Towards Reliable Few-Shot Adaptation of Pathology Foundation Models via Conformal Prediction

Aditya Narendra[1], Subhankar Panda[1,2], Chandresh Kumar Maurya[1]

[1] Department of Computer Science and Engineering, IIT Indore, India
[2] Odisha University of Technology and Research, India
{adityanarendra, chandresh}@iiti.ac.in
subhankarpanda556@gmail.com

Abstract

Recent advances in foundation models have enabled their integration into high-stakes clinical settings, particularly in computational pathology, where domain-specialized FMs demonstrate strong generalization. However, real-world deployment is constrained by their poorly calibrated uncertainty awareness and degraded performance in low-data regimes requiring few-shot adaptation strategies, leading to unreliable and inefficient diagnostic workflows. Conformal Prediction (CP) is an uncertainty quantification framework that offers distribution-free, finite-sample coverage guarantees for ensuring safer deployment in such settings. In this work, we explore the integration of various CP methods with pathology foundation models using three few-shot adaption strategies for classification tasks across two datasets. To assess the clinical effectiveness of these approaches, we propose four novel metrics aimed at improving clinical reliability and alleviating diagnostic workload in few-shot settings. Our results demonstrate that Conformal Prediction methods enhance the reliability of pathology foundation models and offer actionable uncertainty estimates to enable safe and efficient deployment in few-shot pathological classification workflows, with the LAC method achieving the best overall performance. Code is available at https://github.com/AdiNarendra98/Few-Shot-PathCP.

Introduction

Foundation Models (FMs) have emerged as a transformative paradigm, demonstrating remarkable downstream performance across a wide spectrum of applications. Their capabilities have led to their adoption in high-stakes domains, including healthcare (Chen et al. 2023; Xie et al. 2024) where they exhibit strong potential to improve diagnostic accuracy and support clinical decision-making. One such critical area of healthcare application is computational pathology, which leverages machine learning (ML) algorithms to analyze and interpret tissue slides for enhanced cancer diagnosis and treatment. In this field, various Pathology Foundation Models (PFMs) such as Prov-GigaPath (Xu et al. 2024a), UNI (Chen et al. 2024a) and Vichrow (Vorontsov et al. 2024) have been developed in the recent years and exhibit impressive generalization across diverse pathology tasks (Lee et al. 2025; Neidlinger et al. 2025). Despite these advancements, deployment of pathology foundation models into routine clinical practice remains limited.

A significant barrier to real-world deployment is the limited availability of labeled pathology data, which is expensive and often restricted to proprietary institutional repositories. Moreover, fine-tuning such foundation models demands extensive computing resources and expert domain knowledge in pathology due to its fine-grained nature, both of which are impractical in most clinical settings. This necessitates *few-shot adaptation* of foundation models to new tasks using minimal labeled data *without full fine-tuning*, which leads to diminished predictive performance compared to fine-tuned counterparts. Furthermore, these adapted FMs are constrained by their limited capability to produce well-calibrated uncertainty estimates, as they rely solely on their maximum likelihood estimates, like other DL models. This results in unreliable and poorly calibrated probability outputs, forcing clinicians to manually evaluate multiple diagnostic possibilities and ultimately leading to reduced clinical efficiency in such high-stakes workflows.

CP offers a promising approach for uncertainty quantification (UQ) in such scenarios, offering a statistically rigorous framework with distribution-free, finite-sample coverage guarantees. Unlike conventional probabilistic approaches, CP provides actionable uncertainty estimates by producing prediction sets designed to include the true label with a user-specified confidence threshold. These compact and uncertainty-calibrated sets enable clinicians to review lesser diagnostic categories compared to exhaustive reviews prompted by unreliable single-label predictions, particularly in few-shot settings. Further, CP is particularly well-suited to medical applications, as these prediction sets not only include the most probable diagnosis but also safely include or exclude other potentially life-threatening conditions. For example, even when the most likely diagnosis is a common cold, CP methods account for serious conditions like pneumonia, COPD, or Lung cancer by including all possible outcomes in the prediction sets. Despite these advantages, limited works have explored the integration of CP methods in few-shot settings where it can improve diagnostic reliability and clinical efficiency. Building upon the above motivation, the main contributions of this work are as follows:

- To the best of our knowledge, this is the first study focused on integrating split-CP (Lei et al. 2018) methods

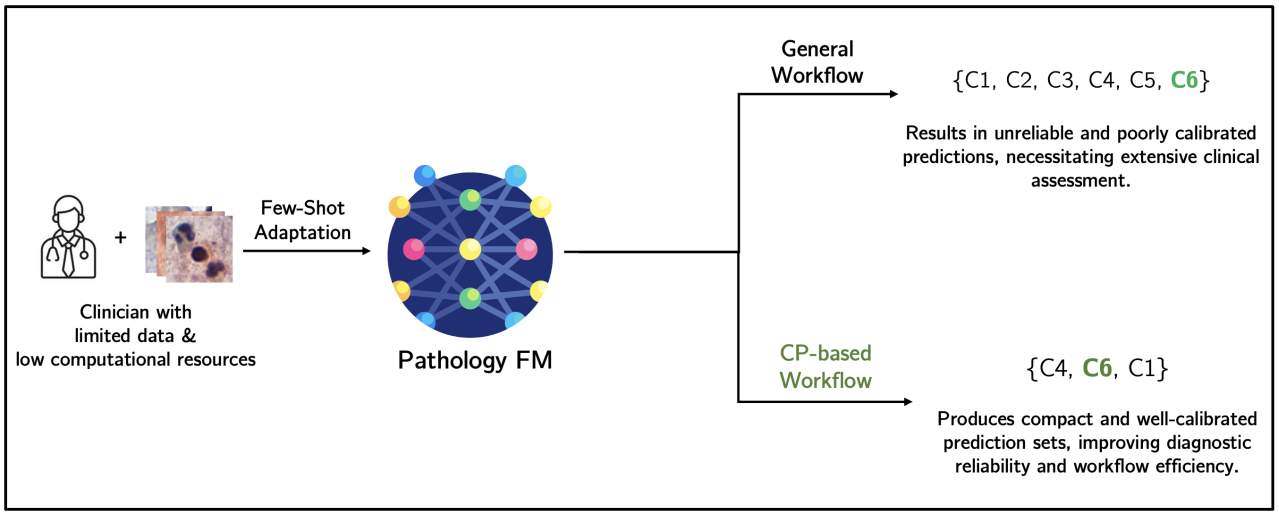

Figure 1: Comparison of the general and Conformal Prediction (CP) based workflows for few-shot adaptation of pathology foundation models in a classification task. The general workflow results in uncertain predictions and extensive clinical verification, whereas CP-based workflow produces compact and calibrated prediction sets which improve diagnostic reliability and reduce clinical workload.

with few-shot adaptation of pathology foundation models to enhance reliability and safety in classification tasks.

- We present four novel, clinically aligned evaluation metrics designed to assess the real-world effectiveness of CP approaches in few-shot diagnostic settings.

Related Works

CP is built on the foundational work by Vovk et al. (Vovk, Gammerman, and Shafer 2005), which has been explored in both regression (Romano, Patterson, and Candes 2019) and classification settings (Sadinle, Lei, and Wasserman 2019; Angelopoulos et al. 2020). Most CP methods use a split approach (Papadopoulos et al. 2002) with a held-out calibration set. CP methods have been utilized across various healthcare domains, including MRI (Lu, Angelopoulos, and Pomerantz 2022), CT scans (Angelopoulos et al. 2024; Zhan et al. 2020), and other clinical areas (Lu et al. 2022; Zhang et al. 2023). Some works have also worked on integrating CP methods into low data settings (Kumar et al. 2022; Fisch et al. 2021)

While prior studies have examined few-shot adaptation of foundation models (Xu et al. 2024b), their application to medical foundation models (Shakeri et al. 2024) remains relatively underexplored. Similarly, limited progress has been made in integrating CP with foundation models (Vishwakarma et al. 2025), particularly those developed for healthcare applications (Silva-Rodríguez, Ben Ayed, and Dolz 2025; Silva-Rodr'iguez et al. 2025). To the best of our knowledge, no existing work has investigated the integration of CP with few-shot adaptation of pathology foundation models or evaluated their effectiveness in clinical classification workflows.

Background Material

In this section, we provide an overview of Conformal Prediction (CP) and various split-CP methods, together with few-shot learning and the techniques employed in this work.

Conformal Prediction and Procedures

Conformal Prediction is a distribution and model-agnostic statistical framework that generates prediction sets containing ground truth labels with a desired coverage guarantee. A general step-by-step workflow for any conformal prediction, given an input x & output y, is outlined below:

1. Define a heuristic notion of uncertainty: For classification tasks, a typical measure for uncertainty is to consider the model's softmax outputs or logits.

2. Define the score function $S(x, y) \in \mathbb{R}$: A standard score function can be the softmax score for the true class.

3. Compute $\hat{q}$ threshold of the calibration scores: $\hat{q}$ is calculated as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the sorted conformity scores from the calibration set, where n is the number of examples in the calibration set, α is the error rate (e.g., $\alpha = 0.05$ for a 95% confidence level), and $\lceil \cdot \rceil$ denotes the ceiling function.

4. Use this $\hat{q}$ threshold to form the prediction sets C for new examples for the prediction set: For a new test input, a prediction set is obtained by including all classes whose score meets or exceeds $\hat{q}$, ensuring the true label is included with confidence $1 - \alpha$.

The various split-CP based methods used in this work and their respective workflows are mentioned below:

Least Ambiguous set-valued Classifier (LAC) : LAC (Sadinle, Lei, and Wasserman 2019) aims to construct prediction sets C from model f and calibration data using the following steps: First, we calculate the conformal score s_i

for each calibration point (X_i, Y_i), defined as $s_i = 1 - f(X_i)_{Y_i}$, the $1-$softmax score for the true class. Next, we compute the threshold $\hat{q}$ as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the sorted conformal scores. Finally, for a new test point $(X_{\text{test}}, Y_{\text{test}})$, we construct the prediction set $C(X_{\text{test}}) = \{y : f(X_{\text{test}})_y \geq 1 - \hat{q}\}$, by including classes with scores meeting or exceeding $1 - \hat{q}$.

Adaptive Prediction Sets (APS) : APS (Romano, Sesia, and Candes 2020) aims to construct prediction sets C from model f and calibration data using the following steps: First, for a calibration point (X_i, Y_i), we sort the softmax scores for all classes in descending order and calculate the conformal score $s_i = \sum_{i=1}^{k} f(X_i)$ as the sum of softmax outputs for all classes up till the true class k. Next, we compute the threshold $\hat{q}$ as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the sorted conformal scores. Finally, for a new test point $(X_{\text{test}}, Y_{\text{test}})$, we construct the prediction set $C(X_{\text{test}}) = \{f_1(X_{\text{test}}), \ldots, f_{k'}(X_{\text{test}})\}$ by including all classes until $\sum_{i=1}^{k'} f(X_{\text{test}})$ or the sum of sorted softmax scores exceeds $\hat{q}$.

Regularized Adaptive Prediction Sets (RAPS) : RAPS (Angelopoulos et al. 2020) aims to constructs prediction sets C from model f and calibration data using the following steps: First, for a calibration point (X_i, Y_i), we sort the softmax scores in descending order and calculate the conformal score as $s_i = \sum_{i=1}^{k} f(X_i) + \lambda$, where a regularization term λ is added for classes beyond a randomized threshold (k_{reg}). Next, we compute the threshold $\hat{q}$ as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the sorted conformal scores. Finally, for a new test point $(X_{\text{test}}, Y_{\text{test}})$, we construct the prediction set $C(X_{\text{test}}) = \{f_1(X_{\text{test}}), \ldots, f_{k'}(X_{\text{test}})\}$ by including all the classes, until $\sum_{i=1}^{k'} f(X_{\text{test}}) + \lambda$ or the sum of class-wise sorted softmax scores for all classes, appended with a regularization term exceeds $\hat{q}$.

Few-Shot Learning and Adaptation Strategies

The term *few-shot adaptation* refers to a specific scenario within the broader few-shot learning (FSL) paradigm, where a model adapts to new tasks without explicit fine-tuning. This is particularly beneficial in data-scarce domains such as healthcare where labeled data are limited or expensive to obtain. In the standard FSL formulation, an N-way K-shot task involves learning from K labeled samples of each of N classes (Vinyals et al. 2016). For each few-shot task, the model is trained on a randomly sampled subset of classes known as the support set (each containing K examples) and evaluated on a corresponding subset of unseen examples called the query set. The various few-shot frameworks explored in this work and their standard workflows are outlined below:

Baseline: The Baseline method (Chen et al. 2019) employs a pre-trained feature extractor that remains frozen while a linear classifier is trained on embeddings derived from a small support set. Classification is performed using a softmax operation over the linear layer outputs, with only the classifier parameters being updated. This setup enables efficient adaptation to novel classes with minimal labeled data.

In a few-shot classification task, a pre-trained pathology foundation model serves as a fixed encoder to extract feature embeddings. For each training trial, K samples per class are randomly selected from a disjoint support pool to construct the support set according to the K-shot configuration. To enhance generalization, random augmentations such as random cropping and horizontal flipping are applied to the support images. These augmented samples are then passed through the frozen encoder to obtain feature representations, upon which a linear classifier is trained using cross-entropy loss. During this stage, the encoder remains fixed and only the classifier weights are updated. Finally, the trained classifier is evaluated on a corresponding query set, randomly sampled from a disjoint query pool to assess few-shot performance.

Baseline++: The Baseline++ method (Chen et al. 2019) extends the standard Baseline framework by replacing the linear classifier with a cosine similarity-based classifier to improve feature discrimination and reduce intra-class variability. Classification is performed by computing the cosine similarity between the input feature embeddings and the class weight vectors, followed by a softmax operation to obtain class probabilities. This cosine-based formulation encourages tighter clustering of features belonging to the same class, resulting in more stable and discriminative representations in few-shot settings.

The Baseline++ method follows the same training procedure as the Baseline, except for utilizng a cosine similarity-based classifier instead of the linear classifier. The predictions are obtained by computing cosine similarities between input embeddings and class weight vectors, followed by a softmax normalization.

Prototypical Networks (ProtoNets): PrototNets (Snell, Swersky, and Zemel 2017) is a metric-based few-shot learning approach that represents each class by the mean of its support embeddings (class prototype) and classifies query samples based on their distance to these prototypes, where smaller distances indicate higher similarity. This distance based formulation allows efficient adaptation to unseen classes using only a few labeled samples, without the need for fine-tuning the encoder.

The ProtoNets method also follows the same training procedure as the Baseline, differing only in the classification strategy. Instead of training a linear classifier, we use the feature embeddings from the frozen encoder to compute the class prototypes which is the mean of the embeddings belonging to each class in the support set. During inference, predictions for query samples are obtained by computing Euclidean distances to these class prototypes, followed by a softmax over the negative distances to derive class probabilities.

Methodology

In this section, we outline the procedure used to integrate CP with few-shot adaptation techniques for classification tasks,

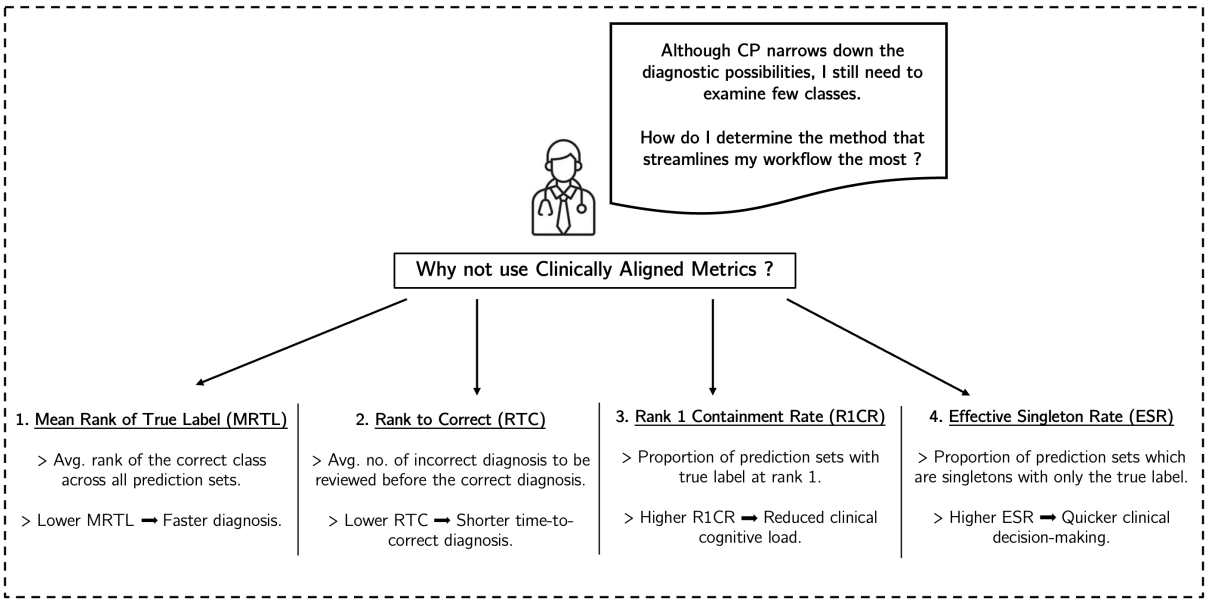

Figure 2: Overview of the proposed clinically aligned metrics (MRTL, RTC, R1CR, ESR) used to evaluate the effectiveness of Conformal Prediction methods in few-shot settings.

along with the proposed novel clinical evaluation metrics.

CP for Few-Shot Adaptation Techniques

In this subsection, we outline the step-by-step procedure for integrating the CP methods with the utilized few-shot techniques for this work.

Baseline and Baseline++ with CP : The CP integration of the Baseline and Baseline++ methods incorporates uncertainty estimation into the few-shot classification framework while retaining the frozen encoder based on a pathology foundation model and standard classifier setup. For each training trial, K samples per class are selected from a disjoint support pool to form the support set according to the K-shot configuration. Meanwhile, additional samples from separate image pools are used to construct the calibration and query sets for conformal threshold estimation and evaluation. The support samples are passed through the frozen encoder to extract feature embeddings, which are then used to train the classifier following their standard training procedures. Each calibration sample is subsequently processed through the frozen encoder and trained classifier to compute nonconformity scores based on class probabilities, which define the threshold for prediction set construction. Finally, each query image is evaluated using this threshold to generate calibrated prediction sets, yielding uncertainty-aware and well-calibrated few-shot classification performance.

Prototypical Networks with CP : The CP extension of Prototypical Networks builds upon the frozen encoder based on a pathology foundation model and prototype-based classifier to incorporate uncertainty estimation into the few-shot classification framework. Each trial begins with the pretrained encoder and class prototypes computed during the standard training procedure. A separate calibration set is sampled from a held-out pool distinct from the support and query sets. The calibration images are passed through the encoder to extract embeddings, and nonconformity scores are computed based on their distances to the corresponding class prototypes. These scores are then used to determine a conformal threshold at the desired confidence level to produce prediction sets. Finally, each query image is evaluated using this threshold to generate calibrated prediction sets, yielding uncertainty-aware and well-calibrated few-shot classification performance.

Clinically Aligned Evaluation Metrics

While standard CP metrics such as coverage and average set size quantify of the statistical guarantees and compactness of the prediction sets, they do not adequately reflect their clinical utility in few-shot diagnostic settings. The clinical effectiveness of a prediction set is critically dependent on the positional rank of the true label, as it directly influences the diagnostic workload and overall patient outcomes. For example, a compact prediction set may position the correct diagnosis at a lower rank, necessitating review of multiple incorrect labels before arriving at the true label. To address this limitation, we propose four novel clinically aligned metrics designed to assess the real-world effectiveness of CP methods in few-shot diagnostic tasks, as detailed below and shown in Figure 2.

Mean Rank of True Label (MRTL): This denotes the average positional index of the correct class label across all prediction sets. If n is the total number of test samples and rank(y_i) is the position (rank) of the true label within the prediction set, MRTL is defined as in eq 1 :

Model	CP Method	Metric	Baseline						Baseline++						ProtoNets					
			Cov:90			Cov:95			Cov:90			Cov:95			Cov:90			Cov:95		
			K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10
UNI	LAC	Accuracy	0.07±0.00	0.10±0.01	0.14±0.02	0.06±0.00	0.08±0.01	0.10±0.01	0.06±0.01	0.07±0.05	0.08±0.02	0.06±0.01	0.06±0.08	0.07±0.03	0.08±0.08	0.09±0.04	0.09±0.04	0.07±0.03	0.06±0.08	0.06±0.08
		Coverage	0.89±0.01	0.90±0.02	0.89±0.01	0.95±0.01	0.95±0.01	0.94±0.01	0.89±0.02	0.90±0.01	0.90±0.02	0.94±0.07	0.95±0.01	0.95±0.08	0.90±0.02	0.90±0.01	0.90±0.01	0.95±0.01	0.95±0.08	0.94±0.08
		Avg. Set Size	17.1±0.5	12.5±0.9	10.2±0.6	18.9±0.4	15.5±0.9	13.3±0.8	16.5±0.7	16.1±1.2	15.8±1.2	18.8±0.7	18.6±0.6	17.9±0.7	18.9±0.2	18.4±0.0	18.3±0.1	19.5±0.5	19.2±0.4	19.1±0.2
	APS	Accuracy	0.06±0.00	0.09±0.01	0.11±0.01	0.06±0.00	0.07±0.01	0.08±0.01	0.06±0.00	0.06±0.01	0.07±0.03	0.06±0.00	0.06±0.03	0.07±0.01	0.08±0.01	0.09±0.01	0.09±0.02	0.07±0.01	0.06±0.03	0.06±0.00
		Coverage	0.93±0.01	0.92±0.01	0.94±0.01	0.98±0.01	0.96±0.01	0.97±0.01	0.92±0.02	0.91±0.02	0.92±0.04	0.97±0.01	0.97±0.02	0.97±0.06	0.92±0.01	0.92±0.03	0.93±0.04	0.95±0.09	0.96±0.08	0.96±0.05
		Avg. Set Size	18.1±0.6	14.1±0.8	12.6±0.6	20.0±0.5	16.7±0.9	15.1±0.6	17.3±0.9	17.4±1.1	17.1±0.2	19.6±0.6	19.3±0.3	19.1±0.2	18.8±0.6	18.4±0.8	18.3±0.2	19.8±0.4	19.6±0.4	19.5±0.6
	RAPS	Accuracy	0.06±0.00	0.08±0.01	0.09±0.01	0.06±0.00	0.07±0.01	0.08±0.01	0.06±0.01	0.06±0.02	0.07±0.01	0.06±0.04	0.06±0.01	0.06±0.08	0.08±0.09	0.09±0.04	0.09±0.02	0.06±0.01	0.06±0.08	0.09±0.02
		Coverage	0.91±0.01	0.90±0.01	0.90±0.02	0.95±0.01	0.95±0.01	0.94±0.01	0.90±0.02	0.90±0.04	0.91±0.01	0.95±0.02	0.91±0.02	0.96±0.05	0.91±0.02	0.90±0.04	0.91±0.02	0.94±0.01	0.96±0.05	0.91±0.02
		Avg. Set Size	17.4±0.6	13.4±0.9	11.1±0.7	19.2±0.5	16.2±0.9	14.3±0.9	17.0±0.7	16.5±1.2	16.2±1.0	18.8±1.0	18.9±1.0	18.8±1.0	18.3±0.2	18.2±0.6	18.2±0.5	19.4±0.2	19.5±0.4	19.4±0.2
Phikon	LAC	Accuracy	0.06±0.00	0.08±0.01	0.11±0.01	0.06±0.00	0.07±0.01	0.08±0.01	0.06±0.01	0.06±0.01	0.06±0.08	0.06±0.03	0.06±0.02	0.06±0.04	0.16±0.02	0.22±0.02	0.24±0.01	0.10±0.06	0.11±0.08	0.11±0.08
		Coverage	0.90±0.01	0.90±0.02	0.90±0.04	0.95±0.00	0.94±0.07	0.95±0.02	0.89±0.08	0.90±0.01	0.90±0.02	0.94±0.07	0.95±0.02	0.95±0.02	0.90±0.01	0.90±0.02	0.90±0.02	0.93±0.01	0.94±0.02	0.95±0.02
		Avg. Set Size	17.0±0.5	13.4±0.6	11.3±0.7	18.8±0.4	14.5±0.7	11.9±0.7	17.4±1.1	16.8±1.6	16.1±1.6	19.2±1.0	18.4±1.4	18.1±0.4	14.8±0.6	14.6±0.4	14.5±0.7	16.8±0.5	16.6±0.7	16.2±0.8
	APS	Accuracy	0.06±0.00	0.07±0.01	0.09±0.01	0.06±0.00	0.06±0.00	0.07±0.01	0.06±0.01	0.06±0.02	0.06±0.06	0.06±0.02	0.06±0.02	0.06±0.02	0.10±0.06	0.11±0.08	0.11±0.08	0.06±0.01	0.06±0.02	0.12±0.02
		Coverage	0.93±0.02	0.92±0.01	0.93±0.01	0.96±0.01	0.97±0.02	0.97±0.01	0.92±0.02	0.93±0.03	0.93±0.02	0.97±0.02	0.97±0.02	0.97±0.05	0.93±0.01	0.94±0.02	0.94±0.01	0.96±0.01	0.97±0.05	0.97±0.01
		Avg. Set Size	18.0±0.7	14.5±0.6	12.7±0.6	19.9±0.5	17.2±0.6	15.6±0.6	18.7±0.8	17.9±0.3	17.5±0.1	19.7±1.0	19.7±0.9	19.7±1.0	16.8±0.5	16.6±0.7	16.2±0.8	18.1±0.5	17.9±0.4	17.5±0.7
	RAPS	Accuracy	0.06±0.00	0.07±0.01	0.08±0.01	0.06±0.00	0.06±0.00	0.07±0.00	0.06±0.04	0.06±0.04	0.06±0.08	0.06±0.03	0.06±0.00	0.06±0.00	0.10±0.06	0.11±0.08	0.12±0.20	0.06±0.00	0.06±0.01	0.12±0.10
		Coverage	0.91±0.01	0.91±0.03	0.92±0.04	0.96±0.03	0.96±0.03	0.96±0.03	0.90±0.02	0.91±0.01	0.91±0.02	0.95±0.03	0.95±0.02	0.95±0.02	0.91±0.02	0.91±0.04	0.92±0.01	0.96±0.03	0.95±0.04	0.95±0.01
		Avg. Set Size	17.3±0.6	13.7±0.6	11.9±0.7	19.2±0.5	16.5±0.6	15.0±0.7	17.9±0.5	17.1±0.2	17.1±0.8	18.8±1.0	19.0±0.7	18.8±1.0	16.2±0.9	15.1±0.8	15.1±0.3	17.2±0.9	16.8±0.3	16.1±0.2
Hibou-B	LAC	Accuracy	0.06±0.01	0.09±0.01	0.13±0.01	0.06±0.00	0.07±0.01	0.09±0.01	0.06±0.01	0.06±0.03	0.06±0.02	0.06±0.00	0.06±0.02	0.06±0.00	0.07±0.02	0.08±0.02	0.08±0.01	0.80±0.03	0.12±0.02	0.14±0.02
		Coverage	0.89±0.08	0.90±0.03	0.90±0.08	0.94±0.07	0.94±0.08	0.95±0.01	0.89±0.02	0.90±0.01	0.90±0.01	0.94±0.08	0.95±0.02	0.95±0.01	0.90±0.01	0.90±0.03	0.90±0.02	0.94±0.08	0.95±0.01	0.95±0.03
		Avg. Set Size	17.0±0.7	12.8±0.8	10.5±0.8	18.9±0.6	15.9±0.9	13.6±0.9	16.54±0.67	16.36±0.14	16.13±0.28	18.64±0.6	18.52±0.7	18.49±0.6	18.4±0.6	18.1±0.8	18.1±0.3	16.9±0.5	15.8±0.2	14.8±0.6
	APS	Accuracy	0.06±0.00	0.08±0.01	0.10±0.01	0.06±0.00	0.07±0.01	0.08±0.01	0.06±0.01	0.06±0.01	0.06±0.02	0.06±0.00	0.06±0.01	0.06±0.00	0.18±0.02	0.28±0.02	0.31±0.02	0.07±0.10	0.07±0.09	0.08±0.04
		Coverage	0.92±0.01	0.93±0.04	0.93±0.06	0.97±0.01	0.96±0.08	0.97±0.03	0.93±0.02	0.93±0.01	0.93±0.01	0.97±0.01	0.97±0.02	0.97±0.02	0.93±0.01	0.94±0.01	0.94±0.20	0.96±0.02	0.96±0.03	0.96±0.05
		Avg. Set Size	18.1±0.8	14.0±0.8	12.4±0.7	20.0±0.6	16.9±0.9	15.1±0.8	17.1±0.67	16.92±0.19	16.18±0.28	19.76±0.63	19.74±0.95	19.68±0.06	18.5±2.4	13.3±0.6	12.1±0.6	19.9±0.5	19.7±0.4	19.5±0.9
	RAPS	Accuracy	0.06±0.00	0.07±0.01	0.08±0.01	0.06±0.00	0.07±0.00	0.08±0.01	0.06±0.01	0.06±0.01	0.06±0.02	0.06±0.01	0.06±0.01	0.06±0.01	0.06±0.40	0.07±0.10	0.08±0.20	0.06±0.00	0.06±0.01	0.06±0.01
		Coverage	0.91±0.01	0.91±0.04	0.91±0.07	0.95±0.03	0.96±0.03	0.95±0.01	0.91±0.02	0.91±0.02	0.91±0.03	0.95±0.03	0.95±0.02	0.95±0.03	0.90±0.05	0.90±0.08	0.91±0.10	0.95±0.02	0.95±0.03	0.95±0.04
		Avg. Set Size	17.4±0.8	13.4±1.0	11.1±0.7	19.2±0.6	16.4±0.8	14.4±0.8	16.95±0.75	16.49±1.20	16.32±1.01	18.98±0.61	18.94±0.07	18.77±0.03	17.9±0.5	17.3±0.9	16.8±0.4	18.6±0.4	16.8±0.3	18.3±0.6

Table 1: Comparison of Accuracy, Emprical Coverage and Average Set Size for Baseline, Baseline++, and ProtoNets methods with three CP methods and pathology foundation models across varying coverage and K-shot settings on the HiCervix dataset.

$$\text{MRTL} = \frac{1}{n}\sum_{i=1}^{n} \text{rank}(y_i) \tag{1}$$

This metric indicates how highly a CP method prioritizes the correct diagnosis within its prediction set, enabling the clinician to view the right diagnosis at the earliest. A lower MRTL implies that, the CP method consistently assigns a higher rank to the true label in the prediction set. This helps to shorten diagnostic turnaround time and streamline clinical decision-making.

Rank to Correct (RTC): This is an accompanying metric with MRTL which denotes the average number of incorrect labels preceding the correct labels within the prediction sets. RTC is defined as in eq 2 :

$$\text{RTC} = \frac{1}{n}\sum_{i=1}^{n} \big(\text{rank}(y_i) - 1\big) \tag{2}$$

This metric represents a measure of diagnostic effort, indicating the average number of incorrect diagnosis a clinician must review before arriving at the correct diagnosis. A lower RTC implies that, the CP method effectively prioritizes the true diagnosis in its prediction set. This helps to shorten time-to-correct diagnosis and results in efficient clinical workflows.

Rank 1 Containment Rate (R1CR): Rank 1 Containment Rate (R1CR) is the proportion of test samples for which the true label appears at highest ranking position (rank 1) within the prediction set. If rank(y_i) is the position (rank) of the true label within the prediction set, R1CR is defined as in eq 3:

$$\text{R1CR} = \frac{1}{n}\sum_{i=1}^{n} \delta[\,\text{rank}(y_i) = 1\,] \tag{3}$$

where n is the total number of test samples and δ denotes an indicator function.

This metric indicates how consistently the CP method presents the correct diagnosis at the highest positional index (rank 1) in the prediction sets, enabling the clinician to identify the right diagnosis with minimal effort. A higher R1CR implies that, the CP method reliably places the true diagnosis at the highest position in the prediction sets. This results to reduce cognitive workload burden and diagnostic ambiguity for clinicians.

Effective Singleton Rate (ESR): Effective Singleton Rate denotes the proportion of test samples for which the prediction set is singleton (i.e., contains exactly one label) which is the true label. For (x_i, y_i) representing a test sample and its true label, C(x_i) representing its predicted conformal set and n is total number of test samples, ESR is defined as in eq 4:

$$\text{ESR} = \frac{1}{n}\sum_{i=1}^{n} \delta[\,|\mathcal{C}(x_i)| = 1 \;\wedge\; y_i \in \mathcal{C}(x_i)\,] \tag{4}$$

where $|\mathcal{C}(x_i)|$ denotes the cardinality of the prediction set for the i^{th} sample, and δ denotes an indicator function.

This metric represents how frequently the CP method produces a singleton prediction set containing only the correct diagnosis, enabling the clinician to arrive at the right diagnosis immediately. A higher ESR implies that the CP method consistently produces precise and reliable prediction sets. This results in faster clinical decision-making and improved patient outcomes.

Experimental Setup

This section outlines the datasets, pathology foundation models, implementation, and standard metrics for evaluation across the two pathological classification tasks in this study.

Model	CP Method	Metric	Baseline						Baseline++						ProtoNets					
			Cov:90			Cov:95			Cov:90			Cov:95			Cov:90			Cov:95		
			K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10
UNI	LAC	MRTL	9.5±0.3	7.3±0.4	6.1±0.3	10.0±0.2	8.4±0.4	7.4±0.4	10.1±0.8	9.7±0.8	9.2±0.1	10.1±0.3	10.2±0.4	10.0±0.3	9.7±0.3	8.7±0.2	8.4±0.2	9.8±0.3	9.9±0.1	9.9±0.1
		RTC	8.5±0.3	6.3±0.4	5.1±0.3	9.0±0.2	7.4±0.4	6.4±0.4	9.1±0.8	8.7±0.8	8.2±0.1	9.1±0.3	9.2±0.4	9.0±0.3	8.7±0.3	7.7±0.2	7.4±0.2	8.8±0.3	8.9±0.1	8.9±0.1
		R1CR	0.07±0.00	0.10±0.01	0.14±0.02	0.06±0.00	0.08±0.01	0.10±0.01	0.06±0.01	0.06±0.04	0.07±0.01	0.06±0.00	0.06±0.00	0.06±0.00	0.05±0.01	0.06±0.03	0.06±0.03	0.06±0.02	0.06±0.03	0.06±0.03
		ESR	0.00±0.01	0.01±0.01	0.03±0.01	0.00±0.01	0.01±0.04	0.02±0.05	0.00±0.01	0.00±0.01	0.00±0.04	0.00±0.00	0.00±0.00	0.01±0.01	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.01
	APS	MRTL	10.0±0.3	7.9±0.4	7.1±0.3	10.5±0.2	9.0±0.4	8.2±0.3	9.9±0.4	9.9±0.3	9.2±0.5	9.9±0.4	10.4±0.8	10.2±0.3	9.8±0.1	8.9±0.4	8.5±0.2	9.8±0.3	9.9±0.8	9.9±0.9
		RTC	9.0±0.3	6.9±0.4	6.1±0.3	9.5±0.2	8.0±0.4	7.2±0.3	8.9±0.4	8.9±0.3	8.2±0.5	8.8±0.4	9.4±0.8	9.2±0.3	8.8±0.1	7.9±0.4	7.5±0.2	8.8±0.3	8.9±0.8	8.9±0.9
		R1CR	0.06±0.00	0.09±0.01	0.11±0.01	0.06±0.00	0.07±0.01	0.08±0.01	0.06±0.00	0.06±0.01	0.07±0.02	0.06±0.00	0.06±0.01	0.06±0.08	0.06±0.01	0.06±0.02	0.06±0.02	0.06±0.01	0.06±0.05	0.06±0.03
		ESR	0.00±0.00	0.01±0.00	0.02±0.01	0.00±0.00	0.02±0.02	0.02±0.08	0.00±0.00	0.00±0.00	0.01±0.02	0.00±0.00	0.00±0.00	0.00±0.01	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.01±0.01	0.00±0.01
	RAPS	MRTL	9.8±0.3	7.8±0.5	6.6±0.4	10.3±0.2	8.9±0.4	8.0±0.4	9.8±0.4	9.6±0.6	10.4±0.6	10.3±0.3	10.3±0.2	10.2±0.5	9.8±0.2	9.6±0.3	9.5±0.1	9.9±0.1	9.8±0.1	9.8±0.1
		RTC	8.8±0.3	6.8±0.5	5.6±0.4	9.3±0.2	7.9±0.4	7.0±0.4	8.8±0.4	8.6±0.6	9.4±0.6	9.3±0.3	9.3±0.2	9.2±0.5	8.8±0.2	8.6±0.3	8.5±0.1	8.9±0.1	8.8±0.1	8.8±0.1
		R1CR	0.06±0.00	0.08±0.01	0.09±0.01	0.06±0.00	0.07±0.01	0.08±0.01	0.06±0.01	0.06±0.03	0.07±0.01	0.06±0.00	0.06±0.00	0.06±0.08	0.06±0.00	0.06±0.03	0.06±0.01	0.06±0.00	0.06±0.08	0.06±0.01
		ESR	0.00±0.00	0.01±0.02	0.02±0.00	0.00±0.00	0.01±0.01	0.01±0.09	0.00±0.00	0.00±0.00	0.00±0.02	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.02	0.00±0.02	0.00±0.00	0.00±0.00	0.00±0.00
Phikon	LAC	MRTL	9.3±0.3	7.6±0.3	6.5±0.3	9.9±0.2	8.7±0.3	7.9±0.3	9.1±0.7	9.1±0.4	9.0±0.1	10.1±0.4	10.0±0.8	10.1±0.2	8.5±1.1	5.4±0.2	4.7±0.1	9.5±0.6	5.6±0.3	4.9±0.1
		RTC	8.3±0.3	6.6±0.3	5.5±0.4	8.9±0.2	7.7±0.3	6.9±0.3	8.1±0.7	8.1±0.4	8.0±0.1	9.1±0.4	9.0±0.8	9.1±0.2	7.5±1.1	4.4±0.2	3.7±0.1	8.5±0.6	4.6±0.3	3.9±0.1
		R1CR	0.06±0.00	0.08±0.01	0.11±0.01	0.06±0.00	0.07±0.01	0.08±0.01	0.06±0.01	0.06±0.01	0.06±0.01	0.06±0.03	0.06±0.01	0.06±0.00	0.10±0.02	0.30±0.02	0.40±0.02	0.17±0.02	0.27±0.02	0.30±0.02
		ESR	0.00±0.00	0.02±0.002	0.06±0.05	0.00±0.00	0.00±0.01	0.02±0.02	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.01±0.01	0.01±0.02	0.02±0.01	0.00±0.00	0.01±0.08	0.04±0.05
	APS	MRTL	9.9±0.3	8.0±0.3	7.2±0.3	10.4±0.2	9.2±0.3	8.4±0.3	10.4±0.4	10.3±0.2	10.2±0.4	10.6±0.4	10.4±0.4	10.4±0.2	7.7±0.4	7.6±0.3	7.2±0.3	9.8±0.2	9.8±0.3	8.5±0.8
		RTC	8.9±0.3	7.0±0.3	6.2±0.3	9.4±0.2	8.1±0.3	7.4±0.3	9.3±0.2	9.3±0.2	9.2±0.4	9.6±0.4	9.4±0.4	9.4±0.2	6.7±0.4	6.6±0.3	6.5±0.6	8.8±0.2	8.8±0.3	8.5±0.8
		R1CR	0.06±0.00	0.07±0.01	0.09±0.01	0.06±0.00	0.06±0.20	0.07±0.01	0.06±0.00	0.06±0.00	0.06±0.01	0.06±0.00	0.06±0.00	0.06±0.00	0.08±0.01	0.11±0.01	0.09±0.01	0.01±0.01	0.01±0.02	0.04±0.02
		ESR	0.00±0.00	0.01±0.02	0.03±0.08	0.00±0.00	0.00±0.00	0.02±0.01	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.02±0.01	0.01±0.08	0.02±0.08	0.00±0.00	0.00±0.01	0.01±0.08
	RAPS	MRTL	9.8±0.3	7.9±0.3	7.0±0.4	10.3±0.2	9.1±0.4	8.3±0.3	10.3±0.9	9.7±0.4	9.4±0.9	10.3±0.6	10.3±0.6	10.3±0.6	10.1±0.4	7.8±0.4	7.0±0.4	10.3±0.2	8.8±0.4	8.3±0.3
		RTC	8.8±0.3	6.9±0.3	6.0±0.4	9.3±0.2	8.0±0.3	7.3±0.3	9.3±0.9	8.7±0.4	8.4±0.9	9.3±0.6	9.3±0.6	9.3±0.6	9.1±0.4	6.8±0.4	6.0±0.4	9.3±0.2	7.8±0.4	7.3±0.3
		R1CR	0.06±0.00	0.07±0.01	0.08±0.01	0.06±0.00	0.07±0.00	0.08±0.01	0.06±0.01	0.06±0.01	0.06±0.01	0.06±0.00	0.06±0.00	0.06±0.00	0.03±0.01	0.00±0.00	0.03±0.01	0.02±0.01	0.00±0.00	0.02±0.01
		ESR	0.00±0.00	0.00±0.00	0.01±0.01	0.00±0.00	0.00±0.00	0.00±0.06	0.00±0.00	0.00±0.00	0.00±0.01	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.01	0.00±0.00	0.00±0.00	0.00±0.00
Hibou-B	LAC	MRTL	9.4±0.4	7.4±0.4	6.2±0.4	10.0±0.3	8.6±0.4	7.5±0.4	9.48±0.41	9.45±0.79	10.05±0.80	10.09±0.34	10.24±0.37	10.43±0.40	9.8±0.3	8.8±0.2	8.6±0.2	9.5±0.3	5.7±0.2	4.9±0.8
		RTC	8.4±0.4	6.4±0.4	5.2±0.4	9.0±0.3	7.6±0.4	6.5±0.4	8.37±0.41	8.33±0.78	8.92±0.78	9.03±0.34	9.18±0.36	9.37±0.39	8.8±0.3	7.8±0.2	7.6±0.2	8.5±0.3	4.7±0.2	3.9±0.8
		R1CR	0.06±0.01	0.09±0.01	0.13±0.01	0.06±0.00	0.07±0.01	0.09±0.01	0.06±0.01	0.06±0.01	0.06±0.01	0.06±0.00	0.06±0.00	0.06±0.00	0.01±0.02	0.02±0.02	0.03±0.02	0.04±0.02	0.07±0.02	0.90±0.02
		ESR	0.00±0.01	0.01±0.05	0.03±0.01	0.00±0.00	0.01±0.03	0.01±0.05	0.000±0.001	0.000±0.000	0.000±0.001	0.000±0.000	0.000±0.000	0.000±0.000	0.00±0.00	0.02±0.02	0.05±0.09	0.00±0.00	0.00±0.00	0.00±0.00
	APS	MRTL	9.9±0.4	7.9±0.4	7.0±0.4	10.4±0.2	9.1±0.4	8.2±0.4	9.48±0.41	9.45±0.79	10.05±0.80	10.39±0.25	10.40±0.34	10.51±0.34	9.1±1.1	5.9±0.7	4.9±0.1	10.1±0.2	9.9±0.04	9.2±0.3
		RTC	8.8±0.4	6.9±0.4	6.0±0.4	9.4±0.2	8.0±0.4	7.1±0.4	8.37±0.41	8.33±0.78	8.92±0.78	9.37±0.25	9.37±0.35	9.49±0.34	8.1±1.1	4.9±0.7	3.9±0.1	9.1±0.2	9.3±0.1	8.2±0.3
		R1CR	0.06±0.00	0.08±0.01	0.10±0.01	0.06±0.00	0.07±0.01	0.10±0.01	0.06±0.01	0.06±0.01	0.06±0.01	0.06±0.00	0.06±0.00	0.06±0.00	0.08±0.02	0.10±0.30	0.11±0.08	0.01±0.02	0.02±0.01	0.02±0.08
		ESR	0.00±0.00	0.01±0.01	0.02±0.06	0.00±0.00	0.00±0.00	0.02±0.01	0.000±0.001	0.000±0.000	0.000±0.001	0.000±0.000	0.000±0.000	0.000±0.000	0.01±0.02	0.04±0.02	0.05±0.09	0.00±0.00	0.00±0.01	0.01±0.01
	RAPS	MRTL	9.8±0.5	7.9±0.5	6.7±0.3	10.3±0.3	9.1±0.4	8.3±0.3	9.78±0.41	9.58±0.61	10.43±0.62	10.29±0.31	10.28±0.52	10.51±0.47	10.1±0.4	8.8±0.4	8.6±0.4	10.4±0.2	10.3±0.1	10.3±0.1
		RTC	8.8±0.4	6.9±0.5	5.7±0.4	9.3±0.3	8.1±0.4	7.1±0.4	8.67±0.40	8.46±0.61	9.31±0.61	9.24±0.31	9.23±0.51	9.45±0.46	9.1±0.4	7.8±0.4	7.6±0.4	9.4±0.2	9.3±0.1	9.3±0.1
		R1CR	0.06±0.00	0.07±0.01	0.08±0.01	0.06±0.00	0.06±0.00	0.07±0.00	0.06±0.01	0.06±0.01	0.06±0.01	0.06±0.00	0.06±0.00	0.06±0.00	0.03±0.01	0.04±0.20	0.05±0.08	0.02±0.03	0.02±0.06	0.02±0.06
		ESR	0.00±0.00	0.01±0.01	0.01±0.03	0.00±0.00	0.00±0.00	0.00±0.01	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.01±0.01	0.05±0.08	0.00±0.00	0.00±0.00	0.00±0.00

Table 2: Comparison of MRTL, RTC, R1CR, and ESR for Baseline, Baseline++, and ProtoNets methods with three CP methods and pathology foundation models across varying coverage and K-shot settings on the HiCervix dataset.

Datasets

We utilize two publicly available pathology datasets namely, HiCervix (Cai et al. 2024) and HMU-GC-HE-30K (Lou et al. 2025) to assess few-shot learning performance across distinct pathological classification tasks derived from different organs.

The HiCervix dataset is the largest publicly-available cervical cytology dataset with over 40,229 images cell images extracted from whole slide images and consists of 21 classes on the second level of annotation which we use for cervical cancer cell classification in this study. The HMU-GC-HE-30K dataset contains over 30,000 Hematoxylin and Eosin (H&E) stained gastric tissue images divided into 9 categories and is used for a gastric tissue classification task.

Pathology Foundation Models

This section provides an overview of the UNI (Chen et al. 2024b), Phikon (Filiot et al. 2023), and Hibou-B (Nechaev, Pchelnikov, and Ivanova 2024) pathology foundation models in this work and their training frameworks.

UNI utilizes the DINOv2 (Oquab et al. 2024) framework, a self-supervised learning (SSL) method that integrates self-distillation with masked image modeling (MIM), to pretrain a ViT-L model on the large-scale Mass-100K dataset. **Phikon** employs the iBOT (Zhou et al. 2022) framework, a self-supervised learning (SSL) paradigm that combines self-distillation with masked image modeling (MIM) to pretrain a ViT-B model on a TCGA based pathology dataset. **Hibou-B** leverages the DINOv2 (Oquab et al. 2024) self-supervised learning (SSL) framework to pretrain a ViT-based model on a proprietary dataset.

Standard CP Evaluation Metrics

In addition to the proposed clinically aligned metrics, we also employ standard CP evaluation metrics in this work. For a CP workflow, let (x_i, y_i) denote a test sample and $\mathcal{C}(x_i)$ its corresponding prediction set; the standard evaluation metrics are defined as follows:

Empirical Coverage: It represents the assurance provided by a CP method that the true label will be included in the prediction set with a probability of 1-α, where α is the error rate. For example, with 95% coverage (α=0.05), the method ensures that the true label is included in the prediction set for at least 95% of test data points.. It is defined in eq 5:

$$\text{Empirical Coverage} = \frac{1}{n}\sum_{i=1}^{n}\delta[y_i \in C(x_i)] \tag{5}$$

where n is the total number of test samples and δ denotes an indicator function that is 1 when its argument is true and 0 otherwise.

Average Set Size: This metric represents the average number of labels in the prediction sets, i.e., the average cardinality of the conformal prediction sets. It is defined in eq 6:

$$\text{Average Set Size} = \frac{1}{n}\sum_{i=1}^{n}|\mathcal{C}(x_i)| \tag{6}$$

where $|\mathcal{C}(x_i)|$ denotes the cardinality of the prediction set for the i^{th} sample, and n is the total number of test samples.

Implementation Details

We assess the integration of CP within the few-shot adaptation of pathology foundation models across two diagnostic

Model	CP Method	Metric	Baseline						Baseline++						ProtoNets					
			Cov:90			Cov:95			Cov:90			Cov:95			Cov:90			Cov:95		
			K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10
UNI	LAC	Accuracy	0.18±0.02	0.30±0.03	0.37±0.03	0.16±0.01	0.24±0.03	0.30±0.03	0.18±0.03	0.26±0.05	0.29±0.05	0.17±0.02	0.21±0.04	0.26±0.05	0.14±0.01	0.16±0.02	0.18±0.03	0.14±0.01	0.15±0.02	0.17±0.03
		Coverage	0.90±0.01	0.90±0.01	0.90±0.00	0.95±0.00	0.95±0.00	0.95±0.00	0.90±0.02	0.90±0.02	0.89±0.02	0.95±0.02	0.95±0.02	0.94±0.01	0.90±0.02	0.91±0.02	0.92±0.02	0.95±0.01	0.96±0.01	0.97±0.01
		Avg. Set Size	5.30±0.40	3.80±0.20	3.20±0.20	6.00±0.30	4.80±0.20	4.20±0.20	5.40±0.40	4.40±0.40	3.70±0.40	6.00±0.40	5.10±0.50	4.50±0.40	9.0±0.0	8.6±0.4	8.2±0.5	9.0±0.0	8.8±0.4	8.5±0.5
	APS	Accuracy	0.15±0.01	0.21±0.02	0.24±0.02	0.14±0.00	0.17±0.01	0.20±0.02	0.15±0.01	0.20±0.04	0.22±0.04	0.14±0.00	0.17±0.02	0.19±0.04	0.15±0.01	0.16±0.02	0.18±0.03	0.15±0.01	0.16±0.02	0.18±0.03
		Coverage	0.98±0.01	0.97±0.01	0.98±0.00	1.00±0.00	0.99±0.00	0.99±0.00	0.97±0.02	0.96±0.01	0.96±0.01	0.99±0.00	0.98±0.01	0.98±0.01	0.93±0.01	0.94±0.02	0.95±0.02	0.96±0.01	0.97±0.01	0.98±0.01
		Avg. Set Size	6.40±0.30	5.30±0.20	5.00±0.10	7.00±0.10	6.10±0.20	5.80±0.10	6.40±0.40	5.30±0.30	5.00±0.40	6.90±0.10	6.20±0.30	5.80±0.50	9.0±0.0	8.7±0.4	8.4±0.5	9.0±0.0	8.8±0.4	8.5±0.5
	RAPS	Accuracy	0.17±0.02	0.25±0.03	0.32±0.03	0.15±0.01	0.21±0.02	0.26±0.03	0.16±0.02	0.23±0.05	0.25±0.03	0.15±0.01	0.19±0.04	0.22±0.04	0.14±0.01	0.15±0.02	0.17±0.03	0.14±0.01	0.15±0.02	0.17±0.03
		Coverage	0.93±0.01	0.93±0.01	0.93±0.01	0.99±0.01	0.97±0.00	0.96±0.00	0.93±0.02	0.93±0.02	0.93±0.01	0.97±0.01	0.96±0.01	0.96±0.01	0.91±0.02	0.92±0.02	0.93±0.02	0.95±0.01	0.96±0.01	0.97±0.01
		Avg. Set Size	5.90±0.40	4.40±0.20	3.70±0.20	6.60±0.30	5.30±0.20	4.70±0.20	5.80±0.40	4.70±0.40	4.40±0.30	6.50±0.30	5.60±0.40	5.10±0.40	8.3±0.3	8.0±0.5	7.7±0.6	8.5±0.2	8.3±0.4	8.0±0.5
Phikon	LAC	Accuracy	0.20±0.03	0.28±0.03	0.34±0.03	0.18±0.02	0.24±0.03	0.28±0.03	0.20±0.03	0.25±0.04	0.30±0.04	0.18±0.02	0.23±0.03	0.26±0.04	0.07±0.02	0.09±0.02	0.11±0.03	0.07±0.02	0.08±0.02	0.09±0.03
		Coverage	0.897±0.005	0.898±0.005	0.897±0.005	0.947±0.004	0.947±0.004	0.947±0.004	0.893±0.024	0.892±0.019	0.892±0.018	0.942±0.017	0.942±0.014	0.939±0.015	0.90±0.01	0.91±0.02	0.92±0.02	0.95±0.01	0.96±0.01	0.97±0.01
		Avg. Set Size	5.20±0.36	3.96±0.23	3.35±0.16	5.89±0.27	4.84±0.24	4.25±0.19	5.24±0.51	4.17±0.39	3.60±0.30	5.89±0.41	4.93±0.34	4.45±0.33	9.0±0.0	7.5±0.4	6.8±0.6	9.0±0.0	8.3±0.3	7.9±0.5
	APS	Accuracy	0.17±0.02	0.21±0.02	0.24±0.02	0.15±0.01	0.18±0.02	0.20±0.02	0.17±0.02	0.20±0.03	0.23±0.03	0.15±0.01	0.18±0.03	0.21±0.03	0.08±0.02	0.09±0.02	0.11±0.03	0.08±0.02	0.09±0.02	0.10±0.03
		Coverage	0.971±0.012	0.960±0.005	0.967±0.004	0.998±0.003	0.986±0.003	0.985±0.003	0.971±0.019	0.959±0.011	0.960±0.011	0.995±0.008	0.985±0.009	0.980±0.008	0.93±0.01	0.94±0.02	0.95±0.02	0.96±0.01	0.97±0.01	0.98±0.01
		Avg. Set Size	6.30±0.33	5.18±0.21	4.83±0.12	6.92±0.12	6.03±0.24	5.58±0.14	6.33±0.47	5.35±0.43	4.92±0.36	6.82±0.25	6.10±0.44	5.59±0.40	9.0±0.0	8.1±0.3	7.6±0.5	9.0±0.0	8.4±0.3	8.1±0.4
	RAPS	Accuracy	0.18±0.02	0.25±0.02	0.30±0.03	0.16±0.02	0.21±0.02	0.25±0.03	0.18±0.03	0.23±0.04	0.26±0.03	0.17±0.02	0.20±0.03	0.23±0.02	0.07±0.02	0.08±0.02	0.10±0.03	0.07±0.02	0.08±0.02	0.09±0.03
		Coverage	0.932±0.006	0.927±0.005	0.926±0.005	0.978±0.011	0.965±0.003	0.963±0.003	0.926±0.021	0.921±0.017	0.929±0.015	0.967±0.016	0.960±0.014	0.961±0.012	0.91±0.01	0.92±0.02	0.93±0.02	0.95±0.01	0.96±0.01	0.97±0.01
		Avg. Set Size	5.73±0.31	4.45±0.23	3.84±0.16	6.44±0.30	5.33±0.23	4.71±0.17	5.67±0.47	4.62±0.41	4.22±0.34	6.28±0.39	5.46±0.45	4.98±0.35	8.5±0.2	7.9±0.4	7.3±0.5	8.8±0.2	8.2±0.3	7.8±0.4
Hibou-B	LAC	Accuracy	0.20±0.03	0.30±0.03	0.37±0.03	0.18±0.02	0.25±0.03	0.30±0.03	0.18±0.02	0.25±0.05	0.27±0.05	0.17±0.02	0.22±0.04	0.25±0.05	0.26±0.04	0.31±0.05	0.34±0.06	0.26±0.04	0.30±0.05	0.33±0.06
		Coverage	0.903±0.005	0.902±0.005	0.902±0.004	0.952±0.005	0.952±0.003	0.951±0.004	0.895±0.02	0.899±0.02	0.904±0.02	0.943±0.018	0.949±0.015	0.950±0.015	0.90±0.01	0.91±0.02	0.92±0.02	0.95±0.01	0.96±0.01	0.97±0.01
		Avg. Set Size	5.23±0.34	3.89±0.26	3.30±0.19	5.94±0.27	4.83±0.29	4.22±0.20	5.41±0.28	4.37±0.44	3.96±0.47	6.01±0.27	5.14±0.43	4.67±0.46	9.0±0.0	8.5±0.4	8.0±0.5	9.0±0.0	8.6±0.3	8.2±0.4
	APS	Accuracy	0.16±0.02	0.22±0.02	0.25±0.02	0.15±0.01	0.18±0.02	0.21±0.02	0.16±0.02	0.21±0.04	0.22±0.04	0.14±0.00	0.18±0.03	0.20±0.03	0.27±0.04	0.31±0.05	0.34±0.06	0.27±0.04	0.31±0.05	0.34±0.06
		Coverage	0.972±0.008	0.967±0.005	0.974±0.005	0.999±0.003	0.989±0.002	0.989±0.003	0.972±0.018	0.964±0.012	0.965±0.010	0.999±0.003	0.991±0.007	0.988±0.007	0.93±0.01	0.94±0.02	0.95±0.02	0.96±0.01	0.97±0.01	0.98±0.01
		Avg. Set Size	6.30±0.30	5.23±0.19	4.94±0.12	6.93±0.13	6.06±0.25	5.70±0.14	6.46±0.29	5.50±0.41	5.02±0.37	6.95±0.09	6.29±0.42	5.89±0.35	9.0±0.0	8.5±0.4	8.0±0.5	9.0±0.0	8.6±0.3	8.2±0.4
	RAPS	Accuracy	0.18±0.02	0.27±0.03	0.33±0.03	0.16±0.01	0.22±0.03	0.26±0.02	0.17±0.02	0.23±0.05	0.25±0.04	0.16±0.02	0.20±0.04	0.23±0.04	0.26±0.04	0.29±0.05	0.32±0.06	0.26±0.04	0.29±0.05	0.32±0.04
		Coverage	0.933±0.005	0.929±0.005	0.929±0.004	0.978±0.007	0.967±0.003	0.965±0.003	0.926±0.019	0.932±0.014	0.934±0.014	0.971±0.015	0.969±0.012	0.968±0.012	0.91±0.01	0.92±0.02	0.93±0.02	0.95±0.01	0.96±0.01	0.97±0.01
		Avg. Set Size	5.70±0.27	4.36±0.24	3.75±0.18	6.44±0.27	5.25±0.27	4.65±0.19	5.84±0.21	4.87±0.41	4.36±0.41	6.45±0.25	5.65±0.38	5.15±0.47	8.5±0.3	8.1±0.4	7.8±0.5	8.5±0.2	8.2±0.3	7.9±0.4

Table 3: Comparison of Accuracy, Emprical Coverage and Average Set Size for Baseline, Baseline++, and ProtoNets methods with three CP methods and pathology foundation models across varying coverage and K-shot settings on the HMU-GC-HE-30K dataset.

tasks: cervical cancer cell classification using the HiCervix dataset and gastric cancer tissue classification using the HMU-GC-HE-30K dataset. We use three CP methods (LAC, APS, and RAPS) with Baseline, Baseline++, and Prototypical Networks methods for few-shot adaptation of UNI, Phikon, and Hibou-B foundation models. Each experiment is conducted over 100 independent trials, with datasets partitioned into three disjoint pools: a training pool (70%) for constructing support sets, a calibration pool (10%) for calibration set generation, and a test pool (20%) for query set formation. The experiments are performed for 1-shot (K=1), 5-shot (K=5), and 10-shot (K=10) configurations. For each trial, K samples per class are drawn from the training pool to form the support set, while additional samples from the calibration and test pools form the calibration and query sets, respectively. The classifiers are optimized using the Adam optimizer with a learning rate of 0.01, and standard data augmentation techniques like color jittering, random horizontal flipping, and random resized cropping in Baseline and Baseline++. The experimental results are evaluated using both standard CP and clinically aligned metrics, with performance reported as the average mean ± standard deviation across all trials.

Results and Discussion

In Tables 1 and 2, we report the standard CP metrics and our proposed clinical aligned metrics for Baseline, Baseline++, and ProtoNets, employing three split-CP frameworks and various pathology foundation models across multiple coverage targets and few-shot settings on the HiCervix dataset. Our results demonstrate that the LAC approach consistently yielded the best-calibrated and most clinically efficient prediction sets, followed by APS as the second best performing CP method.Across all experimental settings, LAC achieved up to 35% lower MRTL and 50% lower RTC, reflecting a substantial reduction in diagnostic turnaround time and facilitating more efficient clinical workflows. Among the few-shot adaptation strategies, ProtoNets demonstrated the most consistent true label ranking and reliable calibration across all configurations. It achieved upto 64% higher R1CR while maintaining comparable ESR levels, , supporting faster clinical decision-making and improved patient outcomes. Overall, the best-performing configurations for this task resulted from the integration of the LAC and ProtoNets methods with the Phikon foundation model.

In Table 3 and 4, we report both the standard CP evaluation and our proposed clinical aligned metrics for LAC, APS and RAPS methods with three few shot adaptation methods and pathology foundation models across varying coverage and few-shot settings on the HMU-GC-HE-30K dataset. Our findings show that the LAC method consistently produced the most well-calibrated and clinically efficient prediction sets across all few-shot configurations, with APS ranking as the next best-performing approach. Across both coverage levels and shot configurations, LAC achieved approximately 15–30% lower MRTL and 25–45% lower RTC compared to APS, indicating a substanial reduction in time-to-correct diagnosus and resulting in efficient clinical workflows. Among the few-shot adaptation methods, Baseline demonstrated the most robust true label ranking consistency and calibration across all K-shot settings. It consistently achieved up to 25% higher R1CR and as much as 250% higher ESR than both Baseline++ and ProtoNets, highlighting its effectiveness in reducing clinical workload and supporting improved patient outcomes. Notably, the most effective configurations for this task were obtained by integrating the LAC and Baseline methods with the Hibou-B foundation model.

Across both the datasets LAC emerged as the best performing CP method with all the evaluation metrics following the expected trends across all configurations, validating the effectiveness of CP frameworks in improving reliability for few-shot pathology classification tasks.

			Baseline						Baseline++						ProtoNets					
			Cov:90			Cov:95			Cov:90			Cov:95			Cov:90			Cov:95		
Model	**CP Method**	**Metric**	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10	K=1	K=5	K=10
UNI	LAC	MRTL	3.40±0.20	2.70±0.20	2.30±0.10	3.70±0.10	3.10±0.10	2.80±0.10	3.50±0.20	2.90±0.20	2.60±0.20	3.70±0.20	3.20±0.20	2.90±0.20	5.1±0.1	4.7±0.3	4.3±0.2	5.0±0.1	4.7±0.3	4.4±0.2
		RTC	2.40±0.20	1.70±0.20	1.30±0.10	2.70±0.10	2.10±0.10	1.80±0.10	2.50±0.20	1.90±0.20	1.60±0.20	2.70±0.20	2.20±0.20	1.90±0.20	4.1±0.1	3.8±0.3	3.5±0.3	4.0±0.1	3.8±0.3	3.5±0.3
		R1CR	0.18±0.02	0.30±0.03	0.37±0.03	0.16±0.01	0.24±0.03	0.30±0.03	0.18±0.03	0.26±0.05	0.29±0.05	0.17±0.02	0.21±0.04	0.26±0.05	0.14±0.01	0.16±0.02	0.18±0.03	0.14±0.01	0.15±0.02	0.17±0.03
		ESR	0.01±0.01	0.08±0.02	0.14±0.02	0.01±0.01	0.05±0.02	0.09±0.02	0.02±0.01	0.03±0.02	0.04±0.03	0.01±0.01	0.01±0.01	0.02±0.02	0.00±0.00	0.02±0.05	0.04±0.08	0.00±0.00	0.01±0.04	0.02±0.05
	APS	MRTL	3.80±0.10	3.30±0.10	3.10±0.10	4.00±0.00	3.60±0.10	3.50±0.10	3.80±0.10	3.30±0.20	3.10±0.30	4.00±0.10	3.70±0.10	3.50±0.30	5.0±0.1	4.8±0.3	4.5±0.3	5.0±0.1	4.7±0.3	4.4±0.2
		RTC	2.80±0.10	2.30±0.10	2.10±0.10	3.00±0.00	2.60±0.10	2.50±0.10	2.80±0.10	2.30±0.20	2.10±0.30	3.00±0.10	2.70±0.10	2.50±0.30	4.0±0.1	3.8±0.3	3.6±0.3	4.0±0.1	3.8±0.3	3.5±0.3
		R1CR	0.15±0.01	0.21±0.02	0.24±0.02	0.14±0.00	0.17±0.01	0.20±0.02	0.15±0.01	0.20±0.04	0.22±0.04	0.14±0.00	0.17±0.02	0.19±0.04	0.15±0.01	0.16±0.02	0.18±0.03	0.15±0.01	0.16±0.02	0.18±0.03
		ESR	0.00±0.00	0.02±0.01	0.04±0.01	0.00±0.00	0.01±0.00	0.02±0.01	0.00±0.00	0.00±0.00	0.00±0.01	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.01±0.04	0.03±0.06	0.00±0.00	0.01±0.03	0.02±0.05
	RAPS	MRTL	3.70±0.20	2.90±0.10	2.60±0.10	3.90±0.10	3.30±0.10	3.00±0.10	3.70±0.20	3.10±0.30	2.90±0.20	3.80±0.10	3.40±0.20	3.20±0.30	5.1±0.1	4.8±0.3	4.5±0.2	5.0±0.1	4.7±0.3	4.4±0.2
		RTC	2.70±0.20	1.90±0.10	1.60±0.10	2.90±0.10	2.30±0.10	2.00±0.10	2.70±0.20	2.10±0.30	1.90±0.20	2.80±0.10	2.40±0.20	2.20±0.30	4.1±0.1	3.8±0.3	3.5±0.3	4.0±0.1	3.8±0.3	3.5±0.3
		R1CR	0.17±0.02	0.25±0.03	0.32±0.03	0.15±0.01	0.21±0.02	0.26±0.03	0.16±0.02	0.23±0.05	0.25±0.03	0.15±0.01	0.19±0.04	0.22±0.04	0.14±0.01	0.15±0.02	0.17±0.03	0.14±0.01	0.15±0.02	0.17±0.03
		ESR	0.00±0.00	0.05±0.02	0.10±0.02	0.00±0.00	0.02±0.01	0.05±0.01	0.00±0.00	0.01±0.01	0.01±0.01	0.00±0.00	0.00±0.00	0.00±0.01	0.00±0.00	0.01±0.04	0.02±0.06	0.00±0.00	0.01±0.03	0.02±0.05
Phikon	LAC	MRTL	3.38±0.20	2.74±0.14	2.42±0.10	3.60±0.15	3.10±0.14	2.81±0.11	3.40±0.23	2.88±0.23	2.56±0.18	3.61±0.17	3.15±0.21	2.92±0.22	5.7±0.1	4.8±0.2	4.3±0.3	5.6±0.1	4.5±0.2	4.3±0.3
		RTC	2.27±0.20	1.64±0.14	1.32±0.10	2.55±0.15	2.04±0.14	1.76±0.11	2.29±0.24	1.77±0.24	1.45±0.18	2.55±0.18	2.09±0.22	1.86±0.22	4.7±0.1	3.8±0.2	3.3±0.3	4.6±0.1	3.5±0.2	3.2±0.3
		R1CR	0.20±0.03	0.28±0.03	0.34±0.03	0.18±0.02	0.24±0.03	0.28±0.03	0.20±0.03	0.25±0.04	0.30±0.04	0.18±0.02	0.23±0.03	0.26±0.04	0.07±0.02	0.09±0.02	0.11±0.03	0.07±0.02	0.08±0.02	0.09±0.03
		ESR	0.01±0.01	0.05±0.02	0.10±0.02	0.00±0.00	0.02±0.01	0.05±0.01	0.02±0.01	0.03±0.02	0.05±0.02	0.01±0.01	0.01±0.01	0.02±0.02	0.00±0.00	0.02±0.04	0.03±0.06	0.00±0.00	0.01±0.03	0.02±0.04
	APS	MRTL	3.75±0.13	3.24±0.12	3.07±0.09	3.97±0.05	3.59±0.12	3.40±0.08	3.76±0.18	3.34±0.22	3.13±0.21	3.93±0.10	3.64±0.20	3.39±0.20	5.5±0.1	4.5±0.2	4.2±0.3	5.5±0.1	4.4±0.2	4.2±0.3
		RTC	2.72±0.14	2.20±0.12	2.04±0.09	2.97±0.05	2.58±0.12	2.38±0.08	2.73±0.20	2.30±0.23	2.09±0.21	2.93±0.11	2.63±0.20	2.37±0.21	4.5±0.1	3.6±0.2	3.2±0.3	4.5±0.1	3.5±0.2	3.2±0.3
		R1CR	0.17±0.02	0.21±0.02	0.24±0.02	0.15±0.01	0.18±0.02	0.20±0.02	0.17±0.02	0.20±0.03	0.23±0.03	0.15±0.01	0.18±0.03	0.21±0.03	0.08±0.02	0.09±0.02	0.11±0.03	0.08±0.02	0.09±0.02	0.10±0.03
		ESR	0.00±0.00	0.01±0.01	0.02±0.01	0.00±0.00	0.00±0.00	0.01±0.00	0.00±0.00	0.00±0.01	0.01±0.01	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.01±0.03	0.02±0.05	0.00±0.00	0.01±0.03	0.01±0.04
	RAPS	MRTL	3.59±0.16	2.95±0.14	2.64±0.10	3.79±0.12	3.31±0.13	3.02±0.10	3.57±0.20	3.05±0.22	2.82±0.20	3.76±0.16	3.37±0.22	3.14±0.18	5.6±0.1	5.0±0.2	4.7±0.3	5.6±0.1	4.9±0.2	4.5±0.3
		RTC	2.52±0.17	1.88±0.14	1.57±0.10	2.77±0.13	2.27±0.13	1.98±0.10	2.50±0.22	1.97±0.23	1.75±0.21	2.72±0.17	2.33±0.23	2.10±0.19	4.6±0.1	4.0±0.2	3.5±0.3	4.6±0.1	3.9±0.2	3.5±0.3
		R1CR	0.18±0.02	0.25±0.02	0.30±0.03	0.16±0.02	0.21±0.02	0.25±0.03	0.18±0.03	0.23±0.04	0.26±0.03	0.17±0.02	0.20±0.03	0.23±0.02	0.07±0.02	0.08±0.02	0.10±0.03	0.07±0.02	0.08±0.02	0.09±0.03
		ESR	0.00±0.00	0.02±0.01	0.06±0.02	0.00±0.00	0.01±0.01	0.03±0.01	0.01±0.01	0.01±0.01	0.02±0.01	0.00±0.00	0.00±0.01	0.01±0.01	0.00±0.00	0.01±0.03	0.02±0.05	0.00±0.00	0.00±0.03	0.01±0.04
Hibou-B	LAC	MRTL	3.36±0.19	2.66±0.15	2.35±0.11	3.61±0.15	3.05±0.16	2.75±0.12	3.49±0.15	2.95±0.25	2.71±0.28	3.68±0.12	3.23±0.24	2.97±0.29	4.5±0.2	3.8±0.3	3.4±0.3	4.4±0.2	3.7±0.3	3.3±0.3
		RTC	2.27±0.19	1.56±0.14	1.25±0.11	2.57±0.15	2.00±0.16	1.70±0.11	2.39±0.16	1.85±0.26	1.61±0.29	2.62±0.13	2.18±0.24	1.92±0.29	3.5±0.2	2.8±0.3	2.4±0.3	3.4±0.2	2.8±0.3	2.4±0.3
		R1CR	0.20±0.03	0.30±0.03	0.37±0.03	0.18±0.02	0.25±0.03	0.30±0.03	0.18±0.02	0.25±0.05	0.27±0.05	0.17±0.02	0.22±0.04	0.25±0.05	0.26±0.04	0.31±0.05	0.34±0.06	0.26±0.04	0.30±0.05	0.33±0.06
		ESR	0.01±0.01	0.06±0.02	0.12±0.02	0.01±0.01	0.03±0.01	0.06±0.02	0.01±0.01	0.02±0.01	0.03±0.02	0.00±0.00	0.01±0.01	0.02±0.02	0.00±0.00	0.03±0.06	0.05±0.08	0.00±0.00	0.01±0.04	0.02±0.05
	APS	MRTL	3.74±0.14	3.22±0.12	3.08±0.08	3.97±0.06	3.58±0.13	3.42±0.08	3.82±0.12	3.33±0.26	3.12±0.22	3.98±0.04	3.65±0.24	3.47±0.24	4.4±0.2	3.7±0.3	3.3±0.3	4.4±0.2	3.7±0.3	3.3±0.3
		RTC	2.71±0.15	2.19±0.11	2.05±0.08	2.97±0.06	2.57±0.13	2.41±0.08	2.79±0.13	2.30±0.27	2.09±0.23	2.98±0.04	2.65±0.24	2.46±0.25	3.4±0.2	2.8±0.3	2.4±0.3	3.4±0.2	2.8±0.3	2.4±0.3
		R1CR	0.16±0.02	0.22±0.02	0.25±0.02	0.15±0.01	0.18±0.02	0.21±0.02	0.16±0.02	0.21±0.04	0.22±0.04	0.14±0.00	0.18±0.03	0.20±0.03	0.27±0.04	0.31±0.05	0.34±0.06	0.27±0.04	0.31±0.05	0.34±0.06
		ESR	0.00±0.00	0.01±0.01	0.03±0.01	0.00±0.00	0.00±0.00	0.01±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.02±0.05	0.04±0.07	0.00±0.00	0.01±0.03	0.02±0.05
	RAPS	MRTL	3.55±0.16	2.86±0.14	2.55±0.11	3.79±0.13	3.23±0.14	2.95±0.11	3.65±0.13	3.10±0.26	2.85±0.27	3.82±0.12	3.41±0.20	3.14±0.23	4.5±0.2	4.2±0.3	3.9±0.3	4.5±0.2	4.2±0.3	4.0±0.3
		RTC	2.48±0.16	1.79±0.14	1.48±0.10	2.77±0.14	2.20±0.14	1.91±0.11	2.58±0.13	2.03±0.27	1.79±0.28	2.79±0.12	2.38±0.20	2.11±0.23	3.5±0.2	3.2±0.3	2.9±0.3	3.5±0.2	3.2±0.3	3.0±0.3
		R1CR	0.18±0.02	0.27±0.03	0.33±0.03	0.16±0.01	0.22±0.03	0.26±0.02	0.17±0.02	0.23±0.05	0.25±0.04	0.16±0.02	0.20±0.04	0.23±0.04	0.26±0.04	0.29±0.05	0.32±0.06	0.26±0.04	0.29±0.05	0.32±0.04
		ESR	0.00±0.00	0.04±0.01	0.08±0.02	0.00±0.00	0.01±0.01	0.04±0.01	0.00±0.00	0.01±0.01	0.01±0.01	0.00±0.00	0.00±0.01	0.00±0.01	0.00±0.00	0.01±0.04	0.02±0.05	0.00±0.00	0.01±0.03	0.02±0.04

Table 4: Comparison of MRTL, RTC, R1CR, and ESR for Baseline, Baseline++, and ProtoNets methods with three CP methods and pathology foundation models across varying coverage and K-shot settings on the HMU-GC-HE-30K dataset.

Conclusion

In this study, we systematically integrated Conformal Prediction methods with few-shot adaptation of pathology foundation models to enhance diagnostic reliability and operational robustness in clinical classification tasks . We evaluated these approaches across 2 datasets, using both standard CP metrics and our proposed clinically aligned metrics designed to assess their clinical applicability in few-shot settings. Our results indicate that the LAC method achieved the most reliable performance through its well-calibrated and compact prediction sets across both datasets. These findings highlight that integrating CP methods with few-shot adaptation of pathology foundation models in clinical workflows can substantially enhance diagnostic reliability and lead to improved patient outcomes.

Acknowledgments

This work was supported by the Department of Science and Technology (DST), Government of India, under the project SP-YO-2021-1679 (C) and (G).

References

Angelopoulos, A.; Bates, S.; Malik, J.; and Jordan, M. I. 2020. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Angelopoulos, A. N.; Pomerantz, S. R.; Do, S.; Bates, S.; Bridge, C. P.; Elton, D. C.; Lev, M. H.; Gonzalez, R. G.; Jordan, M. I.; and Malik, J. 2024. Conformal Triage for Medical Imaging AI Deployment. *medRxiv*, 2024–02.

Cai, D.; Chen, J.; Zhao, J.; Xue, Y.; Yang, S.; Yuan, W.; Feng, M.; Weng, H.; Liu, S.; Peng, Y.; Zhu, J.; Wang, K.; Jackson, C.; Tang, H.; Huang, J.; and Wang, X. 2024. HiCervix: An Extensive Hierarchical Dataset and Benchmark for Cervical Cytology Classification. *IEEE Transactions on Medical Imaging*, 43(12): 4344–4355.

Chen, R. J.; Ding, T.; Lu, M. Y.; Williamson, D. F.; Jaume, G.; Chen, B.; Zhang, A.; Shao, D.; Song, A. H.; Shaban, M.; et al. 2024a. Towards a General-Purpose Foundation Model for Computational Pathology. *Nature Medicine*.

Chen, R. J.; Ding, T.; Lu, M. Y.; Williamson, D. F. K.; Jaume, G.; Song, A. H.; Chen, B.; Zhang, A.; Shao, D.; Shaban, M.; Williams, M.; Oldenburg, L.; Weishaupt, L. L.; Wang, J. J.; Vaidya, A.; Le, L. P.; Gerber, G. K.; Sahai, S.; Williams, W.; and Mahmood, F. 2024b. Towards a general-purpose foundation model for computational pathology. *Nature medicine*.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C.; and Huang, J.-B. 2019. A Closer Look at Few-shot Classification. In *International Conference on Learning Representations*.

Chen, Z.; Cano, A. H.; Romanou, A.; Bonnet, A.; Matoba, K.; Salvi, F.; Pagliardini, M.; Fan, S.; Kopf, A.; Mohtashami, A.; Sallinen, A.; Sakhaeirad, A.; Swamy, V.; Krawczuk, I.; Bayazit, D.; Marmet, A.; Montariol, S.; Hartley, M.-A.; Jaggi, M.; and Bosselut, A. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. *ArXiv*, abs/2311.16079.

Filiot, A.; Ghermi, R.; Olivier, A.; Jacob, P.; Fidon, L.; Kain, A. M.; Saillard, C.; and Schiratti, J.-B. 2023. Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling. In *medRxiv*.

Fisch, A.; Schuster, T.; Jaakkola, T.; and Barzilay, R. 2021. Few-shot Conformal Prediction with Auxiliary Tasks. In *In-*

ternational Conference on Machine Learning, volume 139, 3246–3256. PMLR.

Kumar, B.; Palepu, A.; Tuwani, R.; and Beam, A. 2022. Towards reliable zero shot classification in self-supervised models with conformal prediction. *arXiv preprint arXiv:2210.15805*.

Lee, J.; Lim, J.; Byeon, K.; and Kwak, J. T. 2025. Benchmarking pathology foundation models: Adaptation strategies and scenarios. *Computers in biology and medicine*, 190: 110031.

Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R. J.; and Wasserman, L. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111.

Lou, S.; Ji, J.; Li, H.; Zhang, X.; Jiang, Y.; Hua, M.; Chen, K.; Ge, K.; Zhang, Q.; Wang, L.; Han, P.; and Cao, L. 2025. A large histological images dataset of gastric cancer with tumour microenvironment annotation for AI. *Scientific Data*, 12.

Lu, C.; Angelopoulos, A. N.; and Pomerantz, S. 2022. Improving trustworthiness of AI disease severity rating in medical imaging with ordinal conformal prediction sets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 545–554. Springer.

Lu, C.; Lemay, A.; Chang, K.; Höbel, K.; and Kalpathy-Cramer, J. 2022. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12008–12016.

Nechaev, D.; Pchelnikov, A.; and Ivanova, E. 2024. Hibou: A Family of Foundational Vision Transformers for Pathology. arXiv:2406.05074.

Neidlinger, P.; Nahhas, O. S. M. E.; Muti, H. S.; Lenz, T.; Hoffmeister, M.; Brenner, H.; van Treeck, M.; Langer, R.; Dislich, B.; Behrens, H.-M.; Röcken, C.; Foersch, S.; Truhn, D.; Marra, A.; Saldanha, O. L.; and Kather, J. N. 2025. Benchmarking foundation models as feature extractors for weakly supervised computational pathology. *Nature biomedical engineering*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; HAZIZA, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*. Featured Certification.

Papadopoulos, H.; Proedrou, K.; Vovk, V.; and Gammerman, A. 2002. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, 345–356. Springer.

Romano, Y.; Patterson, E.; and Candes, E. 2019. Conformalized quantile regression. *Advances in neural information processing systems*, 32.

Romano, Y.; Sesia, M.; and Candes, E. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33: 3581–3591.

Sadinle, M.; Lei, J.; and Wasserman, L. 2019. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234.

Shakeri, F.; Huang, Y.; Silva-Rodriguez, J.; Bahig, H.; Tang, A.; Dolz, J.; and Ben Ayed, I. 2024. Few-shot Adaptation of Medical Vision-Language Models . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15012. Springer Nature Switzerland.

Silva-Rodríguez, J.; Ben Ayed, I.; and Dolz, J. 2025. Conformal Prediction for Zero-Shot Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Silva-Rodr'iguez, J.; Fillioux, L.; Cournède, P.-H.; Vakalopoulou, M.; Christodoulidis, S.; Ayed, I. B.; and Dolz, J. 2025. Full Conformal Adaptation of Medical Vision-Language Models. In *Information Processing in Medical Imaging*.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical Networks for Few-shot Learning. In *Neural Information Processing Systems*.

Vinyals, O.; Blundell, C.; Lillicrap, T. P.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching Networks for One Shot Learning. In *Neural Information Processing Systems*.

Vishwakarma, H.; Mishler, A.; Cook, T.; Dalmasso, N.; Raman, N.; and Ganesh, S. 2025. Prune 'n Predict: Optimizing LLM Decision-making with Conformal Prediction. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.

Vorontsov, E.; Bozkurt, A.; Casson, A.; Shaikovski, G.; Zelechowski, M.; Severson, K.; Zimmermann, E.; Hall, J.; Tenenholtz, N.; Fusi, N.; Yang, E.; Mathieu, P.; Eck, A.; Lee, D.; Viret, J.; Robert, E.; Wang, Y.; Kunz, J.; Lee, M.; and Fuchs, T. 2024. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30: 2924–2935.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Xie, Q.; Chen, Q.; Chen, A.; Peng, C.; Hu, Y.; Lin, F.; Peng, X.; Huang, J.; Zhang, J.; Keloth, V. K.; Zhou, X.; He, H.; Ohno-Machado, L.; Wu, Y.; Xu, H.; and Bian, J. 2024. Me-LLaMA: Foundation Large Language Models for Medical Applications. *Research Square*.

Xu, H.; Usuyama, N.; Bagga, J.; Zhang, S.; Rao, R.; Naumann, T.; Wong, C.; Gero, Z.; González, J.; Gu, Y.; Xu, Y.; Wei, M.; Wang, W.; Ma, S.; Wei, F.; Yang, J.; Li, C.; Gao, J.; Rosemon, J.; and Poon, H. 2024a. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630: 1–8.

Xu, Z.; Shi, Z.; Wei, J.; Mu, F.; Li, Y.; and Liang, Y. 2024b. Towards Few-Shot Adaptation of Foundation Models via Multitask Finetuning. In *ICLR*.

Zhan, X.; Wang, Z.; Yang, M.; Luo, Z.; Wang, Y.; and Li, G. 2020. An electronic nose-based assistive diagnostic prototype for lung cancer detection with conformal prediction. *Measurement*, 158: 107588.

Zhang, Y.; Wang, S.; Zhang, Y.; and Chen, D. Z. 2023. Rr-cp: Reliable-region-based conformal prediction for trustworthy medical image classification. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, 12–21. Springer.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2022. Image BERT Pre-training with Online Tokenizer. In *International Conference on Learning Representations*.