Article

NMGrad: Advancing Histopathological Bladder Cancer Grading with Weakly Supervised Deep Learning

Saul Fuster [1,*], Umay Kiraz [2,3], Trygve Eftestøl [1], Emiel A. M. Janssen [2,3] and Kjersti Engan [1]

[1] Department of Electrical Engineering and Computer Science, University of Stavanger, 4021 Stavanger, Norway; trygve.eftestol@uis.no (T.E.); kjersti.engan@uis.no (K.E.)
[2] Department of Pathology, Stavanger University Hospital, 4011 Stavanger, Norway; umay.kiraz@sus.no (U.K.); emilius.adrianus.maria.janssen@sus.no (E.A.M.J.)
[3] Department of Chemistry, University of Stavanger, 4021 Stavanger, Norway
* Correspondence: saul.fusternavarro@uis.no

Abstract: The most prevalent form of bladder cancer is urothelial carcinoma, characterized by a high recurrence rate and substantial lifetime treatment costs for patients. Grading is a prime factor for patient risk stratification, although it suffers from inconsistencies and variations among pathologists. Moreover, absence of annotations in medical imaging renders it difficult to train deep learning models. To address these challenges, we introduce a pipeline designed for bladder cancer grading using histological slides. First, it extracts urothelium tissue tiles at different magnification levels, employing a convolutional neural network for processing for feature extraction. Then, it engages in the slide-level prediction process. It employs a nested multiple-instance learning approach with attention to predict the grade. To distinguish different levels of malignancy within specific regions of the slide, we include the origins of the tiles in our analysis. The attention scores at region level are shown to correlate with verified high-grade regions, giving some explainability to the model. Clinical evaluations demonstrate that our model consistently outperforms previous state-of-the-art methods, achieving an F1 score of 0.85.

Keywords: computational pathology; deep learning; grading; multiscale; urothelial carcinoma; weakly supervised learning

Citation: Fuster, S.; Kiraz, U.; Eftestøl, T.; Janssen, E.A.M.; Engan, K. NMGrad: Advancing Histopathological Bladder Cancer Grading with Weakly Supervised Deep Learning. *Bioengineering* **2024**, *11*, 909. https://doi.org/10.3390/bioengineering11090909

Academic Editors: Brij Gupta, Nadia Nedjah and Kwok Tai Chui

Received: 31 July 2024
Revised: 3 September 2024
Accepted: 9 September 2024
Published: 11 Septemebr 2024

1. Introduction

Bladder cancer, a prevalent urological malignancy, poses significant clinical challenges, in terms of both diagnosis and prognosis [1]. Non-muscle-invasive bladder cancer (NMIBC) accounts for approximately 75% of the newly diagnosed cases of urothelial carcinoma. NMIBC is particularly known for its variable outcomes, necessitating accurate and consistent grading for optimal patient management [2]. The 2022 edition of the European Association of Urology guidelines on NMIBC recommends a stratification of patients into risk groups based on the risk of progression to muscle-invasive disease [3]. Grade, stage, and various other factors contribute to the risk. Precise risk assessment is vital in the management of NMIBC, since treatment strategies do not only rely on the presence of muscle invasion.

Grading is based on assessment of the cellular morphology abnormalities of urothelial tissue. In 2004, the WHO introduced a grading classification system (WHO04) for NMIBC, based on histological features. WHO04 encompasses three categories: papillary urothelial neoplasm of low malignant potential (PUNLMP); non-invasive papillary carcinoma low-grade (LG); and non-invasive papillary carcinoma high-grade (HG), ranging from lower to higher malignancy, respectively [4]. HG is related to lower differentiation, loss of polarity, and pleomorphic nuclei, among others. The intricate evaluation of heterogeneous scenarios contributes to significant inter- and intra-observer variability. Disparities potentially lead to misclassification and, consequently, to inappropriate treatment decisions [5].

The WHO04 grading system was subsequently retained in the updated 2016/2022 WHO classifications [6]. However, according to several multi-institutional analyses of individual patient data, the proportion of tumors classified as PUNLMP (WHO 2004/2016) has markedly decreased to very low levels in the last decade. This trend has resulted in the suggestion to reassess PUNLMP tumors as LG [7,8]. Therefore, the upcoming grading system will reasonably undergo a modification, shifting towards the inclusion of only LG and HG categories. In recent years, the integration of deep learning techniques within the field of computational pathology (CPATH) has offered promising avenues for enhancing the precision of computer-aided diagnosis (CAD) systems and elucidating discrepancies among pathologists [9,10]. Consequently, the synergy between histological expertise and CAD technologies is vital for accurate grading assessments.

CPATH is the field of pathology that leverages the potential of CAD systems to thoroughly analyze high-resolution digital images known as whole slide images (WSIs) for diverse diagnostic and prognostic purposes [11]. WSIs are produced by slide scanners and pre-stored at different magnification levels, emulating the functionality of physical microscopes. Lower magnification is suitable for tissue-level morphology examination, while higher magnification is suitable for cell-level scrutiny [12]. WSIs are characterized by their substantial size, which can introduce adversarial noise. Bladder cancer WSIs present unique challenges, due to their disorganized nature and the presence of diagnostically non-relevant tissue. These slides often include artifacts, such as cauterized or stretched tissue [13]. Moreover, tissue such as blood, muscle, and stroma are less informative for grading a tumor. Therefore, the absence of annotations presents a significant challenge for identifying regions of interest (ROIs) [9]. It is crucial to distinguish between region-based labels (e.g., tissue type, grade) and WSI-based labels, including follow-up information and overall patient grade. Grade exemplifies a label that encompasses both perspectives [14,15]. While clinical reports assign the worst grade observed to a patient, a WSI may exhibit diverse urothelium regions with normal, LG, and HG. This dual nature underscores the complexity of label interpretation when considering both the medical, WSI-focused perspective and the more technical perspective involving regional data analysis and processing.

CPATH is in a transformative era, aiming to reshape the landscape of digital pathology as we know it [16]. Among the diverse practices in the field, imaging methodologies rooted in convolutional neural networks (CNNs) have emerged as the foundation of feature extraction from histological images [11]. These deep learning networks possess a remarkable capacity to automatically discern morphological and cellular patterns within WSIs [17–22]. Ultimately, CNNs contribute to more precise and timely clinical decisions. However, training deep learning models in CPATH presents challenges when only WSI-level labels are available, lacking region-based annotations [9,23–26]. To address these, weakly supervised learning techniques, like attention-based methods and multiple-instance learning (MIL), are employed. However, MIL methods can be susceptible to individual instances dominating the weighted aggregation of the WSI representation [27–29]. In the context of WSIs, the tissue is distributed across the slide, for which reason the regions typically present similar features and pathologists are able to pinpoint ROIs with crucial information [30]. Specifically, while grading, situations may arise where multiple instances in close proximity exhibit HG characteristics, while other regions may concurrently display LG attributes. As a result, constraining instances to specific regions enhances our understanding of the diverse features within WSIs. Consequently, a conventional MIL architecture approach may not be appropriate, because there is a susceptibility to information leakage between regions. A model accommodating the nested structure of WSIs—wherein tissues are part of a region, and regions, in turn, belong to a WSI—may more effectively capture the clinical WSI-level grading label [31–33].

In this study, we introduce a novel pipeline for grading NMIBC using histological slides, referred to as nested multiple grading (NMGrad). The proposed solution starts by tissue segmentation of the WSI, separating urothelium from other tissue types. The next step categorizes extracted tiles of urothelium areas into location-dependent regions for predicting the patient's WHO04 grade. We implemented a weakly supervised learning framework, using attention mechanisms and a nested aggregation architecture for ROI differentiation. Our method offers an innovative approach for generating diagnostic suggestions, with generated heatmaps for highlighting tiles and ROIs independently.

2. Related Work

Numerous studies in the domain of computational pathology for bladder cancer diagnostics have emerged in recent years [34]. In Wetteland et al. [35], a pipeline for grading NMIBC was introduced. This pipeline identifies relevant areas in the WSIs and predicts the cancer grade by considering individual tile predictions and applying a decision threshold to determine the overall patient prediction. Their results demonstrated promising performance, with potential benefits for patient care. In Zheng et al. [36], the authors focused on the development of deep learning-based models for bladder cancer diagnosis and predicting overall survival in muscle-invasive bladder cancer patients. They introduced two deep learning models for diagnosis and prognosis, respectively. They showed that their presented algorithm outperformed junior pathologists. In Jansen et al. [37], the authors proposed a fully automated detection-and-grading network based on deep learning, to enhance NMIBC grading reproducibility. The study employed a U-Net-based segmentation network to automatically detect urothelium, followed by a VGG16 CNN network for classification. Their findings demonstrated that the automated classification achieved moderate agreement with consensus and individual grading from a group of three senior uropathologists. Spyridonos et al. [38] investigated the effectiveness of support vector machines and probabilistic neural networks for urinary bladder tumor grading. The results indicated that both SVM and PNN models achieve a relatively high overall accuracy, with nuclear size and chromatin cluster patterns playing key roles in optimizing classification performance.

Zhang et al. [39] addressed a common limitation of interpretability in CAD methods. To tackle this, they introduced MDNet, a novel approach that established a direct multimodal mapping between medical images and diagnostic reports. This framework consists of an image model and a language model. Through experiments on pathology bladder cancer images and diagnostic reports, MDNet demonstrated superior performance compared to comparative baselines. Zhang et al. [40] proposed a method that leverages deep learning to automate the diagnostic reasoning process through interpretable predictions. Using a dataset of NMIBC WSIs, the study demonstrated that their method achieves diagnostic performance comparable to that of 17 uropathologists.

Two critical challenges we have identified include summarizing information from local image features into a WSI representation and the scarcity of annotated datasets. Effectively translating detailed local information to the WSI level is complex, particularly in tasks like grading NMIBC. Moreover, the limited availability of well-annotated datasets hinders the development and evaluation of robust models. To tackle these issues, weakly supervised methods have emerged as a standout tool in CPATH [9]. While weakly supervised methods are widespread, some studies still rely on annotations and supervised learning. However, there is a growing consensus for the future of CPATH to predominantly embrace weakly supervised approaches. This shift is being driven by the impracticality of obtaining detailed annotations for large datasets covering various cancers and tasks. Among the various weakly supervised methods, attention-based MIL (AbMIL), a popular instance-aggregation method, exploits attention mechanisms, in order to mitigate the uncertainty from individual instances and enhance interpretability [41,42]. AbMIL bridges the gap between limited supervision and the spatial details necessary for accurate analysis and explainability. An evolution of MIL model architectures relies on the arrangement of the data within bags, where instances are further subdivided into finer groups. This concept is referred to as nesting [31,32]. Nested architectures preserve a sense of localization or categorization by selectively processing data instances within individual subgroups. Subsequently, they aggregate summarized information from the subgroups into a final bag representation.

In our work, we aimed to bridge the gap between non-annotated datasets, weakly supervised methods, and the intrinsic categorization of WSI data. Therefore, we leveraged the nested MIL with the attention mechanisms (NMIA) model architecture that we proposed in Fuster et al. [33], for accurate and interpretable NMIBC grading. Finally, in order to overcome the lack of annotations for defining the tissue of interest, a tissue segmentation algorithm TRI-25× -100× -400× was proposed by our research group in [43]. More recent works from our research group on tissue segmentation have found adoption within the scientific literature [44,45]. The utilization of this segmentation algorithm offers the opportunity to extract tiles specifically from the urothelium, contributing to a refined and targeted extraction process.

3. Data Material

The dataset comprised a total of 300 digital whole-slide images (WSIs) derived from 300 patients diagnosed with NMIBC, from the Department of Pathology, Stavanger University Hospital (SUH) [14,46]. The glass slides were digitized using a Leica SCN400 slide scanner and saved in the vendor-specific SCN file format. Collected over the period spanning 2002 to 2011, this dataset encompassed all risk group cases of non-muscle-invasive bladder cancers. The biopsies were processed through formalin fixation and paraffin embedding, and, subsequently, 4 µm thick sections were prepared and stained using hematoxylin, eosin, and saffron (HES). Furthermore, all WSIs underwent meticulous manual quality checks, ensuring the inclusion of only high-quality slides with minimal or no blur. Due to the cauterization process used in the removal of NMIBC, some slides exhibited areas with burned and damaged tissues. All WSIs originated from the same laboratory, resulting in relatively consistent staining color across the dataset.

All WSIs were graded by an expert uropathologist, in accordance with the WHO04 classification system, as either LG or HG, thus providing slide-level diagnostic information. However, the dataset lacked region-based annotations pinpointing the precise areas of LG or HG regions within the WSI. Consequently, the dataset was considered weakly labeled. For WSIs labeled as LG, at least one LG region was expected, with the possibility of presenting non-cancerous tissue in other regions. As for HG slides, at least one region should display HG tissue, while other regions may exhibit an LG appearance or non-cancerous tissue. Given the absence of alternative gold standards, we were compelled to continue utilizing a grading assessment that might have limitations for training and evaluating our algorithms. The dataset employed in this study was divided into three subsets: 220 WSI/patients for training, 30 for validation, and 50 for testing. The split employed ensured that each subset maintained the same proportional representation of diagnostic outcomes. This stratification encompassed factors such as WHO04 grading, cancer stage, recurrence, and disease progression, to best mirror the diversity of the original data material. The distribution of LG and HG WSIs within each dataset is detailed in Table 1, for reference.

Within a subset of the test set, denoted as $\text{Test}_{\text{ANNO}} \in$ Test, 14 WSIs contained either one or two annotated regions of confirmed LG or HG tissue, verified by an expert uropathologist. It is noteworthy that not all regions were annotated. The labels of these regions corresponded to the associated weak label of the WSI.

Table 1. Overview of the distribution of WSIs within each set, in terms of the WHO04 grading system [1].

Subset	Low-Grade	High-Grade
Train	124 (0)	96 (0)
Validation	17 (0)	13 (0)
Test	28 (7)	22 (7)

[1] The number between parenthesis corresponds to slides containing some annotations giving region-based labels.

4. Methods

We propose NMGrad, a pipeline that begins with a tissue segmentation algorithm for extracting urothelium tissue. Subsequently, the urothelium is divided into localized regions. Thereafter, we employ a weakly supervised learning method to predict tumor grade from the segmented urothelium regions. We exploit the sense of region locations by adopting a nested architecture with attention, NMIA [33]. The rationale behind employing this structured data arrangement analysis is to identify relevant instances and regions within the WSI. The attention mechanism and the nested bags/regions also contribute to a more precise and insightful analysis of the data. An overview of NMGrad is visualized in Figure 1:

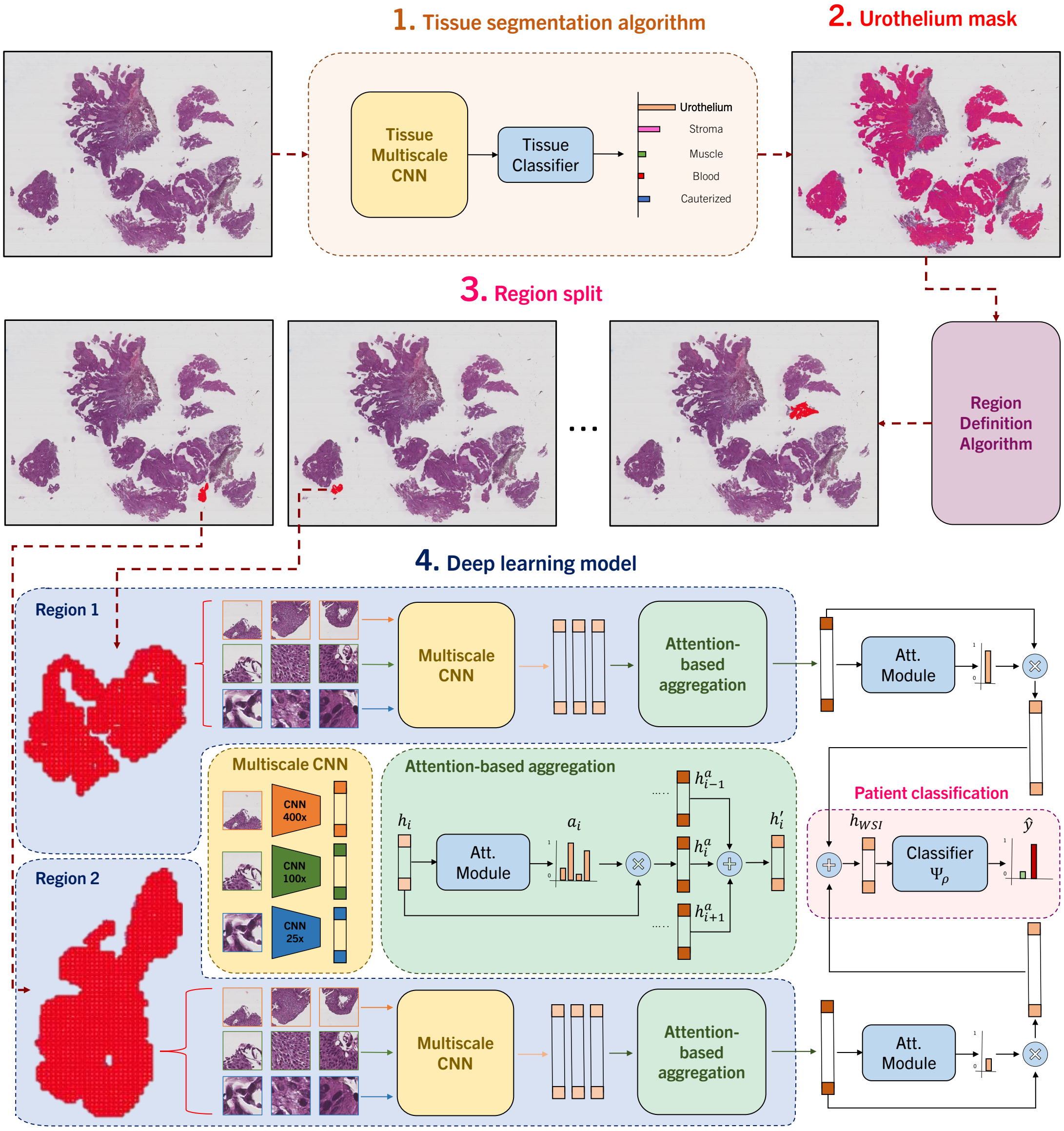

Figure 1. NMGrad pipeline. Initially, we apply a tissue segmentation algorithm for ROI extraction. Then, we pinpoint diagnostically significant urothelium areas within WSIs. Subsequently, we split the urothelium mask into regions, based on proximity and size, and extract tile triplets. In a hierarchical fashion, we further transform these triplets within their corresponding regions into region feature embeddings, using an attention-based aggregation method. All the region representations are then consolidated into a comprehensive WSI-level representation through a weight-independent attention module. Finally, this WSI feature embedding is input into the WHO04 grading classifier, in order to produce accurate WSI grade predictions.

4.1. Automatic Tissue Segmentation and Region Definition

We utilize the tissue segmentation algorithm introduced by Wetteland et al. [43] to automatically generate tissue type masks, facilitating the subsequent extraction of tiles. We define triplets $\mathcal{T}$ of tissue, which consist of a set of three tiles at various magnification levels, namely 25×, 100×, and 400×. An example is shown in Figure 2. We use a tile size of 128 × 128 for all magnification levels. Triplets are formed to maintain consistency, ensuring that the center pixel in every tile accurately represents the same physical point. The tissue segmentation algorithm works at tile level and classifies all triplets $\mathcal{T}$ in the WSI as $y \in \mathcal{Y} = \{$*urothelium, lamina propria, muscle, blood, damage, background*$\}$. As grading relies on *urothelium* alone, we utilize the *urothelium* mask for defining valid areas for tile extraction, as described in [47]. In this work, various magnifications are explored for defining the model's input, either using mono-scale MONO (400×), di-scale DI (400×, 100×) or tri-scale TRI (400×, 100×, 25×). We employ 400× magnification to establish a tight grid of tiles for data extraction purposes. These sets of tiles are later fed to the grading models, where each magnification tile is processed by its respective weight-independent CNN. To preserve the sense of location within the image, we define regions. This results in the following stratified division of data: all urothelium in the WSI, scattered regions of urothelium, and finally, individual tiles of urothelium.

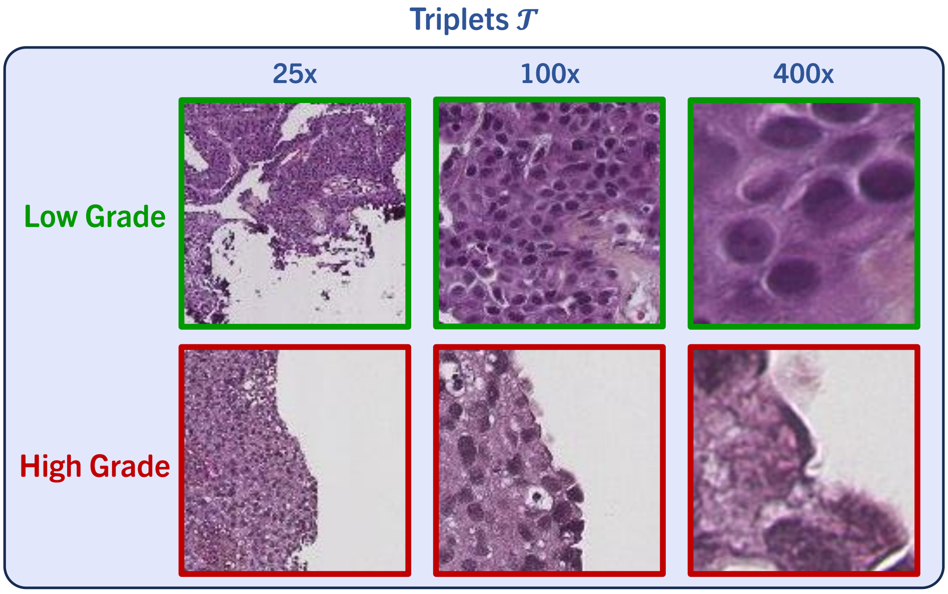

Figure 2. We obtain sets comprised of three tiles at different magnification levels, named triplets $\mathcal{T}$, enabling detailed examination. Tile triplets demonstrate regions associated with low- and high-grade features.

Region Definition

For defining regions out of the extracted urothelium tiles, we define blobs of tiles $\text{URO}_{\text{BLOB}} \subseteq \text{URO}$. URO_{BLOB} is formed when tiles are 8-connected, and this joint set of tiles is the representation of a region $\text{URO}_{\text{BLOB}} = \{\mathcal{T}_1, \mathcal{T}_2 \ldots\}$. A region is eligible for inclusion if the number of tiles N_B is higher than the threshold number T_{LOWER}. Any blob with $N_B < T_{\text{LOWER}}$ tiles is discarded, along with the tiles within. As NMIBC WSI can contain large tissue bundles, resulting in sizable blobs, we also define an upper-limit threshold T_{UPPER}. For blobs where $N_B \geq T_{\text{UPPER}}$, we split the region into several sub-regions for more detailed analysis. We apply KMeans clustering over the coordinates of the tiles x, y within the blob for location sense, defining the number of clusters as $N_C = \lceil N_B / T_{\text{UPPER}} \rceil$. This results in joint regions within a bundle of tissue of consistent size, as observed in regions 5–8 in Figure 3.

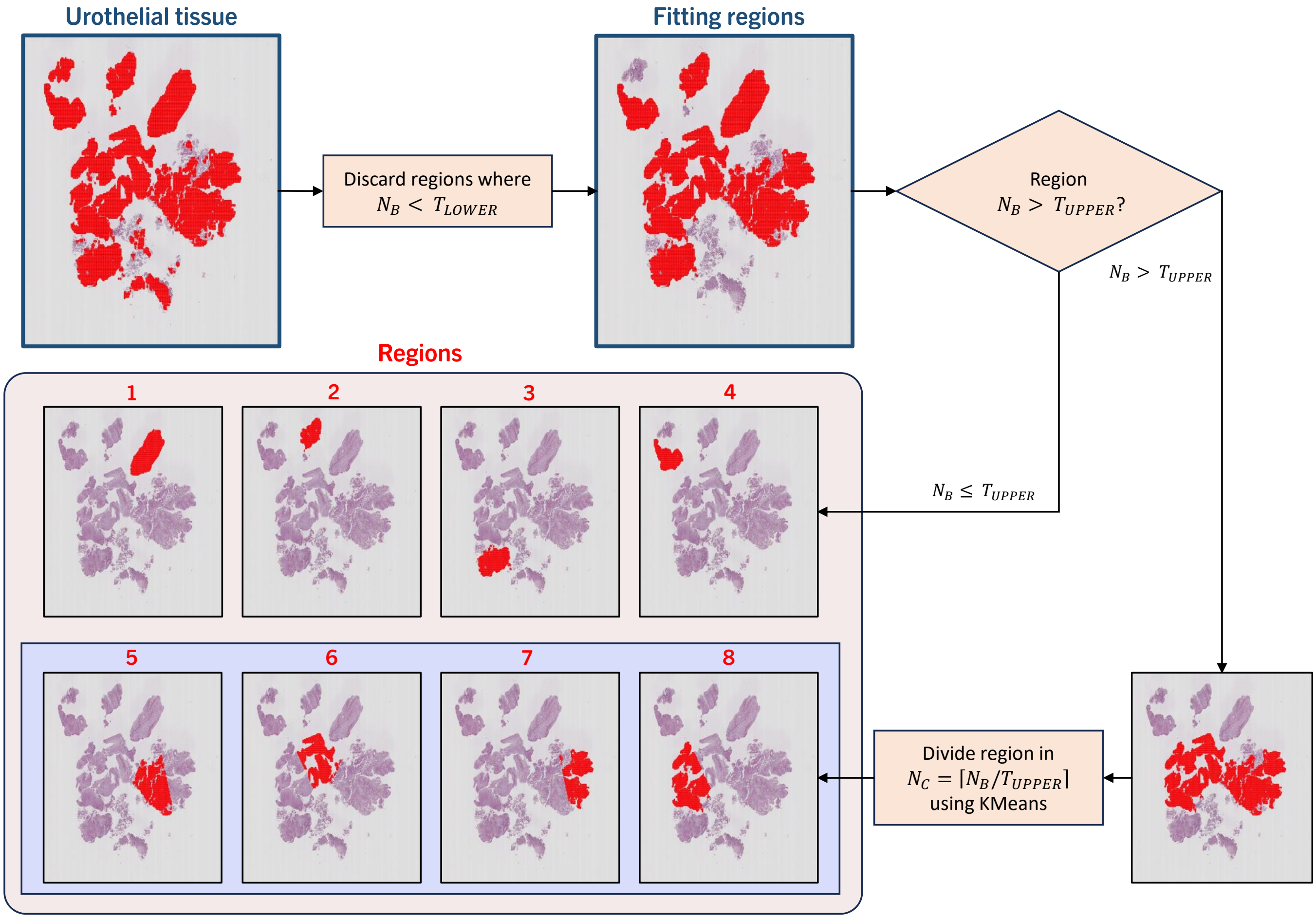
Urothelial tissue
Discard regions where $N_B < T_{LOWER}$
Fitting regions
Region $N_B > T_{UPPER}$?
$N_B > T_{UPPER}$
$N_B \leq T_{UPPER}$
Regions
1
2
3
4
5
6
7
8
Divide region in $N_C = \lceil N_B/T_{UPPER} \rceil$ using KMeans

Figure 3. Region definition. Urothelial tissue within a WSI is eligible for tile extraction. Blobs of tiles are formed, and blobs smaller than a threshold T_{LOWER} are discarded. From the remaining blobs, any smaller than T_{UPPER} are kept and defined as a region. For blobs bigger than T_{UPPER}, the blob is subdivided into smaller pieces, using the location of the individual tiles within and KMeans clustering. The obtained clusters are designated as regions.

4.2. Multiple-Instance Learning in a WSI Context

Multiple instance learning (MIL) is a weakly supervised learning method where unlabeled instances are grouped into bags with known labels [27,28]. A dataset $\mathcal{X}, \mathcal{Y} = \{(\mathbf{X}^i, y^i), \forall i = 1, \ldots, N\}$ is formed of pairs of sample sets **X** and their corresponding labels y, where i denotes a bag index. In the context of WSI, the bag can be one patient or one WSI or one region. In a conventional MIL data arrangement, we consider the bag **X** to be one WSI consisting of instances $\mathbf{x}_l$:

$$\mathbf{X} = \{\mathbf{x}_l, \forall l = 1, \ldots, L\} \tag{1}$$

where L is the number of instances in the bag. In our study, **x** refers to individual tiles in a set of extracted tiles from a patient slide **X**. A feature extractor $G_\theta : \mathcal{X} \rightarrow \mathcal{H}$ transforms image tiles, $\mathbf{x}_l$, into low-dimensional feature embeddings, $\mathbf{h}_l$. At this point, the bag structure previously formed remains intact, as instances have been simply transformed. Given a label y for a WSI, the training objective of the model is to predict the grade observed in the WSI. However, to deduce the specific region(s) within a WSI that leads to the patient's diagnosis of either LG or HG is of the utmost importance. This entails the model's ability to discern and highlight the critical areas within the WSI that play a pivotal role in the diagnosis. In order to accomplish this goal, we adopted attention-based multiple-instance learning (AbMIL) as our MIL framework, using attention-based aggregation, as shown in Figure 1. An attention score a_i for a feature embedding $\mathbf{h}_i$ can be calculated as

$$a_i = \frac{\exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{h_i}^\top) \odot \text{sigm}(\mathbf{U}\mathbf{h_i}^\top))\}}{\sum_{l=1}^{L}\exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{h_l}^\top) \odot \text{sigm}(\mathbf{U}\mathbf{h_l}^\top))\}} \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^{L\times 1}$, $\mathbf{V} \in \mathbb{R}^{L\times M}$, and $\mathbf{U} \in \mathbb{R}^{L\times M}$ are trainable parameters and $\odot$ is an element-wise multiplication. Furthermore, the hyperbolic tangent $\tanh(\cdot)$ and sigmoid $\text{sigm}(\cdot)$ are included, to introduce non-linearity for learning complex applications. The benefit of attention modules extends beyond interpretability for understanding the model's decision-making process, as it also grants enhanced predictive capabilities by prioritizing salient features. This is because attention scores directly influence the forward propagation of the model, allowing it to focus on the most relevant and informative regions within the data. Once the attention scores $\mathbf{A}$ are obtained, we obtain the patient prediction $\hat{y}$, using a patient classifier Ψ_ρ, as

$$\hat{y} = \Psi_\rho(\mathbf{H}^a) = \Psi_\rho(\mathbf{A}\cdot\mathbf{H}) = \Psi_\rho(\mathbf{A}\cdot G_\theta(\mathbf{X})) \tag{3}$$

Nested Multiple-Instance Architecture

An evolution of the conventional AbMIL architecture defines levels of bags within bags where only the innermost bags contain instances. This is referred to as nested multiple instance with attention (NMIA) [33]. A bag-of-bags for a WSI $\mathbf{H}_{\text{WSI}}$ contains a set of inner bags, or regions, $\mathbf{H}_{\text{REG},k}$:

$$\mathbf{H}_{\text{WSI}} = \left\{\mathbf{H}_{\text{REG},k}, \forall k = 1, \ldots, K\right\} \tag{4}$$

where the number of inner bags K varies between different WSI. Ultimately, $\mathbf{H}_{\text{REG},k}$ contains instance-level representations $\mathbf{h}_{\text{TILE},l}$ of tiles located within the physical region. This serves to further stratify into clusters or regions and to accurately represent the arrangement of the scattered data, where tiles belong to particular tissue areas, and the areas themselves belong to the WSI:

$$\mathbf{h}_{\text{REG}} = \mathbf{A}_{\text{TILE}}\cdot\mathbf{H}_{\text{TILE}} = \mathbf{A}_{\text{TILE}}\cdot G_\theta(\mathbf{X}) \tag{5}$$

Finally, the ultimate WSI representation h_{WSI} is fed into the classifier for obtaining the grade prediction, leveraging the region representations $\mathbf{H}_{\text{REG}}$ and the attention scores $\mathbf{A}_{\text{REG}}$ obtained from those same representations:

$$\hat{y} = \Psi_\rho(h_{\text{WSI}}) = \Psi_\rho(\mathbf{A}_{\text{REG}}\cdot\mathbf{H}_{\text{REG}}) \tag{6}$$

5. Experiments

Within the scope of our experimental investigation, we systematically assessed and compared the impact of diverse magnification levels and their combinations, namely MONO, DI, and TRI-scale models. We further investigated the impact of several weakly supervised aggregation techniques in the performance of our deep learning model. These aggregation techniques vary in their ability to distill valuable diagnostic insights from the data, namely, mean, max, and attention-based. We put special emphasis on the comparison between a standard AbMIL architecture and the nested model proposed in NMGrad. Finally, we compared our solution to current state-of-the-art methods. The code is available at https://github.com/Biomedical-Data-Analysis-Laboratory/GradeMIL (accessed on 3 September 2024).

We list the details of the design choice of the models during training. VGG16 was used as the CNN backbone, with ImageNet pre-trained weights [48]. In our preliminary experiments, we explored various architectures and found that VGG16 exhibited favorable performance across multiple tasks on NMIBC WSIs. Stochastic gradient descent (SGD) was set as the optimizer, with a learning rate of 1×10^{-4}, a batch size of 128, and a total of 200 epochs, with 30 epochs for early stopping, based on the AUC score on the validation set.

A total number of 5000 tiles per WSIs were pre-emptively randomly sampled, to be further sub-sampled during training. Focal Tversky loss (FTL) was employed [49]. The Tversky index (TI) leverages false predictions, emphasizing on recall, in case of large class imbalance, tuning parameters α and β. TI is defined as

$$\mathrm{TI}_c(\hat{y}, y) = \frac{1 + \sum_{i=1}^{N} \hat{y}_{i,c} y_{i,c} + \epsilon}{1 + \sum_{i=1}^{N} \hat{y}_{i,c} y_{i,c} + \alpha \sum_{i=1}^{N} \hat{y}_{i,\hat{c}} y_{i,c} + \beta \sum_{i=1}^{N} \hat{y}_{i,c} y_{i,\hat{c}} + \epsilon} \quad (7)$$

where $\hat{y}_{i,\hat{c}} = 1 - \hat{y}_{i,c}$ and $y_{i,\hat{c}} = 1 - y_{i,c}$ are the probability that sample i is not of class $c \in \mathcal{C}$; ϵ is used for numerical stability, preventing zero division operations. FTL employs another parameter γ for leveraging training examples hardship:

$$\mathrm{FTL}_c(\hat{y}, y) = \sum (1 - \mathrm{TI}_c(\hat{y}, y))^{1/\gamma} \quad (8)$$

6. Results and Discussion

The utilization of our proposed pipeline NMGrad using TRI-scale input emerged as a standout performer, as shown in Table 2. TRI-scale models showcased the ability to capture relevant patterns across different magnifications, substantially enhancing the overall grading accuracy. Comparing the effects of scales on the model performance shows a larger gap in performance between MONO and TRI models than structuring the processing of data using NMGrad or a standard AbMIL architecture. Furthermore, our exploration of aggregation techniques extended to mean and max aggregation methods, which did not include in-built attention mechanisms and yielded less promising outcomes. The absence of attention mechanisms rendered these techniques less effective in capturing nuanced features, underscoring the significance of attention mechanisms.

Table 2. Test performance for various aggregation techniques in weakly supervised learning. We provide the average of five runs, with the standard deviation shown in parentheses. The table presents the different approaches employed in the field, including aggregation techniques that involve considering spatial separation of the instances. We also explore the use of multiple magnification levels, considering 400× as the foundation for all magnification analysis. We also show the results from other bladder cancer grading works, although in other datasets.

Model	Accuracy	Precision	Recall	F1 Score	κ	AUC
AbMIL$_{\mathbf{MONO}}$	0.68 (0.07)	0.71 (0.07)	0.68 (0.06)	0.67 (0.07)	0.36 (0.12)	0.81 (0.07)
AbMIL$_{\mathbf{DI}}$	0.79 (0.09)	0.80 (0.10)	0.78 (0.09)	0.78 (0.09)	0.57 (0.18)	0.85 (0.13)
AbMIL$_{\mathbf{TRI}}$	0.82 (0.07)	0.82 (0.07)	0.82 (0.07)	0.82 (0.07)	0.64 (0.14)	0.91 (0.04)
MEAN$_{\mathbf{TRI}}$	0.81 (0.03)	0.83 (0.03)	0.80 (0.03)	0.80 (0.03)	0.61 (0.05)	0.92 (0.03)
MAX$_{\mathbf{TRI}}$	0.80 (0.06)	0.80 (0.06)	0.78 (0.06)	0.79 (0.07)	0.58 (0.13)	0.85 (0.06)
NMGrad$_{\mathbf{MONO}}$	0.68 (0.09)	0.71 (0.08)	0.69 (0.08)	0.68 (0.09)	0.37 (0.16)	0.80 (0.06)
NMGrad$_{\mathbf{DI}}$	0.83 (0.03)	0.85 (0.03)	0.82 (0.03)	0.82 (0.03)	0.65 (0.06)	0.91 (0.04)
NMGrad$_{\mathbf{TRI}}$	**0.86 (0.03)**	**0.87 (0.02)**	**0.85 (0.04)**	**0.85 (0.03)**	**0.71 (0.06)**	**0.94 (0.01)**
Wetteland [35]	0.90 (-)	0.87 (-)	0.80 (-)	0.83 (-)	-	-
Jansen [37]	0.74 (-)	-	0.71 (-)	-	0.48 (0.14)	-
Zhang [40]	0.95 (-)	-	-	-	-	0.95 (-)

NMIA architecture embedded in NMGrad, which employs attention mechanisms for both tile and region aggregation, marked a substantial leap in performance in comparison to mean and max aggregation. This strategy provided the strengths of attention-based aggregation and ROI localization via a nested architecture. The incorporation of attention mechanisms allowed the model to pinpoint and emphasize critical visual cues within WSIs related to urothelial cell differentiation, ultimately resulting in a notable enhancement in predictive accuracy for bladder cancer grading. Finally, in a direct comparison to the previous best-performing model proposed by Wetteland [35], we implemented weakly supervised learning in a naive manner, where all patches were assigned a weak label. Predictions were made at the patch level, and the determination of WSI-level prediction relied somewhat arbitrarily on the summation of patch predictions, neglecting consideration of localized regions. Using our proposed solution NMGrad$_{\mathrm{TRI}}$, the solution aligned more closely with clinical expectations, such as the presence of one or more regions indicative of HG if the WSI was classified as HG, rather than scattered patches. Furthermore, the capacity of NMGrad$_{\mathrm{TRI}}$ to learn attention scores offered interpretability, as opposed to relying on ad hoc rules for post-processing patches into a final prediction. Ultimately, we obtained a slightly better F1 score, with a trade-off in accuracy. We also adhered to the works of Jansen [37] and Zhang [40] for comparing the performance of state-of-the-art grading algorithms, although the results corresponding to their in-house cohorts were different from ours. The development of our deep learning model NMGrad$_{\mathrm{TRI}}$ for predicting the grading of bladder cancer represents a significant advancement in the realm of accurate grading of bladder tumors.

In order to enhance the fidelity of binary decisions, we opted to introduce an uncertainty spectrum, thereby introducing a third class. Given the output predictions of test set WSIs, we defined the uncertainty spectrum as $[\mu_{\hat{y}(y=LG)} + \sigma_{\hat{y}(y=LG)}, \mu_{\hat{y}(y=HG)} - \sigma_{\hat{y}(y=HG)}]$. A plot illustrating the concept and WSI predictions is shown in Figure 4. It was observed that if we were to exclude predictions falling within the uncertainty spectrum then the overall F1 score increased to 0.89. This underscores the potential utility of skepticism regarding non-confident predictions for robust clinical interpretation, which needs to be considered when implementing an algorithm at the inference stage.

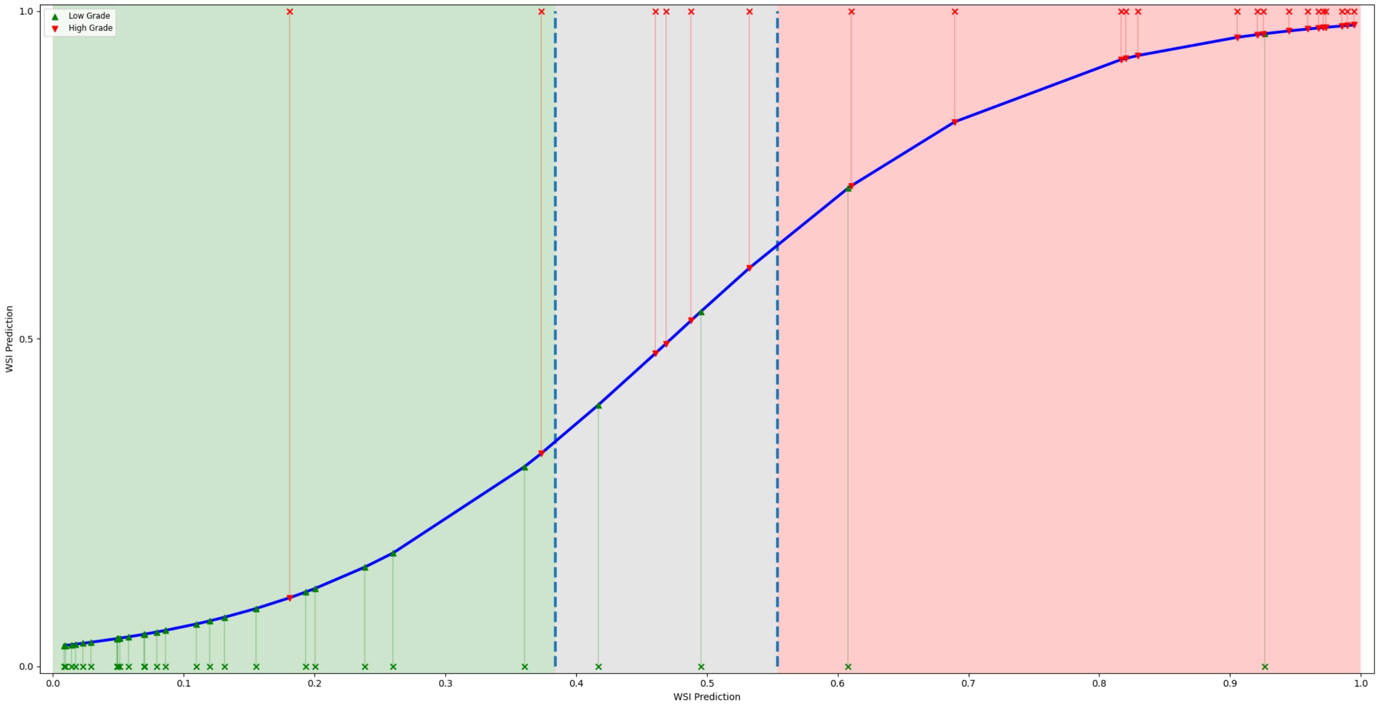

Figure 4. Plot displaying the WSI predictions of the test set, with green shading representing the LG confidence interval, red for HG, and gray denoting the uncertainty interval. Additionally, a blue line depicts the regression line fitting the predictions.

Due to the attention scores generated at the inference stage, we were able to visualize a heatmap, as shown in Figure 5. NMGrad$_{\mathrm{TRI}}$ demonstrated effectiveness in correctly assessing individual tiles and ROIs, despite being trained solely on patient-level weak labels. We consulted the generated heatmaps with experienced pathologists for qualitative analysis. The results on the annotated set of regions, Test$_{\mathrm{ANNO}}$, exhibited competence in discerning LG and HG regions, as shown in Table 3. Region attention scores were considered for direct comparison to region prediction scores utilizing the classifier Ψ_{ρ}, as the values for both scores were restrained between 0 and 1. These two scores were evaluated individually against the annotated areas. It was corroborated that higher attention scores were associated with HG areas in high-grade WSIs. However, the same did not apply for low-grade. For low-grade WSIs, we observed a wide range of possibilities, where high attention was spread across regions. Ideally, one would anticipate a direct correlation between HG and elevated attention and, conversely, a correlation between LG and reduced attention. However, this correlation was primarily observable in the positive class (HG), aligning with the inherent design of MIL, which is tailored for identifying positive instances. In contrast to attention, we observed a high degree of correspondence between the label of annotated ROIs and the output region predictions.

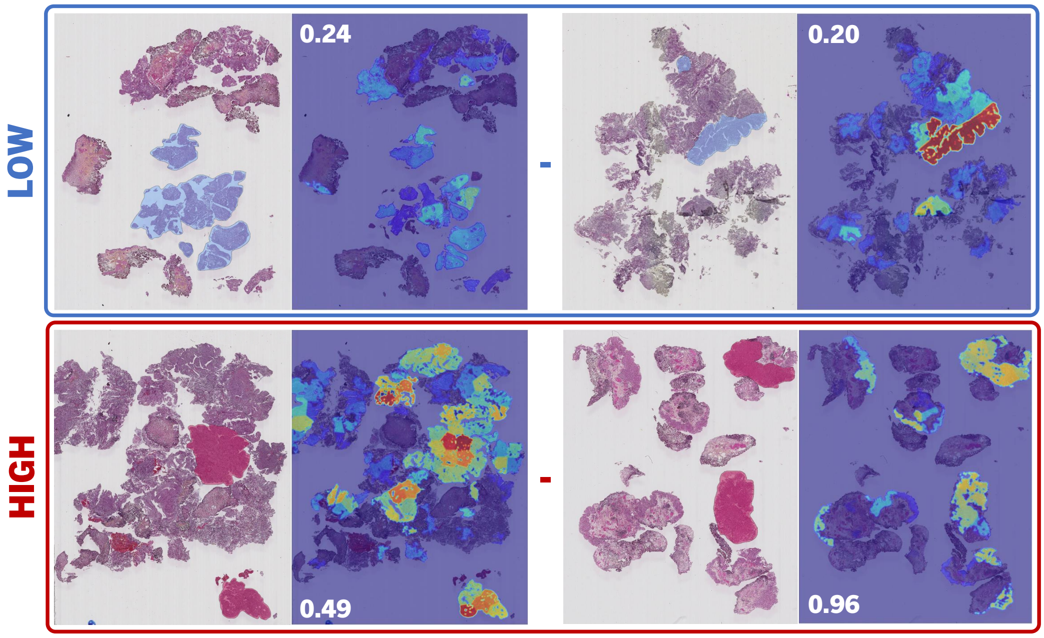

Figure 5. Region-level attention score heatmaps. Example regions of annotations of low- and high-grade ROIs annotated by a uropathologist are compared to the output attention provided by the proposed model NMGrad, left-to-right, respectively. The choice of annotated ROIs corresponded to the highest attention scores; red and blue correspond to low and high attention correspondingly. We have included the WSI-level prediction score for reference.

Table 3. Region-level prediction and attention correspondence on annotated areas, using NMGrad$_{\mathrm{TRI}}$. We individually compared the degree of consensus of annotations with both the highest attributed attention within the WSI and the output region prediction of the classifier Ψ_{ρ}.

Output	Accuracy	Precision	Recall	F1 Score
Attention	0.76	0.81	0.69	0.75
Prediction	0.89	0.83	0.91	0.87

We further investigated the correlation between region attention scores and the output region predictions, as displayed in Figure 6. We observed a generalized reciprocity between low-grade having smaller attention scores and prediction outputs, and vice versa. Moreover, the high-grade range of values was more limited to a lower range compared to the low-grade. For instance, low-grade areas with predictions rounding zero showed the broadest range of values. This observation aligns with the earlier statement, wherein the positive class typically exhibited a more focused distribution of attention scores, predominantly linked with positive HG instances. In contrast, the negative class dispersed attention across various regions within the WSI despite all presenting similar LG features. In regards to misclassified WSIs, we noted that LG WSIs manifested both high attention and prediction scores, whereas HG slides displayed a broad range of values. When examining the regression lines of TP and FP, they exhibited similar trends, as did TN and FN, respectively. Essentially, wrongly predicted WSIs exhibited characteristics contrary to their assigned class.

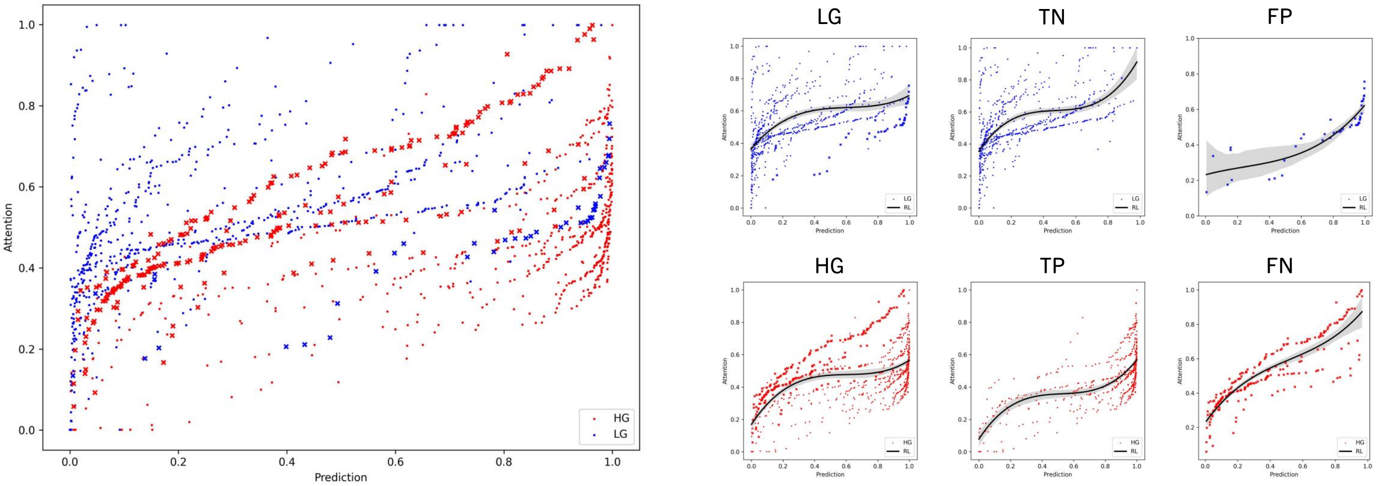

Figure 6. Correlation between region attention scores and output prediction of region embeddings on the test set of WSIs. Regions from accurately predicted WSIs (TN, TP) are denoted by squares, while those from incorrectly predicted WSIs (FN, FP) are marked with crosses. A discernible pattern emerges, where low attention scores align with diminished predictions and, conversely, higher attention scores correlate with elevated predictions. Additionally, an observable trend indicates that wrong predictions tend to manifest on the opposite end of the spectrum, with low-grade instances concentrating high attention and prediction scores, and vice versa. The trend is represented with a polynomial regression line (RL).

To augment the evaluation process, we integrated correlation calculations with follow-up information, thereby ensuring a more thorough assessment of our model's performance. We employed the Cramér's V correlation coefficient φ_c for calculating the intercorrelation between grading and the event of recurrence and progression for the test set [50]. We observed a lack of correlation between grading and recurrence for either the manual or automatic grading, aligning with our expectations. However, for progression, NMGrad exhibited a higher correlation than the uropathologist (0.32 vs. 0.26, respectively). In accordance with [51], a strong correlation between grading and progression was observed. Notably, these correlations suggest that NMGrad's grade may hold greater predictive value for assessing the likelihood of progression in the context of NMIBC.

7. Conclusions

Accurate grading of NMIBC is paramount for patient risk stratification, but it has long suffered from inconsistencies and variations among pathologists. Furthermore, the pathological workload is increasing, as well as its expenses. In response to this challenge, we have introduced the NMGrad pipeline, a pioneering approach in bladder cancer grading using WSIs. NMGrad starts by using a tissue segmentation algorithm, finding areas of urothelium in the slides. Thereafter, it leverages a nested AbMIL model architecture to precisely identify diagnostically relevant regions within WSIs and collectively predict tumor grade. Moreover, through a multiscale CNN model, NMGrad processes urothelium tissue tiles at multiple magnification levels. We observed that in high-grade patients, attention scores pinpointed specific ROIs, while in low-grade patients, attention was more dispersed, deviating from the expected MIL pattern. Our clinical evaluations demonstrated that NMGrad consistently outperformed previous state-of-the-art methods, achieving a 0.94 AUC score. This achievement represents a significant advancement in the field of bladder cancer diagnosis, with the potential to improve patient outcomes, reduce economic burdens, and enhance the quality of care in the management of this challenging disease.

Author Contributions: Conceptualization, S.F., T.E. and K.E.; methodology, S.F., T.E. and K.E.; software, S.F.; validation, S.F. and U.K.; formal analysis, S.F.; investigation, S.F.; resources, E.A.M.J.; data curation, U.K.; writing—original draft preparation, S.F.; writing—review and editing, S.F., U.K., T.E., E.A.M.J. and K.E.; visualization, S.F.; supervision, T.E. and K.E.; project administration, E.A.M.J. and K.E.; funding acquisition, E.A.M.J. and K.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research has received funding from the European Union's Horizon 2020 research and innovation program under grant agreements 860627 (CLARIFY).

Institutional Review Board Statement: This work involved human subjects or animals in its research. Ethical approval of all experimental procedures and protocols was granted by the Regional Committees for Medical and Health Research Ethics (REC) under Application No. 2011/1539. All experimental procedures and protocols were performed in line with the Norwegian Health Research Act.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data underlying this article were generated at Stavanger University Hospital. Derived data supporting the findings may be available from the corresponding author upon request.

Conflicts of Interest: The authors declare no competing interests.

References

1. Teoh, J.Y.C.; Huang, J.; Ko, W.Y.K.; Lok, V.; Choi, P.; Ng, C.F.; Sengupta, S.; Mostafid, H.; Kamat, A.M.; Black, P.C.; et al. Global trends of bladder cancer incidence and mortality, and their associations with tobacco use and gross domestic product per capita. *Eur. Urol.* **2020**, *78*, 893–906. [CrossRef]
2. Burger, M.; Catto, J.W.; Dalbagni, G.; Grossman, H.B.; Herr, H.; Karakiewicz, P.; Kassouf, W.; Kiemeney, L.A.; La Vecchia, C.; Shariat, S.; et al. Epidemiology and risk factors of urothelial bladder cancer. *Eur. Urol.* **2013**, *63*, 234–241. [CrossRef]
3. Babjuk, M.; Burger, M.; Capoun, O.; Cohen, D.; Compérat, E.M.; Escrig, J.L.D.; Gontero, P.; Liedberg, F.; Masson-Lecomte, A.; Mostafid, A.H.; et al. European Association of Urology guidelines on non–muscle-invasive bladder cancer (Ta, T1, and carcinoma in situ). *Eur. Urol.* **2022**, *81*, 75–94. [CrossRef]
4. Eble, J. World Health Organization classification of tumours. In *Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs*; IARC Press: Lyon, France, 2004; pp. 68–69.
5. Tosoni, I.; Wagner, U.; Sauter, G.; Egloff, M.; Knönagel, H.; Alund, G.; Bannwart, F.; Mihatsch, M.J.; Gasser, T.C.; Maurer, R. Clinical significance of interobserver differences in the staging and grading of superficial bladder cancer. *BJU Int.* **2000**, *85*, 48–53. [CrossRef]
6. Netto, G.J.; Amin, M.B.; Berney, D.M.; Compérat, E.M.; Gill, A.J.; Hartmann, A.; Menon, S.; Raspollini, M.R.; Rubin, M.A.; Srigley, J.R.; et al. The 2022 World Health Organization classification of tumors of the urinary system and male genital organs—Part B: Prostate and urinary tract tumors. *Eur. Urol.* **2022**, *82*, 469–482. [CrossRef]
7. Hentschel, A.E.; van Rhijn, B.W.; Bründl, J.; Compérat, E.M.; Plass, K.; Rodríguez, O.; Henríquez, J.D.S.; Hernández, V.; de la Peña, E.; Alemany, I.; et al. Papillary urothelial neoplasm of low malignant potential (PUN-LMP): Still a meaningful histo-pathological grade category for Ta, noninvasive bladder tumors in 2019? In *Urologic Oncology: Seminars and Original Investigations*; Elsevier: Amsterdam, The Netherlands, 2020; Volume 38, pp. 440–448.
8. Jones, T.D.; Cheng, L. Reappraisal of the papillary urothelial neoplasm of low malignant potential (PUNLMP). *Histopathology* **2020**, *77*, 525–535. [CrossRef]
9. Van der Laak, J.; Litjens, G.; Ciompi, F. Deep learning in histopathology: The path to the clinic. *Nat. Med.* **2021**, *27*, 775–784. [CrossRef]
10. Komura, D.; Ishikawa, S. Machine learning methods for histopathological image analysis. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 34–42. [CrossRef]
11. Cui, M.; Zhang, D.Y. Artificial intelligence and computational pathology. *Lab. Investig.* **2021**, *101*, 412–422. [CrossRef]
12. Azam, A.S.; Miligy, I.M.; Kimani, P.K.; Maqbool, H.; Hewitt, K.; Rajpoot, N.M.; Snead, D.R. Diagnostic concordance and discordance in digital pathology: A systematic review and meta-analysis. *J. Clin. Pathol.* **2020**, *74*, 448–455. [CrossRef]
13. Kanwal, N.; Pérez-Bueno, F.; Schmidt, A.; Engan, K.; Molina, R. The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: A review. *IEEE Access* **2022**, *10*, 58821–58844. [CrossRef]

14. Kvikstad, V.; Mangrud, O.M.; Gudlaugsson, E.; Dalen, I.; Espeland, H.; Baak, J.P.; Janssen, E.A. Prognostic value and reproducibility of different microscopic characteristics in the WHO grading systems for pTa and pT1 urinary bladder urothelial carcinomas. *Diagn. Pathol.* **2019**, *14*, 90. [CrossRef]
15. van der Kwast, T.; Liedberg, F.; Black, P.C.; Kamat, A.; van Rhijn, B.W.; Algaba, F.; Berman, D.M.; Hartmann, A.; Lopez-Beltran, A.; Samaratunga, H.; et al. International Society of Urological Pathology expert opinion on grading of urothelial carcinoma. *Eur. Urol. Focus* **2022**, *8*, 438–446. [CrossRef] [PubMed]
16. Berbís, M.A.; McClintock, D.S.; Bychkov, A.; Van der Laak, J.; Pantanowitz, L.; Lennerz, J.K.; Cheng, J.Y.; Delahunt, B.; Egevad, L.; Eloy, C.; et al. Computational pathology in 2030: A Delphi study forecasting the role of AI in pathology within the next decade. *EBioMedicine* **2023**, *88*, 104427. [CrossRef]
17. Ciompi, F.; Geessink, O.; Bejnordi, B.E.; De Souza, G.S.; Baidoshvili, A.; Litjens, G.; Van Ginneken, B.; Nagtegaal, I.; Van Der Laak, J. The importance of stain normalization in colorectal tissue classification with convolutional networks. In Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI), Melbourne, Australia, 18–21 April 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 160–163.
18. Fuster, S.; Khoraminia, F.; Kiraz, U.; Kanwal, N.; Kvikstad, V.; Eftestøl, T.; Zuiverloon, T.C.; Janssen, E.A.; Engan, K. Invasive cancerous area detection in Non-Muscle invasive bladder cancer whole slide images. In Proceedings of the 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), Nafplio, Greece, 26–29 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–5.
19. Cruz-Roa, A.; Gilmore, H.; Basavanhally, A.; Feldman, M.; Ganesan, S.; Shih, N.; Tomaszewski, J.; Madabhushi, A.; González, F. High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection. *PLoS ONE* **2018**, *13*, e0196828. [CrossRef]
20. Litjens, G.; Sánchez, C.I.; Timofeeva, N.; Hermsen, M.; Nagtegaal, I.; Kovacs, I.; Hulsbergen-Van De Kaa, C.; Bult, P.; Van Ginneken, B.; Van Der Laak, J. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* **2016**, *6*, 26286. [CrossRef]
21. Reis, S.; Gazinska, P.; Hipwell, J.H.; Mertzanidou, T.; Naidoo, K.; Williams, N.; Pinder, S.; Hawkes, D.J. Automated classification of breast cancer stroma maturity from histological images. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 2344–2352. [CrossRef] [PubMed]
22. Dundar, M.M.; Badve, S.; Bilgin, G.; Raykar, V.; Jain, R.; Sertel, O.; Gurcan, M.N. Computerized classification of intraductal breast lesions using histopathological images. *IEEE Trans. Biomed. Eng.* **2011**, *58*, 1977–1984. [CrossRef]
23. Andreassen, C.; Fuster, S.; Hardardottir, H.; Janssen, E.A.; Engan, K. Deep Learning for Predicting Metastasis on Melanoma WSIs. In Proceedings of the 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Cartagena, Colombia, 18–21 April 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
24. Tabatabaei, Z.; Colomer, A.; Moll, J.O.; Naranjo, V. Toward More Transparent and Accurate Cancer Diagnosis With an Unsupervised CAE Approach. *IEEE Access* **2023**, *11*, 143387–143401. [CrossRef]
25. Deng, R.; Liu, Q.; Cui, C.; Yao, T.; Long, J.; Asad, Z.; Womick, R.M.; Zhu, Z.; Fogo, A.B.; Zhao, S.; et al. Omni-seg: A scale-aware dynamic network for renal pathological image segmentation. *IEEE Trans. Biomed. Eng.* **2023**, *70*, 2636–2644. [CrossRef]
26. Jiao, P.; Zheng, Q.; Yang, R.; Ni, X.; Wu, J.; Chen, Z.; Liu, X. Prediction of HER2 Status Based on Deep Learning in H&E-Stained Histopathology Images of Bladder Cancer. *Biomedicines* **2024**, *12*, 1583. [CrossRef] [PubMed]
27. Dietterich, T.G.; Lathrop, R.H.; Lozano-Pérez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* **1997**, *89*, 31–71. [CrossRef]
28. Maron, O.; Lozano-Pérez, T. A framework for multiple-instance learning. *Adv. Neural Inf. Process. Syst.* **1997**, *10*, 570–576.
29. Carbonneau, M.A.; Cheplygina, V.; Granger, E.; Gagnon, G. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit.* **2018**, *77*, 329–353. [CrossRef]
30. Mercan, E.; Aksoy, S.; Shapiro, L.G.; Weaver, D.L.; Brunyé, T.T.; Elmore, J.G. Localization of diagnostically relevant regions of interest in whole slide images: A comparative study. *J. Digit. Imaging* **2016**, *29*, 496–506. [CrossRef]
31. Tibo, A.; Frasconi, P.; Jaeger, M. A network architecture for multi-multi-instance learning. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference (ECML PKDD), Skopje, Macedonia, 18–22 September 2017; Springer: Cham, Switzerland, 2017; pp. 737–752.
32. Tibo, A.; Jaeger, M.; Frasconi, P. Learning and interpreting multi-multi-instance learning networks. *J. Mach. Learn. Res.* **2020**, *21*, 7890–7949.
33. Fuster, S.; Eftestøl, T.; Engan, K. Nested multiple instance learning with attention mechanisms. In Proceedings of the 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), Nassau, Bahamas, 12–14 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 220–225.
34. Khoraminia, F.; Fuster, S.; Kanwal, N.; Olislagers, M.; Engan, K.; van Leenders, G.J.; Stubbs, A.P.; Akram, F.; Zuiverloon, T.C. Artificial Intelligence in Digital Pathology for Bladder Cancer: Hype or Hope? A Systematic Review. *Cancers* **2023**, *15*, 4518. [CrossRef]
35. Wetteland, R.; Kvikstad, V.; Eftestøl, T.; Tøssebro, E.; Lillesand, M.; Janssen, E.A.; Engan, K. Automatic diagnostic tool for predicting cancer grade in bladder cancer patients using deep learning. *IEEE Access* **2021**, *9*, 115813–115825. [CrossRef]
36. Zheng, Q.; Yang, R.; Ni, X.; Yang, S.; Xiong, L.; Yan, D.; Xia, L.; Yuan, J.; Wang, J.; Jiao, P.; et al. Accurate Diagnosis and Survival Prediction of Bladder Cancer Using Deep Learning on Histological Slides. *Cancers* **2022**, *14*, 5807. [CrossRef]

37. Jansen, I.; Lucas, M.; Bosschieter, J.; de Boer, O.J.; Meijer, S.L.; van Leeuwen, T.G.; Marquering, H.A.; Nieuwenhuijzen, J.A.; de Bruin, D.M.; Savci-Heijink, C.D. Automated detection and grading of non–muscle-invasive urothelial cell carcinoma of the bladder. *Am. J. Pathol.* **2020**, *190*, 1483–1490. [CrossRef]
38. Spyridonos, P.; Petalas, P.; Glotsos, D.; Cavouras, D.; Ravazoula, P.; Nikiforidis, G. Comparative evaluation of support vector machines and probabilistic neural networks in superficial bladder cancer classification. *J. Comput. Methods Sci. Eng.* **2006**, *6*, 283–292.
39. Zhang, Z.; Xie, Y.; Xing, F.; McGough, M.; Yang, L. Mdnet: A semantically and visually interpretable medical image diagnosis network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6428–6436.
40. Zhang, Z.; Chen, P.; McGough, M.; Xing, F.; Wang, C.; Bui, M.; Xie, Y.; Sapkota, M.; Cui, L.; Dhillon, J.; et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nat. Mach. Intell.* **2019**, *1*, 236–245. [CrossRef]
41. Ilse, M.; Tomczak, J.; Welling, M. Attention-based deep multiple instance learning. In Proceedings of the International Conference on Machine Learning (PMLR), Stockholm, Sweden, 10–15 July 2018; pp. 2127–2136.
42. Campanella, G.; Hanna, M.G.; Geneslaw, L.; Miraflor, A.; Werneck Krauss Silva, V.; Busam, K.J.; Brogi, E.; Reuter, V.E.; Klimstra, D.S.; Fuchs, T.J. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **2019**, *25*, 1301–1309. [CrossRef] [PubMed]
43. Wetteland, R.; Engan, K.; Eftestøl, T.; Kvikstad, V.; Janssen, E.A. A multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides. *Technol. Cancer Res. Treat.* **2020**, *19*, 1533033820946787. [CrossRef]
44. Dalheim, O.N.; Wetteland, R.; Kvikstad, V.; Janssen, E.A.M.; Engan, K. Semi-supervised Tissue Segmentation of Histological Images. In Proceedings of the Colour and Visual Computing Symposium, Gjøvik, Norway, 16–17 September 2020; Volume 2688, pp. 1–15.
45. Fuster, S.; Khoraminia, F.; Eftestøl, T.; Zuiverloon, T.C.; Engan, K. Active Learning Based Domain Adaptation for Tissue Segmentation of Histopathological Images. In Proceedings of the 2023 31st European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 4–8 September 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1045–1049.
46. Kvikstad, V. Better Prognostic Markers for Nonmuscle Invasive Papillary Urothelial Carcinomas. Ph.D. Thesis, Universitetet i Stavanger, Stavanger, Norway, 2022.
47. Wetteland, R.; Engan, K.; Eftestøl, T. Parameterized extraction of tiles in multilevel gigapixel images. In Proceedings of the 2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA), Zagreb, Croatia, 13–15 September 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 78–83.
48. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems 25, Lake Tahoe, NV, USA, 3–6 December 2012.
49. Abraham, N.; Khan, N.M. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI), Venice, Italy, 8–11 April 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 683–687.
50. Bergsma, W. A bias-correction for Cramér's V and Tschuprow's T. *J. Korean Stat. Soc.* **2013**, *42*, 323–328. [CrossRef]
51. Akoglu, H. User's guide to correlation coefficients. *Turk. J. Emerg. Med.* **2018**, *18*, 91–93. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.