Improving the Accuracy of Intrahepatic Cholangiocarcinoma Subtype Classification by Hidden Class Detection via Label Smoothing

Jing Wei Tan [1], *Kanggeun Lee* [2], *Kyoungbun Lee* [3], *Won-Ki Jeong* [1]

[1] Department of Computer Science and Engineering, Korea University, Seoul, Korea
[2] School of Electrical and Computer Engineering, UNIST, Ulsan, Korea
[3] Department of Pathology, Seoul National University Hospital, Seoul, Korea

ABSTRACT

Obtaining ground-truth labels for supervised training is a labor-intensive and time-consuming task. Owning to their large size, only slide-level labels or a handful of coarse annotations are usually provided for pathology images, which makes the training of the classifier challenging. In this study, we propose a conceptually simple, two-stage approach to classify small and large duct types in intrahepatic chalongiocarcinoma using only slide-level labels. Unlike conventional pathology image analysis methods employ multiple instance learning (MIL) applied to overcome the problem of the slide-level label, we introduce a novel label smoothing method to progressively refine the training labels to improve the classification accuracy. The main idea is that we introduce the hidden class, which is assumed to be mutually inclusive of all ground-truth classes and less confident for classification. By iteratively refining (i.e., smoothing) per-patch labels, we can extract and discard the hidden class from the training data. We demonstrate that the proposed label filtering scheme improves the classification accuracy by up to 9% compared to the baseline MIL method and 10% compared to the state-of-the-art noisy label cleaning method. In addition, we demonstrate the effectiveness of gene mutation prior information in the classification of two different duct types. The experimental results suggest that the proposed method may provide pathologists insight into the study of correlations between genetic and histologic subtypes.

Index Terms— intrahepatic cholangiocarcinoma, duct type, hidden class detection, multiple instance learning

1. INTRODUCTION

Biliary tract tumors known as intrahepatic cholangiocarcinomas (IHCCs) can be further divided into peripheral small duct type and perihilar large duct type. The morphological classification between the small duct type and large duct remains unclear and has low reproducibility. In terms of genetic, IHCCs of the small duct type were found to have higher mutation rate in the IDH gene than large duct type IHCC, whereas IHCCs of the large duct type have higher mutation rates in the IDH gene. IDH and KRAS are the two major mutated genes in IHCCs, but the overall mutation rates of KRAS and IDH are not as high as adenocarcinoma of colon or pancreatic cancer [1, 2, 3, 4]. The high cost of annotating a whole slide image (WSI) makes the annotation of patches even more difficult. Patch-wise labeling is even more difficult to obtain because of the lack of clear morphologic criteria between the small and large duct types and the overlapping of features. Low tumor cell density and plenty of stroma of IHCC also make annotation difficult.

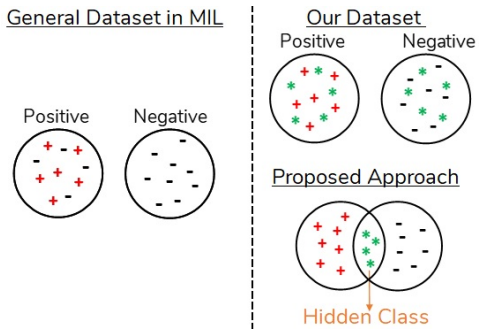

Fig. 1. Differences between general MIL and the proposed method.

Multiple instance learning (MIL) is the most common approach in addressing weak labeling instances where only bag-level (image-wise) and no instance-level (patch-wise) labels are provided [5, 6, 7, 8]. Other approaches have been proposed in previous studies to solve similar weak-label issues such as treating the patch label as a noisy label and hallucinating clean representation [9] and, iterative patch labeling [10].

Generally, MIL assumes at least one positive instance in the positive bag, while all instances in the negative bag are negative. However, our dataset is slightly different from the general dataset in most MIL studies (Fig. 1). Our dataset is composed of two different duct types in IHCCs, in which some of the regions in the WSIs from these two duct types may share similar features. Therefore, we aim to search for those patches that share similar features in order to reduce the false positive and false negative patches in each bag for image-level classification. In this work, we propose a conceptually simple approach to detect the samples from the hid-

den class, with the aim of solving the problem of image-level labels. First, we assume there exists an additional hidden class in which the patches share similar features among the cancer subtypes. Hence, we assign an additional class label (hidden class) to the ground-truth labels and update the label iteratively to reduce the ambiguity in the confidence score, and seek the patches that potentially belong to the hidden class, which will be discarded to improve image-level classification accuracy. We further evaluate the efficiency and effectiveness of hidden class detection by using a simple MIL model because only image-level labels are provided. Our proposed method is applied to detect and discard the hidden class within two subtypes of intrahepatic cholangiocarcinoma, and to identify discriminative patches for small and large duct type classification. To the best of our knowledge, this is the first study to detect the hidden class among the subclasses by assigning an additional class label to classify intrahepatic cholangiocarcinoma into small and large duct types.

2. METHODOLOGY

An overview of the proposed framework is shown in Fig. 2. In this study, our proposed framework consists of two stages: stage 1 is hidden class detection, and stage 2 is image-level classification.

2.1. WSI Patching

First, all of the whole slide images from level 0 are downsampled at the scale of 0.5. The patches are then extracted in a dimension of $256 \times 256 \times 3$ and patches with less than 75% tissue coverage are discarded.

2.2. Hidden Class Detection

It is certain that some patches share the same features among all patches from any two or more subtypes of cancer. For example, patches from the normal tissue area in the WSIs of two different cancer subtypes should share similar information and features, even though their WSIs are labeled differently. Hence, we deduce that these patches can be further classified into an another additional subclass, which we named as *hidden class* in this study. This is as illustrated in Fig. 3 as Class 3 (overlapping region).

The initial ground-truth binary label vectors for the patches from Class 1 and Class 2 are [1,0] and [0,1], respectively. At the initial stage, because no patch-level labels are provided, we assume that all patches x have the potential to be classified as the hidden class. Therefore, we add an additional label, Class 3 (which indicates the hidden class) to the middle of their initial label vectors, and the new label vectors of Class 1 and Class 2 are changed to [1,1,0] and [0,1,1], respectively. For example, a patch with a label [1,1,0] means that this patch belongs to Class 1 and 3. The labels from each class are further expanded into binary labels using two digits per class (first and second digits indicate negative and positive class labels, respectively) as shown in Fig. 3b. This means that the label vectors are expanded from [1,1,0] to [0,1,0,1,1,0] and from [0,1,1] to [1,0,0,1,0,1], respectively. The motivation for using six labels is to train the data under a multitask learning setup that promotes better feature sharing among the classes. These labels are then fed into a CNN model to train for the first iteration.

The predicted output (i.e., probability) of each positive label from each class is then treated as the confidence score for each class. After iteration t, the top $P\%$ patches x from each WSI X_j with the highest confidence score among three classes are relabeled with either [1,0,0] or [0,0,1] (meaning that these patches are highly likely Class 1 or 2). Meanwhile, the other patches carry the same label vector, which is assigned before the first iteration. After relabeling, the model continues to be trained for the next iteration $t+1$ with the new labels. The patches are considered to belong to the hidden class if they have the highest confidence score with respect to the hidden class.

In this stage, we adopted efficientNet [11] with pretrained weight from ImageNet [12], followed by a dropout layer, dense layer, dropout layer, dense layer, dropout layer, and sigmoid layer. The dropout rate was set to 0.5, and each dense layer was set to 256 neurons with Relu activation. The model was trained along with the rectified Adam optimizer [13] with total steps of 10,000, warm-up proportion of 0.2, and minimum learning rate in 10^{-5}. The loss function is formulated as follows:

$$loss = \sum_{i=1}^{N} \parallel y^{(i)} - \hat{y}^{(i)} \parallel_2^2 + \sum_{i=1}^{N} \parallel 1 - \sum_{j \epsilon J}^{M} \hat{y}_j^{(i)} \parallel_2^2 \quad (1)$$

where i indicates the patch number, j represents the label index among the six labels and $J = \{2, 4, 6\}$ which is the index of the positive class label of each class. In Eq. 1, the first term is the mean square error between the ground-truth label $y^{(i)}$ and the predicted output $\hat{y}^{(i)}$. The second term is the sum of one constraint that applies to the probability of the positive class from each class.

2.3. Image-Level Classification

In the second stage, we adopted a conventional multiple instance learning model to evaluate the performance of the first stage (hidden class detection) with the absence of patch-level labels.

As the first step, we perform *patch filtering* using the result from the first stage. The patches h, which were defined as the hidden class, were discarded, and the remaining patches r were assumed to carry well-representative features for Class 1 and Class 2, respectively. These representative patches from each WSI were then used to construct 50 bags, each of which

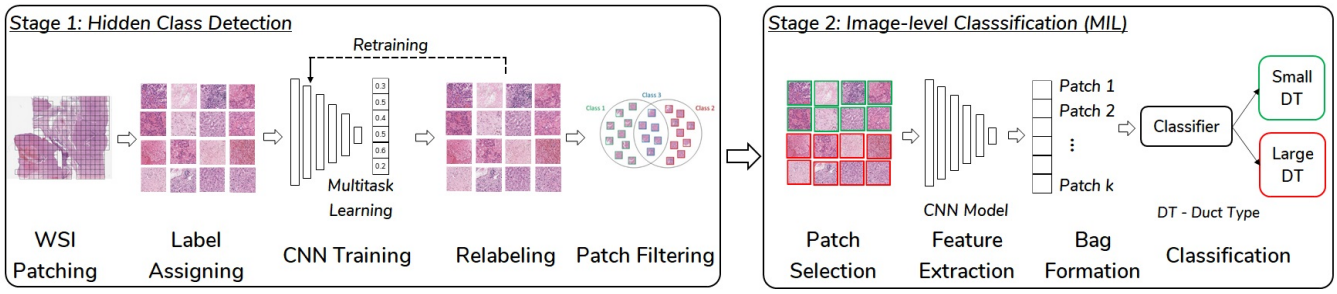

Fig. 2. Overview of the proposed framework.

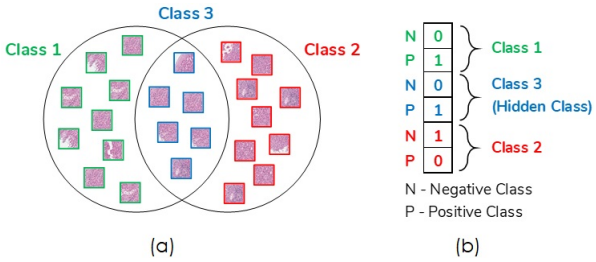

Fig. 3. (a) Hidden class which carries shared or similar features between any two subclasses, shown in overlapping region as Class 3. (b) Label description.

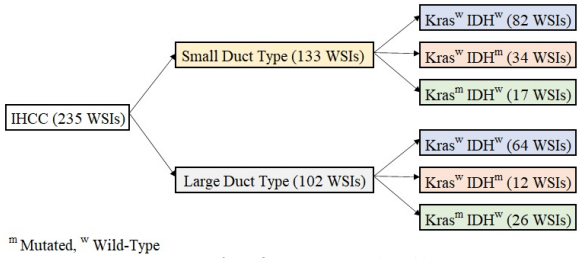

Fig. 4. Dataset details.

consisted of 100 patches. Each bag was generated by concatenating 512-dimensional feature vectors extracted from the pre-trained VGG16 model [12, 14] of 100 patches.

On the other hand, several simple filtering schemes were applied to the patch r, which included the threshold scheme and top $N\%$ patches from each WSI based on the confidence score from stage 1 in order to improve the bag classification performance. Subsequently, these bags were used as the input for training a multi-layer perceptron (MLP) model. The MLP classifier had three fully connected layers with 2048 neurons and a softmax layer. This classifier was trained along with stochastic gradient descent, learning rate of 10^{-5}, and weight decay in 10^{-7}.

3. EXPERIMENTS

3.1. Dataset

A total of 119 patients were enrolled in this dataset with a total of 235 WSIs. The intrahepatic cholangiocarcinoma can further be classified into small and large duct types. Patients with intrahepatic cholangiocarcinoma were found to have mutations in two common genes: IDH and Kras [15]. The Kras mutation study was conducted by a hemi-tested polymerase chain reaction for exon2 and IDH mutation was analyzed by pyrosequencing assay (Qiagen, Hilden, Germany) for IDH1 codon 123 and IDH2 codon 172.

3.2. Results and Discussion

Accuracy was used as the main evaluation metric for this study and the reported results included the average accuracy from 10 iterations.

Because gene mutation information is provided in this dataset, we conducted an experiment with a simple MIL model to determine the effectiveness of gene mutation information on the classification performance by using samples with Kras^{w} IDH^{m} in the small duct type and Kras^{m} IDH^{w} in the large duct type as the training samples. Table 1 shows how the gene mutation prior information affects the classification performance. It can be seen that the model trained with the selected samples based on prior gene mutation information (i.e., selecting only Kras^{w} IDH^{m} from the small duct type and Kras^{m} IDH^{w} from the large duct type) achieved 0.6343 in accuracy, which outperformed the model trained without any prior information (0.6171 in accuracy). The experimental results show that the gene mutation prior information plays a crucial role in classifying the small and large duct types even though the correlation information between the histologic subtype and genetic subtype still remains ambiguous. This data selection scheme was then employed for the rest of the study.

In this study, the model was trained by updating the top 30% patches, top 50%, and combination of top 40% and bottom 10% (the bottom 10% patches were updated as [0,1,0]) for each WSI. Among these updating schemes, the model with the top 30% of label updating showed the best performance. Therefore, the top 30% was employed throughout the whole study. During the first iteration, most of the patches tended to have similar confidence scores in two or more classes. However, as the iteration continued, the ambiguity of the confidence score of a patch was reduced. In other words, confidence score was much higher in a particular class instead of having a similar confidence score in two or three classes. Empirically, the model obtained the best results after

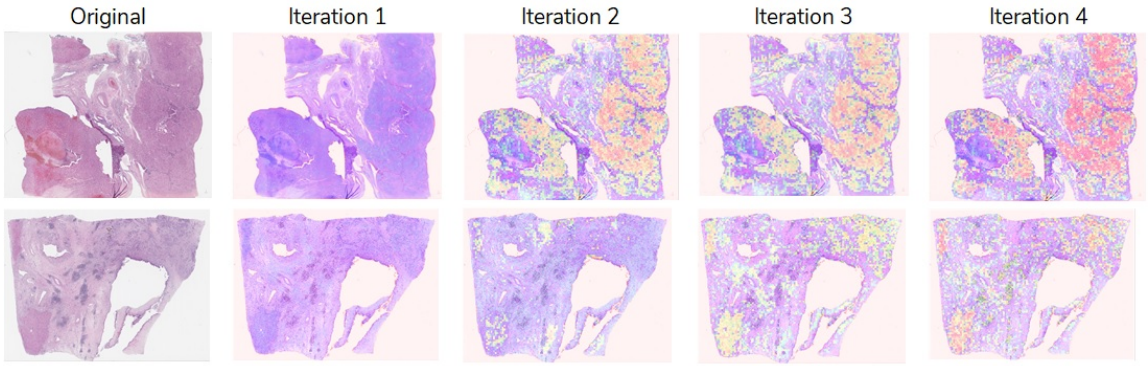

Fig. 5. Heatmap based on probability from hidden class. First Row:Small Duct Type, Second Row: Large Duct Type. [Purple: higher probability; Red: lower probability]

Table 1. Performance of data selection scheme.

Data Selection Scheme	Accuracy
Without any prior information	0.6171
With gene mutation prior information	**0.6343**

Table 2. Performance of different bag formation schemes.

Bag Formation Scheme	Accuracy
None	0.6897
Training sample (>0.7)	0.6857
Training sample (Top 70%)	0.6874
Top70% (WSI >500 patches)	0.6989
Top60% (WSI >500 patches)	**0.7006**
Top50% (WSI >500 patches)	0.6920

four iterations.

The performance of the MIL model for duct type classification was further improved by using simple filtering schemes based on the confidence score from the hidden class detection model. As shown in Table 2, the models in which the training samples with a confidence score of higher than 0.7 and top 70% in each WSI were used to construct the training bags. These showed a comparative performance to the model without a filtering scheme with 0.6857 and 0.6874 in their accuracy, respectively. As the patches of each WSI may be varied, another filtering scheme that included only the top $f\%$ for those WSIs with more than 500 patches after discarding the hidden class was proposed. The models using this scheme outperformed the other models with 0.6989, 0.7006, and 0.6920 for the accuracies of the top 70%, top 60% and top 50%, respectively. It can be deduced that the top $f\%$ filtering scheme is able to collect more discriminative patches for bag construction.

In addition, we plotted a heatmap based on the confidence score of the hidden class (Fig. 5). As shown in Fig. 5, the purple and red regions indicate the areas with high and low confidence for the hidden class, respectively. In other words, the red region (low confidence for the hidden class) was treated as an important region with respect for either the small or large duct type. As the iteration continued, some regions became more red, while the other regions remained purple.

We compared our proposed algorithm to two state-of-the-art methods: baseline MIL information and MetaCleaner [9] with gene mutation prior information (Table 3). For the baseline MIL, we deployed a simple MIL model as described in the Methodology section. MetaCleaner proposed a hallucinate clean representation for noisy-labeled sets, which was similar to our work in cleaning a bag or a subset. We implemented MetaCleaner to clean the noisy-labeled subset and classified it using MLP in order to ensure MetaCleaner was fair compared to the MLP was used in evaluating our proposed method. The proposed hidden class detection with an accuracy of 0.7006 outperformed the MIL and MetaCleaner models which achieved only 0.6343 and 0.6046 in their respective accuracy. When the proposed hidden class detection was applied to the MetaCleaner, it showed an increment of 6% in performance. In summary, the proposed hidden class detection can be an effective way to reduce false positives and false negatives and improve the accuracy of image-level classification tasks.

Table 3. Comparison of state-of-the-art and proposed method with gene mutation prior information.

Method	Accuracy
MetaCleaner [9]	0.6046
MIL	0.6343
Proposed + MetaCleaner [9]	0.6600
Proposed + MIL	**0.7006**

4. CONCLUSION

In this study, we showed that gene mutation prior information is important information in the duct type classification for intrahepatic chalongiocarcinoma. Our proposed hidden class detection method showed its ability and efficiency in boosting the image-level classification. Our experimental results showed that our proposed framework performed better than the baseline model and can be applied to any state-of-the-art models. In our future work, we plan to develop an advanced method to integrate the gene mutation information in our model without manual selection based on the gene mutation information.

5. ETHICAL STANDARDS INFORMATION

This study was approved by the institutional review board of Seoul National University Hospital (IRB NO.H-1011-046-339).

6. ACKNOWLEDGMENTS

This work is partially supported by the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health Welfare, Republic of Korea (HI18C0316) and the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2016M3C4A7952635, NRF-2019M3E5D2A01063819).

7. REFERENCES

[1] Shanshan Zou, Jiarui Li, Huabang Zhou, Christian Frech, Xiaolan Jiang, Jeffrey SC Chu, Xinyin Zhao, Yuqiong Li, Qiaomei Li, Hui Wang, et al., "Mutational landscape of intrahepatic cholangiocarcinoma," *Nature communications*, vol. 5, no. 1, pp. 1–11, 2014.

[2] Ilya G Serebriiskii, Caitlin Connelly, Garrett Frampton, Justin Newberg, Matthew Cooke, Vince Miller, Siraj Ali, Jeffrey S Ross, Elizabeth Handorf, Sanjeevani Arora, et al., "Comprehensive characterization of ras mutations in colon and rectal cancers in old and young patients," *Nature communications*, vol. 10, no. 1, pp. 1–12, 2019.

[3] Jonathan R Brody, Cinthya S Yabar, Mahsa Zarei, Joseph Bender, Lynn M Matrisian, Lola Rahib, Craig Heartwell, Kimberly Mason, Charles J Yeo, Stephen C Peiper, et al., "Identification of a novel metabolic-related mutation (IDH1) in metastatic pancreatic cancer," *Cancer biology & therapy*, vol. 19, no. 4, pp. 249–253, 2018.

[4] Nazanin Fallah-Rad, Philippe L Bedard, Lillian L Siu, Suzanne Kamel-Reid, Helen Chow, Zhang Weijiang, Raymond Jang, Monika K Krzyzanowska, Albiruni RA Razak, and Eric Xueyu Chen, "Association of isocitrate dehydorgenase-1 (IDH-1) mutations with elevated oncometabolite 2-hydroxyglutarate (2hg) in advanced colorectal cancer.," 2016.

[5] Heather D Couture, James Stephen Marron, Charles M Perou, Melissa A Troester, and Marc Niethammer, "Multiple instance learning for heterogeneous images: Training a cnn for histopathology," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 254–262.

[6] Gwenolé Quellec, Guy Cazuguel, Béatrice Cochener, and Mathieu Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE reviews in biomedical engineering*, vol. 10, pp. 213–234, 2017.

[7] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3460–3469.

[8] Jiawen Yao, Xinliang Zhu, and Junzhou Huang, "Deep multi-instance learning for survival prediction from whole slide images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 496–504.

[9] Weihe Zhang, Yali Wang, and Yu Qiao, "Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7373–7382.

[10] Chaoyi Zhang, Yang Song, Donghao Zhang, Sidong Liu, Mei Chen, and Weidong Cai, "Whole slide image classification via iterative patch labelling," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1408–1412.

[11] Mingxing Tan and Quoc V Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[13] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.

[14] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] Darrell R Borger, Kenneth K Tanabe, Kenneth C Fan, Hector U Lopez, Valeria R Fantin, Kimberly S Straley, David P Schenkein, Aram F Hezel, Marek Ancukiewicz, Hannah M Liebman, et al., "Frequent mutation of isocitrate dehydrogenase (IDH) 1 and IDH2 in cholangiocarcinoma identified through broad-based tumor genotyping," *The oncologist*, vol. 17, no. 1, pp. 72, 2012.