White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/248072/

Version: Accepted Version

Article:
Li, Z., Kamnitsas, K., Islam, M. et al. (2026) Estimating performance for safe deployment of machine learning models. Medical Image Analysis, 116. 104364. ISSN: 1361-8415

https://doi.org/10.1016/j.media.2026.104364

© 2026 The Authors. Except as otherwise noted, this author-accepted version of a journal article published in Medical Image Analysis is made available via the University of Sheffield Research Publications and Copyright Policy under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

arXiv:2610.29262v2 [cs.LG] 29 Oct 2026

Estimating Performance for Safe Deployment of Machine Learning Models

Zeju Li[a,b], Konstantinos Kamnitsas[c], Mobarak I. Hoque[d], Mélanie Roschewitz[e], Cheng Ouyang[c,e,g], Chen Chen[f,e], Ben Glocker[e]

[a] *College of Biomedical Engineering, Fudan University, Shanghai, China*
[b] *Oxford Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK*
[c] *Department of Engineering Science, University of Oxford, Oxford, UK*
[d] *Division of Informatics, Imaging and Data Science, University of Manchester, Manchester, UK*
[e] *Department of Computing, Imperial College London, London, UK*
[f] *School of Computer Science, University of Sheffield, Sheffield, UK*
[g] *Institute of Clinical Sciences, Imperial College London, London, UK*

Abstract

The main challenge in applying deep neural networks to real-world scenarios lies in distributional discrepancies between training and test datasets, often caused by variations in data acquisition conditions. Assessing the performance of pre-trained models without labels is crucial for safety, but this is complicated by deep models' overly optimistic probabilistic outputs, often exacerbated by class imbalances, leading to unreliable predictions. To address these issues, we introduce MOVAL, a novel MOdel eVALuation tool that estimates performance using unlabeled test data. MOVAL incorporates class-wise confidence calibration to adjust model outputs, allowing for the estimation of metrics beyond accuracy and mitigating the detrimental effects of class imbalance on confidence calibration. Additionally, we propose an ensemble approach that combines multiple calibrated confidence scores for a more robust solution. We validate MOVAL on natural and medical image datasets with various acquisition shifts, covering tasks from classification to segmentation. It achieves state-of-the-art results across evaluation metrics such as accuracy, F1-Score, AUC, and more. Furthermore, we demonstrate that MOVAL provides accurate performance estimations at the class, instance, and pixel levels, ensuring the transparent and safe deployment of machine learning models. We provide MOVAL as a deployment tool for monitoring and estimating performance under a range of domain shifts.

Keywords: Trustworthy Machine Learning, Interpretability, Quality Control, Deep Learning

1. Introduction

The witnessed success of machine learning is built upon the independently and identically distributed (IID) assumption. However, this assumption does not always hold in real-world applications due to domain shifts, where the training and test data come from different distributions. There are multiple kinds of domain shifts, but we primarily consider acquisition shifts, as they have the most direct and significant impact on model performance, causing models to fail silently. Therefore, trustworthy safeguards are needed to ensure model performance with minimal assumptions and to provide prompt feedback without access to ground truth (Eche et al., 2021). Existing algorithms for model evaluation are either task-specific or rely on globally calibrated probabilities, which do not provide a comprehensive evaluation beyond accuracy or Dice similarity coefficient (DSC) (Zenk et al., 2025). In this work, we combine recent advances in model calibration and confidence score estimation to develop a new model evaluation method that can be applied to various types of tasks, informing model failure with multiple estimated metrics and meaningful indicators at different levels.

Developing reliable, real-time performance assessment systems is crucial for the safe deployment of machine learning in clinical practice. Such systems help identify the limits of model applicability and support practitioners in iteratively refining their training strategies. For instance, when a skin-lesion classification model is trained on data from a source domain, deployment decisions in a new target domain can be guided by the model's predicted accuracy on all unlabeled test images. Additionally, because magnetic resonance imaging (MRI) is typically non-quantitative, it is often unclear whether a model remains valid when applied in a local hospital where acquisition protocols differ and expert annotations may be unavailable. This approach helps prevent overfitting when optimizing results using a small number of labeled test images by using predicted accuracy as a surrogate to validation loss.

1.1. Related Works

Many methods have been proposed for this scenario, but they often lack robustness or have limited applicability. For example, distance metric be-

tween the training and test datasets can be a good indicator for model performance, but requires the access to validation data every time of estimating (Horvitz and Thompson, 1952; Chen et al., 2021) and/or need another regression model (Deng and Zheng, 2021; Robinson et al., 2018). Reference-based methods (Fan and Davidson, 2006; Wang et al., 2020b; Valindria et al., 2017; Li et al., 2022a) estimate model performance by comparing the prediction results to a predefined reference (i.e. shape template for segmentation), but lack the ability to leverage the probability of the model output, and are mostly limited to segmentation tasks with strong shape priors. As a separate but related line of work, conformal prediction methods (Barber et al., 2023) provide prediction sets with formal coverage guarantees, which can increase confidence at deployment but do not directly reflect the underlying model's performance. These are related but distinct goals: conformal prediction provides set-valued uncertainty guarantees for individual predictions under specific assumptions (e.g., exchangeability or its relaxations), whereas MOVAL estimates aggregate, class-, instance-, and pixel-wise performance metrics (accuracy, F1-Score, sensitivity, precision, AUC) without providing per-prediction coverage guarantees.

Confidence-score-based methods, which estimate the accuracy of a trained model using the mean value of the maximum class probability (MCP), are more efficient and are believed to be accurate for model evaluation (Elsahar and Gallé, 2019; Mehrtash et al., 2020; Garg et al., 2022; Islam et al., 2021; Baek et al., 2022; Ouyang et al., 2022b; Roschewitz and Glocker, 2023; Lu et al., 2023; Press et al., 2024). However, this approach is largely limited to accuracy estimation, which can be a misleading metric for imbalanced datasets, a common occurrence in real-world classification tasks, particularly in segmentation. For example, accuracy can approach 100% in imbalanced datasets, even if all samples are classified as belonging to the majority class. In medical applications, it is often preferable to track clinically relevant metrics such as sensitivity or specificity, or metrics that take into account class imbalance, such as area under the curve (AUC) or F1-Score.

Model calibration, also known as confidence calibration, is the process of adjusting the probability outputs of a trained model to better reflect the true likelihood of its predictions. This adjustment is typically performed using validation data, aiming to ensure that the calibrated model's confidence scores align with the observed accuracy on unseen test data. A model is said to be well-calibrated when the predicted confidence aligns with the true probability that the associated model prediction is correct. However, most methods as-

sume that every class performs equally well (Guo et al., 2017; Guillory et al., 2021; Garg et al., 2022; Kivimäki et al., 2024; Mehrtash et al., 2020), thus calibrating the probabilities of all classes in the same way. In practice, this is not true, especially in the presence of class imbalance, where predictions for majority classes tend to be overconfident, while predictions for minority classes are often underconfident (Li et al., 2020). Consequently, after calibration, although global metrics may be aligned, the discrepancy between different classes becomes even larger. This not only increases the estimation error on test data but also leads to misleading results when examining class-wise or instance-wise performance (Li et al., 2022b).

Several methods target multi-class calibration with class-dependent transformations, including Dirichlet calibration (Kull et al., 2019), matrix scaling and its regularized variants, and related approaches (Zong et al., 2024; Le Coz et al., 2024; Zhao et al., 2021; Valk and Kull, 2023). These methods can, in principle, apply different corrections across classes via a full $c \times c$ transformation. However, they are fit by minimizing an aggregate calibration loss (e.g., negative log-likelihood or expected calibration error) over the whole validation set rather than a class-specific objective, so they do not explicitly enforce that each class's calibrated confidence matches that class's true performance; correcting class imbalance is not their stated goal.

1.2. Contributions

We address these open challenges with MOVAL, a new MOdel eVALuation framework, benchmark, and deployable tool for assessing robust and comprehensive model performance under domain shifts without the need for manual labels. Specifically, we propose to enhance confidence score-based approaches through class-wise calibration, which aligns confidence estimates with performance metrics using class-specific parameters. This not only improves accuracy estimation by reducing bias between classes but also naturally extends confidence-based methods to the estimation of other metrics, including F1-Score, sensitivity, precision, and AUC. To our knowledge, MOVAL is the first benchmark to evaluate how well confidence scores estimate metrics beyond accuracy. Rather than introducing a fundamentally new calibration algorithm, our contribution lies in systematically unifying and extending existing calibration, confidence-score, and ensemble ideas across a broader range of metrics and tasks than prior work.

MOVAL is the first unified confidence score-based estimation framework applicable to tasks beyond classification—including 2D and 3D segmenta-

tion; existing works are primarily limited to classification. We systemically extend various confidence scores—beyond MCP—to the problem of performance estimation. This includes scores like Entropy (Steinhardt and Liang, 2016), Doctor (Granese et al., 2021), and Energy (Liu et al., 2020b), which were designed for identifying model failure but have not been extensively discussed in the context of performance estimation. Importantly, we find that an ensemble of multiple confidence scores provides the most robust and accurate performance estimation. We provide this as a ready-to-use tool and anticipate that practitioners will integrate it into their existing deployment pipelines as a standard procedure [1]. MOVAL does not require validation data at the inference stage, impose assumptions on the tasks, or rely on additional regression models, making it lightweight, flexible, and reproducible. Through experiments on six tasks across 6,572 test settings, MOVAL outperforms state-of-the-art methods in several areas, reducing estimation error by 43% in real-world classification tasks and by over 50% in segmentation tasks. Additionally, it provides detailed indicator results at the class-wise, instance-wise, and pixel-wise levels.

We presented our preliminary results at MICCAI 2022 (Li et al., 2022b), which mainly focused on the discussion of class-wise calibration for performance estimation, where it was shown to outperform class-wise calibration baselines such as vector scaling. Building on that, this paper further discusses: 1) the estimation of metrics other than accuracy, 2) estimation for 2D and 3D segmentation tasks, 3) estimation with different confidence scores and their ensemble with additional evidence, and most importantly, 4) the development of an open-source tool for the community to enable the safe deployment of machine learning models.

2. Preliminaries: Confidence-Based Model Evaluation

2.1. Problem Setup

In the context of c-class image classification, our goal is to establish a mapping between the image space $\mathcal{X} \subset \mathbb{R}^d$ and the label space $\mathcal{Y} = \{1, 2, \ldots, c\}$. This framework can be extended to segmentation tasks by assuming that an image patch $\boldsymbol{x}_i$ is sampled from $\mathcal{X}$, with $y_i \in \mathcal{Y}$ representing the label of the central pixels.

[1] https://pypi.org/project/moval/

We assume that two labeled image sets, $\mathcal{D}^{Tr} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{I}$ and $\mathcal{D}^{V} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, drawn from the same source domain, are available for model training and validation. Once trained, the model is applied to a target domain $\mathcal{D}^{Te}$, where the distribution differs from the source domain, a common scenario in real-world applications. Our goal is to estimate its performance, including accuracy, F1-Score, sensitivity, precision, and AUC, on the target domain to support decision-making when only a set of test images $\mathcal{D}^{Te} = \{\boldsymbol{x}_i^{Te}\}_{i=1}^{M}$ is available without any ground truth.

We propose MOVAL, a framework for estimating model performance under domain shifts without requiring ground truth labels, specifically for modern machine learning models like convolutional neural networks (CNNs). MOVAL leverages logits (pre-softmax outputs) to estimate metrics across the dataset. To ensure these probabilities are useful for performance estimation under domain shifts, they are adjusted to match real metrics on a labeled validation set from the source domain.

After training a neural network f on the training dataset $\mathcal{D}^{Tr}$, we aim to develop a calibration model based on its performance on the validation dataset $\mathcal{D}^{V} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$. Specifically, given an input sample $\boldsymbol{x}_i$ from the validation set, the network output $f(\boldsymbol{x}_i)$ (i.e., logits) is obtained. Our objective is to minimize the following equation based on the validation dataset with respect to the parameters in the MOVAL framework $\mathcal{M}$:

$$\left| \mathcal{M}\Big(\{f(\boldsymbol{x}_i)\}_{i=1}^{N}\Big) - \mathcal{C}\Big(\{(f(\boldsymbol{x}_i), y_i)\}_{i=1}^{N}\Big) \right|, \tag{1}$$

where $\mathcal{C}$ refers to the calculation of a metric that evaluates the model's performance across the entire labeled validation dataset. In contrast, $\mathcal{M}$ estimates these metrics based solely on the network outputs and $\left|\cdot\right|$ refers to a desired measurement of the gap, such as L1 distance. After optimization, $\mathcal{M}$ can be applied directly to any test dataset, and we expect the estimated metrics to offer more accurate insights into performance, particularly in the presence of domain shifts.

MOVAL maps a set of logits to the final metrics through three key procedures: confidence score computation, model calibration, and metric calculation (Fig. 2b). To enable the calibration process, a tunable parameter is typically added to the confidence score computation. Different calibration techniques, such as Temperature Scaling (TS) (Guo et al., 2017), Differences of Confidences (DoC) (Guillory et al., 2021), and Average Thresholded Confidence (ATC) (Garg et al., 2022), define this parameter in distinct ways.

MOVAL extends these methods by altering the optimization objectives on a per-class basis—$\mathcal{C}$ and $\mathcal{M}$ calculate the metric for each class individually, applying class-specific parameters. This not only enhances accuracy estimation under class imbalance but is also a key step in expanding the MOVAL framework to metrics beyond accuracy and tasks beyond classification—areas that have been largely unexplored.

MOVAL's flexibility makes it more effective for supporting deployment decisions for machine learning models, by making the trade-offs among different performance metrics explicit. For example, in brain lesion segmentation from MRI data, accuracy alone is insufficient. To properly evaluate segmentation performance, tracking the F1-Score and focusing on pixel-level insights is critical. If minimizing missed lesions is a priority, sensitivity estimation is essential. Alternatively, when assessing post-processing techniques aimed at reducing false positives without significantly affecting overall performance, precision becomes the most appropriate metric to track.

2.2. Accuracy Estimation

The confidence score indicates the likelihood that the predicted label is correct and can estimate accuracy for a specific dataset. We first describe a general accuracy estimation framework based on MCP, which aligns well with existing confidence-based methods.

Given an input $\boldsymbol{x}_i$, the trained model f generates the output logit $\boldsymbol{z}_i = f(\boldsymbol{x}_i) \in \mathbb{R}^c$. The predicted probability vector $\boldsymbol{p}_i$ is then computed by applying the softmax function to $\boldsymbol{z}_i$, where $p_{ij} = \sigma(\boldsymbol{z}_i)_j = \frac{\mathrm{e}^{z_{ij}}}{\sum_{j=1}^{c} \mathrm{e}^{z_{ij}}}$ for the j'th class. The predicted class label y_i' corresponds to the class with the highest probability, given by $y_i' = \mathrm{argmax}_j(p_{ij})$. The confidence score q_i for this sample, modeled by MCP, can be calculated as $q_i = \max_j(p_{ij})$.

The confidence score is a number ranging from 0 to 1; the higher the number, the more likely the sample is correctly classified. Ideally, we want the confidence to have quantitative meaning that informs the decision-making process. Specifically, for the entire validation dataset $\mathcal{D}^V$, we calculate the average confidence with $\alpha' = \frac{1}{N}\sum_{i=1}^{N} q_i$, and expect it to reflect the real accuracy. In practice, we can apply a similar procedure to the test dataset under domain shifts $\mathcal{D}^{Te}$ and obtain the estimated accuracy as $\alpha'^{(Te)} = \frac{1}{M}\sum_{i=1}^{M} q_i^{Te}$ (Fig. 1).

However, directly computing confidence q_i over the softmax function for estimation can be problematic as the predicted probability cannot reflect

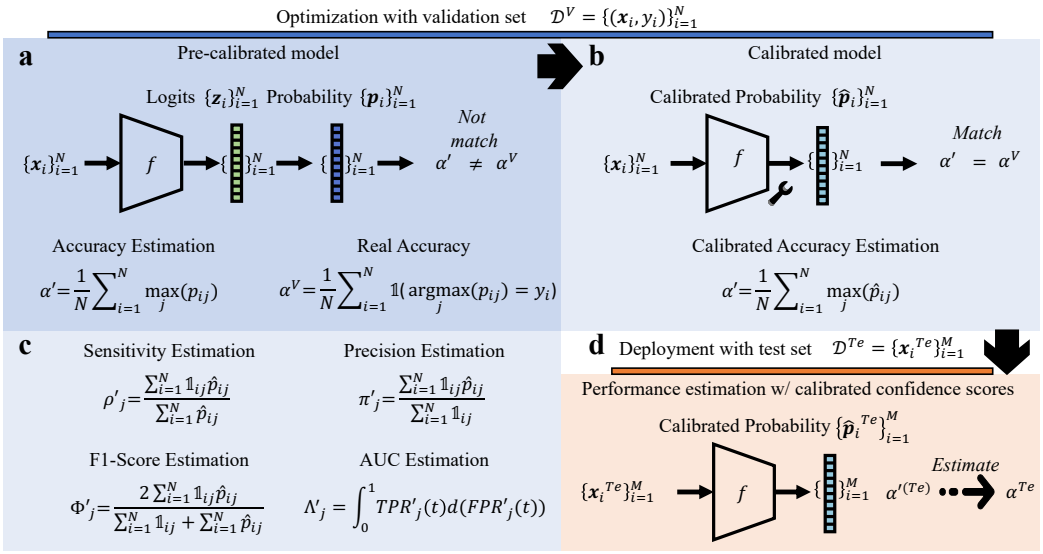


Figure 1: Detailed descriptions of the confidence calibration process. Here p_{ij} denotes the raw (uncalibrated) softmax probability that unit i belongs to class j, where i indexes a sample for classification and a pixel/voxel for segmentation, assuming a single volume is under consideration; after calibration, p_{ij} is replaced by the calibrated probability $\hat{p}_{ij}$ used to compute all reported metric estimates. **(a)**: Based on the model probabilities $\boldsymbol{p}_i$, multiple performance metrics can be estimated. As we illustrate with the example of accuracy, vanilla model probabilities are well known to be overconfident. Therefore, the estimation is prone to being too optimistic and not accurate. **(b)**: The calibration process adjusts the probabilities such that the estimation aligns with the real metrics on the validation dataset. **(c)**: The same principle can be applied to estimating other group-wise metrics. **(d)**: The calibrated probabilities are expected to produce more accurate estimations when exposed to domain shifts.

the real accuracy with modern neural networks (Guo et al., 2017). Therefore, we need to adjust the confidence to obtain the modified confidence $\hat{q}_i$, a process commonly known as model calibration. To achieve this, additional parameters are introduced into the confidence calculation process, enabling fine-tuning. While methods like TS, DoC, and ATC define these parameters differently, they all share the same optimization function. Specifically, a common approach is to use the validation accuracy, defined as $\alpha^V = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(y'_i = y_i)$[2], as a surrogate objective for α' to find the solu-

[2] $\mathbb{1}$ is an indicator function that equals 1 when the underlying criterion is met and 0 otherwise.

tion (Elsahar and Gallé, 2019; Guillory et al., 2021; Garg et al., 2022). Subsequently, the same calibration process is applied to the test dataset, with the expectation that the estimation results using the calibrated confidence $\hat{q}_i^{Te}$ will be more accurate.

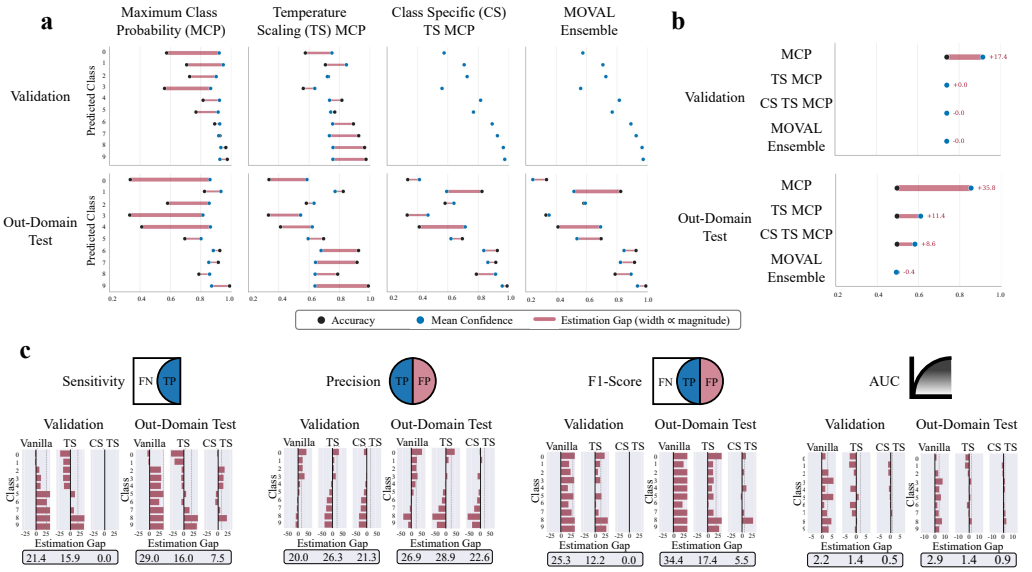


Figure 2: Class-wise calibration and multi-metric estimation with MOVAL. We show an estimation example on CIFAR10-LT, where sample counts decrease from class 0 to class 9. **(a)**: MOVAL facilitates class-wise calibration with class-specific parameters. Class-wise calibration allows for the alignment of estimated class-wise metrics and the true underlying counterparts, which can mitigate model biases caused by class imbalance. **(b)**: Overall accuracy estimation across different methods. By aggregating multiple confidence scores, MOVAL Ensemble achieves the most accurate results. **(c)**: In addition to accuracy estimation, which is provided by existing methods, class-wise alignment enables MOVAL to estimate class-wise metrics, thus providing a comprehensive evaluation.

We argue that existing confidence-based accuracy estimation methods struggle to generalize effectively to imbalanced datasets, which are common in medical image classification and segmentation. Prior studies have demonstrated that models trained on imbalanced data are prone to misclassifying minority class samples as belonging to the majority class, but rarely the reverse (Li et al., 2020). As a result, these models often achieve high overall accuracy but are underconfident in predicting minority class samples, while being overly confident for majority class predictions (Fig. 2c). Existing calibration methods apply a uniform probability-adjustment procedure across all

classes. This uniformity leads to higher estimation error for minority-class samples (e.g., TS MCP exhibits substantial estimation error on minority classes, as shown in Fig. 2a), which in turn limits their ability to generalize effectively (Fig. 2b). To address this, we introduce class-specific parameters into the performance estimation framework using adaptive optimization, reducing the confidence bias of neural networks toward minority classes (Fig. 3). Other group-wise metrics, such as sensitivity, precision, F1-Score and AUC, can be formulated similarly. The details are provided in the Suppl. Note.

3. Proposed Method: MOVAL

The MOVAL model consists of three key components. 1) Confidence Score Calculation: Given the network logits $\boldsymbol{z}_i$, the confidence score q_i is computed. 2) Model Calibration: The confidence score is adjusted to align estimated metrics with target metrics, producing the calibrated confidence score $\hat{q}_i$ from q_i. 3) Metrics Computation: Using $\hat{q}_i$, evaluation metrics for the entire dataset are computed, including calibrated probabilities $\hat{\boldsymbol{p}}_i$ for metrics beyond accuracy.

In following sections, we will present our main contribution, class-wise model calibration, using MCP for accuracy estimation. Subsequent sections will explore its extension to various metrics and types of confidence scores.

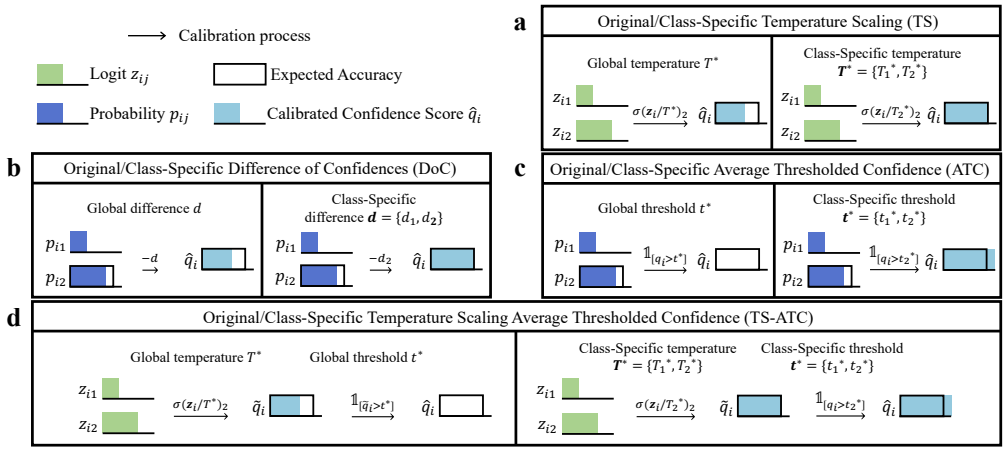


Figure 3: We extend existing model calibration methods to class-specific variants by introducing parameters for different classes. This enables class-wise calibration and improves estimation accuracy, especially when the model is trained with class-imbalanced data.

3.1. Class-Specific Temperature Scaling

The TS method is one of the most widely used calibration methods (Guo et al., 2017; Hinton et al., 2015). It introduces a temperature parameter to rescale the logits, with $q_i = \max_j(\sigma(\boldsymbol{z}_i/T)_j)$. To find the value of T that makes the estimated accuracy closely approximate the true accuracy, the optimal temperature T^* is determined using the validation dataset $\mathcal{D}^V$ as:

$$T^* = \underset{T}{\operatorname{argmin}} \left| \frac{1}{N} \sum_{i=1}^{N} q_i - \alpha^V \right| = \underset{T}{\operatorname{argmin}} \left| \frac{1}{N} \sum_{i=1}^{N} \max_j(\sigma(\frac{\boldsymbol{z}_i}{T})_j) - \alpha^V \right|. \quad (2)$$

To enable class-specific calibration, we extend the temperature parameter to a vector $\boldsymbol{T}$ with c elements, where the j'th element T_j corresponds to the j'th class. The calibrated confidence score is then computed as $q_i = \max_j(\sigma(\boldsymbol{z}_i/T_{y'_i})_j)$, with the optimal T_j^* determined by:

$$T_j^* = \underset{T_j}{\operatorname{argmin}} \left| \frac{1}{N_j} \sum_{i=1}^{N} \mathbb{1}_{ij} q_i - \alpha_j^V \right| = \underset{T_j}{\operatorname{argmin}} \left| \frac{1}{N_j} \sum_{i=1}^{N} \mathbb{1}_{ij} \max_k(\sigma(\frac{\boldsymbol{z}_i}{T_j})_k) - \alpha_j^V \right|, \quad (3)$$

where $N_j = \sum_{i=1}^{N} \mathbb{1}_{ij}$ and α_j^V is the accuracy for the j'th class on the validation set which is calculated as $\alpha_j^V = \frac{1}{N_j} \sum_{i=1}^{N} \mathbb{1}_{ij} \mathbb{1}(y'_i = y_i)$. This approach ensures that the optimized $\boldsymbol{T}^*$ results in $\hat{q}_i$ aligning with the class-wise accuracy (Fig. 3a).

3.2. Class-Specific Difference of Confidences

The DoC method (Guillory et al., 2021) adjusts q_i by subtracting the difference $d = \alpha' - \alpha^V$, which is the gap between the real accuracy α^V and the estimated results α':

$$\hat{q}_i = q_i - d. \quad (4)$$

A clipping operation between 0 and 1 may be necessary to ensure that the calibrated confidence is properly normalized. We introduce a class-specific DoC method, where we define a difference vector $\boldsymbol{d}$ with c elements, with the j'th element d_j representing the difference for the j'th class. The class-wise calibrated confidence score $\hat{q}_i$ is then computed as follows:

$$\hat{q}_i = q_i - d_{y'_i}, \tag{5}$$

where $d_j = \frac{1}{N_j} \sum_{i=1}^{N} \mathbb{1}_{ij} q_i - \alpha_j^V$ is the difference between the average predicted probability for samples predicted that they belong to class j and α_j^V. This approach insulates the probabilities of samples predicted as minority classes from the influence of majority classes, reducing bias (Fig. 3b).

3.3. Class-Specific Average Thresholded Confidence

The ATC method (Garg et al., 2022) is a heuristic yet effective approach for estimating model performance, consistently delivering top results across multiple robustness evaluation benchmarks (Garg et al., 2022; Baek et al., 2022). It predicts accuracy by calculating the proportion of samples with q_i values exceeding a learned threshold t. The threshold t is determined as follows, and the confidence score is then calculated as $\hat{q}_i = \mathbb{1}(q_i > t^*)$, where:

$$t^* = \underset{t}{\operatorname{argmin}} \left| \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(q_i > t) - \alpha^V \right|. \tag{6}$$

In the context of ATC, $\hat{q}_i$ is no longer a standard calibrated confidence score but instead a discrete value $\in \{0, 1\}$. Despite this, it can still be used for accuracy estimation as previously described. For class-specific calibration, we propose learning a set of thresholds represented by the vector $\boldsymbol{t}$, which has c elements, one for each class. The j'th element is determined by:

$$t_j^* = \underset{t_j}{\operatorname{argmin}} \left| \frac{1}{N_j} \sum_{i=1}^{N} \mathbb{1}_{ij} \mathbb{1}(q_i > t_j) - \alpha_j^V \right|. \tag{7}$$

We then calibrate the probabilities as $\hat{q}_i = \mathbb{1}(q_i > t^*_{y'_i})$ based on the predicted class y'_i (Fig. 3c). This approach discretizes the probabilities of samples predicted as different classes using distinct thresholds $t^*_{y'_i}$, effectively compensating for the bias introduced by class imbalance.

3.4. Class-Specific with Average Thresholded Confidence with Temperature Scaling

Confidence scores can be first calibrated using TS before being applied in the ATC process, referred to as TS-ATC (Garg et al., 2022). Specifically, the confidence score is initially calibrated with TS to obtain $\tilde{q}_i$, which satisfies

Eq. 2, and is then used in the ATC calibration process. By incorporating our class-specific modifications to both TS and ATC, we similarly derive the class-specific variant, CS TS-ATC (Fig. 3d).

3.5. Confidence Scores Calculation

In addition to MCP, various other confidence scores can be derived from logits $\boldsymbol{z}_i$, such as Entropy (Steinhardt and Liang, 2016), Doctor (Granese et al., 2021), and Energy (Liu et al., 2020b). While these techniques have not been extensively explored in the context of model evaluation, they have demonstrated improved performance in related tasks like anomaly detection (Chandola et al., 2009) and failure detection (Bernhardt et al., 2022). Therefore, we believe it would be valuable to extend the discussed model calibration methods to incorporate these confidence scores. To achieve this, Entropy and Doctor is calculated with:

$$q_i^{\text{Entropy}} = 1 + \sum_j p_{ij} \log_c p_{ij}, \quad q_i^{\text{Doctor}} = \frac{\sum_j p_{ij}^2 - 1}{\sum_j p_{ij}^2}. \tag{8}$$

The calibration of these two confidence scores is straightforward since they directly involve p_{ij} in their computation. As a result, TS calibration can be applied by adjusting T in the softmax process, where $p_{ij} = \sigma(\boldsymbol{z}_i/T)_j$. In contrast, Energy is calculated as follows:

$$q_i^{\text{Energy}} = T \log \sum_j e^{z_{ij}/T}. \tag{9}$$

It is important to note that we adjust the calculation to ensure alignment with the typical requirements of confidence scores, where a higher value indicates greater confidence. Additionally, we normalize the Doctor and Energy scores using $\max_i q_i$ and $\min_i q_i$[3], then clip the values between 0 and 1 to give them quantitative significance.

3.6. Calibration for Group-Wise Metrics

As previously discussed, calculating group-wise metrics requires calibrated probabilities $\hat{\boldsymbol{p}}_i$. Thus, we need to derive $\hat{\boldsymbol{p}}_i$ from the confidence score

[3]The calibrated confidence scores are normalized based on the range of uncalibrated confidence scores rather than calibrated scores; otherwise, the calibration process would not affect the estimation results.

$\hat{q}_i$, which involves filling in the probabilities for non-maximum classes. To achieve this, we propose two approaches: 1) Generate the probabilities as if they originated from the TS process. Specifically, we find the parameter that best fits the maximum probabilities (i.e., find T^* such that $\hat{q}_i = \max_j(\sigma\left(\frac{\boldsymbol{z}_i}{T^*}\right)_j)$) on a per-case basis and use it to estimate the non-maximum probabilities. 2) Equally distribute the remaining probability across the non-maximum classes, i.e., $\frac{1-q_i}{c-1}$ [4]. Initial experiments suggest that neither solution consistently outperforms the other. We adopt the second approach for all experiments, as it is more efficient.

Since the class-specific parameters may influence the group-wise metrics estimation of other classes, we optimize each parameter sequentially, starting with the class that has the highest number of predictions and proceeding in descending order. This approach helps minimize the overall estimation error on the validation dataset.

3.7. MOVAL Ensemble

Model ensembles, which average the probabilities from multiple models before making decisions, can effectively enhance model performance (Kamnitsas et al., 2018). We apply this concept to our task, referring to it as MOVAL Ensemble. In this approach, L different models are calibrated separately, each yielding $\hat{\boldsymbol{p}}_i^{(l)}$. The ensemble probabilities are then computed as follows:

$$\hat{\boldsymbol{p}}_i^{\text{Ensemble}} = \frac{1}{L}\sum_{l=1}^{L}\hat{\boldsymbol{p}}_i^{(l)}, \tag{10}$$

which are later used on test data to estimate metrics. We select different combinations for various tasks, summarized in Extended Table E.19.

3.8. Performance Estimation for Segmentation Tasks

We propose extending confidence-based model evaluation methods to segmentation. Unlike classification, segmentation evaluation emphasizes group-wise metrics that compare volumes rather than the entire dataset.

To this end, we slightly adjust the notation and denote the validation dataset as consisting of Z segmentation cases, represented as $\mathcal{D}^V = \{\mathcal{S}_z\}_{z=1}^{Z}$,

[4] For calibrated probabilities under ATC, the non-maximal values are simply set to 0.

where $\mathcal{S}_z$ contains all the training pairs for case z, with a total of n pixels. This is defined as $\mathcal{S}_z = \{(\boldsymbol{x}_{zi}, y_{zi})\}_{i=1}^{n}$, where $\boldsymbol{x}_{zi}$ is an image patch sampled from case z and y_{zi} is the ground truth class of its central pixel. We denote p_{zij} as the predicted probability of sample $\boldsymbol{x}_{zi}$ for the j'th class, and its calibrated version as $\hat{p}_{zij}$ (derived from the calibrated confidence score $\hat{q}_{zi}$). The estimated F1-Score is then computed using the predicted probabilities, analogous to the common soft DSC loss (Milletari et al., 2016). Specifically, the real F1-Score of $\mathcal{D}^V$ for the j'th class, denoted as Φ_j^V, is estimated as follows:

$$\Phi_j' = \frac{1}{Z}\sum_{z=1}^{Z}\frac{2\sum_{i=1}^{n}\mathbb{1}_{zij}\hat{p}_{zij}}{\sum_{i=1}^{n}\mathbb{1}_{zij}+\sum_{i=1}^{n}\hat{p}_{zij}}, \tag{11}$$

where now $\mathbb{1}_{zij} = \mathbb{1}(\text{argmax}_k(p_{zik}) = j)$. Notably, we do not calibrate the background class, as its performance remains consistently high due to the relative ease of classifying the majority of background samples. Adjusting the estimation to match this performance is difficult and can interfere with the calibration of foreground classes, as optimizing for the background class was found to destabilize the overall results.

3.9. Group-Wise Calibration

Since our approach performs calibration on a per-class basis, a natural concern is the requirement for a sufficient number of samples in each class, especially for minority classes. To address this limitation, we introduce a group-wise calibration strategy that allows class-wise calibration methods to operate at the group level, sharing class-specific parameters across classes and aligning metrics across multiple predicted classes. This approach, also used in (Li et al., 2024) for a different application, reduces reliance on validation datasets and proves more robust to various advanced training strategies.

As an example, we consider the CS TS of accuracy based on MCP (i.e., Eq. 3). For group-wise calibration, we assume a unique parameter $\widetilde{T}$ for calibrating all the samples belonging to a batch of B classes, $\boldsymbol{b} = [b_1, \cdots, j, \cdots, b_B]^\intercal$. The construction of $\boldsymbol{b}$ can be user-defined, but by default, the classes are grouped based on the number of samples predicted as class j, i.e., N_j. Recall that the accuracy for the j'th class can be estimated as $\alpha_j' = \frac{1}{N_j}\sum_{i=1}^{N}\mathbb{1}_{ij}\max_k(\sigma\left(\frac{\boldsymbol{z}_i}{\widetilde{T}}\right)_k)$. The group-wise calibration of CS TS is optimized with:

$$\widetilde{T}^* = \underset{\widetilde{T}}{\operatorname{argmin}} \left| \sum_{j \in \boldsymbol{b}} N_j (\alpha'_j - \alpha_j^V) \right|. \tag{12}$$

The estimation of other metrics (e.g., replacing α'_j with Φ'_j for F1-Score estimation) or the use of alternative estimation methods can be done in very similar ways. Notably, if B equals 1, group-wise calibration is equivalent to the vanilla CS calibration we introduced beforehand; if B equals the class number C when estimating accuracy, it is equivalent to the original class-agnostic calibration. All results in this paper are reported without group-wise calibration unless specified.

3.10. Optimization

We optimize the parameters of the MOVAL models using bound-constrained minimization algorithms. By default, we apply the Nelder-Mead simplex algorithm (Gao and Han, 2012) through the SciPy toolbox (Virtanen et al., 2020). The parameters are initially set to 1.0, and to improve the robustness of the optimization, we perform 5 additional initializations with different starting points if the initial optimization results are unsatisfactory. In practice, we optimize the class-specific parameter for class j only when there are both predicted instances of class j and ground truth instances labeled as class j in the validation dataset. Otherwise, optimization would be meaningless.

4. Experiments

4.1. Experimental Setup

In our experiments, we first train a classifier f with $\mathcal{D}^{Tr}$, then we make predictions on $\mathcal{D}^V$ and optimize the calibration model based on model output and the validation labels. Afterwards, we evaluate our model on $\mathcal{D}^{Te}$ and estimate the quality of the predicted results based on the pre-softmax network outputs $f(\boldsymbol{x}_i^{Te})$.

We train the classification and segmentation models using a well-established process based on ResNet (He et al., 2016) and U-Net (Ronneberger et al., 2015), following previous successful related efforts (Isensee et al., 2021; Russakovsky et al., 2015; Marrakchi et al., 2021). Detailed training configurations are provided in Suppl. Note.

4.2. Datasets

CIFAR10-LT. We resample the CIFAR10 training set (Krizhevsky et al., 2009) using an exponential class ratio to create an imbalanced dataset with $c = 10$ classes and an imbalance ratio of 100, following (Cao et al., 2019). The validation and test sets remain unchanged. To introduce domain shifts, we apply 95 synthetic corruptions from CIFAR10-C (Hendrycks and Dietterich, 2019a). Dataset details are provided in Extended Fig. C.12, with visualizations shown in Extended Fig. C.13.

CIFAR100-LT. We adopt similar strategies to create an imbalanced CIFAR100 dataset, which also has an imbalance ratio of 100. This dataset has $c = 100$ classes.

Skin lesion classification. The skin lesion datasets have $c = 7$ classes. We randomly select 70% of the data from HAM10000 (Tschandl et al., 2018) for the training dataset and reserve the rest for validation. We additionally collect 7 publicly available skin lesion datasets which were collected in different institutions including BCN (Combalia et al., 2019), VIE (Rotemberg et al., 2021), MSK (Gutman et al., 2016), UDA (Gutman et al., 2016), OTH (Rotemberg et al., 2021), D7P (Kawahara et al., 2018) and PH2 (Mendonça et al., 2013) mainly following (Yoon et al., 2019). The split is performed at the dataset level: training and validation use HAM10000 exclusively, and each of the seven test datasets is an independently collected cohort from a different institution, with no patient or image overlap with the source data.

Brain lesion MRI segmentation. We evaluate the proposed methods with brain lesion segmentation based on T1-weighted MR images using data from Anatomical Tracings of Lesions After Stroke (ATLAS) (Liew et al., 2022). This dataset is for binary segmentation ($c = 2$), with highly imbalanced classes. Images are registered to the MNI152 template (Brett et al., 2002) with a spacing of 1.00 mm $\times$ 1.00 mm $\times$ 1.00 mm.

This dataset contains 653 samples collected using 13 different scanners of different field strengths and manufacturers. The split is performed at the scanner level: all 88 cases from the chosen scanner are used exclusively for training and validation, and the remaining 565 cases, acquired on the other 12 scanners, are used exclusively as target test data, with no patient or scanner overlap between source and test. We create synthetic domains shifts by applying 83 distinct transformations, including spatial, appearance and noise operations to the validation data. We show some visualization of this dataset in Extended Fig. C.14. The detailed dataset statistics and

transformation implementation are summarized in Extended Table E.4 and Extended Table E.20.

Prostate MRI segmentation. We also evaluate on the task of cross-site prostate segmentation based on T2-weighted MR images following (Liu et al., 2020a). This is also a binary segmentation task ($c = 2$). We resample the images to a common resolution of 0.80 mm $\times$ 0.80 mm $\times$ 1.50 mm.

We use 30 cases collected by (Bloch et al., 2015) as the source domain. For the target domain, we utilize five datasets described in (Lemaître et al., 2015; Bloch et al., 2015; Litjens et al., 2014), which were collected from various institutions using scanners of different field strengths and manufacturers. As with the brain lesion data, the split is at the institution/scanner level. Synthetic domain shifts are generated using the same 83 transformations applied to brain lesions.

Cardiac MRI segmentation. We utilize cardiac MR images from the Multi-Centre, Multi-Vendor, and Multi-Disease Cardiac Segmentation (M&Ms) Challenge (Campello et al., 2021). The task is to segment the images into three classes ($c = 3$), including the left and right ventricle cavities and the left ventricle myocardium. The image volume is anisotropic and is therefore evaluated slice by slice, that is, in 2D. We resample the images to a common resolution of 1.25 mm $\times$ 1.25 mm $\times$ 10.00 mm.

This dataset contains short-axis cardiac MR images from 345 samples across 5 different sites. Following the clinical protocol, we segment two frames of the MRI, including the end-diastolic (ED) and end-systolic (ES) phases, totaling 690 images. We use data from one site (190 images) as the source domain and validate the trained model on other domains and synthesized domain shifts. The split is again at the site level. Compared with synthesized domain shifts in 3D segmentation, we exclude those that operate in the z-axis due to its anisotropic nature, resulting in 65 test conditions. We report the average estimation error across all three classes.

4.3. Class-Wise Performance Alignment

We first demonstrate the effectiveness of class-wise calibration by extending existing methods including TS, DoC, ATC and TS-ATC (which is a combination of TS and ATC) to the variants where performances are aligned in class-wise manner with class-specific (CS) parameters. These proposed variants are denoted as CS TS, CS DoC, CS ATC and CS TS-ATC, respectively.

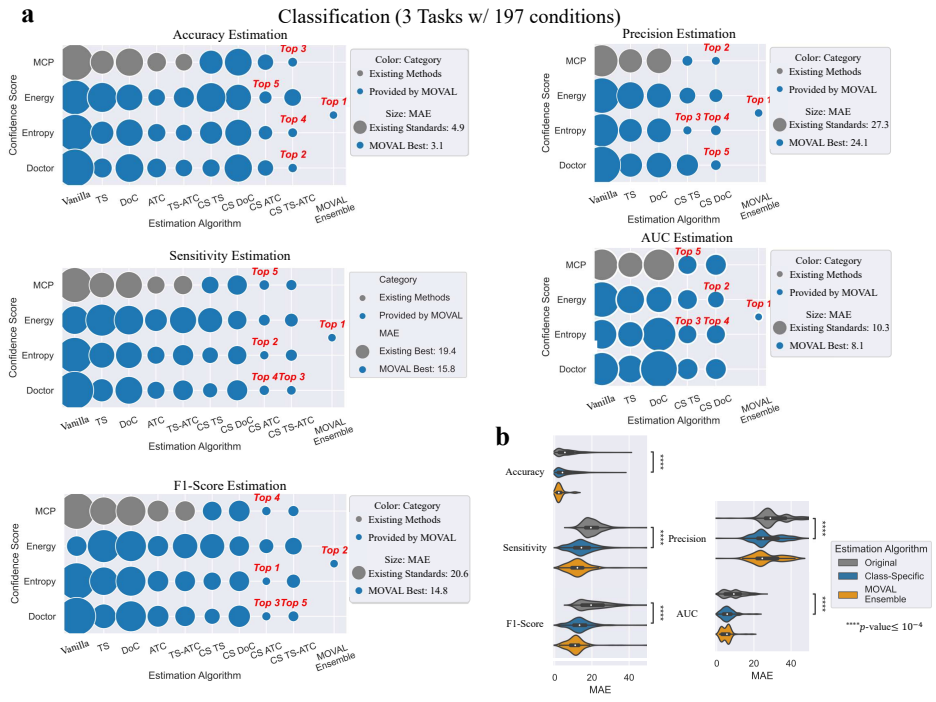


Figure 4: Class-wise calibration improves the estimation results for all metrics in classification tasks. **(a)**: The mean absolute error (MAE) between real metrics and estimated results based on different estimation methods for classification tasks. The size of the circles represents the MAE, with smaller circles indicating better performance. We report the average MAE for various tasks here. **(b)**: MAE comparison between the original estimation methods and our proposed class-specific variants for classification tasks. Paired t-tests were used for statistical analysis.

We use CIFAR10-LT to illustrate the calibration process for CS TS, as shown in Fig. 2b. TS adjusts confidence based on validation performance to align global accuracy with global confidence. However, this is achieved by sacrificing class-wise performance, where the majority classes remain overconfident (i.e., the mean confidence is higher than the actual accuracy), while the minority classes are underconfident.

We evaluate the estimation results across 3 classification tasks, covering 197 types of domain shifts, which include both synthetic shifts for CIFAR (Krizhevsky et al., 2009) and real-world acquisition shifts for skin lesion classification where images are gathered from various sites (Tschandl et al., 2018) (Fig. 7a, Extended Fig. C.17a). To note, we resample the CIFAR

training dataset to create a long-tailed distribution, making the evaluation more realistic, as this reflects the nature of real-world datasets (Liu et al., 2019; Cao et al., 2019). The models are based on a standardized version of ResNet (He et al., 2016). In order to evaluate the estimation results, we calculate the Mean Absolute Error (MAE) between the predicted and the real accuracy and report the overall results as the mean MAE across three tasks (Fig. 4a). Our proposed CS variants significantly improve the estimation results, reducing the overall MAE from 6.64 to 5.81 (Fig. 4b, Extended Fig. C.17b, paired t-test, $p = 1.01 \times 10^{-48}$). Specifically, the CS variants reduce the MAE for TS from 7.33 to 7.06 ($p = 0.02$), from 10.40 to 9.76 for DoC ($p = 1.71 \times 10^{-9}$), from 4.38 to 3.50 for ATC ($p = 6.42 \times 10^{-15}$) and from 4.45 to 2.93 for TS-ATC ($p = 1.94 \times 10^{-42}$). These results clearly highlight the importance of addressing class imbalance in the design of performance-prediction methods to mitigate biases introduced by the data.

Another advantage of class-wise alignment is its robustness to label shifts in test datasets, as it tracks performance separately for each class. This is especially beneficial when there are more samples from minority classes at test time. For example, when deploying the model on the CIFAR10-LT dataset, which includes only the two most minority classes, CS variants can reduce the MAE from 27.0 to 21.7 (Extended Table E.7, paired t-test, $p = 2.40 \times 10^{-120}$). Additionally, it is clear that the proposed CS variants provide more accurate performance estimation on the subsets defined by the model's predictions (Fig 2b).

4.4. Multi-Metric Estimation

The MOVAL framework allows to estimate other metrics beyond accuracy, delivering a thorough assessment based solely on the network outputs. Those metrics like the F1-Score are typically calculated on a per-class basis. Consequently, we compute the MAE for each individual class and report the MAE for a case along with its average MAE across all classes (Extended Fig. C.16a).

Class-wise performance alignment naturally suits the estimation of group-wise metrics, as demonstrated with CIFAR10-LT in Fig. 2a. Existing calibration methods focused on accuracy can reduce overconfidence and, in turn, lower estimation error for group-wise metrics (NannyML), but they do not achieve optimal results for those metrics. Although our CS variants may not fully align performance for precision and AUC, they greatly enhance calibration, thereby reducing estimation errors during testing.

Similarly to the accuracy estimation evaluation, we report results across three classification tasks (Fig. 4a). Our CS variants show clear advantage than existing metrics (Fig. 4b, Extended Fig. C.17b), reducing the MAE from 21.25 to 13.95 for F1-Score (paired t-test, $T(3151) = 79.16, p \leq 10^{-4}$), from 20.40 to 14.96 for sensitivity ($T(3151) = 46.89, p \leq 10^{-4}$), from 31.23 to 26.99 for precision ($T(1575) = 64.76, p \leq 10^{-4}$), and from 9.19 to 6.04 for AUC ($p = 1.25 \times 10^{-125}$).

4.5. MOVAL for Segmentation Tasks

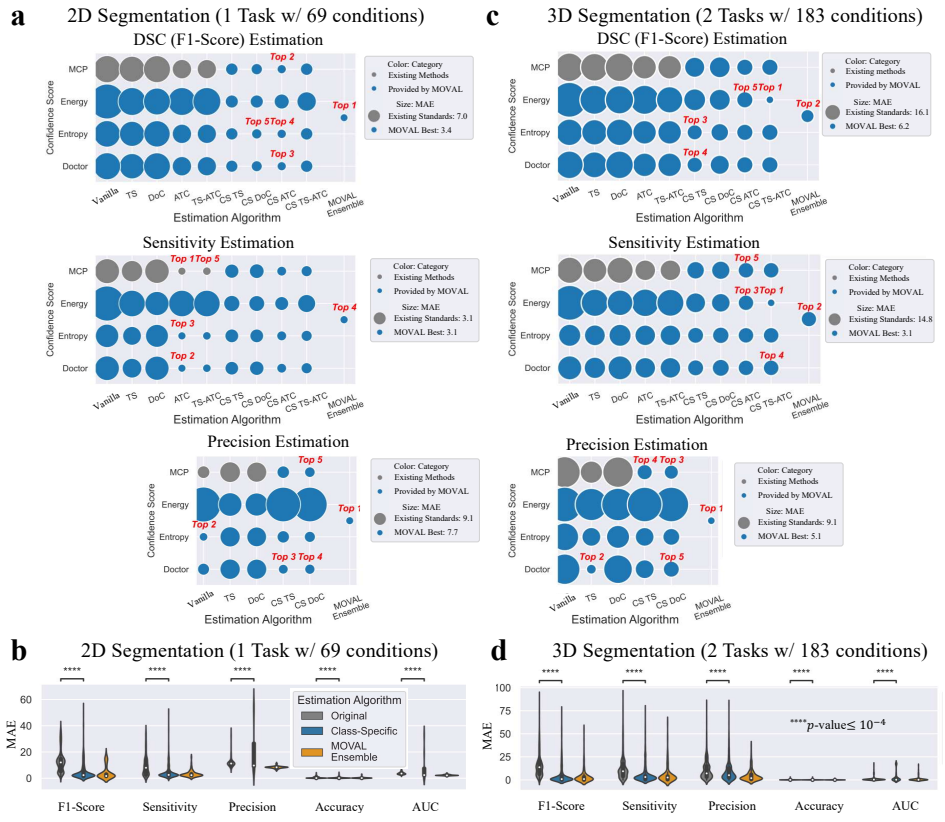


Figure 5: Class-wise calibration is even more crucial for segmentation tasks. **(a, c)**: MAE between real metrics and estimated results from different methods for 2D and 3D segmentation tasks. **(b, d)**: MAE comparison between the original estimation methods and proposed class-specific variants for 2D and 3D segmentation tasks. Statistical tests are conducted using paired t-tests.

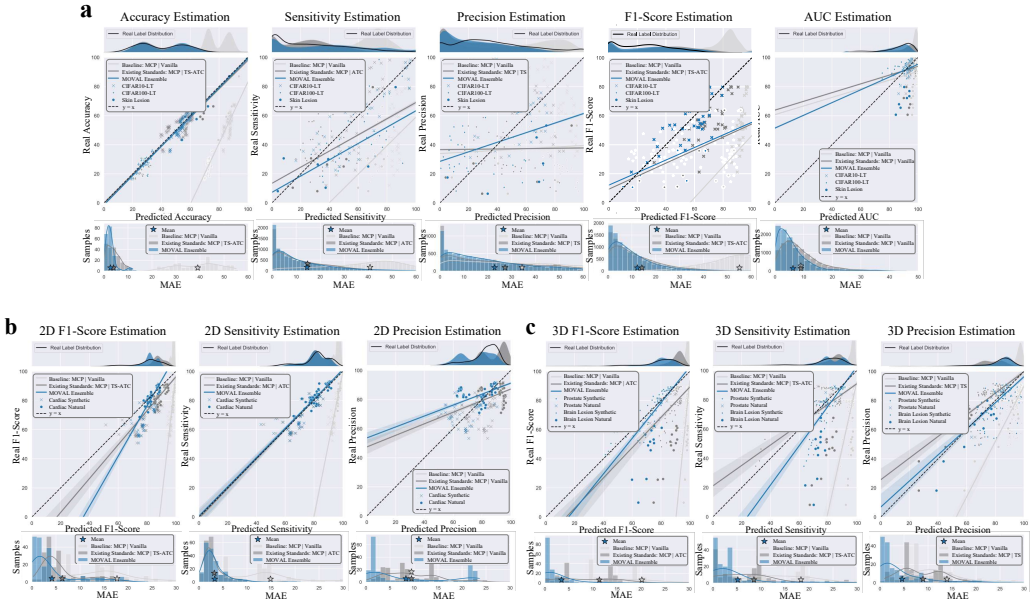


Figure 6: Agreement between predicted and real metrics for different tasks. For each subfigure: Top: Predicted and real label distribution. We benchmark MOVAL ensemble (blue) against MCP baselines (light gray) and the best-performing existing method (dark gray). Middle: Predicted versus real metrics. Each point represents the model's performance on unseen target data under specific conditions. Bottom: Histogram of MAE for different estimation methods. **(a)**: Classification tasks. **(b)**: 2D segmentation tasks. **(c)**: 3D segmentation tasks.

The MOVAL framework supports the evaluation of both 2D and 3D segmentation tasks. It extends to semantic segmentation by treating it as pixel-level dense classification. A key distinction between evaluating segmentation and classification is that segmentation involves comparing 2D slices or 3D volumes, rather than assessing the entire dataset. With this adjustment, all the previously mentioned methods can be easily adapted for segmentation tasks.

We evaluate 2D segmentation on the task of cardiac MRI (Campello et al., 2021) segmentation and 3D segmentation on prostate MRI (Liu et al., 2020a) and brain lesion MRI (Liew et al., 2022) segmentation. For all segmentation tasks, we apply transformations as synthetic domain shifts (Extended Table E.20) and collect images from different scanners to represent real domain shifts. The models are built upon a standardized version of 3D U-Net (Ron-

neberger et al., 2015).

We present the evaluation results of the average MAE between real and synthesized scenarios (Fig. 5a, Fig. 5c). We omit accuracy and AUC evaluations, as these metrics often approach 100% for segmentation tasks (Extended Fig. C.15c, Extended Fig. C.15d). Our CS variants demonstrate even greater advantages over existing standards compared to classification results (Fig. 5b, Fig. 5d). This may be due to previous model calibration methods being heavily biased toward the background class, a consequence of class imbalance.

We focus on evaluating the F1-Score (i.e., DSC), the most commonly used metric for segmentation tasks (Fig 7c, Extended Fig. C.18a, Extended Fig. C.18b). CS variants greatly enhance confidence-based estimation of the F1-Score, lowering the MAE from 12.88 to 4.73 for 2D segmentation (paired t-test, $p = 1.03 \times 10^{-227}$) and from 15.41 to 4.69 for 3D segmentation ($T(2927) = 72.55$, $p \leq 10^{-4}$). Similar conclusions also apply to sensitivity and precision (Extended Fig. C.16b, Extended Fig. C.16c). For sensitivity, CS variants decrease the MAE from 9.55 to 4.31 for 2D segmentation ($p = 2.57 \times 10^{-145}$) and from 12.78 to 5.75 for 3D segmentation($p = 1.59 \times 10^{-320}$). For precision, they reduce the MAE from 11.25 to 9.36 for 2D segmentation ($p = 1.94 \times 10^{-47}$) and from 9.38 to 5.76 for 3D segmentation ($p = 2.97 \times 10^{-29}$). Note that we exclude energy-based evaluation for precision due to its consistently poor performance in segmentation tasks.

4.6. Confidence Scores Ensemble

MOVAL enhances existing model calibration methods by integrating confidence scores beyond MCP and utilizing ensemble techniques for improved results. While traditional model evaluation methods rely on MCP for its simplicity and straightforward quantitative meaning, MOVAL advances this by integrating other confidence scores, such as Entropy (Steinhardt and Liang, 2016), Doctor (Granese et al., 2021), and Energy (Liu et al., 2020b). These confidence scores have shown enhanced outlier detection capabilities due to their improved ability to capture information from non-maximum predictions (Fig. 2a). Additionally, akin to model ensembles in discriminative models, we investigate the ensemble of the best methods to deliver robust and effective solutions.

We begin by investigating the discrepancies between different confidence scores and estimation algorithms, as model ensembles are effective only when the methods exhibit substantial diversity (Kamnitsas et al., 2018; Ganaie

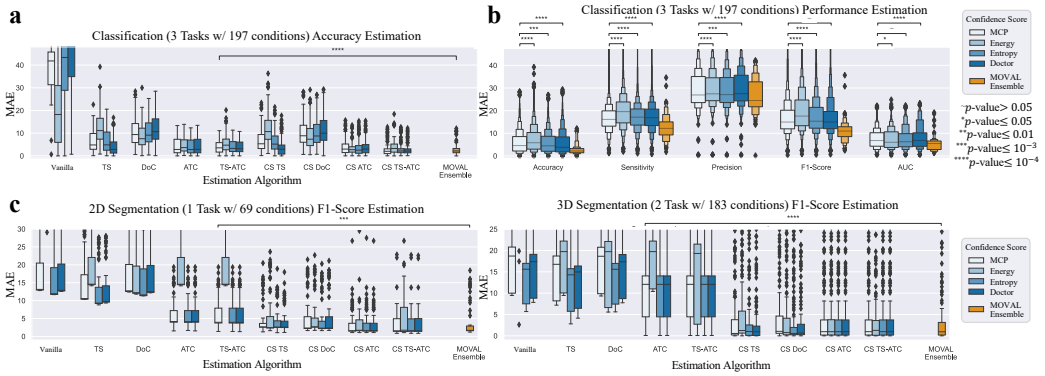


Figure 7: MOVAL Ensemble achieves state-of-the-art performance. **(a, c)**: MAE comparison between proposed algorithms and existing standards (Best-performing existing method versus MOVAL Ensemble, paired t-test, $p = 1.05 \times 10^{-7}$ for accuracy estimation of classification tasks, $p = 3.54 \times 10^{-4}$ for F1-Score estimation of 2D segmentation tasks, $p = 4.28 \times 10^{-27}$ for F1-Score estimation of 3D segmentation tasks, see the results of other metrics in Extended Fig. C.16 and the results of individual datasets metrics in Extended Fig. C.17 and Fig. C.18). **(b)**: MAE comparison between estimation algorithms based on MCP and other confidence scores for classification tasks (see Extended Fig. C.19 for other tasks).

et al., 2022). We find that MCP yields significantly different results compared to other confidence scores in 13 out of 15 conditions for classification tasks (Fig. 7b), 8 out of 9 conditions for 2D segmentation, and 9 out of 9 conditions for 3D segmentation (Extended Fig. C.19b). Additionally, we observe that different existing estimation algorithms also display clear differences under varying conditions (Extended Fig. C.19a).

Next, we combine three estimation methods into an ensemble, known as MOVAL Ensemble, and apply it to all the test conditions previously mentioned. Notably, we select different combinations for various tasks and metrics, with most of the chosen methods being from our proposed CS variants (Extended Table E.19). The resulting MOVAL Ensemble consistently ranks among the top performers across all tasks and metrics, achieving the best results in 7 out of 11 scenarios (Fig. 4a, Fig. 5a, Fig. 5c). Compared to state-of-the-art methods, the MOVAL Ensemble demonstrates significant advantages in most cases (Fig. 7a, Fig. 7c, Extended Fig. C.16a, Extended Fig. C.16b, Extended Fig. C.16c).

For classification tasks, compared to the best-performing existing methods, MOVAL Ensemble reduces the MAE from 4.86 to 3.09 for accuracy

(paired t-test, $p = 1.05 \times 10^{-7}$), from 20.63 to 14.81 for F1-Score ($p = 6.04 \times 10^{-18}$), from 19.37 to 15.84 for sensitivity ($p = 3.54 \times 10^{13}$), from 27.32 to 24.10 for precision ($p = 1.43 \times 10^{-6}$), and from 10.33 to 8.08 for AUC ($p = 8.18 \times 10^{-7}$). Specifically, we observe that MOVAL Ensemble significantly enhances performance especially when there are fewer classes (e.g., 10 classes, Extended Fig. C.17a). For example, MOVAL Ensemble reduces the MAE from 18.85 to 9.45 for sensitivity estimation on CIFAR10-LT (Extended Table E.5, $p = 1.53 \times 10^{-93}$).

For segmentation tasks, MOVAL Ensemble consistently enhances F1-Score estimation accuracy across various tasks (Extended Fig. C.18a). Specifically, compared to the best-performing existing methods, MOVAL Ensemble reduces the MAE from 7.05 to 3.44 for 2D segmentation (paired t-test, $p = 3.54 \times 10^{-4}$) and from 16.13 to 7.74 for 3D segmentation ($p = 4.28 \times 10^{-27}$). For precision estimation, MOVAL Ensemble decreases the MAE from 9.06 to 7.68 for 2D segmentation ($p = 8.54 \times 10^{-7}$) and from 9.09 to 5.12 for 3D segmentation ($p = 1.53 \times 10^{-10}$). MOVAL Ensemble also proves effective for sensitivity estimation in 3D segmentation, reducing the MAE from 14.76 to 10.43 ($p = 1.09 \times 10^{-4}$). Although the improvement in sensitivity for 2D segmentation is not significant (3.07 versus 3.09, $p = 0.80$), MOVAL Ensemble demonstrates notably better results for the most minority class (i.e., left cavities segmentation), reducing the MAE from 2.93 to 2.87 ($p = 1.4 \times 10^{-2}$). In summary, MOVAL Ensemble serves as a reliable tool for estimating key metrics such as F1-Score, sensitivity, and precision in segmentation tasks.

To gain a deeper understanding of the estimation results, we examine the actual metrics and estimation outcomes of MOVAL Ensemble, the best-performing existing methods and the MCP baselines (Fig. 6a, Fig. 6b, Fig. 6c). The predicted performance by the MCP baseline consistently exceeds the actual performance across all metrics due to the common overconfidence in neural networks (Guo et al., 2017). Existing standards can mitigate the discrepancy between estimated and actual metrics, while MOVAL Ensemble performs better at estimating the underlying label distribution (Top Fig. 6a). The estimation results of MOVAL Ensemble are closer to the "predicted metric = real metric" line over a broad range, indicating that it produces less estimation error across varying levels of difficulty (Middle Fig. 6c). The distribution of MAE produced by MOVAL Ensemble is more concentrated at lower values, indicating that it achieves smaller estimation errors in a greater number of cases (Bottom Fig. 6c).

4.7. Robustness under Training Conditions

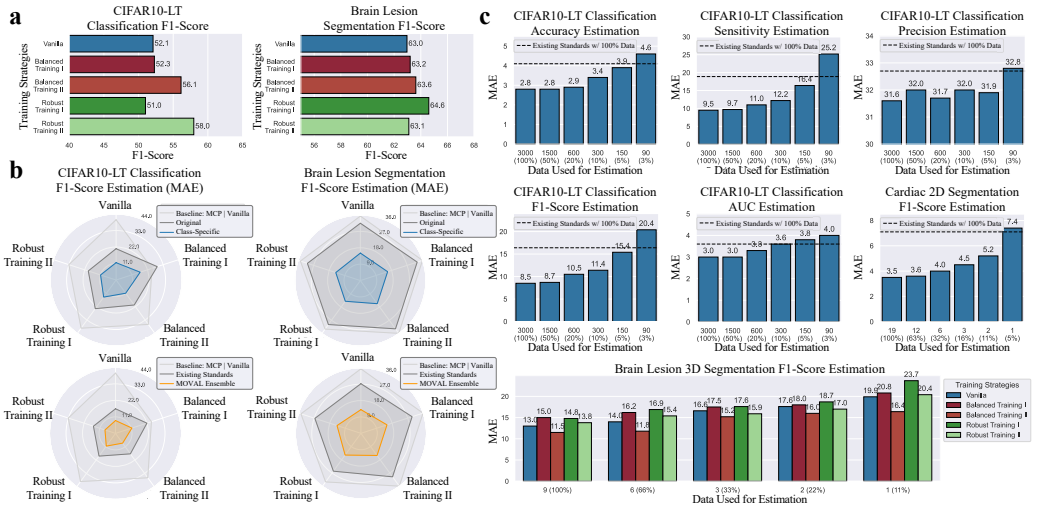


Figure 8: MOVAL Ensemble maintains stability in realistic settings. **(a)**: Classification and segmentation results when models are trained with advanced training strategies designed for class imbalance or domain generalization (Classification, Balanced Training I: label-distribution-aware margin (LDAM) loss (Cao et al., 2019), Balanced Training II: deferred re-weighting (DRW) training schedule (Cao et al., 2019), Robust Training I: cutout (DeVries and Taylor, 2017), Robust Training II: RandAugment (Cubuk et al., 2020); Segmentation, Balanced Training I: asymmetric large margin loss (Li et al., 2020), Balanced Training II: asymmetric focal loss (Li et al., 2020), Robust Training I: global intensity non-linear augmentation (GIN) technique (Ouyang et al., 2022a), Robust Training II: mixup (Zhang et al., 2017)). **(b)**: Top: MAE comparison between the original estimation methods and our proposed class-specific variants with various training strategies. Bottom: MAE comparison between best-performing existing method and MOVAL Ensemble with various training strategies. **(c)**: MAE achieved by MOVAL Ensemble when optimized with different amounts of validation data.

We evaluate the compatibility of MOVAL Ensemble with models trained using advanced strategies, specifically Balanced Training to address class imbalance (Cao et al., 2019; Li et al., 2020) and Robust Training to improve model generalization (DeVries and Taylor, 2017; Cubuk et al., 2020; Ouyang et al., 2022a; Zhang et al., 2017). These strategies are commonly used in real-world applications (Suppl. Note) and are applied to CIFAR10-LT and brain lesion segmentation. Most strategies improve model performance on unseen test data across various domain shifts (Fig. 8a).

We then apply the MOVAL framework to these models to evaluate its effectiveness in these realistic settings (Fig. 8b). Previous conclusions remain valid for most scenarios involving advanced training strategies. For example, the proposed CS variants yield better results than the original estimation algorithms for F1-Score (Top Fig. 8c). Additionally, MOVAL Ensembles consistently outperform best-performing existing methods by a substantial margin when models are trained using various strategies (Bottom Fig. 8c).

In practice, certain balanced training strategies can significantly alter the distribution of predicted probabilities and complicate the confusion matrix. This creates challenges for performance estimation, particularly accuracy estimation, since calibration methods rely on these probabilities and metrics. To address this, we propose group-wise calibration, which aligns accuracy across groups of samples from different classes which share similar statistics instead of focusing on individual classes. This approach can stabilize accuracy estimation in advanced training settings (Extended Table E.13, E.14).

4.8. Stability with Varying Amounts of Data

We evaluate MOVAL Ensemble with varying amounts of available validation data to provide practical application guidelines. We optimize MOVAL Ensemble for default-trained models on CIFAR10-LT and cardiac segmentation with different data quantities, and for all advanced training strategies in brain lesion segmentation. This simulates a realistic setting, and we present the average results from five runs.

We find that MOVAL Ensemble is not sensitive to the amount of validation data, performing well even with limited data (Fig. 8c). For classification tasks like CIFAR10-LT, it requires as little as 5% of the validation data to achieve performance comparable to best-performing existing methods for most metrics. Approximately 60 samples per case are needed to obtain the most stable results. A similar conclusion applies to segmentation tasks, where 12 slices are required for 2D segmentation and 6 volumes for 3D segmentation to achieve stable results. MOVAL Ensemble performs best with Balanced Training II (asymmetric focal loss) (Li et al., 2020), likely due to reduced variance in predicted probabilities for background samples, minimizing probability shifts under domain changes and enhancing estimation effectiveness.

Additionally, as previously mentioned, we provide group-wise calibration for classification, which assumes that classes with similar statistics should share similar optimal calibration parameters, thereby reducing the number

of parameters. This approach could further decrease the need for validation samples (Extended Fig. C.20).

4.9. Informative Indicators at Different Levels

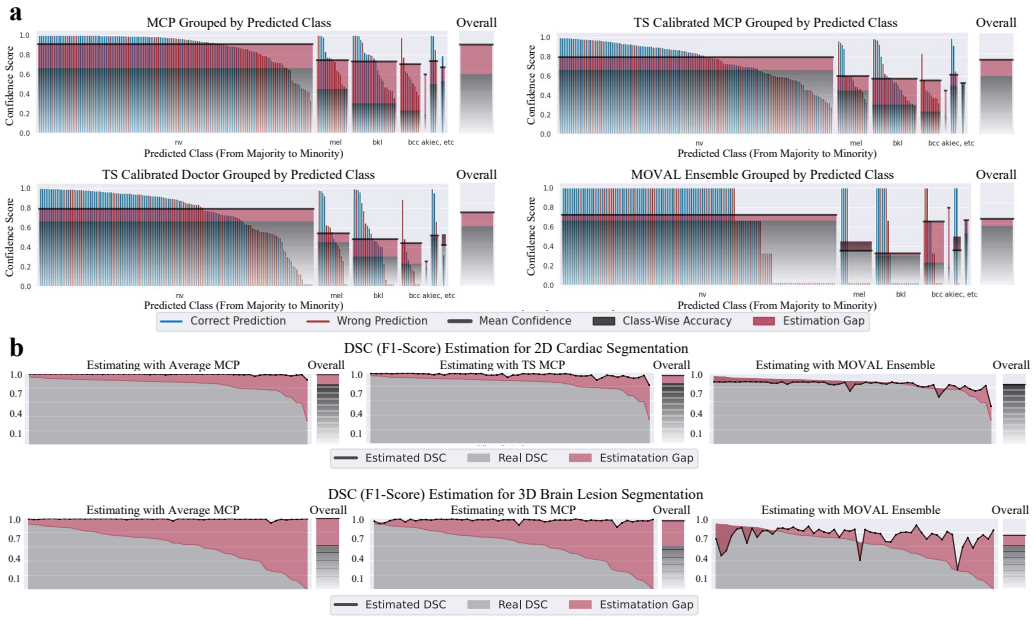


Figure 9: Fine-grained performance indicators at class and instance levels. **(a)**: Alignment of confidence scores and class-wise correctness for skin lesion classification. **(b)**: Top: Alignment of instance-wise DSC estimation and real DSC for 2D cardiac segmentation. Bottom: Alignment for 3D brain lesion segmentation.

Lastly, we test if the calibrated confidence of MOVAL Ensemble can be informative except the use of estimating global metrics. Specifically, the confidence score is calibrated to match accuracy for skin lesion classification and to match the F1-Score for cardiac and brain lesion segmentation.

MOVAL provides more precise feedback at the class-wise level (Fig. 9a). By evaluating model performance on a per-class basis, the MOVAL Ensemble not only offers an overall estimation advantage but also demonstrates improved alignment with class-wise performance. This allows for a more accurate assessment of sample groups and facilitates comparisons between different groups.

MOVAL also provides more precise feedback at the instance-wise level for segmentation tasks (Fig. 9b). By optimizing MOVAL parameters using

averages from various cases, we can also evaluate performance for individual cases. While existing accuracy-calibrated methods tend to be overly optimistic for segmentation tasks, MOVAL Ensemble offers a more accurate reflection of segmentation results across different cases.

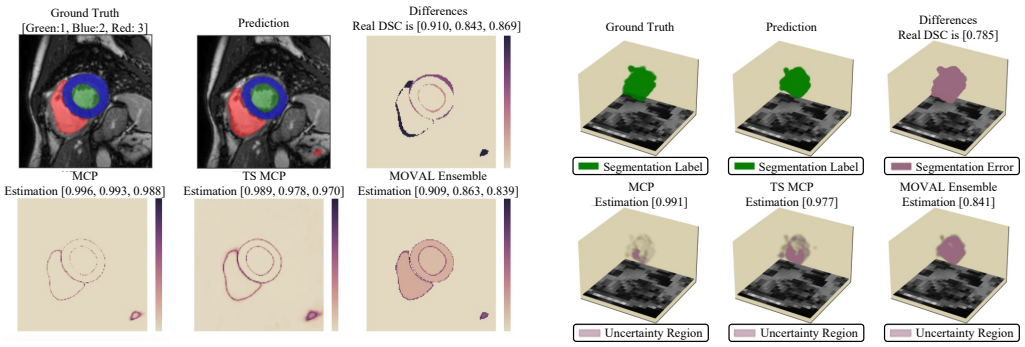


Figure 10: Fine-grained performance indicators at pixel level. Left: Visualization of calibrated confidence scores for 2D cardiac segmentation. Right: 3D scatter plots of segmentation error and uncertainty regions from calibrated confidence scores.

MOVAL provides prompt performance feedback at the pixel-wise level (Fig. 10). We can assess model performance using the calibrated confidence map. For example, MOVAL Ensemble can highlight classes with relatively poor performance (Left Fig. 10). It also offers quantitative insights into uncertain regions and emphasizes areas that need closer attention (Right Fig. 10). This behavior arises because MOVAL Ensemble integrates multiple confidence scores, whose combined confidence values tend to concentrate near 0 and 1 (see Extended Fig. C.21), potentially making it easier to detect anomalies without carefully crafting a threshold.

5. Discussion and Conclusion

Reliable deployment is crucial for fully realizing the potential of modern machine learning models in real-world applications. Efforts to regulate the development of machine learning models to support safer deployment have been on the rise (Rivera et al., 2020; Vasey et al., 2022), yet a standardized tool to easily monitor performance without labels is still lacking. To fill this gap, we introduce MOVAL, a new confidence-based model evaluation tool that estimates model performance without needing ground truth. By extending confidence-based approaches to a variety of confidence scores,

metrics, and tasks, we achieve promising results in diverse scenarios, from cross-site skin lesion classification to cross-scanner MRI segmentation. We believe this tool can enhance existing machine learning workflows with its easy-to-use and transparent design, as a practical post-training monitoring and performance-estimation aid that can inform, rather than by itself guarantee, safe deployment of machine learning in real-world settings. For instance, it can help determine whether a pre-trained model should be deployed by evaluating if its estimated performance on target data meets the acceptance criteria for the deployment environment, complementing rather than replacing standard validation and regulatory processes.

The key feature of MOVAL is its ability to track a wide range of complementary and informative metrics, including accuracy, F1-Score, sensitivity, precision, and AUC, all within a unified framework. Confidence scores have shown notable accuracy results for balanced benchmark datasets like CIFAR10, but they do not address class imbalance, a common issue in medical imaging databases. MOVAL's unique contribution is its extension of the standard calibration process to incorporate class-wise alignment. This approach not only reduces calibration bias caused by class imbalance but also naturally adapts the framework to estimate group-wise metrics beyond accuracy (Fig. 2). Furthermore, unlike existing methods, class-wise calibration enables MOVAL to track performance at the class level, making it less sensitive to variations in label distribution during testing compared to the training distribution.

MOVAL is highly versatile and can be easily applied to a variety of tasks, including classification, 2D segmentation, and 3D segmentation. Through extensive evaluation across six tasks and 6,572 test scenarios (Fig. 4, Fig. 5), we demonstrate that MOVAL offers significant advantages over existing standards. Designed with distinct modular components, the framework is flexible enough to integrate additional confidence scores or other performance estimation methods. In the future, we plan to include more recent successes in confidence-based accuracy estimation, such as Agreement-on-the-Line (Baek et al., 2022), ATC-distance (Roschewitz and Glocker, 2023), Confidence Optimal Transport (COT) (Lu et al., 2023) and Weighted Flips (WF) (Press et al., 2024) into our framework.

Building on our successful approaches, we introduce MOVAL Ensemble, which combines the strengths of various estimation methods by averaging the calibrated probabilities. This ensemble achieves top performance in most of the scenarios we tested. Specifically, compared to the best-performing

existing methods, it reduces classification accuracy estimation error by 36.4%, 2D segmentation F1-Score estimation error by 51.2%, and 3D segmentation F1-Score estimation error by 52.0%. While we provide MOVAL Ensemble as a comprehensive solution, users can also customize it by selecting the confidence scores they wish to combine for their specific tasks.

We illustrate the utility and effectiveness of MOVAL Ensemble in realistic settings and use cases. Our findings indicate that MOVAL Ensemble performs well with various advanced model training strategies designed to enhance model robustness against class imbalance and domain shifts. Despite these strategies significantly altering model behavior, our method consistently produced stable estimation results across most scenarios. We also demonstrate that MOVAL Ensemble has minimal requirements for the validation set and offer practical suggestions for designing experiments in realistic settings (Fig. 8). Furthermore, we show that calibrated confidence can be valuable for fine-grained performance analysis (Fig. 9).

MOVAL is primarily designed to estimate model performance under acquisition shifts in medical imaging, where images are typically obtained under varying conditions (Castro et al., 2020). While it can be used in other domain shift scenarios, users should ensure the following for optimal performance (though not all can be avoided): 1) No annotation shifts between the validation and test sets, meaning the labeling concept remains consistent across datasets. 2) No population shifts between validation and test sets, meaning there should be no population-related covariate shifts in the input features. This requires ensuring that samples are collected from similar regions and demographics, with comparable disease patterns or evolving environments. 3) Although MOVAL is not highly sensitive to prevalence shifts (e.g., differing class ratios between validation and test), performance is generally optimal when these ratios align because the learned calibration parameters are closely tied to the confusion matrix, which is more likely to remain correlated between validation and test with identical prevalence.

The precision estimate is inherently limited as it can only be calibrated from positive predictions, ignoring the informative but large pool of false negatives. This makes it less reliable than sensitivity or F1-Score, which leverage a wider range of prediction data for calibration. Consequently, MOVAL should not be the sole metric for applications prioritizing precision. This limitation matters most in clinical use cases where false positives drive unnecessary interventions or follow-up procedures (e.g., biopsy referral, repeat scans). Based on our experiments (Fig. 4, Fig. 7), MOVAL's precision esti-

mate is reasonably reliable when validation and target prevalence are similar and the positive class is not extremely rare; otherwise, we recommend treating it as indicative only and confirming with a modest labeled sample or the more robust F1-Score or sensitivity estimates.

MOVAL relies on a calibration parameter learned from the validation dataset and applied, without adjustment, to unlabeled test data—an assumption, not a guarantee, since the ideal parameters would require test-domain labels. We therefore describe MOVAL as a practical monitoring and performance-estimation tool that can support deployment decisions, rather than one that ensures safe deployment under arbitrary shift. Its validity depends on the shift type. Under acquisition shifts, where the scanner, protocol, or site changes but the model's input-output relationship is largely preserved, we find empirically that validation-domain calibration remains informative (Section 4). Under prevalence or class-ratio shifts, MOVAL's class-wise design is comparatively robust. Under more severe shifts, where the confidence-correctness relationship changes qualitatively, calibration may not transfer, and we recommend a small labeled test sample to check and, if necessary, recalibrate.

Estimating performance from unlabeled test data is not identifiable in general without assumptions linking validation-domain calibration to test-domain correctness, and no distribution-free guarantee should be expected under arbitrary shift. In the acquisition-shift regime we target, a bounded divergence between validation and target confidence-score distributions, a stable class-conditional label distribution, or an identifiable label-shift structure would let one bound the estimation error; access to multiple labeled source domains would offer a further, more permissive route to a worst-case bound. We leave a formal treatment of these assumptions to future work. Several other directions for future work remain to enhance MOVAL's applicability and robustness. First, integrating test-time calibration, where the model is adjusted on-the-fly using target domain data (Wang et al., 2020a; Hu et al., 2024) or information from multiple domains (Yu et al., 2022), could improve adaptability, though such methods typically require retraining or access to multiple source domains that may not be available at deployment; we view them as complementary rather than as direct substitutes for MOVAL's post-training, single-source setting. Second, more extensive evaluation across a wider range of applications (e.g., ultrasound), tasks (e.g., 3D classification), and model architectures is warranted. In particular, broader validation on modern foundation models, multimodal models, and additional modalities

such as CT, ultrasound, and video remains future work.

For practitioners, we summarize the practical guidance that follows from our ablations (Fig. 8). Calibration set size: as few as 5% of the validation data (approximately 60 samples per class for classification) is sufficient for stable estimates; for segmentation, about 12 slices (2D) or 6 volumes (3D) are typically needed. Class-wise versus group-wise calibration: class-wise calibration is preferred when each class has enough validation samples (rule of thumb, tens of samples per class); when some classes are rare, group-wise calibration, which shares parameters across a small number of groups, achieves comparable accuracy with substantially fewer validation samples per group. Interpreting disagreement across confidence scores: when individual confidence scores (MCP, Entropy, Doctor, Energy) diverge substantially in their estimates, this itself is a useful diagnostic that the shift may be more severe than an acquisition shift, and we recommend defaulting to the MOVAL Ensemble estimate, which is more robust than any single score, in these cases. Triggering manual review: because MOVAL cannot certify that validation-domain calibration transfers, we recommend manual review or additional labeling whenever the estimated performance drops below the acceptance threshold for the application, when confidence-score disagreement is high, or when the estimated prevalence distribution departs substantially from the validation distribution.

MOVAL is parameter-efficient, at most requiring only c (the number of classes) $\times$ L (the number of ensemble choices) parameters. This leads to computationally efficient and reproducible optimization and inference processes, especially compared to methods that require references during inference (Valindria et al., 2017) or an additional neural network (Robinson et al., 2018). We provide a Python API for all the algorithms described above, designed for use and enhancement with every trained model. Given the importance of time efficiency in machine learning applications like autonomous driving and surgical image analysis, we believe our approach can serve as a practical, real-time monitoring aid that supports, but does not by itself guarantee, safe deployment in safety-critical applications.

Acknowledgments

This project received funding from the ERC under the EU's Horizon 2020 research and innovation programme (grant No 757173). MR is funded by an Imperial College President's PhD scholarship and a Google PhD Fellowship.

Author contributions

Conceptualization: ZL, MI, BG. **Methodology:** ZL, MR, CO. **Software:** ZL. **Validation:** ZL. **Formal analysis:** ZL. **Data curation:** ZL, CO. **Writing – original draft:** ZL. **Writing – review & editing:** ZL, KK, CC, MR, BG. **Supervision:** BG. **Project administration:** BG. **Funding acquisition:** BG.

References

Baek, C., Jiang, Y., Raghunathan, A., Kolter, J.Z., 2022. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. Advances in Neural Information Processing Systems 35, 19274–19289.

Barber, R.F., Candes, E.J., Ramdas, A., Tibshirani, R.J., 2023. Conformal prediction beyond exchangeability. The Annals of Statistics 51, 816–845.

Bernhardt, M., Ribeiro, F.D.S., Glocker, B., 2022. Failure detection in medical image classification: A reality check and benchmarking testbed. arXiv preprint arXiv:2205.14094 .

Bloch, N., Madabhushi, A., Huisman, H., Freymann, J., Kirby, J., Grauer, M., Enquobahrie, A., Jaffe, C., Clarke, L., Farahani, K., 2015. Nci-isbi 2013 challenge: automated segmentation of prostate structures. The Cancer Imaging Archive 370.

Brett, M., Johnsrude, I.S., Owen, A.M., 2002. The problem of functional localization in the human brain. Nature reviews neuroscience 3, 243–249.

Campello, V.M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al., 2021. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. IEEE Transactions on Medical Imaging 40, 3543–3554.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T., 2019. Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems 32.

Castro, D.C., Walker, I., Glocker, B., 2020. Causality matters in medical imaging. Nature Communications 11, 3673.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM computing surveys (CSUR) 41, 1–58.

Chen, M., Goel, K., Sohoni, N.S., Poms, F., Fatahalian, K., Ré, C., 2021. Mandoline: Model evaluation under distribution shift. Proceedings of the International Conference on Machine Learning (ICML) , 1617–1629.

Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al., 2019. Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 .

Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V., 2020. Randaugment: Practical automated data augmentation with a reduced search space. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops , 702–703.

Deng, W., Zheng, L., 2021. Are labels always necessary for classifier accuracy evaluation? Proceedings of the IEEE/CVF conference on computer vision and pattern recognition , 15069–15078.

DeVries, T., Taylor, G.W., 2017. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 .

Eche, T., Schwartz, L.H., Mokrane, F.Z., Dercle, L., 2021. Toward generalizability in the deployment of artificial intelligence in radiology: Role of computation stress testing to overcome underspecification. Radiology: Artificial Intelligence 3, e210097.

Elsahar, H., Gallé, M., 2019. To annotate or not? predicting performance drop under domain shift. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) , 2163–2173.

Fan, W., Davidson, I., 2006. Reverse testing: an efficient framework to select amongst classifiers under sample selection bias. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining , 147–156.

Ganaie, M.A., Hu, M., Malik, A.K., Tanveer, M., Suganthan, P.N., 2022. Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence 115, 105151.

Gao, F., Han, L., 2012. Implementing the nelder-mead simplex algorithm with adaptive parameters. Computational Optimization and Applications 51, 259–277.

Gao, W., Jin, R., Zhu, S., Zhou, Z.H., 2013. One-pass auc optimization, in: International conference on machine learning, PMLR. pp. 906–914.

Garg, S., Balakrishnan, S., Lipton, Z.C., Neyshabur, B., Sedghi, H., 2022. Leveraging unlabeled data to predict out-of-distribution performance. arXiv preprint arXiv:2201.04234 .

Granese, F., Romanelli, M., Gorla, D., Palamidessi, C., Piantanida, P., 2021. Doctor: A simple method for detecting misclassification errors. Advances in Neural Information Processing Systems 34, 5669–5681.

Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., Schmidt, L., 2021. Predicting with confidence on unseen distributions. Proceedings of the IEEE/CVF international conference on computer vision , 1134–1144.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. International conference on machine learning , 1321–1330.

Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A., 2016. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1605.01397 .

Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al., 2020. Array programming with numpy. Nature 585, 357–362.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition , 770–778.

Hendrycks, D., Dietterich, T., 2019a. Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the ICLR .

Hendrycks, D., Dietterich, T., 2019b. Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 .

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 .

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association 47, 663–685.

Hu, D., Liang, J., Wang, X., Foo, C.S., 2024. Pseudo-calibration: Improving predictive uncertainty estimation in unsupervised domain adaptation, in: Forty-first International Conference on Machine Learning, pp. 19304–19326.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18, 203–211.

Islam, M., Seenivasan, L., Ren, H., Glocker, B., 2021. Class-distribution-aware calibration for long-tailed visual recognition. arXiv preprint arXiv:2109.05263 .

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al., 2018. Ensembles of multiple models and architectures for robust brain tumour segmentation, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3, Springer. pp. 450–462.

Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G., 2018. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE journal of biomedical and health informatics 23, 538–546.

Kivimäki, J., Białek, J., Nurminen, J.K., Kuberski, W., 2024. Confidence-based estimators for predictive performance in model monitoring. arXiv preprint arXiv:2407.08649 .

Krizhevsky, A., Hinton, G., et al., 2009. Learning multiple layers of features from tiny images. Technical report .

Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., Flach, P., 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. Advances in neural information processing systems 32.

Le Coz, A., Herbin, S., Adjed, F., 2024. Confidence calibration of classifiers with many classes. Advances in Neural Information Processing Systems 37, 77686–77725.

Lemaître, G., Martí, R., Freixenet, J., Vilanova, J.C., Walker, P.M., Meriaudeau, F., 2015. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. Computers in biology and medicine 60, 8–31.

Li, K., Yu, L., Heng, P.A., 2022a. Towards reliable cardiac image segmentation: Assessing image-level and pixel-level segmentation quality via self-reflective references. Medical Image Analysis 78, 102426.

Li, Z., Kamnitsas, K., Glocker, B., 2020. Analyzing overfitting under class imbalance in neural networks for image segmentation. IEEE transactions on medical imaging 40, 1065–1077.

Li, Z., Kamnitsas, K., Islam, M., Chen, C., Glocker, B., 2022b. Estimating model performance under domain shifts with class-specific confidence scores, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 693–703.

Li, Z., Zheng, Y.Q., Chen, C., Jbabdi, S., 2024. Learning label refinement and threshold adjustment for imbalanced semi-supervised learning. arXiv preprint arXiv:2407.05370 .

Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., Jeong, J.N., Barisano, G., Hutton, A., Simon, J.P., Juliano, J.M., Suri, A., et al., 2022.

A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific data 9, 320.

Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al., 2014. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. Medical image analysis 18, 359–373.

Liu, Q., Dou, Q., Heng, P.A., 2020a. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. International conference on medical image computing and computer-assisted intervention , 475–485.

Liu, W., Wang, X., Owens, J., Li, Y., 2020b. Energy-based out-of-distribution detection. Advances in neural information processing systems 33, 21464–21475.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X., 2019. Large-scale long-tailed recognition in an open world, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2537–2546.

Lu, Y., Qin, Y., Zhai, R., Shen, A., Chen, K., Wang, Z., Kolouri, S., Stepputtis, S., Campbell, J., Sycara, K., 2023. Characterizing out-of-distribution error via optimal transport. Advances in Neural Information Processing Systems 36, 17602–17622.

Marrakchi, Y., Makansi, O., Brox, T., 2021. Fighting class imbalance with contrastive learning. International conference on medical image computing and computer-assisted intervention , 466–476.

Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T., 2020. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE transactions on medical imaging 39, 3868–3878.

Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J., 2013. Ph 2-a dermoscopic image database for research and benchmarking. 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC) , 5437–5440.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 fourth international conference on 3D vision (3DV) , 565–571.

NannyML, 2023. NannyML (release 0.11.0). https://github.com/NannyML/nannyml. NannyML, Belgium, OHL.

Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., Rueckert, D., 2022a. Causality-inspired single-source domain generalization for medical image segmentation. IEEE Transactions on Medical Imaging 42, 1095–1106.

Ouyang, C., Wang, S., Chen, C., Li, Z., Bai, W., Kainz, B., Rueckert, D., 2022b. Improved post-hoc probability calibration for out-of-domain mri segmentation, in: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, Springer. pp. 59–69.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in python. the Journal of machine Learning research 12, 2825–2830.

Press, O., Shwartz-Ziv, R., LeCun, Y., Bethge, M., 2024. The entropy enigma: Success and failure of entropy minimization. arXiv preprint arXiv:2405.05012 .

Rivera, S.C., Liu, X., Chan, A.W., Denniston, A.K., Calvert, M.J., Ashrafian, H., Beam, A.L., Collins, G.S., Darzi, A., Deeks, J.J., et al., 2020. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the spirit-ai extension. The Lancet Digital Health 2, e549–e560.

Robinson, R., Oktay, O., Bai, W., Valindria, V.V., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., et al., 2018. Real-time prediction of segmentation quality. International conference on medical image computing and computer-assisted intervention , 578–585.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, Springer. pp. 234–241.

Roschewitz, M., Glocker, B., 2023. Distance matters for improving performance estimation under covariate shift, in: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), IEEE. pp. 4551–4561.

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., et al., 2021. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific data 8, 1–8.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. International journal of computer vision 115, 211–252.

Schneider, S., Lee, J.H., Mathis, M.W., 2023. Learnable latent embeddings for joint behavioural and neural analysis. Nature 617, 360–368.

Steinhardt, J., Liang, P.S., 2016. Unsupervised risk estimation using only conditional independence structure. Advances in Neural Information Processing Systems 29.

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data 5, 1–9.

Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B., 2017. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE transactions on medical imaging 36, 1597–1606.

Valk, K., Kull, M., 2023. Assuming locally equal calibration errors for non-parametric multiclass calibration. Transactions on Machine Learning Research .

Vasey, B., Nagendran, M., Campbell, B., Clifton, D.A., Collins, G.S., Denaxas, S., Denniston, A.K., Faes, L., Geerts, B., Ibrahim, M., et al., 2022. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: Decide-ai. bmj 377.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al., 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods 17, 261–272.

Wang, H., Yu, Z., Yue, Y., Anandkumar, A., Liu, A., Yan, J., 2020a. Learning calibrated uncertainties for domain shift: A distributionally robust learning approach. arXiv preprint arXiv:2010.05784 .

Wang, S., Tarroni, G., Qin, C., Mo, Y., Dai, C., Chen, C., Glocker, B., Guo, Y., Rueckert, D., Bai, W., 2020b. Deep generative model-based quality control for cardiac mri segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23, Springer. pp. 88–97.

Yoon, C., Hamarneh, G., Garbi, R., 2019. Generalizable feature learning in the presence of data bias and domain class imbalance with application to skin lesion classification. International conference on medical image computing and computer-assisted intervention , 365–373.

Yu, Y., Bates, S., Ma, Y., Jordan, M., 2022. Robust calibration with multi-domain temperature scaling. Advances in Neural Information Processing Systems 35, 27510–27523.

Zenk, M., Zimmerer, D., Isensee, F., Traub, J., Norajitra, T., Jäger, P.F., Maier-Hein, K., 2025. Comparative benchmarking of failure detection methods in medical image segmentation: Unveiling the role of confidence aggregation. Medical Image Analysis 101, 103392. URL: https://www.sciencedirect.com/science/article/pii/S1361841524003177, doi:https://doi.org/10.1016/j.media.2024.103392.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 .

Zhao, S., Kim, M., Sahoo, R., Ma, T., Ermon, S., 2021. Calibrating predictions to decisions: A novel approach to multi-class calibration. Advances in Neural Information Processing Systems 34, 22313–22324.

Zong, C.C., Wang, Y.W., Xie, M.K., Huang, S.J., 2024. Dirichlet-based prediction calibration for learning with noisy labels, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 17254–17262.

Appendix A. Quantitative Results

Dataset and Training Strategy			Estimation Algorithm	Confidence Score								MOVAL Ensemble
				Class-Agnostic				Class-Specific				
				MCP	Energy	Entropy	Doctor	MCP	Energy	Entropy	Doctor	
CIFAR10	Synthetic-95	Vanilla Training	Vanilla	32.6 ±6.2	29.4 ±5.2	30.5 ±6.1	36.3 ±6.5	—	—	—	—	
			TS	6.5 ±5.0	13.5 ±6.7	5.5 ±5.0	4.9 ±4.0	7.1 ±5.4	12.4 ±7.6	5.9 ±4.9	4.0 ±3.7	
			DoC	11.7 ±6.2	8.5 ±5.3	10.9 ±6.1	12.6 ±6.5	10.9 ±6.5	8.5 ±5.9	10.3 ±6.1	11.6 ±6.7	**2.8** ±2.1
			ATC	4.1 ±3.7	6.2 ±3.7	4.2 ±3.7	4.1 ±3.7	3.5 ±3.5	3.2 ±1.8	3.3 ±3.2	3.4 ±3.4	
			TS-ATC	5.0 ±2.6	6.7 ±3.7	5.4 ±3.1	5.3 ±2.9	**2.8** ±2.7	3.6 ±2.2	**2.8** ±2.4	<u>3.0</u> ±2.2	
CIFAR100	Synthetic-95	Vanilla Training	Vanilla	46.7 ±4.3	5.4 ±3.1	55.6 ±4.6	64.8 ±4.9	—	—	—	—	
			TS	6.1 ±3.8	10.4 ±4.8	6.9 ±4.2	3.3 ±2.9	5.8 ±3.2	10.5 ±3.9	6.5 ±3.6	2.8 ±2.8	
			DoC	9.0 ±4.3	8.7 ±4.5	9.7 ±4.6	11.0 ±4.9	8.5 ±3.6	7.8 ±3.8	9.0 ±3.8	10.3 ±4.1	<u>2.5</u> ±2.3
			ATC	4.2 ±3.1	3.3 ±2.4	3.8 ±3.0	4.2 ±3.1	3.9 ±2.7	2.9 ±2.6	3.3 ±2.7	3.8 ±2.7	
			TS-ATC	2.7 ±4.2	4.2 ±3.3	3.0 ±2.2	2.7 ±1.7	**2.2** ±2.1	3.5 ±3.2	2.6 ±2.4	**2.2** ±1.8	
HAM10000	Natural-7	Vanilla Training	Vanilla	20.8 ±11.6	41.8 ±12.2	15.9 ±9.4	20.1 ±10.8	—	—	—	—	
			TS	11.0 ±6.6	20.1 ±10.2	9.8 ±6.3	8.9 ±5.4	11.1 ±6.7	19.3 ±10.4	9.8 ±6.3	8.2 ±5.5	
			DoC	15.3 ±8.3	14.0 ±7.8	12.6 ±7.3	14.7 ±8.1	15.1 ±8.5	13.8 ±7.8	12.6 ±7.4	14.5 ±8.2	**3.9** ±3.8
			ATC	8.3 ±5.0	4.9 ±3.0	7.4 ±4.2	8.6 ±5.0	6.0 ±4.3	5.1 ±3.5	5.4 ±4.0	6.2 ±4.3	
			TS-ATC	6.9 ±6.3	7.7 ±6.3	5.5 ±2.8	6.5 ±3.2	4.7 ±4.0	7.4 ±5.8	<u>4.3</u> ±3.8	<u>4.3</u> ±3.9	

Table A.1: Related to Fig. 4. Evaluation of accuracy estimation for different classification tasks under various types of domain shifts based on MAE. The best and second-best results within the same category are in **bold** and <u>underlined</u> for each configuration.

Dataset		Estimation Algorithm	Confidence Score								MOVAL Ensemble
			Class-Agnostic				Class-Specific				
			MCP	Energy	Entropy	Doctor	MCP	Energy	Entropy	Doctor	
Cardiac-2D	Synthetic-65	Vanilla	17.8 ±7.7	52.5 ±9.0	16.3 ±7.2	17.5 ±7.5	—	—	—	—	
		TS	14.8 ±7.1	19.8 ±8.4	13.3 ±6.7	13.7 ±6.7	5.1 ±5.0	6.3 ±7.0	4.7 ±4.2	4.5 ±3.9	
		DoC	17.4 ±7.7	17.0 ±8.0	15.9 ±7.2	17.1 ±7.5	5.6 ±6.0	5.9 ±6.1	5.2 ±5.5	5.5 ±5.9	4.3 ±5.0
		ATC	6.5 ±4.5	19.8 ±8.4	6.4 ±4.4	6.5 ±4.5	**3.6** ±4.2	4.5 ±7.7	**3.6** ±4.3	**3.6** ±4.2	
		TS-ATC	6.2 ±6.7	19.8 ±8.4	6.2 ±4.4	6.2 ±4.4	4.0 ±6.2	6.2 ±10.1	4.0 ±6.2	4.0 ±6.2	
	Natural-4	Vanilla	15.9 ±0.8	52.3 ±1.8	14.8 ±0.8	15.7 ±0.8	—	—	—	—	
		TS	13.8 ±0.8	17.3 ±0.9	12.8 ±0.8	13.1 ±0.8	3.4 ±0.6	2.6 ±0.5	3.8 ±0.4	3.8 ±0.4	
		DoC	15.5 ±0.8	15.4 ±1.0	14.5 ±0.8	15.3 ±0.8	**2.4** ±0.4	2.9 ±0.7	**2.4** ±0.4	**2.4** ±0.4	2.6 ±0.4
		ATC	7.8 ±0.4	17.3 ±0.9	7.8 ±0.4	7.8 ±0.4	3.6 ±0.2	5.5 ±1.2	3.6 ±0.2	3.6 ±0.2	
		TS-ATC	7.9 ±0.8	17.3 ±0.9	7.9 ±0.4	7.9 ±0.4	4.5 ±0.8	8.1 ±2.0	4.5 ±0.8	4.5 ±0.8	
Prostate-3D	Synthetic-83	Vanilla	12.7 ±8.0	61.4 ±5.3	8.6 ±4.8	10.7 ±6.4	—	—	—	—	
		TS	10.4 ±7.0	13.1 ±9.0	6.1 ±3.4	7.2 ±4.2	2.3 ±6.5	4.0 ±9.0	1.4 ±1.6	**1.3** ±1.7	
		DoC	12.5 ±8.0	9.1 ±7.8	8.5 ±4.8	10.6 ±6.4	3.4 ±7.8	2.8 ±7.6	1.7 ±4.2	2.4 ±6.1	1.7 ±2.2
		ATC	3.5 ±1.6	14.1 ±9.1	3.5 ±1.6	3.5 ±1.6	3.1 ±5.3	3.2 ±5.4	3.1 ±5.3	3.1 ±5.3	
		TS-ATC	3.5 ±3.4	3.1 ±3.4	3.5 ±1.6	3.5 ±1.6	3.1 ±5.3	3.6 ±5.9	3.1 ±5.3	3.1 ±5.3	
	Natural-5	Vanilla	18.9 ±6.6	57.2 ±5.3	12.9 ±4.3	16.0 ±5.4	—	—	—	—	
		TS	16.0 ±5.9	20.0 ±7.4	9.7 ±3.5	11.2 ±4.0	7.4 ±5.8	11.0 ±7.4	2.2 ±1.8	2.5 ±2.1	
		DoC	18.8 ±6.6	15.2 ±6.6	12.8 ±4.3	15.9 ±5.3	9.5 ±6.4	8.7 ±6.5	5.4 ±4.0	7.4 ±5.1	**1.6** ±1.4
		ATC	5.2 ±1.9	21.1 ±7.4	5.2 ±1.9	5.2 ±1.9	2.7 ±2.1	2.7 ±2.1	2.7 ±2.1	2.7 ±2.1	
		TS-ATC	5.2 ±3.5	2.1 ±0.9	5.2 ±1.9	5.2 ±1.9	2.7 ±2.1	2.7 ±2.0	2.7 ±2.1	2.7 ±2.1	
Brain Lesion-3D	Synthetic-83	Vanilla	23.2 ±7.7	73.0 ±8.2	18.7 ±5.9	21.1 ±6.7	—	—	—	—	
		TS	20.9 ±7.2	24.8 ±8.4	17.3 ±5.6	18.0 ±5.8	4.4 ±7.2	5.1 ±8.3	2.2 ±4.1	2.3 ±4.2	
		DoC	23.1 ±7.7	24.8 ±8.4	18.7 ±5.9	21.0 ±6.7	4.3 ±7.3	5.1 ±8.3	2.8 ±5.4	3.4 ±6.2	2.1 ±3.3
		ATC	15.3 ±4.6	24.8 ±8.4	15.3 ±4.6	15.3 ±4.6	2.3 ±4.0	2.3 ±4.0	2.3 ±4.0	2.3 ±4.0	
		TS-ATC	15.3 ±5.6	24.8 ±8.4	15.3 ±4.6	15.3 ±4.6	2.3 ±4.0	**1.8** ±2.6	2.3 ±4.0	2.3 ±4.0	
	Natural-12	Vanilla	51.8 ±15.0	39.2 ±16.7	45.5 ±12.6	48.5 ±13.4	—	—	—	—	
		TS	49.3 ±14.8	54.0 ±15.9	44.0 ±12.4	44.9 ±12.6	34.4 ±16.2	35.7 ±17.1	28.7 ±12.7	28.9 ±12.8	
		DoC	51.7 ±15.0	54.0 ±15.9	45.5 ±12.6	48.5 ±13.4	33.9 ±15.8	35.6 ±17.1	30.3 ±13.3	31.8 ±14.1	25.5 ±12.3
		ATC	40.5 ±11.0	54.0 ±15.9	40.5 ±11.0	40.5 ±11.0	28.8 ±12.4	28.7 ±12.3	28.8 ±12.4	28.8 ±12.4	
		TS-ATC	40.5 ±12.4	54.0 ±15.9	40.5 ±11.0	40.5 ±11.0	28.8 ±12.4	**16.8** ±10.7	28.8 ±12.4	28.8 ±12.4	

Table A.2: Related to Fig. 5. Evaluation of DSC (F1-Score) estimation for different segmentation tasks under various types of domain shifts based on MAE. The best and second-best results within the same category are in **bold** and underlined for each configuration.

Appendix B. MOVAL API and Example Usage

We offer a Python implementation of MOVAL that supports input in the form of NumPy arrays (Harris et al., 2020) [5]. The `moval` package (installable via `pip install moval`) implements only the MOVAL estimation algorithms (confidence scores, calibration, metric computation) for use with any trained model; it does not itself contain the paper's experiments. The separate `moval_scripts` repository, together with the dataset and model repositories linked above, contains the scripts and configurations needed to reproduce the main classification and segmentation experiments reported in this paper. The provided API is compatible with scikit-learn (Pedregosa et al., 2011), a widely used machine learning library. This compatibility allows users to leverage scikit-learn tools for downstream processing of the computed probabilities, such as classification, clustering, and dimensionality reduction (Schneider et al., 2023).

In a classification task with $c = 10$ classes, the validation set contains 100 samples, while the test dataset consists of 200 samples:

```
import numpy as np
C = 10
logits = np.random.randn(100, C)
gts = np.random.randint(0, C, (100))
l_test = np.random.randn(200, C)
```

For example, a MOVAL classification model calibrated for accuracy using TS with Entropy can be defined as:

```
import moval
moval_model = moval.MOVAL(
    mode = "classification",
    confidence_scores = "entropy-conf",
    estim_algorithm = "ts-model",
    class_specific = False
)
```

Alternatively, as we suggest, a MOVAL Ensemble model can be constructed with:

[5] Classification datasets and models: https://github.com/ZerojumpLine/Robust-Skin-Lesion-Classification. Segmentation datasets and models: https://github.com/ZerojumpLine/Robust-Medical-Segmentation. Project page: https://zerojumpline.github.io/moval/. Code: https://github.com/ZerojumpLine/MOVAL. Documentation: https://moval.readthedocs.io/en/latest/index.html. Code (and data) to reproduce the figures: https://github.com/ZerojumpLine/moval_scripts.

```
moval_model = moval.MOVAL(
    mode = "classification",
    estim_algorithm = \
    "moval-ensemble-cls-accuracy"
)
```

For a 3D segmentation task with $c = 2$ classes, the validation set consists of 5 samples, and the test set includes 1 sample, defined as:

```
C = 2 ; H = 50 ; W = 50 ; D = 80
logits = [] ; gts = []
for _ in range(5):
    logit = np.random.randn(C, H, W, D)
    gt = np.random.randint(0, C, (H, W, D))
    logits.append(logit)
    gts.append(gt)
l_test = [np.random.randn(C, H, W, D)]
```

Similarly, a MOVAL segmentation model calibrated for F1-Score using CS ATC with Doctor can be defined as:

```
moval_model = moval.MOVAL(
    mode = "segmentation",
    metric = "f1score",
    confidence_scores = "doctor-conf",
    estim_algorithm = "atc-model",
    class_specific = True
)
```

The fitting and inference processes for the MOVAL model in both tasks are identical:

```
moval_model.fit(logits, gts)
estim = moval_model.estimate(l_test)
# estim is of shape (1, ) for accuracy,
# while (c, ) for group-wise metrics
```

After optimization, we can obtain instance-wise estimation for the validation data using:

```
estim_val = moval_model.fitted_perf
# estim_val has N values for classification
# while Z values for segmentation
```

In addition to the overall estimated metrics, we can also derive the calibrated probabilities $\hat{p}_i$:

```
prob = moval_model.get_probability(l_tests)
```

We recommend supplying the MOVAL model with all pre-trained models. To achieve this, the parameters can be saved and loaded using:

```
moval_model.save('./m.pkl')
loaded_model = moval.MOVAL.load('./m.pkl')
```

Appendix C. Supplementary Figures

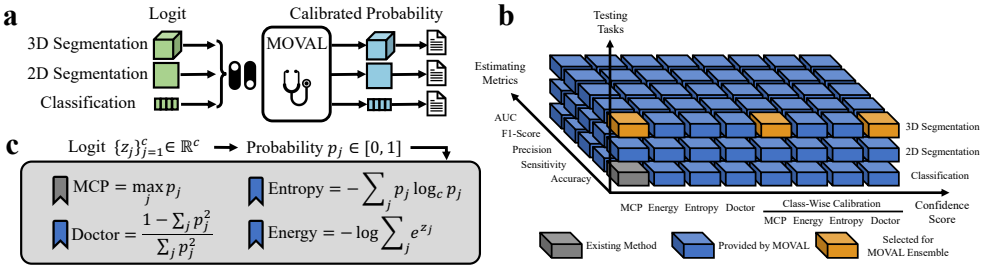


Figure C.11: Comparison of MOVAL with existing standards. **(a)**: Within a unified framework, MOVAL can be readily applied to classification, 2D, and 3D segmentation tasks. **(b)**: After extending existing methods in many dimensions, we find that an ensemble of multiple confidence scores leads to more robust and state-of-the-art results. **(c)**: MOVAL explores performance estimation based on confidence scores rather than maximum class probability (MCP), which is the basis of all existing methods.

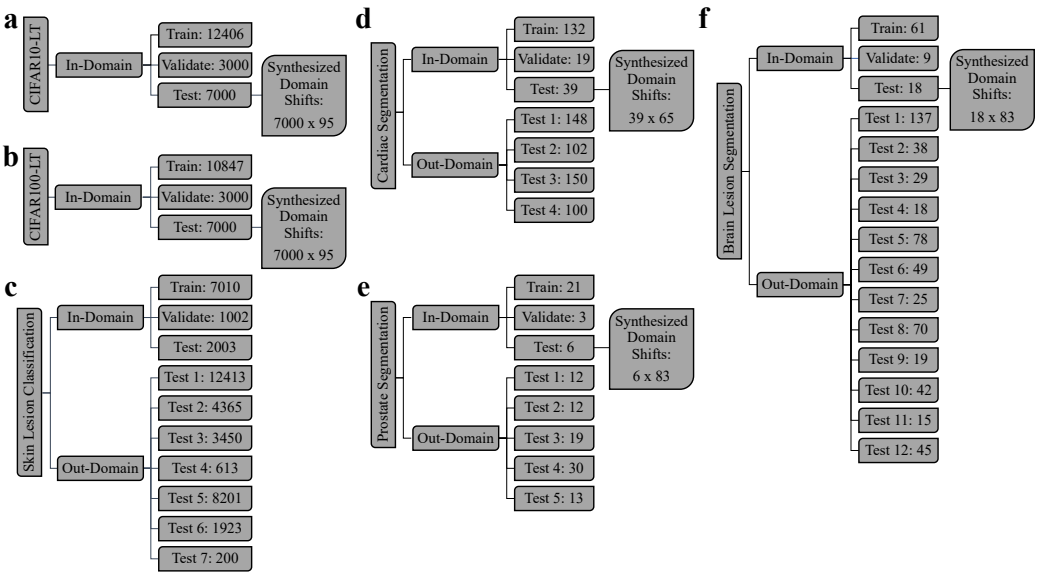


Figure C.12: Summary of the datasets used in this study. **(a)**: Classification task I, Long-tailed version of the CIFAR10 dataset. **(b)**: Classification task II, Long-tailed version of the CIFAR10 dataset. **(c)**: Classification task III, skin lesion classification from 8 sources of data. **(d)**: 2D segmentation task I, cardiac MRI segmentation from 5 different scanners. **(e)**: 3D segmentation task I, prostate MRI segmentation from 6 different scanners. **(f)**: 3D segmentation task II, brain lesion segmentation from 13 different scanners.

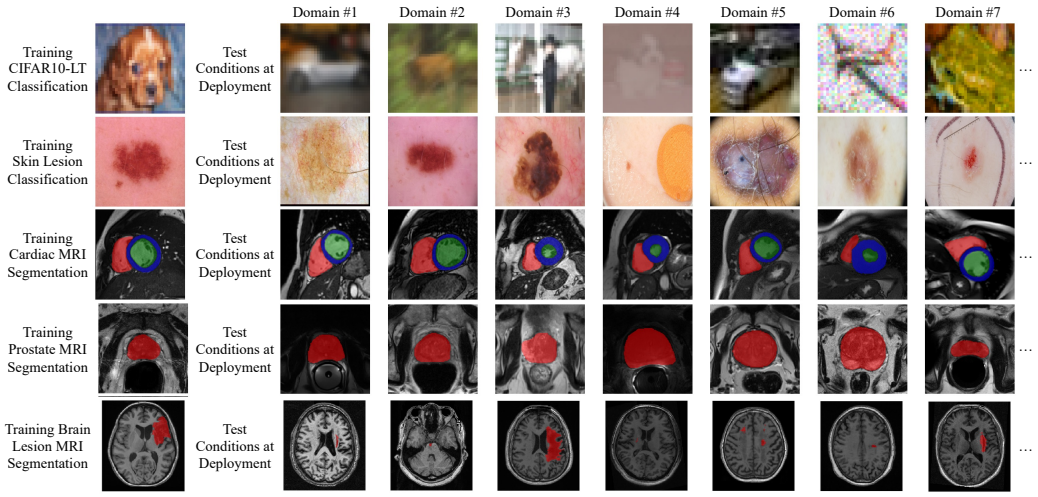


Figure C.13: In real-world scenarios, significant discrepancies often exist between training and testing environments, affecting the distribution of acquired images. For instance, in natural images, such discrepancies can be caused by image corruptions like blur or noise (Hendrycks and Dietterich, 2019b). In medical images, these can arise from scanner effects or variations in imaging protocols (Castro et al., 2020). For segmentation tasks, regions of interest are highlighted with different colors.

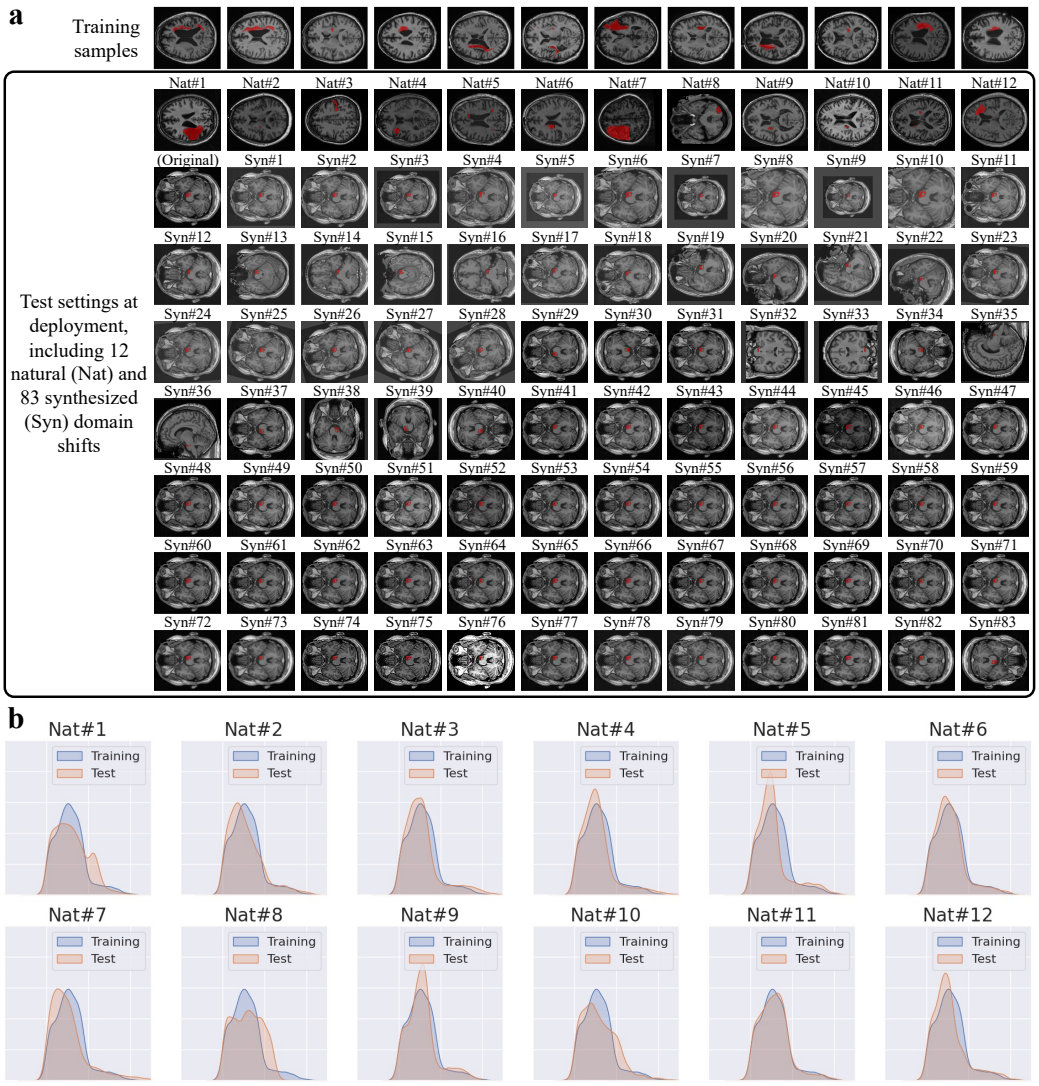


Figure C.14: Examples of brain lesion segmentation datasets. **(a)**: Visualization of Brain MR images of In-Domain and Out-Domain Examples. Top: Brain MR images used for training. Brain lesion regions are highlighted with red color. Bottom: Brain MR images with natural and synthesized domain shifts. **(b)**: Intensity histograms of training and test data for 12 test conditions.

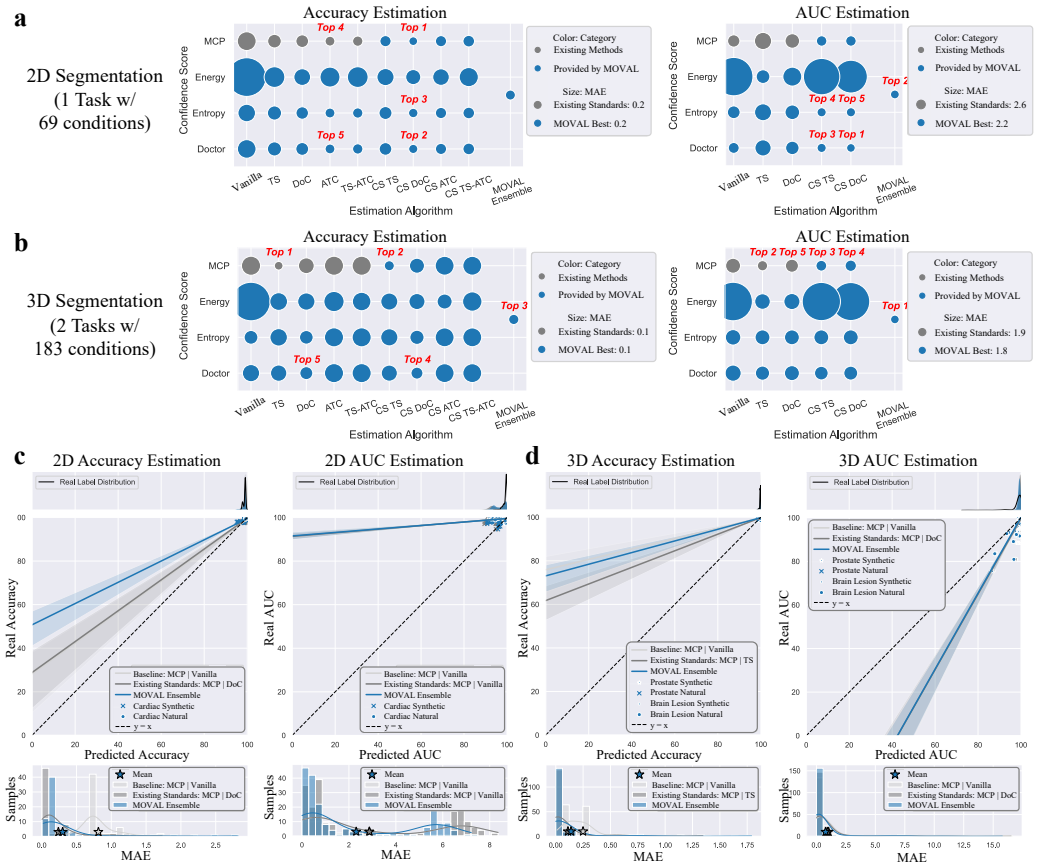


Figure C.15: Accuracy and AUC are less useful for segmentation tasks. **(a, b)**: MAE between real metrics and estimated results based on different estimation methods for 2D and 3D segmentation tasks. We also report the average MAE for natural and synthetic domain shifts here. The smaller the MAE, the better. **(c, d)**: Top: Predicted and real label distribution. We benchmark MOVAL ensemble (blue) against MCP baselines (light gray) and the best-performing existing method (dark gray). Middle: Predicted versus real metrics for 2D and 3D segmentation tasks. Bottom: Histogram of MAE for different estimation methods. As the class ratios are extremely imbalanced for segmentation tasks, accuracy and AUC are less informative as they are close to 100%.

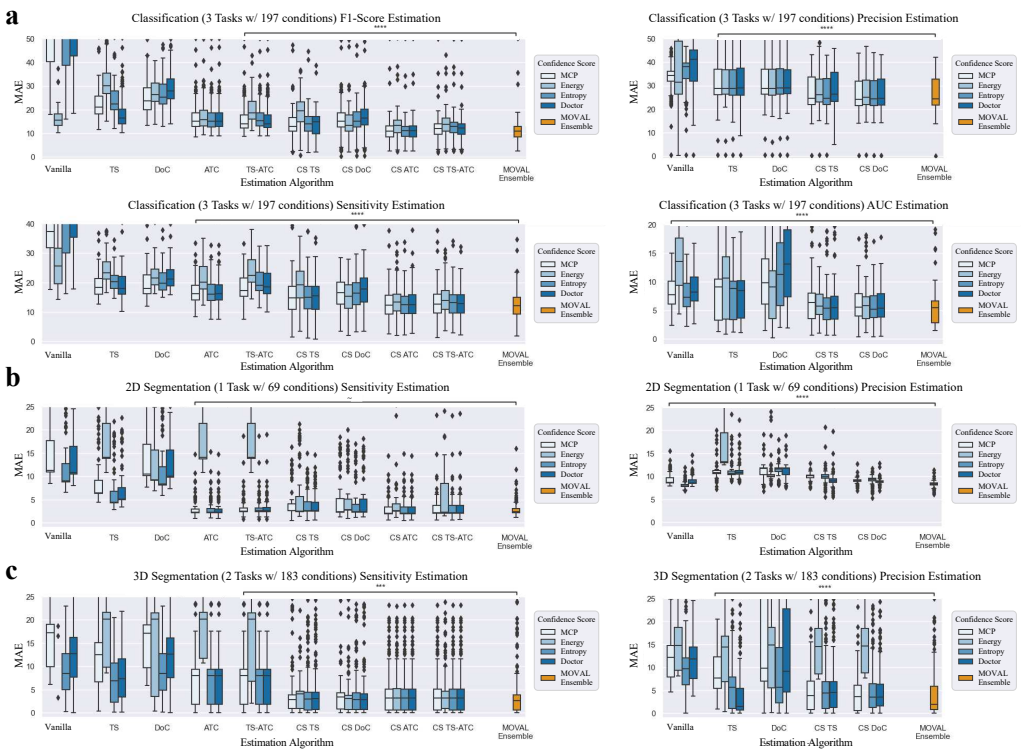


Figure C.16: MOVAL Ensemble achieves significant improvements on major evaluation metrics. **(a, b, c)**: MAE comparison between proposed algorithms and existing methods (Existing best method versus MOVAL Ensemble, paired t-test, $T(196) = 9.54, p = 6.04 \times 10^{-18}$ for F1-Score estimation of classification tasks, $T(196) = 4.97, p = 1.43 \times 10^{-6}$ for precision estimation of classification tasks, $T(196) = 7.80, p = 3.54 \times 10^{13}$ for sensitivity estimation of classification tasks, $T(196) = 5.10, p = 8.18 \times 10^{-7}$ for AUC estimation of classification tasks, $T(68) = -0.26, p = 0.80$ for sensitivity estimation of 2D segmentation tasks, $T(68) = 5.42, p = 8.54 \times 10^{-7}$ for precision estimation of 2D segmentation tasks, $T(182) = 3.96, p = 1.09 \times 10^{-4}$ for sensitivity estimation for 3D segmentation tasks and $T(182) = 6.79, p = 1.53 \times 10^{-10}$ for precision estimation for 3D segmentation tasks.

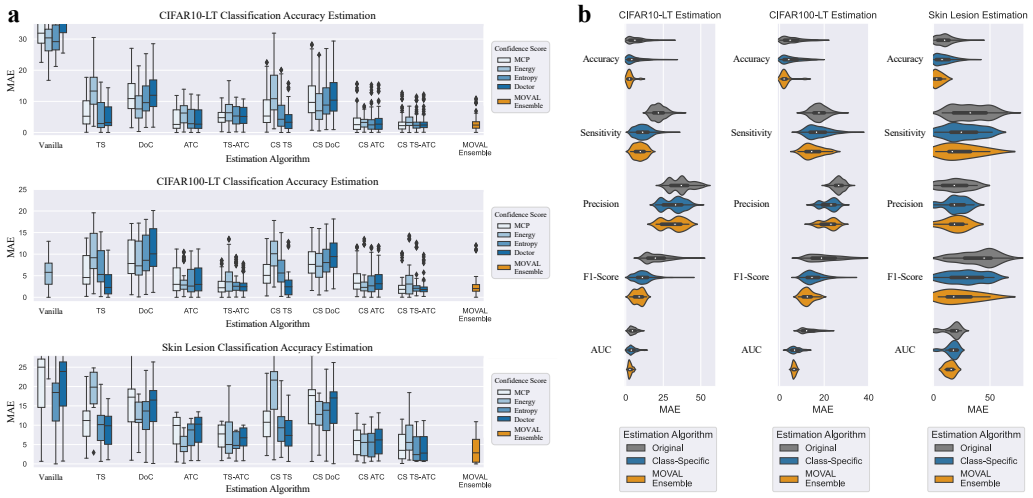


Figure C.17: MOVAL Ensemble achieves state-of-the-art performance for several classification tasks. **(a)**: MAE comparison between proposed algorithms and existing methods for different classification tasks. **(b)**: MAE comparison between the original estimation methods and our proposed class-specific variants for different classification tasks.

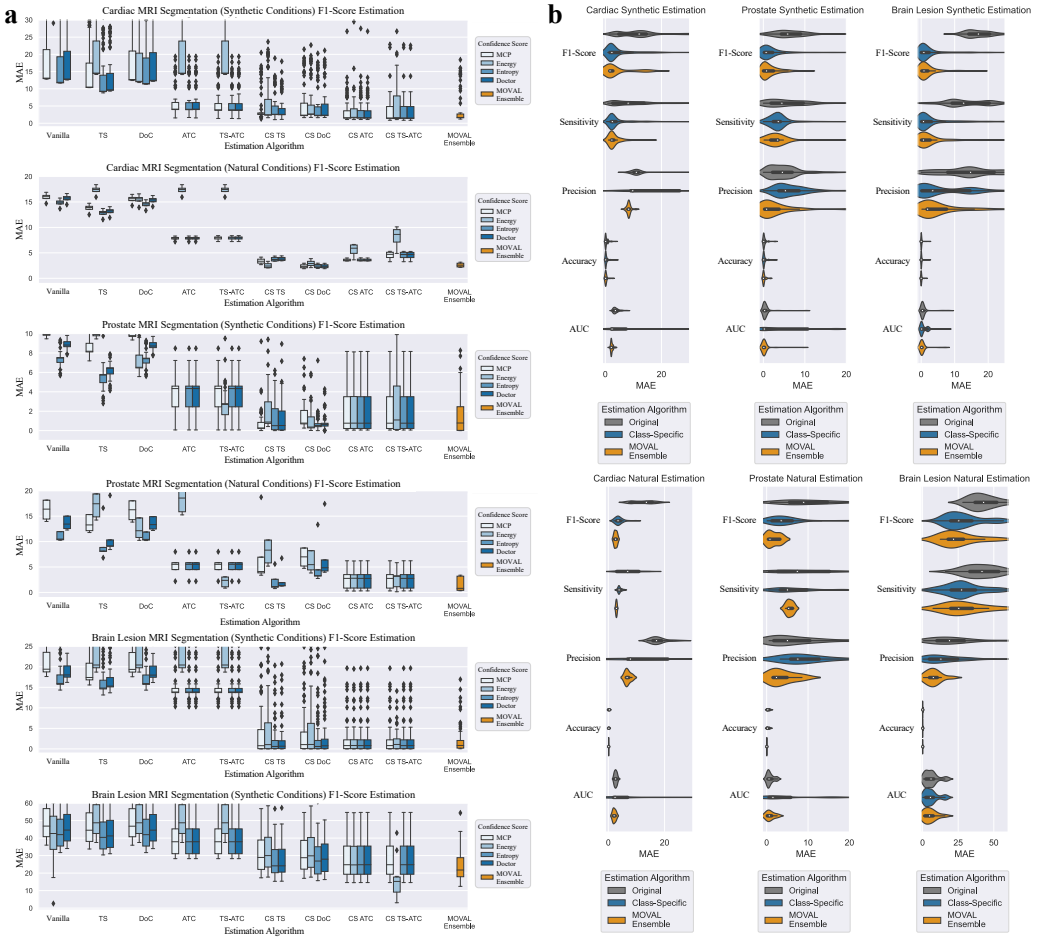


Figure C.18: MOVAL Ensemble achieves state-of-the-art performance for several segmentation tasks. **(a)**: MAE comparison between proposed algorithms and existing methods for different segmentation tasks. **(b)**: MAE comparison between the original estimation methods and our proposed class-specific variants for different segmentation tasks.

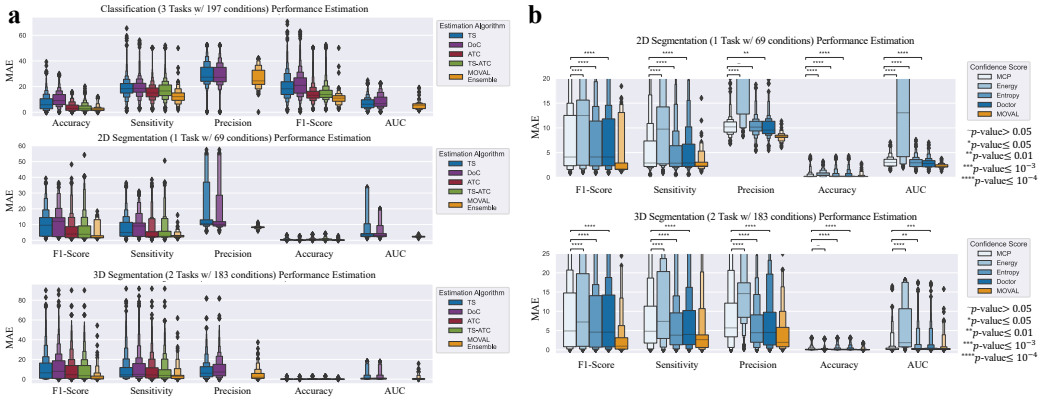


Figure C.19: MOVAL Ensemble combines advanced calibration and confidence scores for better performance. **(a)**: MAE comparison between estimation algorithms based on different confidence calibration algorithms. **(b)**: MAE comparison between estimation algorithms based on MCP and other confidence scores for segmentation tasks.

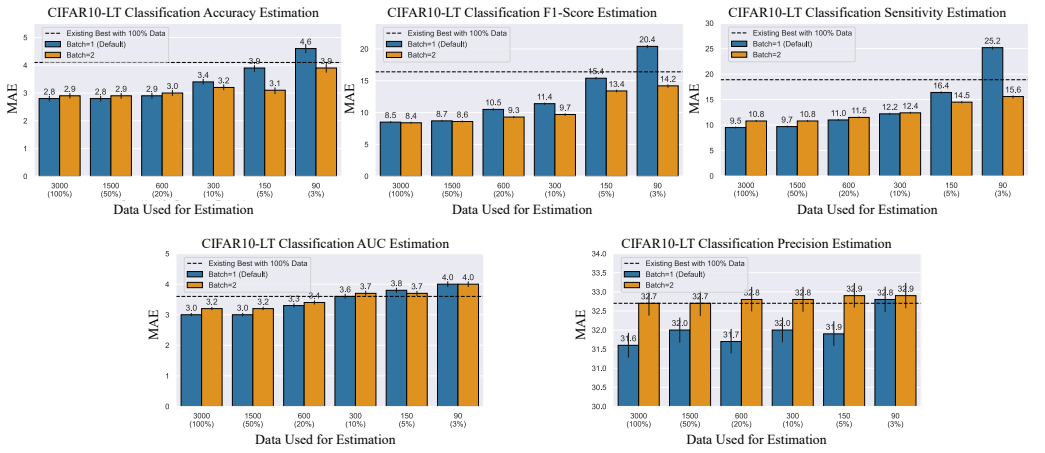


Figure C.20: Group-wise calibration enhances the robustness of the MOVAL Ensemble against a reduction in validation samples. The proposed group-wise calibration method generally requires fewer validation samples because it has fewer learnable parameters. However, it may negatively impact the estimation of false positive predictions, which require class-specific adjustments. This could potentially affect the estimation of AUC and precision when data are sufficient. The standard error is shown alongside the bar plots.

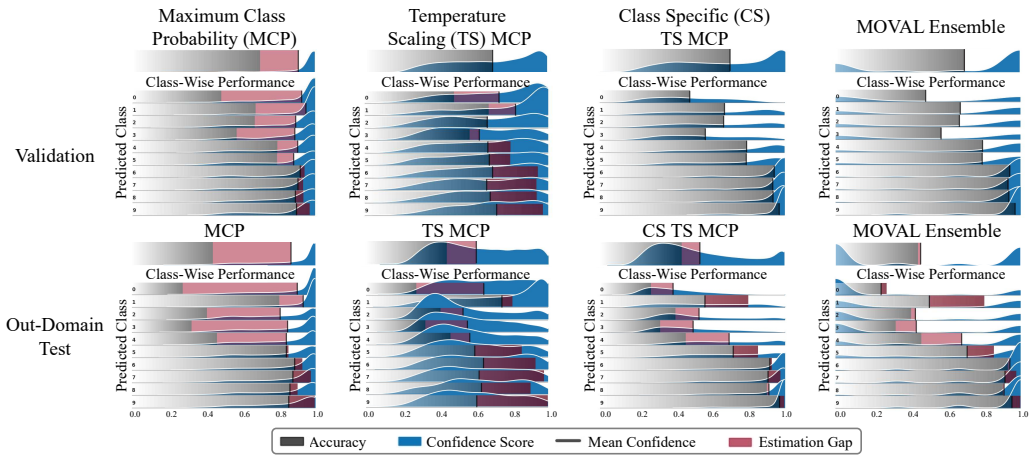


Figure C.21: Confidence score distributions of different calibration methods on CIFAR10-LT. For each class and calibration method, we show the estimated accuracy (gray gradient bar), the true accuracy (black tick), and the gap between them (red), together with the underlying confidence-score distribution (blue). Confidence distributions are shown as kernel density estimates (KDE) of the per-sample confidence values, for visualization only, to make the shift in confidence across classes easier to see. MOVAL Ensemble, which combines multiple calibrated confidence scores, shows a distinctly bimodal distribution concentrated near 0 and 1, providing a more distinct indicator of prediction correctness than the single-score methods.

Appendix D. Supplementary Note

Appendix D.1. Group-Wise and Class-Specific Metrics Estimation

Apart from accuracy, we can also estimate group-wise metrics that measure model performance from different perspectives based on the predicted probability $\boldsymbol{p}_i$ instead of the confidence score q_i. This is because these metrics also incorporate information from non-maximum classes. This approach is related to efforts that formulate these metrics as loss functions (Gao et al., 2013; Milletari et al., 2016). However, to enable estimation without ground truth at test time, we use the predicted label y_i' as a substitute for the ground truth in our calculations. This can be viewed as an extension of the confidence-based accuracy framework, which assumes that all predictions are correct at some level of confidence. Specifically, the estimated F1-Score is defined as:

$$\Phi_j' = \frac{2\sum_{i=1}^{N} \mathbb{1}_{ij} p_{ij}}{\sum_{i=1}^{N} \mathbb{1}_{ij} + \sum_{i=1}^{N} p_{ij}}, \tag{D.1}$$

where $\mathbb{1}_{ij} = \mathbb{1}(y_i' = j)$ is 1 if the predicted most probable class is j and 0 otherwise. In a similar manner, sensitivity and precision can be calculated as follows:

$$\rho_j' = \frac{\sum_{i=1}^{N} \mathbb{1}_{ij} p_{ij}}{\sum_{i=1}^{N} p_{ij}}, \quad \pi_j' = \frac{\sum_{i=1}^{N} \mathbb{1}_{ij} p_{ij}}{\sum_{i=1}^{N} \mathbb{1}_{ij}}. \tag{D.2}$$

The AUC calculation uses the One-vs-Rest strategy, which involves comparing one class against all other classes combined in multi-class problems. Notably, this is equivalent to the One-vs-One strategy for binary classification, which involves comparing every unique pairwise combination. The estimation is given by:

$$\Lambda_j' = \int_0^1 \mathrm{TPR}_j'(t) d(\mathrm{FPR}_j'(t)), \tag{D.3}$$

where the true positive rate (TPR) and false positive rate (FPR) are calculated as follows:

$$\mathrm{TPR}_j'(t) = \frac{\sum_{i=1}^{N} \mathbb{1}(p_{ij} > t) p_{ij}}{\sum_{i=1}^{N} \mathbb{1}(p_{ij} > t)}, \quad \mathrm{FPR}_j'(t) = \frac{\sum_{i=1}^{N} \mathbb{1}(p_{ij} \leq t) \sum_{k \neq j} p_{ik}}{\sum_{i=1}^{N} \mathbb{1}(p_{ij} \leq t)}. \tag{D.4}$$

Throughout, i indexes a sample for classification and a pixel (2D) or voxel (3D) for segmentation; for segmentation, z additionally indexes the case (volume), so the per-pixel probability is written p_{zij}. Eqs. D.1-D.4 are written with the raw softmax output (p_{ij} for classification, p_{zij} for segmentation) to introduce the estimators; as described below, all reported estimates are computed after MOVAL's class-wise calibration, i.e., with the raw probability replaced by its calibrated counterpart ($\hat{p}_{ij}$ or $\hat{p}_{zij}$). In this manner, we can derive all the metrics using only the network outputs. However, like accuracy estimation, metrics derived from uncalibrated probabilities may be inaccurate. Although some studies suggest that accuracy-calibrated models perform better (NannyML), we believe that this approach may be suboptimal, as accuracy calibration does not necessarily improve all metrics.

Instead, we offer a novel solution for calibrating confidence to align with these metrics. For each metric, we optimize a specific model tailored to that metric. It directly adjusts parameters to ensure that the estimated metrics align closely with the actual metrics on the validation set. This alignment is made possible by our unique MOVAL framework, which incorporates class-wise calibration with class-specific parameters.

After obtaining the calibrated probability $\hat{\boldsymbol{p}}_i$, the optimization process closely follows the accuracy calibration procedure, as class-specific parameters can naturally be extended to other metrics. Using F1-Score as an example, we perform separate calibration for the j'th class by adjusting all $\boldsymbol{p}_i$ predicted as j to $\hat{\boldsymbol{p}}_i$ with class-specific parameters, ensuring that the estimation from Eq. D.1 using $\hat{\boldsymbol{p}}_i$ aligns with the actual metrics. This process is carried out for each class individually. All the other group-wise metrics, including sensitivity, precision, and AUC, can be calibrated similarly using Eq. D.2 and Eq. D.3, respectively. Regarding AUC, we compute five thresholds at equal intervals based on the range of probabilities $\hat{p}_{ij}$ for all samples of class j (denoted as t in Eq. D.4). Using these thresholds, we then calculate the areas under the FPR'_j–TPR'_j curve to estimate the AUC of class j, i.e, Λ'_j.

Appendix D.2. Network Configurations

Here, we detail the network configurations and training procedures used to develop the model for classification and segmentation tasks. This includes specifics on neural network training and advanced strategies commonly employed to tackle class imbalance and domain generalization. These methods are established techniques widely adopted as best practices.

Appendix D.2.1. Classification

Network architecture. We train a ResNet-32 (He et al., 2016) for CIFAR10 and CIFAR100 experiments following (Cao et al., 2019). We train a ResNet-50 for skin lesion lesion classification following (Russakovsky et al., 2015; Marrakchi et al., 2021).

Network training. We trained for 200 epochs on the CIFAR experiments and 20,000 epochs for skin lesion classification. The networks are trained using stochastic gradient descent (SGD). The learning rate was set to 0.1 and reduced using step decay. The images are normalized based on the statistics of the training dataset. For the CIFAR baseline experiments, we employed data augmentation techniques such as random flipping and color transformations. For the skin lesion classification baseline experiments, we applied data augmentation strategies including cropping to 224×224 and random flipping.

Given a sample with label $y \in \{1, 2, \ldots, c\}$, the neural network generates $\boldsymbol{z}$, which is subsequently transformed into a probability vector $\boldsymbol{p}$, where the jth element p_j represents the probability of the jth class. Unless specified otherwise, the network is trained using cross-entropy (CE) loss, which is defined as:

$$\mathcal{L}_{CE} = -\log p_y = -\log \frac{e^{z_y}}{\sum_k e^{z_k}}. \tag{D.5}$$

Balanced training I: label-distribution-aware margin (LDAM) loss (Cao et al., 2019). LDAM applies unequal margins to the logits of different classes to enhance the robustness of predictions for minority classes. The calculation is defined as:

$$\mathcal{L}_{CE}^{LDAM} = -\log \frac{e^{z_y - \Delta_y}}{e^{z_y - \Delta_y} + \sum_{k \neq y} e^{z_k}}, \tag{D.6}$$

where the margins for different classes are calculated as:

$$\Delta_j = \frac{C}{n_j^{1/4}}, \tag{D.7}$$

where C is a hyper-parameter and n_j is the number of samples for class j. This encourages the network to focus on minority classes, as they require larger margins.

Balanced training II: deferred re-weighting (DRW) weighting schedule (Cao et al., 2019). DRW adjusts the weight of samples belonging to different classes. To learn good representations and achieve unbiased performance across classes, it applies uniform weighting at the beginning and then reweights the samples based on class frequency (i.e., minority classes receive larger weights) in the later stages of network training. Specifically, it optimizes with:

$$\mathcal{L}_{CE}^{DRW} = -w_y \log p_y, \tag{D.8}$$

where w_y is the weight for class y, which starts at 1 and later changes to $1/n_y$.

Robust training I: cutout (DeVries and Taylor, 2017). Cutout is a data augmentation strategy that modifies the input sample by randomly cropping out a square in the image and replacing it with black patches. This helps networks become robust to local features and improves generalization. Cutout is commonly adopted in neural networks to enhance model generalizability. We found that it significantly improves model accuracy on in-domain samples but fails to enhance performance on most out-of-domain samples, possibly because it causes neural networks to rely more on the original intensity distributions. We choose to report its performance because it is a default technique for training neural networks in classification.

Robust training II: RandAugment (Cubuk et al., 2020). RandAugment is a data augmentation strategy that applies multiple stacked transformations with predefined magnitudes, unlike baseline methods. It introduces transformations from a larger collection, improving network performance under these conditions.

Appendix D.2.2. Segmentation

Network architecture. We train a 3D U-Net (Ronneberger et al., 2015) for segmentation using the nnU-Net (Isensee et al., 2021) framework, with a base feature map size of 30 and 4 downsampling steps at the bottleneck.

Network training. The networks are trained using patches sampled from whole images, as the images are typically too large to be input into the network due to memory limitations. For cardiac MRI segmentation, we train using patches of size $128 \times 128 \times 8$ with a batch size of 30. For prostate MRI segmentation, we use patches of size $64 \times 64 \times 32$ with a batch size of 32. For brain lesion segmentation, we train with patches of size $128 \times 128 \times 128$ and a batch size of 2.

We train neural networks for 1000 epochs, in each epoch we train with patches for 100 iterations. The networks are trained using SGD, with learning rates linearly decreased from 0.01 to 0. By default, we apply Z-score normalization to MR images, ensuring that each MRI volume has a mean of zero and unit variance. Given an image patch, the label y corresponds to the center pixel and is represented as a one-hot vector. In this vector, e_j equals 1 if j is the true class, and 0 otherwise. For baseline experiments, the networks are trained with a balanced sum of cross-entropy (CE) loss and sample-wise soft DSC loss, which are defined as:

$$\mathcal{L}_{CE} = -\sum_{j=1}^{c} e_j \log p_j = -\sum_{j=1}^{c} e_j \log \frac{e^{z_j}}{\sum_k e^{z_k}}, \tag{D.9}$$

and

$$\mathcal{L}_{DSC} = \sum_{j=1}^{c} \frac{(1-e_j)p_j + e_j(1-p_j)}{p_j + e_j}. \tag{D.10}$$

Balanced training I: asymmetric large margin loss (Li et al., 2020). Asymmetric large margin loss, similar to LDAM, is better suited for segmentation tasks and has been validated with soft DSC. Specifically, we represent the rare class with a one-hot vector $\boldsymbol{r}$, where the jth element $r_j \in \{0, 1\}$ is 1 if class j is rare. For example, for brain lesion segmentation, we set $[0, 1]$ to indicate that the brain lesion class is considered rare. Based on this, the CE and soft DSC loss are modified to:

$$\mathcal{L}_{CE}^{ALM} = -\sum_{j=1}^{c} e_j \log h_j, \tag{D.11}$$

and

$$\mathcal{L}_{DSC}^{ALM} = \sum_{j=1}^{c} \frac{(1-e_j)h_j + e_j(1-h_j)}{h_j + e_j}, \tag{D.12}$$

where h_j is the modified probability:

$$h_j = \frac{e^{z_j} - e_j\Delta_j}{\sum_k e^{z_k - e_j\Delta_j}}, \tag{D.13}$$

where

$$\Delta_j = r_j C', \tag{D.14}$$

where C' is a hyper-parameter. The neural network is then optimized using the equal sum of the CE and soft DSC losses, and the same applies to the later asymmetric focal loss.

Balanced training II: asymmetric focal loss (Li et al., 2020). Asymmetric focal loss builds upon the concept of focal loss but is better suited for long-tail learning, as it removes the attenuation term for rare classes. This adjustment helps prevent the network from under-fitting minority classes. Specifically, the CE and soft DSC losses are modified as follows:

$$\mathcal{L}_{CE}^{AF} = -\sum_{j=1}^{c} r_j e_j \log p_j - \sum_{j=1}^{c} (1 - r_j)(1 - p_j)^{\gamma} e_j \log p_j, \tag{D.15}$$

and

$$\mathcal{L}_{DSC}^{AF} = \sum_{j=1}^{c} \frac{(1 - e_j)p_j + r_j e_j (1 - p_j)}{p_j + e_j} + \sum_{j=1}^{c} \frac{(1 - e_j)p_j + (1 - r_j)(1 - p_j)^{\gamma} e_j (1 - p_j)}{p_j + e_j}. \tag{D.16}$$

Robust training I: global intensity non-linear augmentation (GIN) technique (Ouyang et al., 2022a). GIN is a random kernel-based domain generalization method that transforms images into random styles. Given an image patch $\boldsymbol{x}$, it transforms it into $\tilde{\boldsymbol{x}}$ using:

$$\tilde{\boldsymbol{x}} = \frac{\beta g(\boldsymbol{x} + (1 - \beta)\boldsymbol{x})}{\|\beta g(\boldsymbol{x} + (1 - \beta)\boldsymbol{x})\|_F} \|\boldsymbol{x}\|_F, \tag{D.17}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $g(\cdot)$ represents a shadow CNN with randomly initialized parameters, and β is a parameter sampled from a uniform distribution. In this way, the network can be robust to various kinds of intensity transformations. We find that GIN helps neural networks generalize better across different appearances but performs worse with spatial transformations. Therefore, to ensure a fair comparison, we exclude all spatial transformation-based synthetic domain shifts when evaluating advanced training strategies, including GIN.

Robust training II: mixup (Zhang et al., 2017). Mixup extends the distributions of the training dataset by combining two individual samples. In

addition to $\{\boldsymbol{x}, \boldsymbol{e}\}$, it requires another pair $\{\boldsymbol{x}', \boldsymbol{e}'\}$ and creates a new sample with:

$$\tilde{\boldsymbol{x}} = \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{x}', \tilde{\boldsymbol{e}} = \lambda \boldsymbol{e} + (1-\lambda)\boldsymbol{e}', \tag{D.18}$$

where λ is a parameter sampled from a beta distribution.

Appendix D.3. Fine-Grained Performance Analysis

To evaluate the proposed algorithms, we compare the estimated results with the actual metrics and compute the MAE. For class-wise metrics beyond accuracy, we report the average MAE across different classes. To assess the significance between methods, we perform a paired t-test on the MAE values obtained under different conditions using each method. We exclude the evaluation of classes that are absent during testing, particularly in the task of skin lesion classification, where some classes may be missing at test time. For segmentation, we exclude the background class when evaluating the model's performance.

For a comprehensive analysis of the estimation results, we provide examples at the class, instance, and pixel levels. For class-wise performance, we group the confidence scores by the predicted class, calculate their averages, and compare these averages to the corresponding accuracy of the samples. In instance-wise analysis for segmentation tasks, we evaluate performance on a case-by-case basis rather than as a whole. For pixel-wise analysis, we visualize the calibrated confidence scores adjusted by the F1-Score, highlighting uncertain regions as $1 - \hat{q}_i$. For 3D confidence score visualization, we display pixels with uncertainty greater than 0.1 and set the cluster transparency to 0.005.

Appendix E. Supplementary Tables

	Accuracy	F1-Score	Sensitivity	Precision	AUC	Advanced Training	Reduced Data	Total
Classification	197	197	197	197	197	380	2375	3740
2D Segmentation	69	69	69	69	69	0	276	621
3D Segmentation	183	183	183	183	183	216	1080	2211
Total	449	449	449	449	449	596	3731	6572

Table E.3: Summary of experimental results covered in the main body.

	nv (%)	mel (%)	bcc (%)	df (%)	bkl (%)	vasc (%)	akiec (%)	Total
HAM	6705 (67)	1113 (11)	514 (5)	115 (1)	1099 (11)	142 (1)	327 (3)	10015
BCN	4206 (34)	2857 (23)	2809 (23)	124 (1)	1138 (9)	111 (1)	1168 (9)	12413
VIE	4331 (99)	34 (1)	0	0	0	0	0	4365
MSK	2202 (62)	826 (23)	30 (1)	5 (<1)	470 (13)	0	7 (<1)	3540
UDA	408 (67)	193 (31)	3 (<1)	2 (<1)	7 (1)	0	0	613
OTH	4523 (55)	1669 (20)	513 (6)	124 (2)	889 (11)	95 (1)	388 (5)	8201
D7P	1150 (60)	501 (26)	84 (4)	40 (2)	90 (5)	58 (3)	0	1923
PH2	160 (80)	40 (20)	0	0	0	0	0	200

Table E.4: Dataset statistics of skin lesion classification datasets used in this study. nv: Melanocytic nevi, mel: Melanoma, bcc: Basal cell carcinoma, df: Dermatofibroma, bkl: Benign keratosis, vasc: Vascular skin lesions, akiec: Actinic Keratoses (Solar Keratoses) and Intraepithelial Carcinoma (Bowen's disease).

Evaluation Metric and Dataset			Estimation Algorithm	Confidence Score								MOVAL Ensemble
				Class-Agnostic				Class-Specific				
				MCP	Energy	Entropy	Doctor	MCP	Energy	Entropy	Doctor	
Sensitivity	CIFAR10	Synthetic-95	Vanilla	32.3 ±5.2	31.4 ±3.2	30.6 ±5.1	35.9 ±5.7	—	—	—	—	
			TS	20.8 ±3.7	24.6 ±5.5	21.3 ±3.7	21.6 ±3.5	12.2 ±5.5	15.4 ±6.9	11.8 ±5.1	10.9 ±4.4	
			DoC	21.9 ±4.2	22.9 ±4.6	21.7 ±4.2	22.6 ±4.5	13.8 ±6.1	13.1 ±5.8	13.3 ±5.8	14.5 ±6.3	<u>9.5</u> ±3.6
			ATC	18.9 ±3.0	25.0 ±3.9	19.1 ±3.1	18.9 ±3.0	<u>9.5</u> ±3.8	10.8 ±4.2	**9.4** ±3.7	**9.4** ±3.7	
			TS-ATC	21.2 ±3.7	27.7 ±4.9	22.9 ±3.4	22.6 ±3.4	10.2 ±4.4	12.0 ±4.8	10.2 ±3.8	10.0 ±3.6	
	CIFAR100	Synthetic-95	Vanilla	41.6 ±4.6	15.1 ±4.0	52.5 ±4.7	61.9 ±4.9	—	—	—	—	
			TS	17.0 ±2.2	23.4 ±2.3	19.6 ±2.1	15.9 ±2.0	17.1 ±4.6	23.4 ±5.6	18.1 ±4.7	19.0 ±3.9	
			DoC	16.8 ±2.4	21.3 ±2.2	19.1 ±2.4	20.9 ±2.5	18.4 ±4.6	16.8 ±5.2	19.2 ±4.8	20.3 ±5.0	15.1 ±4.0
			ATC	**14.3** ±2.3	18.0 ±2.0	<u>14.4</u> ±2.2	**14.3** ±2.2	15.6 ±4.1	16.1 ±4.2	15.4 ±4.0	15.7 ±4.2	
			TS-ATC	15.4 ±2.1	20.3 ±2.1	17.5 ±2.0	16.4 ±2.0	15.6 ±4.0	16.6 ±4.7	16.0 ±3.9	15.8 ±3.8	
	HAM10000	Natural-7	Vanilla	44.4 ±11.6	34.6 ±22.5	39.5 ±11.1	41.9 ±10.9	—	—	—	—	
			TS	37.1 ±11.6	53.2 ±12.5	35.3 ±11.5	31.0 ±11.8	28.1 ±14.9	34.9 ±15.2	25.7 ±15.4	23.9 ±15.3	
			DoC	41.2 ±11.5	45.3 ±11.6	37.8 ±11.4	39.0 ±11.3	29.8 ±14.5	30.5 ±15.2	27.7 ±14.7	28.8 ±14.4	<u>23.0</u> ±15.1
			ATC	24.9 ±13.0	24.5 ±12.8	24.5 ±13.1	24.8 ±13.1	24.0 ±15.3	23.2 ±14.7	23.2 ±14.8	23.8 ±15.0	
			TS-ATC	23.6 ±11.5	32.1 ±11.3	23.7 ±12.9	23.8 ±13.0	23.6 ±15.5	23.6 ±13.8	24.2 ±14.6	**22.5** ±14.8	
Precision	CIFAR10	Synthetic-95	Vanilla	32.7 ±6.7	49.2 ±3.8	33.2 ±6.9	<u>31.5</u> ±6.2	—	—	—	—	
			TS	37.1 ±7.2	36.3 ±7.1	37.0 ±7.1	37.7 ±6.8	32.4 ±7.3	32.5 ±6.9	31.8 ±7.0	34.9 ±6.2	
			DoC	36.6 ±7.1	36.6 ±7.2	36.6 ±7.1	36.5 ±7.1	**31.3** ±7.1	32.2 ±6.9	**31.3** ±7.1	31.7 ±7.0	31.6 ±7.0
			ATC	—	—	—	—	—	—	—	—	
			TS-ATC	—	—	—	—	—	—	—	—	
	CIFAR100	Synthetic-95	Vanilla	34.6 ±1.4	26.3 ±2.4	38.8 ±1.6	45.3 ±1.8	—	—	—	—	
			TS	26.3 ±2.2	26.8 ±2.3	26.2 ±2.3	26.0 ±2.1	22.0 ±3.6	23.3 ±3.1	22.1 ±3.6	22.7 ±3.1	
			DoC	26.7 ±2.2	26.4 ±2.3	26.7 ±2.2	27.0 ±2.2	**21.8** ±3.4	22.1 ±3.4	<u>21.9</u> ±3.4	22.1 ±3.4	22.0 ±3.5
			ATC	—	—	—	—	—	—	—	—	
			TS-ATC	—	—	—	—	—	—	—	—	
	HAM10000	Natural-7	Vanilla	24.1 ±12.5	21.2 ±11.1	20.9 ±12.0	23.6 ±12.2	—	—	—	—	
			TS	18.6 ±11.7	22.5 ±13.6	<u>18.3</u> ±11.5	**18.2** ±10.8	18.8 ±9.4	20.2 ±10.6	18.6 ±9.3	22.5 ±9.9	
			DoC	20.0 ±12.4	19.7 ±12.4	19.1 ±11.9	19.9 ±12.0	19.3 ±9.5	20.0 ±10.3	19.8 ±9.8	19.2 ±9.4	18.7 ±9.4
			ATC	—	—	—	—	—	—	—	—	
			TS-ATC	—	—	—	—	—	—	—	—	

Table E.5: Evaluation of sensitivity and precision estimation for different classification tasks under various types of domain shifts based on MAE. The best and second-best results within the same category are in **bold** and <u>underlined</u> for each configuration.

Evaluation Metric and Dataset			Estimation Algorithm	Confidence Score								MOVAL Ensemble
				Class-Agnostic				Class-Specific				
				MCP	Energy	Entropy	Doctor	MCP	Energy	Entropy	Doctor	
F1-Score	CIFAR10	Synthetic-95	Vanilla	40.2 ±7.4	16.0 ±2.9	38.9 ±7.3	42.4 ±7.7	—	—	—	—	
			TS	22.3 ±6.0	28.7 ±8.1	21.6 ±6.0	19.4 ±5.9	11.8 ±5.5	15.1 ±6.8	11.3 ±5.0	10.2 ±4.1	
			DoC	26.4 ±7.1	24.9 ±6.8	25.9 ±6.9	27.3 ±7.5	14.0 ±6.4	13.1 ±5.8	13.5 ±6.2	14.9 ±6.9	**8.5** ±3.1
			ATC	17.1 ±4.4	17.2 ±6.1	17.3 ±4.5	17.1 ±4.4	**8.5** ±3.1	11.6 ±4.8	8.7 ±3.2	**8.5** ±3.1	
			TS-ATC	16.4 ±6.0	19.4 ±6.8	16.3 ±5.4	16.3 ±5.4	9.8 ±3.8	13.2 ±5.8	10.7 ±4.4	10.0 ±3.7	
	CIFAR100	Synthetic-95	Vanilla	56.6 ±4.2	15.9 ±3.1	64.6 ±4.3	70.4 ±4.4	—	—	—	—	
			TS	21.5 ±3.9	31.7 ±4.6	25.1 ±4.2	15.5 ±3.1	14.6 ±3.4	22.7 ±4.5	15.9 ±3.6	17.2 ±2.5	
			DoC	22.8 ±4.2	28.4 ±4.5	26.4 ±4.5	29.2 ±4.6	16.7 ±3.5	14.5 ±4.0	17.1 ±3.8	18.2 ±4.1	12.7 ±2.6
			ATC	14.7 ±3.1	16.5 ±2.9	14.3 ±3.0	14.7 ±3.1	12.8 ±2.6	14.1 ±2.9	**12.6** ±2.5	12.9 ±2.6	
			TS-ATC	13.9 ±4.2	19.8 ±3.4	15.9 ±2.7	14.4 ±2.6	13.3 ±2.6	14.3 ±3.0	14.2 ±2.8	13.4 ±2.6	
	HAM10000	Natural-7	Vanilla	52.6 ±15.1	29.9 ±20.2	48.3 ±14.0	50.8 ±14.1	—	—	—	—	
			TS	44.6 ±14.3	56.5 ±17.2	43.0 ±13.6	39.2 ±12.5	33.5 ±15.6	38.0 ±17.3	29.7 ±15.5	25.6 ±15.3	
			DoC	48.6 ±15.0	50.6 ±15.6	45.6 ±14.1	46.9 ±14.2	34.2 ±16.5	34.3 ±16.1	31.7 ±16.2	33.5 ±16.3	23.2 ±15.1
			ATC	33.2 ±12.7	29.9 ±11.2	32.7 ±12.7	33.5 ±12.7	23.7 ±15.5	25.5 ±13.3	23.1 ±14.4	23.5 ±15.1	
			TS-ATC	31.6 ±13.6	37.6 ±12.2	30.6 ±12.1	31.3 ±12.1	23.5 ±14.5	28.3 ±13.9	25.4 ±13.6	**22.9** ±14.9	
AUC	CIFAR10	Synthetic-95	Vanilla	6.8 ±2.8	9.4 ±1.9	6.5 ±2.8	7.2 ±2.9	—	—	—	—	
			TS	3.6 ±1.8	3.6 ±0.8	3.7 ±1.9	4.0 ±2.1	3.8 ±2.3	5.1 ±2.6	3.7 ±2.3	3.8 ±2.3	
			DoC	6.3 ±1.7	3.9 ±1.7	5.9 ±1.6	7.3 ±1.9	4.6 ±2.8	3.7 ±1.6	4.4 ±2.7	4.8 ±2.9	**3.0** ±1.1
			ATC	—	—	—	—	—	—	—	—	
			TS-ATC	—	—	—	—	—	—	—	—	
	CIFAR100	Synthetic-95	Vanilla	9.2 ±2.2	17.4 ±1.6	8.9 ±2.5	10.3 ±2.8	—	—	—	—	
			TS	10.3 ±0.6	14.0 ±1.7	9.9 ±0.7	10.3 ±1.5	7.9 ±1.7	7.1 ±2.5	6.8 ±1.7	7.0 ±1.9	
			DoC	13.7 ±1.8	11.6 ±1.3	16.2 ±2.2	18.4 ±2.5	7.2 ±2.3	7.3 ±1.4	6.6 ±2.4	6.8 ±2.6	**6.5** ±0.7
			ATC	—	—	—	—	—	—	—	—	
			TS-ATC	—	—	—	—	—	—	—	—	
	HAM10000	Natural-7	Vanilla	20.2 ±6.9	**13.7** ±5.9	19.1 ±6.6	20.3 ±6.9	—	—	—	—	
			TS	17.0 ±6.6	15.6 ±8.5	17.6 ±6.4	18.6 ±6.2	16.0 ±2.6	15.8 ±7.1	17.0 ±1.5	17.9 ±2.4	
			DoC	16.1 ±6.7	16.1 ±7.5	16.6 ±6.6	16.4 ±6.7	16.8 ±7.0	15.7 ±2.4	16.6 ±6.8	16.9 ±7.1	14.7 ±3.4
			ATC	—	—	—	—	—	—	—	—	
			TS-ATC	—	—	—	—	—	—	—	—	

Table E.6: Evaluation of F1-Score and AUC estimation for different classification tasks under various types of domain shifts based on MAE. The best and second-best results within the same category are in **bold** and underlined for each configuration.

Dataset and Training Strategy			Estimation Algorithm	Confidence Score								MOVAL Ensemble
				Class-Agnostic				Class-Specific				
				MCP	Energy	Entropy	Doctor	MCP	Energy	Entropy	Doctor	
CIFAR10	Synthetic-95	Vanilla Training	Vanilla	57.0 ±10.4	8.6 ±5.3	54.5 ±10.4	61.1 ±10.5	—	—	—	—	
			TS	28.8 ±9.6	39.4 ±9.5	27.2 ±9.9	24.2 ±11.1	27.4 ±4.4	35.9 ±6.1	24.3 ±3.7	18.9 ±3.4	
			DoC	36.1 ±10.4	32.4 ±8.5	35.0 ±10.4	37.4 ±10.5	34.1 ±5.5	28.9 ±4.5	32.9 ±5.3	35.2 ±5.6	—
			ATC	23.1 ±11.3	20.2 ±10.1	22.9 ±11.4	23.2 ±11.3	15.7 ±4.1	11.2 ±5.2	14.7 ±4.2	15.2 ±4.1	
			TS-ATC	21.5 ±9.9	18.1 ±9.1	21.4 ±11.3	21.4 ±11.4	14.3 ±3.4	**10.5** ±6.4	14.0 ±3.4	13.7 ±3.5	

Table E.7: Evaluation of accuracy estimation for different classification tasks under various types of domain shifts based on MAE. The test datasets include only samples from the 2 most minority classes. The best and second-best results within the same category are in **bold** and underlined for each configuration.

Evaluation Metric and Dataset		Estimation Algorithm	Confidence Score								MOVAL Ensemble
			Class-Agnostic				Class-Specific				
			MCP	Energy	Entropy	Doctor	MCP	Energy	Entropy	Doctor	
Sensitivity	Cardiac-2D Synthetic-65	Vanilla	15.1 ±6.5	62.3 ±9.2	12.4 ±5.9	14.5 ±6.3	—	—	—	—	
		TS	9.8 ±5.8	18.8 ±7.8	7.4 ±5.3	8.1 ±5.3	4.4 ±3.7	6.0 ±6.3	4.2 ±3.2	4.1 ±3.1	
		DoC	14.4 ±6.5	13.5 ±7.0	11.8 ±5.9	13.8 ±6.3	5.1 ±5.1	5.5 ±5.1	4.6 ±4.5	4.9 ±4.9	3.3 ±2.5
		ATC	3.2 ±2.4	18.8 ±7.8	3.3 ±2.5	3.2 ±2.4	**3.1** ±2.5	4.5 ±5.9	**3.1** ±2.5	**3.1** ±2.5	
		TS-ATC	3.5 ±5.3	18.8 ±7.8	3.5 ±2.7	3.5 ±2.8	3.4 ±3.4	6.4 ±9.1	3.4 ±3.4	3.4 ±3.5	
	Cardiac-2D Natural-4	Vanilla	11.1 ±1.5	66.2 ±2.6	9.1 ±1.5	10.7 ±1.5	—	—	—	—	
		TS	7.1 ±1.6	13.7 ±1.7	5.6 ±1.4	6.1 ±1.6	4.3 ±0.4	4.2 ±0.9	4.3 ±0.3	4.3 ±0.3	
		DoC	10.3 ±1.5	10.1 ±1.8	8.5 ±1.5	9.9 ±1.5	4.0 ±0.8	4.5 ±0.6	3.8 ±0.7	3.9 ±0.8	2.9 ±0.2
		ATC	2.9 ±0.1	13.7 ±1.7	2.9 ±0.1	2.9 ±0.1	3.6 ±0.1	4.0 ±0.1	3.6 ±0.1	3.6 ±0.1	
		TS-ATC	2.8 ±1.4	13.7 ±1.7	2.8 ±0.3	**2.7** ±0.3	3.7 ±0.1	5.5 ±0.4	3.7 ±0.1	3.7 ±0.1	
	Prostate-3D Synthetic-83	Vanilla	11.9 ±8.6	71.2 ±8.3	5.2 ±4.3	8.4 ±6.4	—	—	—	—	
		TS	7.7 ±7.1	12.8 ±10.3	**3.0** ±2.9	3.7 ±3.5	4.7 ±6.6	7.2 ±10.3	3.8 ±3.2	3.8 ±3.5	
		DoC	11.7 ±8.6	5.3 ±8.3	5.1 ±4.3	8.2 ±6.4	6.1 ±8.6	5.4 ±8.3	4.1 ±4.1	4.8 ±6.2	4.1 ±5.0
		ATC	3.7 ±4.7	14.7 ±10.5	3.7 ±4.7	3.7 ±4.7	3.5 ±2.3	3.5 ±2.3	3.5 ±2.3	3.5 ±2.3	
		TS-ATC	3.7 ±2.9	8.3 ±6.4	3.7 ±4.7	3.7 ±4.7	3.5 ±2.3	3.7 ±2.6	3.5 ±2.3	3.5 ±2.3	
	Prostate-3D Natural-5	Vanilla	11.6 ±3.8	71.9 ±4.2	6.6 ±1.7	6.8 ±4.5	—	—	—	—	
		TS	6.3 ±4.7	13.8 ±3.3	6.4 ±5.9	6.4 ±3.6	5.4 ±1.8	8.2 ±3.3	6.5 ±3.8	6.4 ±3.3	
		DoC	11.4 ±3.8	**5.2** ±3.5	6.5 ±1.7	6.7 ±4.4	5.8 ±3.8	5.3 ±3.5	6.2 ±2.5	5.6 ±1.5	5.6 ±0.7
		ATC	10.4 ±10.0	15.8 ±3.2	10.4 ±10.0	10.4 ±10.0	7.1 ±6.4	7.1 ±6.4	7.1 ±6.4	7.1 ±6.4	
		TS-ATC	10.4 ±5.9	15.4 ±10.3	10.4 ±10.0	10.4 ±10.0	7.1 ±6.4	6.6 ±6.1	7.1 ±6.4	7.1 ±6.4	
	Brain Lesion-3D Synthetic-83	Vanilla	20.2 ±6.5	75.6 ±6.9	12.6 ±5.9	16.3 ±6.4	—	—	—	—	
		TS	15.9 ±6.3	23.3 ±7.0	10.7 ±5.4	11.5 ±5.7	2.9 ±5.6	3.1 ±6.9	3.7 ±6.1	3.6 ±6.0	
		DoC	20.1 ±6.5	23.3 ±7.0	12.6 ±5.9	16.3 ±6.4	**2.8** ±6.1	3.0 ±6.9	3.2 ±5.7	3.0 ±5.7	**2.8** ±4.8
		ATC	9.4 ±4.8	23.3 ±7.0	9.4 ±4.8	9.4 ±4.8	4.4 ±7.3	4.4 ±7.3	4.4 ±7.3	4.4 ±7.3	
		TS-ATC	9.4 ±5.4	23.3 ±7.0	9.4 ±4.8	9.4 ±4.8	4.4 ±7.3	4.1 ±6.8	4.4 ±7.3	4.4 ±7.3	
	Brain Lesion-3D Natural-12	Vanilla	52.4 ±14.8	37.8 ±17.0	42.4 ±12.4	47.2 ±13.0	—	—	—	—	
		TS	47.9 ±14.5	56.5 ±16.2	40.0 ±12.3	41.4 ±12.4	34.8 ±15.5	37.8 ±17.4	29.3 ±12.7	29.6 ±12.8	
		DoC	52.3 ±14.8	56.5 ±16.2	42.4 ±12.4	47.1 ±13.0	35.5 ±15.9	37.6 ±17.3	31.1 ±13.4	32.8 ±14.2	29.3 ±14.3
		ATC	35.6 ±11.3	56.5 ±16.2	35.6 ±11.3	35.6 ±11.3	27.4 ±11.9	27.2 ±11.9	27.4 ±11.9	27.4 ±11.9	
		TS-ATC	35.6 ±12.3	56.5 ±16.2	35.6 ±11.3	35.6 ±11.3	27.4 ±11.9	**15.6** ±11.0	27.4 ±11.9	27.4 ±11.9	

Table E.8: Evaluation of sensitivity estimation for different segmentation tasks under various types of domain shifts based on MAE. The best and second-best results within the same category are in **bold** and underlined for each configuration.

Evaluation Metric and Dataset		Estimation Algorithm	Confidence Score								MOVAL Ensemble
			Class-Agnostic				Class-Specific				
			MCP	Energy	Entropy	Doctor	MCP	Energy	Entropy	Doctor	
Precision	Cardiac-2D Synthetic-65	Vanilla	9.5 ±1.5	55.0 ±3.7	**8.2** ±0.9	9.3 ±1.3	—	—	—	—	
		TS	10.8 ±1.5	17.7 ±7.7	11.0 ±2.0	11.0 ±1.9	10.1 ±0.6	52.1 ±3.6	10.1 ±1.9	9.3 ±1.9	
		DoC	10.8 ±1.6	10.3 ±1.4	11.0 ±1.3	10.9 ±1.5	9.2 ±0.6	52.1 ±3.6	9.2 ±1.0	9.0 ±0.7	8.4 ±0.8
		ATC	—	—	—	—	—	—	—	—	
		TS-ATC	—	—	—	—	—	—	—	—	
	Cardiac-2D Natural-4	Vanilla	8.6 ±0.5	56.4 ±2.0	7.5 ±0.5	8.4 ±0.3	—	—	—	—	
		TS	17.8 ±1.5	17.9 ±0.7	16.6 ±1.5	16.8 ±1.5	7.9 ±0.6	54.8 ±2.0	7.8 ±2.0	**7.0** ±1.6	
		DoC	16.7 ±1.6	24.0 ±1.6	15.7 ±1.5	16.5 ±1.6	7.5 ±0.4	54.8 ±2.0	7.8 ±2.1	7.5 ±0.4	**7.0** ±0.9
		ATC	—	—	—	—	—	—	—	—	
		TS-ATC	—	—	—	—	—	—	—	—	
	Prostate-3D Synthetic-83	Vanilla	9.3 ±3.4	11.9 ±5.8	6.5 ±2.8	8.8 ±3.2	—	—	—	—	
		TS	6.3 ±3.4	6.3 ±2.1	**2.6** ±3.1	2.7 ±3.2	7.7 ±8.4	11.2 ±5.8	6.5 ±5.4	6.6 ±6.1	
		DoC	6.3 ±2.5	9.1 ±7.4	3.3 ±2.6	4.6 ±2.6	7.4 ±9.0	11.4 ±5.8	5.8 ±4.8	6.0 ±5.6	3.1 ±4.5
		ATC	—	—	—	—	—	—	—	—	
		TS-ATC	—	—	—	—	—	—	—	—	
	Prostate-3D Natural-5	Vanilla	17.4 ±9.3	22.3 ±11.7	11.9 ±7.3	16.1 ±8.8	—	—	—	—	
		TS	6.4 ±2.6	9.1 ±7.4	4.6 ±5.7	5.1 ±5.3	8.3 ±3.0	21.4 ±11.6	6.6 ±1.0	6.1 ±1.5	
		DoC	9.2 ±8.1	8.3 ±4.6	6.7 ±7.9	8.6 ±8.7	7.4 ±3.0	21.6 ±11.6	6.4 ±1.6	5.9 ±1.4	**3.7** ±2.7
		ATC	—	—	—	—	—	—	—	—	
		TS-ATC	—	—	—	—	—	—	—	—	
	Brain Lesion-3D Synthetic-83	Vanilla	16.9 ±8.4	20.6 ±9.9	13.7 ±7.4	16.3 ±8.2	—	—	—	—	
		TS	11.3 ±5.0	20.6 ±9.9	9.9 ±7.6	5.8 ±8.6	**3.2** ±5.9	20.4 ±9.9	4.6 ±7.5	5.1 ±7.8	
		DoC	22.5 ±7.6	20.6 ±9.9	12.9 ±6.1	19.4 ±6.8	**3.2** ±5.7	20.4 ±9.9	4.6 ±7.3	4.7 ±7.1	5.7 ±7.3
		ATC	—	—	—	—	—	—	—	—	
		TS-ATC	—	—	—	—	—	—	—	—	
	Brain Lesion-3D Natural-12	Vanilla	23.6 ±8.4	38.5 ±17.3	20.1 ±8.6	23.0 ±8.6	—	—	—	—	
		TS	12.4 ±9.5	38.5 ±17.3	13.1 ±8.2	7.9 ±5.9	**6.5** ±4.9	38.3 ±17.4	12.2 ±8.7	12.4 ±8.5	
		DoC	26.1 ±13.8	38.5 ±17.3	15.2 ±12.6	23.0 ±13.7	6.8 ±4.9	38.3 ±17.4	10.8 ±8.1	9.9 ±7.4	7.9 ±5.3
		ATC	—	—	—	—	—	—	—	—	
		TS-ATC	—	—	—	—	—	—	—	—	

Table E.9: Evaluation of precision estimation for different segmentation tasks under various types of domain shifts based on MAE. The best and second-best results within the same category are in **bold** and underlined for each configuration.

Evaluation Metric and Dataset		Estimation Algorithm	Confidence Score								MOVAL Ensemble
			Class-Agnostic				Class-Specific				
			MCP	Energy	Entropy	Doctor	MCP	Energy	Entropy	Doctor	
Accuracy	Cardiac-2D Synthetic-65	Vanilla	0.8 ±0.3	80.9 ±0.2	0.7 ±0.3	0.8 ±0.3	—	—	—	—	
		TS	0.2 ±0.4	1.2 ±0.5	0.3 ±0.6	0.3 ±0.6	0.2 ±0.4	0.7 ±0.5	0.3 ±0.6	0.3 ±0.6	
		DoC	0.2 ±0.3	**0.1** ±0.2	0.3 ±0.5	0.3 ±0.5	0.2 ±0.3	0.2 ±0.2	0.3 ±0.5	0.3 ±0.5	0.3 ±0.5
		ATC	0.4 ±0.7	1.0 ±0.3	0.4 ±0.8	0.4 ±0.7	0.4 ±0.8	0.6 ±0.6	0.4 ±0.8	0.4 ±0.8	
		TS-ATC	0.4 ±0.6	1.2 ±0.5	0.4 ±0.8	0.4 ±0.8	0.5 ±0.8	0.9 ±0.4	0.5 ±0.8	0.5 ±0.8	
	Cardiac-2D Natural-4	Vanilla	0.5 ±0.0	81.9 ±0.1	0.4 ±0.0	0.5 ±0.0	—	—	—	—	
		TS	0.4 ±0.0	0.6 ±0.0	0.4 ±0.0	0.4 ±0.0	0.3 ±0.0	0.3 ±0.0	0.3 ±0.0	0.2 ±0.0	
		DoC	0.3 ±0.0	0.8 ±0.0	0.3 ±0.0	0.3 ±0.0	0.2 ±0.0	0.6 ±0.0	0.2 ±0.0	0.2 ±0.0	0.1 ±0.0
		ATC	**0.0** ±0.0	0.6 ±0.0	**0.0** ±0.0	**0.0** ±0.0	0.1 ±0.0	0.4 ±0.0	0.1 ±0.0	0.1 ±0.0	
		TS-ATC	**0.0** ±0.0	0.6 ±0.0	**0.0** ±0.0	**0.0** ±0.0	0.1 ±0.0	0.5 ±0.0	0.1 ±0.0	0.1 ±0.0	
	Prostate-3D Synthetic-83	Vanilla	0.3 ±0.2	73.9 ±0.3	0.3 ±0.6	0.4 ±0.5	—	—	—	—	
		TS	0.2 ±0.3	**0.1** ±0.2	0.4 ±0.6	0.3 ±0.6	0.2 ±0.3	**0.1** ±0.2	0.4 ±0.6	0.3 ±0.6	
		DoC	0.2 ±0.3	0.2 ±0.2	0.3 ±0.6	0.3 ±0.6	0.2 ±0.3	0.2 ±0.2	0.3 ±0.6	0.3 ±0.6	0.2 ±0.4
		ATC	0.4 ±0.7	0.4 ±0.2	0.4 ±0.7	0.4 ±0.7	0.4 ±0.7	**0.1** ±0.1	0.4 ±0.7	0.4 ±0.7	
		TS-ATC	0.4 ±0.6	0.4 ±0.6	0.4 ±0.7	0.4 ±0.7	0.4 ±0.7	0.4 ±0.6	0.4 ±0.7	0.4 ±0.7	
	Prostate-3D Natural-5	Vanilla	0.5 ±0.2	74.1 ±0.2	0.1 ±0.0	0.3 ±0.1	—	—	—	—	
		TS	**0.0** ±0.0	0.3 ±0.2	0.4 ±0.3	0.3 ±0.2	0.1 ±0.0	0.3 ±0.2	0.4 ±0.3	0.3 ±0.2	
		DoC	0.3 ±0.2	0.3 ±0.2	0.1 ±0.1	0.1 ±0.1	0.3 ±0.1	0.2 ±0.1	0.2 ±0.1	0.1 ±0.0	**0.0** ±0.0
		ATC	0.6 ±0.4	0.7 ±0.3	0.6 ±0.4	0.6 ±0.4	0.6 ±0.4	0.2 ±0.1	0.6 ±0.4	0.6 ±0.4	
		TS-ATC	0.6 ±0.3	0.4 ±0.3	0.6 ±0.4	0.6 ±0.4	0.6 ±0.4	0.6 ±0.4	0.6 ±0.4	0.6 ±0.4	
	Brain Lesion-3D Synthetic-83	Vanilla	**0.1** ±0.2	96.8 ±0.3	**0.1** ±0.4	0.2 ±0.4	—	—	—	—	
		TS	**0.1** ±0.2	**0.1** ±0.0	**0.1** ±0.5	**0.1** ±0.4	**0.1** ±0.2	**0.1** ±0.0	**0.1** ±0.5	**0.1** ±0.4	
		DoC	**0.1** ±0.2	**0.1** ±0.0	**0.1** ±0.4	**0.1** ±0.4	**0.1** ±0.2	**0.1** ±0.0	**0.1** ±0.4	**0.1** ±0.4	**0.1** ±0.3
		ATC	**0.1** ±0.5	**0.1** ±0.0	**0.1** ±0.5	**0.1** ±0.5	**0.1** ±0.5	**0.1** ±0.0	**0.1** ±0.5	**0.1** ±0.5	
		TS-ATC	**0.1** ±0.5	**0.1** ±0.0	**0.1** ±0.5	**0.1** ±0.5	**0.1** ±0.5	**0.1** ±0.0	**0.1** ±0.5	**0.1** ±0.5	
	Brain Lesion-3D Natural-12	Vanilla	0.2 ±0.1	96.5 ±0.2	**0.1** ±0.1	0.2 ±0.1	—	—	—	—	
		TS	**0.1** ±0.1	0.2 ±0.1	**0.1** ±0.1	**0.1** ±0.1	**0.1** ±0.1	0.2 ±0.1	**0.1** ±0.1	**0.1** ±0.1	
		DoC	**0.1** ±0.1	0.2 ±0.1	**0.1** ±0.1	**0.1** ±0.1	**0.1** ±0.1	0.2 ±0.1	**0.1** ±0.1	**0.1** ±0.1	**0.1** ±0.1
		ATC	**0.1** ±0.1	0.2 ±0.1	**0.1** ±0.1	**0.1** ±0.1	**0.1** ±0.1	0.2 ±0.1	**0.1** ±0.1	**0.1** ±0.1	
		TS-ATC	**0.1** ±0.1	0.2 ±0.1	**0.1** ±0.1	**0.1** ±0.1	**0.1** ±0.1	0.2 ±0.1	**0.1** ±0.1	**0.1** ±0.1	

Table E.10: Evaluation of accuracy estimation for different segmentation tasks under various types of domain shifts based on MAE. The best and second-best results within the same category are in **bold** and underlined for each configuration.

Evaluation Metric and Dataset		Estimation Algorithm	Confidence Score								MOVAL Ensemble
			Class-Agnostic				Class-Specific				
			MCP	Energy	Entropy	Doctor	MCP	Energy	Entropy	Doctor	
AUC	Cardiac-2D Synthetic-65	Vanilla	2.9 ±0.6	45.6 ±0.8	3.0 ±0.7	2.7 ±0.6	—	—	—	—	
		TS	4.3 ±1.0	3.1 ±0.8	4.3 ±1.1	4.1 ±1.0	2.5 ±0.5	33.1 ±0.6	2.4 ±0.4	**2.3** ±0.4	
		DoC	3.4 ±0.6	5.0 ±1.3	3.5 ±0.7	3.2 ±0.6	2.5 ±0.5	19.2 ±0.3	2.4 ±0.4	**2.3** ±0.4	**2.3** ±0.4
		ATC	—	—	—	—	—	—	—	—	
		TS-ATC	—	—	—	—	—	—	—	—	
	Cardiac-2D Natural-4	Vanilla	2.3 ±0.5	45.3 ±0.2	2.3 ±0.5	2.2 ±0.4	—	—	—	—	
		TS	2.8 ±0.3	2.6 ±0.5	2.7 ±0.3	2.6 ±0.3	2.2 ±0.5	32.5 ±0.2	2.2 ±0.5	2.2 ±0.5	
		DoC	2.5 ±0.4	3.2 ±0.3	2.5 ±0.4	2.4 ±0.4	2.2 ±0.5	19.1 ±0.3	2.2 ±0.5	**2.1** ±0.5	**2.1** ±0.5
		ATC	—	—	—	—	—	—	—	—	
		TS-ATC	—	—	—	—	—	—	—	—	
	Prostate-3D Synthetic-83	Vanilla	0.4 ±1.3	18.1 ±1.6	0.4 ±1.3	0.4 ±1.3	—	—	—	—	
		TS	0.6 ±1.0	0.9 ±1.1	0.5 ±1.2	0.5 ±1.2	0.4 ±1.3	17.5 ±1.5	0.4 ±1.3	0.4 ±1.3	
		DoC	**0.3** ±1.3	1.0 ±0.6	**0.3** ±1.3	**0.3** ±1.3	0.4 ±1.3	17.0 ±1.5	0.4 ±1.3	0.4 ±1.3	0.4 ±1.1
		ATC	—	—	—	—	—	—	—	—	
		TS-ATC	—	—	—	—	—	—	—	—	
	Prostate-3D Natural-5	Vanilla	1.3 ±0.9	16.6 ±0.7	1.1 ±0.9	1.3 ±0.9	—	—	—	—	
		TS	0.8 ±0.7	**0.7** ±0.6	0.8 ±0.9	0.8 ±0.8	1.3 ±0.9	16.0 ±0.7	1.2 ±0.9	1.3 ±0.9	
		DoC	1.1 ±0.9	0.8 ±0.6	1.0 ±0.9	1.1 ±0.9	1.3 ±0.9	15.5 ±0.6	1.2 ±0.9	1.3 ±0.9	0.9 ±0.9
		ATC	—	—	—	—	—	—	—	—	
		TS-ATC	—	—	—	—	—	—	—	—	
	Brain Lesion-3D Synthetic-83	Vanilla	0.9 ±1.5	2.6 ±0.5	0.7 ±1.5	0.9 ±1.5	—	—	—	—	
		TS	0.6 ±1.3	1.1 ±1.5	0.6 ±1.5	**0.5** ±1.5	**0.5** ±1.4	2.2 ±0.7	**0.5** ±1.5	**0.5** ±1.5	
		DoC	0.8 ±1.5	1.1 ±1.5	0.6 ±1.5	0.8 ±1.5	**0.5** ±1.5	2.1 ±0.7	**0.5** ±1.5	**0.5** ±1.5	**0.5** ±1.4
		ATC	—	—	—	—	—	—	—	—	
		TS-ATC	—	—	—	—	—	—	—	—	
	Brain Lesion-3D Natural-12	Vanilla	6.0 ±4.4	6.0 ±5.0	6.7 ±4.9	6.5 ±4.7	—	—	—	—	
		TS	**5.4** ±4.4	6.2 ±4.5	6.7 ±5.0	6.6 ±4.9	**5.4** ±4.3	6.2 ±5.1	6.5 ±4.9	6.1 ±4.7	
		DoC	5.9 ±4.4	6.2 ±4.5	6.6 ±4.9	6.4 ±4.7	5.5 ±4.4	6.3 ±5.1	6.5 ±4.9	6.1 ±4.7	**5.4** ±4.4
		ATC	—	—	—	—	—	—	—	—	
		TS-ATC	—	—	—	—	—	—	—	—	

Table E.11: Evaluation of AUC estimation for different segmentation tasks under various types of domain shifts based on MAE. The best and second-best results within the same category are in **bold** and underlined for each configuration.

Dataset and Training Strategy		Estimation Algorithm	Confidence Score								MOVAL Ensemble
			Class-Agnostic				Class-Specific				
			MCP	Energy	Entropy	Doctor	MCP	Energy	Entropy	Doctor	
CIFAR10 Synthetic-95	Balanced Training I	Vanilla	25.2 ±8.2	21.8 ±8.5	49.8 ±10.3	13.0 ±2.7	—	—	—	—	
		TS	27.9 ±8.0	32.9 ±11.0	27.3 ±7.8	27.7 ±8.2	15.7 ±7.7	23.5 ±11.9	14.3 ±6.9	14.4 ±7.4	
		DoC	31.5 ±10.2	32.8 ±11.0	31.9 ±10.3	29.6 ±9.1	19.9 ±9.9	23.5 ±11.9	20.4 ±10.2	16.0 ±7.8	11.5 ±6.4
		ATC	22.0 ±5.6	36.5 ±10.3	28.3 ±8.8	26.7 ±8.1	9.4 ±6.2	29.0 ±14.3	13.3 ±7.1	12.0 ±6.6	
		TS-ATC	22.3 ±7.8	36.5 ±10.3	22.2 ±5.1	25.5 ±7.4	**8.8** ±5.0	29.1 ±14.3	**8.8** ±5.2	11.6 ±6.7	
	Balanced Training II	Vanilla	36.6 ±8.2	17.3 ±4.7	35.3 ±7.9	39.0 ±8.5	—	—	—	—	
		TS	21.9 ±6.3	27.6 ±9.0	21.3 ±6.0	19.6 ±5.2	11.1 ±5.2	15.7 ±8.1	10.4 ±4.5	9.3 ±3.8	
		DoC	25.5 ±7.9	24.2 ±7.4	25.0 ±7.6	26.4 ±8.3	13.9 ±7.0	12.8 ±6.1	13.4 ±6.6	15.0 ±7.7	**8.0** ±3.2
		ATC	18.5 ±4.7	16.3 ±4.8	18.3 ±4.6	18.5 ±4.7	8.1 ±3.2	9.8 ±3.2	**8.0** ±3.2	8.1 ±3.2	
		TS-ATC	16.9 ±6.0	17.2 ±5.7	16.1 ±4.2	16.2 ±4.3	8.4 ±3.2	12.6 ±4.9	9.1 ±3.3	8.8 ±3.3	
	Robust Training I	Vanilla	39.7 ±9.1	17.3 ±4.7	38.2 ±8.9	44.4 ±10.1	—	—	—	—	
		TS	25.1 ±7.4	32.5 ±10.8	24.6 ±7.2	23.4 ±6.7	14.2 ±6.6	20.0 ±10.1	13.6 ±6.3	12.8 ±6.0	
		DoC	28.9 ±8.7	28.7 ±9.1	28.3 ±8.4	31.2 ±9.8	16.9 ±8.0	16.8 ±8.0	16.3 ±7.6	19.3 ±9.3	10.5 ±5.8
		ATC	20.8 ±5.3	18.0 ±5.3	20.2 ±5.3	20.7 ±5.3	10.8 ±6.2	12.9 ±5.9	10.6 ±6.3	**10.3** ±5.7	
		TS-ATC	18.7 ±7.2	21.5 ±6.2	17.8 ±5.8	17.4 ±6.3	11.5 ±7.0	15.2 ±6.3	11.7 ±6.5	12.2 ±6.7	
	Robust Training II	Vanilla	33.1 ±6.0	21.5 ±4.5	31.6 ±5.8	36.2 ±6.4	—	—	—	—	
		TS	20.2 ±4.9	26.1 ±7.1	19.6 ±5.0	17.7 ±5.3	10.3 ±4.6	14.9 ±6.4	9.9 ±4.4	9.5 ±4.1	
		DoC	23.2 ±5.8	22.9 ±5.9	22.5 ±5.6	24.3 ±6.2	12.2 ±5.1	12.4 ±5.4	11.6 ±4.9	13.3 ±5.7	**7.4** ±3.8
		ATC	17.5 ±4.3	17.4 ±5.5	16.4 ±4.5	17.4 ±4.3	**7.4** ±3.9	12.5 ±6.4	**7.4** ±3.8	**7.4** ±3.9	
		TS-ATC	15.5 ±5.0	19.4 ±6.3	15.9 ±4.6	16.0 ±4.7	9.2 ±5.0	14.9 ±6.9	10.6 ±6.2	10.5 ±6.2	

Table E.12: Evaluation of F1-Score estimation for CIFAR10 classification under different training strategies based on MAE. The best and second-best results within the same category are in **bold** and underlined for each configuration.

Dataset and Training Strategy		Estimation Algorithm	Confidence Score								MOVAL Ensemble
			Class-Agnostic				Class-Specific				
			MCP	Energy	Entropy	Doctor	MCP	Energy	Entropy	Doctor	
CIFAR10 Synthetic-95	Balanced Training I	Vanilla	37.3 ±9.3	8.3 ±7.2	52.4 ±9.4	20.2 ±8.9	—	—	—	—	
		TS	12.4 ±7.7	22.5 ±10.9	11.2 ±7.7	10.5 ±9.4	13.6 ±10.3	20.2 ±13.6	12.7 ±10.3	11.1 ±10.7	
		DoC	18.4 ±9.3	22.4 ±10.9	19.0 ±9.4	14.0 ±9.3	18.2 ±11.9	20.2 ±13.6	18.8 ±12.2	14.3 ±11.0	12.4 ±11.7
		ATC	**4.5** ±6.0	26.4 ±12.9	8.1 ±9.8	8.0 ±8.7	6.4 ±8.8	23.1 ±17.1	8.3 ±9.8	8.2 ±9.5	
		TS-ATC	5.8 ±7.7	26.5 ±12.9	5.3 ±5.4	7.6 ±8.0	7.5 ±8.9	23.1 ±17.1	7.3 ±9.1	8.0 ±9.1	
	Balanced Training II	Vanilla	30.5 ±7.9	31.9 ±6.6	28.3 ±7.6	34.6 ±8.2	—	—	—	—	
		TS	8.1 ±6.2	15.0 ±8.4	7.4 ±5.6	5.9 ±4.6	9.7 ±9.3	15.5 ±10.7	8.8 ±9.0	7.4 ±8.5	
		DoC	13.0 ±7.9	10.3 ±6.8	12.3 ±7.6	14.1 ±8.2	13.6 ±9.9	11.5 ±9.3	12.9 ±9.6	14.6 ±10.3	5.5 ±6.9
		ATC	5.9 ±5.4	4.3 ±2.6	5.7 ±5.2	5.9 ±5.4	6.7 ±8.6	5.2 ±5.7	6.3 ±7.8	6.6 ±8.4	
		TS-ATC	4.4 ±5.6	4.8 ±3.3	**3.9** ±2.9	**3.9** ±3.0	5.8 ±8.0	5.1 ±4.6	5.5 ±7.2	5.6 ±7.3	
	Robust Training I	Vanilla	30.8 ±7.1	30.4 ±7.0	28.3 ±6.8	38.8 ±8.3	—	—	—	—	
		TS	9.2 ±5.4	18.6 ±9.2	8.4 ±5.3	6.9 ±5.1	10.1 ±7.1	17.8 ±10.6	9.1 ±6.9	7.7 ±6.4	
		DoC	14.4 ±7.1	13.0 ±7.1	13.5 ±6.8	17.3 ±8.3	14.2 ±8.1	13.1 ±8.7	13.4 ±7.8	16.5 ±9.2	4.0 ±3.1
		ATC	5.4 ±4.4	4.8 ±3.4	4.6 ±4.4	5.2 ±4.4	4.7 ±4.1	5.0 ±2.9	4.4 ±3.7	4.3 ±3.7	
		TS-ATC	3.9 ±5.3	4.3 ±4.2	4.4 ±2.6	4.8 ±2.9	4.0 ±3.6	6.0 ±3.6	**3.5** ±2.9	**3.5** ±2.4	
	Robust Training II	Vanilla	24.9 ±4.8	36.8 ±4.5	22.2 ±4.7	29.9 ±5.1	—	—	—	—	
		TS	5.6 ±4.3	12.8 ±5.9	4.9 ±4.3	5.1 ±4.0	4.4 ±2.6	10.5 ±5.2	3.5 ±2.3	3.0 ±2.8	
		DoC	9.6 ±4.8	8.0 ±4.6	8.5 ±4.7	10.7 ±5.1	7.8 ±3.7	6.6 ±3.6	6.9 ±3.4	8.7 ±4.0	5.5 ±4.2
		ATC	4.6 ±3.7	8.8 ±6.8	4.2 ±3.5	4.5 ±3.6	**2.4** ±2.1	7.8 ±6.0	3.2 ±2.5	2.5 ±2.2	
		TS-ATC	5.1 ±4.3	10.8 ±8.2	6.3 ±4.7	6.4 ±4.8	3.9 ±3.0	8.3 ±6.3	4.9 ±3.7	4.7 ±3.5	

Table E.13: Evaluation of accuracy estimation for CIFAR10 classification under different training strategies based on MAE. The proposed CS variants underperform in accuracy estimation compared to their counterparts when models are trained using advanced balanced training strategies. We observe that balanced training strategies largely affect the model's predicted probabilities, resulting in a more significant difference in the confusion matrix when exposed to domain shifts. These changes, in particular, create difficulties for CS variants, as they rely on predictions to choose class parameters. In such cases, we recommend performing class-wise calibration with group-wise parameters (c.f. Table E.14) or alternatively tracking the F1-Score. The best and second-best results within the same category are in **bold** and underlined for each configuration.

Dataset and Training Strategy		Estimation Algorithm	Confidence Score								MOVAL Ensemble
			Class-Agnostic				Class-Specific				
			MCP	Energy	Entropy	Doctor	MCP	Energy	Entropy	Doctor	
CIFAR10 Synthetic-95	Balanced Training I	Vanilla	37.3 ±9.3	8.3 ±7.2	52.4 ±9.4	20.2 ±8.9	—	—	—	—	
		TS	12.4 ±7.7	22.5 ±10.9	11.2 ±7.7	10.5 ±9.4	12.2 ±7.4	20.7 ±9.8	11.1 ±7.5	10.6 ±9.3	
		DoC	18.4 ±9.3	22.4 ±10.9	19.0 ±9.4	14.0 ±9.3	17.6 ±8.7	20.7 ±9.8	18.2 ±8.9	13.8 ±9.1	11.5 ±6.9
		ATC	**4.5** ±6.0	26.4 ±12.9	8.1 ±9.8	8.0 ±8.7	<u>4.7</u> ±6.0	24.2 ±11.2	8.3 ±10.2	8.1 ±9.1	
		TS-ATC	5.8 ±7.7	26.5 ±12.9	5.3 ±5.4	7.6 ±8.0	5.4 ±5.2	24.2 ±11.2	5.3 ±5.4	7.8 ±8.4	
	Balanced Training II	Vanilla	30.5 ±7.9	31.9 ±6.6	28.3 ±7.6	34.6 ±8.2	—	—	—	—	
		TS	8.1 ±6.2	15.0 ±8.4	7.4 ±5.6	5.9 ±4.6	8.0 ±5.8	14.5 ±8.0	7.1 ±5.4	5.5 ±4.6	
		DoC	13.0 ±7.9	10.3 ±6.8	12.3 ±7.6	14.1 ±8.2	12.8 ±7.5	10.0 ±6.4	12.0 ±7.2	13.7 ±7.8	<u>3.7</u> ±3.0
		ATC	5.9 ±5.4	4.3 ±2.6	5.7 ±5.2	5.9 ±5.4	5.7 ±5.5	3.9 ±2.3	5.1 ±5.1	5.6 ±5.4	
		TS-ATC	4.4 ±5.6	4.8 ±3.3	3.9 ±2.9	3.9 ±3.0	4.2 ±3.9	4.6 ±3.0	**3.6** ±3.0	**3.6** ±3.0	
	Robust Training I	Vanilla	30.8 ±7.1	30.4 ±7.0	28.3 ±6.8	38.8 ±8.3	—	—	—	—	
		TS	9.2 ±5.4	18.6 ±9.2	8.4 ±5.3	6.9 ±5.1	8.6 ±4.8	17.7 ±8.7	7.7 ±4.7	6.3 ±4.5	
		DoC	14.4 ±7.1	13.0 ±7.1	13.5 ±6.8	17.3 ±8.3	13.8 ±6.6	12.4 ±6.7	12.9 ±6.2	16.5 ±7.7	3.7 ±3.0
		ATC	5.4 ±4.4	4.8 ±3.4	4.6 ±4.4	5.2 ±4.4	4.0 ±3.5	4.3 ±3.3	<u>3.5</u> ±3.4	3.9 ±3.5	
		TS-ATC	3.9 ±5.3	4.3 ±4.2	4.4 ±2.6	4.8 ±2.9	**3.3** ±2.8	4.4 ±4.1	3.9 ±2.4	4.0 ±2.5	
	Robust Training II	Vanilla	24.9 ±4.8	36.8 ±4.5	22.2 ±4.7	29.9 ±5.1	—	—	—	—	
		TS	5.6 ±4.3	12.8 ±5.9	4.9 ±4.3	5.1 ±4.0	5.1 ±3.8	12.0 ±5.4	4.6 ±3.9	5.0 ±3.6	
		DoC	9.6 ±4.8	8.0 ±4.6	8.5 ±4.7	10.7 ±5.1	8.8 ±4.4	7.6 ±4.3	7.8 ±4.3	9.9 ±4.7	6.4 ±5.1
		ATC	4.6 ±3.7	8.8 ±6.8	4.2 ±3.5	4.5 ±3.6	**3.7** ±3.0	8.6 ±6.7	4.0 ±3.1	<u>3.8</u> ±3.0	
		TS-ATC	5.1 ±4.3	10.8 ±8.2	6.3 ±4.7	6.4 ±4.8	5.0 ±3.7	10.7 ±8.2	6.0 ±4.8	6.1 ±4.9	

Table E.14: Related to Extended Table E.13. Evaluation of accuracy estimation for CIFAR10 classification under different training strategies based on MAE. To achieve stable estimation in complex scenarios, we employ a group-wise optimization strategy. Specifically, we estimate two parameters for the CS variants: one shared by the 8 classes with the highest number of predicted samples, and another shared by the remaining 2 classes. Compared to CS variants applied to individual classes, this technique reduces the number of required validation samples and mitigates the complexity of the confusion matrix (e.g., misclassifications within a group have less impact on the estimation results). In this ways, the CS variants can effectively generalize with advanced balanced training strategies. The best and second-best results within the same category are in **bold** and <u>underlined</u> for each configuration.

Dataset and Training Strategy		Estimation Algorithm	Confidence Score								MOVAL Ensemble
			Class-Agnostic				Class-Specific				
			MCP	Energy	Entropy	Doctor	MCP	Energy	Entropy	Doctor	
Brain Lesion-3D	Synthetic-42 Balanced Training I	Vanilla	24.8 ±9.7	70.2 ±10.0	21.6 ±7.5	23.3 ±8.5	—	—	—	—	
		TS	22.5 ±8.9	25.9 ±10.5	19.6 ±6.8	20.2 ±7.1	5.8 ±9.0	6.6 ±10.4	2.8 ±4.7	2.9 ±4.8	
		DoC	24.8 ±9.7	25.9 ±10.5	21.5 ±7.5	23.2 ±8.5	5.8 ±9.2	6.5 ±10.2	4.0 ±6.8	4.8 ±7.9	2.6 ±4.2
		ATC	17.1 ±5.2	25.9 ±10.5	17.1 ±5.2	17.1 ±5.2	2.7 ±4.0	2.7 ±4.0	2.7 ±4.0	2.7 ±4.0	
		TS-ATC	17.1 ±6.8	25.9 ±10.5	17.1 ±5.2	17.1 ±5.2	2.7 ±4.0	**2.5** ±3.4	2.7 ±4.0	2.7 ±4.0	
		Vanilla	26.4 ±14.6	52.4 ±12.2	20.5 ±10.0	23.4 ±12.1	—	—	—	—	
	Balanced Training II	TS	24.0 ±13.4	28.6 ±16.3	18.7 ±9.0	19.7 ±9.7	8.9 ±13.7	10.3 ±16.1	3.3 ±5.1	3.6 ±5.5	
		DoC	26.4 ±14.6	24.7 ±15.1	20.5 ±10.0	23.4 ±12.1	8.9 ±13.9	9.4 ±14.5	5.6 ±8.9	7.1 ±11.2	2.5 ±3.8
		ATC	15.9 ±6.5	28.5 ±16.3	15.9 ±6.5	15.9 ±6.5	**1.6** ±2.4	**1.6** ±2.3	**1.6** ±2.4	**1.6** ±2.4	
		TS-ATC	15.9 ±9.0	28.6 ±16.3	15.9 ±6.5	15.9 ±6.5	**1.6** ±2.4	**1.6** ±2.1	**1.6** ±2.4	**1.6** ±2.4	
		Vanilla	27.6 ±15.4	67.6 ±18.4	15.2 ±6.8	21.7 ±9.8	—	—	—	—	
	Robust Training I	TS	26.7 ±15.0	31.3 ±18.7	20.9 ±9.0	22.1 ±10.0	10.0 ±15.8	11.9 ±18.6	**2.7** ±4.8	2.9 ±5.2	
		DoC	27.5 ±15.3	31.3 ±18.7	19.6 ±8.2	21.8 ±9.8	8.9 ±14.0	11.8 ±18.5	3.0 ±5.6	5.0 ±8.3	2.8 ±4.5
		ATC	13.6 ±6.2	31.3 ±18.7	13.6 ±6.2	13.6 ±6.2	3.1 ±4.8	3.1 ±4.8	3.1 ±4.8	3.1 ±4.8	
		TS-ATC	13.6 ±9.0	31.3 ±18.7	13.6 ±6.2	13.6 ±6.2	3.1 ±4.8	3.7 ±5.8	3.1 ±4.8	3.1 ±4.8	
		Vanilla	30.5 ±17.9	63.3 ±19.3	21.2 ±11.3	26.0 ±14.3	—	—	—	—	
	Robust Training II	TS	27.5 ±16.8	33.6 ±20.3	21.8 ±11.6	22.7 ±12.4	11.7 ±17.6	13.5 ±20.2	4.5 ±7.1	4.9 ±7.6	
		DoC	30.4 ±17.9	32.1 ±20.0	21.5 ±11.4	26.0 ±14.3	11.3 ±17.0	13.2 ±19.9	6.5 ±9.9	8.7 ±13.0	3.5 ±5.5
		ATC	17.7 ±7.7	33.6 ±20.3	17.7 ±7.7	17.7 ±7.7	**2.4** ±4.3	**2.4** ±4.3	**2.4** ±4.3	**2.4** ±4.3	
		TS-ATC	17.7 ±11.6	33.6 ±20.3	17.7 ±7.7	17.7 ±7.7	**2.4** ±4.3	2.7 ±3.7	**2.4** ±4.3	**2.4** ±4.3	
	Natural-12 Balanced Training I	Vanilla	52.5 ±15.3	38.3 ±18.7	48.7 ±13.3	50.7 ±14.2	—	—	—	—	
		TS	50.2 ±14.8	53.8 ±15.9	46.6 ±12.7	47.3 ±13.0	34.5 ±15.6	35.1 ±16.6	30.3 ±12.2	30.5 ±12.3	
		DoC	52.5 ±15.3	53.8 ±15.9	48.7 ±13.3	50.6 ±14.2	33.9 ±15.7	35.0 ±16.5	31.4 ±13.4	32.5 ±14.3	29.6 ±11.5
		ATC	43.6 ±11.1	53.8 ±15.9	43.6 ±11.1	43.6 ±11.1	29.7 ±11.9	29.5 ±11.8	29.7 ±11.9	29.7 ±11.9	
		TS-ATC	43.6 ±12.7	53.8 ±15.9	43.6 ±11.1	43.6 ±11.1	29.7 ±11.9	**27.4** ±10.3	29.7 ±11.9	29.7 ±11.9	
		Vanilla	51.6 ±15.4	27.5 ±13.5	44.2 ±12.4	47.7 ±13.7	—	—	—	—	
	Balanced Training II	TS	48.7 ±14.8	54.2 ±16.5	42.1 ±11.8	43.2 ±12.3	33.9 ±15.6	36.6 ±16.9	24.5 ±10.4	25.1 ±10.6	
		DoC	51.5 ±15.4	49.8 ±15.9	44.2 ±12.4	47.6 ±13.7	34.5 ±15.4	34.8 ±16.0	29.7 ±12.1	31.8 ±13.5	22.7 ±9.7
		ATC	39.0 ±10.0	54.2 ±16.5	39.0 ±10.0	39.0 ±10.0	18.1 ±8.3	**18.0** ±8.3	18.1 ±8.3	18.1 ±8.3	
		TS-ATC	39.0 ±11.8	54.2 ±16.5	39.0 ±10.0	39.0 ±10.0	18.1 ±8.3	**18.0** ±8.3	18.1 ±8.3	18.1 ±8.3	
		Vanilla	47.7 ±14.5	45.1 ±18.9	38.1 ±12.6	43.1 ±12.8	—	—	—	—	
	Robust Training I	TS	47.0 ±14.4	50.8 ±15.3	42.5 ±12.6	43.3 ±12.8	31.6 ±15.9	32.0 ±16.5	27.6 ±13.6	27.7 ±13.6	
		DoC	47.7 ±14.5	50.8 ±15.3	41.5 ±12.8	43.1 ±12.8	30.0 ±15.1	32.0 ±16.5	26.6 ±13.3	27.5 ±13.5	28.6 ±15.0
		ATC	37.4 ±11.6	50.7 ±15.3	37.4 ±11.6	37.4 ±11.6	**25.5** ±15.3	**25.5** ±15.3	**25.5** ±15.3	**25.5** ±15.3	
		TS-ATC	37.4 ±12.6	50.8 ±15.3	37.4 ±11.6	37.4 ±11.6	**25.5** ±15.3	28.9 ±18.2	**25.5** ±15.3	**25.5** ±15.3	
		Vanilla	50.5 ±14.3	41.8 ±14.2	41.9 ±13.0	45.9 ±13.4	—	—	—	—	
	Robust Training II	TS	47.7 ±14.2	53.4 ±14.8	42.4 ±13.0	43.1 ±13.2	32.7 ±15.1	34.0 ±15.7	26.7 ±13.1	27.0 ±13.1	
		DoC	50.4 ±14.3	52.0 ±14.8	42.2 ±13.0	45.9 ±13.4	32.0 ±14.9	33.8 ±15.6	28.1 ±13.5	29.6 ±14.0	26.9 ±13.2
		ATC	39.6 ±12.2	53.4 ±14.8	39.6 ±12.2	39.6 ±12.2	**25.6** ±13.1	**25.6** ±13.1	**25.6** ±13.1	**25.6** ±13.1	
		TS-ATC	39.6 ±13.0	53.4 ±14.8	39.6 ±12.2	39.6 ±12.2	**25.6** ±13.1	26.3 ±14.2	**25.6** ±13.1	**25.6** ±13.1	

Table E.15: Evaluation of F1-Score estimation for brain lesion segmentation under different training strategies based on MAE. The best and second-best results within the same category are in **bold** and underlined for each configuration.

Estimating Metric	Group-Wise Calibration	Samples Used for Estimating					
		3000 (100%)	1500 (50%)	600 (20%)	300 (10%)	150 (5%)	90 (3%)
Accuracy	No	2.8 ±2.1	2.8 ±2.3	2.9 ±2.3	3.4 ±2.1	3.9 ±3.0	4.6 ±3.6
Sensitivity	No	9.5 ±3.6	9.7 ±3.5	11.0 ±3.6	12.2 ±4.0	16.4 ±4.4	25.2 ±8.1
Precision	No	31.6 ±7.0	32.0 ±7.2	31.7 ±6.8	32.0 ±7.0	31.9 ±6.8	32.8 ±7.2
F1-Score	No	8.5 ±3.1	8.7 ±3.0	10.5 ±3.8	11.4 ±4.2	15.4 ±4.4	20.4 ±6.2
AUC	No	3.0 ±1.1	3.0 ±1.1	3.3 ±1.4	3.6 ±1.6	3.8 ±1.8	4.0 ±2.0
Accuracy	5 Groups	2.9 ±2.1	2.9 ±2.2	3.0 ±2.4	3.2 ±2.4	3.1 ±2.2	3.9 ±3.0
Sensitivity	5 Groups	10.8 ±4.3	10.8 ±4.4	11.5 ±4.1	12.4 ±4.4	14.5 ±4.4	15.6 ±4.6
Precision	5 Groups	32.7 ±7.1	32.7 ±7.1	32.8 ±7.2	32.8 ±7.2	32.9 ±7.2	32.9 ±7.3
F1-Score	5 Groups	8.4 ±3.4	8.6 ±3.4	9.3 ±3.6	9.7 ±3.7	13.4 ±4.4	14.2 ±4.7
AUC	5 Groups	3.2 ±1.2	3.2 ±1.3	3.4 ±1.5	3.7 ±1.8	3.7 ±1.8	4.0 ±2.0

Table E.16: MAE of estimation results for CIFAR10 classification when MOVAL Ensemble is optimized with different amounts of validation data. For group-wise calibration, we split the 10 cases into 5 groups, with each group containing 2 classes.

Dataset	Samples Used for Estimating					
	19 (100%)	12 (63%)	6 (32%)	3 (16%)	2 (11%)	1 (5%)
Synthetic-65	4.3 ±5.0	4.6 ±4.4	4.7 ±4.6	5.4 ±4.2	6.2 ±3.3	8.3 ±2.4
Natural-4	2.6 ±0.4	2.6 ±0.3	3.3 ±0.3	3.6 ±0.3	4.2 ±0.2	6.5 ±0.4
Average	3.5	3.6	4.0	4.5	5.2	7.4

Table E.17: MAE of F1-Score estimation results for cardiac segmentation when MOVAL Ensemble is optimized with different amounts of validation data.

Dataset	Training Strategy	Samples Used for Estimating 9 (100%)	6 (66%)	3 (33%)	2 (22%)	1 (11%)
Synthetic-42	Vanilla	0.5 ±0.6	2.9 ±0.3	4.0 ±0.4	5.8 ±0.1	13.5 ±0.3
	Balanced Training I	0.5 ±0.5	3.4 ±0.1	3.8 ±0.1	6.5 ±0.2	14.3 ±0.4
	Balanced Training II	0.2 ±0.3	1.5 ±0.1	3.5 ±0.4	5.4 ±0.2	8.4 ±0.2
	Robust Training I	0.9 ±0.7	4.4 ±0.3	4.3 ±0.3	8.8 ±0.4	21.0 ±0.8
	Robust Training II	0.6 ±0.6	3.8 ±0.4	3.6 ±0.2	6.8 ±0.1	15.0 ±0.2
Natural-12	Vanilla	25.5 ±12.3	25.1 ±12.5	29.2 ±12.0	29.3 ±12.4	26.3 ±12.1
	Balanced Training I	29.6 ±11.5	29.1 ±11.6	31.3 ±11.6	29.5 ±11.9	27.2 ±11.3
	Balanced Training II	22.7 ±9.7	22.0 ±9.8	26.9 ±9.6	26.6 ±9.8	24.3 ±10.2
	Robust Training I	28.6 ±15.0	29.4 ±14.5	31.0 ±14.4	28.5 ±14.3	26.5 ±11.4
	Robust Training II	26.9 ±13.2	26.9 ±13.0	28.1 ±13.2	27.3 ±13.0	25.8 ±12.1
Average	Vanilla	13.0	14.0	16.6	17.6	19.9
	Balanced Training I	15.0	16.2	17.5	18.0	20.8
	Balanced Training II	11.5	11.8	15.2	16.0	16.4
	Robust Training I	14.8	16.9	17.6	18.7	23.7
	Robust Training II	13.8	15.4	15.9	17.0	20.4

Table E.18: MAE of F1-Score estimation results for brain lesion segmentation under different training strategies when MOVAL Ensemble is optimized with different amounts of validation data.

Estimating Metric	Classification Choice 1	Choice 2	Choice 3
Accuracy	MCP | CS TS-ATC	ENT | CS TS-ATC	ENE | CS ATC
Sensitivity	DOC | CS TS-ATC	ENT | CS ATC	MCP | CS ATC
Precision	DOC | CS TS	MCP | CS DoC	DOC | CS DoC
F1-Score	DOC | CS ATC	ENT | CS ATC	MCP | CS ATC
AUC	ENT | CS TS	DOC | CS TS	DOC | CS DoC
Estimating Metric	Segmentation-2D Choice 1	Choice 2	Choice 3
F1-Score	ENT | CS DoC	DOC | CS DoC	MCP | CS ATC
Sensitivity	MCP | ATC	MCP | TS-ATC	DOC | ATC
Precision	ENT | CS TS	MCP | CS DoC	ENT | CS DoC
Accuracy	MCP | ATC	ENT | CS DoC	MCP | CS TS
AUC	ENE | TS	MCP | CS TS	ENE | CS DoC
Estimating Metric	Segmentation-3D Choice 1	Choice 2	Choice 3
F1-Score	DOC | CS TS	ENT | CS TS	ENE | CS TS-ATC
Sensitivity	ENE | CS DoC	MCP | CS TS	ENE | CS TS-ATC
Precision	MCP | CS TS	MCP | CS DoC	DOC | TS
Accuracy	MCP | TS	MCP | CS TS	DOC | CS DoC
AUC	MCP | CS TS	MCP | TS	MCP | CS DoC

Table E.19: The confidence scores chosen for creating MOVAL Ensemble.

Category	Operation ID	Operation Name	Description	Range of magnitudes
Spatial transformations	1,2,3,4,5	Scaling down	Scale down the image by factor $1 + M$	[0.05, 0.15, 0.25, 0.35, 0.45]
	6,7,8,9,10	Scaling up	Scale up the image by factor $1 + M$	[0.05, 0.15, 0.25, 0.35, 0.45]
	11,12,13,14,15	RotateFrontal ACW	Rotate the image along frontal axis anticlockwise	[5°, 15°, 25°, 90°, 180°]
	16,17,18,19	RotateFrontal CCW	Rotate the image along frontal axis clockwise	[5°, 15°, 25°, 90°]
	20,21,22,23,24	RotateSagittal ACW	Rotate the image along sagittal axis anticlockwise	[5°, 15°, 25°, 90°, 180°]
	25,26,27,28	RotateSagittal CCW	Rotate the image along sagittal axis clockwise	[5°, 15°, 25°, 90°]
	29,30,31,32,33	RotateLongitudinal ACW	Rotate the image along longitudinal axis anticlockwise	[5°, 15°, 25°, 90°, 180°]
	34,35,36,37	RotateLongitudinal CCW	Rotate the image along longitudinal axis clockwise	[5°, 15°, 25°, 90°]
	38,39,40,41	Mirroring	Flip the sample in different planes	[Sagittal, Frontal, Axial, All]
Intensity transformations	42,43,44	Gamma expansion	Scale $\boldsymbol{x}$ to [0,1], then $\mathcal{O}(\boldsymbol{x}) = \boldsymbol{x}^{1+\gamma}$, and scale it back	[0.1, 0.3, 0.5]
	45,46,47	Gamma compression	Scale $\boldsymbol{x}$ to [0,1], then $\mathcal{O}(\boldsymbol{x}) = \boldsymbol{x}^{1/(1+\gamma)}$, and scale it back	[0.1, 0.3, 0.5]
	48,49,50	Inverted gamma expansion	Do the gamma compression with the inverted image	[0.1, 0.3, 0.5]
	51,52,53	Inverted gamma compression	Do the gamma expansion with the inverted image	[0.1, 0.3, 0.5]
	54,55,56	Adding intensity	$\mathcal{O}(\boldsymbol{x}) = \boldsymbol{x}$ + shift	[0.05, 0.15, 0.25]
	57,58,59	Subtracting intensity	$\mathcal{O}(\boldsymbol{x}) = \boldsymbol{x}$ - shift	[0.05, 0.15, 0.25]
	60,61,62	Scaling up intensity histogram	$\mathcal{O}(\boldsymbol{x}) = \boldsymbol{x}$ * (1 + scale)	[0.05, 0.15, 0.25]
	63,64,65	Scaling down intensity histogram	$\mathcal{O}(\boldsymbol{x}) = \boldsymbol{x}$ / (1 + scale)	[0.05, 0.15, 0.25]
	66,67,68	Increasing contrast	Reduce the image mean, then $\mathcal{O}(\boldsymbol{x}) = \boldsymbol{x}$ * (1 + scale), and add the mean value back	[0.05, 0.15, 0.25]
	69,70,71	Decreasing contrast	Reduce the image mean, then $\mathcal{O}(\boldsymbol{x}) = \boldsymbol{x}$ / (1 + scale), and add the mean value back	[0.05, 0.15, 0.25]
Noise transformations	72,73,74	Blurring	Blur the image using Gaussian filter with different standard deviations	[0.5, 0.7, 0.9]
	75,76,77	Sharpening	Sharpen the image using Laplacian of Gaussian filter with different standard deviations	[0.9, 0.7, 0.5]
	78,79,80	Adding Gaussian noise	Add Gaussian noise with different standard deviations	[0.025, 0.075, 0.125]
	81,82,83	Simulating low resolution	Scale down the image, and scale it back	[0.9, 0.7, 0.5]

Table E.20: List of all transformations that we use to generate synthetic domain shifts for segmentation. There are 24 types of operations with a total of 83 choices. For 2D segmentation, we exclude spatial transformations in the frontal and sagittal axes. For 3D segmentation with advanced training strategies, we exclude all spatial transformations as some algorithms impose strong spatial priors.