IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 19, 2021, accepted August 3, 2021, date of publication August 13, 2021, date of current version August 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3104724

Automatic Diagnostic Tool for Predicting Cancer Grade in Bladder Cancer Patients Using Deep Learning

RUNE WETTELAND[1], VEBJØRN KVIKSTAD[2,3], TRYGVE EFTESTØL[1], (Senior Member, IEEE), ERLEND TØSSEBRO[1], MELINDA LILLESAND[2], EMIEL A. M. JANSSEN[2,3], AND KJERSTI ENGAN[1], (Senior Member, IEEE)

[1]Department of Electrical Engineering and Computer Science, University of Stavanger, 4021 Stavanger, Norway
[2]Department of Pathology, Stavanger University Hospital, 4011 Stavanger, Norway
[3]Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, 4021 Stavanger, Norway

Corresponding author: Rune Wetteland (rune.wetteland@uis.no)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Regional Committees for Medical and Health Research Ethics (REC) under Application No. 2011/1539, and performed in line with the Norwegian Health Research Act.

ABSTRACT The most common type of bladder cancer is urothelial carcinoma, which is among the cancer types with the highest recurrence rate and lifetime treatment cost per patient. Diagnosed patients are stratified into risk groups, mainly based on grade and stage. However, it is well known that correct grading of bladder cancer suffers from intra- and interobserver variability and inconsistent reproducibility between pathologists, potentially leading to under- or overtreatment of the patients. The economic burden, unnecessary patient suffering, and additional load on the health care system illustrate the importance of developing new tools to aid pathologists. We propose a pipeline, called $\text{TRI}_{\text{grade}}$, that will identify diagnostic relevant regions in the whole-slide image (WSI) and collectively predict the grade of the current WSI. The system consists of two main models, trained on weak slide-level grade labels. First, a WSI is segmented into the different tissue classes (urothelium, stroma, muscle, blood, damaged tissue, and background). Next, tiles are extracted from the diagnostic relevant urothelium tissue from three magnification levels (25x, 100x, and 400x) and processed sequentially by a convolutional neural network (CNN) based model. Ten models were trained for the slide-level grading experiment, where the best model achieved an F1-score of 0.90 on a test set consisting of 50 WSIs. The best model was further evaluated on a smaller segmentation test set, consisting of 14 WSIs where low- and high-grade regions were annotated by a pathologist. The $\text{TRI}_{\text{grade}}$ pipeline achieved an average F1-score of 0.91 for both the low-grade and high-grade classes.

INDEX TERMS Automated cancer grading, bladder cancer, convolutional neural networks, multiscale classification, urothelial carcinoma, weakly labeled data, whole-slide image.

I. INTRODUCTION

Bladder cancer is the 10$^{\text{th}}$ most commonly diagnosed cancer disease worldwide, with 573 278 new cases in 2020 [1]. The most common type of bladder cancer is urothelial carcinoma, in which men are overrepresented. It is among the cancer types with the highest recurrence rate, approximatly 50 to 70%, which makes it especially challenging [2]. It requires an intensive treatment and follow-up plan, which results in it being one of the cancer types with the highest lifetime treatment cost per patient [3], [4]. In the case of muscle-invasive bladder cancer (MIBC), where the cancer has invaded the muscle wall of the bladder, a cystectomy is often required. However, cancers that stay confined in the bladder mucosa are referred to as non-muscle-invasive bladder cancer (NMIBC) and are easier to treat.

In histopathological diagnostics, pathologists use grading and staging to describe the tumor. These parameters are used to stratify patients into risk groups and form a suitable treatment and follow-up plan. The grade of a tumor describes the differentiation state of the tumor cells under a microscope. Different cancers have different grading scales, but in general, if the cancer cells are similar to that of healthy non-cancerous cells, the grade will be low, and the cancer will have a lower likelihood of spreading. On the other hand, if the cells have a more abnormal appearance and are disorganized, the grade will be higher. In addition to the grade, tumor stage is also important and is determined by the size of the primary

The associate editor coordinating the review of this manuscript and approving it for publication was Kin Fong Lei.

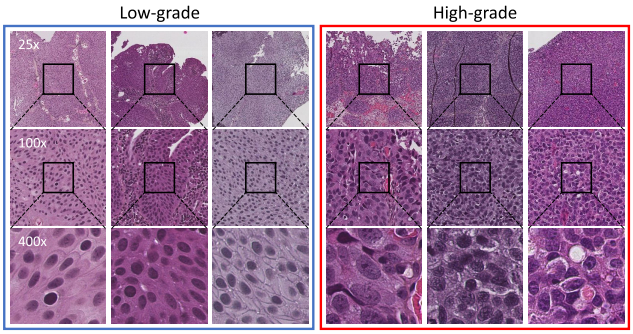

FIGURE 1. Examples of low-grade and high-grade tiles extracted from a WSI. The tiles are extracted from three magnification levels (25x, 100x, and 400x) and all have the same size of 256 × 256 pixels.

tumor, how far it has spread into the surrounding tissue, and the number of primary tumors present. In this paper, we focus on grading of NMIBC. However, it is well known that correct grading of bladder cancer suffers from intra- and interobserver variability and inconsistent reproducibility between pathologists [5], [6], which can lead to both under- or overtreatment of the patients. New tools to aid pathologists are therefore desired.

The World Health Organization (WHO) has proposed three grading systems for bladder cancer. The first grading system was introduced in 1973, referred to as WHO73, which is still somewhat used today. It consists of three categories, grade 1, grade 2, and grade 3, where grade 3 is the most severe state. A revised edition of the grading system was introduced in 2004 called WHO04, and further updated in 2016 as WHO16. In these versions, cases are split into low- and high-grade carcinoma. Some examples of low- and high-grade areas are shown in Fig. 1. Grade 1 patients are referred to as low-grade patients, and grade 3 patients are high-grade patients. Patients diagnosed as grade 2, however, are now split into either the low- or high-grade case. This might seem like a minor change, but for a patient to be diagnosed as low- or high-grade may result in very different follow-up regimes and local treatment with potential adverse events. A patient falsely diagnosed as a high-risk patient is an example of unnecessary patient suffering by overtreatment, additional load on the health care system, and extra cost. The data material used in this paper was collected and diagnosed prior to 2016 and will therefore focus on the WHO04 grading system.

After the tumor is removed, it is placed on an object glass and stained before a pathologist examines it. This is usually done through a microscope; however, with the introduction of digital pathology, digital versions of the stained specimen are also available in the form of whole-slide images (WSI). This has multiple advantages, such as remote access, storage and sharing cases between institutes, cloud computing, improved workflow, as well as computational pathology, which enables the use of new tools to process and interpret the tissue samples. All of which can improve the diagnostic accuracy and the clinical outcome of the patients [7]–[11].

Recent years have seen a rapid increase in both interest and usage of machine learning applications. Such tools could potentially be used to assist pathologists and help reduce the increasing workload. Also, because the errors made by a machine learning system may be different from that of a pathologist, the two may be combined for improved accuracy by the pathologist, as shown by Wang *et al.* [12]. Low reproducibility and variability in interpretations may also be reduced if a trustworthy computer-aided diagnosis (CAD) system could be implemented in a clinical setting.

With a CAD system, we want to map a WSI input to one of the disease output categories. The traditional machine-learning method to achieve this is by supervised learning. A set of known image and label pairs are shown to the model, which uses a gradient descent algorithm to optimize its parameters. For these algorithms to work efficiently and create robust models, a large set of image-label pairs are needed. Within digital pathology, we have access to a large amount of image data in the form of WSIs. However, annotated data is limited, challenging the practicability of supervised learning approaches. The nature of the images also calls for expert input to be able to annotate them. This is a time-consuming and, in some cases, challenging task. To create enough of the image-label pairs necessary to train these models and avoid the expensive annotation process, one possibility is to utilize data already available in the form of the slide-level diagnosis information. The WSIs are split into smaller images in the form of tiles, and the slide-level diagnosis will be assigned to each of the tiles.

For patients diagnosed with NMIBC, the tumor is usually removed through transurethral resection of bladder tumor (TURBT). During this process, parts of the tissue get damaged, either heating damage from the cauterization process or physical damage from tearing. Other tissue types, like stroma or muscle, as well as blood, are also often present in the slides of urothelial carcinoma. For the purpose of grading NMIBC, urothelium is the most diagnostic relevant tissue. For staging, both urothelium and stroma, and particularly the border between them, is essential. The presence of muscle tissue also has importance for correct staging. However, cauterized tissue from the TURBT process, as well as areas containing blood, have no diagnostic relevance. Feeding a deep learning model with these irrelevant tissue classes, e.g., blood or damaged tissue, may harm the diagnostic model's accuracy. To avoid this, we have previously proposed a method based on convolutional neural networks (CNN), which automatically segments NMIBC slides into background and five foreground classes (urothelium, stroma, muscle, blood, and damaged tissue). This tissue classification model is referred to as the $\text{TRI}_{\text{tissue}}$-model in the following and is explained in detail in Wetteland *et al.* [13].

In the current paper, we propose a system called $\text{TRI}_{\text{grade}}$ for automatically grading WSI according to the WHO04 grading system. The proposed system uses the $\text{TRI}_{\text{tissue}}$-model as a first-stage network for preprocessing the WSI to find regions of urothelium tissue. The extracted

urothelium tissue is then fed through a second-stage network called the $\text{TRI}_{\text{WHO04}}$-model for automatic grading.

The large size of the gigapixel images causes some challenges. It is not possible to feed the entire image into a deep learning algorithm; instead, tiles of a suitable size are extracted from the WSI and fed to the algorithm sequentially. The CNN-based model assigns a prediction score to every tile. These predictions are used to create a heatmap showing which regions were predicted with low- or high-grade carcinoma. The final decision can further be aggregated from the micro predictions into a slide-level prediction.

A WSI is stored in a pyramid format with multiple magnification levels, where the different levels will give different information. An example of such a pyramidal WSI is shown in Fig. 2. A pathologist will typically zoom in and out of a WSI to gather information at several scales before reaching a final decision. Our proposed method mimics this behavior by combining global context information and local details by utilizing a multiscale model architecture.

A. PREVIOUS WORK

With the introduction of digital pathology, there has been an increase in medical application research utilizing machine learning and deep learning approaches. Most research is related to cancer diseases such as breast, lung, prostate, brain, and skin cancer [14]. By looking at the list of US Food & Drugs Administration (FDA) approved artificial intelligence (AI) based medical technologies, most are in the fields of radiology, cardiology, and Internal Medicine/General Practice [15]. Still, a lot of effort is also aimed towards histological images [16]–[20].

The majority of CAD research conducted on histological images utilize two or more seperate models in their methods [16], [21]–[24]. First, a segmentation algorithm or region of interest (ROI) selection step is performed to narrow down the area which needs additional processing. This is an important step that helps in several ways. Compared to standard images, the WSIs are very large in size, and it is computationally expensive to process the entire WSI. By limiting the number of extracted tiles, the classification runtime is reduced, speeding up the classification step. Also, by removing the unwanted and diagnostically irrelevant areas, the extracted datasets will consist of higher quality tiles, which aids the classification algorithm in the following steps. After segmentation, tiles from the ROI are processed, usually by a classification model, which will predict the class of the tiles. Examples of tile classes can be cancer vs. non-cancer, recurrence vs. no recurrence, cancer grading or staging, or other classes related to cancer diagnosis. After all the selected tiles have been classified, the predictions are aggregated into a final slide-level prediction, usually using statistical or machine-learning methods.

Some research has been aimed towards urothelial carcinoma, otherwise known as bladder cancer. In Jansen *et al.* [22], they utilized two individual single-scale neural networks to detect and grade 328 cases of bladder cancer collected from 232 patients. A U-net-based segmentation network was trained to detect and segment the urothelium tissue, used as input to a second network trained to grade the urothelium tissue according to the WHO04 grading system. The classification network assessed the WHO04 grading on slide-level, using the majority vote of all classified tiles. The predictions were compared with the grading of three experienced pathologists. According to the consensus reading, the classification model achieved an accuracy score of 74%. The included whole-slide images were all exported at 20x magnification (0.5 μm per pixel).

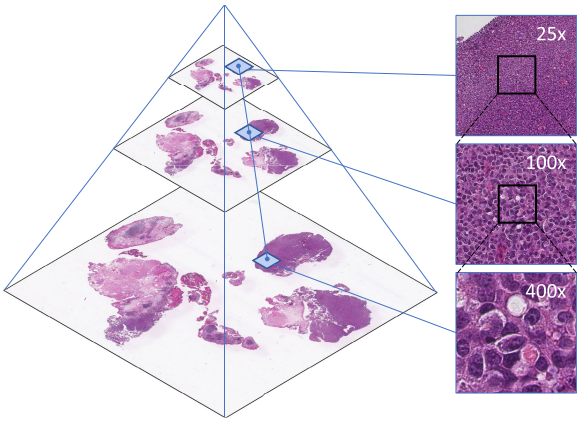

FIGURE 2. WSI images are stored in a pyramidal format, where the base image corresponds to the highest magnification level. The right-hand side shows a set of three tiles extracted so that the center of the tile corresponds to the same physical area in the WSI, forming a triplet.

From the same research group, the work of Lucas *et al.* [24] utilized the same urothelium segmentation model as presented in [22]. Regions of urothelium were then fed into a selection network which classified tiles into recurrence vs. no recurrence. A strategy was applied to select features from 200 tiles fed into a final bidirectional gated recurrent unit (GRU) classification network that predicts 1-year and 5-year recurrence-free survival (RFS) in bladder cancer patients.

The work of Zhang *et al.* [23] was also performed on bladder cancer. They used three different neural networks referred to as s-net, d-net, and a-net. The s-net model is a U-net-like architecture that classifies each pixel as tumor vs. non-tumor. The d-net then characterizes the tumor ROIs and generates an interpretable diagnosis and low-dimensional encodings. Finally, the a-net uses the ROI encodings and predicts a slide-level WHO04 grading.

Multiscale cancer subtype classification, where two or more different magnification scales are fed to the classification model, has been shown to improve the accuracy compared to single-scale models [13], [25]. This mimics the pathologist's process, which will zoom in and out to investigate the tissue area at several scales.

In Skrede *et al.* [21] the WSI is first segmented, before tiles are extracted at 10x and 40x resolution. The tiles from each scale are fed to an ensemble of 5 models, using a total of ten CNN-based models. The average score from the ensembles is used to predict the prognosis of colorectal patients.

TABLE 1. Overview of how the data material in this study is distributed into training, validation, and test sets. For triplets in the training dataset, see Table 2.

	Low-grade WSIs	High-grade WSIs	Total WSIs	Total triplets
Training	124	96	220	Table 2
Validation	17	13	30	301 775
Test	28	22	50	473 678

TABLE 2. Extracted triplets for the training dataset.

N	Total triplets before aug.	Total triplets after aug.	Percentage increase
250	54 564	55 000	0.8%
500	106 577	110 000	3.1%
1 000	202 904	219 560	7.6%
3 000	534 734	647 368	17.4%
5 000	812 588	1 051 752	22.7%

In the work of Hashimoto *et al.* [26] WSIs from malignant lymphoma were fed to a multiscale CNN-based model. They compared the results of models using tiles extracted at 10x or 20x resolution. However, the best result was achieved by combining the two scales into a multiscale model. The authors of this study also confirm that class-specific features exist at different magnification scales.

Previous work from our group, on bladder cancer, included tissue segmentation [13], [27], [28], and prediction of recurrence in NMIBC patients [29]. In Wetteland *et al.* [13], we experimented with three magnification scales and any combination of these. We proposed three MONO-models (Mono-25x, Mono-100x, and Mono-400x), three DI-models (DI-25x-100x, DI-25x-400x, and DI-100x-400x), and finally a model utilizing all three magnification scales, TRI-25x-100x-400x. All models used the VGG16 network as a feature extractor and were trained and evaluated on six tissue classes. The MONO-models performed worst, and the best result was achieved with the TRI-model utilizing all scales, supporting the claim that multiscale models achieve better results. Both frozen and unfrozen weights were experimented with, but the TRI-model trained with frozen weights in the VGG16 models performed best and achieved an average F1-score of 96.5% when evaluated on all six classes, and an average F1-score of 97.6% for the urothelium class alone.

Based on this result, we continued with the TRI-model and VGG16 as feature extractors in the current paper. We have not evaluated the MONO- or DI-models on the diagnostic data. The model referred to as TRI-25x-100x-400x in [13] is in the current paper referred to as the $\text{TRI}_{\text{tissue}}$-model. It is used for tissue extraction as shown in Fig. 4. The name, architecture, and base model have also been carried over to this paper and are the basis for the $\text{TRI}_{\text{WHO04}}$-model we propose here.

B. OUR CONTRIBUTIONS

The current study's main contributions is listed below.

- A novel, fully automated pipeline called $\text{TRI}_{\text{grade}}$ is proposed. The system consists of a tissue segmentation model and a diagnostic WHO04 grade model. The system's output consists of a tissue segmentation map, a WHO04 heatmap, and a predicted slide-level WHO04 grade. The proposed $\text{TRI}_{\text{grade}}$ system correctly predicted 45 of the 50 WSIs in the test set, achieving an accuracy of 90%.
- The $\text{TRI}_{\text{grade}}$ system-generated heatmaps are both visualized and evaluated against a segmentation test set. This helps to demonstrate the usage of such a system for a pathologist in a clinical setting.
- An algorithm for finding the optimal value of a decision threshold for classifying WSIs at slide-level is proposed.
- We trained models on differently sized training sets. The results for this provide insight on how dataset sizes affect the performance of the models, training time per epoch, and trained epochs before reaching stopping criteria during early stopping.
- Source code for this paper is accessible at the following URL address https://git.io/J3OdW.

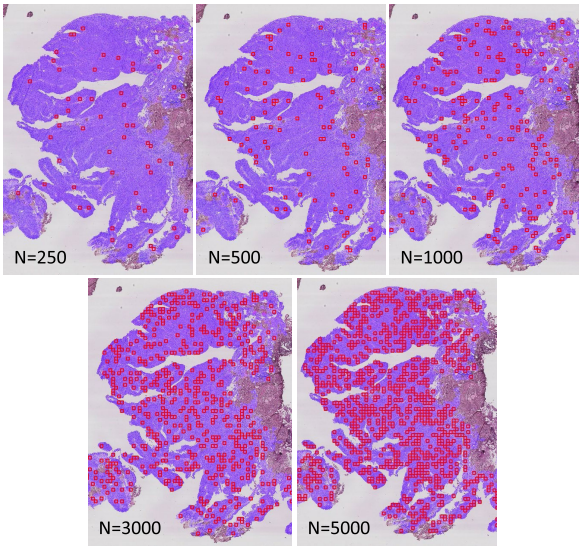

FIGURE 3. A close-up image from a WSI with a superimposed urothelium ROI mask (semi-purple). As *N* increases, the density of the tiles (red squares) also increases. The illustrated tiles are shown on 400x magnification level, but tiles from 25x and 100x are also extracted.

II. METHODS

The proposed $\text{TRI}_{\text{grade}}$ system presented in this paper utilizes multiscale models, which use tiles extracted at multiple magnification levels as input. For improved readability, we define these tiles as a *triplet*. A triplet is denoted T_i and is defined as a set of three tiles extracted from a WSI at three different magnification levels (25x, 100x, and 400x). Let $\mathcal{T}$ denote a set of triplets in a WSI, where $\mathcal{T} = \{T_1, T_2 \dots T_i \dots T_{max}\}$, and the number of elements in the set is given by the cardinality $|\mathcal{T}|$. An example of a triplet is shown in Fig. 2.

A. DATA MATERIAL

The data material consists of 300 digital whole-slide images from patients diagnosed with NMIBC, where the tissue is

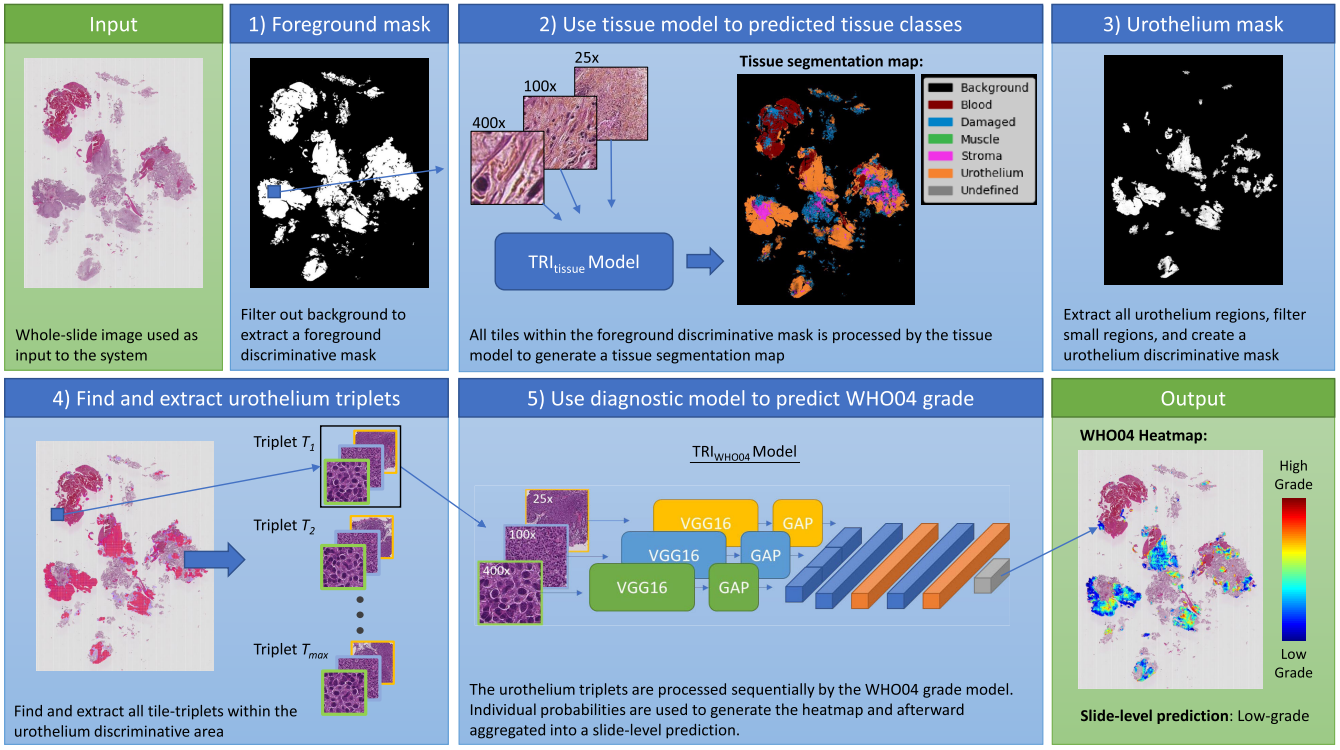

FIGURE 4. **This figure presents the pipeline for our proposed system, TRI_{grade}. Input) A WSI of urothelial carcinoma is used as input. 1) A foreground discriminative mask is found by evaluating the pixel intensity values and used as a reference to extract tiles from the WSI. 2) The TRI_{tissue}-model is used to generate a tissue segmentation map. 3) The urothelium regions are used to create a urothelium discriminative mask. 4) Using the urothelium mask, triplets consisting of tiles from three magnification levels are extracted from the input WSI. 5) The urothelium triplets are fed sequentially to the TRI_{WHO04}-model, which outputs a probabilistic score for the two classes, low- and high-grade carcinoma. Output) The system will output a WHO04 grade heatmap and a slide-level WHO04 prediction.**

removed from the patient through transurethral resection of bladder tumor. The data was collected at the University Hospital of Stavanger, Norway, in the period 2002-2011. All non-muscle invasive bladder cancers are included in the dataset, making it a true population based dataset. The biopsies were formalin-fixed and paraffin-embedded, from which 4 μm thick sections were cut and stained with Hematoxylin, Eosin, and Saffron (HES).

The slides were diagnosed and graded according to WHO73 and WHO04 [30]. All slides have the label low-grade or high-grade in the WHO04 system. In addition, cancer stage and follow-up data on recurrence and disease progression are recorded, and all patients have stage Ta or T1, i.e., non-muscle invasive. We have, however, no annotated regions with healthy non-cancerous urothelium available. All WSI have gone through a manual quality check at the department of pathology, Stavanger University Hospital, and only high-quality slides, with little or no blur, have been included in the dataset. However, as mentioned, NMIBC is removed by cauterization, which will leave burned and damaged tissue areas. All WSI are from the same laboratory, and the variation in staining color is relatively low. Ethical approval from Regional Committees for Medical and Health Research Ethics (REC), Norway, ref.no.: 2011/1539, regulated according to the Norwegian Health Research Act.

The glass slides were digitized using a Leica SCN400 slide scanner, producing WSI images in the vendor-specific scn file format. These WSI images are gigapixel images with a typical resolution of 100 000 × 100 000 pixels, stored as a pyramidal tiled image with several down-sampled versions of the base image in the same file to accommodate for rapid panning and zooming. The pyramidal structure of the WSI is depicted in Fig. 2. The Vips library [31] can extract the base image and the down-sampled versions, making it easy to extract the dataset at each resolution.

Tiles are extracted from the image pyramid at levels corresponding to 25x, 100x and 400x magnification, which is equivalent to a spatial resolution of 4 μm/pixel, 1 μm/pixel and 0.25 μm/pixel, respectively. For the TRI_{tissue}-model, we used a tile size of 128 × 128 pixels, which for the three magnification levels correspond to (512 μm × 512 μm), (128 μm × 128 μm), and (32 μm × 32 μm). For the TRI_{WHO04}-model, we had access to a much larger library of WSIs, and thus a larger tile size of 256 × 256 pixels was chosen. For the three magnification levels, this corresponds to (1 024 μm × 1 024 μm), (256 μm × 256 μm), and (64 μm × 64 μm).

The 300 WSIs included in this study were split into 220/30/50 WSIs for training, validation, and testing, respectively. Demographic characteristics of the data material were not used when splitting the data material into the different

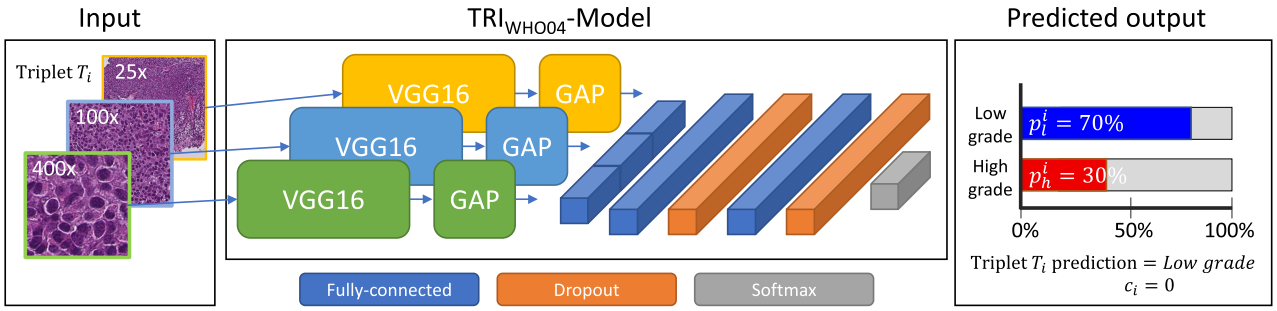

FIGURE 5. Architecture of the TRI$_{WHO04}$-model. Three separate VGG16 networks are used to extract features from each magnification scale. The global average pooling layer (GAP) is used to flatten the features into feature vectors, which are concatenated. The classification network consists of fully-connected layers and dropout layers. The output uses a softmax activation function to predict the input tiles to the two classes, low-grade and high-grade carcinoma.

datasets. Instead, the WSIs were randomly selected and stratified to include the same ratio of all diagnostic outcomes based on the WHO73 and WHO04 grading, stage, recurrence, and disease progression, to represent the data material best. The distribution of low- and high-grade WSIs in each dataset, as well as the number of triplets in the validation and test set, can be seen in Table 1.

The 50 WSIs in the test set will use the slide-level label as ground truth to evaluate the TRI$_{WHO04}$-model. In addition, a pathologist has carefully annotated low- and high-grade regions in 14 of the 50 WSIs. The 14 WSIs are a sub-set of the test set and are referred to as the *segmentation test set* and will be used to evaluate the low- and high-grade segmentation performance of the best TRI$_{WHO04}$-model.

From the 220 WSIs used for training, five datasets were extracted with a different number of triplets extracted from each WSI. A set of N triplets was selected randomly from the predicted urothelium regions in each WSI, where N was set to 250, 500, 1 000, 3 000, and 5 000.

Some of the WSIs in the data material contain only small amounts of urothelium, either because the tissue sample itself is small or because most of the tissue sample consists of damaged tissue or other tissue classes. For these WSIs, an augmentation strategy was employed, where a randomly selected set of triplets were augmented. The aim of this process is for each WSI to contribute equally, or as close as possible, to the number of triplets specified by N. Augmentation was performed by rotation and vertical/horizontal mirroring of the individual tiles in the triplet. All tiles in the triplet were augmented in the same manner. By combining rotation and mirroring, a tile can be oriented in eight uniquely defined ways, making this the maximum number a particular tile can be augmented. For $N \geq 1\,000$, some WSIs did not reach the desired number of triplets, even with 8x augmentation. No augmentation was performed on the validation or test datasets. Table 2 shows a list of total triplets extracted, before and after augmentation, for each value of N.

Fig. 3 shows a region from one WSI with the extracted tiles superimposed. The semi-transparent purple color indicates the predicted urothelium region. From this region, N randomly selected tiles are extracted as indicated by the red tiles on the image. As N increase, the density of extracted tiles also increases. Also, note that only the tile extracted at magnification level 400x is visualized in the figure. At each tile position, tiles from all three magnification levels (25x, 100x, and 400x) are extracted in such a manner that the center position of each tile corresponds to the same physical location, as illustrated in Fig. 2.

For preprocessing, all pixel intensity values were normalized from 0-255 values into 0-1 values, and the order of the color channels was altered from RGB to BGR. These steps ensure that the input data is presented to the VGG16 network in the same fashion as when it was pre-trained on the ImageNet data. No stain normalization was performed on the extracted tiles.

Our data material contains slide-level diagnostic information; however, no location annotations exist, showing where in the WSI the low- or high-grade regions are found, except on our segmentation test set, as explained. As manual annotation is time-consuming, expensive, and requires expert input, it is not feasible to get this type of detailed annotations on large datasets as needed for training such models, particularly considering both the size of each WSI and the total number of WSIs in the data material. Instead, each extracted tile inherits the slide-level WHO04 grade as its label. This is not ideal, as high-grade slides may contain regions with low-grade tissue. Consequently, all the extracted datasets are thus regarded as weakly labeled due to the inaccurate labels, which is consistent with what is called a weak label in many tasks [32]. The segmentation test set is considered strongly labeled.

B. PROPOSED SYSTEM

We propose a pipeline, called TRI$_{grade}$, that takes a WSI as input and outputs a tissue segmentation map, a WHO04 grading heatmap, and a slide-level WHO04 grade prediction. The pipeline consists of two main models, denoted as TRI$_{tissue}$-model and TRI$_{WHO04}$-model. The task of the TRI$_{tissue}$-model is to classify an input triplet as a tissue type which then can be used to make a tissue segmentation map. The task of the TRI$_{WHO04}$-model is predicting the cancer grade, i.e., low- or high-grade, based on the urothelium tissue. The TRI$_{grade}$ pipeline is depicted in Fig. 4 and explained in detail below.

Algorithm 1 Find Optimal Threshold Value D_t

Initialize: $\mathcal{Y}$, $\hat{\mathcal{Y}}$, $\mathcal{R}$, $\mathcal{D}_{c_{best}}$ are empty lists
Initialize: $Acc_{max} = 0$
for *WSI* $\leftarrow$ training set **do**
  Feed *WSI* through pipeline in Fig. 4
  $R_{high} = \frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|} c_i$
  Append R_{high} to the list $\mathcal{R}$
  Append the true grade Y of *WSI* to the list $\mathcal{Y}$
end for
for $D_c \leftarrow 0$ to 50 **do**
  for $R_{high} \leftarrow \mathcal{R}$ **do**
    $\hat{Y} = \begin{cases} \text{High-grade}, & \text{if } R_{high} \geq D_c \\ \text{Low-grade}, & \text{otherwise} \end{cases}$
    Append the slide-level prediction $\hat{Y}$ to the list $\hat{\mathcal{Y}}$
  end for
  $Acc_{D_c} = \text{sklearn.metrics.accuracy\_score}(\mathcal{Y}, \hat{\mathcal{Y}})$
  if $Acc_{D_c} > Acc_{max}$ **then**
    $Acc_{max} \leftarrow Acc_{D_c}$
    Clear list $\mathcal{D}_{c_{best}}$
  end if
  if $Acc_{D_c} \geq Acc_{max}$ **then**
    Append D_c to list $\mathcal{D}_{c_{best}}$
  end if
end for
$D_t = \lceil \frac{1}{|\mathcal{D}_{c_{best}}|} \sum \mathcal{D}_{c_{best}} \rceil$

1) $\text{TRI}_{\text{grade}}$ PIPELINE

The $\text{TRI}_{\text{grade}}$ pipeline depicted in Fig. 4 contains five steps explained here. The input to the pipeline consists of a WSI file in the vendor-specific.scn file format. First, in step 1, a foreground discriminative mask is found on the 400x level by evaluating the pixel intensity values as grey background or not. Using the foreground mask as reference, tiles with dimension 128 $\times$ 128 pixels were extracted from the WSI with 87.5% overlap, resulting in the inner 16 $\times$ 16 pixels being classified for each tile. Three tiles were extracted in the WSI (25x, 100x, and 400x) for each location, forming a triplet. All tiles in each triplet have the same dimension of 128 $\times$ 128 pixels and are extracted such as the center point corresponds to the same physical location in the WSI for all three tiles, as shown in Fig. 2.

In step 2, triplets are sequentially fed into the $\text{TRI}_{\text{tissue}}$-model we proposed in Wetteland *et al.* [13]. This model will evaluate the triplets and predict which of the six tissue classes (urothelium, stroma, muscle, blood, damaged tissue, and background) the current triplet belongs. In our case, the class of damaged tissue is a collection of all tissue that is not one of the other classes, and in our dataset, this is mainly cauterized or torn tissue areas. If blurred regions are a problem in the dataset, this can be made as a separate class or included in the damaged tissue class. After predicting all triplets, a segmented tissue map is created, visualizing all tissue regions in the WSI. This tissue map can also be presented to the clinician to help guide them more efficiently to the specific tissue types in the WSI.

From the generated tissue map, all urothelium regions are extracted in step 3. Small regions are filtered to suppress noise, and a urothelium discriminative mask is created on the 400x level. In step 4, a grid of non-overlapping tiles is overlayed on the WSI at the 400x level, this time using tiles of dimension 256 $\times$ 256 pixels. The individual tiles in the grid are checked against the discrimination mask. If 80% or more of a tile lay within the discriminate mask, the position is saved, while the remaining tiles are discarded. For the validation and test sets, triplets from all the saved positions are extracted. Whereas for the training set, N randomly selected triplets are extracted from the saved positions, where training sets are formed with N set to 250, 500, 1 000, 3 000, and 5 000. If fewer than N positions are saved, the augmentation strategy explained in the data material section is employed. The total number of extracted triplets for each dataset is shown in Tables 1 and 2.

A comprehensive description of how triplets are extracted from the WSI is given in Wetteland *et al.* [33], where a parameterized method for extracting tiles in multilevel images is given. The parameters used in this paper are the tile size parameter $L_T = 256$. The overlap-ratio between a tile and the discriminative mask is set to 80%, which corresponds to a value of $\phi = 0.8$. Tiles are checked at the 400x level by setting $\alpha = 0$, and the corresponding tiles in the triplets are found at level 25x and 100x, i.e., $\mathcal{S}_\beta = \{1, 2\}$. The binary mask B^k is set as the urothelium discriminative mask, and the starting coordinate of the grid is at position (0, 0). With these parameters and the methods described in [33], extraction of the triplets in the WSIs is repeatable and reproducible.

In step 5, the extracted urothelium triplets are fed to the $\text{TRI}_{\text{WHO04}}$-model, which outputs a probabilistic score for the two classes, low- and high-grade carcinoma. Finally, all scores are used to generate a heatmap which is overlayed on the WSI, and the aggregated micro-predictions are measured against the decision threshold D_t to get the final slide-level prediction.

2) MODEL ARCHITECTURE

The proposed pipeline in Fig. 4 contains two CNN-based models used for different tasks; the $\text{TRI}_{\text{tissue}}$-model is used for tissue classification and the $\text{TRI}_{\text{WHO04}}$-model for grading of urothelium tissue. The models are built upon the same architecture but have different inputs and outputs. The architecture consists of three separate VGG16 networks, one for each input scale. Both the model architecture and the TRI-terminology comes from our previous work on the tissue model in Wetteland *et al.* [13].

The input to the $\text{TRI}_{\text{tissue}}$-model is a triplet consisting of three 128 $\times$ 128 pixel tiles (25x, 100x, and 400x). The model can predict triplets extracted from anywhere in the WSI, but a foreground discriminative mask is usually used to save processing time by removing the background. The output of the $\text{TRI}_{\text{tissue}}$-model is a probability distribution over the six predicted classes (urothelium, stroma, muscle, blood, damaged tissue, and background). The input to the $\text{TRI}_{\text{WHO04}}$-model

is a triplet consisting of three 256 × 256 pixel tiles (25x, 100x, and 400x) extracted from urothelium tissue regions. The model outputs a probability distribution over the two predicted classes, low- and high-grade carcinoma. A block diagram of the $\text{TRI}_{\text{WHO04}}$-model architecture is depicted in Fig. 5. The $\text{TRI}_{\text{tissue}}$-model has almost the same architecture but has six output classes instead of two.

The individual tiles in the input triplet are fed to separate VGG16 networks. The VGG16 networks are used as base models with weights pre-trained on the ImageNet dataset, a large dataset containing annotated photographs used for computer vision research. Each VGG16 network acts as a feature extractor and takes a high dimensional tile as input (128 × 128 × 3 or 256 × 256 × 3 pixels) and compresses it down to a feature volume (8 × 8 × 512). A global average pooling (GAP) layer is used as the output layer for each VGG16 network, transforming the feature volume into a feature vector of length 512. The three feature vectors, one for each scale, are concatenated into one final feature vector of length 1 536 and fed to the classification network.

The classification network consists of two fully-connected (FC) layers using a rectified linear unit (ReLU) activation function, each followed by a dropout layer for regularisation. Lastly, an output layer with a softmax activation function is used to provide the prediction of the model. The two FC-layers and the two dropout layers each have a dimension of 4 096 neurons, and the output layer has one output neuron for each class. The $\text{TRI}_{\text{WHO04}}$-model consists of 67M parameters, where 23M of the parameters are trainable parameters belonging to the classification network.

3) TILE-LEVEL PREDICTION

When a triplet T_i is fed to the $\text{TRI}_{\text{WHO04}}$-model, the model outputs a list of probabilities for the two classes, low- and high-grade. These probabilities are denoted as $[p_l^i, p_h^i]$. To find the class with the largest predicted probability, the argmax function is used.

$$c_i = argmax([p_l^i, p_h^i]) \tag{1}$$

where c_i is the index to the predicted class for the triplet at position T_i. The low-grade class has an index of 0, and the high-grade class has an index of 1.

The proposed system can also produce a heatmap from the individual triplet probabilities, which indicates the location of low- and high-grade regions. This is useful for pathologists who can focus their limited per-patient investigation time on the diagnostic relevant areas in the WSI. A color mapping function converts the high-grade probability p_h^i into a color based on its value. This color is then superimposed on the WSI at the current triplet's position, covering the same area as the 400x magnification tile in the triplet. This results in the heatmap, as seen in the bottom-right of Fig. 4. Only the model's probabilistic score for the high-grade class is used to generate the heatmaps. However, because there are only two classes, a low probabilistic score of the high-grade class implicitly means a high score for the low-grade class. I.e., red highlighted regions in the heatmaps are associated with the high-grade class, and blue highlights indicate the low-grade class.

4) SLIDE-LEVEL PREDICTION

In addition to predicting the individual triplets, we also output a WHO04 slide-level prediction. A pathologist will often assign the worst case to a slide during a clinical examination, meaning that if a high-grade region exists in the WSI, the WHO04 grading should be high-grade. However, we must assume some misclassification in the WSI from both the $\text{TRI}_{\text{tissue}}$-model and $\text{TRI}_{\text{WHO04}}$-model, so there must be a minimum amount of high-grade triplets before the slide-level prediction becomes high-grade, and we would like to find a decision threshold, D_t, which maximizes correct prediction of the WSIs.

By summing over c_i, the number of triplets predicted as high-grade is counted, since triplets predicted as low-grade is at index 0 and thus not adding to the sum. By dividing by the total number of triplets in the WSI, we get the ratio of high-grade triplets referred to as R_{high} in this paper:

$$R_{high} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} c_i \tag{2}$$

If R_{high} exceeds the decision threshold D_t, the slide is given the slide-level prediction of high-grade; else, it is considered low-grade.

$$\hat{Y} = \begin{cases} \text{High-grade}, & \text{if } R_{high} \geq D_t \\ \text{Low-grade}, & \text{otherwise} \end{cases} \tag{3}$$

Algorithm 1 describes how to find the optimal threshold value D_t. Y is considered the ground truth grading of a slide and consists of a single value, whereas $\mathcal{Y}$ is a list of all the ground truth values. The same holds for $\hat{Y}$ and $\hat{\mathcal{Y}}$, which holds a single slide-level prediction and a list of all the predictions, respectively. First, all WSIs are processed, and the ratio R_{high} for each WSI is appended to the list $\mathcal{R}$. The true grade Y of each WSI is also saved in the list $\mathcal{Y}$. All WSIs in the dataset are processed before proceeding to the next step. A set of candidate threshold values, D_c, between 0-50% are tested one at a time. For each candidate threshold, the slide-level prediction $\hat{Y}$ for all WSIs is saved to the list $\hat{\mathcal{Y}}$. The total accuracy score is then calculated for the dataset. The decision threshold D_t is chosen as the candidate threshold with the highest score, or, if more than one D_c value yielded the same maximum result, the average integer value is selected as the decision threshold D_t.

5) TRAINING PARAMETERS

The $\text{TRI}_{\text{WHO04}}$-model was trained using a stochastic gradient descent (SGD) optimizer with a learning rate of 1×10^{-3}, learning rate decay of 1×10^{-6}, and momentum set to 0.9. The batch size used during training was set to 128. Both dropout layers had a dropout rate of 0.5. The cross-entropy loss function was used to optimize the model during training.

TABLE 3. Slide-level prediction results for automatic WHO04 grading tested on the 50 WSIs of the test set. Precision, recall, and F1-score is the weighted average score for the two classes across all 50 WSIs in the test set. D_t is the decision threshold found using Algorithm 1. The column trained epochs show how many epochs each model was trained before the early stopping criteria were reached. Training times are shown as hours:minutes.

Model	Trained epochs	Time per epoch	Training time	Precision	Recall	F1-Score	D_t
TRI_{WHO04}-250	23	1:22	31:39	0.86	0.84	0.84	49
TRI_{WHO04}-250-AUG	33	1:19	43:53	0.89	0.86	0.85	47
TRI_{WHO04}-500	21	1:42	35:59	0.89	0.86	0.85	43
TRI_{WHO04}-500-AUG	21	1:44	36:39	0.77	0.76	0.76	49
TRI_{WHO04}-1000	15	1:52	28:03	0.83	0.82	0.82	49
TRI_{WHO04}-1000-AUG	18	1:55	34:45	0.80	0.80	0.80	49
TRI_{WHO04}-3000	15	3:24	51:01	0.89	0.86	0.85	49
TRI_{WHO04}-3000-AUG	12	3:29	41:54	0.78	0.78	0.78	49
TRI_{WHO04}-5000	16	4:10	66:42	0.85	0.84	0.84	48
TRI_{WHO04}-5000-AUG	17	5:18	90:20	**0.92**	**0.90**	**0.90**	48

The pre-trained weights of the VGG16 networks were held frozen during training. To avoid overfitting the models on the training set, an early-stopping rule monitored the validation loss and stopped the training when no improvements were seen for ten epochs. The best epoch was restored when testing the models on the test set.

To train the models, a program was written in Python 3.6 using Keras 2.2.4 together with the Tensorflow 1.14 as backend [34], [35]. The PyVips 2.1 library was used for handling the WSI [31], and Scikit-learn 0.19 for evaluation [36]. The models were training on a Ubuntu 18.04 server, running on dual Xeon E5-2650 v5 @ 2.2GHz with a total of 48 cores. An Nvidia Tesla P100 16GB GPU was used for the training. Training parameters for the TRI_{tissue}-model can be found in Wetteland *et al.* [13].

III. EXPERIMENTS

We have conducted two experiments, listed here.

Experiment 1: is for slide-level prediction of WHO04 grade and is tested on the test set of 50 WSIs. As training of the TRI_{WHO04}-model is very time-consuming, we wanted to see if it is preferable to utilize more of the available urothelium data from each WSI as training data at the cost of additional training time or if a smaller dataset could perform equally well. This is interesting, both for our research group as well as other researchers working with large WSI datasets. If the optimal number of tiles used from each WSI during training can be lowered, then time can be saved in future experiments. To investigate this, we created several datasets where we extracted N triplets per WSI, as shown in Table 2. In this experiment, ten versions of the TRI_{WHO04}-model, all with the same architecture, were trained on training sets of various sizes, listed in Table 2. The micro predictions from the individual triplets were aggregated into a slide-level prediction of the WHO04 grading. A decision threshold D_t was found for each model using Algorithm 1; then, equation 3 was used to provide the final predicted grade.

Experiment 2: is testing the tile-level prediction and compare that in detail with the 14 WSIs of the segmentation test set. This set contains pathologist annotated regions belonging to either low- or high-grade which are considered the ground truth. The best model from experiment 1 is used for this, and the model's performance will be visualized as heatmaps. Calculation of recall and F1-score will be presented for each WSI, in addition to a total score across all WSIs.

IV. RESULTS

In experiment 1, slide-level test results for the ten models are listed in Table 3, showing trained epochs, time, precision, recall, F1-score, and the threshold value D_t. For precision, recall, and F1-score, the weighted average score is presented as reported by the *classification report* function from the scikit-learn library [36].

For experiment 2, the TRI_{WHO04}-5000-AUG model was used, as it performed best in experiment 1. The predicted heatmaps for each WSI in the segmentation test set are shown in Fig. 6 together with the ground truth. Recall, and F1-score for each WSI is listed in Table 4. As each ground truth WSIs only contain annotations for one of the two classes, the precision score will always be 1.00 because whenever the model predicts the ground truth class, it will be correct. The precision column in Table 4 is thus discarded. The last row in Table 4 shows the average value of all scores for each class together with the standard deviation. Table 5 shows the total aggregated results for all 14 WSIs. Here, the predictions for all WSIs are accumulated before the score is calculated.

A slide-level comparison between the proposed TRI_{grade} system and the model presented in Jansen *et al.* [22] is shown in Table 6. The TRI_{grade} system consists of the TRI_{tissue}-model followed by the TRI_{WHO04}-5000-AUG model. Values for sensitivity, specificity, and accuracy are shown for easier comparison with the reported results from [22]. These values are unweighted and calculated using values for true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Note that these results are based on models trained and evaluated on different datasets.

V. DISCUSSION

The three VGG16 networks are identical copies as we have used frozen (pre-trained) weights in this work. Thus, it would be possible to use only one copy of the model, with the appropriate change in the architecture, keeping in mind that

TABLE 4. Tile-level prediction for each individual WSI in the segmentation test set, using the TRI_{WHO04}-5000-AUG model. The WSI numbering is referring to the WSIs in Fig. 6. The last row shows the average value and standard deviation for its respective column.

	Low-grade		High-grade	
WSI	Recall	F1-score	Recall	F1-Score
WSI A	-	-	0.79	0.88
WSI B	-	-	0.90	0.95
WSI C	-	-	0.86	0.92
WSI D	0.87	0.93	-	-
WSI E	-	-	0.94	0.97
WSI F	0.83	0.91	-	-
WSI G	0.86	0.92	-	-
WSI H	-	-	0.90	0.95
WSI I	0.85	0.92	-	-
WSI J	0.79	0.88	-	-
WSI K	-	-	0.92	0.96
WSI L	0.92	0.96	-	-
WSI M	0.68	0.81	-	-
WSI N	-	-	0.58	0.73
Average	0.83 ± 0.07	0.91 ± 0.04	0.84 ± 0.12	0.91 ± 0.08

the feature vectors from the different magnifications are concatenated before the classification network. However, utilizing three versions of the VGG16 network allows us to train the entire multiscale model end-to-end and allows unfreezing the weights if a larger training set is available. We have experimented with unfreezing weights, but we quickly get overfitting problems with the available data material, this is therefore omitted from the paper.

Experiment 1 was conducted using ten training sets with a different number of triplets extracted from the same 220 WSI. From the result in Table 3, we see that the best performing model is trained on the largest dataset. However, the other models are not far behind. Even with a small value of N, the models do a good job at correctly predicting the WHO04 grade of WSIs.

Regarding overfitting, we tried training the models using unfrozen weights in the VGG16 networks, but this led to instantaneous overfitting of the model and had no improvements on the validation set. However, by freezing the weights, we see that all models improve on the validation dataset before reaching a plateau and eventually triggering the early stopping trigger. E.g., as shown in Fig. 7, the best model, TRI_{WHO04}-5000-AUG, improved its performance for seven epochs before training stopped after epoch 17. The weights from epoch seven were restored when using the model on the test sets. The number of trained epochs before the early stopping criteria is triggered decreases as the training dataset increases. This can be explained by the models trained on the larger datasets having more parameter updates per epoch than that of the smaller dataset models, thus reaching the plateau faster. Similarly, we see that the duration of one epoch is increasing as the dataset size increases. There is about a 60-hour difference in the smallest and largest model by comparing the total training time. Even though we would advise utilizing the most data to train a production model, it could be helpful to do an extended hyperparameter search and train multiple models on a smaller dataset.

TABLE 5. Aggregated tile-level result for all WSIs in the segmentation test set using the TRI_{WHO04}-5000-AUG model.

	Precision	Recall	F1-score
Low-grade	0.83	0.79	0.81
High-grade	0.90	0.81	0.85
Weighted Average	0.87	0.80	0.83

TABLE 6. Comparison table for automatic slide-level grading between our proposed method and the method presented in Jansen *et al.* [22]. Note that these results are based on models trained and evaluated on different datasets.

Model	Sensitivity	Specificity	Accuracy
TRI_{grade}	0.85	1.00	0.90
Jansen *et al.* [22]	0.71	0.76	0.74

Experiment 2, tile-level prediction, was conducted using the TRI_{WHO04}-5000-AUG model, which had a slide-level F1-score of 0.90. As seen in Fig. 6, Table 4 and 5, the results are overall excellent. The model does a very good job at correctly identifying both the low-grade and high-grade regions in the different WSIs. Table 4 shows that the model achieved an average F1-score of 91% for both the low-grade and high-grade classes. The aggregated score for all WSIs in Table 5 shows a small decrease in performance, with an F1-score of 81% and 85% for the two classes, respectively.

The largest misclassification in Fig. 6 is one of the regions in WSI-N, where the ground truth is high-grade, but the model predicts low-grade. When reevaluated by the pathologist, the misclassified area was found to be heterogenous, showing mixed low- and high-grade features, consequently regarded as high-grade initially. This illustrates one of the challenges with automatic grading of urothelial carcinoma, that grading between low- and high-grade is not two distinct binary classes but rather a continuous spectrum with a floating transition, making it difficult to set a hard threshold between the two.

To correct such misclassifications, and also avoid the costly task of annotating a large dataset, one possible solution is human-assisted learning. For example, the proposed TRI_{grade} system could be used to find and predict urothelium regions into the low-grade and high-grade classes, e.g., like the regions seen in Fig. 6. Then, a pathologist could verify the regions in each WSI and correct misclassified regions. This way, a large, strongly labeled dataset could be created, and the TRI_{WHO04}-model could be fine-tuned on the new dataset.

A direct comparison of results with others reported in the literature is not straightforward, as the experiments performed in this paper are conducted on a private dataset, which is often the case in many medical applications. To our knowledge, there exists no publically available NMIBC dataset or any publically available models from other researchers that we can evaluate on our dataset. The work of Jansen *et al.* [22] is based on the same labels but evaluated on a private dataset using different methods. Unfortunately, their models are not available for us to evaluate, and we do not have access to labels to train a Unet segmentation model from scratch, hence we cannot test the same approach by training the models ourselves. However, even though the dataset or model used in

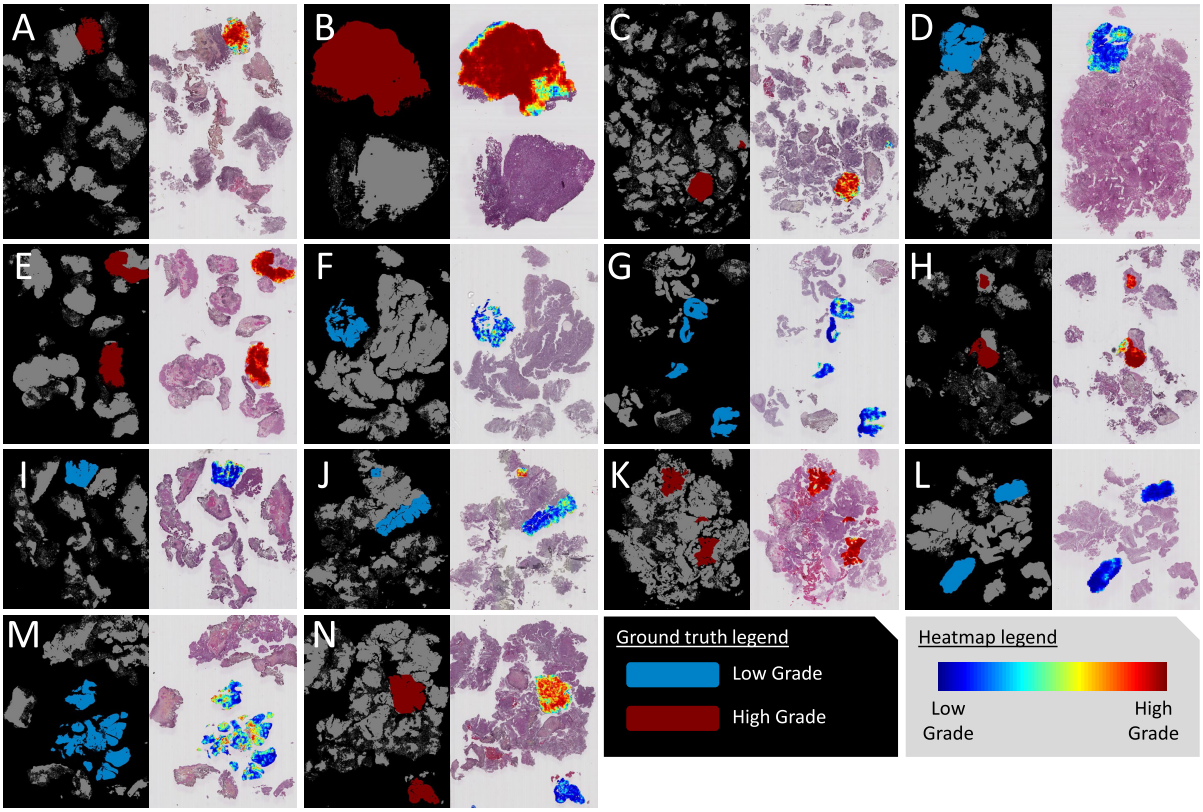

FIGURE 6. Ground truth annotations vs. model prediction. The WSI with a black background is the ground truth images with low- and high-grade annotations. The WSI with a grey background has superimposed a heatmap from the same area as the ground truth and highlights the predictions from the TRI_{WHO04}-model. For quantitative results, see Table 4 and 5.

Jansen *et al.* [22] are not publically available, a comparison is still included as both research results are based on an NMIBC dataset of similar size (328 WSIs from 232 patients vs. our dataset of 300 WSIs), a similar split of the dataset into training, validation, and test, and the use of the same labels (WHO04). The results in Table 6 compare the slide-level sensitivity, specificity, and accuracy for our proposed TRI_{grade} pipeline, to the results reported in table 3 from [22]. We achieve better results on all metrics, and with 45 of the 50 WSIs correctly predicted, we achieve an accuracy of 0.90.

Training and validation accuracy from the training of the TRI_{WHO04}-5000-AUG model is shown in Fig. 7. The model uses frozen pre-trained weights for the VGG16 networks, and only the last layers in the model have random weights which are being optimized. The model uses the largest training dataset from Table 2 with a mini-batch size of 128, resulting in a large number of weight updates per epoch, and the majority of the accuracy is achieved from the first epoch. After the initial epoch, the validation accuracy is not improving too much. This is most likely because the datasets use imprecise weak labels (e.g., all urothelium triplets extracted from a high-grade WSI will have the class label high-grade, but not all triplets from this WSI will represent high-grade tissue). Note also that all the urothelium triplets from all the WSIs in the validation set are predicted before Tensorflow computes the accuracy score for the validation set.

A. USAGE SCENARIOS

The automatic TRI_{grade} system presented in this paper has many potential applications. The tissue model we presented in Wetteland *et al.* [13] provides the tissue segmentation maps, which clinicians can use to discriminate urothelium regions from other tissue classes. This can be a valuable tool to aid pathologists in examining the whole-slide images by focusing their attention on the diagnostic relevant areas of the stained specimen. With the addition of the TRI_{WHO04}-model presented in this paper, the focus can not only be aimed towards the urothelium regions in general but be further narrowed down to the most *severe* urothelium regions.

The automated slide-level prediction can potentially be used to prioritize high-grade patients for earlier examination. Also, it can be used as input to an automatic prognostic tool and output a measure of the patient's overall clinical outcome, such as the risk of recurrence, 1-yr and 5-yr survival rate, and mortality. In the future, it is also a possibility to use it in an automatic system that predicts how a patient will respond to a given treatment and therapy program.

B. LIMITATIONS

In the paper, we train a model to classify urothelium tissue into two classes, low- and high-grade carcinoma. However, it is also a possibility that the urothelium tissue can be healthy non-cancerous tissue. Since our models are dependent on the weak slide-level label, and all cases in the data material are

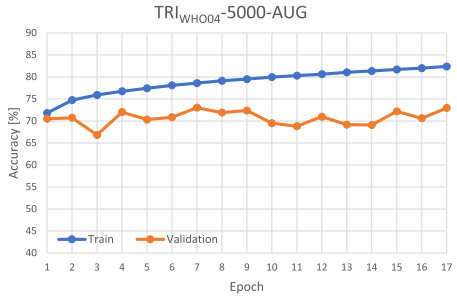

FIGURE 7. Training and validation accuracy for the TRI$_{WHO04}$-5000-AUG model. The model is trained on imprecise weak labels, using the largest training set in Table 2. Results are shown for tile-level prediction on the entire training and validation sets. Validation accuracy is computed at the end of each epoch.

diagnosed with cancer, we currently do not have any training material containing non-cancerous samples.

All WSIs in this study are collected from the same laboratory and consists of high quality with relatively small variations in stain colors and little blur. This is both a strength in the sense that we have produced good models and reliable predictions, but also a limitation in the sense that we do not know how the system will perform on slides of lower quality.

C. FUTURE WORK

In future work, preprocessing steps might be added to deal with color variations, blur, and folded tissue, or the tissue segmentation model can be updated with a new class for blur, providing a more generalized system.

From [13] it was concluded that for the tissue segmentation task, the multiscale TRI-25x-100x-400x model (which is used as the TRI$_{tissue}$-model in this work) provided the best performance. Following, a multiscale model was adopted for the grading task as well, with the masking of the urothelium tissue performed at the 400x level. However, the large field-of-view provided by the 25x and 100x magnification will bring neighboring tissue types into the triplet, like, for example, damaged tissue, which might affect the performance in such areas. In future work, we would like to use the tissue segmentation maps and not only extract the urothelium tissue but also mask out unwanted regions of damaged tissue and blood. Incorporating attention modules is also something we will try, which would further help explain what parts of the WSI are responsible for the predictions.

Cells of low-grade cancer often resemble that of non-cancerous cells, and high-grade cells have a more abnormal appearance and are disorganized. Thus, we expect that non-cancerous tissue would be predicted as low-grade carcinoma. However, this is our expectation as we do not have verified material to test this on. To better detect these non-cancerous regions in the future, we would have to expand our training dataset to include examples of non-cancerous urothelium. The TRI$_{WHO04}$-model architecture must be updated to include one additional class on the output and then be trained on the updated dataset.

The proposed model uses three VGG16 networks as feature extractors. In the future, we would like to experiment with other deep learning networks for our base model. Newer deep learning models continuously improve the results on datasets like ImageNet, and could potentially improve feature extraction of urothelium tissue. We also plan to look into different ways of fusing the multiscale information, both for the tissue classifier (TRI$_{tissue}$) and grade-classifier (TRI$_{WHO04}$).

VI. CONCLUSION

In this paper, we have proposed a TRI$_{grade}$ pipeline for automatic grading of urothelial carcinoma slides based on the WHO04 grading system. First, the slide is segmented into the tissue classes (urothelium, stroma, muscle, blood, damaged tissue, and background). Next, tiles are extracted at three magnification levels (25x, 100x, and 400x) from the urothelium regions. The three tiles form a triplet, which is fed sequentially to a multiscale CNN-based WHO04 grading model.

The proposed method will generate a tissue segmentation map, helpful for the clinicians to easier find diagnostic relevant regions during an examination. The system will also output a WHO04 grade heatmap, highlighting the most severe urothelium tissue regions, beneficial for the pathologists who can focus their limited per-patient time on the most important regions in the WSI. Finally, the system produces a slide-level WHO04 grade that could potentially be used to prioritize high-grade patients for earlier examination, as well as suggest the diagnosis to the pathologist.

Ten WHO04 grade models were trained on datasets of varying sizes. Note that all the same number of WSI were used all the time, but a different number of triplets were extracted from each WSI, constituting the training set. The model trained on the largest training dataset achieved the best result, a weighted average F1-score of 0.90 on the test set. This model was further evaluated on a segmentation test set, where low- and high-grade regions were annotated by a pathologist. On this task, the model got an average F1-score of 0.91 on both the low-grade and high-grade classes.

The system as a whole can be used by clinicians and pathologists to potentially improve their decision-making and further help patients by receiving correct diagnoses and treatment.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA A, Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] O. M. Mangrud, "Identification of patients with high and low risk of progresson of urothelial carcinoma of the urinary bladder stage Ta and T1," Ph.D. dissertation, Dept. Clin. Med., Fac. Med. Dentistry, Univ. Bergen, Bergen, Norway, 2014.

[3] K. D. Sievert, B. Amend, U. Nagele, D. Schilling, J. Bedke, M. Horstmann, J. Hennenlotter, S. Kruck, and A. Stenzl, "Economic aspects of bladder cancer: What are the benefits and costs?" *World J. Urol.*, vol. 27, no. 3, pp. 295–300, Jun. 2009, doi: 10.1007/s00345-009-0395-z.

[4] M. F. Botteman, C. L. Pashos, A. Redaelli, B. Laskin, and R. Hauser, "The health economics of bladder cancer," *PharmacoEconomics*, vol. 21, no. 18, pp. 1315–1330, Dec. 2003, doi: 10.1007/bf03262330.

[5] V. Kvikstad, O. M. Mangrud, E. Gudlaugsson, I. Dalen, H. Espeland, J. P. A. Baak, and E. A. M. Janssen, "Prognostic value and reproducibility of different microscopic characteristics in the WHO grading systems for pTa and pT1 urinary bladder urothelial carcinomas," *Diagnostic Pathol.*, vol. 14, no. 1, pp. 1–8, Dec. 2019, doi: 10.1186/s13000-019-0868-3.

[6] O. M. Mangrud, R. Waalen, E. Gudlaugsson, I. Dalen, I. Tasdemir, E. A. M. Janssen, and J. P. A. Baak, "Reproducibility and prognostic value of WHO1973 and WHO2004 grading systems in TaT1 urothelial carcinoma of the urinary bladder," *PLoS ONE*, vol. 9, no. 1, Jan. 2014, Art. no. e83192, doi: 10.1371/journal.pone.0083192.

[7] L. Browning, E. Fryer, D. Roskell, K. White, R. Colling, J. Rittscher, and C. Verrill, "Role of digital pathology in diagnostic histopathology in the response to COVID-19: Results from a survey of experience in a UK tertiary referral hospital," *J. Clin. Pathol.*, vol. 74, no. 2, pp. 129–132, Feb. 2021, doi: 10.1136/jclinpath-2020-206786.

[8] M. G. Hanna, V. E. Reuter, O. Ardon, D. Kim, S. J. Sirintrapun, P. J. Schüffler, K. J. Busam, J. L. Sauter, E. Brogi, L. K. Tan, and B. Xu, "Validation of a digital pathology system including remote review during the COVID-19 pandemic," *Modern Pathol.*, vol. 33, no. 11, pp. 2115–2127, Nov. 2020, doi: 10.1038/s41379-020-0601-5.

[9] M. K. K. Niazi, A. V. Parwani, and M. N. Gurcan, "Digital pathology and artificial intelligence," *Lancet Oncol.*, vol. 20, no. 5, pp. e253–e261, 2019, doi: 10.1016/S1470-2045(19)30154-8.

[10] K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi, "Artificial intelligence in digital pathology—New tools for diagnosis and precision oncology," *Nature Rev. Clin. Oncol.*, vol. 16, no. 11, pp. 703–715, Nov. 2019, doi: 10.1038/s41571-019-0252-y.

[11] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Med. Image Anal.*, vol. 33, no. 6, pp. 170–175, 2016, doi: 10.1016/j.media.2016.06.037.

[12] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," 2016, *arXiv:1606.05718*. [Online]. Available: http://arxiv.org/abs/1606.05718

[13] R. Wetteland, K. Engan, T. Eftestøl, V. Kvikstad, and E. A. M. Janssen, "A multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides," *Technol. Cancer Res. Treatment*, vol. 19, Jan. 2020, Art. no. 153303382094678, doi: 10.1177/1533033820946787.

[14] K. Munir, H. Elahi, A. Ayub, F. Frezza, and A. Rizzi, "Cancer diagnosis using deep learning: A bibliographic review," *Cancers*, vol. 11, no. 9, p. 1235, Aug. 2019, doi: 10.3390/cancers11091235.

[15] S. Benjamens, P. Dhunnoo, and B. Meskó, "The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–8, Dec. 2020, doi: 10.1038/s41746-020-00324-0.

[16] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Med.*, vol. 25, no. 8, pp. 1301–1309, Aug. 2019, doi: 10.1038/s41591-019-0508-1.

[17] S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, and N. Rajpoot, "Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101563, doi: 10.1016/j.media.2019.101563.

[18] Y. Celik, M. Talo, O. Yildirim, M. Karabatak, and U. R. Acharya, "Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images," *Pattern Recognit. Lett.*, vol. 133, pp. 232–239, May 2020, doi: 10.1016/j.patrec.2020.03.011.

[19] P. Ström, K. Kartasalo, H. Olsson, L. Solorzano, B. Delahunt, D. M. Berney, D. G. Bostwick, A. J. Evans, D. J. Grignon, P. A. Humphrey, and K. A. Iczkowski, "Pathologist-level grading of prostate biopsies with artificial intelligence," 2019, *arXiv:1907.01368*. [Online]. Available: http://arxiv.org/abs/1907.01368

[20] J. D. Ianni, R. E. Soans, S. Sankarapandian, R. V. Chamarthi, D. Ayyagari, T. G. Olsen, M. J. Bonham, C. C. Stavish, K. Motaparthi, C. J. Cockerell, T. A. Feeser, and J. B. Lee, "Tailored for real-world: A whole slide image classification system validated on uncurated multi-site data emulating the prospective pathology workload," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Dec. 2020, doi: 10.1038/s41598-020-59985-2.

[21] O.-J. Skrede, S. De Raedt, A. Kleppe, T. S. Hveem, K. Liestøl, J. Maddison, H. A. Askautrud, M. Pradhan, J. A. Nesheim, F. Albregtsen, and I. N. Farstad, "Deep learning for prediction of colorectal cancer outcome: A discovery and validation study," *Lancet*, vol. 395, no. 10221, pp. 350–360, 2020, doi: 10.1016/S0140-6736(19)32998-8.

[22] I. Jansen, M. Lucas, J. Bosschieter, O. J. de Boer, S. L. Meijer, T. G. van Leeuwen, H. A. Marquering, J. A. Nieuwenhuijzen, D. M. de Bruin, and C. D. Savci-Heijink, "Automated detection and grading of non–muscle-invasive urothelial cell carcinoma of the bladder," *Amer. J. Pathol.*, vol. 190, no. 7, pp. 1483–1490, Jul. 2020, doi: 10.1016/j.ajpath.2020.03.013.

[23] Z. Zhang, P. Chen, M. McGough, F. Xing, C. Wang, M. Bui, Y. Xie, M. Sapkota, L. Cui, J. Dhillon, N. Ahmad, F. K. Khalil, S. I. Dickinson, X. Shi, F. Liu, H. Su, J. Cai, and L. Yang, "Pathologist-level interpretable whole-slide cancer diagnosis with deep learning," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 236–245, May 2019, doi: 10.1038/s42256-019-0052-1.

[24] M. Lucas, I. Jansen, T. G. van Leeuwen, J. R. Oddens, D. M. de Bruin, and H. A. Marquering, "Deep learning–based recurrence prediction in patients with non–muscle-invasive bladder cancer," *Eur. Urol. Focus*, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2405456920303102, doi: 10.1016/j.euf.2020.12.008.

[25] K. Sirinukunwattana, N. K. Alham, C. Verrill, and J. Rittscher, "Improving whole slide segmentation through visual context—A systematic study," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 192–200, doi: 10.1007/978-3-030-00934-2_22.

[26] N. Hashimoto, D. Fukushima, R. Koga, Y. Takagi, K. Ko, K. Kohno, M. Nakaguro, S. Nakamura, H. Hontani, and I. Takeuchi, "Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3852–3861, doi: 10.1109/CVPR42600.2020.00391.

[27] X. Zhang, F. Dong, G. Clapworthy, Y. Zhao, and L. Jiao, "Semi-supervised tissue segmentation of 3D brain MR images," in *Proc. 14th Int. Conf. Inf. Visualisation*, vol. 2688, Jul. 2010, pp. 623–628. [Online]. Available: http://CEUR-WS.org/Vol-2688/paper14.pdf

[28] R. Wetteland, K. Engan, T. Eftestøl, V. Kvikstad, and E. Janssen, "Multiclass tissue classification of whole-slide histological images using convolutional neural networks," in *Proc. 8th Int. Conf. Pattern Recognit. Appl. Methods*, vol. 1, 2019, pp. 320–327, doi: 10.5220/0007253603200327.

[29] J. Urdal, K. Engan, V. Kvikstad, and E. A. M. Janssen, "Prognostic prediction of histopathological images by local binary patterns and RUSBoost," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 2349–2353, doi: 10.23919/EUSIPCO.2017.8081630.

[30] M. Babjuk, A. Böhle, M. Burger, O. Capoun, D. Cohen, E. M. Compérat, V. Hernández, E. Kaasinen, J. Palou, M. Rouprêt, B. W. G. van Rhijn, S. F. Shariat, V. Soukup, R. J. Sylvester, and R. Zigeuner, "EAU guidelines on non–muscle-invasive urothelial carcinoma of the bladder: Update 2016," *Eur. Urol.*, vol. 71, no. 3, pp. 447–461, Mar. 2017, doi: 10.1016/j.eururo.2016.05.041.

[31] K. Martinez and J. Cupitt, "VIPS—A highly tuned image processing software architecture," in *Proc. IEEE Int. Conf. Image Process.*, vol. 2, Sep. 2005, pp. 2–574, doi: 10.1109/ICIP.2005.1530120.

[32] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.*, vol. 54, pp. 280–296, May 2019, doi: 10.1016/j.media.2019.03.009.

[33] R. Wetteland, K. Engan, and T. Eftestøl, "Parameterized extraction of tiles in multilevel gigapixel images," in *Proc. 12th Int. Symp. Image Signal Process. Anal. (ISPA)*, 2021.

[34] F. Chollet. (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[35] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, vol. 16, 2016, pp. 265–283.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.