

Deep Bayesian Active Learning-to-Rank with Relative Annotation for Estimation of Ulcerative Colitis Severity

Takeaki Kadota^{a,*}, Hideaki Hayashi^b, Ryoma Bise^{a,c}, Kiyohito Tanaka^d, Seiichi Uchida^{a,c}

^aDepartment of Advanced Information Technology, Kyushu University, 744, Motoooka, Nishi-ku, Fukuoka-shi, Fukuoka, 819-0395, Japan

^bInstitute for Dataability Science, Osaka University, 2-8, Yamadaoka, Suita-shi, Osaka, 565-0871, Japan

^cResearch Center for Medical Bigdata, National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan

^dDepartment of Gastroenterology, Kyoto Second Red Cross Hospital, 355-5, Haruobicho Kamigyo-ku, Kyoto-shi, Kyoto, 602-8026, Japan

Abstract

Automatic image-based severity estimation is an important task in computer-aided diagnosis. Severity estimation by deep learning requires a large amount of training data to achieve a high performance. In general, severity estimation uses training data annotated with discrete (i.e., quantized) severity labels. Annotating discrete labels is often difficult in images with ambiguous severity, and the annotation cost is high. In contrast, relative annotation, in which the severity between a pair of images is compared, can avoid quantizing severity and thus makes it easier. We can estimate relative disease severity using a learning-to-rank framework with relative annotations, but relative annotation has the problem of the enormous number of pairs that can be annotated. Therefore, the selection of appropriate pairs is essential for relative annotation. In this paper, we propose a deep Bayesian active learning-to-rank that automatically selects appropriate pairs for relative annotation. Our method preferentially annotates unlabeled pairs with high learning efficiency from the model uncertainty of the samples. We prove the theoretical basis for adapting Bayesian neural networks to pairwise learning-to-rank and demonstrate the efficiency of our method through experiments on endoscopic images of ulcerative colitis on both private and public datasets. We also show that our method achieves a high performance under conditions of significant class imbalance because it automatically selects samples from the minority classes.

Keywords: Computer-aided diagnosis, Learning to rank, Active learning, Relative annotation, Endoscopic image dataset

1. Introduction

Automatic image-based severity estimation is important to assist medical doctors in clinical practice. Deep learning has been applied to many disease severity estimations (Cho et al., 2019; Klang et al., 2021; Takenaka et al., 2020). Severity estimation using deep learning requires the collection of a large amount of training data annotated with severity labels by medical experts. Since medical experts need to carefully identify disease severity for many images, creating training data demands laborious efforts.

Standard annotations (hereafter referred to as *absolute annotations*) represent disease severity as discretized severity labels. Figure 1(a) shows the absolute annotations for endoscopic images of ulcerative colitis (UC). Absolute annotation can be difficult even for medical experts. This is because disease severity is inherently continuous, and when expressed in discrete severity levels, the levels can be ambiguous in intermediate cases. For example, when medical experts annotate medical images to be classified into four severity levels (0, 1, 2, and 3), they frequently encounter images near the intermediate severity levels (0.5, 1.5, and 2.5). Assigning discrete severity levels to these

images is a time-consuming task that potentially leads to variability in decision outcomes. It has also been reported that absolute annotation has high variability not only between different medical experts but also within the same medical expert (Hirai and Matsui, 2008).

Relative annotation is a promising alternative to absolute annotation in that it offers an easier process. Figure 1(b) shows the relative annotations for UC endoscopic images, where we compare the severity of the two images and attach relative labels that indicate the result of the comparison. Relative annotation tends to be easier for annotators compared to absolute annotation, and it helps reduce subjective bias. Consequently, relative annotation leads to less variability in decision outcomes across different annotators. Relative annotation has been used for pairwise learning-to-rank (LTR) methods, mainly for ranking tasks in information retrieval (Carterette and Petkova, 2006; Liu, 2009; Leaman et al., 2013; Hofmann et al., 2013). In computer vision, relative annotation is also used in image analysis applications because it is easy to perform and is stable when labeling continuously changing data (Parikh and Grauman, 2011). Recently, relative annotation has been applied to medical images, and reports have indicated that it reduces annotation costs and labeling errors (Kadota et al., 2022a; Saibro et al., 2022).

Image pair datasets with relative annotation can be used for ranking tasks to estimate the order of image severity. Given an image pair (x_i, x_j) , where x_i has a higher severity than x_j ,

*Corresponding author: Takeaki Kadota

Email address: takeaki.kadota@human.ait.kyushu-u.ac.jp
(Takeaki Kadota)



Figure 1: Absolute and relative annotations.

a ranking function $f(x)$, which outputs a scalar value called a rank score, is trained such that $f(x_i) > f(x_j)$ is satisfied. Since $f(x)$ gives a higher rank score to an image with higher severity, it can be used as the severity of x . Therefore, we can estimate the order of severity for multiple images by comparing their rank scores. Furthermore, a calibration method using a small number of absolute annotations allows $f(x)$ to be used in the classification task as the absolute severity instead of the relative severity (Kadota et al., 2022a). In addition, $f(x)$ can be used for two-class classification when medical experts determine the threshold for $f(x)$ (Saibro et al., 2022).

A critical challenge in LTR with relative annotation is the need to carefully select and annotate highly effective pairs from all possible pairs for learning. In relative annotation, there are $N(N - 1)/2$ possible pairs for N image samples. Even though individual relative annotation is easy, it is practically difficult to annotate all possible pairs. In addition, even if all pairs were used for training, the training time would be considerably longer, and less effective pairs for learning could reduce performance. Therefore, it is essential to preferentially select and annotate pairs with high learning effectiveness suitable for severity estimation from all pairs.

In this paper, we address the problem of pair selection for relative annotations by using an active learning framework based on a Bayesian convolutional neural network (Bayesian CNN). For a trained model, Bayesian CNN estimates the uncertainty of the sample. We employed a Bayesian CNN because it can estimate uncertainty by applying Monte Carlo (MC) dropout (Gal and Ghahramani, 2016b) without changing the network structure for the pairwise LTR framework. The proposed method features *active learning*, in which the Bayesian CNN is introduced into the LTR framework to find pairs with high uncertainty and gradually add relative annotations to them. The experiments described in later sections show that the proposed method achieves high performance on training data with significantly fewer pairs than $N(N - 1)/2$.

We also provide theoretical justification for the uncertainty estimation in LTR based on the Bayesian CNN. Our method applies MC dropout to CNNs with a Siamese network structure of pairwise ranking approaches. The application of MC dropout to CNNs in regression and classification tasks has already been demonstrated (Gal et al., 2017). Here, we prove that MC dropout is equally effective for CNNs in a Siamese network structure in the ranking task. Our findings indicate that the proposed method may be effective not only for severity estimation with relative annotation but also for various other applications. Furthermore, through experiments using both pri-

vate and public UC endoscopic image datasets, we demonstrate the effectiveness of the proposed method in severity estimation. Medical image datasets generally have class imbalance because there are more images with normal or mild disease than those with severe disease. We found that the proposed method preferentially selects important samples from the minority classes and thereby reduces the class imbalance.

The main contributions of this paper are as follows:

1. We propose an active learning method that introduces Bayesian CNN into a learning-to-rank framework to address the pair selection problem, an essential challenge in relative annotation.
2. We theoretically demonstrate the applicability of MC dropout in estimating uncertainty for a pairwise LTR task using a Bayesian Siamese neural network (NN).
3. We experimentally demonstrate that the proposed method improves the severity estimation performance by selecting pairs with high learning efficiency on a UC endoscopy image dataset. In addition, we verify the generalization ability of the proposed method using a public dataset. We also confirm that the proposed method is robust to class imbalances by selecting minority but important samples with high priority.
4. We prove the usefulness of the proposed method in a severity classification task that classifies each image into one of the discretized severity levels, which is a common task in medical image diagnosis.

A preliminary version of this paper was published as a conference paper at MIUA2022 (Kadota et al., 2022b). In the current version, we add the following new and significant contributions.

- We prove that MC dropout can be applied to pairwise LTR by deriving the variational lower bounds in pairwise LTR with Siamese network structures. We then show how this allows us to treat the Siamese network structure with dropout as a Bayesian model and obtain uncertainty estimates for its features (corresponds to Contribution 2).
- We experimentally verify the effectiveness of our method in the task of relative severity estimation using a public dataset of UC endoscopic images (as a part of Contribution 3).
- To use the proposed method in the multi-class classification tasks commonly used in clinical practice to make treatment decisions, we integrate a technique to estimate the discrete severity levels by calibrating the rank score to the UC severity score into the proposed method (corresponds to Contribution 4).

2. Related work

2.1. Disease severity estimation

2.1.1. Severity estimation based on endoscopic images

Many methods have been proposed to estimate disease severity based on endoscopic images. Liu et al. introduced a classification method that estimates three endoscopic severity levels of

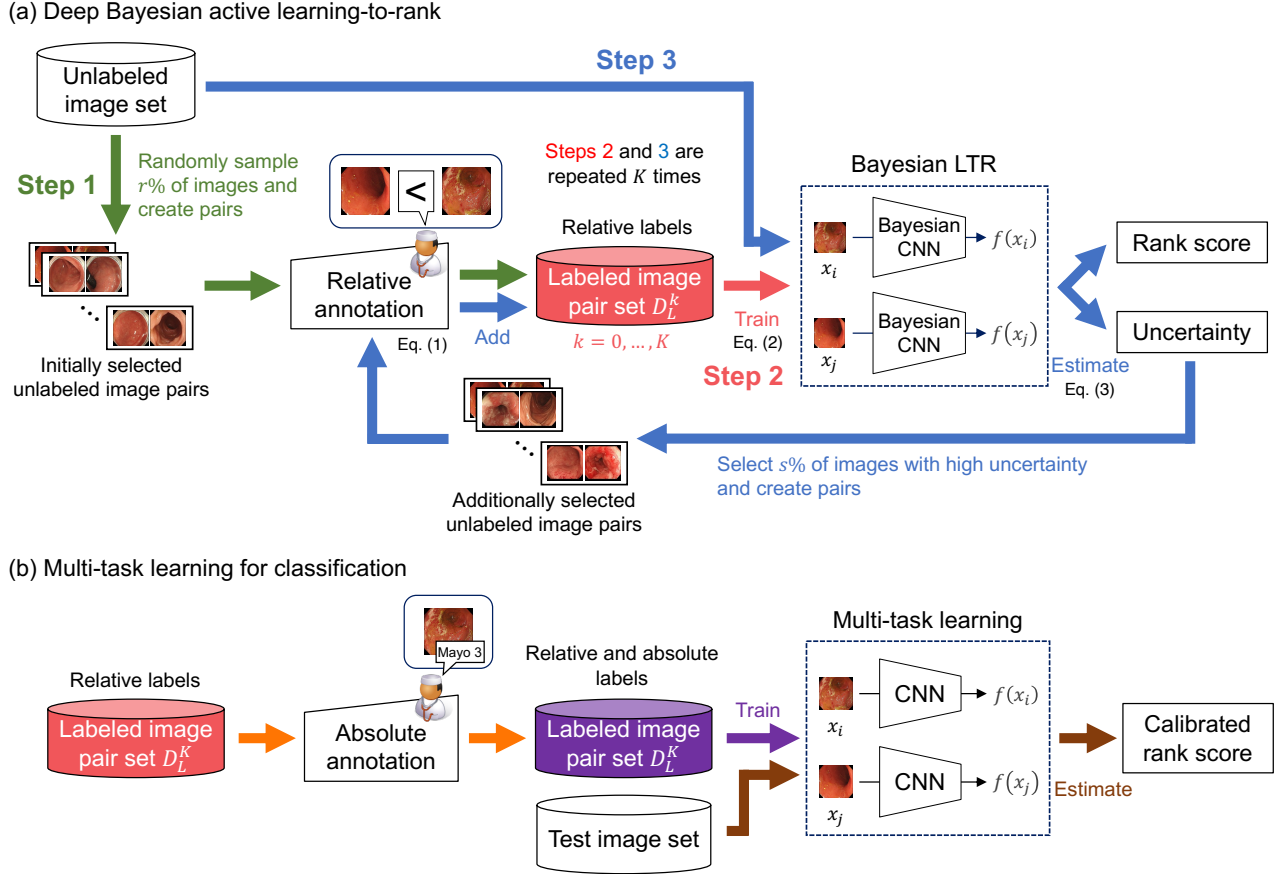


Figure 2: (a) Deep Bayesian active learning-to-rank for relative severity estimation; step 1 (green arrows): generating a small number of pairs using randomly selected images from an unlabeled image set and annotating these pairs for the initial training; step 2 (red arrow): training the Bayesian CNN using the labeled image pair set; step 3 (blue arrows): selecting high-uncertainty images from the unlabeled image set to create pairs and attaching relative labels to the pairs. (b) Multi-task learning for severity classification.

esophageal cancer: normal, premalignant, and cancerous (Liu et al., 2020). Klang et al. used an NN to estimate five severity levels of Crohn’s disease using capsule endoscopy (Klang et al., 2021). However, these methods come with high annotation costs due to the need for absolute annotation that attaches discrete levels to continuously changing lesions on endoscopic images. Cho et al. used endoscopic images of pathologically confirmed gastric lesions to estimate five severity levels of gastric cancer (Cho et al., 2019). Although they demonstrated the feasibility of estimating pathologically diagnosed severity based on endoscopic images, their annotation process involved obtaining pathological images through biopsies, leading to considerable annotation costs.

2.1.2. Ulcerative colitis (UC) severity estimation

Several studies have investigated deep learning applications to UC severity estimation. Takenaka et al. used a dataset including the Ulcerative Colitis Endoscopic Index of Severity (UCEIS) for severity estimation (Takenaka et al., 2020). UCEIS is a discrete severity score ranging from zero to eight points based on vascular pattern, bleeding, and erosion/ulceration, but it incurs a very high annotation cost. Palot et al. proposed an ordinal regression-based method to estimate the Mayo

score, which categorizes UC severity into four levels (Polat et al., 2022a). They proposed a unique loss function that focuses on severity order, but the annotation cost is high because they use discrete severity levels as ground truth. To avoid the costly annotation of all captured images, Schwab et al. proposed a weakly supervised learning method that estimates UC severity (Stidham et al., 2019). Becker et al. proposed a method to reduce annotation costs by automatically extracting images suitable for UC severity scoring from endoscopic video frames (Becker et al., 2021). Although these studies tackle the issue of annotation costs in medical image datasets, they do not inherently solve the quantization error in disease severity, which is the root cause of rising annotation costs.

2.2. Learning to rank (LTR) with relative annotation

2.2.1. LTR for general image analysis

LTR with relative annotation has been applied to image analyses such as attribute evaluation and image quality assessment (IQA). Parikh et al. proposed a relative attribute model that predicts attribute similarity to images instead of a general classification model that predicts the category of an attribute (Parikh and Grauman, 2011). In addition, Souri et al. proposed using a CNN-based architecture to predict the strength of relative

attributes (Souri et al., 2016). Ma et al. used a CNN-based pairwise LTR architecture for IQA (Ma et al., 2017), and Liu et al. used a Siamese network structure to predict the quality ranking of images using a dataset featuring relative annotations for IQA (Liu et al., 2017). These studies focus on IQA but do not address the critical issue of relative annotation, which causes an enormous increase in the number of annotations due to pair creation.

2.2.2. LTR for medical image analysis

Recent studies have reported the successful application of LTR with relative annotation to the image analysis of continuously changing lesions. Kalpathy-Cramer et al. modeled relative severity using relative annotations and proposed continuous severity scores for the retinopathy of prematurity (Kalpathy-Cramer et al., 2016). They found poor absolute agreements on classification but good relative agreements on disease severity in expert diagnoses. Li et al. developed a Siamese NN approach to assess changes between disease severity at a single time point and longitudinal patient visits, focusing on continuous disease changes (Li et al., 2020). Lyu et al. proposed automatically selecting high-quality images based on image quality ranking using pairwise LTR (Lyu et al., 2021). As mentioned above, using LTR with relative annotation, Kadota et al. found that annotation costs could be reduced (Kadota et al., 2022a), while Saibro et al. found that labeling errors could be reduced (Saibro et al., 2022). However, although these studies show the effectiveness of relative annotation, they do not solve the issue of pair selection in relative annotation.

2.3. Active learning

2.3.1. Diversity-based sampling

Diversity-based sampling, one of the main techniques in active learning, is a selection strategy to efficiently find samples representative of the entire data distribution in the feature space. Dasgupta et al. proposed efficient sample selection using hierarchical clustering (Dasgupta and Hsu, 2008). Sener et al. reformulated active learning as core-set selection and examined the core-sets by solving a greedy k-center problem (Sener and Savarese, 2018). Thapa et al. applied an active learning method based on core-set to semantic segmentation and depth estimation of endoscopic images (Thapa et al., 2022). Sourati et al. proposed a sampling based on Fisher information for CNNs (Sourati et al., 2018). Smailagic et al. proposed a selection method for unlabeled samples using a distance function in the feature space (Smailagic et al., 2020). The sample selection implemented in these studies assumes absolute annotation and does not consider relative annotation that creates pairs.

2.3.2. Uncertainty-based sampling

Many techniques for uncertainty-based sampling have been proposed in medical image analysis to select samples with high uncertainty as informative samples. Yang et al. and Gorriz et al. used pixel-wise sample uncertainty to determine effective annotation regions in their segmentation tasks (Yang et al., 2017; Gorriz et al., 2017). Tang et al. used active learning with uncertainty selection and pseudo labels for the classification and

segmentation of endoscopic images (Tang et al., 2023). Wen et al. proposed a sample selection method for pathology images using patch-wise uncertainty (Wen et al., 2018). Nair et al. used voxel-wise uncertainty measures for 3D lesion segmentation (Nair et al., 2020). These methods do not assume any LTRs with relative annotations because they deal with segmentation or detection tasks.

In the field of natural language processing, Wang et al. have recently reported a method for applying a Bayesian CNN to LTR (Wang et al., 2021). Although their method seems similar to ours, their purpose, structure, and application are quite different. Specifically, their objective is to rank sentences (answers) for a given sentence (query) according to their relevance. For this purpose, their network always takes two inputs (e.g., $f(x_i, x_j)$), whereas our network takes one input (e.g., $f(x_i)$). These differences make it impossible to use their method for active relative annotation tasks and thus to compare our method with theirs.

3. Deep Bayesian active learning-to-rank

In relative annotation, labeled image pair sets $\mathcal{D}_L^k = \{(x_i, x_j, C_{i,j})\}$, where x_i, x_j are input images, $C_{i,j}$ is the ground truth of the pair (relative label), and k is the number of repetitions, are obtained by repeatedly selecting images from the unlabeled image set $\mathcal{D}_U = \{x_i\}$ to create pairs that are then annotated by medical experts. Our purpose is to achieve a high performance with a small number of pairs by selecting images from the unlabeled image set that are highly effective in learning.

We employed a Bayesian CNN for uncertainty estimation in pairwise LTR to select highly effective samples for training. This is because MC dropout with the Bayesian CNN can estimate uncertainty without changing the network structure of the pairwise LTR, which uses relative annotation. While ensemble-based methods can also be used to estimate uncertainty in deep neural networks (Beluch et al., 2018), they require substantial computational resources due to the need to train multiple networks, and the cost would be further amplified in Siamese network structures. For these reasons, we adopted the Bayesian CNN for uncertainty estimation in active learning for pairwise LTR.

As shown in Fig. 2(a), the proposed method consists of Bayesian inference based on the LTR algorithm for disease severity estimation and uncertainty-based active learning. While the Bayesian CNN is trained on the basis of the LTR algorithm, medical experts gradually add relative labels to the selected image pairs based on the uncertainty the Bayesian CNN provides. The proposed method consists of three steps. In step 1, we generate a small number of pairs using randomly selected images from an unlabeled image set. Medical experts annotate these pairs for the initial training. In step 2, we train the Bayesian CNN using the labeled image pair set to estimate the rank score and uncertainty of the individual training samples. In step 3, high-uncertainty images are selected from the unlabeled image set based on the estimated uncertainty to create pairs, and the medical expert attaches relative labels to the

pairs. We repeat steps 2 and 3 K times to train the Bayesian CNN while gradually increasing the training data by adding labeled image pairs.

3.1. Preparing pairs for initial training (step 1)

Relative annotation is generally performed by randomly selecting the number of pairs that can be annotated since this number increases quadratically with the image data. However, random selection results in many images of the majority class and can thus lead to a dataset with low learning efficiency due to many similar images. Therefore, the proposed method randomly selects as small a number of images as possible for the dataset in the initial training and prepares a labeled image pair set \mathcal{D}_L^0 as follows. First, given a set of N unlabeled images, randomly sample R images, namely, $r\%$ of all images (i.e., $R = rN/100$ image samples). Then, the set of R pairs is formed by randomly selecting from $R - 1$ samples for each R sample. The strategy for pair formation is arbitrary. Here, we form R pairs instead of all possible $R(R - 1)/2$ pairs to limit the number of pairs to be annotated. This strategy is used to avoid annotating $O(R^2)$ and thus to annotate all samples of R at least once. Medical experts attach relative labels to these image pairs. Relative labels $C_{i,j}$ are defined for image pairs (x_i, x_j) as follows:

$$C_{i,j} = \begin{cases} 1, & \text{if } x_i \text{ has more severity than } x_j, \\ 0.5, & \text{else if } x_i \text{ and } x_j \text{ have equal severity,} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Through this step, we obtain the annotated image pair sets $\mathcal{D}_L^0 = \{(x_i, x_j, C_{i,j})\}$ and $|\mathcal{D}_L^0| = R$ for initial training.

3.2. Training a Bayesian CNN (step 2)

In active learning, uncertainty-based sampling is generally used to reduce the annotation cost for training data. Uncertainty-based sampling preferentially selects highly effective samples for learning. Gal et al. proposed approximate Bayesian inference with MC dropout for applying this sampling method to deep learning (Gal and Ghahramani, 2016b; Gal et al., 2017). Their sampling strategy is to obtain the model uncertainty of samples in regression and classification tasks through a Bayesian CNN. While they provide theoretical proof that MC dropout functions as a Bayesian CNN, their study did not consider it in Siamese network structures for ranking tasks.

LTR with relative annotations generally uses a pairwise LTR algorithm with a Siamese network structure. We propose an uncertainty-based active learning method that applies Bayesian CNN to Siamese network structures for ranking tasks. A Bayesian CNN is trained with a labeled image pair set \mathcal{D}_L^0 and obtained as a ranking function. The obtained CNN outputs a rank score that represents the input image's severity order and the uncertainty of the rank score. We used RankNet (Burges et al., 2005), a pairwise ranking method with a Siamese network structure, for the LTR algorithm. We employed RankNet because it utilizes a neural network as the model, and uncertainty can be easily estimated by applying MC dropout. RankNet uses a probabilistic ranking cost function in training. The proposed

method performs approximate Bayesian inference by applying MC dropout to the Siamese neural network.

Let $f(\cdot)$ be a ranking function that is a CNN with L -weighted layers. Given a single image x , the CNN returns a scalar value $f(x)$ as the ranking score for x . Let \mathbf{W}_l be the l -th weight tensor of the CNN. The Bayesian CNN is trained on the mini-batch \mathcal{M} sampled from \mathcal{D}_L^0 , with dropout performed using the loss function \mathcal{L}_M defined as follows:

$$\mathcal{L}_M^{\text{rank}} = - \sum_{(i,j) \in \mathcal{I}_M} \{C_{i,j} \log P_{i,j} + (1 - C_{i,j}) \log(1 - P_{i,j})\} + \lambda \sum_{l=1}^L \|\mathbf{W}_l\|_F^2, \quad (2)$$

where \mathcal{I}_M is a set of index pairs by the elements in mini-batch \mathcal{M} , $P_{i,j} = \text{sigmoid}(f(x_i) - f(x_j))$ is a probability obtained from output values, λ is a constant value for weight decay, and $\|\cdot\|_F$ is a Frobenius norm. The first term in Eq. (2) is a probabilistic ranking loss function (Burges et al., 2005) for the CNN to learn rank scores. The loss function is a cross-entropy loss defined by the target probability $C_{i,j}$ as ground truth and the probability $P_{i,j}$ obtained from the output values. The second term in Eq. (2) is a weight regularization term that can be derived from the Kullback-Leibler divergence between the approximate posterior and the posterior of the CNN weights (Gal and Ghahramani, 2016b). The CNN is trained to minimize the loss function \mathcal{L}_M for each mini-batch while performing dropout. With a probability of p_{dropout} , if the binary random variable takes one, it is sampled for every unit in the CNN at each forward calculation, and if the corresponding binary variable takes zero, the output of the unit is set to zero.

The rank score of the unlabeled image \mathbf{x}^* is the average of the predictions by the trained Bayesian CNN calculated as $y^* = \frac{1}{T} \sum_{t=1}^T f(\mathbf{x}^*; \omega_t)$ where T is the number of sampling operations by MC dropout, ω_t is the t -th realization of the set of CNN weights obtained by MC dropout, and $f(\cdot; \omega_t)$ is the output of $f(\cdot)$ given a set of weights ω_t .

The prediction uncertainty obtained from the Bayesian CNN is defined as the variance of the posterior distribution of y^* . This uncertainty is used to select images for annotation in active learning and plays an important role in obtaining a highly effective dataset for learning. The variance of the posterior distribution, $\text{Var}_{q(y^*|\mathbf{x}^*)}[y^*]$, which represents the uncertainty, is approximately obtained using MC dropout as follows:

$$\text{Var}_{q(y^*|\mathbf{x}^*)}[y^*] = \mathbb{E}_{q(y^*|\mathbf{x}^*)}[(y^*)^2] - (\mathbb{E}_{q(y^*|\mathbf{x}^*)}[y^*])^2 \approx \frac{1}{T} \sum_{t=1}^T (f(\mathbf{x}^*; \omega_t))^2 - \left(\frac{1}{T} \sum_{t=1}^T f(\mathbf{x}^*; \omega_t) \right)^2 + \text{const.}, \quad (3)$$

where $q(y^*|\mathbf{x}^*)$ is the posterior distribution estimated by the model. In the next step, the absolute value of the uncertainty is not required, and thus, the constant term can be ignored.

3.3. Uncertainty-based sample selection (step 3)

The estimated uncertainty and relative annotations provide a new set of annotated image pairs. We use a trained Bayesian

CNN to estimate the rank scores and associated uncertainties for the unlabeled images and select the top $s\%$ of images with high uncertainty. As in step 1, image pairs are generated from the selected images, and the medical expert annotates the image pairs with relative labels. The image pairs with newly attached relative labels are added to the current set of annotated image pairs \mathcal{D}_L^0 . The updated \mathcal{D}_L^1 is used to retrain the Bayesian CNN. We repeat steps 2 and 3 K times to increase the size of the annotated set \mathcal{D}_L^k ($k = 0, \dots, K$).

3.4. Additional absolute annotation for multi-class classification

As shown in Fig. 2(b), to apply the proposed method to a multi-class classification task, the training set with relative labels is additionally annotated with absolute labels for multi-task learning. The training set obtained in step 3 has relative labels attached to the pairs but no absolute labels attached to the individual images. For a classification task, absolute labels need to be used as ground truth, so we perform absolute annotation on each image to the obtained training set. In addition, to calibrate the rank score to the disease severity score, LTR and regression are trained simultaneously with absolute and relative labels by multi-task learning. The regression loss function is defined by the squared error loss function for each sample of the pairs as follows:

$$L_M^{\text{reg}} = \sum_{(i,j) \in \mathcal{I}_M} \{(f(x_i) - A_i)^2 + (f(x_j) - A_j)^2\}, \quad (4)$$

where A_i and A_j are absolute labels of x_i and x_j , respectively. The loss function of multi-task learning is defined as the sum of the LTR loss function L_M^{rank} in Eq. (2) and the regression loss function L_M^{reg} .

4. Theoretical analysis of Bayesian learning-to-rank

4.1. Evaluating log evidence lower bound for ranking

In this section, we demonstrate that model uncertainty can be estimated by utilizing MC dropout in the context of LTR employing a Siamese network. The validity of the MC dropout-based uncertainty estimation for an NN was originally shown by Gal and Ghahramani on general classification and regression tasks (Gal and Ghahramani, 2016a). Their approach involved obtaining an approximate distribution of the posterior distribution over the weights of the NN through variational inference, which is attributed to maximizing the log-evidence lower bound. This was then shown to be equivalent to minimizing the loss function (cross-entropy for classification or mean squared error for regression) with L_2 regularization and MC dropout. Our proof process follows that of the aforementioned general classification and regression cases, and we primarily establish the equivalency between a log-evidence lower bound and the loss function with L_2 regularization and MC dropout in the context of LTR with a Siamese network.

We consider the adaptation of MC dropout to the Siamese network structure for the case of an NN with a single hidden

layer. We use the Siamese network structure consisting of single hidden layer NNs to simplify the explanations in this section, but the generalization to multi-layer NNs is straightforward. Let \mathbf{W}_1 and \mathbf{W}_2 be the weight matrices connecting the first layer to the hidden layer and the hidden layer to the output layer, respectively, and let \mathbf{b} be the bias. Let \mathbf{X}_1 and \mathbf{X}_2 be training data matrices with the n -th training sample $(x_1^n)^\top$ and $(x_2^n)^\top$ ($n = 1, \dots, N$) in the n -th row, respectively. The output data matrices for the inputs \mathbf{X}_1 and \mathbf{X}_2 are denoted by \mathbf{Y}_1 and \mathbf{Y}_2 that contain the n -th output vector $(y_1^n)^\top$ and $(y_2^n)^\top$ in the n -th row, respectively. Given data matrices \mathbf{X}_1 , \mathbf{X}_2 , \mathbf{Y}_1 , and \mathbf{Y}_2 , we estimate the ranking function $y = f(\mathbf{x})$ obtained by the pairwise LTR of the Siamese network structures. Let $\mathbf{c} = [c^1, \dots, c^N]^\top$ be the relative label matrix with the n -th relative label c^n for x_1^n and x_2^n . Then, we can write the generative model for the ranking task as follows:

$$\begin{aligned} p(c | \mathbf{X}_1, \mathbf{X}_2) &= \int p(c | \mathbf{Y}_1, \mathbf{Y}_2) p(\mathbf{Y}_1, \mathbf{Y}_2 | \mathbf{X}_1, \mathbf{X}_2) d\mathbf{Y}_1 d\mathbf{Y}_2 \\ &= \int p(c | \mathbf{Y}_1, \mathbf{Y}_2) p(\mathbf{Y}_1, \mathbf{Y}_2 | \mathbf{X}_1, \mathbf{X}_2, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}) \\ &\quad \cdot p(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}) d\mathbf{W}_1 d\mathbf{W}_2 d\mathbf{b} d\mathbf{Y}_1 d\mathbf{Y}_2, \end{aligned} \quad (5)$$

where \mathbf{W}_1 is a $Q \times U$ matrix derived from Q -dimensional inputs and U hidden units, \mathbf{W}_2 is a $U \times D$ matrix derived from U hidden units and D -dimensional outputs, and \mathbf{b} is a U -dimensional vector of bias terms.

From Eq. (5), the log-evidence lower bound of the pairwise LTR can be written as follows (the calculation process is described in the Supplementary Materials):

$$\begin{aligned} \mathcal{L}_{\text{GP-VI}} &:= \int p(\mathbf{Y}_1, \mathbf{Y}_2 | \mathbf{X}_1, \mathbf{X}_2, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}) q(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}) \\ &\quad \cdot \log(p(c | \mathbf{Y}_1, \mathbf{Y}_2)) d\mathbf{W}_1 d\mathbf{W}_2 d\mathbf{b} d\mathbf{Y}_1 d\mathbf{Y}_2 \\ &\quad - \text{KL}(q(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}) || p(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b})), \end{aligned} \quad (6)$$

where $q(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b})$ is the approximating variational distribution, and KL is the Kullback-Leibler divergence.

The integrand of the first term in Eq. (6) can be rewritten as a sum:

$$\log(p(c | \mathbf{Y}_1, \mathbf{Y}_2)) = \sum_{n=1}^N \log(p(c^n | y_1^n, y_2^n)). \quad (7)$$

As a result, Eq. (6) is expressed as follows:

$$\begin{aligned} \mathcal{L}_{\text{GP-VI}} &:= \sum_{n=1}^N \int p(y_1^n, y_2^n | x_1^n, x_2^n, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}) q(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}) \\ &\quad \cdot \log(p(c^n | y_1^n, y_2^n)) d\mathbf{W}_1 d\mathbf{W}_2 d\mathbf{b} dy_1^n dy_2^n \\ &\quad - \text{KL}(q(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}) || p(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b})). \end{aligned} \quad (8)$$

The integrands in the sum of Eq. (8) can be re-parameterized not to depend directly on \mathbf{W}_1 , \mathbf{W}_2 , and \mathbf{b} but rather on the standard normal and Bernoulli distributions. Let \mathbf{z}_1 and \mathbf{z}_2 be binary vectors whose element follows the Bernoulli distribution, as $q(z_{1,q}) = \text{Bernoulli}(p_1)$ with $p_1 \in [0, 1]$ for $q = 1, \dots, Q$ and

$q(z_{2,u}) = \text{Bernoulli}(p_2)$ with $p_2 \in [0, 1]$ for $u = 1, \dots, U$. Let $\epsilon_1 \in \mathbb{R}^{Q \times U}$, $\epsilon_2 \in \mathbb{R}^{U \times D}$, and $\epsilon \in \mathbb{R}^U$ be random matrices and a vector whose element independently follows the standard normal distribution. We re-parameterize the integrands as follows:

$$\begin{aligned} \mathbf{W}_1 &= \mathbf{M}_1 \text{diag}(\mathbf{z}_1) + \sigma \epsilon_1, \\ \mathbf{W}_2 &= \mathbf{M}_2 \text{diag}(\mathbf{z}_2) + \sigma \epsilon_2, \\ \mathbf{b} &= \mathbf{m} + \sigma \epsilon, \\ y_1^n &= \sqrt{\frac{1}{U}} \mathbf{W}_2^\top \phi(\mathbf{W}_1^\top \mathbf{x}_1^n + \mathbf{b}), \\ y_2^n &= \sqrt{\frac{1}{U}} \mathbf{W}_2^\top \phi(\mathbf{W}_1^\top \mathbf{x}_2^n + \mathbf{b}), \end{aligned} \quad (9)$$

where $\mathbf{M}_1 = [m_q]_{q=1}^Q$, $\mathbf{M}_2 = [m_u]_{u=1}^U$, and \mathbf{m} are variational parameters, $\text{diag}(\mathbf{z})$ is an operation that returns a diagonal matrix with the elements of vector \mathbf{z} on the main diagonal, $\sigma > 0$ is a scalar, and $\phi(\cdot)$ is an element-wise nonlinear function.

Next, using Monte Carlo integration with a distinct single sample, we estimate each integral for all pair samples as follows:

$$\begin{aligned} \mathcal{L}_{\text{GP-MC}} &:= \sum_{n=1}^N \log(p(c^n | \hat{y}_1^n(\mathbf{x}_1^n, \hat{\mathbf{W}}_1^n, \hat{\mathbf{W}}_2^n, \hat{\mathbf{b}}^n), \hat{y}_2^n(\mathbf{x}_2^n, \hat{\mathbf{W}}_1^n, \hat{\mathbf{W}}_2^n, \hat{\mathbf{b}}^n))) \\ &\quad - \text{KL}(q(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}) || p(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b})), \end{aligned} \quad (10)$$

where \hat{y}_1^n , \hat{y}_2^n , $\hat{\mathbf{W}}_1^n$, $\hat{\mathbf{W}}_2^n$, and $\hat{\mathbf{b}}^n$ are the realizations of y_1^n , y_2^n , \mathbf{W}_1^n , \mathbf{W}_2^n , and \mathbf{b}^n sampled on the basis of Eq. (9).

The first term of the sum in Eq. (10) can be written from the prediction probabilities of RankNet as follows:

$$\begin{aligned} \log(p(c^n | \hat{y}_1^n, \hat{y}_2^n)) &= \log(\text{sigmoid}(\hat{y}_1^n - \hat{y}_2^n)) \\ &= \log\left(\frac{1}{1 + \exp(\hat{y}_2^n - \hat{y}_1^n)}\right). \end{aligned} \quad (11)$$

Using Monte Carlo integration, we can approximate the KL divergence term with the variational parameters \mathbf{M}_1 , \mathbf{M}_2 , and \mathbf{m} and the probabilities p_1 , p_2 (the details of the approximation are described in the Supplementary Materials). Furthermore, we can scale the objective by $1/N$ and optimize it to yield the maximization objective as follows:

$$\begin{aligned} \mathcal{L}_{\text{GP-MC}} &\propto \frac{1}{N} \sum_{n=1}^N \log(p(c^n | \hat{y}_1^n, \hat{y}_2^n)) \\ &\quad - \frac{p_1}{2N} \|\mathbf{M}_1\|_2^2 - \frac{p_2}{2N} \|\mathbf{M}_2\|_2^2 - \frac{1}{2N} \|\mathbf{m}\|_2^2. \end{aligned} \quad (12)$$

In the training of an NN, a regularization term is often added to the loss function. The L_2 regularization weighted by some weight decays λ is often used, and we can obtain the following equation to minimize the objective:

$$\mathcal{L}_{\text{dropout}} := E + \lambda_1 \|\mathbf{W}_1\|_2^2 + \lambda_2 \|\mathbf{W}_2\|_2^2 + \lambda_3 \|\mathbf{b}\|_2^2, \quad (13)$$

where E is the loss function.

The optimal parameters for maximizing Eq. (12) lead to the same as those for minimizing Eq. (13) if the weight decays in

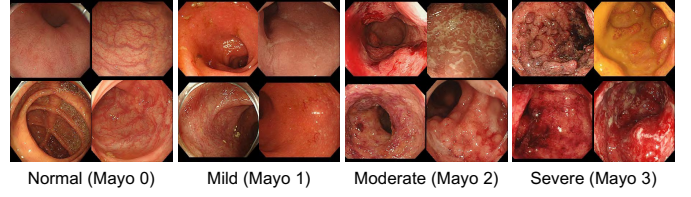


Figure 3: Examples of endoscopic images of ulcerative colitis at each Mayo (severity).

Eq. (13) are properly determined. The two equations converge to the same limit with the correct stochastic optimizer. The above results prove that MC dropout can be applied to pairwise LTR with the Siamese network structure.

4.2. Acquisition function for pairwise ranking

The predictive distribution of the Siamese network model is represented as a joint probability, such as $p(\mathbf{Y}_1, \mathbf{Y}_2 | \mathbf{X}_1, \mathbf{X}_2)$ in Eq. (5). In addition, \mathbf{Y}_1 and \mathbf{Y}_2 are independent of each other, and thus the distribution can be expressed as follows:

$$\begin{aligned} p(\mathbf{Y}_1, \mathbf{Y}_2 | \mathbf{X}_1, \mathbf{X}_2, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}) \\ = p(\mathbf{Y}_1 | \mathbf{X}_1, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}) p(\mathbf{Y}_2 | \mathbf{X}_2, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}). \end{aligned} \quad (14)$$

Therefore, we define the variance of the posterior distribution of the output value, y^* , for the single input x^* , as shown in Eq. (3), as an acquisition function of the model uncertainty.

5. Experiment

To evaluate the effectiveness of our method, we conducted experiments on the disease severity estimation task using one private and one public dataset for UC severity. In these experiments, we evaluated the estimation performance of the proposed method for two use cases. In the first use case, we examine the accuracy of estimating the relative labels, that is, correctly identifying the image with higher severity in a given endoscopic image pair. The objective of this evaluation is to determine the effectiveness of treatment and to assess changes in severity during follow-up in clinical practice. We quantitatively compare the proposed method to baseline methods and also analyze the relationship between uncertainty and class prior distribution for the datasets obtained by active learning.

In the second use case, we evaluate the performance of estimating UC severity levels in classification tasks. By attaching additional absolute labels to the dataset obtained by active learning, we estimate UC severity classification by multi-task learning (Kadota et al., 2022a) combining LTR and regression. The performance and annotation cost of the proposed method are compared with those of conventional multi-class classification methods.

Table 1: Quantitative performance evaluation of the accuracy of estimating relative labels. The labeling ratio shows the percentage of relative labels used in training. The labeling ratio of 100% indicates that the labels were created from all training data using the pairing method described in Section 5.1.2. ‘**’ indicates a statistically significant difference between the proposed method and each compared method at $p < 0.05$ by multiple statistical comparisons using McNemar’s test.

Data	Method	Labeling ratio	Overall	Neighboring			
				0–1	1–2	2–3	Mean
Private	Baseline	50%	0.861*	0.827*	0.837*	0.628*	0.763*
	Baseline (all data)	100%	0.875	0.855*	0.870	0.635*	0.785
	Core-set	50%	0.851*	0.785	0.827*	0.632*	0.747*
	Proposed w/o UBS	50%	0.856*	0.818*	0.842*	0.634*	0.763*
	Proposed	50%	0.880	0.787	0.871	0.736	0.797
Public	Baseline	50%	0.838*	0.803	0.748*	0.731*	0.760*
	Baseline (all data)	100%	0.878	0.827*	0.806	0.778*	0.804
	Core-set	50%	0.836*	0.802	0.741*	0.705*	0.749*
	Proposed w/o UBS	50%	0.857*	0.804	0.776*	0.749*	0.776*
	Proposed	50%	0.882	0.793	0.813	0.806	0.804

5.1. Relative severity estimation

5.1.1. Dataset

To examine the performance of the proposed method for relative label estimation, we used two datasets (one private and one public) with absolute severity labels for UC.

The private dataset contains 10,265 endoscopic images from 388 ulcerative colitis (UC) patients at Kyoto Second Red Cross Hospital. The Ethical Review Committee of Kyoto Second Red Cross Hospital approved the experiments using the private dataset. Multiple medical experts carefully annotated a Mayo score, which determines UC severity on a four-point scale (Mayo 0–3) for each image. Figure 3 shows examples of endoscopic images in which the Mayo score was determined. Schroeder et al. (Schroeder et al., 1987) defined a scoring system to assess UC activity as Mayo scores in which the endoscopic findings of UC in the system are divided into four stages: Mayo 0 is a normal or inactive disease, Mayo 1 is a mild disease with erythema, decreased vascular pattern, and mild friability, Mayo 2 is a moderate disease with marked erythema, absent vascular pattern, collapse, and erosion, and Mayo 3 is a severe disease with spontaneous bleeding and ulceration. The private dataset has a large class imbalance and contains 6,678, 1,995, 1,395, and 197 samples for Mayo 0, 1, 2, and 3, respectively. Note that medical imaging datasets usually have class imbalances because the number of patients with normal or mild disease is typically larger than the number of patients with severe disease.

The public dataset we used is the LIMUC dataset (Polat et al., 2022b) annotated with Mayo scores for UC. This dataset contains 11,276 images from 564 patients, all annotated by at least two medical experts. The public dataset also has class imbalance: Mayo 0, Mayo 1, Mayo 2, and Mayo 3 are 6,105, 3,052, 1,254, and 865, respectively. In principle, the same settings were used in the evaluation of all experiments with the private dataset.

5.1.2. Evaluation metrics

We defined the accuracy of relative label estimation, namely, the percentage of relative labels correctly estimated for pairs,

as a performance metric. In this experiment, we created pairs of two images and attached a relative label to each pair based on the Mayo score. Estimated relative labels were attached by comparing the rank scores of the image pairs. According to Li et al., the severity rank score obtained from the Siamese network can provide a continuous measure of the change in disease severity between images (Li et al., 2020).

The number of pairs ($O(N^2)$) that can be created from the N samples is too large when creating image pairs, so we limited the number of pairs for random sampling as follows. By selecting one sample from $N - 1$ samples for each N sample, we created N pairs (instead of $O(N^2)$ pairs). We also utilized this pair creation method in Section 3 for selecting the initial set of R samples. This setting, which is typical for pairwise LTR evaluations (Kadota et al., 2022a; Xu et al., 2021; You et al., 2019), was used for all pair creations.

We performed five-fold cross-validation on all methods. The datasets were divided into training (60%), validation (20%), and test (20%) sets using patient-based sampling to ensure that images from the same patient were not included in different sets. For a fair evaluation, we created the image pairs within each set after dividing the training, validation, and test sets. Therefore, there were no duplicate images or image pairs between each set.

5.1.3. Implementation details

The experimental environment was Ubuntu 18.04 and two NVIDIA TITAN RTX 24GB GPUs. We implemented our method with Tensorflow 1.13.1 and Keras 2.2.4. We used the DenseNet-169 structure (Huang et al., 2017) as the backbone of the Bayesian CNN. We trained the CNN with dropout ($p_{\text{dropout}} = 0.2$) and weight decay ($\lambda = 1 \times 10^{-4}$) settings in the convolutional and fully connected layers. We used the Adam optimizer to optimize weight parameters. The initial learning rates were set to 1×10^{-5} for the private dataset and 2×10^{-5} for the public dataset. All images in the datasets were resized to 224×224 pixels and normalized to values between 0 and 255.

The hyperparameters for active learning in all experiments were set to $K = 6$ for the number of iterations, $s = 5\%$ for the selection rate from the training data at each sampling, $r = 20\%$

for the selection rate from the training data when sampling the initial training data, and $T = 30$ for the number of estimates for uncertainty estimation. The initial annotation ratio $r = 20\%$ was selected as a realistically low-cost option in the relative annotation scenario. The annotation ratio of the labeled dataset after all iterations (6 in this case) was 50% ($r + sK = 20 + 5 \times 6$), assuming 100% for the case with all training data.

5.1.4. Compared methods

The proposed method was compared with four methods, which are referred to as ‘baseline,’ ‘baseline (all data),’ ‘proposed w/o UBS,’ and ‘Core-set’ (Sener and Savarese, 2018). The baseline was trained by randomly sampling $r + sK$ pairs of training data (the same number as for the proposed method). The baseline (all data) was trained using N pairs created using all training samples. The annotation ratio of the baseline (all data) was 100%, which is twice the maximum annotation ratio of the proposed method (that is, $r + sK = 50\%$). The proposed w/o UBS was trained with increasing training data by K iterations using random sampling, instead of uncertainty-based sampling (UBS) to confirm the effect of UBS. Core-set represents a diversity-based sampling strategy that was used in place of UBS to assess its efficacy. Core-set utilized 1664-dimensional features extracted from DenseNet-169 trained on ImageNet. Note that all methods shared the same backbone architecture, DenseNet-169-based CNN, to ensure a fair comparison. For all methods, rank scores were calculated based on the average of the predictions estimated $T = 30$ times using Bayesian CNN.

5.1.5. Quantitative evaluation with test data

We evaluated the accuracy of the relative severity estimation between the image pairs by comparing the estimated rank scores. Specifically, given a pair (x_i, x_j) where x_i is more severe than x_j , the estimate is considered “accurate” if $f(x_i) > f(x_j)$.

Relative severity estimation is typically utilized in clinical practice to assess changes in disease severity. In general, it is difficult to compare severity when the difference in severity between pairs is small. Therefore, we prepared two types of test sets according to the difference in severity between pairs. Note that these test sets are obtained from virtual relative annotations with relative labels attached using Mayo labels.

“Overall” case: We prepared a randomly paired test set from all Mayo scores. Specifically, we selected images so that the number of samples for each Mayo score was equal and randomly created pairs from the selected images. This test set contains many easy pairs, such as pairs with Mayo 0 (normal images) and Mayo 3 (severe images), which have very different degrees of severity, as shown in Fig. 3. This test set was used to evaluate the overall performance of each method.

“Neighboring” case: We created a neighboring pair test set with Mayo 0–1, Mayo 1–2, and Mayo 2–3, paired from neighboring Mayo score images. Neighboring pairs are similar in severity, making it difficult to determine their relative severity. In clinical applications, it is important to achieve high accuracy in cases where severity comparison is difficult. This test set evaluates the performance of each method in difficult cases.

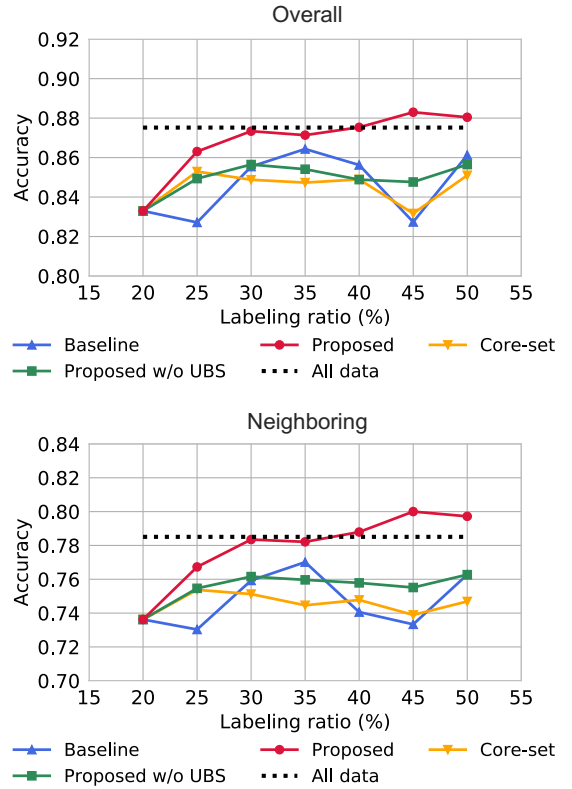


Figure 4: Accuracy of relative label estimates for baseline (blue), Core-set (orange), proposed w/o UBS (green), and proposed method (red) at each labeling ratio. The black dotted line indicates the result of baseline (all data).

Table 1 shows the mean accuracy of relative label estimation by each method in five-fold cross-validation. The labeling ratio column shows the percentage of relative labels used in training. The labeling ratio of 100% indicates that the labels were created from all training data using the pairing method described in Section 5.1.2. ‘*’ denotes a statistically significant difference at $p < 0.05$ by McNemar’s test using Holm’s method of multiple statistical comparisons.

In the results for the “Overall” test set, the proposed method achieved a higher performance than all other methods. In particular, it outperformed the baseline (all data) despite using half the amount of training data. The performances of the baseline and the proposed w/o UBS were lower than those of the baseline (all data). These methods used a smaller size (by half) of the training data than the baseline (all data). This difference is due to class imbalances in the training data, as described in Section 5.1.6 below. We used the two datasets with large class imbalances, where the number of samples decreases as the severity increases. The proposed method mitigated class imbalances in the training data and improved the performance by automatically selecting samples from minority classes.

In the results for “Neighboring” (difficult case), the proposed method outperformed all other methods in terms of mean accuracy for neighbor pairs. The performance comparison for each pair showed that the proposed method performed better than the other methods on “Mayo 1–2” and “Mayo 2–3”. In particular,

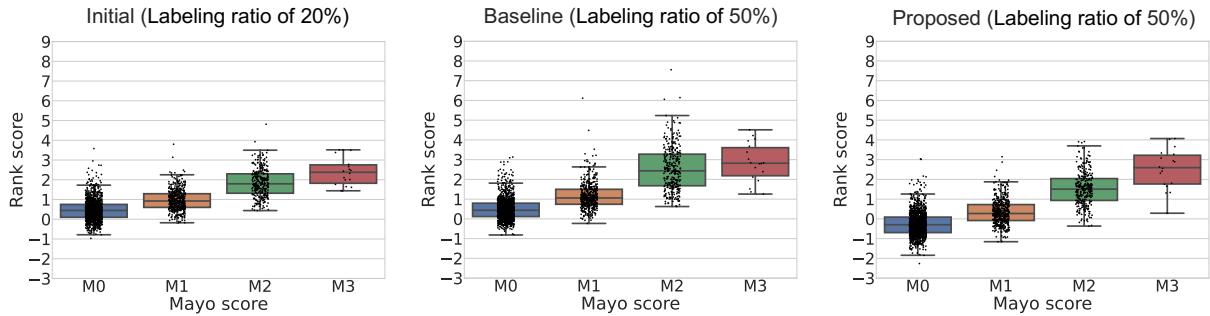


Figure 5: Box plots of estimated rank scores at each Mayo score. The initial labeling ratio was measured with 20% (iteration $K = 0$). The results of the baseline and the proposed method estimates were measured with a labeling ratio of 50% (iteration $K = 6$). The estimate is considered reasonable if there is little overlap in the distribution of rank scores for each Mayo score.

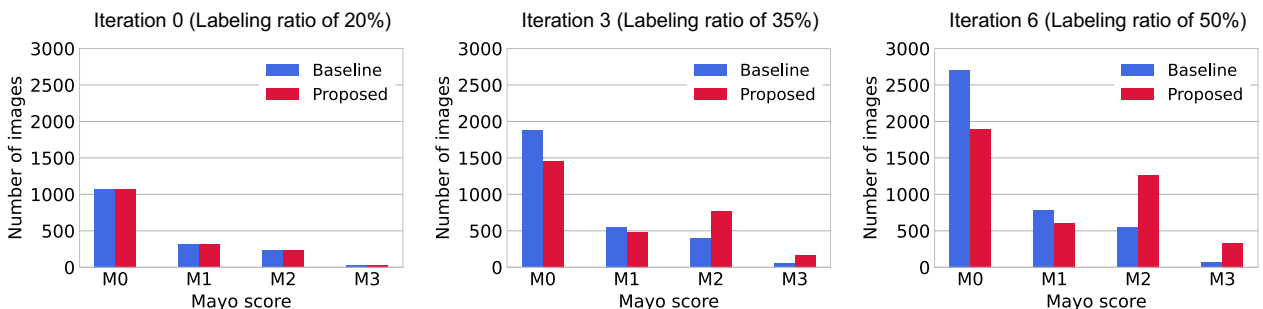


Figure 6: Class proportions of each Mayo score for accumulated sampled images at iterations $K = 0$ (labeling ratio of 20%), $K = 3$ (labeling ratio of 35%), and $K = 6$ (labeling ratio of 50%). For a labeling ratio of 20%, the class proportions were the same for the baseline and the proposed method. The proposed method mitigates the class imbalance problem by selecting more samples from minority classes (Mayo 2 and 3).

the proposed method improved the accuracy of “Mayo 2–3” by more than 10% compared to the other methods on the private dataset. The more severe the disease, the greater the need for treatment, so evaluating treatment effects on severe diseases is critical. The proposed method performed better on severe image pairs and is thus considered superior to the other methods in terms of clinical application. However, the results of “Mayo 0–1” showed that the accuracy of the proposed method was lower than that of the other methods. This is because the proposed method mitigates the class imbalance in the training data and thus has fewer images labeled “Mayo 0–1” in the training data.

Figure 4 shows the change in the accuracy of estimating relative labels at each iteration (each labeling ratio) for the “Overall” and “Neighboring” test sets. The horizontal axis is the labeling ratio, the vertical axis is the mean accuracy for five-fold cross-validation, and the black dotted line is the baseline (all data) result with a labeling ratio of 100%. As we can see, the accuracy of the proposed method (red) improved as the amount of training data increased, and the improvement was more pronounced than that of the other methods. The accuracy of the proposed method was better than that of the baseline (all data) when the labeling ratio was 40% in the “Neighboring” test set. In contrast, the baseline (blue), the proposed w/o UBS (green), and Core-set (orange) showed only marginal improvement in accuracy as the amount of training data increased, with limited gains from the initial training. As shown in Table 1, the accuracies of the compared methods were lower than that of

the proposed method, especially for Mayo 2–3, which is a pair of minority classes. The lack of significant accuracy increase over iterations in the compared methods can be attributed to the effect of class imbalance in the pair creation. These results indicate that active learning with uncertainty-based sampling effectively selects highly effective pairs for learning.

Figure 5 shows box plots of the estimated rank scores at each Mayo score from the baseline and the proposed method. The horizontal axis shows each Mayo score, and the vertical axis shows the mean estimated rank scores. The left figure, “Initial”, shows the method where no iterations were performed, and the network was trained using only the initial training data (labeling ratio of 20%). The baseline and the proposed method refer to the results obtained using the training data with a labeling ratio of 50%. In the box plots, we consider the rank scores to be reasonable when there is little overlap in the distribution of rank scores at each Mayo score, as the estimated difference in severity is clear. In the initial and the baseline results, the rank score distributions of Mayo 2 and 3 overlapped significantly, while in contrast, the proposed method reduced the overlap. These results indicate that the estimated rank score and the Mayo score are highly correlated in the proposed method.

5.1.6. Relationship between uncertainty and class imbalance

As discussed in Section 5.1.5, the proposed method significantly improved the accuracy of relative label estimation by automatically sampling from minority classes. These results

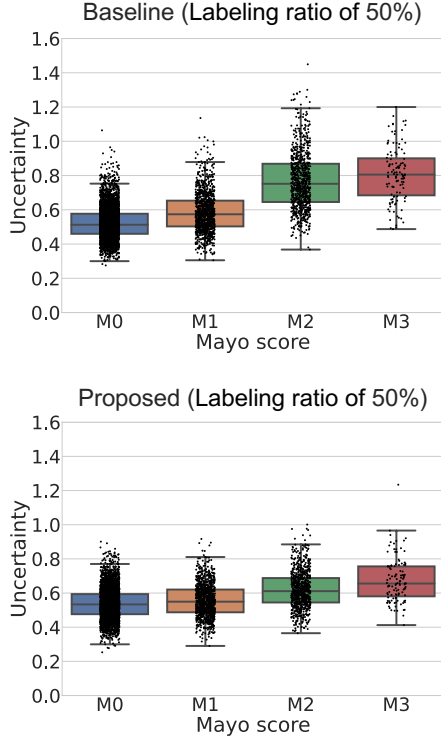


Figure 7: Box plots of model uncertainty for each Mayo score for the baseline and the proposed method. The performance of each method was measured at a labeling rate of 50%. In the baseline, the minority classes (Mayo 2 and 3) had higher uncertainty than the majority class. The proposed method mitigated the class imbalance and reduced the uncertainty in Mayo 2 and 3.

suggest a relationship between the model uncertainty using the Bayesian CNN and the minority classes in the dataset. Therefore, we examined the number of samples of each Mayo score and the uncertainty of the sampled images in the training data updated by iterations.

Figure 6 shows the class proportions of the Mayo score of the samples in the training data of the baseline and the proposed method when the labeling ratio is 20% ($k = 0$), 35% ($k = 3$), and 50% ($k = K = 6$). The horizontal axis shows the Mayo scores, and the vertical axis shows the average number of images in the five-fold cross-validation. We analyzed the difference between uncertainty-based and random sampling by the number of images in each Mayo score. The initial training data in each method (labeling ratio of 20%) had the same class imbalance as the entire dataset due to random sampling. In addition, at the labeling ratios of 35% and 50% in the baseline, the training data had class imbalances similar to the initial training data. In the baseline, this class imbalance in the training data affected the performance. Thus, the performance improvement was limited despite increasing the number of images in the training data. In contrast, the proposed method gradually mitigated the class imbalance by selecting more minority class images (Mayo 2 and 3) as the number of iterations increased. As a result, the accuracy of relative label estimation was significantly improved, even though the amount of training data was half that of the baseline (all data).

Figure 7 shows the model uncertainty distribution of the sample at each Mayo score on the training data for the baseline and the proposed method. These results were obtained using the training data with a labeling ratio of 50%. The horizontal axis shows the Mayo scores, and the vertical axis shows the uncertainty of the samples in the training data. In the baseline, the uncertainty in the minority classes (Mayo 2 and 3) is higher than in the majority classes (Mayo 0 and 1). In the proposed method, the uncertainty of the sample at each Mayo score is not much different, and the uncertainty for Mayo 2 and 3 is lower than the baseline uncertainty. These results indicate that the uncertainty is correlated with the class imbalance, and the smaller the sample size of the class, the higher the uncertainty. Therefore, the proposed method achieved a high performance thanks to using class-balanced training data obtained by uncertainty-based sampling as discussed in Section 5.1.5. The proposed method is thus useful for ranking tasks in medical images because serious class imbalance problems often occur in medical image datasets.

5.1.7. Qualitative evaluation

We examined the distribution of the samples selected by UBS in the feature space. Figure 8 shows the distribution of selected samples at the first iteration just after the initial training and Mayo scores in the feature space using t -SNE. In this visualization, we extracted 1664-dimensional features from images using DenseNet-169 trained on ImageNet and compressed these features into a two-dimensional map with t -SNE. The left panel shows the distribution of all training data and the corresponding Mayo scores. The middle and right panels show the differences in the distribution of selected samples using the proposed method, based on the presence or absence of UBS. The proposed method without UBS selected samples uniformly, predominantly choosing samples from Mayo 0, the majority class. In contrast, with UBS, the proposed method selected samples from minority classes more frequently, thereby reducing class imbalance.

5.2. Multi-class classification

5.2.1. Dataset

To investigate the effectiveness of the proposed method in classification tasks, we conducted experiments to evaluate its performance for multi-class classification. In these experiments, we used the private dataset with large class imbalances, and the test data preserved the class imbalance. We also used five-fold cross-validation, and the dataset was divided into training (60%), validation (20%), and test (20%) sets with patient-based sampling.

5.2.2. Evaluation metrics

We used precision, recall, and F1-score as performance metrics for multi-class classification. We evaluated classification performance primarily on F1-scores because the dataset has a large class imbalance. The rank scores obtained from the trained CNN are not categorical scales and cannot be used in classification evaluations without modification. Therefore, we

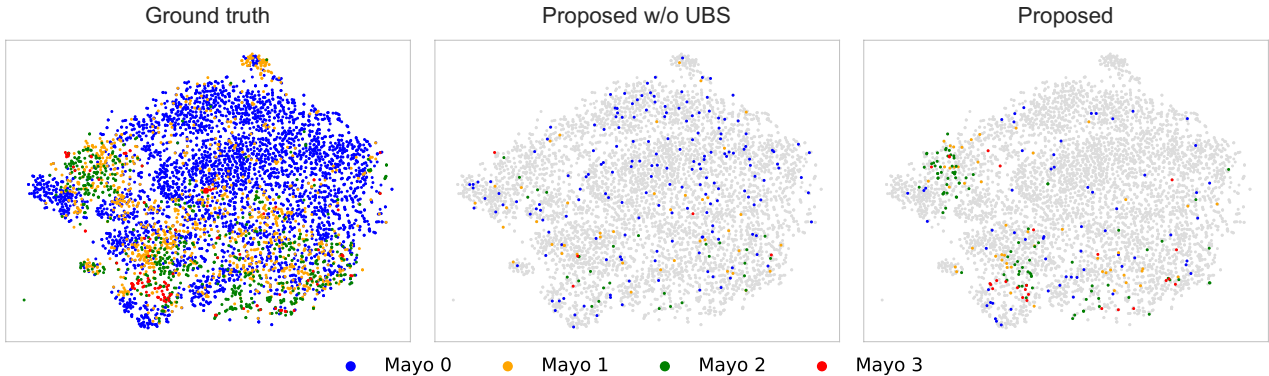


Figure 8: t -SNE visualization of the feature distribution and the sampling results after initial training. In the middle and right panels, colored dots represent selected images.

Table 2: Comparison of the number of relative and absolute labels and the annotation time for training data after additional absolute annotations.

Method	Labels		Time (s)*
	Relative	Absolute	
Conventional	0	8,214	164,280
Baseline**	4,106	4,106	86,226
Baseline (all data)	8,214	8,214	172,494
Proposed w/o UBS**	4,106	3,378	71,666
Proposed**	4,106	2,475	53,606

* Relative: 1 (s/pair), Absolute: 20 (s/image)

** Relative labeling ratio of 50%

Table 3: Classification performance evaluation on test data.

Method	Precision	Recall	F1-score
Conventional	0.626	0.642	0.629
Baseline	0.661	0.623	0.627
Baseline (all data)	0.632	0.655	0.640
Proposed w/o UBS	0.629	0.634	0.620
Proposed	0.682	0.641	0.649

determined the class by quantizing the rank score to the nearest integer for converting the rank scores to the discrete severity classes (Mayo scores). For example, a rank score of 1.7 is classified as Mayo 2.

5.2.3. Compared methods

We compared the proposed method with baseline, baseline (all data), and proposed w/o UBS in Section 5.1.4, and a conventional CNN-based method. In the conventional method, the CNN backbone used DenseNet-169 (the same as the proposed method) and was trained with categorical cross-entropy as the loss function. All 8,214 training samples with absolute labels (Mayo scores) were used to train the conventional method.

5.2.4. Annotation efficiency evaluation

Table 2 shows the relative and absolute labels and annotation costs of the training data for each method in the classification tasks. We used the training data with a relative labeling ratio of

50% in the baseline, the proposed w/o UBS, and the proposed method. Annotation times were calculated as follows. The relative annotation takes one second per pair, and the absolute annotation takes 20 seconds per image (Kadota et al., 2022a). As shown in Table 2, the annotation time of the proposed method was less than that of all other methods and was as low as about one-third of that of the conventional method. Interestingly, the proposed method had about 900 fewer absolute labels than the proposed w/o UBS. The proposed method preferentially selects minority class images with the highest learning effect to create pairs. We believe the number of absolute labels was reduced because the proposed method repeatedly selected the same minority class images during active learning. Note that the proposed method always generates non-duplicate pairs and performs relative annotation, even if active learning selects the same image repeatedly.

5.2.5. Classification performance evaluation

Table 3 shows the classification performance of each method. As we can see, the F1-score of the proposed method was higher than all other methods. In particular, it performed better than the baseline (all data) trained with all training images. We found that the proposed method can preferentially select the more effective images for learning in classification. In addition, the proposed method achieved a higher F1-score than the conventional method. These findings demonstrate the superiority of the proposed method in terms of significantly reducing the annotation cost for disease severity classification.

6. Conclusion

We proposed a deep Bayesian active learning-to-rank for efficient relative annotation by uncertainty-based sampling. We theoretically proved that MC dropout can be applied to estimate the model uncertainty of a pairwise LTR using a Bayesian Siamese neural network. The proposed method actively determines effective sample pairs for additional relative annotation by estimating the model uncertainty of the samples using Bayesian CNN. Experimental results showed that the proposed

method achieves high performance with a small number of samples by selecting highly effective images for learning in ranking and classification tasks. We also found that it is robust to class imbalances because it selects a minority but significant samples. The proposed method has two drawbacks. First, it is computationally expensive because it requires multiple estimations of output values to obtain uncertainty. In future work, we plan to investigate an active learning-to-rank approach that allows for multiple estimations in a short time. Second, uncertainty may become inaccurate for datasets from domains that are different from the training dataset. We will explore domain generalization for an active learning-to-rank approach that calibrates uncertainty across multiple domains.

Declaration of Competing Interest

None.

CRedit authorship contribution statement

Takeaki Kadota: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft. **Hedeaki Hayashi:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Supervision, Writing - review & editing. **Ryoma Bise:** Funding acquisition, Methodology, Supervision, Writing - review & editing. **Kiyohito Tanaka:** Resources, Data curation. **Seiichi Uchida:** Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - review & editing.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP20H04211, JP21H03511, and JP23K18509, AMED Grant Number JP201k1010036h0002, and JST SPRING Grant Number JPMJSP2136.

References

Becker, B.G., Arcadu, F., Thalhammer, A., Serna, C.G., Feehan, O., Drawnel, F., Oh, Y.S., Prunotto, M., 2021. Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data. *Therapeutic advances in gastrointestinal endoscopy* 14.

Beluch, W.H., Genewein, T., Nurnberger, A., Kohler, J.M., 2018. The power of ensembles for active learning in image classification, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9368–9377.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G., 2005. Learning to rank using gradient descent, in: Proceedings of the 22nd international conference on Machine learning (ICML), pp. 89–96.

Carterette, B., Petkova, D., 2006. Learning a Ranking from Pairwise Preferences, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 629–630.

Cho, B.J., Bang, C.S., Park, S.W., Yang, Y.J., Seo, S.I., Lim, H., Shin, W.G., Hong, J.T., Yoo, Y.T., Hong, S.H., et al., 2019. Automated classification of gastric neoplasms in endoscopic images using a convolutional neural network. *Endoscopy* 51, 1121–1129.

Dasgupta, S., Hsu, D., 2008. Hierarchical sampling for active learning, in: Proceedings of the 25th international conference on Machine learning, pp. 208–215.

Gal, Y., Ghahramani, Z., 2016a. Dropout as a bayesian approximation: Appendix. arXiv preprint arXiv:1506.02157.

Gal, Y., Ghahramani, Z., 2016b. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, in: Proceedings of the 33rd International Conference on Machine Learning (ICML), pp. 1050–1059.

Gal, Y., Islam, R., Ghahramani, Z., 2017. Deep Bayesian active learning with image data, in: Proceedings of the 34th International Conference on Machine Learning (ICML), pp. 1183–1192.

Gorritz, M., Carlier, A., Faure, E., Giro-i Nieto, X., 2017. Cost-effective active learning for melanoma segmentation, in: Machine Learning for Health Workshop at NIPS (ML4H), pp. 1–5.

Hirai, F., Matsui, T., 2008. A critical review of endoscopic indices in ulcerative colitis: inter-observer variation of the endoscopic index. *Clinical J. Gastroenterology* 1, 40–45.

Hofmann, K., Whiteson, S., de Rijke, M., 2013. Balancing exploration and exploitation in listwise and pairwise online learning to rank for information retrieval. *Information Retrieval* 16, 63–90.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269.

Kadota, T., Abe, K., Bise, R., Kawamura, T., Sakiyama, N., Tanaka, K., Uchida, S., 2022a. Automatic Estimation of Ulcerative Colitis Severity by Learning to Rank With Calibration. *IEEE Access* 10, 25688–25695.

Kadota, T., Hayashi, H., Bise, R., Tanaka, K., Uchida, S., 2022b. Deep Bayesian Active-Learning-to-Rank for Endoscopic Image Data, in: Medical Image Understanding and Analysis (MIUA), pp. 609–622.

Kalpathy-Cramer, J., Campbell, J.P., Erdogmus, D., Tian, P., Kedarisetti, D., Moleta, C., Reynolds, J.D., Hutcheson, K., Shapiro, M.J., Repka, M.X., et al., 2016. Plus disease in retinopathy of prematurity: improving diagnosis by ranking disease severity and using quantitative image analysis. *Ophthalmology* 123, 2345–2351.

Klang, E., Grinman, A., Soffer, S., Margalit Yehuda, R., Barzilay, O., Amitai, M.M., Konen, E., Ben-Horin, S., Eliakim, R., Barash, Y., et al., 2021. Automated detection of Crohn's disease intestinal strictures on capsule endoscopy images using deep neural networks. *Journal of Crohn's and Colitis* 15, 749–756.

Leaman, R., Islamaj Doğan, R., Lu, Z., 2013. DNORM: disease name normalization with pairwise learning to rank. *Bioinformatics* 29, 2909–2917.

Li, M.D., Chang, K., Bearce, B., Chang, C.Y., Huang, A.J., Campbell, J.P., Brown, J.M., Singh, P., Hoebel, K.V., Erdoğmuş, D., et al., 2020. Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. *NPJ digital medicine* 3, 1–9.

Liu, G., Hua, J., Wu, Z., Meng, T., Sun, M., Huang, P., He, X., Sun, W., Li, X., Chen, Y., 2020. Automatic classification of esophageal lesions in endoscopic images using a convolutional neural network. *Annals of translational medicine* 8.

Liu, T.Y., 2009. Learning to Rank for Information Retrieval. *Foundations and Trends® in Information Retrieval* 3, 225–331.

Liu, X., van de Weijer, J., Bagdanov, A.D., 2017. RankIQA: Learning From Rankings for No-Reference Image Quality Assessment, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1040–1049.

Lyu, J., Ling, S.H., Banerjee, S., Zheng, J., Lai, K.L., Yang, D., Zheng, Y.P., Bi, X., Su, S., Chamoli, U., 2021. Ultrasound volume projection image quality selection by ranking from convolutional RankNet. *Computerized Medical Imaging and Graphics* 89, 101847.

Ma, K., Liu, W., Liu, T., Wang, Z., Tao, D., 2017. dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing* 26, 3951–3964.

Nair, T., Precup, D., Arnold, D.L., Arbel, T., 2020. Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation. *Medical Image Analysis* 59, 101557.

Parikh, D., Grauman, K., 2011. Relative Attributes, in: Proceedings of the 2011 International Conference on Computer Vision (ICCV), pp. 503–510.

Polat, G., Ergenc, I., Kani, H.T., Alahdab, Y.O., Atug, O., Temizel, A., 2022a. Class distance weighted cross-entropy loss for ulcerative colitis severity estimation, in: Medical Image Understanding and Analysis (MIUA), pp. 157–171.

Polat, G., Kani, H., Ergenc, I., Alahdab, Y., Temizel, A., Atug, O., 2022b. Labeled images for ulcerative colitis (limuc) dataset. Accessed March URL: <https://doi.org/10.5281/zenodo.5827695>.

- Saibro, G., Diana, M., Sauer, B., Marescaux, J., Hostettler, A., Collins, T., 2022. Automatic Detection of Steatosis in Ultrasound Images with Comparative Visual Labeling, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 408–418.
- Schroeder, K.W., Tremaine, W.J., Ilstrup, D.M., 1987. Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. *The New England Journal of Medicine* 317, 1625–1629.
- Sener, O., Savarese, S., 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach, in: International Conference on Learning Representations (ICLR), pp. 1–13.
- Smailagic, A., Costa, P., Gaudio, A., Khandelwal, K., Mirshekari, M., Fagert, J., Walawalkar, D., Xu, S., Galdran, A., Zhang, P., et al., 2020. O-MedAL: Online active deep learning for medical image analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, e1353.
- Sourati, J., Gholipour, A., Dy, J.G., Kurugol, S., Warfield, S.K., 2018. Active deep learning with fisher information for patch-wise semantic segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (ML-CDS), pp. 83–91.
- Souri, Y., Noury, E., Adeli, E., 2016. Deep relative attributes, in: Asian conference on computer vision, pp. 118–133.
- Stidham, R.W., Liu, W., Bishu, S., Rice, M.D., Higgins, P.D., Zhu, J., Nallamothu, B.K., Waljee, A.K., 2019. Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. *JAMA network open* 2, e193963.
- Takenaka, K., Ohtsuka, K., Fujii, T., Negi, M., Suzuki, K., Shimizu, H., Oshima, S., Akiyama, S., Motobayashi, M., Nagahori, M., Saito, E., Matsuoka, K., Watanabe, M., 2020. Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis. *Gastroenterology* 158, 2150–2157.
- Tang, S., Yu, X., Cheang, C.F., Liang, Y., Zhao, P., Yu, H.H., Choi, I.C., 2023. Transformer-based multi-task learning for classification and segmentation of gastrointestinal tract endoscopic images. *Computers in Biology and Medicine* 157, 106723.
- Thapa, S.K., Poudel, P., Bhattarai, B., Stoyanov, D., 2022. Task-aware active learning for endoscopic image analysis. *arXiv preprint arXiv:2204.03440*.
- Wang, Q., Wu, W., Qi, Y., Zhao, Y., 2021. Deep Bayesian Active Learning for Learning to Rank: A Case Study in Answer Selection. *IEEE Transactions on Knowledge and Data Engineering*.
- Wen, S., Kurc, T., Hou, L., Saltz, J., Gupta, R., Batiste, R., Zhao, T., Nguyen, V., Samaras, D., Zhu, W., 2018. Comparison of Different Classifiers with Active Learning to Support Quality Control in Nucleus Segmentation in Pathology Images. *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science 2017*, 227–236.
- Xu, Q., Yang, Z., Chen, Z., Jiang, Y., Cao, X., Yao, Y., Huang, Q., 2021. Deep Partial Rank Aggregation for Personalized Attributes, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 678–688.
- Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z., 2017. Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 399–407.
- You, Y., Lu, C., Wang, W., Tang, C.K., 2019. Relative CNN-RNN: Learning Relative Atmospheric Visibility From Images. *IEEE Transactions on Image Processing* 28, 45–55.