Medical Image Analysis 73 (2021) 102206

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Joint categorical and ordinal learning for cancer grading in pathology images

Trinh Thi Le Vuong[a], Kyungeun Kim[b], Boram Song[b], Jin Tae Kwak[a,*]

[a] School of Electrical Engineering, Korea University, Seoul 02841, Republic of Korea
[b] Department of Pathology, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul 03181, Republic of Korea

ARTICLE INFO

Article history:
Received 3 November 2020
Revised 26 July 2021
Accepted 28 July 2021
Available online 8 August 2021

Keywords:
Cancer grading
Multi-task learning
Categorical classification
Ordinal classification

ABSTRACT

Cancer grading in pathology image analysis is one of the most critical tasks since it is related to patient outcomes and treatment planning. Traditionally, it has been considered a categorical problem, ignoring the natural ordering among the cancer grades, i.e., the higher the grade is, the more aggressive it is, and the worse the outcome is. Herein, we propose a joint categorical and ordinal learning framework for cancer grading in pathology images. The approach simultaneously performs both categorical classification and ordinal classification and aims to leverage the distinctive features from the two tasks. Moreover, we propose a new loss function for the ordinal classification task that offers an improved contrast between the correctly classified examples and misclassified examples. The proposed method is evaluated on multiple collections of colorectal and prostate pathology images that underwent different acquisition and processing procedures. Both quantitative and qualitative assessments of the experimental results confirm the effectiveness and robustness of the proposed method in comparison to other competing methods. The results suggest that the proposed approach could permit improved histopathologic analysis of cancer grades in pathology images.

1. Introduction

Cancer grading is one of the most crucial tasks in cancer pathology since it is known to be associated with patient outcomes. Trained pathologists are capable of analyzing complex tissue structures and determining tumor grades, but the decisions are known to be subjective, qualitative, and time-consuming, leading to substantial intra- and inter-observer variability (Egevad et al., 2013; Elmore et al., 2015). Due to the increase in the workload (Metter et al., 2019), pathologists' fatigue and burnout could contribute to diagnostic errors, decreasing the overall quality of pathology service. For such reasons, numerous machine learning methods (Niazi et al., 2019; Madabhushi and Lee, 2016) have been applied to pathology images in order to improve the accuracy, efficiency, and robustness of cancer pathology.

The factors that determine tumor grade can vary between different types of cancer, but it is in general categorized into distinct grades, depending on the amount of abnormality of tumor cells; for instance, low, mid, and high-grade tumor. The higher the tumor grade is, the more aggressive the tumor cells are. This indicates that there is a natural ordering among tumor grades, i.e., cancer grading is intrinsically an ordinal problem. However, it has been mainly studied as a multi-class categorical classification problem (Arvaniti et al., 2018; Nagpal et al., 2019; Albarqouni et al., 2016; Araújo et al., 2017; Coudray et al., 2018) where machine learning methods are trained to classify tissue samples into different grades regardless of the ordering among them both in the conventional approaches of utilizing hand-crafted features (Kwak et al., 2011; Kather et al., 2016; Kwak and Hewitt, 2017; Turki and Wei, 2018; Krawczyk et al., 2016; Gorelick et al., 2013) and in the recent deep learning-based methods (Arvaniti et al., 2018; Nagpal et al., 2019; Albarqouni et al., 2016; Araújo et al., 2017; Coudray et al., 2018; Serag et al., 2019). Due to the ordinal characteristics, cancer grading can be cast as an ordinal classification or regression problem. The ordinal classification or regression methods aim to predict class labels on an ordinal scale. Most of such methods have been built based upon a ranking algorithm (Cao et al., 2012), a series of binary classification algorithms (Li and Lin, 2007), or deep neural networks (Fu et al., 2018). These have been applied to age estimation (Cao et al., 2012) and depth estimation (Fu et al., 2018). As for pathology image analysis, disease prognosis (Mobadersany et al., 2018) and cell detection (Xie et al., 2018) have been formulated as a regression problem, predicting a real or continuous value, but not as an ordinal classification problem. Furthermore, multi-task learning, which jointly learns multiple related tasks at the same time,

* Corresponding author.
E-mail address: jkwak@korea.ac.kr (J.T. Kwak).

https://doi.org/10.1016/j.media.2021.102206
1361-8415/

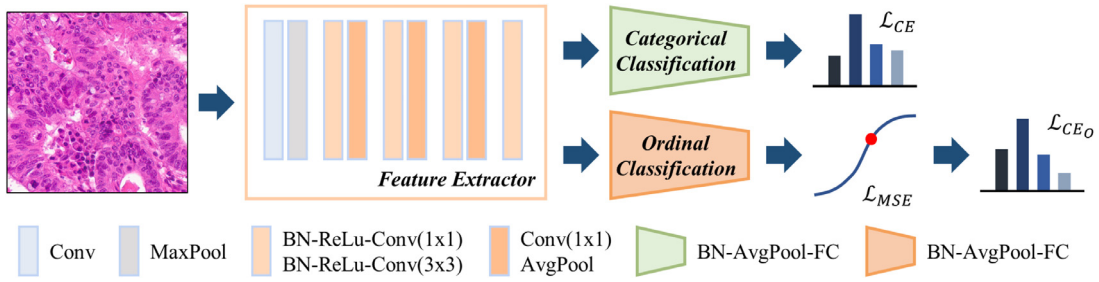

Fig. 1. Overview of the proposed network. The network consists of three components: one feature extractor and categorical and ordinal classification branches. Three loss functions are adopted to train the network.

has been applied to pathology image analysis. Most of such works simultaneously conduct the segmentation and classification tasks; for instance, the segmentation and classification of breast tissues (Mehta et al., 2018), nuclei (Graham et al., 2019), and glaucoma (Zhao et al., 2019). (Liao et al., 2017) proposed a multi-task learning approach that jointly conducts the classification of multiple cancer types. However, to the best of our knowledge, there has been no such prior work for cancer grading in pathology images.

Herein, we propose a unified framework of deep learning that simultaneously performs both categorical and ordinal classification (Fig. 1). For the categorical classification, we exploit the existing deep learning models that have already shown to be effective in pathology image analysis. As for the ordinal classification, we formulate it as both a regression problem and an ordinal regression problem by introducing a new loss function that converts the regression output into probability measures. The proposed model consists of three major parts: a shared feature extractor and two subbranches a categorical classification branch and an ordinal classification branch. The intuition behind it is quite straightforward: useful information that is specific to cancer grading (extracted by the feature extractor) can be utilized to identify its status (categorical classification) and to determine its relative order (ordinal classification); the two tasks are related to each other and thus can benefit from each other, assuming that there exists a general model that can jointly learn multiple tasks together, leading to a superior performance to the models that learn the single tasks independently (Zhang and Yang, 2018).

To summarize, our main contributions are as follows:

- We propose a unified deep learning model for pathology image grading that takes advantage of both categorical classification and ordinal classification. This exploits the natural property of pathology images, i.e., tumor grades, as well as the recent advances in deep learning for image analysis and classification.
- We introduce a new loss function for the ordinal classification problem that can further improve the accuracy and generalizability of the model. The concept and computation of the loss function is rather simple but effective in joint learning.
- We systematically evaluate the performance of the proposed method for cancer grading in pathology images using multiple sets of cancer images from two different organs. The proposed method outperforms 11 other competing methods.
- For each organ, we employ tissue image datasets that underwent different acquisition and processing procedures. Under various conditions, the proposed method is able to generalize to unseen images.

2. Related work

2.1. Pathology image analysis and grading

In recent years, machine learning techniques have been widely applied to pathology image grading and analysis. Conventional models mainly sought to mimic the diagnostic process of the human experts. For instance, the appearance and formation of glandular structure and shape and arrangement of cells and nuclei are known to be associated with tumor grades. Typical approaches utilized hand-crafted features, including color histogram (Gorelick et al., 2013), morphology (Kwak et al., 2011), wavelet transform (Tabesh et al., 2007), gray level co-occurrence matrix (GLCM) (Kather et al., 2016; Doyle et al., 2012), local binary pattern (Kather et al., 2016), and Gabor filters (Kather et al., 2016; Doyle et al., 2012) to extract useful image features. These image features, in combination with conventional machine learning algorithms, were used to identify tumor cells or tumor grades.

Moreover, deep learning, especially convolutional neural networks (CNNs), known as an efficient, automatic feature extractor with minimal human processing, has been applied to various problems in pathology images (Litjens et al., 2017); for instance, detection of mitosis (Cireşan et al., 2013) and invasive ductal carcinoma (Cruz-Roa et al., 2014) in the breast, detection of cancer in the prostate (Kwak and Hewitt, 2017; Turki and Wei, 2018; Krawczyk et al., 2016), and classification of tissue sub-types in the colon (Kather et al., 2016). To further improve such deep learning models, a number of approaches have been proposed. Some adopted an approach of multi-scale image analysis. In (Duong et al., 2019), a network that can extract and utilize features from multiple scales was proposed to detect cancers in the prostate. In (Shaban et al., 2020), the local representation of an image is encoded into high dimensional features. Utilizing the spatial organization of the features, the grade of colorectal cancers was determined. Some others proposed a graph-based method. In (Zhou et al., 2019), a cell-graph CNN for colorectal cancer grading was built from segmented nuclei to capture both cell-level information and morphological information of the gland. Along with these developments, another approach, so called multi-task learning, has been also applied to pathology image analysis.

2.2. Multi-task learning

Multi-task learning is a branch of machine learning that jointly learns multiple related tasks at the same time to leverage the useful information obtained from the multiple tasks and to improve generalizability of the model (Caruana, 1997). Multi-task learning can be useful for the case where training data is limited like in medical imaging and has shown an improved learning capability in comparison to single-task learning (Zhang and Yang, 2017). Although there are various variants of multi-task learning approaches, many of them adopt a single target task with one or more other tasks as auxiliary constraints (Zhang et al., 2014; Girshick, 2015). Moreover, a majority of these approaches conduct classification and segmentation tasks together. For example, in (Mehta et al., 2018), breast tissue segmentation was performed in conjunction with tissue classification by using a modified U-Net (Ronneberger et al., 2015). In (Graham et al., 2019), a network that enables simultaneous nuclei segmentation and classification in

Table 1
Details of colorectal and prostate tissue datasets.

Type	Status	Training	Validation	Testing I	Testing II
Colorectal	Benign	773	374	453	27986
Tissue	WD	1866	264	192	8394
	MD	2997	370	738	61985
	PD	1397	234	205	11895
Prostate	Benign	2076	666	127	1284
Tissue	Grade 3	6303	923	2121	5852
	Grade 4	4541	573	1784	9682
	Grade 5	2383	320	387	248

pathology images was proposed. In this work, we propose to conduct both categorical and ordinal classifications in order to exploit the ordering of tumor grades and to improve the overall classification performance. To the best of our knowledge, there has been no such prior work for cancer grading in pathology images.

2.3. Ordinal classification

Ordinal classification or regression aims to predict class labels on an ordinal scale. Most of such methods have been built based upon well-studied classification algorithms, including SVM (Herbrich et al., 1999; Shashua and Levin, 2003) and decision tree (Frank and Hall, 2001). Recently, CNN-based approaches have been proposed. For instance, age estimation was formulated as a series of binary classification sub-problems that were conducted by a multiple output CNN (Niu et al., 2016). An end-to-end CNN model was also proposed for both depth estimation (Fu et al., 2018) and age estimation (Cao et al., 2020). Moreover, a CNN model was built for prostate cancer detection and grading in MRI that converts class labels into an ordinal vector (Cao et al., 2019). In (De Vente et al., 2020), an approach of a soft-label ordinal regression was proposed for prostate cancer detection and grading in MRI. However, ordinal classification has not been well-studied in the context of pathology image classification.

3. Methods

3.1. Dataset

3.1.1. Colorectal tissue dataset

We collected three sets of colorectal tissue samples from Kangbuk Samsung Hospital. The first and second set contains 3 whole slide images (WSIs) from 3 patients and 6 colorectal tissue microarrays (TMAs) from 340 patients, respectively, acquired between 2006 and 2008. The 3 WSIs and 6 TMAs were scanned at 40x magnification using an Aperio digital slide scanner (Leica Biosystems). The size of the WSIs is ~100,000 × 80,000 pixels (0.2465 μm×0.2465 μm per pixel). The size of a tissue core in the TMAs is ~8,400 × 8,400 pixels (0.2518 μm×0.2518 μm per pixel). The third set includes 45 WSIs from 45 patients, acquired between 2016 and 2017. Using a NanoZoomer digital slide scanner (Hamamatsu Photonics K.K.), the 45 WSIs, of size ~100,000 × 100,000 pixels (0.2253 μm×0.2253 μm per pixel), were digitized at 40x magnification. Upon the pathologic review of the tissue samples (K. Kim and B. Song), we obtained benign (BN) and three cancer ROIs, including well-differentiated (WD) tumor, moderately-differentiated (MD) tumor, and poorly-differentiated (PD) tumor.

From BN and three cancer ROIs, a number of image patches were obtained. Excluding the image patches with a large luminal and/or unannotated regions, the remaining image patches were used as the training and test datasets. Table 1 shows the details of the colorectal training and test datasets. 9857 image patches (1600 BN, 2322 WD, 4105 MD, and 1830 PD) of size 1024x1024 pixels (~258 μm×258 μm) were gained from the first and second sets. The image patches were divided into three distinct datasets: 1) training dataset (C_{Train}), 2) validation dataset ($C_{Validation}$), and 3) test dataset I (C_{TestI}). From the third set, 110170 image patches (27986 BN, 8394 WD, 61985 MD, and 11985 PD) of size 1144 × 1144 pixels (~258μm×258 μm) were acquired and designated as the test dataset II (C_{TestII}).

3.1.2. Prostate tissue dataset

Two sets of prostate tissue samples and annotations that are publicly available were employed. The first set was obtained from the Harvard dataverse (https://dataverse.harvard.edu/), including 5 TMAs with 886 tissue cores. Each core has a size of 3,100 × 3,100 pixels (0.23 μm×0.23 μm per pixel). The 5 TMAs were digitized at 40x magnification at the University Hospital Zurich using a NanoZoomer digital slide scanner (Hamamatsu Photonics K.K.). Similar to the colorectal tissue datasets, BN and three cancer ROIs (Gleason grade 3, grade 4, and grade 5) were delineated by an experienced pathologist (Arvaniti et al., 2018). The first 4 TMAs were used to generate the training dataset (P_{Train}) and the validation dataset ($P_{Validation}$). The remaining TMA was utilized to produce the prostate test dataset I (P_{TestI}). Using the code from (Arvaniti et al., 2018), 15141 image patches, in total, of size 750 × 750 pixels (172.5 μm×172.5 μm) were generated as excluding the patches with a large portion of luminal and/or unannotated regions. The training dataset is composed of 2076, 6303, 4541, and 2383 image patches for BN, grade 3, grade 4, and grade 5, respectively, while P_{TestI} includes 127 BN, 1602 grade 3, 2121 grade 4, and 351 grade 5 image patches.

The second set was acquired from the training set of Gleason2019 challenge (https://gleason2019.grand-challenge.org/). This involves a set of 244 prostate tissue cores that were digitized at 40x magnification using an Aperio digital slide scanner (Leica Biosystems) at the Vancouver Prostate Centre. Each tissue core has a size of 4,400 × 4,400 pixels (0.25 μm×0.25 μm per pixel). Tissue cores were annotated by 6 pathologists (Nir et al., 2018) and the ground truth label was determined by Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm (Warfield et al., 2004). Applying the same strategy with the first set, 17,066 image patches (1284 BN, 5852 grade 3, 9682 grade 4, and 248 grade 5), of which each has a size of 690 × 690 pixels (172.5 μm×172.5 μm), were acquired, forming the test dataset II (P_{TestII}). The details of the prostate tissue datasets are available in Table 1.

3.2. Problem formulation

Suppose that we are given a set of pathology image-ground truth label pairs $\{x_i, y_i\}_{i=1}^{N}$ where $x_i \in \mathbb{R}^{w\times h\times c}$ is the ith pathology image, y_i is the ground truth label ($y_i \in$ {BN, WD, MD, PD} if x_i is a colorectal tissue or $y_i \in$ {BN, grade 3, grade 4, grade 5} if x_i is a prostate tissue), and N is the number of image-ground truth label pairs. w, h, and c denote the width, height, and the number of channels, respectively. Let a shared feature extractor be f. Then, f maps an input pathology image x_i to a high-dimensional feature space $v_i \in \mathbb{R}^{w'\times h'\times c'}$ as follows: $f(x_i) = W_f^T x_i + b_F = v_i$ where $w' \ll w, h' \ll h, c' \ll c$. Let multi-task learning tasks be $\{T_j\}_{j=1}^{M}$ where T_j is the jth task and M is the number of tasks. The learning function for the task T_j is defined as $g_j(v_i) = W_j^T v_i + b_j = z_{j,i} \in \mathbb{R}^{D_j}$ where D_j is the cardinality of the output of the function g_j. Multi-task learning can be formulated as follows:

$$\min_{W_f, W_{1,\dots,M}} \sum_{i=1}^{N}\sum_{j=1}^{M} \mathcal{L}(g_j(f(x_i; W_f); W_j)) \quad (1)$$

where $\mathcal{L}$ denotes the loss function.

3.3. Joint categorical and ordinal learning

For joint categorical and ordinal learning, we propose a network that consists of three major parts: 1) a shared feature extractor f; 2) a categorical classification branch g_c; 3) an ordinal classification branch g_o. The network is illustrated in Fig. 1. Given an input pathology image, f generates a high dimensional feature map, which is, in turn, fed into g_c and g_o independently. The dimensionality of the output of g_c is determined by the cardinality of the ground truth class labels D_c (here, D_c is 4), of which each denotes the probability of an input image to be assigned to the corresponding class label. The output of g_o is a continuous value that is assumed to be relevant to the degree of tumor aggressiveness, i.e., tumor grades. The objective of both g_c and g_o is to predict the tumor grade but in a different manner.

3.3.1. Shared feature extractor

The shared feature extractor f conducts a sequence of convolution, pooling, activation, and normalization operations to extract a high-level feature map which serves as the input to the following specific tasks. In this study, a well-known neural network architecture, EfficientNet-B0 (Tan and Le, 2019), is adopted to build the shared feature extractor. Based upon the compound scaling method, EfficientNet-B0 (Tan and Le, 2019) scales the width, depth, and resolution of the network using a set of fixed scaling factors. It adopts one convolution layer and 16 mobile inverted bottleneck convolution (MBConv) blocks of MobileNet-V2 (Sandler et al., 2018). Similar to MobileNet-V2 (Sandler et al., 2018), each MBConv block consists of one 1×1 pointwise convolution, one depthwise convolution, and one 1×1 pointwise convolution. Depthwise convolutions are implemented with either a 3×3 or 5×5 kernel.

3.3.2. Categorical classification branch

The categorical classification branch g_c contains a batch normalization (BN), an average pooling layer (AvgPool), and a fully connected layer (FC) with 4 neurons. Finally, a softmax function is employed to transform the outputs into the probability distribution as follows:

$$p_i^k = \frac{\exp\left(z_{c,i}^k\right)}{\sum_{j=1}^{D_c} \exp\left(z_{c,i}^j\right)} \tag{2}$$

where p_i^k and $z_{c,i}^k$ represent the probability of an inputbelonging to a class k and the output of the kth neuron in the categorical classification branch g_c, respectively, for an input pathology image x_i.

3.3.3. Ordinal classification branch

Similar to the categorical classification branch, the ordinal classification branch g_o has a series of BN-AvgPool-FC. The FC only contains one neuron that generates an output in the range of 0 to 3. Here, 0, 1, 2, and 3 are corresponding to BN and three cancer grades (WD, MD, and PD or grade 3, grade 4, and grade 5), respectively. The higher the output is, the more abnormal (or aggressive) the pathology image is.

3.4. Loss functions

The proposed model includes three distinct sets of weights: W_f, W_c, and W_o, of which each denotes the weight of the shared feature extractor, categorical classification branch, and ordinal classification branch, respectively. In order to optimize these weights, we employ a number of loss functions that are tailored to the specific tasks as described below.

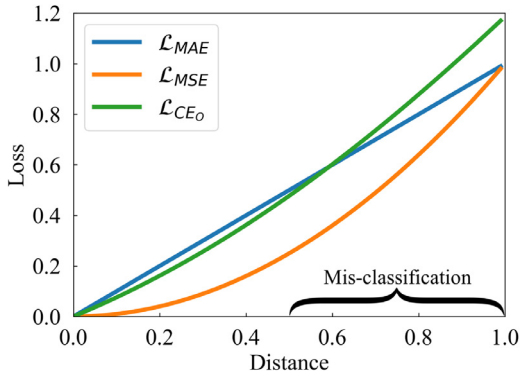

Fig. 2. Loss functions for ordinal classification. CEo permits a clear distinction between correctly classified examples and mis-classified examples.

3.4.1. Categorical classification loss

Cross entropy (CE) loss has been widely used for the classification task. This loss measures the total entropy between the true and predicted probability distribution as follows:

$$\mathcal{L}_{CE}(y, p) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{D_c} y_i^k \log(p_i^k) \tag{3}$$

where y and p are the ground truth label and the output of the categorical classification branch, respectively. Hence, $\mathcal{L}_{CE}$ is computed for the output of the categorical classification branch and is utilized to optimize the weights W_f and W_c.

3.4.2. Ordinal classification loss

For the ordinal classification branch, we define a mean square error (MSE) loss as:

$$\mathcal{L}_{MSE}(y, z_o) = \frac{1}{N}\sum_{i=1}^{N} d_i^2 \tag{4}$$

where $d_i = y_i - z_{o,i}$ is the distance between the ground truth label y_i and the output of the ordinal classification branch $z_{o,i}$. Computing $\mathcal{L}_{MSE}$ for the output of the ordinal classification branch, we optimize the weights W_f and W_o. In addition to $\mathcal{L}_{MSE}$, a mean absolute error (MAE) loss is also considered for the ordinal classification branch. MAE is the mean absolute distance between true and predicted values:

$$\mathcal{L}_{MAE}(y, z_o) = \frac{1}{N}\sum_{i=1}^{N} |d_i| \tag{5}$$

3.4.3. Ordinal cross entropy loss

Fig. 2 shows the $\mathcal{L}_{MSE}$ and $\mathcal{L}_{MAE}$ curves. As we assess these curves, we note that those samples that are correctly classified ($d_i \ll 0.5$) have an insignificant magnitude of loss. For such samples, the model may not strive to further improve the performance. Even for those samples that are closer to the adjacent label ($d_i \gg 0.5$), the magnitude of loss is relatively small. There must be a clearer distinction between the two cases where one example is correctly classified and the other is mis-classified. Ordinal cross entropy is designed to address these issues and to further improve the learning capability of the ordinal classification branch. We propose to convert the distance between the ground truth label and the output of the ordinal classification branch into probability measures by using a softmax function:

$$q_i^k = \frac{\exp\left(-|k - z_{o,i}|\right)}{\sum_{j=1}^{D_c} \exp\left(-|j - z_{o,i}|\right)} \tag{6}$$

where q_i^k refers to the probability of an input pathology image x_i that belongs to a class k and $z_{o,i}$ denotes the output of the ordinal

classification branch for x_i. Then, we compute the CE loss for the ground truth label and the probability measures for the output of the ordinal classification branch, which is designated as $\mathcal{L}_{CE_O}$, as follows:

$$\mathcal{L}_{CE_O} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{D_c} y_i^k \log(q_i^k). \tag{7}$$

As shown in Fig. 2, the curve of $\mathcal{L}_{CE_O}$ is not only able to make a clearer distinction between the correctly classified examples and mis-classified samples but also gives a non-trivial magnitude of loss for those samples that are close to the ground truth label, leading to an improved accuracy and robustness of the model. The full loss function $\mathcal{L}$ for the proposed method is defined as:

$$\mathcal{L} = \lambda_C\mathcal{L}_C + \lambda_R\mathcal{L}_R + \lambda_O\mathcal{L}_O \tag{8}$$

where $\mathcal{L}_C$, $\mathcal{L}_R$, and $\mathcal{L}_O$ denote categorical classification loss ($\mathcal{L}_{CE}$), ordinal classification loss, and ordinal cross entropy loss ($\mathcal{L}_{CE_O}$), respectively. The ordinal classification loss can be either $\mathcal{L}_{MSE}$ or $\mathcal{L}_{MAE}$. Hence, two joint learning models are considered: 1) $\mathcal{M}_{MSE-CE_O}$ and 2) $\mathcal{M}_{MAE-CE_O}$, depending on the choice of the ordinal classification loss. λ_C, λ_R, and λ_O are weighting factors for the associated loss functions.

4. Experiments

4.1. Implementation details

4.1.1. Data augmentation

Several data augmentation techniques are adopted during the training phase as follows: 1) a random horizontal flip; 2) a random vertical flip; 3) an affine transformation with a random rotation in the range [-45°, 45°] and shear in the range [-16°, 16°]; 4) image blurring with a Gaussian, average, or median filter; 5) a random additive Gaussian noise or drop out up to 10% of pixels; 6) a random color change in hue, saturation, and contrast in the range [-20, 20]. 1), 2), and 3) are applied to every image patch. 4), 5), and 6) are applied to an image patch with a 50% chance. All the augmentation techniques are implemented using Aleju library (https://github.com/aleju/imgaug).

4.1.2. Training details

We train all the models in this work with Adam optimizer using default parameter values ($\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 1.0\text{e}^{-8}$) for 60 epochs and *cosine annealing warm restarts schedule* (Loshchilov and Hutter, 2016) with initial learning rate of 1.0e^{-3} and $T_0 = 20$ (learning rate restarts after 20 epochs). The pre-trained weights on ImageNet dataset are utilized to initialize EfficientNet-B0. The layers in the categorical and ordinal classification branches are initialized using K.He method. Following data augmentation, colorectal tissue patches are resized to 512 × 512 pixels while prostate tissue patches are resized to 375 × 375 pixels and then center cropped to 350 × 350 pixels. λ_C, λ_R, and λ_O are set to 1. All the models are implemented on PyTorch platform and executed on a workstation with four TITAN XP GPUs.

4.2. Comparative experiments

To assess the effectiveness of the proposed method, a series of comparative experiments are conducted. The proposed method is compared with three types of single task models (classification, regression, and ordinal regression) and one previously published multi-task learning model. Ablation experiments are also conducted to further analyze and investigate the proposed method.

4.2.1. Classification model

The identical shared feature extractor f and categorical classification branch g_c with the proposed model are employed to form a baseline classification model. EfficientNet-B0 is used as the shared feature extractor. To train the classification model, two loss functions are independently adopted, including $\mathcal{L}_{CE}$ and focal loss ($\mathcal{L}_{Focal}$) (Lin et al., 2017). Trained with $\mathcal{L}_{CE}$ and $\mathcal{L}_{FOCAL}$, we obtain two classification models: 1) $\mathcal{C}_{CE}$ and 2) $\mathcal{C}_{FOCAL}$, respectively.

4.2.2. Regression model

Similarly to the classification model, we build a regression model using the shared feature extractor f (EfficientNet-B0) and the ordinal regression branch g_o. Utilizing two popular loss functions, i.e., $\mathcal{L}_{MAE}$ and $\mathcal{L}_{MSE}$, and a soft-label regression loss ($\mathcal{L}_{SL}$) in (De Vente et al., 2020), three regression-based models are acquired, designated as $\mathcal{R}_{MAE}$, $\mathcal{R}_{MSE}$, and $\mathcal{R}_{SL}$, respectively. $\mathcal{L}_{SL}$, built for prostate cancer detection and grading in bi-parametric MRI, utilizes the normalization function to convert the class label to the ordinal label. We note that the three regression models ($\mathcal{R}_{MAE}$, $\mathcal{R}_{MSE}$, and $\mathcal{R}_{SL}$) use the regression loss functions only.

4.2.3. Ordinal regression model

Three recent ordinal regression models are employed and compared to the proposed model: 1) deep ordinal regression network ($\mathcal{O}_{DORN}$) (Fu et al., 2018), 2) consistent rank logits ($\mathcal{O}_{CORAL}$) (Cao et al., 2020), and 3) focal loss for ordinal regression $\mathcal{O}_{FOCAL}$ (Cao et al., 2019). $\mathcal{O}_{DORN}$ is built for monocular depth estimation. It utilizes a spacing-increasing discretization (SID) strategy that discretizes a given interval in log space to reduce the loss in the regions with a larger depth, and thus the model focuses more on the regions with a smaller depth. As it is applied to the colorectal and prostate tissue datasets, the larger depth corresponds to the higher cancer grade. $\mathcal{O}_{CORAL}$ is proposed to achieve the rank consistency among the class labels. It extends the class label (or rank) into binary labels and enforces rank-monotonicity for consistent predictions. In (Cao et al., 2019), $\mathcal{O}_{FOCAL}$, which adopts focal loss for ordinal regression, is built for cancer detection and Gleason score prediction in mp-MRI. They use focal loss for ordinal regression along with $\mathcal{L}_{MSE}$ to maximize mutual findings between single and multiple components of mp-MRI. As it is used in this study, focal loss is applied to ordinal regression alone.

To make fair comparison, we implement these three methods using EfficientNet-B0 i.e., the same feature extractor with the proposed model. This allows us to make a fair comparison between different approaches regardless of the effect of the feature extractor.

4.2.4. Multi-task model

Multi-task deep model with margin ranking loss ($\mathcal{M}_{MTMR}$) (Liu et al., 2019a) is an approach that explicitly integrates lung nodule classification with eight attributes (subtlety, internal structure, calcification, sphericity, margin, spiculation, lobulation, and texture) score regression. Built based upon a Siamese network, the model contains two modules - the classification module and regression module. The classification module is optimized by utilizing $\mathcal{L}_{CE}$ and the margin ranking loss, which aims at capturing the relationship between pairs of samples. $\mathcal{L}_{MSE}$ is used to adjust the regression module. In our experiments, EfficientNet-B0 is adopted as the feature extractor in $\mathcal{M}_{MTMR}$, similar to the ordinal regression-based models. The regression module is modified to predict the ordinal label of tissues.

4.2.5. Ablation experiments

To investigate the effect of the proposed $\mathcal{L}_{CE_O}$ for cancer grading, ablation experiments are performed. The two proposed models ($\mathcal{M}_{MSE-CE_O}$ and $\mathcal{M}_{MAE-CE_O}$) are separately employed, trained, and tested without $\mathcal{L}_{CE_O}$, resulting in $\mathcal{M}_{MSE}$ and $\mathcal{M}_{MAE}$, respectively.

Table 2
Results of Colorectal Cancer Classification on C_{TestI} and C_{TestII}.

Method	C_{TestI}				C_{TestII}			
	Acc (%)	Acc_G (%)	F1	κ_w	Acc (%)	Acc_G (%)	F1	κ_w
$\mathcal{O}_{DORN}$ (Fu et al., 2018)	79.2	70.9	0.618	0.889	78.5	71.7	0.564	0.872
$\mathcal{O}_{CORAL}$(Cao et al., 2020)	81.9	79.6	0.792	0.912	73.5	67.9	0.695	0.852
$\mathcal{O}_{FOCAL}$(Cao et al., 2019)	86.2	80.9	0.837	0.935	73.4	67.6	0.694	0.859
$\mathcal{C}_{CE}$	84.9	78.9	0.815	0.924	75.1	67.6	0.703	0.854
$\mathcal{C}_{FOCAL}$(Lin et al., 2017)	86.1	80.7	0.820	0.934	76.9	70.7	0.703	0.857
$\mathcal{R}_{SL}$ (De Vente et al., 2020)	64.9	50.8	0.585	0.813	57.9	44.4	0.531	0.788
$\mathcal{R}_{MAE}$	86.0	80.4	0.832	0.935	75.3	67.4	0.706	0.874
$\mathcal{R}_{MSE}$	86.2	80.8	0.826	0.936	74.9	68.4	0.689	0.870
$\mathcal{M}_{MTMR}$(Liu et al., 2019a)	85.9	80.3	0.833	0.931	72.8	66.6	0.691	0.826
$\mathcal{M}_{MAE}$	86.7	81.4	0.835	0.938	75.5	68.0	0.706	0.868
$\mathcal{M}_{MSE}$	87.0	81.9	0.839	0.940	75.3	68.2	0.700	0.857
$\mathcal{M}_{MAE-CE_O}$(Ours)	87.7	82.7	0.843	0.940	**80.3**	**74.0**	**0.744**	**0.891**
$\mathcal{M}_{MSE-CE_O}$(Ours)	**88.4**	**83.7**	**0.854**	**0.943**	78.1	72.0	0.729	0.870

4.3. Evaluation metrics

To assess the performance of cancer grading, 4 evaluation metrics are employed: 1) accuracy (Acc) is the fraction of correctly classified tissue samples over all tissue samples; 2) cancer classification accuracy (Acc_G) is an accuracy of classifying cancer samples into different cancer grades; 3) macro-averaged F1 (F1) is the harmonic mean of the averaged precision and averaged recall; 4) quadratic weighted kappa (κ_w) (Cohen, 1968) is defined as:

$$\kappa_w = 1 - \frac{\sum_i^K \sum_i^K w_{ij} x_{ij}}{\sum_i^K \sum_i^K w_{ij} m_{ij}} \tag{9}$$

where $w_{ij} = \frac{(i-j)^2}{(K-1)^2}$, m_{ij} is the expected proportion for the predicted class i and ground truth class j, x_{ij} is the proportion of C_{ij}, and C is a confusion matrix.

4.4. Visual illustration

We visually illustrate the result of cancer grading by the proposed models and the other classification, regression, ordinal regression, and multi-task models. Using each of the models, cancer classification is performed on the entire tissue images in TMAs and WSIs, not just on the selected patches, to further assess the generalizability of the models. For a tissue image, we slide a rectangular window using the step size of one-eighth and a half of the window size for WSIs and TMAs, respectively. The window size is set to ∼258 μm× 258 μm for colorectal tissue images in TMAs (1024 pixels ×1024 pixels) and WSIs (1144 pixels×1144 pixels) and to 172.5 μm×172.5 μm for prostate tissue images from the first set (750 pixels × 750 pixels; Harvard dataverse) and the second set (690 pixels × 690 pixels; Gleason2019 challenge), respectively.

5. Experimental results

5.1. Colorectal cancer classification

Table 2 and Fig. 3a demonstrate the results of colorectal cancer classification on C_{TestI}. We trained the proposed model and other competing models on C_{Train} and tested on C_{TestI}. In this experiment, the proposed models successfully classified the colorectal tissue samples into benign and 3 cancer grades. $\mathcal{M}_{MSE-CE_O}$ achieved the best performance over the 4 evaluation metrics, and $\mathcal{M}_{MAE-CE_O}$ was the second best model. In the comparative experiments, the classification performance varied among the single- and multi-task models. Among the single-task models, the two regression-based models ($\mathcal{R}_{MAE}$ or $\mathcal{R}_{MSE}$) were superior to both the classification-based models ($\mathcal{C}_{CE}$ or $\mathcal{C}_{FOCAL}$) and the ordinal regression-based models ($\mathcal{O}_{DORN}$, $\mathcal{O}_{CORAL}$, or $\mathcal{O}_{FOCAL}$). Meanwhile, $\mathcal{R}_{SL}$ was substantially inferior to all others. This clearly shows the importance of the choice of the objective function during training. As for the multi-task models, even though the performance of $\mathcal{M}_{MTMR}$ was similar or slightly inferior to the regression-based models, the other multi-task models ($\mathcal{M}_{MAE}$ or $\mathcal{M}_{MSE}$) outperformed the single-task models and $\mathcal{M}_{MTMR}$. Moreover, the addition of $\mathcal{L}_{CE_O}$, i.e., the proposed models, further improved the overall performance. This demonstrates the effectiveness of multi-task learning as well as the proposed $\mathcal{L}_{CE_O}$ in cancer grading, i.e., multi-class classification.

Fig. 4 shows the probability maps, generated by $\mathcal{M}_{MSE-CE_O}$, for colorectal tissue samples in TMAs. Taking the sliding window scheme, the proposed model was able to predict histopathologic class labels for the entire tissue cores, even though the model was trained on image patches. Moreover, the proposed model was successfully applied to tissue cores that possess multiple ROIs with distinct histopathologic class labels. This suggests that the proposed model is capable of coping with complex tissue structures.

5.2. Colorectal cancer classification on independent test dataset

Trained on C_{Train}, the proposed models and other competing models were tested on C_{TestII}, which was obtained at different period of time using different digital scanners from C_{Train} and C_{TestI}. Hence, this experiment could show the generalizability of the models. The results are available in Table 2 and Fig. 3b. Similar to the observations in the above experiment on C_{TestI}, the proposed models ($\mathcal{M}_{MAE-CE_O}$ and $\mathcal{M}_{MSE-CE_O}$) achieved the best and second best performance on C_{TestII}, respectively, suggesting that the proposed models are able to generalize to the independent or unseen tissue images. However, the results of other single- and multi-task models were inconsistent with those on C_{TestI}, i.e., the performance of the models is sensitive to the dataset. For example, the ordinal regression-based model, $\mathcal{O}_{DORN}$, obtained relatively high Acc, Acc_G, and κ_w but the second worst F1; the performance of $\mathcal{M}_{MTMR}$ was, in general, inferior to other single- and multi-task models. Moreover, $\mathcal{M}_{MAE}$ and $\mathcal{M}_{MSE}$ were comparable to other single- and multi-task models but there was a substantial performance drop in comparison to the proposed models. This emphasizes the importance of the proposed $\mathcal{L}_{CE_O}$ in improving the generalizability of the models in cancer grading.

The prediction results by the proposed models and other competing models are visualized using the WSIs on C_{TestII}. The prediction maps are illustrated in Fig. 5. The prediction maps of the pro-

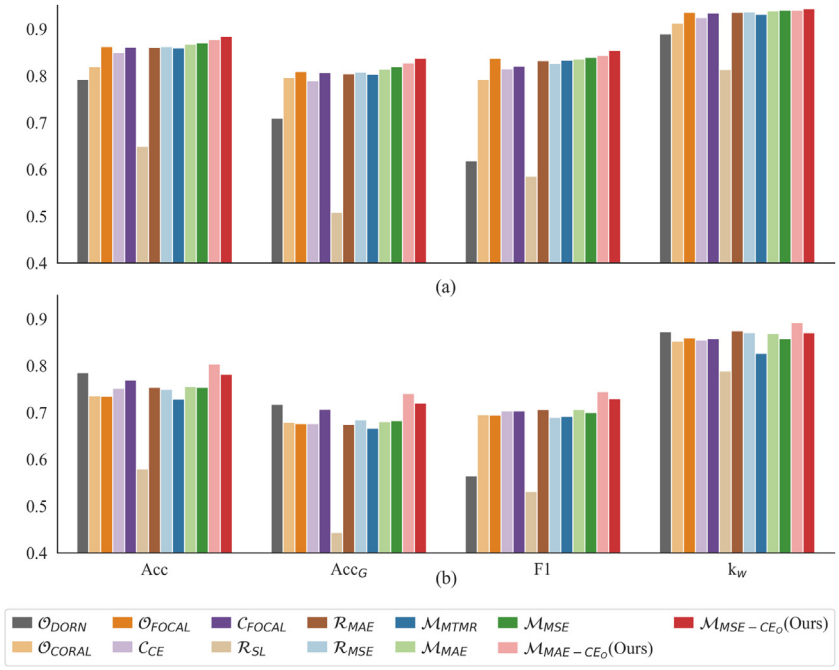

Fig. 3. Plots of classification results on colorectal datasets. (a) C_{TestI} and (b) C_{TestII}.

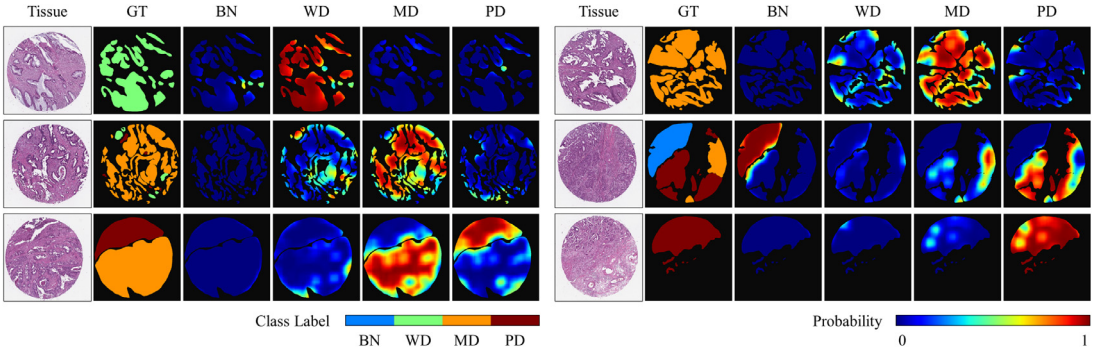

Fig. 4. Probability maps for colorectal tissue samples on C_{TestI}. Given a tissue image, the ground truth (GT) map and the four probability maps for BN, WD, MD, and PD classes, respectively, are shown.

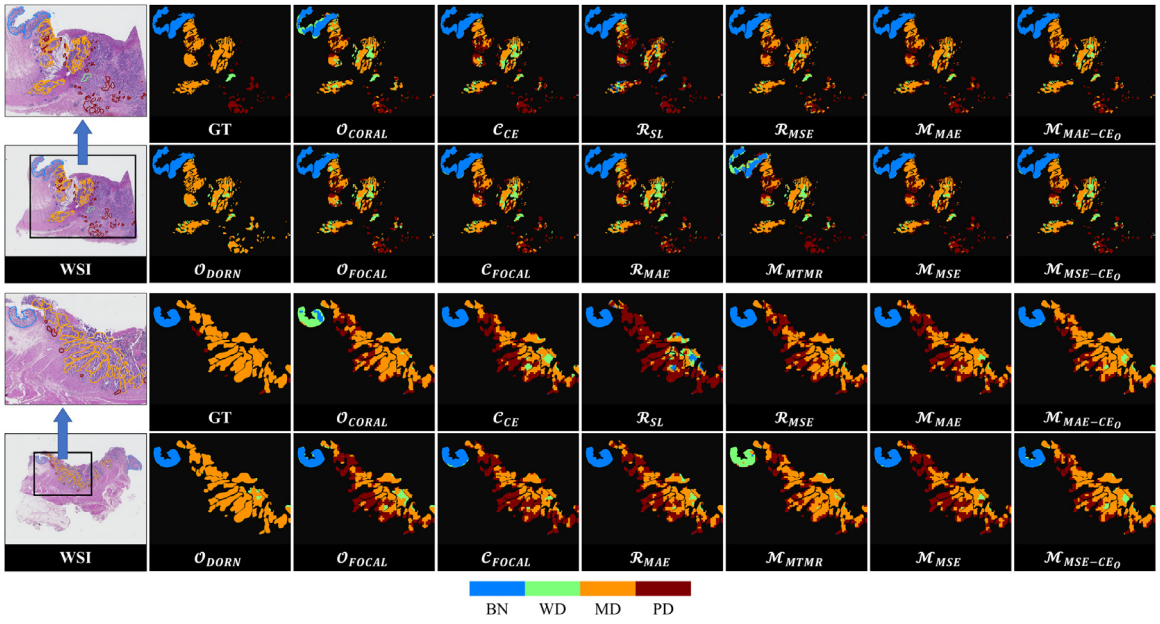

Fig. 5. Prediction maps for colorectal tissue samples on C_{TestII}. Given a WSI image, the ground truth (GT) map and the prediction maps, generated by the proposed method and other competing methods, are illustrated.

Table 3
Results of Prostate Cancer Classification on P_{TestI} and P_{TestII}.

Method	P_{TestI}				P_{TestII}			
	Acc (%)	Acc_G (%)	F1	κ_w	Acc (%)	Acc_G (%)	F1	κ_w
Arvaniti et al. (2018)	-	-	-	0.550	-	-	-	-
$\mathcal{O}_{DORN}$ (Fu et al., 2018)	0.681	0.689	0.484	0.521	**80.4**	**82.0**	0.584	**0.723**
$\mathcal{O}_{CORAL}$(Cao et al., 2020)	0.674	0.682	0.587	0.549	74.3	76.6	0.602	0.643
$\mathcal{O}_{FOCAL}$(Cao et al., 2019)	0.677	0.681	0.625	0.597	76.3	77.4	0.646	0.696
$\mathcal{C}_{CE}$	0.658	0.666	0.592	0.572	73.3	75.4	0.606	0.665
$\mathcal{C}_{FOCAL}$(Lin et al., 2017)	0.671	0.680	0.600	0.599	71.1	72.7	0.589	0.657
$\mathcal{R}_{SL}$ (De Vente et al., 2020)	0.311	0.294	0.299	0.508	29.1	23.7	0.253	0.521
$\mathcal{R}_{MAE}$	0.671	0.672	0.623	0.611	72.5	72.5	0.629	0.654
$\mathcal{R}_{MSE}$	0.671	0.675	0.626	0.606	72.1	73.3	0.589	0.664
$\mathcal{M}_{MTMR}$(Liu et al., 2019a)	0.673	0.683	0.588	0.594	73.7	76.2	0.598	0.665
$\mathcal{M}_{MAE}$	0.659	0.662	0.606	0.572	72.5	72.6	0.616	0.676
$\mathcal{M}_{MSE}$	0.657	0.664	0.600	0.583	75.3	77.1	0.623	0.688
$\mathcal{M}_{MAE-CE_O}$(Ours)	**0.696**	**0.702**	**0.633**	0.616	77.5	79.3	0.651	0.706
$\mathcal{M}_{MSE-CE_O}$(Ours)	0.688	0.695	0.622	**0.617**	**77.7**	**79.5**	**0.658**	**0.707**

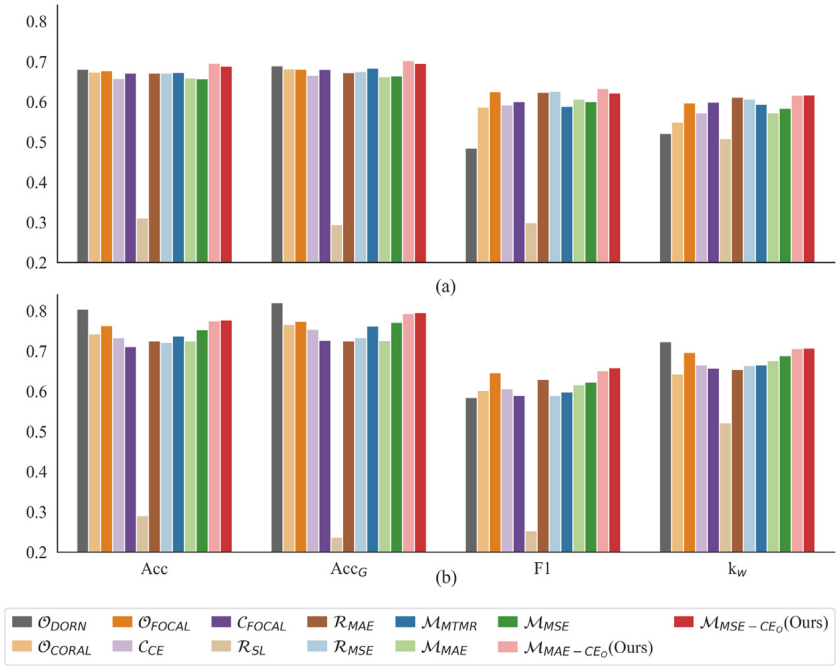

Fig. 6. Plots of classification results on prostate datasets. (a) P_{TestI} and (b) P_{TestII}.

posed models are well corresponding to the ground truth maps. This visually confirms that the proposed models, trained on the TMA-dominant dataset, are generalizable to unseen WSIs under different acquisition and processing conditions. Other competing models are partially successful in classifying cancer grades. For example, $\mathcal{O}_{DORN}$ misses the entire PD, $\mathcal{O}_{CORAL}$, and $\mathcal{M}_{MTMR}$ tend to mis-classify BN as WD, and the rest, in general, overpredicts PD.

5.3. Prostate cancer classification

The results of prostate cancer classification on P_{TestI} are illustrated in Table 3 and Fig. 6a. Similar to colorectal cancer classification, the proposed $\mathcal{M}_{MAE-CE_O}$ and $\mathcal{M}_{MSE-CE_O}$ were the top two models among all the models under consideration. However, the two multi-task models $\mathcal{M}_{MAE}$ and $\mathcal{M}_{MSE}$ were inferior to that of the single-task models (except for $\mathcal{R}_{SL}$) and multi-task models with additional loss functions ($\mathcal{M}_{MTMR}$, $\mathcal{M}_{MAE-CE_O}$, and $\mathcal{M}_{MSE-CE_O}$).

Furthermore, the proposed models substantially outperformed the previously published results (Arvaniti et al. (2018)) on the same datasets; 0.067 increase in κ_w on P_{TestI}. We note that the exact same procedure with Arvaniti et al. (2018) was adopted to generate the image patches for both training and test datasets. Hence, no selection or sampling bias affects the results.

Fig. 7 demonstrates the prediction maps for P_{TestI}. The proposed models, in general, are able to generate prediction maps that are well matched with the ground truth maps. However, the quality of the prediction maps, generated by other models, varies depending on the tissue types. For instance, $\mathcal{R}_{SL}$ showed the worst performance in grade 3 and grade 4 while $\mathcal{O}_{DORN}$ only performed well in benign and low cancer grades.

5.4. Prostate cancer classification on independent test dataset

To evaluate the generalizability of the prostate cancer classification models, the same models tested on P_{TestI} were applied to P_{TestII}, which was obtained and processed at a different institute. Table 3 shows the results of prostate cancer classification on P_{TestII}. In regard with Acc, Acc_G, and κ_w, the proposed models, $\mathcal{M}_{MSE-CE_O}$ and $\mathcal{M}_{MAE-CE_O}$, were the second and third best models, respectively, meanwhile the ordinal regression-based model $\mathcal{O}_{DORN}$

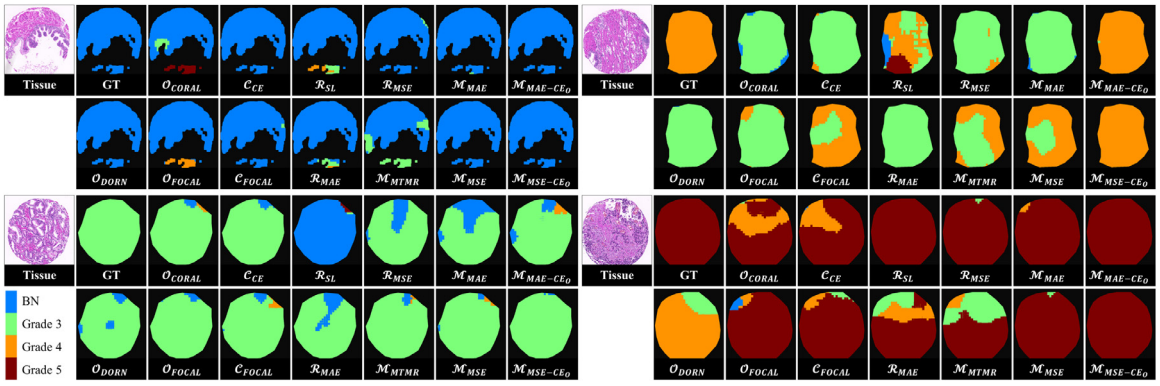

Fig. 7. Prediction results for prostate tissue samples on P_{TestI}. Given a tissue image, the ground truth (GT) map and the prediction maps for the proposed models and the other competing models are shown.

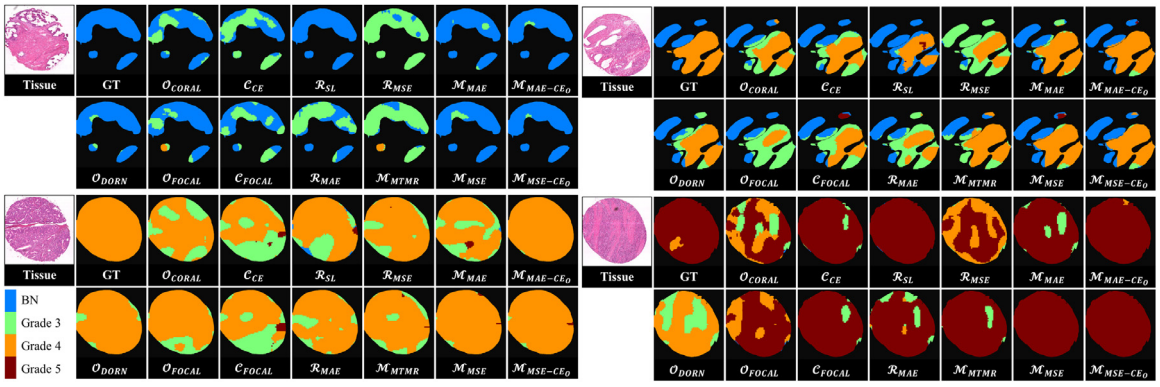

Fig. 8. Prediction results for prostate tissue samples on P_{TestII}. Given a tissue image, the ground truth (GT) map and the prediction maps for the proposed models and the other competing models are shown.

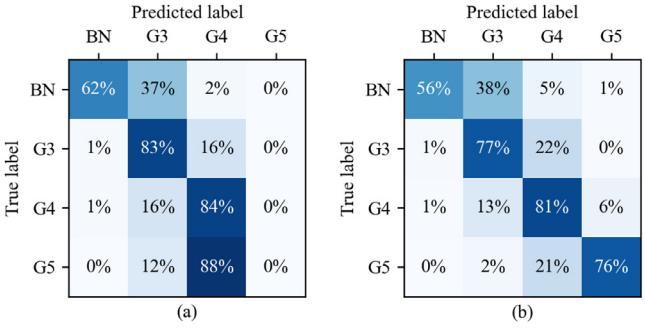

Fig. 9. Confusion matrices for (a) $\mathcal{O}_{DORN}$ and (b) $\mathcal{M}_{MSE-CE_O}$ on P_{TestII}.

gained the best performance for the three evaluation metrics. However, $\mathcal{O}_{DORN}$ obtained the second-worst F1 (0.584), whereas the proposed models showed the best F1 ($\geq$ 0.651). Upon closer examination of the classification results (Fig. 9), we found that the predictions of $\mathcal{O}_{DORN}$ were biased towards mid and low grade cancers and no high grade cancer was predicted. On the contrary, the proposed models were able to predict not only mid and low grade cancers but also high grade cancers.

Fig. 8 illustrates the prediction maps of the proposed models and other competing models. The proposed models permit accurate and robust cancer grading for the unseen tissue samples obtained under different conditions. The prediction results of the other models are less reliable and much sensitive to the type of tissue images and cancer grades. $\mathcal{O}_{DORN}$, in particular, mis-predicts the entire cancers for grade 5.

6. Discussion

CNN models have shown their effectiveness in different tasks of pathology image analysis. To further exploit the pathology images and to enhance the learning capability of the models, multi-task learning or joint learning approaches have been recently adopted in many applications. In this work, cancer grading is cast as both categorical classification and ordinal classification problems, and thus the proposed method conducts two types of classification tasks. Although the objective of the two tasks is identical, i.e., cancer grading, the experimental results on two types of tissue organs demonstrate that there is a synergy between the two tasks, leading to an improved classification of cancer grades. Moreover, a substantial performance gain was obtained by the proposed loss function, i.e., $\mathcal{L}_{CE_O}$. It is simple, generic, and computationally inexpensive, suggesting that it is applicable to other applications and datasets at a minimal cost.

In digital and computational pathology, one of the unmet needs is to achieve good generalizability of the computational methods across datasets that underwent different acquisition and processing procedures. In this work, we employed a set of tissue images from two organs. For each organ, the tissue images were obtained from two different institutes using two different digital slide scanners. As commonly observed, the performance of the proposed models and other competing models substantially varied across different datasets for both organs. Though generally observed, it is surprising that the overall performance of all models on P_{TestII} was better than that on P_{TestI} since the two datasets were collected from different institutes and using different scanners. The difference in the performance may be ascribable to the ground truth annotations. P_{TestI} was annotated by a single pathologists and P_{TestII} was anno-

Table 4
Comparative Results of the regression task with and without benign samples.

Colorectal Dataset	C_{TestI}				C_{TestII}			
Method	Acc (%)	Acc_G (%)	F1	κ_w	Acc (%)	Acc_G (%)	F1	κ_w
$\mathcal{M}_{MAE-CE_O}$	87.7	82.7	0.843	0.816	**80.3**	**74.0**	**0.744**	**0.716**
$\mathcal{M}_{MSE-CE_O}$	**88.4**	**83.7**	**0.854**	**0.828**	78.1	72.0	0.729	0.695
$\mathcal{M}_{MAE-CE_O-Cancer}$	87.5	82.6	0.834	0.813	78.7	72.1	0.737	0.709
$\mathcal{M}_{MSE-CE_O-Cancer}$	88.0	83.3	0.843	0.820	80.3	74.2	0.753	0.724
Prostate Dataset	P_{TestI}				P_{TestII}			
Method	Acc (%)	Acc_G (%)	F1	κ_w	Acc (%)	Acc_G (%)	F1	κ_w
$\mathcal{M}_{MAE-CE_O}$	**0.696**	**0.702**	**0.633**	0.616	77.5	79.3	0.651	0.706
$\mathcal{M}_{MSE-CE_O}$	0.688	0.695	0.622	**0.617**	**77.7**	**79.5**	**0.658**	**0.707**
$\mathcal{M}_{MAE-CE_O-Cancer}$	67.6	69.0	0.574	0.451	76.0	79.4	0.625	0.585
$\mathcal{M}_{MSE-CE_O-Cancer}$	66.9	68.1	0.583	0.449	75.5	78.9	0.601	0.574

tated by six pathologists and aggregated by the STAPLE algorithm, in which the class label that has the highest probability among the six pathologists is assigned to each pixel in an image. It has been well-known that there exists a substantial inter- and intra-observer variability in cancer grading (Egevad et al., 2013; Elmore et al., 2015). Aggregating the annotations among multiple pathologists could provide more consistent labels for P_{TestII}, reducing the effect of the variability in cancer grading on the performance, which, in turn, resulting in the better performance on P_{TestII}. Nonetheless, the proposed models, in general, achieved better performance than other competing models except $\mathcal{O}_{DORN}$ on P_{TestII}. The superior performance of $\mathcal{O}_{DORN}$ is misleading since the model misses all the high grade (grade 5) cancers. This may be ascribable to the design of the model that focuses more on smaller depths, which are corresponding to benign or lower grade tumors. Since the number of grade 5 only amounts to 1.5% of the entire dataset, the misclassification on grade 5 has minimal effect on those evaluation metrics that incline to total number of true positive samples. We also found that $\mathcal{O}_{DORN}$ is not able to correctly classify any grade 5 or PD samples in other test datasets (P_{TestI}, C_{TestI}, and C_{TestII}), which is corresponding to its poor performance in F1. This certainly diminishes its clinical utility and implications. The proposed models, on the other hand, show well-balanced classification results on three cancer grades, and thus these are more likely transferable to the clinics. Please refer to Appendix A for the confusion matrices of all models in the entire test experiments.

The proposed joint learning approach takes advantage of a new loss function, $\mathcal{L}_{CE_O}$. In comparison to the common loss functions for a regression task ($\mathcal{L}_{MAE}$ and $\mathcal{L}_{MSE}$), $\mathcal{L}_{CE_O}$ not only downweights the magnitude of the loss assigned to well-classified samples but also upweights the magnitude associated with mis-classified samples. In this manner, it increases the contrast between correct and wrong classifications, in particular for the samples on the borderline. Similar approaches have been proposed to improve the accuracy and robustness of deep learning models. For instance, (Lin et al., 2017) addressed class imbalance in an object detection task by focal loss ($\mathcal{L}_{Focal}$). It downweights the magnitude of the loss for well-classified samples but does not consider the ordinal relationship among the classes. For depth estimation, (Fu et al., 2018) proposed $\mathcal{O}_{DORN}$ that downweights the magnitude of the loss for larger depths in log scale, degrading the classification performance for high grade cancers. Moreover, (De Vente et al., 2020) proposed the soft-label ordinal regression loss ($\mathcal{L}_{SL}$) for cancer detection and grading in MRI. (Cao et al., 2020) and (Liu et al., 2019a) proposed ranking-based loss functions for the age estimation and lung nodule classification, respectively. $\mathcal{L}_{SL}$ simply assigns a higher probability to a higher grade cancer and the two ranking-based methods utilize the relative order of the class labels, whereas $\mathcal{L}_{CE_O}$ estimates the probability distribution based upon the relative distance to all the class labels, leading to an improved characterization of the ordinal relationship.

Benign samples have a substantial effect on the regression task. Computing the ordinal losses ($\mathcal{L}_{MAE}$, $\mathcal{L}_{MSE}$, and $\mathcal{L}_{CE_O}$) for cancer samples only, the overall performance ($\mathcal{M}_{MAE-CE_O-Cancer}$ and $\mathcal{M}_{MSE-CE_O-Cancer}$) was comparable to that of the proposed joint learning approach, which utilizes both benign and cancer samples, for the colorectal tissue datasets (C_{TestI} and C_{TestII}); however, there was a consistent performance drop for the prostate tissue datasets (P_{TestI} and P_{TestII}) (Table 4). As for ACC_G, there was, in general, a smaller performance drop for both datasets in comparison to other evaluation metrics. This implies that the models are able to learn the ordinal relationship among cancer samples as the regression task utilizes them only. Due to the absence of benign samples, the regression task had no chance to handle them on an ordinal scale, leading to a reduction in the overall classification performance.

The final objective function of the proposed approach contains three loss functions. To fully exploit these three loss functions, their contributions could be adjusted by hand (Liu et al., 2019a), uncertainty (Kendall et al., 2018), or dynamic averaging (Liu et al., 2019b). Although such methods could aid in further optimizing the objective function and improving the capability of the proposed method, this is beyond the scope of this study. We leave this for the future study.

7. Conclusion

Herein, we present an approach of joint categorical and ordinal learning for cancer grading in pathology images. Conducting both categorical classification and ordinal classification tasks with a new loss function, the proposed joint learning model not only achieves accurate classification performance but also shows good generalizability on both colorectal and prostate datasets, leading to an improved pathology image analysis that could facilitate an automated and robust cancer diagnosis in clinics. The proposed joint learning framework as well as the new loss function should be applicable to other types of cancers and tissues in pathology image analysis. However, the additive value of the proposed method to the clinics still remains to be investigated. The routine pathology review determines a single grade per patient, which is used for decision-making on treatment planning and patients' prognosis. Although the accurate classification results on the entire tissue samples suggest that it could aid in improving the patient-level decision-makings, the proposed method has been built and evaluated via region-level experiments. The future direction is to further investigate its impact on the clinical decisions with respect to patient outcomes as well as to improve the joint learning approach to

enhance and stabilize its classification performance under various conditions.

Declaration of Competing Interest

The authors confirm that there are no conflicts of interest.

CRediT authorship contribution statement

Trinh Thi Le Vuong: Methodology, Software, Investigation, Writing – original draft, Visualization. **Kyungeun Kim:** Data curation, Validation, Resources, Writing – review & editing. **Boram Song:** Data curation, Validation, Resources, Writing – review & editing. **Jin Tae Kwak:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Acknowledgments

This work was supported by the Korea University grant (No. K2021531) and the National Research Foundation of Korea (NRF) grant (No. 2016R1C1B2012433 and No. 2021R1A2C2014557).

Appendix A. Confusion matrices of all models

For each of the test datasets, we provide the confusion matrices of all models. Figs. A1 and A2 show the confusion matrices for colorectal cancer classification on C_{TestI} and C_{TestII}, respectively. For prostate cancer classification, Figs. A3 and A4 demonstrate the corresponding confusion matrices on P_{TestI} and P_{TestII}, respectively.

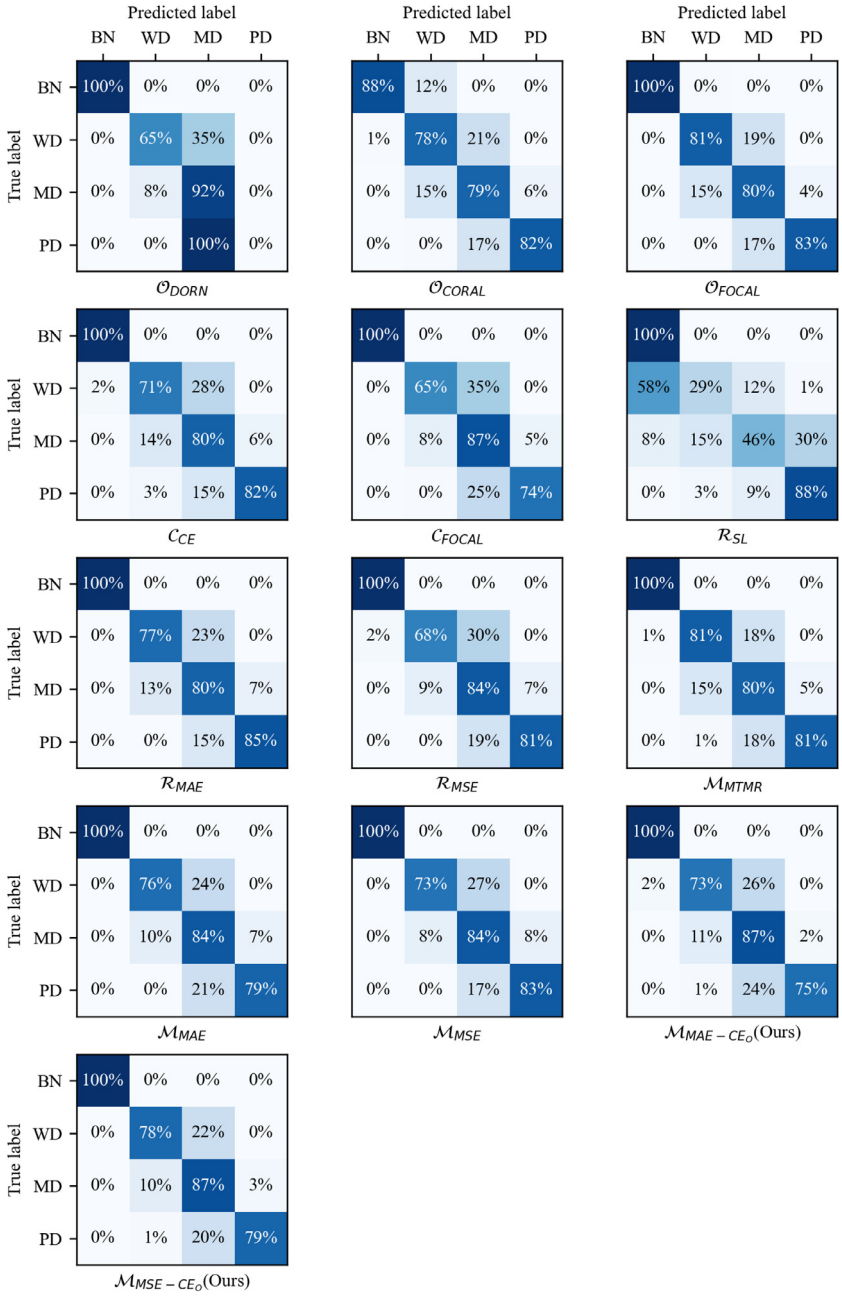

Fig. A1. Confusion matrices on C_{TestI}.

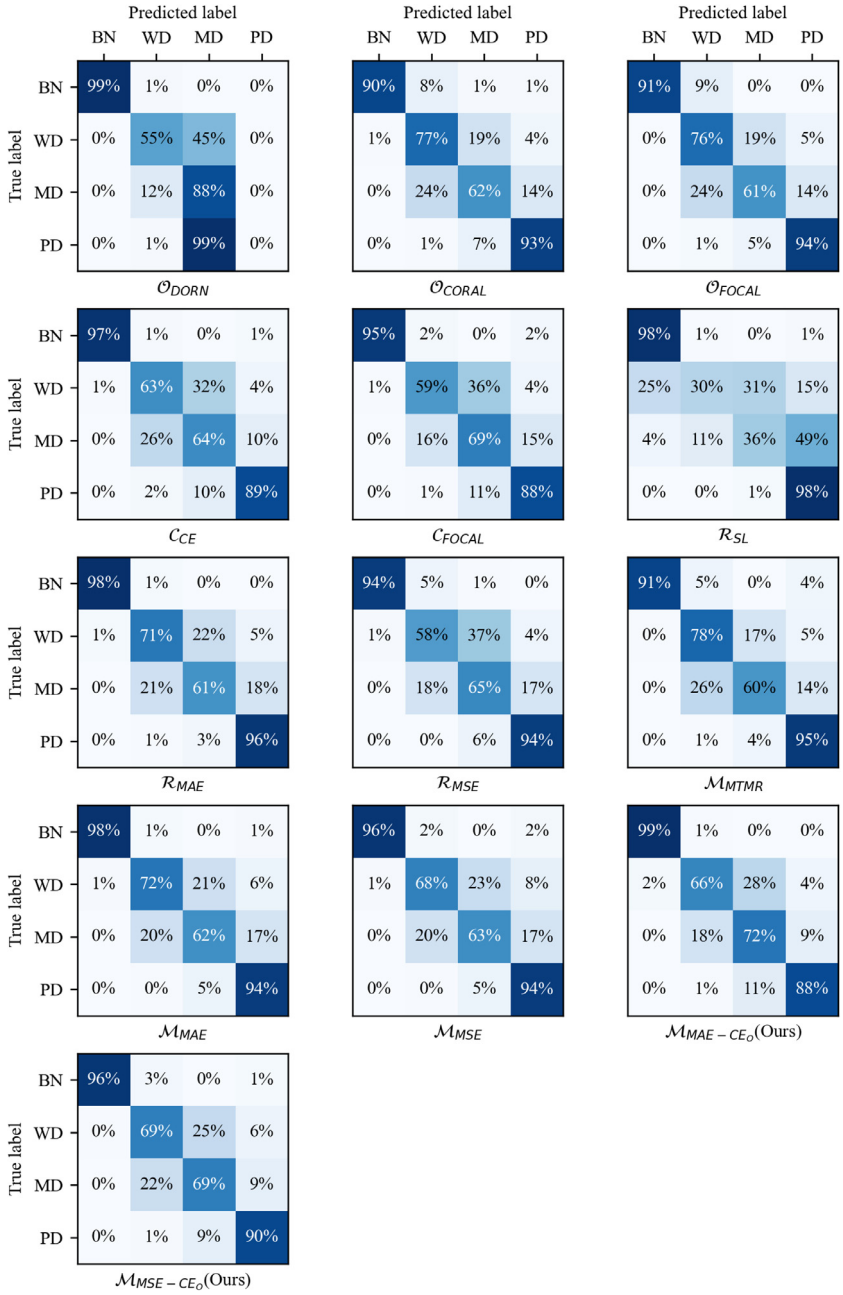

Fig. A2. Confusion matrices on C_{TestII}.

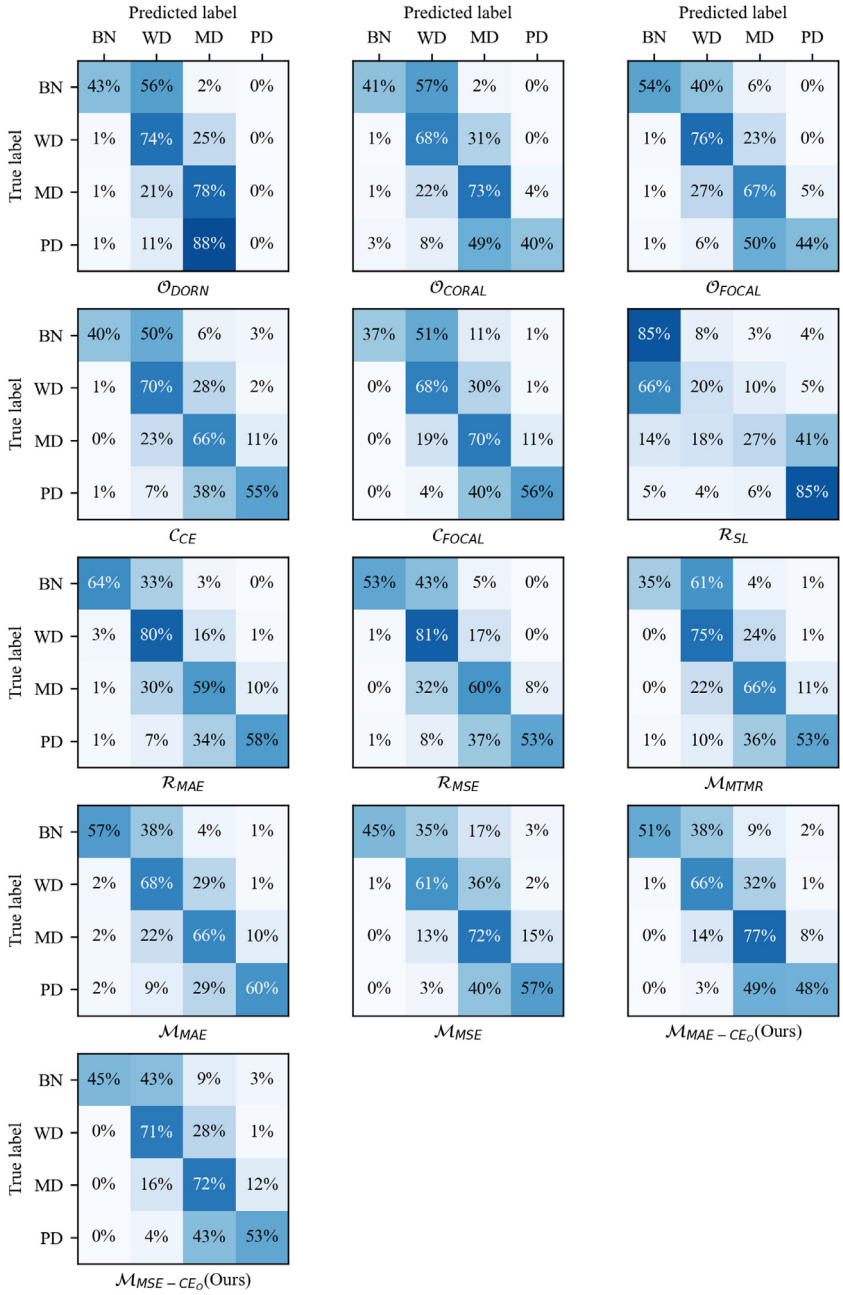

Fig. A3. Confusion matrices on P_{Test1}.

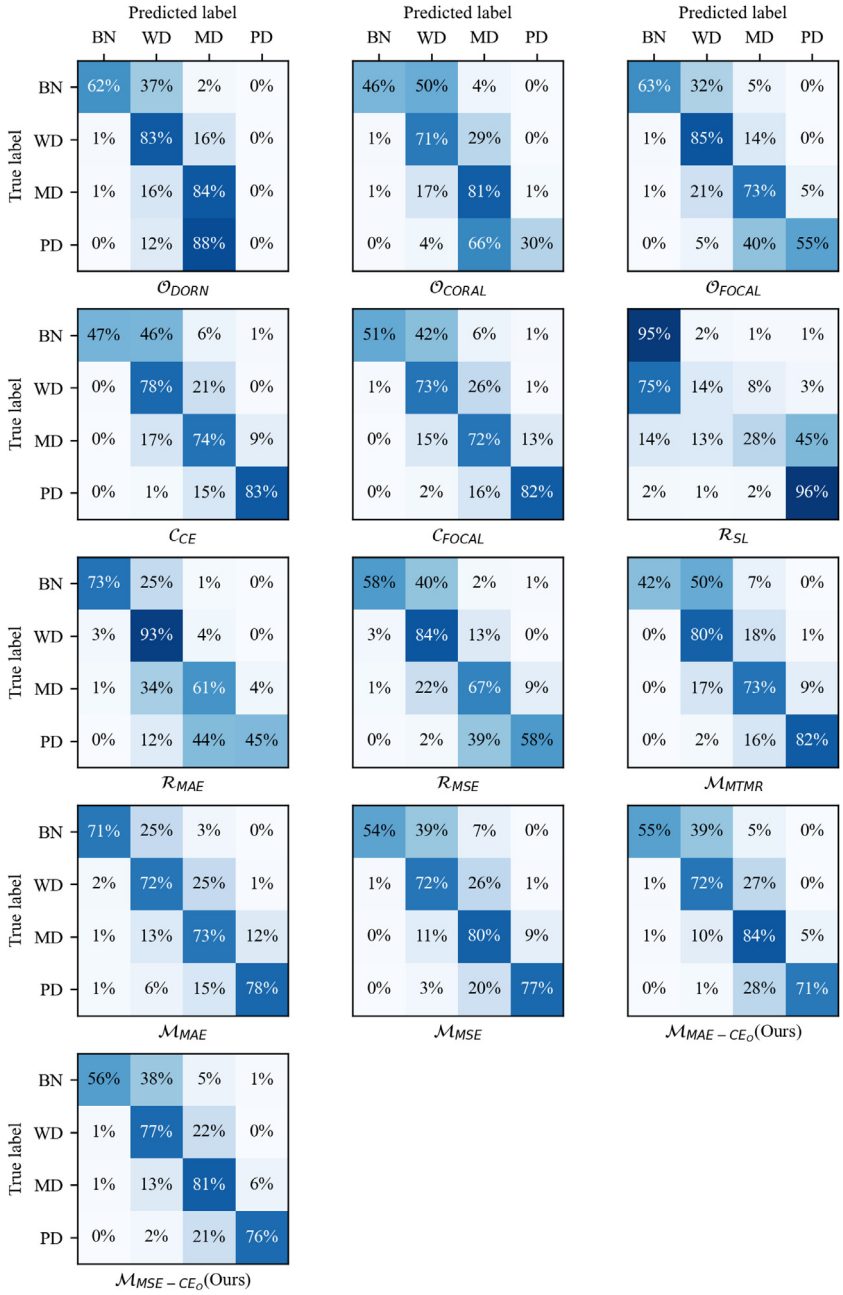

Fig. A4. Confusion matrices on P_{TestII}.

References

Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., Navab, N., 2016. AggNet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Trans. Med. Imaging 35 (5), 1313–1321.

Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., Campilho, A., 2017. Classification of breast cancer histology images using convolutional neural networks. PLoS ONE 12 (6), e0177544.

Arvaniti, E., Fricker, K.S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P.J., Rueschoff, J.H., Claassen, M., 2018. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Sci. Rep. 8 (1), 1–11.

Cao, D., Lei, Z., Zhang, Z., Feng, J., Li, S.Z., 2012. Human age estimation using ranking SVM. In: Chinese Conference on Biometric Recognition. Springer, pp. 324–331.

Cao, R., Bajgiran, A.M., Mirak, S.A., Shakeri, S., Zhong, X., Enzmann, D., Raman, S., Sung, K., 2019. Joint prostate cancer detection and Gleason score prediction in mp-MRI via FocalNet. IEEE Trans. Med. Imaging 38 (11), 2496–2506.

Cao, W., Mirjalili, V., Raschka, S., 2020. Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognit. Lett. 140, 325–331.

Caruana, R., 1997. Multitask learning. Mach. Learn. 28 (1), 41–75.

Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 411–418.

Cohen, J., 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit.. Psychol Bull. 70 (4), 213.

Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A., 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat. Med. 24 (10), 1559–1567.

Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., Madabhushi, A., 2014. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In: Medical Imaging 2014: Digital Pathology, vol. 9041. International Society for Optics and Photonics, p. 904103.

De Vente, C., Vos, P., Hosseinzadeh, M., Pluim, J., Veta, M., 2020. Deep learning regression for prostate cancer detection and grading in Bi-parametric MRI. IEEE Trans. Biomed. Eng..

Doyle, S., Feldman, M.D., Shih, N., Tomaszewski, J., Madabhushi, A., 2012. Cascaded discrimination of normal, abnormal, and confounder classes in histopathology: Gleason grading of prostate cancer. BMC Bioinf. 13 (1), 282.

Duong, Q.D., Vu, D.Q., Lee, D., Hewitt, S.M., Kim, K., Kwak, J.T., 2019. Scale embedding shared neural networks for multiscale histological analysis of prostate cancer. In: Medical Imaging 2019: Digital Pathology, vol. 10956. International Society for Optics and Photonics, p. 1095606.

Egevad, L., Ahmad, A.S., Algaba, F., Berney, D.M., Boccon-Gibod, L., Compérat, E., Evans, A.J., Griffiths, D., Grobholz, R., Kristiansen, G., et al., 2013. Standardization of Gleason grading among 337 European pathologists. Histopathology 62 (2), 247–256.

Elmore, J.G., Longton, G.M., Carney, P.A., Geller, B.M., Onega, T., Tosteson, A.N., Nelson, H.D., Pepe, M.S., Allison, K.H., Schnitt, S.J., et al., 2015. Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA 313 (11), 1122–1132.

Frank, E., Hall, M., 2001. A simple approach to ordinal classification. In: European Conference on Machine Learning. Springer, pp. 145–156.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D., 2018. Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2002–2011.

Girshick, R., 2015. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448.

Gorelick, L., Veksler, O., Gaed, M., Gómez, J.A., Moussa, M., Bauman, G., Fenster, A., Ward, A.D., 2013. Prostate histopathology: learning tissue component histograms for cancer detection and classification. IEEE Trans. Med. Imaging 32 (10), 1804–1818.

Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N., 2019. HoVer-Net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med. Image Anal. 58, 101563.

Herbrich, R., Graepel, T., Obermayer, K., 1999. Support vector learning for ordinal regression.

Kather, J.N., Weis, C.-A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A., Zöllner, F.G., 2016. Multi-class texture analysis in colorectal cancer histology. Sci. Rep. 6, 27988.

Kendall, A., Gal, Y., Cipolla, R., 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7482–7491.

Krawczyk, B., Galar, M., Jeleń, Ł., Herrera, F., 2016. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Appl. Soft. Comput. 38, 714–726.

Kwak, J.T., Hewitt, S.M., 2017. Multiview boosting digital pathology analysis of prostate cancer. Comput. Methods Programs Biomed. 142, 91–99.

Kwak, J.T., Hewitt, S.M., Sinha, S., Bhargava, R., 2011. Multimodal microscopy for automated histologic analysis of prostate cancer. BMC Cancer 11 (1), 62.

Li, L., Lin, H.-T., 2007. Ordinal regression by extended binary classification. In: Advances in neural information processing systems, pp. 865–872.

Liao, Q., Jiang, L., Wang, X., Zhang, C., Ding, Y., 2017. Cancer classification with multi-task deep learning. In: 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC). IEEE, pp. 76–81.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Liu, L., Dou, Q., Chen, H., Qin, J., Heng, P.-A., 2019. Multi-task deep model with margin ranking loss for lung nodule analysis. IEEE Trans. Med. Imaging 39 (3), 718–728.

Liu, S., Johns, E., Davison, A.J., 2019. End-to-end multi-task learning with attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1871–1880.

Loshchilov, I., Hutter, F., 2016. SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983.

Madabhushi, A., Lee, G., 2016. Image analysis and machine learning in digital pathology: challenges and opportunities.

Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J.G., Shapiro, L., 2018. Y-Net: joint segmentation and classification for diagnosis of breast biopsy images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 893–901.

Metter, D.M., Colgan, T.J., Leung, S.T., Timmons, C.F., Park, J.Y., 2019. Trends in the US and Canadian pathologist workforces from 2007 to 2017. JAMA Netw. open 2 (5), e194337.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D.A., Barnholtz-Sloan, J.S., Vega, J.E.V., Brat, D.J., Cooper, L.A., 2018. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc. Natl. Acad. Sci. 115 (13), E2970–E2979.

Nagpal, K., Foote, D., Liu, Y., Chen, P.-H.C., Wulczyn, E., Tan, F., Olson, N., Smith, J.L., Mohtashamian, A., Wren, J.H., et al., 2019. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. NPJ Digit. Med. 2 (1), 1–10.

Niazi, M.K.K., Parwani, A.V., Gurcan, M.N., 2019. Digital pathology and artificial intelligence. Lancet Oncol. 20 (5), e253–e261.

Nir, G., Hor, S., Karimi, D., Fazli, L., Skinnider, B.F., Tavassoli, P., Turbin, D., Villamil, C.F., Wang, G., Wilson, R.S., et al., 2018. Automatic grading of prostate cancer in digitized histopathology images: learning from multiple experts. Med. Image Anal. 50, 167–180.

Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G., 2016. Ordinal regression with multiple output CNN for age estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4920–4928.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520.

Serag, A., Ion-Margineanu, A., Qureshi, H., McMillan, R., Saint Martin, M.-J., Diamond, J., O'Reilly, P., Hamilton, P., 2019. Translational AI and deep learning in diagnostic pathology. Front. Med. 6.

Shaban, M., Awan, R., Fraz, M.M., Azam, A., Tsang, Y.-W., Snead, D., Rajpoot, N.M., 2020. Context-aware convolutional neural network for grading of colorectal cancer histology images. IEEE Trans. Med. Imaging.

Shashua, A., Levin, A., 2003. Ranking with large margin principle: two approaches. In: Advances in Neural Information Processing Systems, pp. 961–968.

Tabesh, A., Teverovskiy, M., Pang, H.-Y., Kumar, V.P., Verbel, D., Kotsianti, A., Saidi, O., 2007. Multifeature prostate cancer diagnosis and gleason grading of histological images. IEEE Trans. Med. Imaging 26 (10), 1366–1378.

Tan, M., Le, Q.V., 2019. Efficientnet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, pp. 6105–6114.

Turki, T., Wei, Z., 2018. Boosting support vector machines for cancer discrimination tasks. Comput. Biol. Med. 101, 236–249.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921.

Xie, W., Noble, J.A., Zisserman, A., 2018. Microscopy cell counting and detection with fully convolutional regression networks. Comput. Methods Biomech. Biomed.Eng. 6 (3), 283–292.

Zhang, Y., Yang, Q., 2017. A survey on multi-task learning. arXiv preprint arXiv:1707.08114.

Zhang, Y., Yang, Q., 2018. An overview of multi-task learning. Natl. Sci. Rev. 5 (1), 30–43.

Zhang, Z., Luo, P., Loy, C.C., Tang, X., 2014. Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision. Springer, pp. 94–108.

Zhao, R., Liao, W., Zou, B., Chen, Z., Li, S., 2019. Weakly-supervised simultaneous evidence identification and segmentation for automated glaucoma diagnosis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 809–816.

Zhou, Y., Graham, S., Alemi Koohbanani, N., Shaban, M., Heng, P.-A., Rajpoot, N., 2019. CGC-Net: cell graph convolutional network for grading of colorectal cancer histology images. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, p. 0.