Leveraging Model Soups to Classify Intangible Cultural Heritage Images from the Mekong Delta

Quoc-Khang Tran[1], Minh-Thien Nguyen[1], Nguyen-Khang Pham[2]
[1] College of Information and Communication Technology, Can Tho University
[2] School of Graduate, Can Tho University
Correspondence: Nguyen-Khang Pham, pnkhang@ctu.edu.vn
Communication: received 30 June 2025, revised 28 August 2025, accepted 01 October 2025
Digital Object Identifier: https://doi.org/10.32913/mic-ict-research.v2025.n3.1395

Abstract: The classification of Intangible Cultural Heritage (ICH) images in the Mekong Delta poses unique challenges due to limited annotated data, high visual similarity among classes, and domain heterogeneity. In such low-resource settings, conventional deep learning models often suffer from high variance or overfit to spurious correlations, leading to poor generalization. To address these limitations, we propose a robust framework that integrates the hybrid CoAtNet architecture with *model soups*, a lightweight weight-space ensembling technique that averages checkpoints from a single training trajectory – *without increasing inference cost*. CoAtNet captures both local and global patterns through stage-wise fusion of convolution and self-attention. We apply two ensembling strategies – *greedy* and *uniform* soup – to selectively combine diverse checkpoints into a final model. Beyond performance improvements, we analyze the ensembling effect through the lens of bias–variance decomposition. Our findings show that *model soups* reduces variance by stabilizing predictions across diverse model snapshots, while introducing minimal additional bias. Furthermore, using cross-entropy-based distance metrics and Multidimensional Scaling (MDS), we show that *model soups* selects geometrically diverse checkpoints, unlike Soft Voting, which blends redundant models centered in output space. Evaluated on the ICH-17 dataset (7,406 images across 17 classes), our approach achieves state-of-the-art results with 72.36% top-1 accuracy and 69.28% macro F1-score, outperforming strong baselines including ResNet-50, DenseNet-121, and ViT. These results underscore that diversity-aware checkpoint averaging provides a principled and efficient way to reduce variance and enhance generalization in culturally rich, data-scarce classification tasks.

Keywords: *ICH Classification, Model Soups, CoAtNet, Hybrid Model, Ensemble Learning, Multidimensional Scaling.*

Class 4 – Lễ hội cúng biển Mỹ Long
(Mỹ Long Sea Worship Festival)

Class 8 - Đại lễ Kỳ yên Đình Tân Phước Tây
(Tân Phước Tây Temple Kỳ Yên Ceremony)

Figure 1. An illustrative example of the challenging nature of this task: class 4 (left) and class 8 (right) share highly similar visual contexts, making them difficult to distinguish.

I. Introduction

The classification of Intangible Cultural Heritage (ICH) images in the Mekong Delta plays a crucial role in cultural preservation, documentation, and digital dissemination. However, building accurate image classifiers for ICH is inherently challenging due to the diversity of cultural practices, subtle inter-class variations, and the limited scale of high-quality annotated data. To address these issues, we propose a novel classification framework based on the CoAtNet architecture and model ensembling via *model soups* [14]. CoAtNet is a hybrid vision model that combines convolutional operations with self-attention mechanisms, effectively capturing both local patterns and long-range de-

pendencies in images [2]. Its flexibility and representational power make it a suitable backbone for our task.

Rather than training multiple independent models with different seeds or configurations, we generate an ensemble by collecting multiple checkpoints from a single CoAtNet training process. These checkpoints, representing different stages of model optimization, are aggregated through the *model soups* technique — a method that averages the weights of multiple checkpoints without requiring additional joint training [14]. This simple yet effective strategy enhances generalization by interpolating between distinct solutions in the weight space, thereby reducing overfitting and leveraging the complementary strengths of intermediate model states.

Our experimental results on a curated 17-class ICH dataset from the Mekong Delta show that the proposed framework significantly outperforms standalone fine-tuned models, highlighting its potential as a practical and scalable solution for heritage-related image classification.

II. Related Work

The classification of intangible cultural heritage (ICH) imagery has received increasing attention with the advancement of deep learning techniques in cultural informatics. In early studies, conventional convolutional neural networks (CNNs) and traditional machine learning approaches were employed to recognize ICH categories in Vietnam. Notably, Do et al. [4] introduced a 17-class ICH image dataset from the Mekong Delta and benchmarked various feature extractors such as VGG19, ResNet50, Inception-v3, and Xception in combination with support vector machines (SVMs), attaining an accuracy of 65.32%. Subsequently, Tran et al. [12] improved upon this result by fusing deep features and classifier outputs through logistic regression, achieving 66.76% accuracy. Despite these efforts, model performance remained modest, and little attention was paid to ensemble strategies or architectural advancements.

Ensemble learning has long been recognized as an effective strategy for improving generalization in classification tasks. A recent and notable contribution is *model soups* [14], a method that combines the weights of independently fine-tuned models to produce a single, more robust model without requiring additional inference time. This approach leverages the diversity of multiple checkpoints of a model to create a solution that generalizes better than any single model alone.

At the architectural level, hybrid models have emerged as strong alternatives to traditional CNNs. Among them, CoAtNet [2] integrates convolutional operations with self-attention mechanisms, offering a unified design that balances local feature extraction with global context modeling. CoAtNet has demonstrated state-of-the-art performance across a wide range of vision benchmarks and is particularly well-suited for tasks involving limited annotated data, such as ICH image classification.

In this work, we build upon these recent advances by applying the *model soups* technique to an ensemble of CoAtNet models for multi-class ICH classification. To the best of our knowledge, this is the first study that leverages CoAtNet-based *model soups* to improve classification performance on cultural heritage datasets.

III. Research Methodology

1. Classification of Intangible Cultural Heritage Images

The Mekong Delta is renowned for its rich repository of Intangible Cultural Heritage (ICH), encompassing a diverse array of cultural expressions. These include traditional practices such as *Đờn ca Tài tử Nam Bộ (Art of Đờn ca tài tử music and song in southern Vietnam)* and *Văn hoá Chợ nổi Cái Răng (Cái Răng Floating Market Culture)*, traditional handicrafts like *Nghề dệt chiếu (Mat weaving)* and *Nghề đan tre (Bamboo weaving)*, as well as traditional festivals, including *Lễ hội Ok Om Bok của người Khmer (Khmer Ok Om Bok festival)* and *Lễ hội Vía Bà Chúa Xứ Núi Sam (Festival of Bà Chúa Xứ Goddess at Sam Mountain)*. These intangible cultural heritages play a pivotal role in preserving Vietnam's cultural values and enhancing the diversity of intangible cultural heritage in the Mekong Delta region.

In this study, we focus on the classification of images of intangible cultural heritage. Specifically, our goal is to improve classification performance on a dataset comprising images from 17 ICH categories (referred to as the ICH-17 dataset). Do et al. [4] constructed this dataset by collecting images via Google search using text-based queries, followed by manual post-processing to remove noisy or irrelevant images. Detailed information about the dataset is provided in Table I.

Despite manual post-processing, the ICH-17 dataset still contains a notable proportion of noisy, irrelevant images . Additionally, the images exhibit significant contextual diversity, and certain categories share relatively similar contexts, such as *Lễ hội cúng biển Mỹ Long (Mỹ Long Sea Worship Festival)* (class 4) and *Đại lễ Kỳ yên Đình Tân Phước Tây (Tân Phước Tây Temple Kỳ Yên Ceremony)* (class 8) (see Fig. 1). These characteristics of the ICH-17 dataset pose significant challenges to achieving high classification performance, as evidenced in prior studies [4, 5, 12].

2. CoAtNet: Hybrid Convolution-Attention Design

CoAtNet [2] is adopted in this study as a hybrid neural architecture unifying convolutional operations with self-attention in a stage-wise design to capitalize on the strengths of both paradigms. Specifically, the network consists of five stages, where the initial stage (S_0) is a convolutional stem composed of standard convolutional layers, while S_1 and S_2 employ MBConv blocks [11] with depthwise separable convolutions and squeeze-and-excitation mechanisms [8] to efficiently capture local features. The later stages (S_3 and S_4) transition to Transformer blocks [13] incorporating relative self-attention and convolutional feed-forward networks to model global dependencies. This C–C–T–T configuration enables CoAtNet to maintain strong inductive biases for spatial generalization while enhancing its capacity to learn long-range interactions.

Comparative experiments were conducted against three representative baselines: **ResNet-50** [7], a classic CNN architecture featuring residual connections that facilitate effective deep learning; **DenseNet-121** [9], which enhances information flow through dense, inter-layer connectivity; and **Vision Transformer (ViT)** [6], a purely attention-based model that processes image patches directly and operates without convolutional priors. These comparisons assess the architecture's performance in intangible cultural heritage classification.

3. Weight-Space Ensembling via Model Soups

The *model soups* technique [14] is adopted to enhance generalization without increasing inference cost. A greedy soup procedure is employed: the checkpoint achieving the highest validation accuracy is selected first, and additional candidates are incorporated only if they yield improved validation performance.

Let $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}\}$ denote the set of saved checkpoints at different epochs. The greedy selection identifies a subset $\mathcal{S} \subseteq \{1, \ldots, T\}$ of checkpoint indices. The final ensemble model, referred to as a *uniform soup*, is then computed as:

$$\theta_{\text{soup}} = \frac{1}{|\mathcal{S}|} \sum_{k \in \mathcal{S}} \theta^{(k)}, \tag{1}$$

where $\theta^{(k)}$ denotes the parameter vector of the k-th selected checkpoint. By integrating greedy selection with uniform averaging, the method filters out suboptimal checkpoints while retaining the simplicity and generalization benefits of weight-space ensembling [13, 14]. The detailed procedure is outlined in Algorithm 1.

TABLE I
DETAILS FOR THE ICH-17 DATASET

No.	Category	# Images
1	Đờn ca Tài tử Nam Bộ *Art of Don ca Tai tu music and song in southern Viet Nam*	513
2	Nghệ thuật Chầm riêng chà pây của người Khmer *Khmer Cham rieng cha pay art*	185
3	Nghề dệt chiếu *Mat weaving*	642
4	Lễ hội cúng biển Mỹ Long *My Long Sea worship festival*	398
5	Nghệ thuật sân khấu Dù Kê của người Khmer *Du Ke (or Lakhon Bassac) Theater Arts of the Khmer people*	404
6	Lễ hội Ok Om Bok của người Khmer *Khmer Ok Om Bok festival*	465
7	Lễ hội Vía Bà Chúa Xứ núi Sam *Festival of Ba Chua Xu Goddess at Sam Mountain*	405
8	Đại lễ Kỳ yên Đình Tân Phước Tây *Tan Phuoc Tay temple Ky Yen ceremony*	223
9	Lễ hội vía Bà Ngũ hành *Festival of the Five Elements Goddess*	569
10	Lễ làm chay *Tam Vu's Alms Giving festival/Tam Vu's Making Offerings festival*	365
11	Nghề đóng xuồng ghe Long định *The craft of building wooden boats in Long Dinh*	281
12	Nghề đan tre *Bamboo weaving*	639
13	Lễ cúng Việc *Ancestor worship ceremony/Ancestral ritual*	447
14	Lễ hội Đua bò Bảy Núi *Bay Nui bull racing festival*	449
15	Lễ hội Nghinh Ông *Nghinh Ong festival/Whale worship festival*	522
16	Lễ hội anh hùng Trương *The Death Anniversary of National Hero Truong Dinh*	361
17	Văn hóa Chợ nổi Cái Răng *Cai Rang floating market culture*	538
	Total images	7406

Empirically, we observed that this greedy-then-uniform ensembling approach consistently outperforms any individual checkpoint, offering improved robustness and better generalization – particularly in the context of multi-class classification for intangible cultural heritage images.

IV. EXPERIMENTS

1. Dataset

We conducted experiments on the ICH-17 dataset, which consists of 17 different categories with a total of 7,406 images. Specific details for each class are presented in Table II. The dataset initially consisted of two sets: a training set (6,657 images) and a test set (749 images) (this test set is the same as the one used in the previous study [4]). The original training set was randomly partitioned into a new training subset comprising 6,057 images and a validation subset containing 600 images. The training subset was utilized for model training, the validation subset was employed to monitor performance during training and to construct *model soups*, and the test set was reserved exclusively for final evaluation.

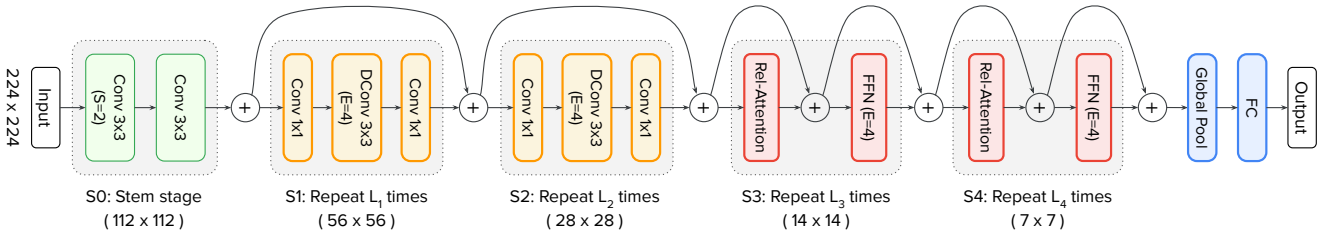

Figure 2. Overview of CoAtNet architecture. The model comprises five stages, gradually transitioning from convolutional blocks (MBConv) to Transformer blocks (self-attention). Each stage reduces spatial resolution while increasing the number of channels. Image credit: [2].

Algorithm 1: Greedy Soup Selection Algorithm

Input: Set of candidate checkpoints $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}\}$;
Validation set $\mathcal{D}_{\text{val}}$;
Output: Selected index set $\mathcal{S}$ and final soup weights θ_{soup}

Initialize $\mathcal{S} \leftarrow \{\arg\max_k \text{Acc}(\theta^{(k)}; \mathcal{D}_{\text{val}})\}$;
// Start with best checkpoint
for $k = 1$ **to** T **do**
  if $k \notin \mathcal{S}$ **then**
    Compute temporary average:
    $\theta_{\text{temp}} = \frac{1}{|\mathcal{S}|+1}\left(\sum_{j\in\mathcal{S}} \theta^{(j)} + \theta^{(k)}\right)$;
    Evaluate temporary accuracy:
    $\text{Acc}_{\text{temp}} = \text{Accuracy}(\theta_{\text{temp}}; \mathcal{D}_{\text{val}})$;
    if $Acc_{temp} \geq Acc\left(\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \theta^{(j)}\right)$ **then**
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{k\}$; // Accept checkpoint
    end
  end
end
Compute final average: $\theta_{\text{soup}} = \frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}} \theta^{(k)}$;
return $\mathcal{S}, \theta_{soup}$

2. Implementation Details

Experiments were conducted on the **CoAtNet** [2] model trained on the ImageNet-1k dataset[1] (referred to as CoAtNet-0) and a more parameter-heavy version pre-trained on the ImageNet-12k dataset, then fine-tuned on the ImageNet-1k dataset[2] (referred to as CoAtNet-2). Experiments were also performed on the **ResNet-50**[3] and **DenseNet-121**[4] models as baselines for comparison, both initialized with pre-trained weights from their respective models on ImageNet-1k. Additionally, the **ViT** model[5] was evaluated as a baseline for comparison with a full attention-based model. The number of parameters for each model is presented in Table III.

TABLE II
NUMBER OF IMAGES PER CLASS IN THE TRAINING, VALIDATION, AND TEST SETS OF THE ICH-17 DATASET

Class	Train	Validation	Test	Total
1	420	41	52	513
2	151	15	19	185
3	525	52	65	642
4	326	32	40	398
5	330	33	41	404
6	380	38	47	465
7	331	33	41	405
8	182	18	23	223
9	466	46	57	569
10	298	30	37	365
11	229	23	29	281
12	522	52	65	639
13	366	36	45	447
14	368	36	45	449
15	428	42	52	522
16	295	29	37	361
17	440	44	54	538
Total	**6057**	**600**	**749**	**7406**

All images were resized to a fixed size of 224×224 pixels and normalized based on ImageNet channel-wise statistics with mean $\mu = [0.485, 0.456, 0.406]$ and standard deviation $\sigma = [0.229, 0.224, 0.225]$. For data augmentation, horizontal flip combined with MixUp [16] and CutMix [15] was applied to improve the generalization capacity of the model.

During the model training phase, all models are fine-tuned in an end-to-end manner (i.e., all parameters are fine-tuned) with a batch size of 128 and were trained for 50 epochs. The loss function used was cross-entropy, and parameters were optimized using the AdamW [10] optimizer with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-3}. For the learning rate scheduler, cosine annealing with warm restarts was employed, setting $min_{lr} = 0.0$, an initial value for cycle length $T_0 = 3$, and a multiplicative factor $T_{mult} = 1$. The threshold for gradient

[1] https://huggingface.co/timm/coatnet_0_rw_224.sw_in1k
[2] https://huggingface.co/timm/coatnet_2_rw_224.sw_in12k_ft_in1k
[3] https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html
[4] https://docs.pytorch.org/vision/main/models/generated/torchvision.models.densenet121.html
[5] https://huggingface.co/timm/vit_base_patch16_224.augreg2_in21k_ft_in1k

norm clipping was set to 1.0 to ensure more stable training. Mixed precision with *fp16* was used for all experiments.

Instead of saving only a single checkpoint for final evaluation, the top $k = 8$ checkpoints were saved for each metric, including *loss*, *accuracy*, and *F1* score, based on the validation set, resulting in a total of 24 best checkpoints saved (a checkpoint from a single epoch may be saved multiple times). The results are reported based on the checkpoint with the highest *F1* score on the test set; in cases where *F1* scores are equal, *accuracy* is used for comparison (see Appendix A for a more detailed explanation).

Rationale for $k = 8$: We selected $k = 8$ checkpoints per metric (loss, accuracy, F1) for *model soups* construction. This choice reflects a balance between diversity and stability: fewer checkpoints would limit the potential benefit of weight averaging, while too many may introduce noisy or suboptimal snapshots. Empirically, we observed that $k = 8$ yields the most stable improvements on the validation and test sets, whereas smaller k led to higher variance and larger k offered no further gains. Moreover, storing eight checkpoints per criterion remains computationally and storage-efficient while providing a sufficiently diverse pool for greedy and uniform soup selection.

Experimental Setup: All experiments were conducted on a system equipped with an AMD EPYC 7702 64-Core Processor (16 physical cores allocated), 31 GB of RAM, and an NVIDIA GeForce RTX 3090 GPU. This configuration provided sufficient computational power for efficient model training and inference.

3. Evaluation Metrics

To evaluate the performance of our models on the multi-class classification task, the following standard metrics are employed: **Accuracy**, **Precision**, **Recall**, and **F1-score**. These metrics are computed over the test set and averaged across all classes to account for potential class imbalance (i.e., *macro-averaging*, except for **Accuracy**, which is calculated by considering the total number of correctly predicted labels across all classes, relative to the total number of test samples).

- **Accuracy** measures the proportion of correctly predicted labels over the total number of test samples:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(y_i = \hat{y}_i), \tag{2}$$

where y_i is the ground-truth label, $\hat{y}_i$ is the predicted label, and $\mathbb{1}(\cdot)$ is the indicator function.

- **Precision**, **Recall**, and **F1-score** are computed for each class and then averaged (macro-average) to ensure equal weighting regardless of class frequency:

$$\text{Precision}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c}, \tag{3}$$

$$\text{Recall}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c}, \tag{4}$$

$$\text{F1}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C} \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c}, \tag{5}$$

where TP_c, FP_c, and FN_c denote the number of true positives, false positives, and false negatives for class c, respectively, and C is the total number of classes.

These metrics allow for the assessment of not only overall model performance, but also class-wise behavior, which is critical in applications with class imbalance or semantic diversity.

4. Experiment Results

a) Main Results and Analysis

Table IV summarizes the classification performance of various baseline models and our proposed CoAtNet-based framework on the ICH-17 test set. Among the conventional architectures, ViT outperforms ResNet-50 and DenseNet-121, achieving an F1-score of 67.41%, highlighting the potential of attention-based models for ICH classification.

CoAtNet-0 achieves a solid performance with an accuracy of 70.63% and F1-score of 65.98%, outperforming all baselines and previous studies without ensembling including ViT, despite having approximately three times fewer parameters (see Table III). This demonstrates the efficiency of CoAtNet's hybrid design, which combines convolutional and attention mechanisms.

TABLE III
NUMBER OF PARAMETERS FOR EACH MODEL

Model	# Params (M)
ResNet-50	23.54
DenseNet-121	6.97
ViT	85.81
CoAtNet-0	26.68
CoAtNet-2	72.86

Applying *uniform soup* to CoAtNet-0 leads to improvements in accuracy (+0.67%), recall (+0.97%), and F1-score (+0.40%), though precision decreases slightly. Interestingly, *greedy soup* also improves performance, achieving a slightly higher F1-score (+0.44%) than uniform soup, with a modest gain in accuracy (+0.40%) and recall (+0.77%), albeit with lower precision. These results indicate that even for smaller

backbones like CoAtNet-0, weight-space ensembling can be beneficial, though metric trade-offs may vary across strategies.

When scaling up to CoAtNet-2, the performance further improves across all metrics. The base model reaches 71.43% accuracy and 68.58% F1-score. Applying uniform soup boosts these values to 72.36% and 69.28%, respectively, achieving the best overall results. Greedy soup also yields substantial gains (72.23% accuracy, 69.05% F1-score), slightly below uniform soup but consistently better than the base model.

These findings collectively validate the effectiveness of weight-space ensembling via *model soups*. In particular, the improvements seen with both uniform and greedy soups – especially in the larger CoAtNet-2 model – underscore the advantages of checkpoint averaging in enhancing generalization without increasing inference complexity.

b) Comparison between Uniform and Greedy Soup

To better understand the effects of different weight-space ensembling strategies, we compare *uniform soup* and *greedy soup* applied to the CoAtNet-2 model. As shown in Table IV, both approaches improve upon the base model across all evaluation metrics, confirming the robustness of *model soups* in enhancing generalization.

Uniform soup achieves the best overall results, with 72.36% accuracy and 69.28% F1-score, outperforming greedy soup by +0.13% in accuracy and +0.23% in F1. It also shows slightly higher precision and recall. This suggests that when checkpoint diversity is sufficient, uniform averaging can yield strong and stable performance gains.

Greedy soup, while slightly behind in absolute performance, still improves over the base CoAtNet-2 by +0.80% in accuracy and +0.47% in F1-score. Unlike uniform soup, the greedy strategy incrementally selects checkpoints based on validation performance, resulting in a leaner and more selective ensemble. This can be advantageous in scenarios with resource constraints or limited checkpoint quality.

Figure 3 illustrates the accuracy of the individual models (CoAtNet-2) and the ensemble models generated via *model soups* techniques, including greedy soup and uniform soup. It is evident that the individual CoAtNet-2 models consistently yield lower accuracy compared to the *model soups* ensembles.

Interestingly, on the smaller CoAtNet-0 model, the performance gap narrows: uniform soup gains +0.40% in F1-score, while greedy soup yields a slightly higher improvement of +0.44%, despite lower precision in both cases. This variation highlights that the effectiveness of each ensembling strategy may depend on model capacity and the diversity of saved checkpoints.

Overall, while uniform soup offers the best results in our setting, greedy soup remains a competitive alternative, especially when computational constraints or checkpoint sparsity are present.

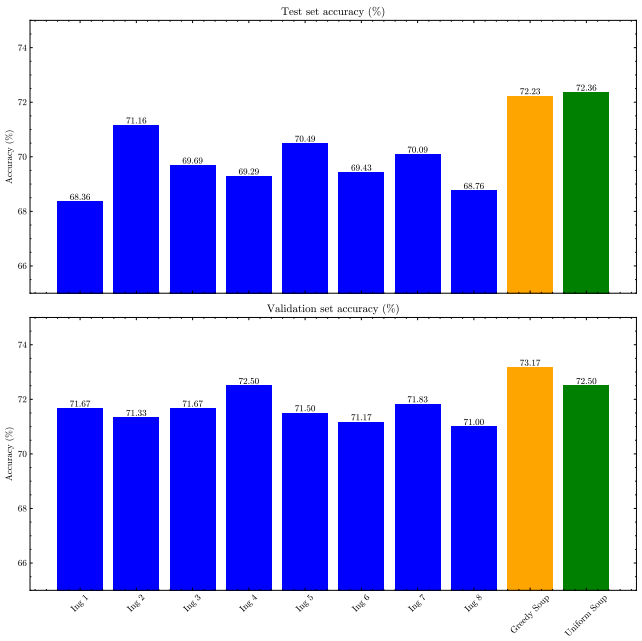

Figure 3. Comparison of validation and test accuracy (%) between individual CoAtNet-2 models and their combinations via *model soups* techniques. The ensemble models – greedy soup and uniform soup – outperform all individual models, demonstrating the effectiveness of *model soups* in enhancing generalization

c) Per-Class Performance Analysis

To further understand the behavior of our model, we conduct a per-class analysis of CoAtNet-2 in both its best individual checkpoint and the *model soups* variant, as shown in Table V. The results indicate that *model soups* yields performance improvements in the majority of classes across all evaluation metrics.

Notably, *model soups* enhances classification accuracy in 7 out of 17 classes and improves F1-score in 11 classes. For example, class 6 (*Lễ hội Ok Om Bok của người Khmer*) sees accuracy rise from 61.70% to 74.47% and F1-score from 59.79% to 69.31%, suggesting that ensemble averaging stabilizes predictions in complex or ambiguous categories. Similar improvements are observed in classes 4 (*Lễ hội cúng biển Mỹ Long*), 7 (*Lễ hội Vía Bà Chúa Xứ núi Sam*), 10 (*Lễ làm chay*), and 13 (*Lễ cúng Việc*), which benefit from consistent gains in both precision and recall.

In some cases, performance remains unchanged (e.g., classes 1, 2, 5, 11, 12, 14, 15, and 17), or shows slight drops in certain metrics. For instance, class 8 (*Đại lễ Kỳ yên Đình Tân Phước Tây*) exhibits a drop in accuracy and recall, though precision slightly increases. Similarly, class

TABLE IV
CLASSIFICATION RESULTS ON THE ICH-17 TEST SET. ASTERISK (*) INDICATES RESULTS REPORTED IN THE ORIGINAL PAPER, AND – DENOTES MISSING RESULTS. THE VALUES LOCATED IN THE UPPER RIGHT CORNER OF THE ENTRIES IN THE ROWS CORRESPONDING TO UNIFORM SOUP AND GREEDY SOUP INDICATE THE DIFFERENCE COMPARED TO THE CORRESPONDING MODEL WITHOUT APPLYING MODEL SOUPS.

Model	Accuracy	Precision	Recall	F1
ResNet-50	65.55	64.72	62.33	62.80
DenseNet-121	64.35	60.33	60.33	60.02
ViT	70.09	68.79	66.99	67.41
Do et al. [5]	65.32*	–	–	–
Tran et al. [12]	66.76*	–	–	–
CoAtNet-0	70.63	67.66	66.17	65.98
CoAtNet-0 + uniform soup	$71.30^{\uparrow 0.67}$	$66.65^{\downarrow 1.01}$	$67.14^{\uparrow 0.97}$	$66.38^{\uparrow 0.40}$
CoAtNet-0 + greedy soup	$71.03^{\uparrow 0.40}$	$66.85^{\downarrow 0.81}$	$66.94^{\uparrow 0.77}$	$66.42^{\uparrow 0.44}$
CoAtNet-2	71.43	69.34	68.40	68.58
CoAtNet-2 + uniform soup	$\mathbf{72.36}^{\uparrow 0.93}$	$\mathbf{69.98}^{\uparrow 0.64}$	$\mathbf{69.09}^{\uparrow 0.69}$	$\mathbf{69.28}^{\uparrow 0.70}$
CoAtNet-2 + greedy soup	$72.23^{\uparrow 0.80}$	$69.91^{\uparrow 0.57}$	$68.81^{\uparrow 0.41}$	$69.05^{\uparrow 0.47}$

TABLE V
RESULTS FOR EACH CLASS OF THE COATNET-2 MODEL, WHERE *best* CORRESPONDS TO THE HIGHEST RESULT AMONG THE SAVED CHECKPOINTS (SEE IV.2), AND *soups* CORRESPONDS TO APPLYING A MODEL SOUP (EITHER UNIFORM OR GREEDY). THE **BOLDED** VALUES CORRESPOND TO THE HIGHEST VALUE IN THE COLUMN FOR THE RESPECTIVE CLASS.

Class	Accuracy		Precision		Recall		F1	
	best	soups	best	soups	best	soups	best	soups
1	**88.46**	**88.46**	80.70	**85.19**	**88.46**	**88.46**	84.40	**86.79**
2	**47.37**	**47.37**	**75.00**	69.23	**47.37**	**47.37**	**58.06**	56.25
3	**93.85**	92.31	**95.31**	95.24	**93.85**	92.31	**94.57**	93.75
4	22.50	**25.00**	28.12	**33.33**	22.50	**25.00**	25.00	**28.57**
5	**65.85**	**65.85**	64.29	**71.05**	**65.85**	**65.85**	65.06	**68.35**
6	61.70	**74.47**	58.00	**64.81**	61.70	**74.47**	59.79	**69.31**
7	78.05	**80.49**	74.42	**76.74**	78.05	**80.49**	76.19	**78.57**
8	**43.48**	34.78	40.00	**44.44**	**43.48**	34.78	**41.67**	39.02
9	50.88	**52.63**	**50.00**	43.07	50.88	**52.63**	**50.43**	47.24
10	54.05	**59.46**	55.56	**57.89**	54.05	**59.46**	54.79	**58.67**
11	**86.21**	**86.21**	**89.29**	86.20	**86.21**	**86.21**	**87.72**	86.20
12	**87.69**	**87.69**	79.17	**81.43**	**87.69**	**87.69**	83.21	**84.44**
13	57.78	**64.44**	68.42	**72.50**	57.78	**64.44**	62.65	**68.23**
14	**86.67**	**86.67**	86.67	**95.12**	**86.67**	**86.67**	86.67	**90.70**
15	**80.77**	**80.77**	72.41	**75.00**	**80.77**	**80.77**	76.36	**77.78**
16	64.86	**70.27**	**70.59**	66.66	64.86	**70.27**	67.61	**68.42**
17	**92.59**	**92.59**	**90.91**	**90.91**	**92.59**	**92.59**	**91.74**	**91.74**

16 (*Lễ hội anh hùng Trương*) shows a modest increase in accuracy but a slight decrease in precision. These variations may stem from overfitting in some checkpoints or limited representation during soup selection.

Overall, in terms of F1-score, *model soups* improves or maintains performance in 12 out of 17 classes, reinforcing its robustness and reliability across diverse cultural heritage categories with varying visual complexity and sample sizes. This per-class evaluation confirms that weight-space ensembling not only improves average-case metrics but also contributes to more stable and generalized class-wise performance.

V. ANALYTICALLY COMPARING MODEL SOUPS TO SOFT VOTING

Given a dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ of N samples, each sample is passed through a model f, followed by a softmax layer, resulting in a probability vector $\mathbf{p}_i^{(f)} \in \Delta^{C-1}$ for each input x_i, where Δ^{C-1} denotes the $(C-1)$-dimensional probability simplex. Collectively, the model f produces a matrix of softmax outputs:

$$\mathbf{P}^{(f)} = \begin{bmatrix} (\mathbf{p}_1^{(f)})^T \\ (\mathbf{p}_2^{(f)})^T \\ \vdots \\ (\mathbf{p}_N^{(f)})^T \end{bmatrix} \in \mathbb{R}^{N \times C}$$

To compare two models f and g, the mean cross-entropy between their corresponding softmax outputs across the dataset is computed:

$$\begin{aligned} \text{Dist}(f, g) &= \frac{1}{N} \sum_{i=1}^{N} H_{\text{sym}} \left(\mathbf{p}_i^{(f)}, \mathbf{p}_i^{(g)} \right), \\ \text{where} \quad H_{\text{sym}}(\mathbf{p}, \mathbf{q}) &= \left[H(\mathbf{p}, \mathbf{q}) + H(\mathbf{q}, \mathbf{p}) \right], \\ H(\mathbf{p}, \mathbf{q}) &= -\sum_{j=1}^{C} p_j \log q_j \end{aligned} \tag{6}$$

This yields a symmetric distance matrix $D \in \mathbb{R}^{|\mathcal{F}| \times |\mathcal{F}|}$ over a set of models $\mathcal{F}$. To visualize the relational structure between models, *Multidimensional Scaling (MDS)* [3] is applied to project the distance matrix D into a two-dimensional space. MDS is particularly well-suited for this task because pairwise distances are preserved in the low-dimensional embedding, enabling the relative similarity between models to be captured and interpreted based purely on their output behaviors. Unlike other nonlinear techniques such as t-SNE [1], which may distort global geometry, both local and global distances are maintained by MDS, rendering it more appropriate for preserving structural fidelity in model comparison tasks.

Model Soups Diversity Justification

The aforementioned pairwise distance matrix D and its corresponding MDS projection are also utilized as a basis for analyzing model diversity in the context of *model soups*. Specifically, the 2D embedding space is employed to visualize the distribution of candidate models and to evaluate whether the models selected for ensembling (i.e., those included in the soup) are well-spread in output space.

Let $\mathcal{S} \subset \mathcal{F}$ denote the subset of models chosen for *model soups*. If *model soups* is effective at capturing a diverse ensemble, the points corresponding to $\mathcal{S}$ in the MDS embedding are expected to exhibit broad coverage of the model space – avoiding clustering and instead spanning distinct regions.

In the experiments conducted, the selected models $\mathcal{S}$ are highlighted in the MDS scatter plot and observed to be well-separated, confirming that the soup does not merely average redundant models, but rather aggregates models with heterogeneous prediction behaviors. This diversity in softmax outputs contributes to the ensemble's robustness and improved generalization.

Thus, the analysis provides empirical evidence that the *model soups* construction process implicitly favors diverse members in the model space, a property known to be beneficial for ensembling in both theory and practice.

Visualizing Model Diversity on the Validation Set

To evaluate the diversity of models selected by *model soups*, the softmax outputs (i.e., probabilities) of all candidate and ensemble models are projected using Multidimensional Scaling (MDS) based on pairwise cross-entropy distances. The validation set is chosen for visualization (Figure 4), as it reflects the basis on which checkpoint selection is performed by both greedy and uniform soup strategies.

The resulting 2D embedding reveals that the ingredient models (blue) are well-dispersed across the output space, indicating a high degree of predictive diversity. In contrast, the ensemble models – Greedy (red), Uniform (green), and Soft Voting (orange) – are observed to be clustered near the center of this distribution. This geometry suggests that *model soups* draws from a broad range of behaviors and aggregates diverse checkpoints, while Soft Voting averages across more redundant predictions.

The geometric separation among the ingredients confirms that *model soups* does not merely average similar models but actively selects complementary ones. Such diversity is critical for reducing variance in ensemble predictions and improving generalization. This visualization provides compelling evidence that *model soups* functions as a diversity-aware selection strategy, offering both theoretical and empirical benefits over traditional voting-based ensembles.

In Soft Voting, the ensemble prediction is computed by averaging the output probability vectors from all individual models in the ensemble. Given a set of M trained models $\{f^{(1)}, f^{(2)}, \ldots, f^{(M)}\}$, the ensemble output for an input x is defined as:

$$\mathbf{p}^{\text{soft}}(x) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{p}^{(f^{(m)})}(x)$$

where $\mathbf{p}^{(f^{(m)})}(x) \in \Delta^{C-1}$ denotes the softmax output of model $f^{(m)}$ for input x, and Δ^{C-1} is the $(C-1)$-dimensional probability simplex over C classes.

The final predicted class is then:

$$\hat{y} = \arg\max_j \left[\mathbf{p}^{\text{soft}}(x) \right]_j$$

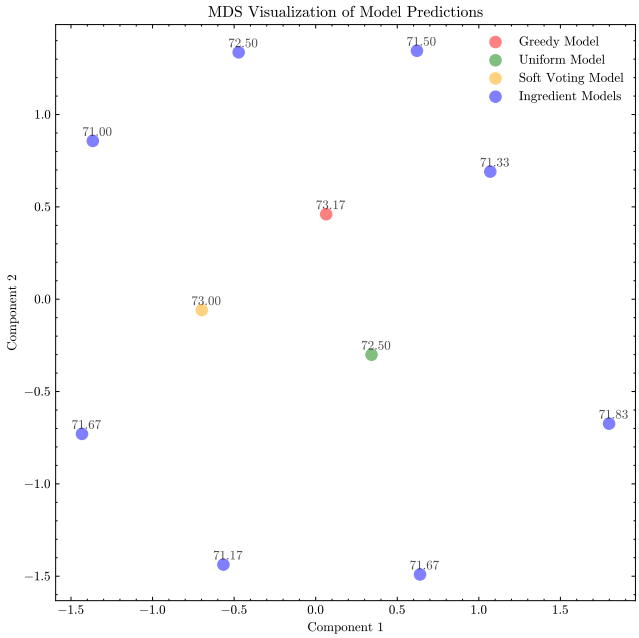

Figure 4. **MDS visualization of model predictions on the validation set.** Each point represents a model projected via MDS using cross-entropy-based pairwise distances. The ingredient models (blue) are broadly distributed, while ensemble models – Greedy (red), Uniform (green), and Soft Voting (orange) – cluster near the center. This spatial structure confirms that *model soups* leverages predictive diversity more effectively than Soft Voting. The value above each point represents the *accuracy* on the validation set for the corresponding model.

This method treats all models as equally informative, regardless of their individual accuracy or diversity, which may limit its effectiveness when model predictions are highly correlated.

Comparison with Soft Voting and Bias-Variance Analysis

To better understand the role of *model soups* in improving generalization, it is compared to the classical Soft Voting ensemble. In Soft Voting, all models contribute equally to the final prediction through uniform averaging of their softmax outputs, regardless of their individual performance or redundancy. This approach does not explicitly account for diversity among models, and can lead to ensembling over highly similar checkpoints. As shown in Figure 4, the Soft Voting ensemble (orange) is located near the center of the output-space embedding, reflecting its averaging over clustered, potentially redundant models.

In contrast, *model soups* – particularly in its *greedy* and *uniform* forms – constructs the ensemble by selectively averaging a subset of validation-optimal checkpoints that are geometrically diverse in the output space. This selective diversity is not incidental; from a bias–variance perspective, averaging over complementary models with uncorrelated errors is known to reduce prediction variance while introducing minimal additional bias. The analysis demonstrates that *model soups* leverages this property effectively, functioning as a variance-reducing, diversity-aware ensemble strategy.

Overall, *model soups* goes beyond naive output averaging: it dynamically adapts to the geometry of model outputs and the validation objective, yielding more robust predictions under low-resource and high-variance conditions compared to Soft Voting.

VI. Ablation Study: Effect of Removing Pre-training

To further investigate the impact of pre-trained initialization, an ablation study was conducted where both CoAtNet-0 and CoAtNet-2 were trained from random initialization rather than being fine-tuned from ImageNet pre-trained weights. The models were trained under the same experimental settings as in the main experiments, and their performance was evaluated on the ICH-17 test set.

As shown in Table VII, both models trained from scratch perform significantly worse compared to their ImageNet-pretrained counterparts (CoAtNet-0: 70.63% accuracy, 65.98% F1; CoAtNet-2: 71.43% accuracy, 68.58% F1). Training without pre-training results in drops of approximately 20–22 percentage points in accuracy and 21–23 percentage points in macro F1-score. These results highlight that pre-training provides critical inductive biases and transferable representations, which are especially important in the low-resource and noisy ICH-17 dataset.

VII. Conclusion and Future Work

In this study, we introduced a robust framework for classifying images of Intangible Cultural Heritage (ICH) in the Mekong Delta by combining the hybrid CoAtNet architecture with weight-space ensembling via *model soups*. By leveraging the representational power of CoAtNet and the generalization benefits of checkpoint averaging, our method achieves state-of-the-art performance on the ICH-17 dataset. We adopted both uniform and greedy soup strategies, outperforming strong baselines such as ResNet-50, DenseNet-121, and ViT across global accuracy and per-class F1-score metrics.

Beyond performance gains, a comprehensive analysis of model diversity was conducted using cross-entropy-based distances and Multidimensional Scaling (MDS). This analysis reveals that the models selected by the soup strategy span diverse regions in output space – unlike classical Soft Voting, which averages over tightly clustered models. The findings empirically confirm that *model soups* acts

TABLE VI
COMPARISON OF MODEL SOUPS AND SOFT VOTING ON KEY ENSEMBLE PROPERTIES.

Aspect	Soft Voting	Model Soups
Inference Cost	High (all models active)	Low (single averaged model)
Memory Footprint	Scales with number of models	Constant (single model)
Deployment Simplicity	Requires parallel model storage	Simple (one model checkpoint)
Bias Reduction	Limited	Slight increase (depends on soup strategy)
Variance Reduction	Moderate (correlated models)	High (diverse checkpoint averaging)

TABLE VII
PERFORMANCE OF COATNET MODELS TRAINED FROM RANDOM INITIALIZATION ON THE ICH-17 TEST SET. THE VALUES LOCATED IN THE UPPER RIGHT CORNER OF THE ENTRIES IN THE ROWS CORRESPONDING TO UNIFORM SOUP AND GREEDY SOUP INDICATE THE DIFFERENCE COMPARED TO THE CORRESPONDING MODEL WITHOUT APPLYING *model soups*.

Model	Accuracy	Precision	Recall	F1
CoAtNet-0	49.13	46.36	45.31	44.85
CoAtNet-2	51.27	48.87	48.21	47.81
CoAtNet-2 + uniform	$52.07^{\uparrow 0.80}$	$49.27^{\uparrow 0.40}$	$48.82^{\uparrow 0.61}$	$47.99^{\uparrow 0.18}$
CoAtNet-2 + greedy	$51.80^{\uparrow 0.53}$	$48.72^{\downarrow 0.15}$	$48.77^{\uparrow 0.56}$	$48.11^{\uparrow 0.30}$

not merely as an averaging scheme, but as a diversity-aware selection mechanism that enhances robustness and generalization across training, validation, and test splits.

For future work, this framework is planned to be extended in two directions. First, integrating semantic priors and multi-modal signals (e.g., textual metadata) will be investigated to further enhance performance in low-resource cultural datasets. Second, the methodology is intended to be scaled to broader ICH datasets across other regions, contributing to inclusive and AI-driven preservation of cultural heritage at both regional and global levels.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to the AniAge project for providing the ICH dataset used in our experiments. This valuable resource enabled a fair and rigorous evaluation of the proposed methods in a culturally grounded context.

DATA AVAILABILITY

The dataset used in this study is not publicly available due to institutional or licensing restrictions. However, it can be made available for academic use upon reasonable request. Interested researchers may contact the authors for further information.

REFERENCES

[1] Cai, T.T., Ma, R.: Theoretical foundations of t-sne for visualizing high-dimensional clustered data (2022), https://arxiv.org/abs/2105.07536

[2] Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes (2021), https://arxiv.org/abs/2106.04803

[3] Demaine, E., Hesterberg, A., Koehler, F., Lynch, J., Urschel, J.: Multidimensional scaling: Approximation and complexity. In: International conference on machine learning. pp. 2568–2578. PMLR (2021)

[4] Do, T.N., Pham, T.P., Nguyen, H.H., Pham, N.K.: Visual classification of intangible cultural heritage images in the mekong delta. Data Analytics for Cultural Heritage: Current Trends and Concepts pp. 71–89 (2021)

[5] Do, T.N., Pham, T.P., Pham, N.K., Nguyen, H.H., Tabia, K., Benferhat, S.: Stacking of svms for classifying intangible cultural heritage images. In: Le Thi, H.A., Le, H.M., Pham Dinh, T., Nguyen, N.T. (eds.) Advanced Computational Methods for Knowledge Engineering. pp. 186–196. Springer International Publishing, Cham (2020)

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://arxiv.org/abs/2010.11929

[7] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), https://arxiv.org/abs/1512.03385

[8] Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks (2019), https://arxiv.org/abs/1709.01507

[9] Huang, G., Liu, Z., van der Maaten, L., Wein-

berger, K.Q.: Densely connected convolutional networks (2018), https://arxiv.org/abs/1608.06993
[10] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
[11] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks (2019), https://arxiv.org/abs/1801.04381
[12] Tran, M.T., Pham, T.P., Thai-Nghe, N., Do, T.N.: Fusing models for classifying intangible cultural heritage images in the mekong delta. In: International Conference on Intelligent Systems and Data Science. pp. 202–212. Springer (2024)
[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023), https://arxiv.org/abs/1706.03762
[14] Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., Schmidt, L.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time (2022), https://arxiv.org/abs/2203.05482
[15] Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)
[16] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)

Quoc-Khang Tran received the B.Eng. degree in Computer Science from Can Tho University (CTU), Vietnam, in 2023. He is currently a Teaching Assistant at the Faculty of Computer Science, College of Information and Communication Technology (CICT), CTU, and pursuing the M.Sc. degree in Computer Science at CTU. His research interests include deep learning, natural language processing, and algorithm design.
Email: tqkhang@ctu.edu.vn

Minh-Thien Nguyen He is currently a final-year undergraduate student in Information Technology at Can Tho University. His research focuses on deep learning and natural language processing.
Email: thienb2111822@student.ctu.edu.vn

Nguyen-Khang Pham received the Ph.D. degree in Computer Science from the University of Rennes 1, France, in 2009, the M.Sc. degree in Computer Science from the Francophone Institute of Informatics, Vietnam, in 2004, and the B.Eng. degree in Computer Science from Can Tho University (CTU), Vietnam, in 1999. He is currently an Associate Professor in the Faculty of Computer Science, College of Information and Communication Technology, CTU, where he also serves as the Dean of the Graduate School. His research interests include information retrieval, content-based image retrieval, data mining, and information visualization.
Email: pnkhang@ctu.edu.vn

APPENDIX A
THE METHOD FOR SELECTING THE BEST CHECKPOINT IN OUR APPROACH

Instead of selecting only a single checkpoint, such as the one with the highest accuracy or lowest loss based on the validation set, the top $k = 8$ checkpoints for each metric, including loss, accuracy, and F1 score, were saved based on the validation set, resulting in a total of 24 best checkpoints saved. These checkpoints were then evaluated on the test set, and the results of the checkpoint with the highest F1 score were reported. In cases where multiple checkpoints had equal F1 scores, they were compared based on accuracy. This approach was adopted because, during the experimental phase, it was observed that some results on the validation set were significantly high, yet a relatively large gap was found on the test set (in some cases, up to a 4% difference in accuracy). This could lead to a scenario where the checkpoint with the highest validation performance is selected, resulting in relatively poor test set performance.

We hypothesize that this phenomenon is due to the characteristics of the ICH-17 dataset, where the context of the images is highly diverse, and the data remains relatively noisy, reducing the model's generalization capability. Furthermore, the number of images per class in both the test and validation sets is limited when considered individually, combined with the diverse contexts of the images, which may lead to an inconsistent distribution between the validation and test sets.

However, it should be noted that this does not affect the application of *model soups*, as evaluation and checkpoint selection were performed entirely based on the validation set results, and the test set was only evaluated on the final checkpoint aggregated from the uniform soup and greedy soup algorithms.