IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 10, 2022, accepted February 23, 2022, date of publication March 2, 2022, date of current version March 10, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3155769

Automatic Estimation of Ulcerative Colitis Severity by Learning to Rank With Calibration

TAKEAKI KADOTA[1], KENTARO ABE[1], RYOMA BISE[1,2], (Member, IEEE), TAKUJI KAWAMURA[3], NAOKUNI SAKIYAMA[3], KIYOHITO TANAKA[3], AND SEIICHI UCHIDA[1,2], (Member, IEEE)

[1]Department of Advanced Information Technology, Kyushu University, Fukuoka 819-0395, Japan
[2]Research Center for Medical Bigdata, National Institute of Informatics, Tokyo 101-8430, Japan
[3]Department of Gastroenterology, Kyoto Second Red Cross Hospital, Kyoto 602-8026, Japan

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

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP20H04211, and in part by the Japan Agency for Medical Research and Development (AMED) under Grant JP20lk1010036h0002.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Kyoto Second Red Cross Hospital.

ABSTRACT For automatic disease-severity-level estimation, a large-scale medical image dataset with level annotations is generally necessary. However, attaching absolute-level annotations (such as levels 0, 1, and 3) is very costly and even inaccurate due to the level ambiguity. In this study, we proved experimentally that using a ranking function for level estimation can relax this difficulty. We propose a multi-task learning method for automatically estimating disease-severity levels that combine learning to rank with regression. The ranking function of the proposed method is trainable by relative-level and a small number of absolute-level annotations. For relative-level annotation, an annotator only needs to specify that one image has a higher disease level than another—this is much easier than absolute-level annotation. The proposed method enables disease-severity classification by calibrating the ranking function based on relative-level annotation through regression. The effectiveness of the method was proved through a large-scale experiment of ulcerative colitis-severity estimation with colonoscopy images.

INDEX TERMS Computer-aided diagnosis, deep learning, endoscopic image dataset, learning to rank, relative-level annotation.

I. INTRODUCTION

To realize automatic disease-severity-level estimation, we often prepare a dataset with level annotation. Fig. 1 (a) shows an absolute-level annotation, where an annotator attaches an absolute disease level to each image. Using the annotated dataset, we can estimate the disease-severity level by using a regression method or classification method.

Even for medical specialists, attaching accurate absolute-level annotations is difficult. This is because the level of disease is inherently continuous with gradual tissue and organ changes; thus, discrete levels such as absolute levels always have quantization errors. For example, even when a four-level annotation (0, 1, 2, 3) is requested, they easily find medical images that should be ''level 1.5''. Moreover, the level itself easily fluctuates among annotators (e.g., [1]) or even with the same annotator.

The purpose of this study was to relax this difficulty by using *relative-level annotation* instead of the above absolute-level annotation. Fig. 1 (b) shows the idea of the relative-level annotation. Given a pair of images (x_i, x_j), the annotator just specifies the image with a higher severity level. This task is far easier and even more accurate than the absolute-level annotation, especially when the paired images have a clear severity-level difference. Therefore, the difficulty in annotating a large number of images can be greatly reduced by using relative-level annotation.

Given a dataset with the relative-level annotation, we can automatically estimate severity level by training a *ranking function* $f(x)$. Fig. 1 (c) shows the idea of the so-called bipartite ranking problem. The basic objective of this problem is to train a function $f(x)$ that maximizes the number of sample pairs whose the relative-level annotation

The associate editor coordinating the review of this manuscript and approving it for publication was Sudipta Roy.

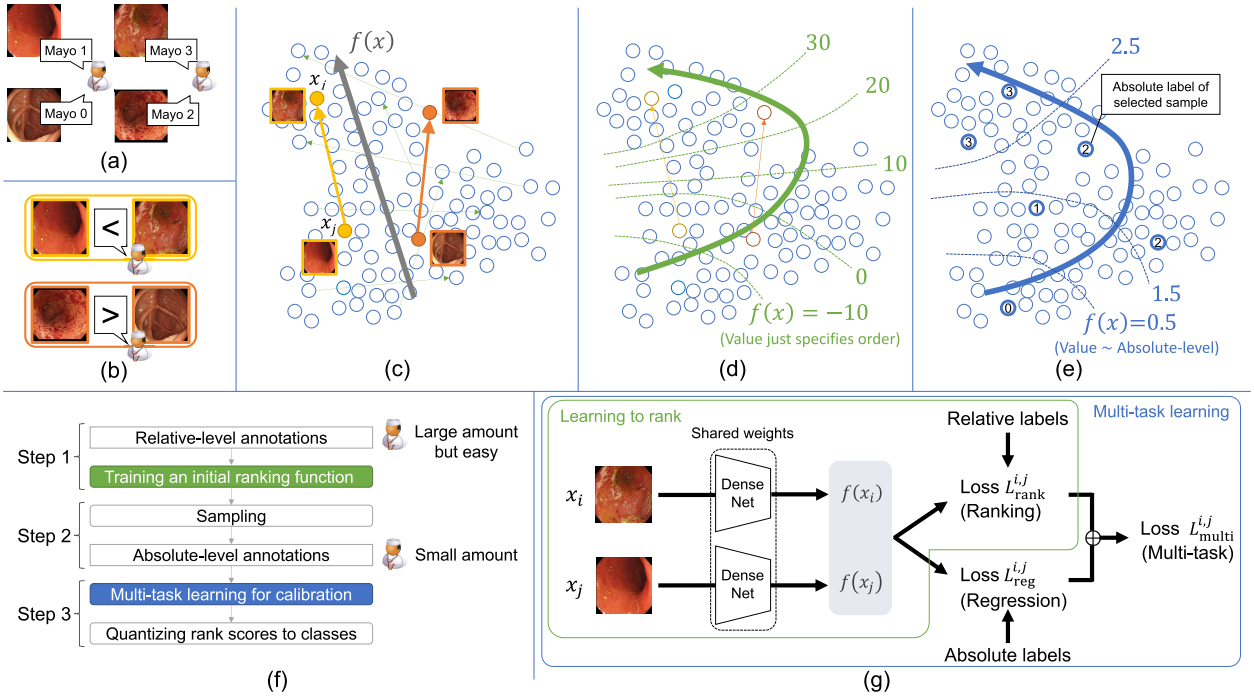


FIGURE 1. **(a) Absolute-level annotation. (b) Relative-level annotation. (c) Bipartite ranking problem with relative-level annotation and linear ranking function. (d) Nonlinear ranking function using learning to rank. (e) Nonlinear ranking function calibrated from small number of samples with absolute-level annotation. (f) Overview of the proposed method. (g) The multi-task learning in the proposed method.**

is "satisfied." More specifically, assume a pair of images x_i and x_j and their relative-level annotation stating that x_i has a higher level (i.e., rank) than x_j. The annotation is then satisfied when $f(x_i) > f(x_j)$. The trained function f is expected to be relative to the original (continuous) disease level.

Since there is a nonlinear relationship between the original image features and the disease levels, a function f should be highly nonlinear to satisfy the relative level annotations as many as possible. Therefore, we use representation learning by a convolutional neural network (CNN) to obtain a nonlinear f that satisfies the relative annotations as many as possible. Fig. 1 (d) shows a nonlinear f with representation learning by using a CNN. The thick green arrow curve shows the nonlinear f. The dotted curve shows the isoline where the samples on it have the same rank values.

It should be emphasized that the above ranking function is still not enough for practical diagnosis. This is because f satisfies only the order between the samples, and its value has no specific meaning for diagnosis. For example, if we realize a ranking function f for or ulcerative colitis (UC) diagnosis with colonoscopy images, the value of f does not have a clear relationship with a common severity level, such as the Mayo score [2]. Colony images x_i and x_j with Mayo levels 0 and 2 might have the "satisfactory" rank value -100 and 35, although it is impossible to guess the Mayo levels from the rank values.

In this paper, we propose a multi-task learning method for obtaining a *calibrated* ranking function $f(x)$ by using a large amount of (easy) relative annotations and a small amount of (costly) absolute-level annotations. Roughly speaking, we train a ranking function f to satisfy the relative-level annotations while satisfying $y \sim f(x)$ for the sample x with the absolute-level annotation y, as shown in Fig. 1 (e). By the calibration, the ranking function f can estimate the *real-valued* absolute levels (such as Mayo scores) of *all* samples.

Fig. 1 (f) shows the three steps of the proposed method. At the first step, an initial ranking function is obtained through a training process with only using relative-level annotations like Fig. 1 (d). At the second step, several samples are selected based on the estimated rank scores, and their absolute level annotation is attached by human (e.g., medical experts). At the final step, the ranking function is calibrated by multi-task learning of ranking and regression, as shown in Fig. 1 (g). The calibrated ranking function f will give a real-valued severity score. If a target severity score is expected as a discrete one, the score can be quantized into several levels, like Mayo 0, 1, 2. In other words, the calibrated ranking function can be seen as a severity classifier.

We applied the proposed method to a UC-level classification task. Specifically, we obtained a f that estimates the Mayo level of the given endoscopic image x. The Mayo level ranges from 0 (normal) and 3 (the most severe) with discrete values. We prove that f trained with the proposed method achieves high classification performance (accuracy and

F1-score) with far less annotation effort, aiding in supporting UC diagnosis.

The main contributions of this paper are summarized as follows:

1) To the best of our knowledge, this is the first trial of using the learning to rank framework for drastically reducing the annotation effort for a medical image dataset through relative-level annotation.
2) We developed a new multi-task learning method that calibrates the rank score to the absolute disease level.
3) Through an experiment to estimate the UC severity, the proposed method achieved even higher performance than the conventional classification methods trained with fully absolute-level annotations. This means that our method increases the estimation performance with much less annotation effort.

II. RELATED WORK

In gastrointestinal diseases, various lesions exist in different parts of the digestive organs, and endoscopy is used for lesion detection. Research on supporting endoscopic imaging diagnosis using machine learning is currently being conducted. There have been many investigations on automating classification tasks, such as classification of gastric cancer [3], [4], gastric precancerous disease [5], colorectal cancer using narrow-band imaging (NBI) images [6], and severity using endoscopic and biopsy histological images of UC [7]. An automatic abnormality detection task on capsule endoscope images has also been investigated [8]–[10]. These machine-learning applications aim to support diagnosis through classification, segmentation, and abnormality detection but do not focus on reducing the annotation cost of training data.

Learning to rank is widely used for recommendation systems and has been used for several image-analysis problems. For example, the ranking function has been applied to image-quality assessment and image attractiveness [11]–[15] because it is difficult to give an absolute quality evaluation for each image in these tasks.

Learning to rank is not common in medical image analysis, despite its usefulness in drastically reducing annotation effort. UC-level estimation is still often formulated as a classification task [16]–[18] and requires a dataset with absolute-level annotation. To the best of our knowledge, only a few studies [19]–[22] used the bipartite ranking problem for medical image analysis. However, none focused on the advantage of the ranking function for annotation-cost reduction. Moreover, some of these studies [19]–[21] just used simple or handcrafted features and thus did not use representation learning, although it drastically enhances the performance of the ranking function.

On the basis of a previous study [23], a ranking task is often converted into a multi-task learning problem (instead of the original bipartite ranking problem) then used in age estimation [24]–[26] and medical analysis [27]–[30]. Each multi-task learning is a binary classification to determine if the input sample is larger than a certain level. This approach requires absolute-level annotation, thus cannot use the benefit of relative-level annotation.

III. TWO ANNOTATION TYPES

In this study, there are the two types of ground-truth labels: absolute labels (ALs) and relative labels (RLs). In the proposed method, RLs are initially given, and then ALs are given to a small number of samples shown in Fig. 1 (f).

A. ABSOLUTE LABELS

AL is a disease severity level. In this study, it corresponds to one of the four-level Mayo scores. As noted in Section 1, giving accurate AL for a medical image is a difficult task even for experts. This is because of large image appearance variations within each level, and ambiguous samples that fall in the middle of two levels, say Mayo 1 and 2. These difficulties increase the annotation costs and thus prevent the realization of a large medical image dataset with ALs.

B. RELATIVE LABELS

RL is attached by comparing the severity of the disease between the two images as shown in Fig. 1 (b). A set of labeled paired images $\{(x_i, x_j, \bar{P}_{ij})\}$, $i, j \in [1, N]$ is defined, where $\bar{P}_{ij}$ is an RL of the image pair (x_i, x_j). The $\bar{P}_{ij}$ takes one of three values according to the following equation:

$$\bar{P}_{i,j} = \begin{cases} 1, & \text{if } x_i \text{ has a higher level than } x_j, \\ 0.5, & \text{else if } x_i \text{ and } x_j \text{ have the same level,} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The annotation for RLs is much easier than that for ALs because annotators do not need to identify the level of difficult samples that have a middle level of severity, such as level 1.5.

IV. LEARNING TO RANK WITH CALIBRATION

The proposed method consists of three steps. In step 1, we first train the initial ranking function by using learning to rank with RLs. In step 2, we then select a small number of samples from the training data to annotate them regarding ALs. The samples are selected using the ranking function trained using the RLs. In step 3, we finally carry out multi-task learning with RLs and additionally and adaptively prepared ALs, for calibrating the ranking function to more meaningful disease-severity levels. An overview of the proposed method is shown in Fig. 1 (f).

Before providing further details of the above steps, two important aspects should be clarified. First, the rank score by the ranking function in step 1 is not a disease-severity level, and thus the calibration of step 3 is necessary. Second, an AL is not given in advance but given after the ranking function is trained by RLs. This provides a more appropriate choice of samples where ALs should be attached, resulting in more accurate severity-level estimation with less AL annotation cost.

A. LEARNING TO RANK

In step 1, we train the initial ranking function using learning to rank. The ranking function $f(x)$ is trained with a CNN for the representation (i.e., feature extraction) that is suitable for the ranking. A CNN is composed of multiple convolutional layers, a single fully connected layer, and a single output node to give a single scalar value $f(x)$. This can be considered a powerful extension of the classical RankNet [31] where a linear ranking function is trained using a very shallow neural network.

The CNN is trained using sample pairs with RLs. For training, we input two images to two CNNs with shared weights and then minimize the loss for the pair. Specifically, the CNN is trained with the loss function $L_{\text{rank}} = \sum_{(i,j)\in\mathcal{P}} L_{\text{rank}}^{i,j}$, where $\mathcal{P}$ is the set of sample pairs. The function $L_{\text{rank}}^{i,j}$ is defined as a cross-entropy,

$$L_{\text{rank}}^{i,j} = -\bar{P}_{i,j} \log P_{i,j} - (1 - \bar{P}_{i,j}) \log(1 - P_{i,j}), \tag{2}$$

where $P_{ij} = \text{sigmoid}(f(x_i) - f(x_j))$ and $\bar{P}_{ij}$ is an RL of the image pair (x_i, x_j).

B. SAMPLING FOR ABSOLUTE-LEVEL ANNOTATION

In step 2, a small number of samples are selected from the training data and attached ALs. As noted above, the proposed method assumes that the samples to which ALs are attached are selected *after* $f(x)$ is estimated. This is more reasonable than, for example, a random selection because we can select samples that are expected to be more necessary for the calibration step by using the clues from $f(x)$.

To select a small number of samples, we first obtain the rank score of the training samples using $f(x)$ and represent the rank score of the training data as a point on a number line. Next, we select M ($\ll N$) samples at equal intervals on the number line within the maximum and minimum rank scores. Finally, ALs are attached to the selected M samples by absolute-level annotations.

C. MULTI-TASK LEARNING

In the final step 3, the proposed method calibrates the ranking function to give the absolute severity score. This calibration process can be seen as a fine-tuning process of $f(x)$ so that the output of $f(x)$ becomes closer to the AL of x. At the same time, we need to be careful that the fine-tuning process does not destroy the sample ranks learned in $f(x)$. These two requirements result in a multi-task learning to fine-tune $f(x)$.

As shown in Fig. 1 (g), the multi-task learning combines regression to make $y \sim f(x)$ for the sample x with AL and learning to rank for the pairs (x_i, x_j) with RL. The loss function of learning to rank is cross entropy of Eq.(2). The loss function for regression, $L_{\text{reg}} = \sum_{(i,j)\in\mathcal{P}} L_{\text{reg}}^{i,j}$ is defined by adding the mean squared error (MSE) loss function for each sample pair,

$$L_{\text{reg}}^{i,j} = (f(x_i) - y_i)^2 + (f(x_j) - y_j)^2, \tag{3}$$

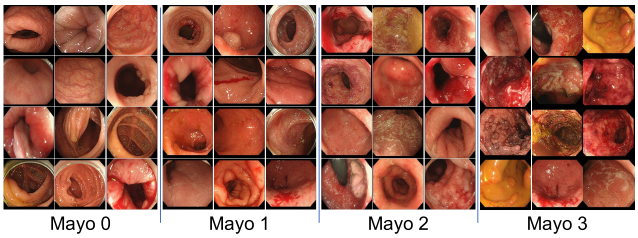


FIGURE 2. Colonoscopy image examples with different UC levels.

where y_i and y_j are the ALs attached by the absolute-level annotation of x_i and x_j, respectively. Furthermore, the multi-task loss function L_{multi} is defined as the sum of the loss functions of learning to rank and regression,

$$L_{\text{multi}} = L_{\text{rank}} + \lambda L_{\text{reg}}, \tag{4}$$

where λ is a hyper-parameter to balance the losses. The trained multi-task $f(x)$ is expected to be a ranking function corrected for the region of severity levels in the feature space optimized by representation learning, as shown in Fig. 1 (e). Note that, in Step 3, we only use M samples with ALs. Therefore, L_{rank} in (4) is minimized with the RLs of $M(M-1)$ pairs.

Finally, the calibrated rank score $f(x)$ is quantized into the nearest discrete disease-severity level as the classification result of x. For example, with the severity levels $\in \{0, 1, 2, 3\}$, the level becomes 3 for x whose $f(x) = 2.7$.

V. EXPERIMENTS AND RESULTS

A. DATASET

We used 10,265 colonoscopy images of UC from 388 patients at Kyoto Second Red Cross Hospital as the dataset. These images were taken from multiple patients (including healthy participants). The images have different sizes and therefore were resized to 224×224 pixels.

Fig. 2 shows several examples of each of four levels of Mayo, which is the standard disease severity score for UC. According to Schroeder *et al.* [2], Mayo 0 is normal or endoscopic remission. Mayo 1 is a mild level showing erythema (i.e., abnormal redness), a decreased vascular pattern, and mild friability. Mayo 2 is a moderate level showing marked erythema, an absent vascular pattern, friability, and erosions. Mayo 3 is a severe level with spontaneous bleeding and ulceration.

Although our method does not require a dataset with full ALs, we attached ALs to all samples for a quantitative performance evaluation. Specifically, a four-level Mayo score is carefully attached to each colonoscopy image by multiple medical experts. The dataset contains 6,678, 1,995, 1,395, and 197 samples for Mayo 0, 1, 2, and 3, respectively. Note that it is common to have such a heavily imbalanced dataset for colonoscopy, as well as other medical image diagnosis tasks.

In the following experiments, five-fold cross-validation was performed. The colonoscopy images were divided into

TABLE 1. Comparison of ground-truth labels and annotation time for each method.

Method	RL	AL	Time (sec)*
Conventional	-	8,212	164,240
Proposed	8,212$-M^{**}$	M^{**}	8,212$+19M^{**}$

* RL: 1 (sec/pair), AL: 20 (sec/image)
** $M = 50, 100, 200, 300, 400$

60%, 20%, and 20% for patient-disjoint training, validation, and test sets, respectively. Note that we divide all images into training, validation, and test sets to have the same severity proportion for keeping a fair and practical evaluation scenario.

For a fair comparison with a conventional method (detailed later), we carefully control several conditions. First, we used the same number of annotations for the conventional and proposed methods. More precisely, the conventional method uses 8,212 images-(80% of the entire data) with AL for training, and the proposed method uses 8,212$-M$ pairs with RL at step 1, and M images with AL at step 3. Since AL has more information than RL, this condition is a handicap for the proposed method. Nevertheless, we adopted this condition so that the conventional method would not be disadvantaged.

Second, we allow over/under sampling for class imbalance removal to the conventional method but not to the proposed method. This is because the conventional method has ALs for all samples and thus, such sampling is possible, whereas the proposed method does not. This condition will be another and large handicap for the proposed method.

B. TIME EFFICIENCY IN ANNOTATION PROCESS

Table 1 shows the number of ground-truth labels (RL and AL) and the annotation time for each method. In our interview with endoscopists, AL labeling takes 20 seconds per image, and RL labeling only takes (less than but roughly) one second. For the case $M = 400$, this indicates that the proposed method requires just 10% of the annotation time of the conventional method.

C. IMPLEMENTATION

The implementation environment is shown as follows. We used an Intel(R) Core(TM) i9-10980XE 3.00 GHz as the CPU and two NVIDIA TITAN RTX 24 GB as GPUs for training. We wrote the code in Python 3.6 and used Tensorflow 1.13.1 and Keras 2.2.4 as the deep learning library. The CUDA version was 10.0. We used Adam as the optimizer to train the weight parameters. The learning rate was set to 5×10^{-6}. The learning was terminated by the early stopping rule (no decrease in validation loss for 20 epochs). For λ in Eq.(4), we examined the range of 0.001 to 1 and had the highest F1-score at $\lambda = 0.01$ for the validation set.

We used DenseNet [32] as the CNN. DenseNet has been widely used in various medical-image classification and analysis tasks due to its state-of-the-art performance (e.g., [33], [34]).

TABLE 2. Classification performance evaluation of the conventional and proposed methods.

Method	Class	Precision	Recall	F1-score
Conventional	Mayo 0	0.901	0.747	0.817
	Mayo 1	0.394	0.675	0.498
	Mayo 2	0.764	0.443	0.561
	Mayo 3	0.252	0.635	0.360
	Overall	0.578	0.625	0.559
Proposed	Mayo 0	0.864	0.840	0.852
	Mayo 1	0.416	0.463	0.438
	Mayo 2	0.643	0.606	0.624
	Mayo 3	0.368	0.432	0.397
	Overall	0.572	0.585	**0.578**

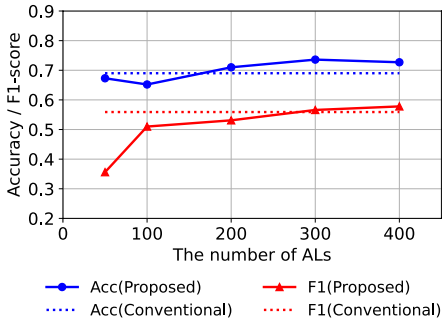


FIGURE 3. Classification performance evaluation of the proposed method using different numbers of ALs. In the conventional method, 8,212 ALs are used.

D. EVALUATION METRICS

The proposed method is evaluated in four-Mayo class classification performance by accuracy, recall, precision, and F1-score. Recall that the class is determined by quantizing the rank score into its neighboring level, e.g., $2.7 \rightarrow 3$. We leave the test samples imbalanced to mimic realistic medical situations. To avoid the under/over-estimation risk of the accuracy values in the imbalanced situation, F1-score is also employed.

E. COMPARISON METHOD

The performance of the proposed method was compared with the conventional CNN-based multi-class classification method. DenseNet-169 trained by the standard categorical cross entropy is used for this comparative method. As the training data, all 8,212 training samples are used with their absolute-level annotations. This means that it uses all of the absolute-level annotations.

F. CLASSIFICATION PERFORMANCE

Table 2 shows the classification performance of the proposed method at $M = 400$. The proposed method achieves higher F1-score than the conventional method. This result shows that the proposed method achieves even higher classification performance than the conventional method, although the proposed method only needs 1/10 annotation cost of the conventional method.

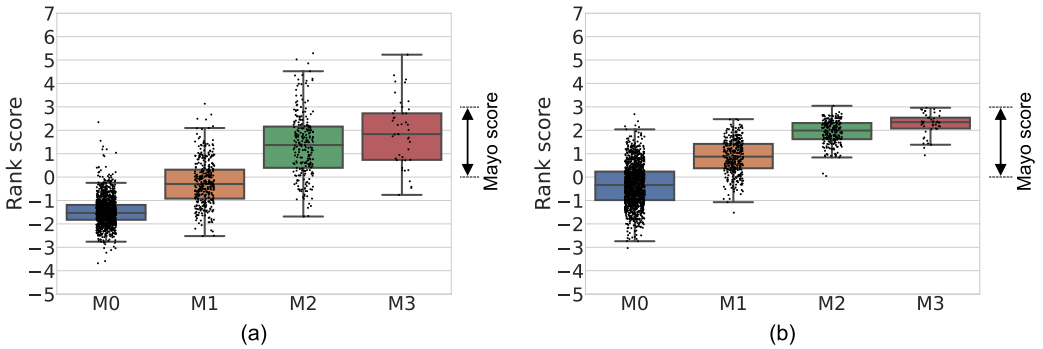


FIGURE 4. Box-plots of test set's rank score obtained with (a) the uncalibrated case and (b) the proposed method. 'M0' denotes for Mayo 0.

We evaluated the performance of the proposed method for severity classification with various numbers of ALs ($M = 50$, 100, 200, 300, and 400). Fig. 3 shows the results of the performance evaluation with the proposed method using different numbers of ALs. *Acc* and *F1* represent accuracy and F1-score, respectively. The F1-score increases as the number of ALs increases. From $M = 300$, the F1-score of the proposed method is higher than that of the conventional method. The accuracy of the proposed method is higher than that of the conventional method for $M = 200$ and over. Therefore, the proposed method achieved higher performance than the conventional method when the number of ALs is more than $M = 300$.

G. ABLATION STUDY

We examined the effect of calibration on rank scores by multi-task learning with the proposed method. Specifically, we verified the effect by comparing the classification performance between calibrated and uncalibrated cases. To determine the classification result for the uncalibrated case, we defined the range of the rank score for each Mayo score by logistic regression using $M = 400$ ALs, which were used for multi-task learning, for a fair comparison.

Fig. 4 shows box-plots for each correct Mayo score of test samples by (a) the uncalibrated case and (b) the proposed method. The horizontal and vertical axes correspond to the correct Mayo score attached by the annotators and the rank scores, respectively. The rank scores obtained with the proposed method are located nearer to the Mayo score range of 0 to 3 than those with the uncalibrated case. Therefore, these results indicate that the rank scores are calibrated with the ALs as the anchor by using regression as the anchor task in multi-task learning.

Table 3 shows the results of the performance evaluation for the uncalibrated case. The overall precision, recall, and F1-score of the uncalibrated case were lower than those of the proposed method. Comparing the F1-score for each class, Mayo 3 was particularly low with the uncalibrated case, indicating an imbalance in the classification performance for each class.

TABLE 3. Classification performance evaluation of the uncalibrated case.

Method	Class	Precision	Recall	F1-score
Uncalibrated	Mayo 0	0.892	0.873	0.882
	Mayo 1	0.439	0.625	0.516
	Mayo 2	0.728	0.422	0.535
	Mayo 3	0.229	0.086	0.124
	Overall	0.570	0.502	0.514

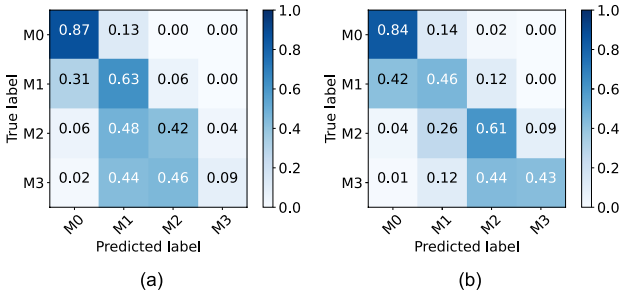


FIGURE 5. Confusion matrices of (a) the uncalibrated case and (b) the proposed method.

Fig. 5 shows the confusion matrices of the classification results using the uncalibrated case and the proposed method. Compared with the proposed method, the uncalibrated case had a higher rate of incorrectly predicting Mayo 2 and Mayo 3 as Mayo 1 and could not accurately classify images with high severity. These results indicate that the calibration effect improves the performance of classifying images with high severity and that the proposed method has higher performance than the uncalibrated case.

VI. CONCLUSION

We proposed a multi-task learning method that combines learning to rank with regression for automatically estimating UC severity levels (Mayo scores). The proposed method has a strong advantage in that it can substantially reduce annotation costs by using relative-level annotation instead of costly absolute-level annotations. Our experimental result shows that the proposed method achieved even higher classification

performance (accuracy and F1-score) than the conventional classification method while requiring just 1/10 annotation cost.

The limitation of the proposed method is that it requires more training time than the conventional method because the number of pairs increases with the number of images to which AL is attached. We will investigate ways to make effective pairs for learning with as few AL combinations as possible.

Future work will involve the proposal of a new continuous-valued UC severity level. Currently, we discretize the regression result into four levels to follow the traditional Mayo-based evaluation. However, the regression result can show an intermediate score, such as Mayo 1.75 by itself. Our continuous severity score can be an accurate and precise alternative to Mayo through discussion with the medical expert committee.

ACKNOWLEDGMENT

All of the endoscopic images used in this article are approved by the ethical review committee at the Kyoto Second Red Cross Hospital.

REFERENCES

[1] F. Hirai and T. Matsui, "A critical review of endoscopic indices in ulcerative colitis: Inter-observer variation of the endoscopic index," *Clin. J. Gastroenterol.*, vol. 1, no. 2, pp. 40–45, Jun. 2008, doi: 10.1007/s12328-008-0018-z.

[2] K. W. Schroeder, W. J. Tremaine, and D. M. Ilstrup, "Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis," *New England J. Med.*, vol. 317, no. 26, pp. 1625–1629, Dec. 1987, doi: 10.1056/nejm198712243172603.

[3] J. H. Lee, Y. J. Kim, Y. W. Kim, S. Park, Y.-I. Choi, Y. J. Kim, D. K. Park, K. G. Kim, and J.-W. Chung, "Spotting malignancies from gastric endoscopic images using deep learning," *Surgical Endoscopy*, vol. 33, no. 11, pp. 3790–3797, Nov. 2019, doi: 10.1007/s00464-019-06677-2.

[4] Y. Zhu, Q.-C. Wang, M.-D. Xu, Z. Zhang, J. Cheng, Y.-S. Zhong, Y.-Q. Zhang, W.-F. Chen, L.-Q. Yao, P.-H. Zhou, and Q.-L. Li, "Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy," *Gastrointestinal Endoscopy*, vol. 89, no. 4, pp. 806–815, 2019, doi: 10.1016/j.gie.2018.11.011.

[5] X. Zhang, W. Hu, F. Chen, J. Liu, Y. Yang, L. Wang, H. Duan, and J. Si, "Gastric precancerous diseases classification using CNN with a concise model," *PLoS ONE*, vol. 12, no. 9, pp. 1–10, 2017, doi: 10.1371/journal.pone.0185508.

[6] T. Tamaki, J. Yoshimuta, M. Kawakami, B. Raytchev, K. Kaneda, S. Yoshida, Y. Takemura, K. Onji, R. Miyaki, and S. Tanaka, "Computer-aided colorectal tumor classification in NBI endoscopy using local features," *Med. Image Anal.*, vol. 17, no. 1, pp. 78–100, Jan. 2013, doi: 10.1016/j.media.2012.08.003.

[7] K. Takenaka, K. Ohtsuka, T. Fujii, M. Negi, K. Suzuki, H. Shimizu, S. Oshima, S. Akiyama, M. Motobayashi, M. Nagahori, E. Saito, K. Matsuoka, and M. Watanabe, "Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis," *Gastroenterology*, vol. 158, no. 8, pp. 2150–2157, Jun. 2020, doi: 10.1053/j.gastro.2020.02.012.

[8] H. Alaskar, A. Hussain, N. Al-Aseem, P. Liatsis, and D. Al-Jumeily, "Application of convolutional neural networks for automated ulcer detection in wireless capsule endoscopy images," *Sensors*, vol. 19, no. 6, p. 1265, Mar. 2019, doi: 10.3390/s19061265.

[9] T. Aoki, A. Yamada, K. A. M. Math, H. Saito, A. Tsuboi, A. Nakada, R. Niikura, M. Fujishiro, S. Oka, S. Ishihara, T. Matsuda, S. Tanaka, K. Koike, and T. Tada, "Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network," *Gastrointestinal Endoscopy*, vol. 89, no. 2, pp. 357–363, 2019, doi: 10.1016/j.gie.2018.10.027.

[10] M. K. Bashar, T. Kitasaka, Y. Suenaga, Y. Mekada, and K. Mori, "Automatic detection of informative frames from wireless capsule endoscopy images," *Med. Image Anal.*, vol. 14, no. 3, pp. 449–470, Jun. 2010, doi: 10.1016/j.media.2009.12.001.

[11] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017, doi: 10.1109/TIP.2017.2708503.

[12] X. Jiang, L. Shen, L. Yu, M. Jiang, and G. Feng, "No-reference screen content image quality assessment based on multi-region features," *Neurocomputing*, vol. 386, pp. 30–41, Apr. 2020, doi: 10.1016/j.neucom.2019.12.027.

[13] J. Yan, S. Lin, S. B. Kang, and X. Tang, "A learning-to-rank approach for image color enhancement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2987–2994, doi: 10.1109/CVPR.2014.382.

[14] N. Ma, A. Volkov, A. Livshits, P. Pietrusinski, H. Hu, and M. Bolin, "An universal image attractiveness ranking framework," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 657–665, doi: 10.1109/WACV.2019.00075.

[15] Y. Murata and Y. Dobashi, "Automatic image enhancement taking into account user preference," in *Proc. Int. Conf. Cyberworlds (CW)*, 2019, pp. 374–377, doi: 10.1109/CW.2019.00070.

[16] A. Dahal, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen, "Detection of ulcerative colitis severity in colonoscopy video frames," in *Proc. 13th Int. Workshop Content-Based Multimedia Indexing (CBMI)*, Jun. 2015, pp. 1–6, doi: 10.1109/CBMI.2015.7153617.

[17] A. Alammari, A. R. Islam, J. Oh, W. Tavanapong, and J. Wong, "Classification of ulcerative colitis severity in colonoscopy videos using vascular pattern detection," in *Proc. Int. Conf. Inf. Manage. Eng. (ICIME)*, 2017, pp. 139–144.

[18] R. W. Stidham, W. Liu, S. Bishu, M. D. Rice, P. D. R. Higgins, J. Zhu, B. K. Nallamothu, and A. K. Waljee, "Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis," *JAMA Netw. Open*, vol. 2, no. 5, May 2019, Art. no. e193963, doi: 10.1001/jamanetworkopen.2019.3963.

[19] W. Huang, K. L. Chan, H. Li, J. H. Lim, J. Liu, and T. Y. Wong, "A computer assisted method for nuclear cataract grading from slit-lamp images using ranking," *IEEE Trans. Med. Imag.*, vol. 30, no. 1, pp. 94–107, Jan. 2011, doi: 10.1109/TMI.2010.2062197.

[20] F. Pedregosa, E. Cauvet, G. Varoquaux, C. Pallier, B. Thirion, and A. Gramfort, "Learning to rank from medical imaging data," in *Proc. Int. Workshop Mach. Learn. Med. Imag. (MLMI)*, 2012, pp. 234–241, doi: 10.1007/978-3-642-35428-1_29.

[21] B. Peng, X. Yao, S. L. Risacher, A. J. Saykin, L. Shen, and X. Ning, "Prioritization of cognitive assessments in Alzheimer's disease via learning to rank using brain morphometric data," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, May 2019, pp. 1–4, doi: 10.1109/BHI.2019.8834618.

[22] J. Lyu, S. H. Ling, S. Banerjee, J. J. Y. Zheng, K.-L. Lai, D. Yang, Y.-P. Zheng, and S. Su, "3D ultrasound spine image selection using convolution learning-to-rank algorithm," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 4799–4802, doi: 10.1109/EMBC.2019.8857182.

[23] L. Li and H. T. Lin, "Ordinal regression by extended binary classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 865–872, doi: 10.5555/2976456.2976565.

[24] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5183–5192, doi: 10.1109/cvpr.2017.86.

[25] K. Y. Chang, C. S. Chen, and Y. P. Hung, "A ranking approach for human age estimation based on face images," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2010, pp. 3396–3399, doi: 10.1109/ICPR.2010.829.

[26] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4920–4928, doi: 10.1109/CVPR.2016.532.

[27] B. Liu, Y. Zhang, M. Chu, X. Bai, and F. Zhou, "Bone age assessment based on rank-monotonicity enhanced ranking CNN," *IEEE Access*, vol. 7, pp. 120976–120983, 2019, doi: 10.1109/ACCESS.2019.2937341.

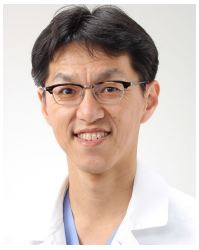
[28] X. Liu, Y. Zou, Y. Song, C. Yang, J. You, and B. V. Kumar, "Ordinal regression with neuron stick-breaking for medical diagnosis," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 335–344, doi: 10.1007/978-3-030-11024-6_23.

[29] T. J. Jun, Y. Eom, D. Kim, C. Kim, J.-H. Park, H. M. Nguyen, and D. Kim, "TRk-CNN: Transferable ranking-CNN for image classification of glaucoma, glaucoma suspect, and normal eyes," 2019, *arXiv:1905.06509*.
[30] B. Wu, X. Sun, L. Hu, and Y. Wang, "Learning with unsure data for medical image diagnosis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10589–10598, doi: 10.1109/ICCV.2019.01069.
[31] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 89–96, doi: 10.1145/1102351.1102363.
[32] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.
[33] Y. Yuan, W. Qin, B. Ibragimov, B. Han, and L. Xing, "RIIS-DenseNet: Rotation-invariant and image similarity constrained densely connected convolutional network for polyp detection," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, 2018, pp. 620–628, doi: 10.1007/978-3-030-00934-2_69.
[34] R. Bise, K. Abe, H. Hayashi, K. Tanaka, and S. Uchida, "Efficient soft-constrained clustering for group-based labeling," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, vol. 5, 2019, pp. 421–430, doi: 10.1007/978-3-030-32254-0_47.

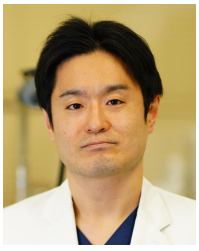
TAKEAKI KADOTA received the B.E. and M.Eng. degrees from Kyoto University, Kyoto, Japan, in 2007 and 2009, respectively, and the M.D. degree from the Shiga University of Medical Science, Shiga, Japan, in 2020. He is currently pursuing the Ph.D. degree with the Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan. From 2009 to 2012, he joined Kansai Electric Power Company, Inc., Japan. His current research interests include medical image analysis and machine learning.

KENTARO ABE received the B.E. and M.E. degrees from Kyushu University, in 2018 and 2020, respectively. His research theme in the master's degree was medical image analysis.

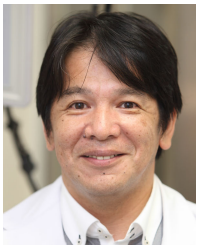
RYOMA BISE (Member, IEEE) received the M.S. degree from the Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan, in 2002, and the Ph.D. degree in interdisciplinary information studies from The University of Tokyo, in 2015. He was engaged in the research and development on informatics at Dai Nippon Printing Company, Ltd., Japan, from 2002 to 2015. He was at the National Institute of Informatics, from2015 to 2017. He joined the Faculty of Information Science and Electrical Engineering, Kyushu University as an Associate Professor, in 2017. His research interest includes computer vision, particularly biomedical image analysis.



TAKUJI KAWAMURA received the M.D. degree. He is an Endoscopist working as the Vice Director of the Gastroenterological Department, Kyoto Second Red Cross Hospital. His specialty is colonoscopy. He received the Distinguished Paper Award of Japan Gastroenterological Endoscopy Society (JGES), in 2018. Currently, he is an Associate Editor of *Digestive Endoscopy*—official journal of JGES.



NAOKUNI SAKIYAMA received the M.D. degree. He is an Endoscopist working with the Kyoto Second Red Cross Hospital. He treats a wide range of diseases from biliopancreatic to gastrointestinal diseases, especially specializes in inflammatory bowel disease, such as ulcerative colitis and Crohn's disease. He is currently engaged in clinical practice and clinical research in this area.



KIYOHITO TANAKA received the M.D. degree. He is an Endoscopist for gastorointestinal endoscopy and pancreatobiliary endoscope. He is the Chief Information Officer with the Kyoto Second Red Cross Hospital (K2RCH). In K2RCH, international tele-conference and live demonstration were performed over ten times by year.

SEIICHI UCHIDA (Member, IEEE) received the B.E., M.E., and Dr.Eng., degrees from Kyushu University, in 1990, 1992, and 1999, respectively. He is currently a Distinguished Professor with Kyushu University. His research interests include pattern recognition and image processing. He received the 2007 IAPR/ICDAR Best Paper Award, the 2010 ICFHR Best Paper Award, and many domestic awards. Currently, he is an Associate Editor of *Pattern Recognition* (Elsevier).