



ARTICLE

Two-Stage Teacher-Student Learning with Confidence-Based Label Curation for Multi-Annotator Endoscopic Anatomical Landmark Classification

Daesung Kang*

School of Bio-Health Convergence, College of Natural Sciences, Sungshin Women's University, Seoul, Republic of Korea

*Corresponding Author: Daesung Kang, Email: mpkds@sungshin.ac.kr

Received: 06 July 2026; Accepted: 18 September 2026; Published: 21 October 2026

ABSTRACT: Background: Public medical image datasets increasingly contain annotations from multiple experts, resulting in substantial inter-observer disagreement. Conventional supervised learning typically aggregates these annotations into a single hard label or discards ambiguous samples, potentially introducing label noise or reducing the amount of informative training data. This study proposes a confidence-based teacher-student framework to utilize disagreement samples while reducing the influence of potentially unreliable annotations. Methods: We propose a two-stage confidence-guided label curation (TS-CL) framework for anatomical landmark classification in upper gastrointestinal endoscopy using the GastroHUN dataset. In the first stage, a teacher model was trained exclusively on images with complete agreement among four endoscopists. The trained teacher then assigned pseudo-labels to disagreement samples whose prediction confidence exceeded a predefined threshold, while low-confidence samples were excluded from student training. In the second stage, student models were trained using the resulting curated dataset. Four representative backbone architectures (ConvNeXt-Tiny, ResNet50, Vision Transformer, and Swin Transformer) were evaluated under different confidence thresholds. Each configuration was trained using three independent random seeds. Results: Across all evaluated architectures, TS-CL achieved higher mean classification performance than the corresponding baseline models trained using only complete agreement samples. At the selected confidence threshold of $\tau = 0.9$, macro F1-score increased by 1.72, 2.25, 1.23, and 1.27 percentage points for ConvNeXt-Tiny, ResNet50, ViT-B, and Swin-B, respectively. ResNet50 showed the largest improvement, with macro F1-score increasing from 82.57% to 84.82% and MCC from 82.29% to 84.53%. The common operating threshold of $\tau = 0.9$ was selected based exclusively on validation macro F1-score across the four backbones. In additional comparisons, TS-CL achieved higher mean macro F1-scores than the one-stage variant, plurality-vote hard-label baseline, and annotator-specific multi-head model across all four backbones, and than Co-teaching in three of the four backbones. Conclusions: Confidence-based label curation can utilize selected disagreement samples as additional supervision while reducing the influence of potentially unreliable annotations. The proposed TS-CL framework provides a simple and practical strategy for improving supervised learning from multi-annotator medical image datasets and may be applicable to a wide range of medical imaging tasks beyond upper gastrointestinal endoscopy.

KEYWORDS: Gastric endoscopy; label curation; multi-annotator learning; pseudo-labeling

1 Introduction

Gastrointestinal (GI) cancer, particularly gastric cancer, remains a major global public health concern, with patient prognosis depending heavily on early detection [1]. Esophagogastroduodenoscopy (EGD) is the standard screening and diagnostic procedure for upper GI diseases and plays a crucial role in detecting early gastric cancer (EGC), thereby reducing gastric cancer-related mortality [2]. However, EGD is highly

operator-dependent; previous studies have reported that approximately 20%–25% of EGC lesions are missed during routine examinations [3]. To improve examination quality and minimize blind spots, standardized photodocumentation protocols have been proposed. Among them, the Systematic Screening Protocol for the Stomach (SSS), introduced by Yao, recommends acquiring 22 representative endoscopic images covering the entire stomach [4]. Although the SSS is intended to standardize gastric photodocumentation and improve examination completeness, manual verification of protocol adherence is labor-intensive and impractical in routine clinical practice, motivating the development of automated quality assessment systems based on artificial intelligence (AI).

To support AI-based quality assessment, the recently introduced GastroHUN dataset provides a public benchmark for anatomical landmark classification in upper GI endoscopy [5]. The dataset contains 8834 endoscopic images collected from 387 patients following the SSS protocol and includes 22 gastric anatomical landmarks together with one Not Applicable (NA) category. Unlike conventional datasets, GastroHUN was independently annotated by four endoscopists through a quadruple-blind annotation process, making it a realistic multi-annotator benchmark. While this annotation strategy better reflects clinical practice, it also introduces substantial inter-observer disagreement, particularly for anatomically ambiguous images. Conventional supervised learning typically resolves annotator disagreement by aggregating multiple annotations into a single hard label, most commonly through majority voting. However, reducing multiple annotations to a single hard label may obscure clinically meaningful disagreement among experts, whereas excluding disagreement samples decreases the amount of potentially informative training data [6–9]. These challenges highlight the need for learning strategies that can effectively utilize disagreement information while limiting the influence of potentially unreliable annotations.

To address annotation uncertainty, previous studies have investigated various learning strategies, including annotator-specific multi-head architectures, soft-label learning methods that preserve annotator uncertainty, and approaches that explicitly learn from disagreement [10–13]. Related work on uncertainty-aware pseudo-labeling has further demonstrated the importance of considering prediction uncertainty when selecting pseudo-labeled samples [14]. Although these approaches can improve robustness to annotation variability, many of them retain disagreement samples throughout optimization or explicitly model annotation uncertainty within the learning objective, rather than constructing a curated training subset before final model training. However, these strategies do not explicitly address the complementary question of whether a reliable subset of disagreement samples can be identified before final model training. Therefore, an effective strategy is needed to exploit informative disagreement samples while reducing the influence of potentially unreliable annotations.

In this study, we propose a two-stage confidence-guided label curation (TS-CL) framework for multi-annotator endoscopic image classification. In the first stage, a teacher model is trained exclusively using the complete agreement subset of the GastroHUN dataset, where all four endoscopists assigned identical labels. The trained teacher model then evaluates the remaining disagreement samples and assigns pseudo-labels to samples with prediction confidence above a predefined threshold (τ), while low-confidence samples are excluded from training. In the second stage, a student model is trained using the resulting curated dataset, which combines the original complete agreement samples with the confidently pseudo-labeled disagreement samples. This teacher-student framework is designed to utilize informative disagreement samples while reducing the influence of potentially unreliable annotations.

Experimental results on the GastroHUN benchmark demonstrated that the proposed TS-CL framework achieved higher mean performance than the corresponding complete agreement baselines across both convolutional neural networks (ConvNeXt-Tiny and ResNet50) and transformer-based architectures (ViT-B and Swin-B). Furthermore, the effect of confidence-based curation was architecture-dependent, and $\tau = 0.9$

was selected as the common operating threshold based exclusively on validation performance across the evaluated backbones.

The main contributions of this study are summarized as follows:

- We propose a novel two-stage confidence-guided label curation (TS-CL) framework that utilizes multi-annotator disagreement through confidence-based pseudo-labeling while reducing the influence of potentially unreliable annotations.
- We systematically investigate the trade-off between pseudo-label reliability and data utilization across different confidence thresholds and show that threshold-dependent curation behavior varies across teacher architectures, with $\tau = 0.9$ selected as the common operating threshold based exclusively on validation performance.
- We evaluate TS-CL across four representative backbone architectures (ConvNeXt-Tiny, ResNet50, ViT-B, and Swin-B) and three independent runs, demonstrating higher mean performance than the corresponding complete agreement baselines and competitive performance relative to alternative training and multi-annotator learning strategies.

2 Proposed Method

2.1 Overview of the Proposed TS-CL Framework

The proposed TS-CL framework is designed to utilize images with inter-observer disagreement through offline confidence-based label curation. Let $D_C = \{(x_i, y_i)\}_{i=1}^{N_C}$ denote the complete agreement subset, where x_i represents an endoscopic image, y_i is the consensus label assigned by all four annotators, and N_C denotes the number of complete agreement samples. Similarly, $D_D = \{(x_j, \mathbf{a}_j)\}_{j=1}^{N_D}$ denotes the disagreement subset, where $\mathbf{a}_j = (a_{j,1}, \dots, a_{j,R})$ represents the labels independently assigned by $R = 4$ annotators to image x_j , and N_D denotes the number of disagreement samples.

The proposed framework consists of two sequential stages. In Stage 1, a teacher model is trained using the complete agreement subset D_C and subsequently applied to the disagreement subset D_D for confidence-based label curation. High-confidence disagreement samples are retained and assigned pseudo-labels, whereas low-confidence samples are excluded from student training. In Stage 2, a student model is trained using the resulting curated dataset. Fig. 1 illustrates the overall workflow of the proposed TS-CL framework.

In this study, confidence is defined as the maximum softmax probability predicted by the teacher and is used only for pseudo-label selection. Unlike classical confident learning, which estimates a confident joint between observed and latent labels, the proposed framework does not estimate a confident joint or a class-conditional noise matrix [15].

2.2 Stage 1: Teacher Model Training and Confidence-Based Label Curation

The teacher model $f(\cdot; \theta_T)$ is trained using the complete agreement subset. Because class frequencies differ across the anatomical categories, the teacher is optimized using class-weighted cross-entropy:

$$L_T = -\frac{1}{N_C} \sum_{i=1}^{N_C} w_{y_i}^{CA} \log P(y_i | x_i; \theta_T), \quad (1)$$

where $w_{y_i}^{CA}$ denotes the class weight for class y_i , calculated from the complete agreement training subset as $w_c = N / (C n_c)$, where N is the total number of training samples, $C = 23$ is the number of classes, and n_c is the number of training samples belonging to class c . Thus, the weights correspond to balanced inverse-frequency weighting. θ_T represents the teacher model parameters.

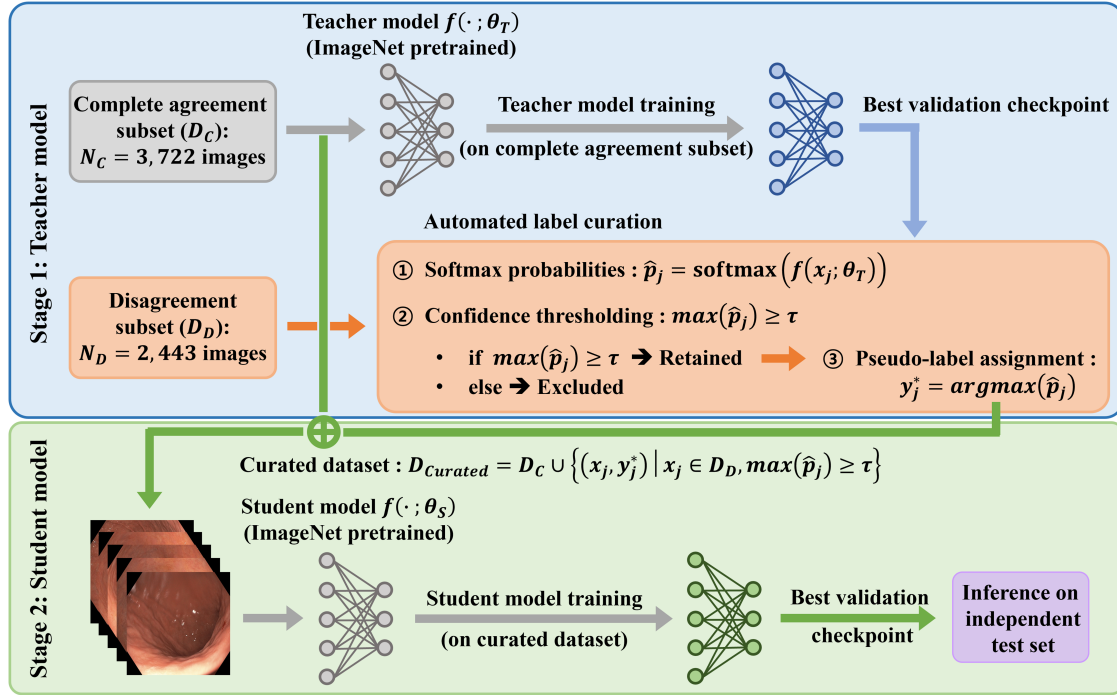


Figure 1: Overview of the proposed Two-Stage Confidence-Guided Label Curation (TS-CL) framework. In Stage 1, a teacher model is trained using only the complete agreement subset (D_C). The trained teacher predicts class probabilities for samples in the disagreement subset (D_D). Samples satisfying the confidence criterion are retained and assigned pseudo-labels, whereas low-confidence samples are excluded from student training. The retained pseudo-labeled samples are combined with D_C to construct the curated dataset ($D_{Curated}$). In Stage 2, a student model is initialized independently from ImageNet-pretrained weights and trained on $D_{Curated}$. The checkpoint achieving the highest macro F1-score on the complete agreement validation subset is selected, and only the selected student model is used for final inference.

After training, the teacher model is applied to each image $x_j \in D_D$. For each disagreement sample, the teacher outputs the class probability vector

$$\hat{\mathbf{p}}_j = \text{softmax}(f(x_j; \theta_T)) = [\hat{p}_{j,1}, \hat{p}_{j,2}, \dots, \hat{p}_{j,C}], \quad (2)$$

where C denotes the number of classes and $\hat{p}_{j,c}$ is the predicted probability for class c . The confidence score s_j and teacher-generated pseudo-label y_j^* are defined as $s_j = \max_c \hat{p}_{j,c}$ and $y_j^* = \text{argmax}_c \hat{p}_{j,c}$, respectively. Let τ denote the predefined confidence threshold used for confidence-based label curation. A disagreement sample is retained when its confidence score satisfies $s_j \geq \tau$. The retained pseudo-labeled subset is defined as

$$D_D^\tau = \{(x_j, y_j^*) \mid (x_j, \mathbf{a}_j) \in D_D, s_j \geq \tau\}. \quad (3)$$

The final curated dataset is then constructed as

$$D_{Curated} = D_C \cup D_D^\tau. \quad (4)$$

The original annotator labels are retained in the source dataset but are not used to train the student on the selected disagreement samples. Samples with $s_j < \tau$ are excluded only from student training; they are not removed from the original dataset.

2.3 Stage 2: Student Learning from the Curated Dataset

The student model $f(\cdot; \theta_s)$ is initialized independently from ImageNet-pretrained weights and optimized using the curated dataset $D_{Curated}$.

Because confidence-based sample selection may alter the class distribution, the class weights are recalculated for each curated training set using the same balanced inverse-frequency weighting defined above. The student objective is defined as

$$L_S = -\frac{1}{|D_{Curated}|} \sum_{(x_k, t_k) \in D_{Curated}} w_{t_k}^{Curated} \log P(t_k | x_k; \theta_s), \quad (5)$$

where θ_s denotes the student model parameters, t_k is the training target assigned to sample x_k , and $w_{t_k}^{Curated}$ denotes the inverse-frequency class weight computed from the curated training dataset.

For complete agreement samples, the original unanimous labels are used as training targets. For retained disagreement samples, the teacher-generated pseudo-labels are used.

The teacher is used only for offline sample selection and pseudo-label generation. No teacher parameters are transferred to the student, and no knowledge-distillation or teacher-student consistency loss is used. At inference, only the student model is required.

3 Experimental Setup

3.1 Dataset and Experimental Protocol

The proposed framework was evaluated using the publicly available GastroHUN dataset, which contains 8834 upper gastrointestinal endoscopic images collected from 387 patients following the SSS protocol [4,5]. The dataset contains 22 predefined gastric anatomical landmarks and one additional Not Applicable (NA) category. Detailed information regarding data acquisition, the annotation protocol, and the anatomical categories is available in the original GastroHUN publication [5].

Unlike conventional medical image datasets that provide a single reference label, GastroHUN was independently annotated by four endoscopists. Images for which all four annotators assigned the same label were categorized as complete agreement samples, whereas images with at least one differing annotation were categorized as disagreement samples. We used the official patient-level training, validation, and test partitions provided with the GastroHUN dataset [5]. The training partition contained 270 patients and 6165 images, the validation partition contained 58 patients and 1316 images, and the test partition contained 59 patients and 1353 images. The use of the publicly available, de-identified GastroHUN dataset for this study was reviewed and granted exemption from ethical review by the Institutional Review Board of Sungshin Women's University (Approval No. SSWUIRB-2026-048).

Across the entire dataset, 5318 images received complete agreement from all four annotators, whereas the remaining 3516 images contained annotation disagreement. In the training partition, the 3722 complete agreement images were used to train the teacher model, while the 2443 disagreement images were processed by the proposed confidence-based label curation procedure. Model selection and confidence-threshold selection were performed exclusively using the complete agreement validation subset. Final performance evaluation was conducted on the 803 complete agreement images in the independent test partition. The experimental data partitions are summarized in Table 1.

Table 1: Patient-level data partition and annotation-agreement subsets used in the experiments.

Dataset	Patients	Complete Agreement	Disagreement	Total
Training	270	3722	2443	6165
Validation	58	793	523	1316
Test	59	803	550	1353
Total	387	5318	3516	8834

3.2 Network Configurations and Implementation Details

Experiments were conducted using four representative backbone architectures: ConvNeXt-Tiny, ResNet50, Vision Transformer (ViT-B), and Swin-B [16–19]. All models were implemented using PyTorch 2.7.0 and torchvision 0.22.0. For all four backbones, the ImageNet-1K V1 pretrained weights distributed with torchvision were used for initialization [20]. The teacher and student models were initialized independently, and the original classification layer of each backbone was replaced with a randomly initialized linear layer with 23 outputs corresponding to the GastroHUN classes. The student did not inherit any parameters from the corresponding teacher. Experiments were performed on an NVIDIA GeForce RTX 5090 GPU using CUDA 12.8 and cuDNN 9.7.1.

All images were resized to 224×224 pixels using Lanczos interpolation with torchvision transforms. During training, color jittering was applied with brightness and contrast factors independently sampled from the range [0.9, 1.1], followed by conversion to a tensor and normalization using the dataset-specific mean ([0.5990, 0.3664, 0.2769]) and standard deviation ([0.2847, 0.2190, 0.1772]). Geometric transformations such as horizontal or vertical flipping, large rotations, and random cropping were not used because the classification labels correspond to specific anatomical locations and viewing orientations defined by the SSS protocol. Flipping or large rotations may alter clinically meaningful orientation cues, whereas random cropping may remove the global anatomical context needed to distinguish adjacent gastric landmarks. Therefore, only mild brightness and contrast augmentation was used to simulate illumination variability while preserving the anatomical information of the original images. For validation and test images, the same resizing, tensor conversion, and normalization were applied without color jittering.

Both the teacher and student models were optimized using AdamW with a batch size of 32, using $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1 \times 10^{-8}$, and a weight decay of 0.01 [21]. Training began with a 10-epoch head-only warm-up phase, during which the backbone was frozen and only the randomly initialized classification head was optimized using a learning rate of 0.001. No learning-rate scheduling was applied during this phase. Subsequently, all network parameters were unfrozen and fine-tuned for 100 epochs using an initial learning rate of 5×10^{-5} . During fine-tuning, a cosine annealing schedule with $T_{max} = 100$ and a minimum learning rate of 0 was applied [22]. The model checkpoint was updated whenever the macro F1-score on the complete agreement validation subset exceeded the previous best value, and the checkpoint with the highest validation macro F1-score was used for subsequent evaluation.

Four confidence thresholds, $\tau = 0.6, 0.7, 0.8, 0.9$, were evaluated. For each threshold, the validation macro F1-score was averaged across three independent runs for each backbone and subsequently across the four backbone architectures. The threshold achieving the highest overall mean validation macro F1-score was selected as the common operating threshold before evaluation on the independent test set. This procedure selected $\tau = 0.9$, with the corresponding validation results provided in Supplementary Table S1.

3.3 Baseline Construction and Comparison Settings

For each backbone architecture, a teacher model and a student model with the same architecture were used to construct an architecture-specific TS-CL pipeline. Specifically, the curated dataset used to train each student model was generated by the corresponding teacher model (e.g., a ConvNeXt-Tiny teacher generated the curated dataset for a ConvNeXt-Tiny student, and the same strategy was applied to ResNet50, ViT-B, and Swin-B).

For each backbone, the baseline and TS-CL student were trained using the same architecture, ImageNet initialization, preprocessing, optimization settings, and model-selection criteria. The baseline was trained using only the 3722 complete agreement images in the training partition, with their unanimous annotations used as training targets. In contrast, the TS-CL student was trained using the curated dataset, which combined the same complete agreement images with disagreement samples retained by the corresponding teacher model and assigned teacher-generated pseudo-labels. No disagreement samples were used to train the baseline models. Therefore, the primary difference between the baseline and TS-CL settings was whether confidence-based label curation was used to incorporate disagreement samples into student training.

3.4 Alternative Comparison Methods

To provide broader comparisons with the proposed TS-CL framework, four additional training strategies were evaluated using the same backbone architectures, preprocessing, data partitions, and evaluation protocol.

First, a one-stage variant was implemented to directly evaluate the effect of separating label curation from independent student optimization. For each backbone and run, the teacher model first generated the curated dataset using the same confidence threshold ($\tau = 0.9$) as the proposed TS-CL framework. Instead of independently initializing a student model, the same teacher model was then further fine-tuned on its curated dataset. Thus, the one-stage variant used the same confidence-based curation procedure as TS-CL but did not employ a separately initialized student model.

Second, a plurality-vote hard-label baseline was implemented to evaluate the use of disagreement samples through conventional label aggregation without teacher-based curation. Complete agreement samples retained their unanimous labels, whereas disagreement samples with a unique plurality vote were assigned the most frequently selected label among the four annotators. Samples with tied plurality votes were excluded from training [13].

Third, an annotator-specific multi-head model was implemented based on the concept of annotator-specific classification heads [10]. A shared backbone was connected to four classification heads corresponding to the four endoscopists. Each head was trained using the labels provided by its corresponding annotator, thereby preserving individual annotator supervision without aggregating the annotations into a single training target. The four annotator-specific cross-entropy losses were averaged to obtain the overall training objective. At inference, the softmax probability distributions from the four heads were averaged, and the class with the highest averaged probability was selected as the final prediction.

Finally, Co-teaching was implemented as a representative noise-robust learning baseline [23]. Because Co-teaching requires a single hard target for each image, complete agreement samples retained their unanimous labels, whereas disagreement samples were assigned the class receiving the unique plurality of the four annotator votes. Samples with tied plurality votes were excluded from training. Two peer networks with the same backbone architecture were independently initialized from ImageNet-pretrained weights. For each mini-batch, each peer computed the unweighted cross-entropy loss for each sample and selected the fraction of samples with the smallest losses according to the current remember rate. The selected samples

were exchanged between the peers, such that each network was updated using the samples selected by the other peer; class-weighted cross-entropy was used for these parameter updates. The forget rate was linearly increased from 0 to 0.20 during the first 10 epochs and was maintained at 0.20 thereafter. Co-teaching followed the same 10-epoch head-only warm-up and subsequent 100-epoch fine-tuning schedule described in [Section 3.2](#). At each fine-tuning epoch, both peers were evaluated on the complete agreement validation subset, and the single peer achieving the highest validation macro F1-score across epochs was retained for final evaluation. The ensemble of the two peer networks was not used.

3.5 Evaluation Protocol and Statistical Reporting

Classification performance was evaluated using accuracy (ACC), macro precision (PRE), macro sensitivity (SEN), macro F1-score (F1), Matthews correlation coefficient (MCC), and one-vs-rest macro- and weighted-average areas under the receiver operating characteristic curve (AUC). Multiclass AUC was computed across all 23 classes using the one-vs-rest approach, as implemented by `roc_auc_score` in `scikit-learn` 1.6.1. Macro AUC was calculated as the unweighted mean of class-wise AUC values, whereas weighted AUC was calculated by weighting the class-wise AUC values according to the number of test samples in each class. Weighted precision, weighted sensitivity, and weighted F1-score were additionally reported in the Supplementary Materials.

Each configuration was trained three times using random seeds 42, 420, and 1004. For each paired teacher-student run, the same seed value was independently applied to the teacher and student training procedures to control Python, NumPy, and PyTorch random-number generation, including model initialization, stochastic training augmentation, and training-data ordering. For TS-CL, each teacher run independently generated its own curated dataset, which was used only to train the independently initialized student from the corresponding run; curated datasets were not shared across runs. Results were reported as mean $\pm$ standard deviation, and the corresponding 95% confidence intervals were calculated using Student's t -distribution. Model selection was based on macro F1-score on the complete agreement validation subset, and final performance was evaluated on the independent complete agreement test subset consisting of 803 images.

4 Results

4.1 Performance of Confidence-Based Label Curation

Before evaluating the final classification performance, we first analyzed the outcome of the confidence-based label curation procedure. The teacher models were applied to the 2443 disagreement training images, and confidence-based label curation was performed using four confidence thresholds ($\tau = 0.6, 0.7, 0.8, 0.9$). [Fig. 2](#) illustrates the exclusion rates across confidence thresholds for the four teacher architectures, while the corresponding numbers and rates of retained and excluded samples are provided in Supplementary Table S2.

As shown in [Fig. 2](#), the exclusion rate increased consistently with the confidence threshold for all teacher architectures, reflecting progressively stricter selection of disagreement samples. At $\tau = 0.6$, the mean exclusion rates ranged from 6.43% for Swin-B to 14.24% for ResNet50. When the threshold was increased to $\tau = 0.9$, the exclusion rates increased to 37.67% for ConvNeXt-Tiny, 40.50% for ResNet50, 40.58% for ViT-B, and 24.11% for Swin-B. Accordingly, the proportion of disagreement samples retained for student training decreased as the confidence criterion became more restrictive.

The extent of sample exclusion also differed across teacher architectures. Swin-B consistently exhibited the lowest exclusion rate at all evaluated thresholds, retaining a larger proportion of disagreement samples than the other teacher architectures. In contrast, ResNet50 and ViT-B showed relatively higher exclusion

rates, particularly at higher confidence thresholds. These findings indicate that confidence-based label curation produced architecture-dependent curated training sets even when the same confidence threshold was applied to the same pool of disagreement samples.

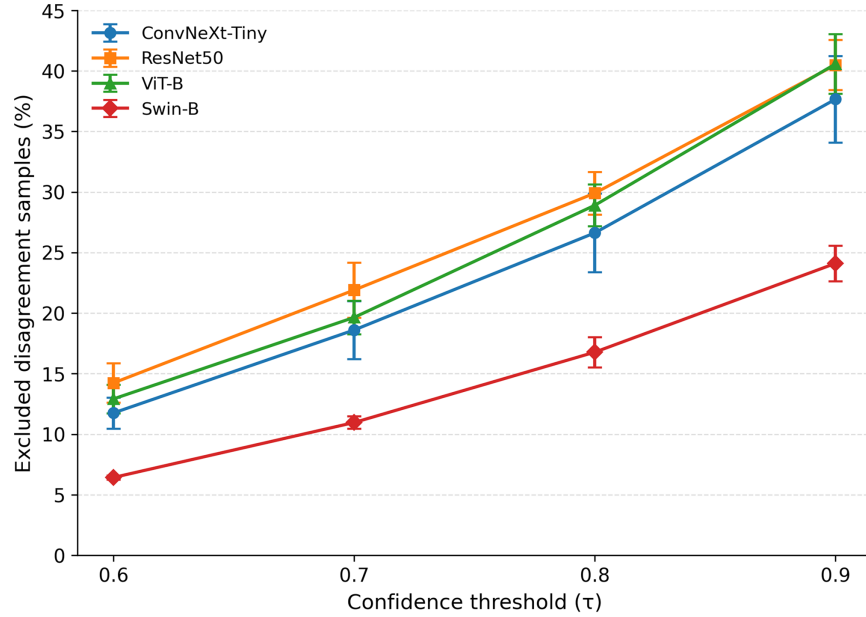


Figure 2: Effect of the confidence threshold (τ) on the exclusion of disagreement samples across four teacher architectures. The exclusion rate represents the proportion of disagreement samples that did not satisfy the teacher confidence criterion at each threshold. Points and error bars indicate the mean and standard deviation, respectively, over three independent runs. All experiments were performed using the same training disagreement pool of 2443 images.

To further characterize the low-confidence samples excluded by the proposed curation procedure, we compared annotator-vote patterns and normalized vote entropy between retained and excluded disagreement samples at the selected threshold ($\tau = 0.9$). Across all four teacher architectures, excluded samples exhibited higher mean normalized vote entropy than retained samples. In addition, excluded samples showed a lower proportion of 3-1 vote patterns and higher proportions of the other disagreement patterns. These findings indicate that confidence-based curation preferentially excluded samples exhibiting greater inter-observer disagreement. Detailed results are provided in Supplementary Table S3.

4.2 Quantitative Comparison with Baseline Models

To evaluate the effectiveness of the proposed TS-CL framework, the TS-CL models were compared with the corresponding complete agreement baselines. Table 2 summarizes the results at the selected confidence threshold ($\tau = 0.9$).

As shown in Table 2, the TS-CL models achieved higher mean classification performance than the corresponding complete agreement baselines across all evaluated backbone architectures. For ConvNeXt-Tiny, the macro F1-score increased from 83.28% to 85.00% (+1.72 percentage points), accompanied by improvements in accuracy from 83.73% to 85.39% (+1.66 percentage points) and MCC from 82.99% to 84.71% (+1.72 percentage points). ResNet50 showed the largest improvement in macro F1-score, increasing from 82.57% to 84.82% (+2.25 percentage points), while accuracy and MCC increased by 2.16 and 2.24 percentage points, respectively.

Table 2: Quantitative performance comparison between the complete agreement baseline and TS-CL at the selected confidence threshold ($\tau = 0.9$) on the independent complete agreement test set. Values are presented as mean $\pm$ standard deviation over three independent runs, with the corresponding 95% confidence intervals shown in parentheses.

Model Architecture	Configuration	ACC (%)	Macro PRE (%)	Macro SEN (%)	Macro F1 (%)	MCC (%)
ConvNeXt-Tiny	Baseline	83.73 $\pm$ 0.44	84.45 $\pm$ 0.54	83.17 $\pm$ 0.55	83.28 $\pm$ 0.61	82.99 $\pm$ 0.46
	(complete agreement)	(82.64–84.81)	(83.11–85.79)	(81.80–84.55)	(81.77–84.79)	(81.85–84.12)
	TS-CL ($\tau = 0.9$)	85.39 $\pm$ 0.29 (84.67–86.10)	85.79 $\pm$ 0.38 (84.85–86.73)	84.82 $\pm$ 0.38 (83.88–85.75)	85.00 $\pm$ 0.38 (84.06–85.93)	84.71 $\pm$ 0.30 (83.96–85.46)
ResNet50	Baseline	83.06 $\pm$ 1.68	83.37 $\pm$ 1.67	82.57 $\pm$ 1.72	82.57 $\pm$ 1.75	82.29 $\pm$ 1.75
	(complete agreement)	(78.90–87.23)	(79.22–87.51)	(78.29–86.84)	(78.22–86.91)	(77.95–86.63)
	TS-CL ($\tau = 0.9$)	85.22 $\pm$ 0.26 (84.58–85.87)	85.64 $\pm$ 0.23 (85.06–86.23)	84.55 $\pm$ 0.37 (83.63–85.48)	84.82 $\pm$ 0.30 (84.09–85.55)	84.53 $\pm$ 0.27 (83.85–85.21)
ViT-B	Baseline	84.77 $\pm$ 0.80	84.91 $\pm$ 0.92	83.99 $\pm$ 0.89	84.16 $\pm$ 0.87	84.06 $\pm$ 0.84
	(complete agreement)	(82.78–86.75)	(82.61–87.21)	(81.78–86.20)	(82.01–86.31)	(81.96–86.15)
	TS-CL ($\tau = 0.9$)	85.93 $\pm$ 0.45 (84.81–87.04)	86.12 $\pm$ 0.36 (85.23–87.01)	85.32 $\pm$ 0.53 (84.01–86.63)	85.39 $\pm$ 0.47 (84.23–86.55)	85.28 $\pm$ 0.47 (84.11–86.45)
Swin-B	Baseline	85.10 $\pm$ 0.63	85.80 $\pm$ 0.45	84.39 $\pm$ 0.87	84.56 $\pm$ 0.61	84.42 $\pm$ 0.66
	(complete agreement)	(83.54–86.65)	(84.68–86.92)	(82.23–86.56)	(83.04–86.08)	(82.77–86.07)
	TS-CL ($\tau = 0.9$)	86.38 $\pm$ 0.80 (84.40–88.37)	86.89 $\pm$ 0.76 (85.00–88.78)	85.61 $\pm$ 1.14 (82.76–88.45)	85.83 $\pm$ 1.06 (83.20–88.45)	85.76 $\pm$ 0.84 (83.68–87.84)

For the transformer-based models, ViT-B showed an increase in macro F1-score from 84.16% to 85.39% (+1.23 percentage points), together with improvements of 1.16 percentage points in accuracy and 1.22 percentage points in MCC. Swin-B also demonstrated consistently higher mean performance, with the macro F1-score increasing from 84.56% to 85.83% (+1.27 percentage points), accuracy from 85.10% to 86.38% (+1.28 percentage points), and MCC from 84.42% to 85.76% (+1.34 percentage points). Overall, at the selected confidence threshold of $\tau = 0.9$, TS-CL achieved higher mean accuracy, macro precision, macro sensitivity, macro F1-score, and MCC than the corresponding complete agreement baseline across all four backbone architectures.

The confusion matrices of the four TS-CL models at the selected confidence threshold ($\tau = 0.9$) are provided in Supplementary Fig. S1. In addition, the corresponding weighted precision, weighted sensitivity, and weighted F1-score for the baseline and TS-CL models are reported in Supplementary Table S4. Macro- and weighted-averaged AUC values were also comparable between the baseline and TS-CL models across all four architectures and are provided in Supplementary Table S5.

4.3 Effect of Confidence Threshold Selection

To investigate the influence of confidence threshold selection on the proposed TS-CL framework, we evaluated student models trained using four confidence thresholds ($\tau = 0.6, 0.7, 0.8, 0.9$). Fig. 3 summarizes the mean performance trends of five evaluation metrics across different confidence thresholds for each backbone architecture. The corresponding numerical results, including the baseline and all evaluated confidence thresholds with standard deviations and 95% confidence intervals, are provided in Supplementary Table S6.

For ConvNeXt-Tiny, the TS-CL models achieved higher mean values than the baseline across all five evaluation metrics at every confidence threshold. Performance generally improved at $\tau = 0.7$, decreased at $\tau = 0.8$, and increased again at $\tau = 0.9$. The highest mean accuracy, macro sensitivity, macro F1-score, and MCC were obtained at $\tau = 0.9$, whereas macro precision was marginally higher at $\tau = 0.7$. ResNet50 showed

a clearer threshold-dependent improvement, with performance generally increasing as the confidence threshold became more restrictive. Among the evaluated thresholds, $\tau = 0.9$ yielded the highest mean values for all five evaluation metrics. For ViT-B, relatively small changes were observed at $\tau = 0.6$ and $\tau = 0.7$, followed by a decrease at $\tau = 0.8$. Performance subsequently increased at $\tau = 0.9$, where all five evaluation metrics reached their highest mean values. Swin-B showed improved mean performance over the baseline across all evaluated thresholds. The highest macro sensitivity and macro F1-score were observed at $\tau = 0.7$, whereas accuracy, macro precision, and MCC reached their highest mean values at $\tau = 0.9$. The differences between these two thresholds were relatively small, indicating that Swin-B maintained strong performance across a broader range of confidence thresholds.

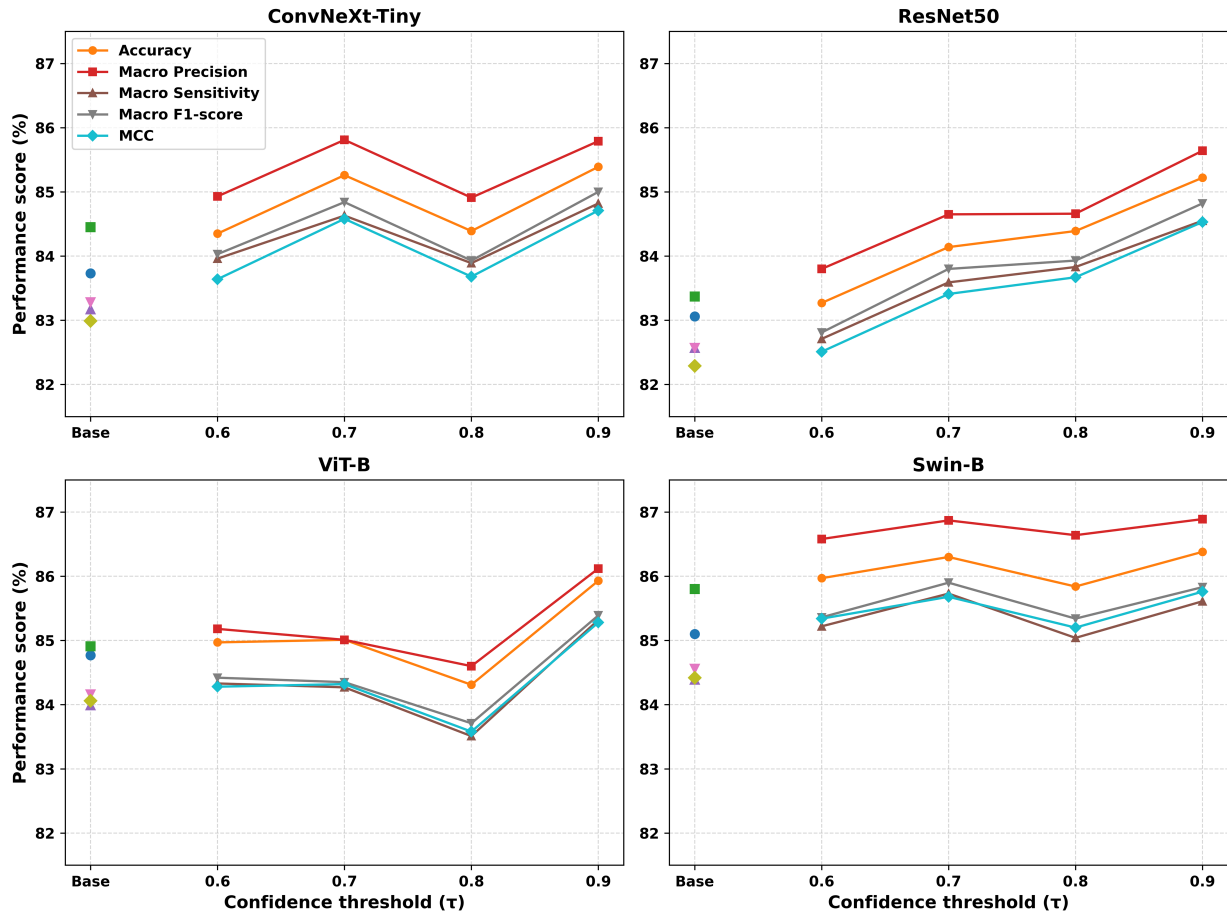


Figure 3: Effect of the confidence threshold (τ) on TS-CL student model performance across ConvNeXt-Tiny, ResNet50, ViT-B, and Swin-B. Performance was evaluated using accuracy, macro precision, macro sensitivity, macro F1-score, and MCC. The “Base” point represents the corresponding baseline model trained using only complete agreement training samples, whereas the remaining points represent TS-CL models trained at different confidence thresholds. Each point represents the mean performance over three independent runs.

Overall, the effect of confidence threshold selection varied across backbone architectures, and the test-set performance trends were not strictly monotonic. Nevertheless, $\tau = 0.9$, which had been selected using validation performance as described in [Section 3.2](#), yielded strong overall performance across the four architectures.

4.4 Comparison with Alternative Training and Multi-Annotator Learning Strategies

To further assess the effectiveness of TS-CL, we compared the proposed framework with several alternative strategies under the same evaluation protocol. These included a one-stage variant, in which the teacher model was further fine-tuned on its own curated dataset without independently initializing a student model; a plurality-vote hard-label baseline, in which disagreement samples with a unique plurality vote were assigned the most frequently selected annotator label [13]; Co-teaching, a representative noise-robust learning method [23]; and an annotator-specific multi-head model, in which a shared backbone was connected to four annotator-specific classification heads [10]. Table 3 summarizes the performance of these methods together with the complete agreement baseline and the proposed TS-CL using accuracy, macro F1-score, and MCC. All models were evaluated on the same independent complete agreement test set, and the results are reported as mean $\pm$ standard deviation over three independent runs.

Table 3: Comparison of the proposed TS-CL framework with alternative training and multi-annotator learning strategies. Values are presented as mean $\pm$ standard deviation over three independent runs.

Model Architecture	Configuration	ACC (%)	Macro F1 (%)	MCC (%)
ConvNeXt-Tiny	Baseline (complete agreement)	83.73 $\pm$ 0.44	83.28 $\pm$ 0.61	82.99 $\pm$ 0.46
	One-stage variant ($\tau = 0.9$)	84.72 $\pm$ 0.47	84.21 $\pm$ 0.46	84.03 $\pm$ 0.49
	Plurality-vote hard-label [13]	84.68 $\pm$ 0.69	84.25 $\pm$ 0.67	83.98 $\pm$ 0.73
	Co-teaching [23]	85.22 $\pm$ 0.73	84.83 $\pm$ 0.56	84.55 $\pm$ 0.77
	Annotator-specific multi-head [10]	85.31 $\pm$ 0.50	84.88 $\pm$ 0.49	84.63 $\pm$ 0.52
	TS-CL ($\tau = 0.9$)	85.39 $\pm$ 0.29	85.00 $\pm$ 0.38	84.71 $\pm$ 0.30
ResNet50	Baseline (complete agreement)	83.06 $\pm$ 1.68	82.57 $\pm$ 1.75	82.29 $\pm$ 1.75
	One-stage variant ($\tau = 0.9$)	84.23 $\pm$ 0.38	83.75 $\pm$ 0.35	83.50 $\pm$ 0.39
	Plurality-vote hard-label [13]	81.78 $\pm$ 1.99	81.02 $\pm$ 1.92	80.94 $\pm$ 2.08
	Co-teaching [23]	83.06 $\pm$ 1.71	82.43 $\pm$ 1.67	82.28 $\pm$ 1.79
	Annotator-specific multi-head [10]	83.56 $\pm$ 0.50	83.14 $\pm$ 0.56	82.81 $\pm$ 0.53
	TS-CL ($\tau = 0.9$)	85.22 $\pm$ 0.26	84.82 $\pm$ 0.30	84.53 $\pm$ 0.27
ViT-B	Baseline (complete agreement)	84.77 $\pm$ 0.80	84.16 $\pm$ 0.87	84.06 $\pm$ 0.84
	One-stage variant ($\tau = 0.9$)	84.97 $\pm$ 0.72	84.48 $\pm$ 0.81	84.28 $\pm$ 0.75
	Plurality-vote hard-label [13]	85.18 $\pm$ 0.37	84.65 $\pm$ 0.48	84.50 $\pm$ 0.39
	Co-teaching [23]	86.34 $\pm$ 0.63	85.83 $\pm$ 0.65	85.72 $\pm$ 0.66
	Annotator-specific multi-head [10]	85.55 $\pm$ 0.78	85.01 $\pm$ 0.83	84.89 $\pm$ 0.81
	TS-CL ($\tau = 0.9$)	85.93 $\pm$ 0.45	85.39 $\pm$ 0.47	85.28 $\pm$ 0.47
Swin-B	Baseline (complete agreement)	85.10 $\pm$ 0.63	84.56 $\pm$ 0.61	84.42 $\pm$ 0.66
	One-stage variant ($\tau = 0.9$)	85.72 $\pm$ 0.56	85.23 $\pm$ 0.54	85.07 $\pm$ 0.59
	Plurality-vote hard-label [13]	84.27 $\pm$ 1.74	83.62 $\pm$ 1.83	83.56 $\pm$ 1.81
	Co-teaching [23]	84.43 $\pm$ 0.94	84.00 $\pm$ 1.16	83.72 $\pm$ 0.98
	Annotator-specific multi-head [10]	84.93 $\pm$ 0.94	84.40 $\pm$ 0.92	84.25 $\pm$ 0.99
	TS-CL ($\tau = 0.9$)	86.38 $\pm$ 0.80	85.83 $\pm$ 1.06	85.76 $\pm$ 0.84

TS-CL achieved higher mean macro F1-scores than the one-stage variant across all four backbone architectures: 85.00% vs. 84.21% for ConvNeXt-Tiny, 84.82% vs. 83.75% for ResNet50, 85.39% vs. 84.48%

for ViT-B, and 85.83% vs. 85.23% for Swin-B. Similar mean differences were observed for accuracy and MCC. TS-CL also achieved higher mean macro F1-scores than the plurality-vote hard-label baseline across all four architectures, with differences of 0.75, 3.80, 0.74, and 2.21 percentage points for ConvNeXt-Tiny, ResNet50, ViT-B, and Swin-B, respectively. This comparison suggests that the observed mean performance differences cannot be explained solely by incorporating disagreement samples into training, as the two approaches differed in how disagreement samples were selected and assigned training targets. Compared with the annotator-specific multi-head model, TS-CL achieved higher mean macro F1-scores across all four backbone architectures, although the difference was relatively small for ConvNeXt-Tiny. TS-CL also achieved higher mean macro F1-scores than Co-teaching for ConvNeXt-Tiny, ResNet50, and Swin-B, whereas Co-teaching achieved a higher value for ViT-B (85.83% vs. 85.39%). Overall, TS-CL achieved the highest mean macro F1-score among the evaluated methods for three of the four backbone architectures.

For additional comparison, we considered the image-classification results originally reported by Bravo et al. for the GastroHUN dataset [5]. Under their complete agreement setting, ConvNeXt-Tiny achieved a macro precision of 87.58%, macro sensitivity of 86.92%, and macro F1-score of 86.79%. These values were higher than the corresponding ConvNeXt-Tiny results obtained with TS-CL in the present study. However, the published results are not directly comparable with ours because the training and evaluation protocols differed. Bravo et al. used Adam optimization, fine-tuned only the final 40% of the feature layers after warm-up, and employed a different learning-rate scheduling strategy, whereas our models were fine-tuned end-to-end using AdamW with a cosine annealing learning-rate schedule. In addition, Bravo et al. estimated performance variability using 100 bootstrap resamples of the test set, whereas our results were obtained from three independent training runs with different random seeds.

A direct numerical comparison with the other backbone architectures was not possible from the published results. Although Bravo et al. evaluated multiple ResNet and ViT variants, detailed numerical results were reported only for the five highest-performing models, and Swin-B was not included in their experiments. Therefore, the published ConvNeXt-Tiny result is presented as a reference benchmark rather than as a directly controlled comparison with TS-CL.

5 Discussion

The present study addressed one of the fundamental challenges in medical image analysis: how to utilize multi-annotator datasets containing substantial inter-observer disagreement. Instead of discarding disagreement samples or directly using their heterogeneous annotations as training targets, the proposed TS-CL framework leveraged a teacher trained on complete agreement samples to identify disagreement samples for which high-confidence pseudo-labels can be assigned. Across four representative CNN- and transformer-based architectures and three independent runs, confidence-based label curation resulted in higher mean classification performance on the complete agreement test set compared with the corresponding baseline models trained using only complete agreement samples. The improvements in macro F1-score were 1.72, 2.25, 1.23, and 1.27 percentage points for ConvNeXt-Tiny, ResNet50, ViT-B, and Swin-B, respectively. These findings suggest that, for the GastroHUN dataset, carefully curated pseudo-labels can provide useful additional supervision by incorporating informative disagreement samples that would otherwise remain unused. Unlike conventional label-noise learning methods that primarily attempt to increase robustness during optimization, TS-CL explicitly curates the training labels before student learning. This separation between label curation and model optimization enables the student model to learn from a more consistent supervision signal while retaining a substantial proportion of disagreement samples.

Although the absolute performance gains were moderate, the practical value of TS-CL lies in its ability to reuse disagreement samples that are already available in a multi-annotator dataset without requiring

additional expert relabeling. The proposed method should not be interpreted as reducing the initial annotation burden because the GastroHUN images had already been annotated by four experts. Rather, it reduces the need for a subsequent manual review of disagreement samples. Moreover, label curation is performed offline, and the teacher is not required after the curated dataset has been constructed. Only a single student model is used at inference, resulting in no additional inference-time model complexity compared with the complete agreement baseline. However, TS-CL requires the additional training of a teacher model and therefore should not be considered more computationally efficient during model development.

One of the major findings of this study is that the confidence threshold substantially influences the effectiveness of the proposed label curation strategy. When the threshold is relatively low (e.g., $\tau = 0.6$), the teacher accepts a larger proportion of disagreement samples. Although this maximizes data utilization, it may also increase the probability that potentially incorrect pseudo-labels are introduced into the curated dataset, thereby reducing the overall label quality. Conversely, a more conservative threshold (e.g., $\tau = 0.9$) excludes a larger proportion of uncertain samples, potentially including images that contain useful anatomical information. Excessive filtering may therefore decrease the diversity of the training data and limit the student's ability to learn representative decision boundaries. These observations suggest that confidence threshold selection reflects a balance between label reliability and data utilization, a trade-off also relevant to confidence-based pseudo-labeling methods [24].

In the present experiments, however, the relationship between the confidence threshold and classification performance was architecture-dependent and not strictly monotonic. Increasing τ progressively increased the proportion of excluded disagreement samples for all teacher architectures. At $\tau = 0.9$, the mean exclusion rates were 37.67% for ConvNeXt-Tiny, 40.50% for ResNet50, 40.58% for ViT-B, and 24.11% for Swin-B. Importantly, the same numerical threshold did not result in the same degree of sample filtering across teacher architectures. Swin-B consistently retained a larger proportion of disagreement samples, whereas ResNet50 and ViT-B applied more restrictive selection at higher thresholds. This observation suggests that the distributions of maximum softmax probabilities differ across architectures and that a fixed threshold does not necessarily represent an equivalent level of pseudo-label reliability for every teacher model. The additional analysis of excluded samples further showed that low-confidence samples exhibited higher normalized vote entropy and more dispersed annotator-vote patterns than retained samples. This finding suggests that teacher confidence was related, at least in part, to the degree of inter-observer disagreement present in the original annotations.

The performance trends further support this architecture-dependent interpretation. ResNet50 and ViT-B achieved their highest mean values for all five primary metrics at $\tau = 0.9$. ConvNeXt-Tiny also obtained its highest mean accuracy, macro sensitivity, macro F1-score, and MCC at $\tau = 0.9$, although its macro precision was marginally higher at $\tau = 0.7$. For Swin-B, accuracy, macro precision, and MCC were highest at $\tau = 0.9$, whereas macro sensitivity and macro F1-score were slightly higher at $\tau = 0.7$. Although the optimal threshold varied across individual metrics and backbones, the validation-based selection procedure identified $\tau = 0.9$ as the common operating threshold. The subsequent test-set results showed strong overall performance at this threshold, while confirming that no single threshold was uniformly optimal for every metric and backbone.

Among the evaluated architectures, ResNet50 exhibited the largest improvement in macro F1-score at the selected threshold, increasing from 82.57% to 84.82%. ConvNeXt-Tiny, ViT-B, and Swin-B also benefited from confidence-based label curation, but the magnitude of improvement differed across architectures. The results do not support a simple distinction between convolutional and transformer-based networks. Instead, they suggest that the benefit of label curation may depend on the interaction among teacher confidence characteristics, pseudo-label distribution, and the learning characteristics of the student model.

In particular, Swin-B maintained relatively strong performance across a broad range of thresholds while retaining substantially more disagreement samples than the other teacher architectures.

Existing approaches for learning from unreliable annotations generally follow two major strategies. The first focuses on noise-robust optimization, such as robust loss functions, sample reweighting, or Co-teaching, where the optimization process attempts to reduce the influence of potentially unreliable samples during training [23,25,26]. The second utilizes soft-label learning, in which multiple annotator opinions are aggregated into probabilistic targets to explicitly model annotation uncertainty [11,12]. A third line of research attempts to improve label quality before model optimization through explicit label correction or refinement. For example, meta-learning-based frameworks have been proposed to identify and correct unreliable annotations prior to network training, thereby reducing the adverse effects of inconsistent supervision [27]. The proposed TS-CL framework differs from these paradigms. Rather than modifying the optimization objective, modeling annotation uncertainty, or iteratively correcting labels during optimization, TS-CL performs a simple confidence-based label curation before student training. The teacher model identifies high-confidence disagreement samples to receive deterministic pseudo-labels while excluding low-confidence samples from student training. Consequently, the student network is trained on a curated dataset with improved label consistency instead of directly confronting conflicting supervision throughout optimization. This explicit separation between label curation and representation learning offers a practical alternative for multi-annotator medical datasets, particularly when a reliable complete agreement subset is available for teacher training.

The additional experimental comparisons further characterize the performance of the proposed framework relative to alternative strategies for utilizing disagreement samples. Across all four backbones, TS-CL showed higher mean macro F1-scores than the corresponding one-stage variant. TS-CL also showed higher mean macro F1-scores than the plurality-vote hard-label baseline and the annotator-specific multi-head model across all four backbones. Compared with Co-teaching, TS-CL showed higher mean macro F1-scores for ConvNeXt-Tiny, ResNet50, and Swin-B, whereas Co-teaching showed a higher mean value for ViT-B. However, these comparisons should be interpreted descriptively because only three independent runs were performed and no formal statistical comparisons between methods were conducted. Thus, the results suggest that the strategy used to incorporate disagreement samples and the backbone architecture may influence performance, but they do not establish statistical superiority of TS-CL over the alternative methods.

For additional context, we considered the image-classification results originally reported with the GastroHUN dataset [5]. Bravo et al. reported a macro F1-score of 86.79% for ConvNeXt-Tiny trained on complete agreement samples, which was higher than the 85.00% obtained by the corresponding TS-CL model in the present study. However, this result should be regarded as a reference benchmark rather than a directly controlled comparison. The original study used Adam optimization and fine-tuned only the final 40% of the feature layers after warm-up, whereas the present study used AdamW, end-to-end fine-tuning, and a cosine annealing learning-rate schedule. In addition, Bravo et al. estimated variability using 100 bootstrap resamples of the test set, whereas our results were obtained from three independently trained models with different random seeds. Detailed numerical results corresponding exactly to our ResNet50 and ViT-B configurations were not reported in the published performance tables, and Swin-B was not included in the original experiments. Therefore, a direct architecture-wise comparison with the published study was possible only for ConvNeXt-Tiny.

This study has several limitations. First, all experiments were conducted exclusively on the GastroHUN dataset [5]. Although the dataset represents a large publicly available multi-annotator gastric endoscopy benchmark, the generalizability of the proposed framework to other medical imaging modalities remains to be validated. Second, the final evaluation was conducted on the complete agreement test

subset, which provides a reliable reference standard but does not directly evaluate model performance on test samples with inter-observer disagreement. Therefore, the current results show that utilizing disagreement samples during training was associated with higher mean classification performance on high-consensus test samples, rather than directly demonstrating improved performance on ambiguous test cases. Third, the confidence threshold was selected based on validation performance from a limited set of predefined values. Although the validation-based procedure selected $\tau = 0.9$ as the common operating threshold, test-set performance remained architecture- and metric-dependent. Therefore, the optimal threshold may depend on dataset characteristics, annotation quality, and model architecture. Fourth, the proposed pseudo-labeling strategy relies solely on the teacher's prediction confidence without explicitly estimating predictive uncertainty or model calibration. Consequently, some confidently incorrect pseudo-labels may still be introduced into the curated dataset. Finally, external validation using independent institutions and different endoscopy systems was not performed, limiting conclusions regarding real-world clinical deployment.

Several directions may further improve the proposed framework. Instead of using a fixed confidence threshold, future studies could investigate adaptive, architecture-specific, or class-specific thresholding that dynamically adjusts the acceptance criterion according to model confidence, class imbalance, or prediction uncertainty [28]. Integrating uncertainty estimation techniques, such as Monte Carlo Dropout or deep ensembles, may also improve pseudo-label reliability by distinguishing epistemic uncertainty from prediction confidence [29,30]. Uncertainty-aware pseudo-label selection strategies may further help identify reliable pseudo-labels [14]. Another promising direction is calibration-aware pseudo-labeling, in which probability calibration techniques are incorporated before confidence filtering [31]. Such an approach may improve the reliability of confidence-based sample selection and reduce the influence of confidently incorrect pseudo-labels. Finally, validating the framework across multiple multi-annotator datasets from different imaging modalities—including colonoscopy, pathology, retinal imaging, and radiology—would help establish the general applicability of confidence-based label curation for medical image analysis.

6 Conclusion

This study proposed a two-stage confidence-guided label curation (TS-CL) framework to utilize multi-annotator medical image datasets containing substantial inter-observer disagreement. By training a teacher model on complete agreement samples and applying confidence-based label curation to disagreement samples, the proposed framework constructed a curated training dataset for subsequent student learning.

Experiments on the GastroHUN dataset demonstrated that TS-CL achieved higher mean classification performance across four representative deep learning architectures compared with baseline models trained using only complete agreement samples. These results indicate that confidence-based label curation can leverage selected disagreement samples as additional supervision. Additional comparisons showed higher mean macro F1-scores for TS-CL than the one-stage variant across all four backbones and competitive mean performance relative to other strategies for utilizing disagreement samples.

Overall, TS-CL provides a simple and practical strategy for improving supervised learning from multi-annotator medical image datasets without requiring additional expert relabeling or additional inference-time model complexity. Future work will investigate adaptive confidence thresholding, uncertainty-aware pseudo-label selection, and validation on additional multi-annotator medical imaging datasets.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the Sungshin Women's University Research Grant of 2026.

Availability of Data and Materials: This study used the publicly available GastroHUN dataset. The data are openly available at <https://doi.org/10.6084/m9.figshare.27308133>.

Ethics Approval: This study used only the publicly available GastroHUN dataset, which contains de-identified data. The study was reviewed and granted an exemption from ethical review by the Institutional Review Board of Sungshin Women's University (Approval No. SSWUIRB-2026-048). No participants were recruited, and no identifiable personal information was collected or accessed by the authors. The requirement to obtain informed consent was formally waived by the Institutional Review Board of Sungshin Women's University.

Conflicts of Interest: The authors declare no conflicts of interest.

Supplementary Materials: The supplementary material is available online at <https://www.techscience.com/doi/10.32604/cmes.2026.088580/s1>. Fig. S1: Confusion matrices of the TS-CL models using the selected confidence threshold ($\tau = 0.9$); Table S1: Validation macro F1-scores (%) across confidence thresholds used for selection of the common operating threshold; Table S2: Distribution of retained pseudo-labeled and excluded disagreement samples across confidence thresholds for four teacher architectures; Table S3: Characteristics of retained and excluded disagreement samples at the selected confidence threshold ($\tau = 0.9$); Table S4: Weighted classification performance of the baseline and TS-CL models ($\tau = 0.9$) on the independent complete agreement test set; Table S5: Macro-averaged and weighted-averaged AUC performance of the complete agreement baseline and TS-CL models evaluated on the independent complete agreement test set using the selected confidence threshold ($\tau = 0.9$); Table S6: Detailed classification performance of the baseline and TS-CL models across different confidence thresholds ($\tau = 0.6, 0.7, 0.8, 0.9$) for four backbone architectures.

Abbreviations

The following abbreviations are used in this manuscript:

ACC	Accuracy
AI	Artificial Intelligence
AUC	Area under the Receiver Operating Characteristic Curve
CI	Confidence Interval
CNN	Convolutional Neural Network
EGC	Early Gastric Cancer
EGD	Esophagogastroduodenoscopy
F1	F1-Score
GI	Gastrointestinal
MCC	Matthews Correlation Coefficient
NA	Not Applicable
PRE	Precision
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SEN	Sensitivity
SSS	Systematic Screening Protocol for the Stomach
Swin-B	Swin Transformer-Base
TS-CL	Two-Stage Confidence-Guided Label Curation
ViT	Vision Transformer
ViT-B	Vision Transformer-Base

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;71(3):209–49. doi:10.3322/caac.21660.
2. Jun JK, Choi KS, Lee HY, Suh M, Park B, Song SH, et al. Effectiveness of the Korean national cancer screening program in reducing gastric cancer mortality. *Gastroenterology.* 2017;152(6):1319–28.e7. doi:10.1053/j.gastro.2017.01.029.
3. Kaise M. Advanced endoscopic imaging for early gastric cancer. *Best Pract Res Clin Gastroenterol.* 2015;29(4):575–87. doi:10.1016/j.bpg.2015.05.010.
4. Yao K. The endoscopic diagnosis of early gastric cancer. *Ann Gastroenterol.* 2013;26(1):11–22.
5. Bravo D, Frias J, Vera F, Trejos J, Martínez C, Gómez M, et al. GastroHUN an endoscopy dataset of complete systematic screening protocol for the stomach. *Sci Data.* 2025;12(1):102. doi:10.1038/s41597-025-04401-5.
6. Tanno R, Saeedi A, Sankaranarayanan S, Alexander DC, Silberman N. Learning from noisy labels by regularized estimation of annotator confusion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA.* p. 11236–45. doi:10.1109/CVPR.2019.01150.
7. Zhang L, Tanno R, Xu M, Huang Y, Bronik K, Jin C, et al. Learning from multiple annotators for medical image segmentation. *Pattern Recognit.* 2023;138:109400. doi:10.1016/j.patcog.2023.109400.
8. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Med Image Anal.* 2020;65(5):101759. doi:10.1016/j.media.2020.101759.
9. Cheplygina V, Pluim JP. Crowd disagreement about medical images is informative. In: *Proceedings of the International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS); 2018 Sep 16; Granada, Spain.* p. 105–11. doi:10.1007/978-3-030-01364-6_12.
10. Wang X, Plank B. ACTOR: active learning with annotator-specific classification heads to embrace human label variation. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2023 Dec 6–10; Singapore.* p. 2046–52. doi:10.18653/v1/2023.emnlp-main.126.
11. Wright D, Augenstein I. Aggregating soft labels from crowd annotations improves uncertainty estimation under distribution shift. *PLoS One.* 2025;20(6):e0323064. doi:10.1371/journal.pone.0323064.
12. Wu B, Li Y, Mu Y, Scarton C, Bontcheva K, Song X. Don't waste a single annotation: improving single-label classifiers through soft labels. In: *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023; 2023 Dec 6–10; Singapore.* p. 5347–55. doi:10.18653/v1/2023.findings-emnlp.355.
13. Uma AN, Fornaciari T, Hovy D, Paun S, Plank B, Poesio M. Learning from disagreement: a survey. *J Artif Intell Res.* 2021;72:1385–470. doi:10.1613/jair.1.12752.
14. Rizve MN, Duarte K, Rawat YS, Shah M. In defense of pseudo-labeling: an uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv:2101.06329.* 2021. doi:10.48550/arXiv.2101.06329.
15. Northcutt C, Jiang L, Chuang I. Confident learning: estimating uncertainty in dataset labels. *J Artif Intell Res.* 2021;70:1373–411. doi:10.1613/jair.1.12125.
16. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 19–24; New Orleans, LA, USA.* p. 11966–76. doi:10.1109/CVPR52688.2022.01167.
17. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA.* p. 770–8. doi:10.1109/CVPR.2016.90.
18. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv:2010.11929.* 2020. doi:10.48550/arXiv.2010.11929.
19. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 11–17; Montreal, QC, Canada.* p. 9992–10002. doi:10.1109/ICCV48922.2021.00986.

20. Deng J, Dong W, Socher R, Li LJ, Kai L, Li FF. ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2009 Jun 20–25; Miami, FL, USA. p. 248–55. doi:10.1109/CVPR.2009.5206848.
21. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv:1711.05101. 2017. doi:10.48550/arXiv.1711.05101.
22. Loshchilov I, Hutter F. SGDR: stochastic gradient descent with warm restarts. arXiv:1608.03983. 2016. doi:10.48550/arXiv.1608.03983.
23. Han B, Yao Q, Yu X, Niu G, Xu M, Hu W, et al. Co-teaching: robust training of deep neural networks with extremely noisy labels. Adv Neural Inf Process Syst. 2018;31:8536–46.
24. Sohn K, Berthelot D, Carlini N, Zhang Z, Zhang H, Raffel C, et al. FixMatch: simplifying semi-supervised learning with consistency and confidence. Adv Neural Inf Process Syst. 2020;33:596–608.
25. Zhang Z, Sabuncu M. Generalized cross entropy loss for training deep neural networks with noisy labels. Adv Neural Inf Process Syst. 2018;31(8):8792–802. doi:10.1007/978-3-030-88013-2_7.
26. Ren M, Zeng W, Yang B, Urtasun R. Learning to reweight examples for robust deep learning. In: Proceedings of the 35th International Conference on Machine Learning (ICML); 2018 Jul 10–15; Stockholm, Sweden. p. 4334–43.
27. Jiang K, Cao B, Fan J. A robust framework for multimodal sentiment analysis with noisy labels generated from distributed data annotation. Comput Model Eng Sci. 2024;139(3):2965–84. doi:10.32604/cmcs.2023.046348.
28. Zoha S, Lee JG, Ko YW. CAT: class-aware adaptive-thresholding for robust semi-supervised domain generalization. PLoS One. 2025;20(9):e0329799. doi:10.1371/journal.pone.0329799.
29. Nguyen-Trang T, Ha-Hoang H. A stochastic ensemble physics-informed neural networks via bagging and Monte Carlo dropout. Comput Model Eng Sci. 2026;147(2):33. doi:10.32604/cmcs.2026.080808.
30. Yoon T, Kang D. Efficient ensemble learning with curriculum-based masked autoencoders for retinal OCT classification. Diagnostics. 2026;16(2):179. doi:10.3390/diagnostics16020179.
31. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML); 2017 Aug 6–11; Sydney, Australia. p. 1321–30.