Uncertainty-based Data-wise Label Smoothing for Calibrating Multiple Instance Learning in Histopathology Image Classification

Hyeongmin Park, Sungrae Hong, Chanjae Song, Jongwoo Kim, Mun Yong Yi†
Korea Advanced Institute of Science and Technology
{mike980409, sr5043, chan4535, gsds4885, munyi}@kaist.ac.kr

Abstract

Deep neural networks (DNNs) have transformed biomedical image analysis, particularly in histopathology with Whole Slide Images (WSIs) classification. However, training DNNs requires large annotated datasets, which is challenging due to the high heterogeneity and high resolution of WSIs. Multiple Instance Learning (MIL) has become a popular method for weakly supervised classification in this context, training with only slide-level labels. Despite the advancements, ensuring the reliability of model performance is crucial in safety-critical domains including healthcare. Deep learning models in real-world decision-making systems must accurately predict probability estimates to reflect the true likelihood of correctness, known as confidence calibration. This study introduces a novel calibration framework, UDLS, which uses data-wise label smoothing based on predictive uncertainty to improve the calibration of MIL frameworks. This approach involves augmenting WSIs with PatchFeatureDropout, computing predictive uncertainty estimates for original data, and applying these estimates to each sample for label smoothing during model training. Experimental results on benchmark histopathology datasets show noticeable improvements in both calibration and classification performance, highlighting UDLS's potential for enhancing the reliability of predictions from deep learning models in clinical settings.

1. Introduction

In recent years, the application of deep neural networks (DNNs) has expanded across various fields, showing remarkable success. Specifically, the advancement of DNNs has significantly influenced biomedical image analysis such as image classification and segmentation, reducing the reliance on human expertise [15]. One key resource in this area is a set of whole slide images (WSIs), which contain extensive histopathology information motivating the development of automated analysis tools [6]. Accurate classification of WSIs is crucial for effective disease diagnosis, presenting a clear need for clinical applications of deep learning models.

However, the effective training of DNNs for image analysis requires large, annotated datasets, which presents a significant challenge due to high heterogeneity and high resolution (billions of pixels) of WSI [29]. Annotating training data for DNNs is a resource-intensive task that demands considerable time and pathological expertise. Given these constraints, the classification of WSIs has increasingly relied on weakly supervised learning approaches due to their high resolution and the scarcity of pixel-wise annotations [35]. In this regard, multiple-instance learning (MIL) [20] has emerged as a promising method, enabling the training of DNNs using bag (slide)-level labels only without instance (patch)-wise labels for weakly supervised classification. MIL is a method used to train DNNs when full data annotation is unavailable. This approach has been effective in the field of digital pathology diagnosis, where WSIs are labeled, but patches are not. Consequently, WSI classification now largely depends on MIL frameworks [13,18,28,40] for its efficiency and scalability.

Despite the advancements, there is a critical need to ensure the reliability of the model's predictions, especially in safety-critical domains such as healthcare. Deep learning models in real-world decision-making systems are expected not only to be accurate but also to be able to indicate when they might be inaccurate. It is crucial for classification models to accurately predict probability estimates that reflect the true likelihood of correctness, which is known as confidence calibration [9]. In other words, the likelihood of its correctness should be reflected in the probability associated with the predicted class label. A well-calibrated model produces a high probability estimate when its prediction is accurate, and low accuracy when the prediction is uncertain.

Recently, some studies have applied uncertainty quantification to the MIL models of histopathology images [12,36]. However, to effectively utilize the methods when testing, the trained model needs to be well-calibrated [8].

†Corresponding Author

In addition, although uncertainty estimation methods [2, 7, 17] have a calibration effect in itself, these methods require multiple forward runs during the inference, which is impractical for real-time applications. Most existing methods that calibrate deterministic models during training apply common hyper-parameters across the whole dataset. However, these methods do not leverage the variation of inherent uncertainty among different samples. It may lead to degradation in classification performance because it can make easy samples uncertain.

Recently, there have been studies [11, 14, 33] on model reliability to avoid making wrong predictions when uncertainty is high. While regularization methods are well-established, we propose a novel calibration approach, "Uncertainty-based Data-wise Label Smoothing", shortened as UDLS, specifically suited for MIL training used in histopathology image classification, where each instance within a WSI carries different significance. The proposed method adjusts data-wise label smoothing on each training sample (WSI) separately with different values using predictive uncertainty estimates for model calibration. First, we apply PatchFeatureDropout for data augmentation on the bag of patch features and train any MIL model using the augmented dataset. Next, the predictive uncertainty estimates are computed separately for each original input WSI based on multiple predictions. Subsequently, we apply data-wise label smoothing using the uncertainty estimates and the MIL model is retrained using the original WSI dataset along with the newly assigned labels.

Our proposed calibration method can be utilized for histopathology image classification tasks within various state-of-the-art MIL frameworks, improving calibration and classification performance across different datasets. By integrating the proposed calibration method into existing MIL models, we can enhance the applicability of these models in clinical settings, ensuring that pathologists are supported with accurate and trustworthy computational tools. We summarize our contributions as follows:

- We propose a novel data-wise label smoothing method in the MIL framework used in histopathology image classification utilizing predictive uncertainty estimates for model calibration.
- Our plug-in calibration method can be applied to any state-of-the-art MIL framework. The proposed method does not require additional time in practical use as it is integrated into the training process.
- Experiment results show that our framework significantly enhances the calibration performance and the classification accuracy of various MIL models in two different benchmark datasets, attesting its generalizability.

2. Related Work

2.1. Multiple Instance Learning

Multiple instance learning (MIL) is a weakly supervised learning algorithm in which training examples are grouped into collections known as bags, and a label is assigned to each bag.

The standard MIL assumption states that all negative bags contain only negative instances, while positive bags contain at least one positive instance. Let X be a bag of instances $X = \{x_1, x_2, ..., x_B\}$. The goal is to find a mapping from a bag X to a label $Y \in \{0, 1\}$. Moreover, there is a binary label for each instance $x_i \in X$, *i.e.* $y_1, ..., y_B, y_i \in \{0, 1\}$. Though we don't have access to the instance labels when training, each instance x_i can be mapped into a class, where negative and positive classes correspond to 0 and 1 respectively. The main assumption of MIL is defined as follows:

$$Y = \begin{cases} 0, \text{ if and only if } \sum_{b=1}^{B} y_b = 0 \\ 1, \text{ if } \exists y_b : y_b = 1 \end{cases} \tag{1}$$

In the studies on MIL, two approaches are mainly used: instance-based methods and embedding-based methods. The instance-based MIL gives scores in a scalar between 0 and 1 to individual instances using a trained neural network shared among instances. The scalar represents a confidence score that the instance belongs to a class. In the next step, a MIL aggregator merges the scores to determine a label for each bag separately. On the other hand, embedding-based MIL involves creating a feature vector for each instance and combining them using a MIL aggregator to form a single latent embedding representing the entire bag. The MIL pooling layer operates on vector inputs rather than scalar values for a bag of instances. Recently, embedding-based approaches using attention module [13, 28, 40] shows high classification performance. Attention-based pooling operator measures the contribution of each instance to the final bag embedding by weighting the instance feature vectors using attention scores.

The MIL approach is widely applied in digital pathology since it provides a solution when annotated data is unavailable, which is common in digital pathology, where the original WSIs represent bags and the patches in a WSI correspond to instances. After obtaining patches from the initial WSI, every patch is processed through a feature extractor. Subsequently, the features are combined and processed through the MIL aggregator to produce a label corresponding to a WSI. Recent studies [12, 36] quantified the uncertainty of the MIL models in histopathology image classification for model reliability. However, to effectively utilize the methods when testing, the trained model needs to be well-calibrated. To address the issue, we developed a calibration method specifically designed for MIL training.

2.2. Calibrating Neural Networks

Despite the high predictive performance of modern neural networks on different benchmarks, they produce unreliable predictions due to poor calibration. Specifically, deterministic deep neural networks do not give uncertainty estimates and may experience issues with being either over- or under-confident. In this regard, there have been several recent advancements in calibrating models [7, 9, 16, 17, 30].

In supervised classification tasks with neural networks, there are an input X and a label $Y \in \{y_1, ..., y_k\}$, which are random variables that follow a ground truth joint distribution $\pi(X,Y) = \pi(Y|X)\pi(X)$. Let g be a neural network model that maps the input X to a categorical distribution $p = \{p_1, .., p_k\}$ over k classes with $g(X) = (\hat{Y}, \hat{P})$ where $\hat{Y}$ is the class prediction and $\hat{P}$ is its associated confidence. Calibration measures the degree of the match between the predicted confidence estimate $\hat{P}$ and the true correctness likelihood. We define a model g is perfectly calibrated on if and only if:

$$P(\hat{Y} = y_i|\hat{P} = p_i) = p_i, \quad \forall p_i \in [0,1] \tag{2}$$

The state-of-the-art calibration methods can be categorized into post-hoc methods, regularization methods, and uncertainty estimation methods. Post-hoc calibration methods adjust a model after training. These methods include non-parametric calibration histogram binning [38], temperature scaling [9] and Dirichlet calibration [16]. To calibrate confidence scores, these methods necessitate a validation set. Although the post-hoc calibration methods can be easily implemented with scaling, they may not perfectly optimize the calibration of the model, and the outcome is highly affected by the selection of the validation set.

Regularization plays a crucial role in preventing overfitting in neural network models. In general, regularization methods [21, 25, 31] adjust the optimization procedures to produce well-calibrated models. Regularization techniques commonly require the incorporation of hyper-parameters, such as the smoothing factor in label smoothing. Therefore, these methods are sensitive to hyper-parameters, and choosing the right ones requires extensive hyper-parameter tuning. Also, since the same hyper-parameters are applied to all data, the individual data uncertainty cannot be reflected.

Uncertainty estimation methods are designed to improve model calibration by adding variability. These approaches reduce uncertainty in a confidence prediction and yield a more reliable predictor. Common approaches include Bayesian techniques such as Bayesian neural networks [2], Monte-Carlo(MC) dropout [7], and deep ensembles [17]. These methods involve multiple inference runs and perform averaging on predictions, which is impractical for real-time applications. Running multiple inferences for testing is less efficient for application than multiple runs during training.

3. Method

In this section, we present a novel calibration framework named UDLS, uncertainty-based data-wise label smoothing for MIL training. Our proposed approach can be applied to any MIL model for histopathology image classification. First, at Stage-1 Train, bags of instance features generated from a feature extractor are augmented using PatchFeatureDropout, which randomly masks patch features to create diverse data points for training a MIL aggregator and a bag classifier, resulting in multiple predictions from each original input WSI. Next, the predictive entropy is computed from these predictions to measure uncertainty, which is then scaled and applied as a data-wise label smoothing factor, ensuring that each training sample's uncertainty is reflected in its label. Finally, the MIL pooling layer and the classifier are retrained during Stage-2 Train using the original un-augmented bag of feature embedding and the newly assigned labels. The overall framework of our proposed method is represented in Fig. 1.

3.1. Stage-1 Train

Given only the WSI labels, typical MIL frameworks utilize limited information during training, and therefore the trained model poses a high risk of uncertainty. Compared to other uncertainty quantification methods that estimate the uncertainty of a trained model during the test phase, our approach estimates the uncertainty during training.

First, the WSI input is randomly augmented for training a MIL classifier with diverse samples, producing multiple predictions to compute predictive uncertainty. We apply a data augmentation technique called PatchFeatureDropout, inspired by PatchDropout [19], which randomly drops input image patches before processing through a model for memory and computation efficiency. PatchFeatureDropout randomly drops patch features after a feature extractor in order to save computation. The augmented bag of features uses the ground-truth label of the un-augmented original input WSI. After augmenting T times for each slide of a dataset with N WSIs, we get $T \times N$ bags of instance features. Specifically, for an original bag X_i with K instances, a feature extractor (*e.g.* convolutional neural networks) extracts the bag of feature embedding $F_i \in R^{K \times D}$.

After applying PatchFeatureDropout for each input WSI with a PatchDropout rate of r, we get T augmented bag of feature embedding (*i.e.* $F_{i,1}, ..., F_{i,T}$, $F_{i,t} \in R^{K(1-r) \times D}$). The MIL classifier is then trained with the augmented bag of instances to get the final bag representation and create a classification for each augmented bag separately. In the Stage-1 Train, a cross-entropy is adopted as:

$$\mathscr{L}_1 = -\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{c=1}^{C} y_c(X_i) log(p(F_{i,t})[c]) \tag{3}$$

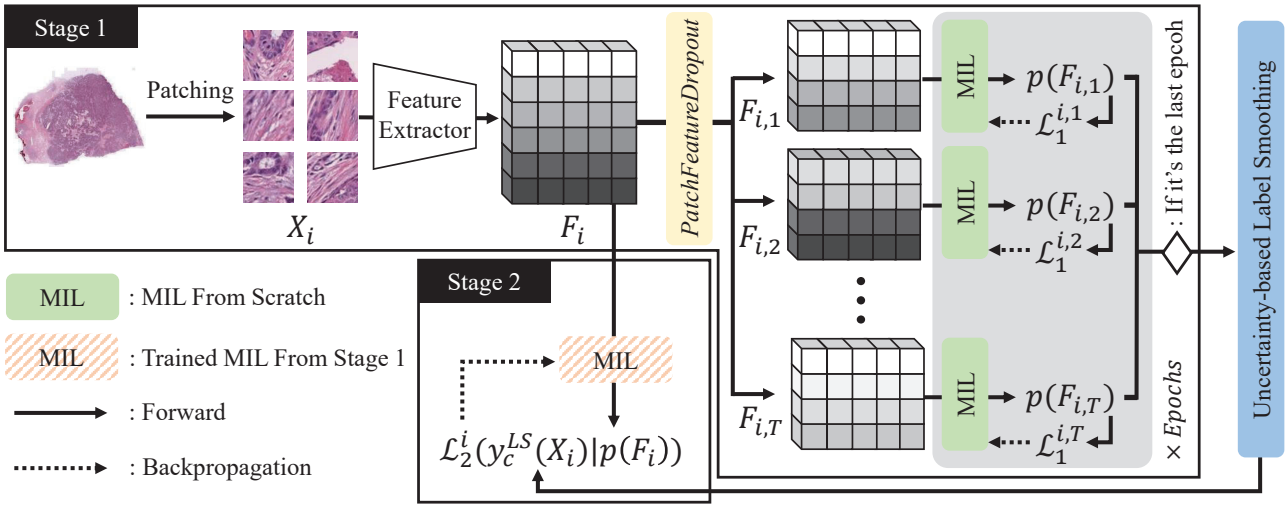


Figure 1. The overall framework of MIL training using UDLS.

where each prediction $p(F_{i,t})$ is a softmax vector for the C different classes, and $p(F_i)[c]$ is the c-th element of the vector $p(F_i)$. The class label of the original bag $y_c(X_i)$ is "1" for the correct class and "0" for the rest.

3.2. Uncertainty-based Label Smoothing

The uncertainty estimation is measured based on different predictions produced by different augmentations of the initial input WSI. For an original bag of instance features F_i, T different predictions are generated after the Stage-1 Train. Then the mean prediction for the T augmented samples can be calculated as follows:

$$p_T(F_i) = \frac{1}{T}\sum_{t=1}^{T} p(F_{i,t}) \tag{4}$$

Inspired by the prior studies [3, 24, 26], we compute the predictive entropy for measuring uncertainty estimates. The predictive entropy represents the uncertainty in a random variable's potential outcomes [27]. It is calculated by determining the entropy of the distribution of the predictive probability vector:

$$H(p_T(F_i)) = -\sum_{c=1}^{C} p_T(F_i)[c] log(p_T(F_i)[c]) \tag{5}$$

The predictive entropy $H(p_T(F_i))$ is scaled into $LS(X_i;\alpha) \in [0,\alpha]$, where α is a label smoothing factor, to avoid overdominant effect that may alter the ground truth.

$$LS(X_i;\alpha) = \frac{(H(p_T(F_i)) - H_{min})\cdot\alpha}{H_{max} - H_{min} + \varepsilon} \tag{6}$$

$$\text{where } \begin{cases} H_{min} = \min_{1\le n\le N} H(p_T(F_n)) \\ H_{max} = \max_{1\le n\le N} H(p_T(F_n)) \end{cases}$$

Then, the label of each bag X_i to be class c is modified using the corresponding data-wise label smoothing factor $LS(X_i)$.

$$y_c^{LS}(X_i) = y_c(X_i)(1 - LS(X_i;\alpha)) + LS(X_i;\alpha)/C \tag{7}$$

where C is the number of classes.

In the original label smoothing method [21], the label smoothing factor α is a pre-defined global hyper-parameter (0.05 or 0.1) that is fixed and uniformly applied to every training sample. Meanwhile, the proposed UDLS method calculates predictive uncertainty estimates for each sample, which is then scaled by α to generate data-wise smoothing factor $LS(X_i;\alpha)$, allowing the label smoothing factor to vary for each data point based on its predictive uncertainty.

3.3. Stage-2 Train

After the data-wise label smoothing, the class label of an un-augmented original bag $y_c(X_i)$ is updated to $y_c^{LS}(X_i)$. Finally, the MIL classifier is retrained using the N original bag of feature embedding F_i and the newly assigned corresponding class label $y_c^{LS}(X_i)$. The weights of the MIL classifier are initialized with the weights of the network from the last epoch of the Stage-1 Train. The loss function for the Stage-2 Train using cross-entropy is defined as:

$$\mathcal{L}_2 = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_c^{LS}(X_i) log(p(F_i)[c]) \tag{8}$$

Since the proposed method does not modify the training process of MIL pooling aggregator, it can be applied to any state-of-the-art MIL framework. The pseudo-code representing the overall process of the UDLS is presented in Appendix A.

4. Experiments

4.1. Experimental Setup

4.1.1 Dataset

The proposed calibration methods are experimented on two public histopathology WSI datasets: Camelyon16 [1] and the Cancer Genome Atlas (TCGA) lung cancer. Both datasets are for binary classification, "0" for negative slides, and "1" for positive slides. Camelyon16 is a public dataset suggested for metastasis detection in breast cancer, consisting of 270 training slides and 130 test slides. TCGA dataset is randomly split into 729 training slides and 183 testing slides. Compared to Camelyon16, tumor slides in the TCGA dataset contain larger tumor regions, leading to a high portion of positive instances in positive bags.

4.1.2 Baselines and Comparison Methods

Our calibration method is experimented on various state-of-the-art MIL frameworks, AB-MIL [13], Trans-MIL [28], and DTFD-MIL [40]. Label Smoothing [30], temperature scaling [9], MC dropout [7], and deep ensembles [17] were used as calibration baselines. Following the original work [30], label smoothing is applied with a factor of 0.1 for AB-MIL and Trans-MIL. Because DTFD-MIL is relatively better calibrated than the other models, the network is trained with a label smoothing factor of 0.05. In order to optimize the temperature parameter for temperature scaling, 50 slides of the Camelyon16 train set and 104 slides of the TCGA train set were utilized as validation sets. In addition, a reasonable estimation from MC dropout [7] is obtained with 10 iterations, with a dropout rate of 0.5. Finally, the deep ensembles are trained with 10 independent networks.

4.1.3 Evaluation Metrics

Evaluating calibration involves assessing the statistical alignment between the predicted distributions and observations [32]. Most calibration metrics estimate model calibration errors from finite samples by grouping N predictions into M interval bins $b_1, \cdots, b_M$ and calculating the classification accuracy of each bin. The predictions are sorted by the predicted confidence $\hat{p}_i$ and grouped into M bins.

Let B_m be the set of indices of samples whose predicted confidence falls into a bin interval b_m [9]. Then, the calibration of a single bin is evaluated by calculating the average bin confidence in relation to the average bin accuracy. The average bin accuracy of B_m is defined as:

$$acc(b_m) = \frac{1}{|b_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i) \tag{9}$$

where $\hat{y}_i$ is the predicted label and y_i is the true class label for sample i.

The average bin confidence of B_m is defined as:

$$conf(b_m) = \frac{1}{|b_m|} \sum_{i \in B_m} \hat{p}_i \tag{10}$$

where $\hat{p}_i$ refers to the predicted confidence for sample i.

Reliability Diagram For a visual representation of model calibration, the reliability diagram [4] [23] is widely used, showing whether a model is over- or under-confident on bins. The diagram plots the expected sample accuracy $acc(b_m)$ against the average bin confidence $conf(b_m)$, enabling a comprehensive examination of how well the predicted probabilities align with the observed frequencies of outcomes. It plots close to the identity function for a well-calibrated model. Deviation from the diagonal line represents a calibration error. Points above the diagonal indicate that the predicted probabilities are overconfident while points below the diagonal represent underconfidence. However, the proportion of samples in each bin is not presented in the diagrams.

Expected Calibration Error (ECE) Expected Calibration Error (ECE) [22] is a scalar summary statistic of calibration for binning-based calibration measures to approximate the calibration error in expectation. ECE calculates the deviation between $acc(b_m)$ and $conf(b_m)$ by partitioning the predictions into M equally-spaced bins $b_1, \cdots, b_M$ in the same way as the reliability diagrams. It quantifies the average absolute difference between the mean of the predicted probabilities and the fraction of observed frequencies within each bin. The ECE is computed by taking the weighted average of the calibration errors across all bins:

$$ECE = \sum_{m=1}^{M} \frac{|b_m|}{N} |acc(b_m) - conf(b_m)| \tag{11}$$

where $|b_m|$ indicates the number of predictions in bin b_m and N is the total number of samples. ECE is used as the primary calibration measure in our experiments.

4.1.4 Implementation Details

The patch features are extracted in a 1024-dimensional vector with a ResNet50 [10] pre-trained on ImageNet-1k [5]. MIL aggregators are trained with a learning rate of $1e-4$, and the mini-batch size for training MIL models is 1 bag. AB-MIL, Trans-MIL, and DTFD-MIL are trained for 100, 40, and 200 epochs, respectively. For the UDLS, PatchFeatureDropout is applied to each bag of instance features with a PatchDropout rate r of 0.3 and augmented 10 times. Since DTFD-MIL possesses relatively less calibration error compared with AB-MIL and Trans-MIL, we use the label smoothing factor α of 0.1 for AB-MIL and Trans-MIL, and 0.05 for DTFD-MIL.

Method	Multiple Inference	AB-MIL [13]						Trans-MIL [28]						DTFD-MIL [40]					
		Camelyon16			TCGA			Camelyon16			TCGA			Camelyon16			TCGA		
		R@10	R@30	R@50	R@10	R@30	R@50	R@10	R@30	R@50	R@10	R@30	R@50	R@10	R@30	R@50	R@10	R@30	R@50
w/o calibration	✗	0.9	0.87	0.82	0.6	0.80	0.76	**1**	**1**	**0.94**	**1**	0.97	0.94	0.9	0.87	0.71	**1**	**1**	0.97
Temp Scaling [9]	✗	0.9	0.57	0.54	0.9	0.93	**0.96**	**1**	**1**	0.92	0.9	0.93	0.96	**1**	**0.93**	0.72	**1**	**1**	0.98
Label Smoothing [21]	✗	0.9	0.83	0.76	0.5	0.67	0.70	0.9	0.97	0.92	**1**	0.97	0.96	**1**	**0.93**	0.73	**1**	**1**	0.90
MC dropout [7]	✓	0.9	0.67	0.63	**1**	0.93	0.95	**1**	0.97	0.92	**1**	**1**	**1**	**1**	**0.93**	0.72	**1**	**1**	0.94
Deep Ensembles [17]	✓	0.9	0.73	0.80	0.8	0.93	0.88	**1**	0.97	0.92	**1**	0.97	0.95	**1**	**0.93**	0.72	**1**	**1**	0.93
UDLS(Ours)	✗	**1**	**0.97**	**0.86**	**1**	**0.97**	0.95	**1**	**1**	0.92	**1**	**1**	**1**	**1**	**0.93**	**0.73**	**1**	**1**	**0.99**

Table 1. Comparison results on top-N positive predictions. R@N stands for recall@N, which is the recall of top-N positive predictions.

Method	AB-MIL [13]						Trans-MIL [28]						DTFD-MIL [40]					
	Camelyon16			TCGA			Camelyon16			TCGA			Camelyon16			TCGA		
	Acc ↑	AUC ↑	ECE ↓	Acc ↑	AUC ↑	ECE ↓	Acc ↑	AUC ↑	ECE ↓	Acc ↑	AUC ↑	ECE ↓	Acc ↑	AUC ↑	ECE ↓	Acc ↑	AUC ↑	ECE ↓
w/o calibration	0.724	0.762	0.152	0.705	0.795	0.139	0.905	**0.927**	0.147	0.798	0.916	0.162	**0.850**	0.880	0.141	0.820	0.916	0.165
Temp Scaling [9]	0.637	0.630	0.088	0.814	0.881	0.202	0.913	0.916	0.088	0.852	0.935	**0.079**	0.756	0.839	0.121	0.825	0.913	0.077
Label Smoothing [21]	0.732	0.749	0.151	0.628	0.726	0.136	0.890	0.906	0.068	0.830	0.896	0.105	0.771	0.837	0.138	0.803	0.917	0.089
MC dropout [7]	0.787	0.757	0.224	0.765	0.807	0.234	**0.927**	0.915	0.073	0.885	0.945	0.115	0.836	**0.879**	0.148	**0.830**	0.916	0.098
Deep Ensembles [17]	0.811	0.860	0.175	0.852	0.885	0.210	0.851	0.900	0.091	0.872	0.925	0.102	0.717	0.840	0.129	0.803	0.911	0.093
UDLS(Ours)	**0.850**	**0.922**	**0.060**	**0.863**	**0.904**	**0.093**	0.898	0.914	**0.064**	**0.886**	**0.945**	0.082	0.803	0.857	**0.115**	0.824	**0.920**	**0.076**

Table 2. Comparison results of classification and calibration performance.

4.2. Measuring Reliability of Positive Predictions

In a histopathology image classification task, reducing type II error (false negatives) is crucial. If a model produces an inaccurate prediction for a patient, the diagnosis could be postponed, possibly causing a missed chance for treatment. In order to verify the reliability of a MIL model from the practical perspective, the positive samples that the classification model determines to be highly confident are examined. We sorted the samples with top-N predictions and calculated the recall (1 − type II error). Tab. 1 shows that the MIL model calibrated with UDLS has the highest or second-highest recall across various MIL frameworks on both datasets, indicating that the proposed method is effective when used in conjunction with those diverse frameworks. In addition, since UDLS does not require multiple inference runs, it is efficient in real-time applications.

4.3. Quantitative Results on Various Models

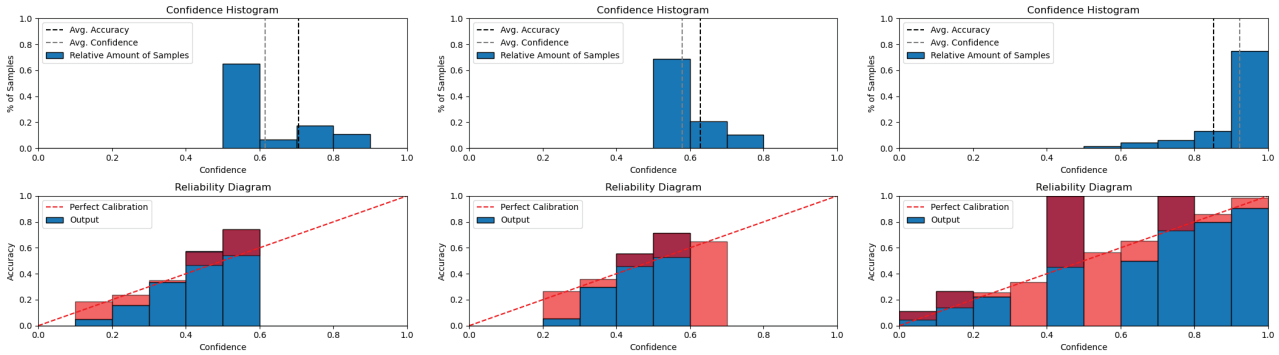
The overall classification and calibration results of our proposed UDLS and the baseline calibration methods on different MIL models are shown in Tab. 2. The results demonstrate that the proposed UDLS outperforms existing calibration methods in most cases. Specifically, UDLS consistently improves accuracy, AUC, and ECE, highlighting its effectiveness in enhancing model calibration and classification performance. Since the proposed data-wise label smoothing approach is applied based on the predictive uncertainty estimates for each sample after the Stage-1 Train, the hard samples would be penalized with high label smoothing factors. Therefore, the model can then be generalized with new data with more accurate and calibrated results during inference. The results also indicate that UDLS is particularly powerful on AB-MIL, suggesting that the method is more effective on under-confident models.

4.4. Qualitative Results on Various Models

Fig. 2, Fig. 3, and Fig. 4 present the confidence histograms and the reliability diagrams that compare the calibration results of the backbone MIL models without calibration, with label smoothing, and with the proposed uncertainty-based data-wise label smoothing on the TCGA image classification. The top rows show the histograms of the predicted confidence, where the dashed black lines indicate the average accuracy, and the dashed gray lines indicate the average confidence. The bottom rows show the reliability diagrams, which plot the average bin accuracy against the average bin confidence of the positive labels. The gaps indicate over- or under-confidence in the corresponding bins. The results of the Camelyon16 classification are presented in Appendix C.

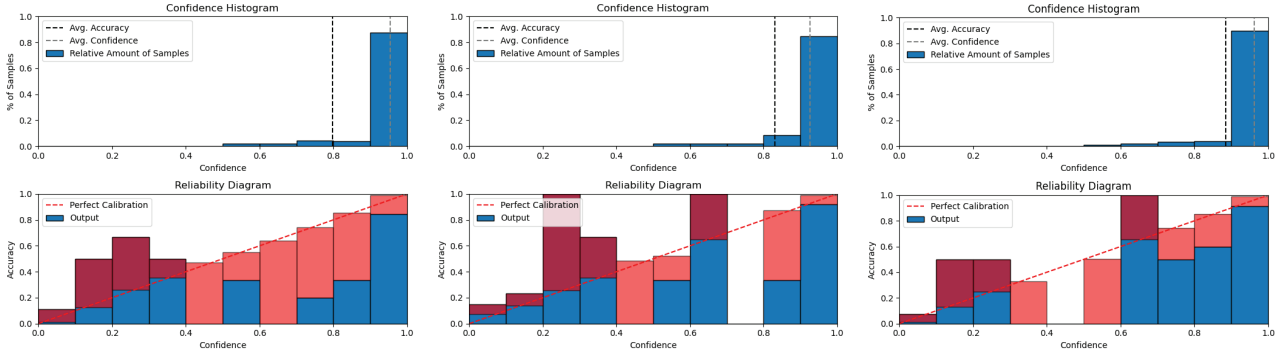
The confidence histograms and the reliability diagrams from Fig. 2 indicate that the AB-MIL is under-confident, and the under-confidence is not resolved with label smoothing. On the other hand, results on UDLS show a notable improvement, with a reduced gap between the average confidence and the average accuracy in the confidence histograms, and positive confidence appearing in every bin in the reliability diagrams, suggesting that the proposed UDLS is more effective than the original label smoothing for the under-confident models in identifying easy samples.

Fig. 3 and Fig. 4 show that Trans-MIL and DTFD-MIL are over-confident, and both the UDLS and the original label smoothing are effective for calibrating the over-confident models. The gap between the average accuracy and average confidence in the confidence histograms and the calibration gap in each bin of reliability diagrams are reduced. However, the confidence histograms indicate that the average accuracy of the label smoothing is always lower than the average accuracy of the UDLS. According to the



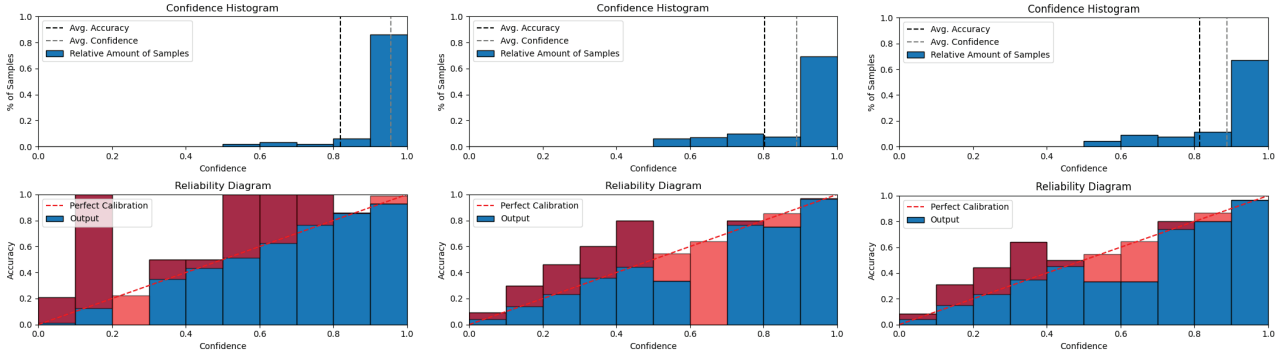
(a) Without Calibration (b) Label Smoothing (c) UDLS

Figure 2. The confidence histograms and the reliability diagrams of AB-MIL on TCGA dataset.



(a) without calibration (b) Label Smoothing (c) UDLS

Figure 3. The confidence histograms and the reliability diagrams of Trans-MIL on TCGA dataset.



(a) without calibration (b) Label Smoothing (c) UDLS

Figure 4. The confidence histograms and the reliability diagrams of DTFD-MIL on TCGA dataset.

memorization effect [39], DNNs typically learn simpler samples early in training and then fit harder ones later. In a prior study [34], it was discovered that DNNs start to fit challenging samples after a few epochs, whereas regularization approaches such as label smoothing reduce the certainty of easy samples. On the other hand, since our proposed method performs data-wise label smoothing based on the predictive uncertainty, the model is trained with the information of the easy and hard samples, preventing the reduction of the certainty of easy ones and resulting in improved classification compared to the original label smoothing method.

4.5. Further Analysis

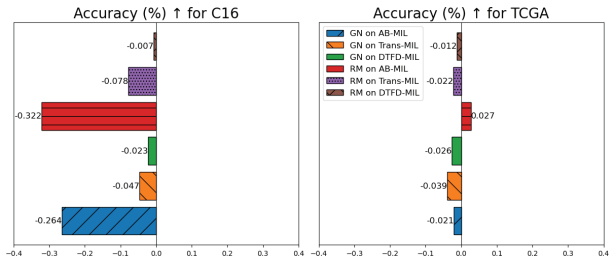
4.5.1 Comparison with Different Data Augmentations

We compare the impact of PatchFeatureDropout on estimating predictive uncertainty with different WSI augmentation methods for classification and calibration across backbone MIL models. Our key idea is to transform data in a latent space for efficient training. The basic transformation involves adding random noise from a Gaussian distribution with a mean of zero and a standard deviation calculated across patch features in a bag. A recent WSI augmentation method builds on the technique proposed in ReMix [37]. It first reduces the number of instances in bags by replacing instances with clustering centroids. Then, latent space augmentation is applied by generating a new representation from the key covariance matrix and appending it to the bag.

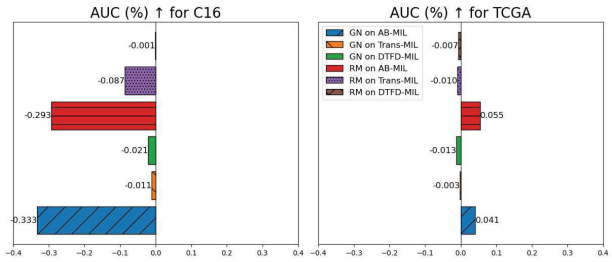
Fig. 5 shows the classification and calibration performance of WSI augmentation methods, compared to the performance of our method when set to 0. Results indicate that PatchFeatureDropout improves model performance in most cases compared to other methods. It suggests that dropping a set of features is more effective in training MIL models based on UDLS, while other augmentation methods may inadvertently distort important instance features.

4.5.2 Ablation Studies

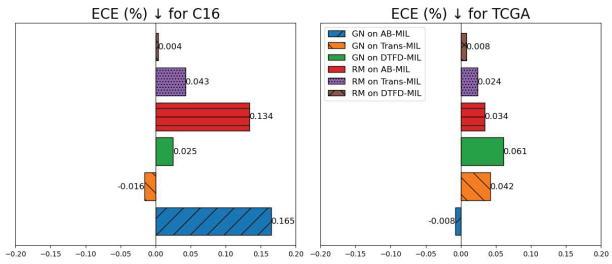
In this section, we report the outcomes of the experiments conducted with different versions of the UDLS to find out how the various settings of its main components affect classification and calibration performance. In Fig. 6, we include ablation studies with our method, where the same model is trained (1) without calibration, (2) without utilizing predictive uncertainty, (3) without scaling the smoothing factor, and (4) with the UDLS on the TCGA dataset. The results indicate that label smoothing may harm classification performance, but utilizing predictive uncertainty estimates can make models more accurate. In addition, scaling the uncertainty estimates further improves the classification and calibration performance in all MIL models since it can avoid the overdominant effects that may modify the ground truth.



(a) Accuracy ↑ compared with Ours (using PatchFeatureDropout)

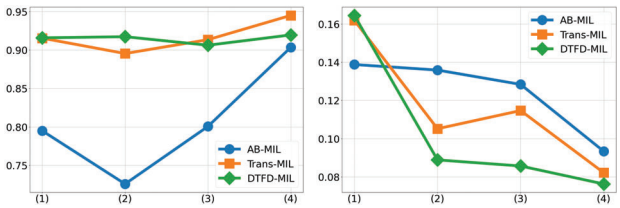


(b) AUC ↑ compared with Ours (using PatchFeatureDropout)



(c) ECE ↓ compared with Ours (using PatchFeatureDropout)

Figure 5. Experiment results using different WSI augmentation methods. GN: Adding Gaussian noise, RM: ReMix [37]



(a) AUC ↑ (b) ECE ↓

Figure 6. Ablation studies for applying the data-wise predictive uncertainty and scaling the smoothing factor. (1) w/o calibration (2) w/o uncertainty estimation (3) w/o scaling (4) UDLS

5. Conclusion

In this paper, we propose a novel calibration method for MIL frameworks in histopathology image classification. By leveraging data-wise label smoothing based on uncertainty estimates, our approach effectively addresses the challenges of miscalibration in model predictions. This method adjusts label smoothing based on predictive entropy for each sample, ensuring that the predictive uncertainty is appropriately reflected during training. The experimental results demonstrate that the proposed method enhances both model calibration and classification performance, making it a robust approach for deploying reliable MIL models in histopathology image analysis.

Acknowledgment

This research was supported by National Research Foundation of Korea (NRF-2022M3J6A1063021).

References

[1] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 5

[2] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015. 2, 3

[3] Marc Combalia, Ferran Hueto, Susana Puig, Josep Malvehy, and Veronica Vilaplana. Uncertainty estimation in deep neural networks for dermoscopic image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 744–745, 2020. 4

[4] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983. 5

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[6] Neofytos Dimitriou, Ognjen Arandjelović, and Peter D Caie. Deep learning for whole slide image analysis: an overview. *Frontiers in medicine*, 6:264, 2019. 1

[7] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2, 3, 5, 6

[8] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023. 1

[9] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 1, 3, 5, 6

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[11] Jay Heo, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, and Sung Ju Hwang. Uncertainty-aware attention for reliable interpretation and prediction. *Advances in neural information processing systems*, 31, 2018. 2

[12] Julianna D Ianni, Rajath E Soans, Sivaramakrishnan Sankarapandian, Ramachandra Vikas Chamarthi, Devi Ayyagari, Thomas G Olsen, Michael J Bonham, Coleman C Stavish, Kiran Motaparthi, Clay J Cockerell, et al. Tailored for real-world: a whole slide image classification system validated on uncurated multi-site data emulating the prospective pathology workload. *Scientific reports*, 10(1):3217, 2020. 1, 2

[13] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 2, 5, 6

[14] Mahdi Khodayar, Saeed Mohammadi, Mohammad E Khodayar, Jianhui Wang, and Guangyi Liu. Convolutional graph autoencoder: A generative deep neural network for probabilistic spatio-temporal solar irradiance forecasting. *IEEE Transactions on Sustainable Energy*, 11(2):571–583, 2019. 2

[15] Minjeong Kim, Chenggang Yan, Defu Yang, Qian Wang, Junbo Ma, and Guorong Wu. Deep learning in biomedical image analysis. In *Biomedical information technology*, pages 239–263. Elsevier, 2020. 1

[16] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019. 3

[17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2, 3, 5, 6

[18] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 1

[19] Yue Liu, Christos Matsoukas, Fredrik Strand, Hossein Azizpour, and Kevin Smith. Patchdropout: Economizing vision transformers using patch dropout. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3953–3962, January 2023. 3

[20] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997. 1

[21] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019. 3, 4, 6

[22] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015. 5

[23] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005. 5

[24] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019. 4

[25] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017. 3

[26] Alicja Rączkowska, Marcin Możejko, Joanna Zambonelli, and Ewa Szczurek. Ara: accurate, reliable and active histopathological image classification framework with bayesian deep learning. *Scientific reports*, 9(1):14347, 2019. 4

[27] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001. 4

[28] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021. 1, 2, 5, 6

[29] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical image analysis*, 67:101813, 2021. 1

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3, 5

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[32] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *The 22nd international conference on artificial intelligence and statistics*, pages 3459–3467. PMLR, 2019. 5

[33] Nurali Virani, Naresh Iyer, and Zhaoyuan Yang. Justification-based reliability in machine learning. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 6078–6085, 2020. 2

[34] Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. *Advances in Neural Information Processing Systems*, 34:11809–11820, 2021. 7

[35] Xi Wang, Hao Chen, Caixia Gan, Huangjing Lin, Qi Dou, Efstratios Tsougenis, Qitao Huang, Muyan Cai, and Pheng-Ann Heng. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE transactions on cybernetics*, 50(9):3950–3962, 2019. 1

[36] Xi Wang, Fangyao Tang, Hao Chen, Luyang Luo, Ziqi Tang, An-Ran Ran, Carol Y Cheung, and Pheng-Ann Heng. Ud-mil: uncertainty-driven deep multiple instance learning for oct image classification. *IEEE journal of biomedical and health informatics*, 24(12):3431–3442, 2020. 1, 2

[37] Jiawei Yang, Hanbo Chen, Yu Zhao, Fan Yang, Yao Zhang, Lei He, and Jianhua Yao. Remix: A general and efficient framework for multiple instance learning based whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2022. 8

[38] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616, 2001. 3

[39] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 7

[40] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022. 1, 2, 5, 6