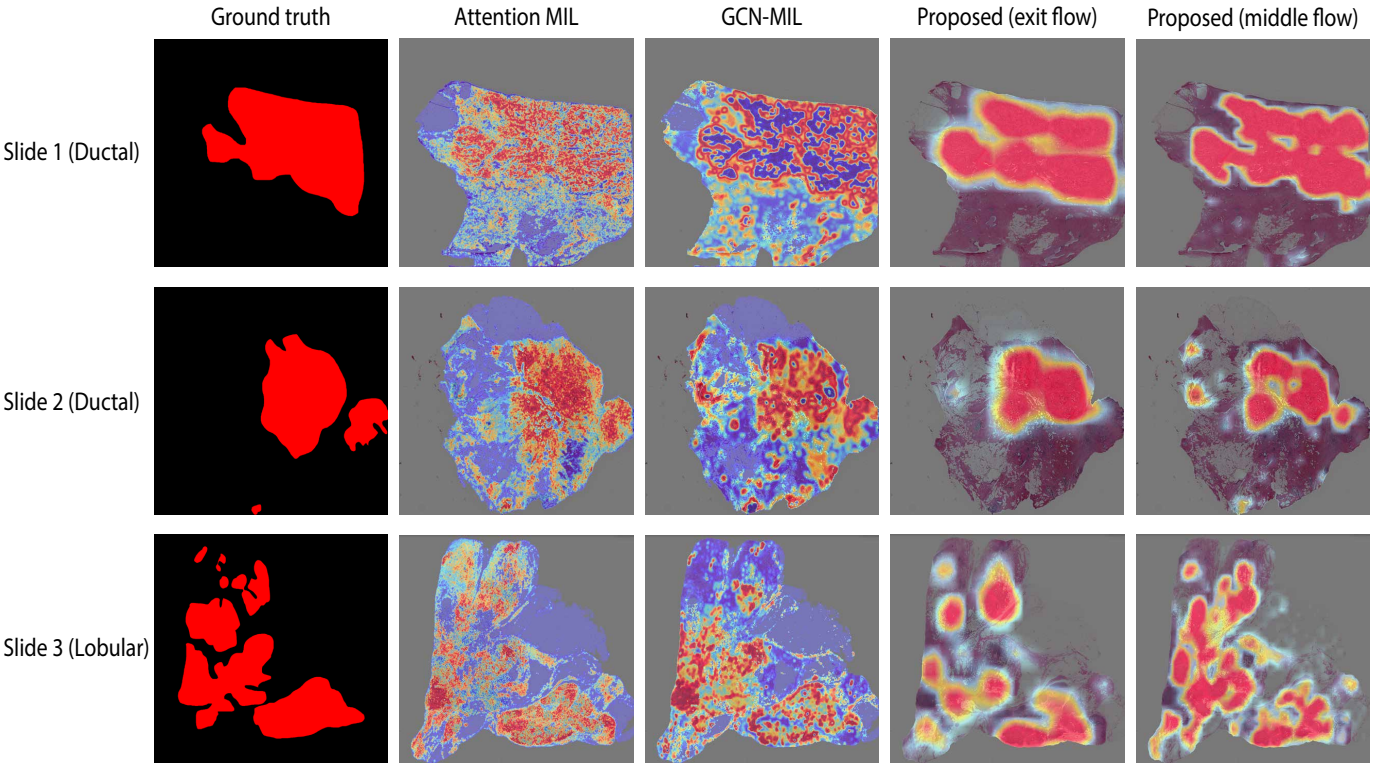

Fig. 3. Heatmaps of attention scores generated for various benchmarked methods from BRCA sample slides. The attention scores were generated with the models on the best split for breast subtyping. We juxtaposed the ground truth tumor masks sourced from Gao *et al.* [36]. The color gradient from blue to red signifies increased attention scores, with red indicating the highest attention.

which models utilize relevant information for predictions while penalizing methods focusing on non-relevant tissue.

For this analysis, we extracted the attention scores from Attention MIL and GCN-MIL using their original implementation to retrieve the attention weight. We were unable to include TransMIL in this analysis, as computing attention weights for all slides proved infeasible under our hardware constraints. We selected the best-performing models from the 10-fold cross-validation splits that were trained without instance sampling. This approach allowed us to evaluate the effectiveness of attention mechanisms in these models under conditions similar to those in which they were originally designed. As SparseXMIL does not incorporate any attention modules, we implemented the GradCAM algorithm to create the corresponding heatmaps. GradCAM offers the advantage of visualizing activations at varying scales. Aiming to explore SparseXMIL's interpretative capabilities across different scales, we chose to analyze activations from two distinct layers: the final layer of the middle flow and the final layer of the exit flow within the sparse convolutional architecture. Here, we also selected the model that performed best across the 10 splits but trained with instance sampling, and we then averaged the heatmaps across the multiple test time augmentations.

To determine the optimal threshold for precision calculations, we set aside 10% of the annotations to form a validation set and computed the AUC and precision metrics on the remaining slides.

The results of this interpretation analysis are summarized in Table IV. We observe that the heatmaps produced with our proposed method at both analyzed scales exhibit higher F1-score and precision compared to the other benchmarked methods, underscoring our approach's efficacy in focusing on pertinent tissue regions for accurate predictions. An example of the different heatmaps produced for the selected methods is given in Fig. 3. Interestingly, we can observe that the heatmaps derived from the middle flow scale offer more detailed segmentation of the tumor, closely aligning with the actual ground truth, as opposed to the exit flow activations that tend to focus on broader aspects.

VI. Conclusion

In this paper, we introduce a new MIL method, SparseXceptionMIL (SparseXMIL). At the core of SparseXMIL is a novel representation of WSI as multi-dimensional sparse images, coupled with a specialized sparse convolutional architecture designed to efficiently handle sparsity and spatial interactions at multiple scales. Across comprehensive experiments in various classification tasks, we have showcased the ability of SparseXMIL to enhance classification performance significantly. Through ablation studies and interpretation analyses, we have further validated the effectiveness of SparseXMIL in leveraging spatial context and focusing on pertinent information for making accurate predictions. These findings underscore the potential of our approach to not only improve upon existing MIL methodologies but also to provide deeper insights into the spatial dynamics of WSI. In future works, we aim to expand upon the foundation laid by this work in several ways. One avenue involves investigating alternative

sampling strategies beyond simple uniform distribution. By tailoring the sampling approach, we aim to sample fewer tiles while preserving spatial context, which could help unfreeze the tile embedder. Additionally, we are interested in exploring new data augmentation techniques enabled by our framework, particularly those that can introduce perturbations to the global tissue structure. Given our model's sensitivity to tissue shape, such augmentations could serve as a powerful tool for enhancing model robustness and adaptability.

REFERENCES

[1] G. Campanella *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, Aug. 2019.

[2] M. Signaevsky *et al.*, "Antemortem detection of Parkinson's disease pathology in peripheral biopsies using artificial intelligence," *Acta Neuropathologica Communications*, vol. 10, no. 1, p. 21, Feb. 2022.

[3] M. Y. Lu *et al.*, "AI-based pathology predicts origins for cancers of unknown primary," *Nature*, vol. 594, no. 7861, pp. 106–110, Jun. 2021.

[4] V. Baxi, R. Edwards, M. Montalto, and S. Saha, "Digital pathology and artificial intelligence in translational medicine and clinical practice," *Modern Pathology*, vol. 35, no. 1, pp. 23–32, Jan. 2022.

[5] I. Garberis *et al.*, "Deep learning allows assessment of risk of metastatic relapse from invasive breast cancer histological slides," *bioRxiv*, 2022.

[6] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, Jan. 1997.

[7] Z. Shao *et al.*, "TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 2136–2147.

[8] R. J. Chen *et al.*, "Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 16 123–16 134.

[9] H. Pinckaers, B. van Ginneken, and G. Litjens, "Streaming convolutional neural networks for end-to-end learning with multi-megapixel images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1581–1590, Mar. 2022, arXiv:1911.04432 [cs].

[10] T. Zhao *et al.*, "Graph data augmentation for graph machine learning: A survey," *arXiv preprint arXiv:2202.08871*, 2022.

[11] M. Lerousseau, M. Vakalopoulou, E. Deutsch, and N. Paragios, "SparseConvMIL: Sparse Convolutional Context-Aware Multiple Instance Learning for Whole Slide Image Classification," in *Proceedings of the MICCAI Workshop on Computational Pathology*. PMLR, Sep. 2021, pp. 129–139, iSSN: 2640-3498.

[12] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, Jul. 2017, pp. 1800–1807.

[13] R. C. Bunescu and R. J. Mooney, "Multiple instance learning for sparse positive bags," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 105–112.

[14] O. Maron and T. Lozano-Pérez, "A Framework for Multiple-Instance Learning," in *Advances in Neural Information Processing Systems*, vol. 10. MIT Press, 1997.

[15] O. Z. Kraus, J. L. Ba, and B. J. Frey, "Classifying and segmenting microscopy images with deep multiple instance learning," *Bioinformatics*, vol. 32, no. 12, pp. i52–i59, Jun. 2016.

[16] M. Ilse, J. Tomczak, and M. Welling, "Attention-based Deep Multiple Instance Learning," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, Jul. 2018, pp. 2127–2136, iSSN: 2640-3498.

[17] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, Jun. 2021, number: 6 Publisher: Nature Publishing Group.

[18] K. Thandiackal *et al.*, "Differentiable Zooming for Multiple Instance Learning on Whole-Slide Images," in *Computer Vision – ECCV 2022*, ser. Lecture Notes in Computer Science, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 699–715.

[19] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning," *Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Workshops*, vol. 2021, pp. 14 318–14 328, Jun. 2021.

[20] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-i.i.d. samples," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 1249–1256.

[21] D. Ahmedt-Aristizabal, M. A. Armin, S. Denman, C. Fookes, and L. Petersson, "A survey on graph-based deep learning for computational histopathology," *Computerized Medical Imaging and Graphics*, vol. 95, p. 102027, Jan. 2022.

[22] Y. Zhao *et al.*, "Predicting Lymph Node Metastasis Using Histopathological Images Based on Multiple Instance Learning With Deep Graph Convolution," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, Jun. 2020, pp. 4836–4845.

[23] R. Bazargani, L. Fazli, L. Goldenberg, M. Gleave, A. Bashashati, and S. Salcudean, "Multi-Scale Relational Graph Convolutional Network for Multiple Instance Learning in Histopathology Images," Aug. 2023, arXiv:2212.08781 [cs].

[24] M. Liang *et al.*, "Interpretable classification of pathology whole-slide images using attention based context-aware graph convolutional neural network," *Computer Methods and Programs in Biomedicine*, vol. 229, p. 107268, Feb. 2023.

[25] Y. Zhao *et al.*, "SETMIL: Spatial Encoding Transformer-Based Multiple Instance Learning for Pathological Image Analysis," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, ser. Lecture Notes in Computer Science, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 66–76.

[26] C. Xiong, H. Chen, J. Sung, and I. King, "Diagnose Like a Pathologist: Transformer-Enabled Hierarchical Attention-Guided Multiple Instance Learning for Whole Slide Image Classification," Jan. 2023, arXiv:2301.08125 [cs].

[27] Y. Huang, W. Zhao, S. Wang, Y. Fu, Y. Jiang, and L. Yu, "ConSlide: Asynchronous Hierarchical Interaction Transformer with Breakup-Reorganize Rehearsal for Continual Whole Slide Image Analysis," Aug. 2023, arXiv:2308.13324 [cs].

[28] K. Tian, Y. Jiang, Q. Diao, C. Lin, L. Wang, and Z. Yuan, "Designing BERT for Convolutional Networks: Sparse and Hierarchical Masked Modeling," in *2023 International Conference on Learning Representations*, Jan. 2023.

[29] B. Graham, "Sparse 3D convolutional neural networks," Aug. 2015, arXiv:1505.02890 [cs].

[30] C. Choy, J. Gwak, and S. Savarese, "4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Jun. 2019, pp. 3070–3079.

[31] B. Graham, M. Engelcke, and L. v. d. Maaten, "3D Semantic Segmentation with Submanifold Sparse Convolutional Networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 9224–9232, iSSN: 2575-7075.

[32] S. Sharma and S. Kumar, "The Xception model: A potential feature extractor in breast cancer histology images classification," *ICT Express*, vol. 8, no. 1, pp. 101–108, Mar. 2022.

[33] T. Lazard, M. Lerousseau, E. Decencière, and T. Walter, "Giga-ssl: Self-supervised learning for gigapixel images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023, pp. 4305–4314.

[34] J. Guinebretière, E. Menet, A. Tardivon, P. Cherel, and D. Vanel, "Normal and pathological breast, the histological basis," *European Journal of Radiology*, vol. 54, no. 1, pp. 6–14, 2005, breast Imaging: Radiohistological Correlation.

[35] A. Tarkhan, T. K. Nguyen, N. Simon, and J. Dai, "Investigation of Training Multiple Instance Learning Networks with Instance Sampling," in *Resource-Efficient Medical Image Analysis*, ser. Lecture Notes in Computer Science, X. Xu, X. Li, D. Mahapatra, L. Cheng, C. Petitjean, and H. Fu, Eds. Cham: Springer Nature Switzerland, 2022, pp. 95–104.

[36] Z. Gao *et al.*, "A semi-supervised multi-task learning framework for cancer classification with weak annotation in whole-slide images," *Medical Image Analysis*, vol. 83, p. 102652, 2023.