# Optimizing Conformal Prediction Sets for Pathological Image Classification

**Shubham Ojha***
Division of Biomedical Informatics
Cincinnati Children's Hospital Medical Center
Cincinnati OH 45229
shubham.ojha1000@gmail.com

**Aditya Narendra***
Division of Biomedical Informatics
Cincinnati Children's Hospital Medical Center
Cincinnati OH 45229
adinarendra0108@gmail.com

**Abhay Kshirsagar**
Dept. of Chemical and Biomolecular Engineering
University of Illinois Urbana-Champaign
Champaign IL
abhaysk2@illinois.edu

**Shyam Sundar Debsarkar**
Dept. of Pediatrics, College of Medicine
University of Cincinnati
Cincinnati OH 45257
debsarss@mail.uc.edu

**Surya Prasath**
Division of Biomedical Informatics
Cincinnati Children's Hospital Medical Center
Cincinnati OH 45229
surya.prasath@cchmc.org

## Abstract

The intersection of Deep Learning (DL) and pathology has gained significant attention, encompassing cell classification, detection, segmentation, and whole-slide image (WSI) analysis. Further works at this intersection have increasingly focused on integrating uncertainty quantification (UQ) with DL methods for pathology to address their occasional unreliability in clinical settings. Conformal Prediction (CP) is one of the UQ methods deployed for various settings, including pathology. CP methods are computationally efficient and generate prediction sets to include the true label with a user-defined coverage guarantee. However, CP methods lack inherent control over the compositionality of prediction sets, which restricts their clinical utility. This study presents a novel training method using Hinge loss for the underlying models used in CP methods. This approach aims to provide effective control over the compositionality of prediction sets, aligning more closely with the specific needs of pathologists. We evaluate the effectiveness of this training approach using three application specific metrics tailored to enhance the integration of CP methods into clinical pathology workflows. Our results show that the Hinge loss based training approach outperforms the traditional Cross Entropy training method across all evaluation metrics, leading to effective management of the compositionality of prediction sets.

## 1 Introduction

The integration of computational algorithms for analyzing microscopic images of human tissue or cells, known as computational pathology, has emerged as a critical area of research with significant

*Equal Contribution

Preprint. Under review.

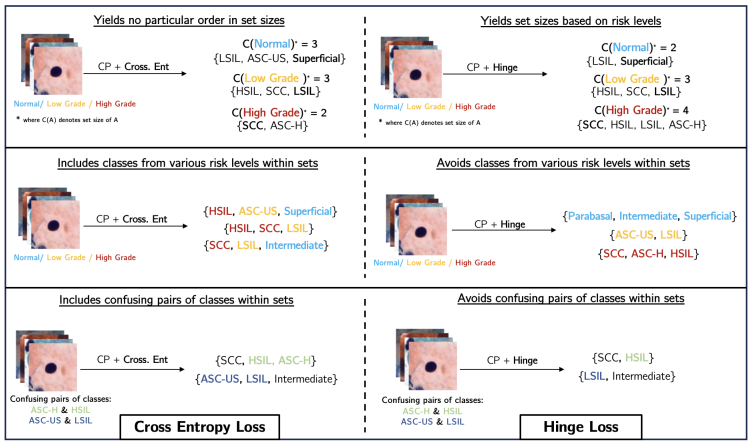


Figure 1: Comparison of Cross-Entropy and Hinge Loss functions used with a CP method in a pathological cell classification workflow, highlighting their impact on the composition of prediction sets. As shown on the left, Cross-Entropy loss results in unordered set sizes with mixed risk and confusing class pairs. In contrast, Hinge loss produces risk-aligned ordered set sizes with minimal mixed risk and confusing class pairs.

clinical implications [10, 27]. This field leverages advanced Machine Learning (ML) techniques to analyze and interpret tissue samples, leading to enhanced efficiency in cancer diagnosis and treatment. [45, 10, 21, 36].

Extensive research in computational pathology, spanning both histopathology and cytology [17, 3, 51], has focused on morphology-based cell classification [37, 31, 26, 23], cell detection tasks [19, 58, 60], cellular segmentation [46, 50, 56, 62], and enhancing tissue-level assessments using whole-slide images (WSI) [6, 8, 30, 43]. Collectively, these advancements have significantly improved diagnostic accuracy and contributed to more effective patient care. However, translating these developments into practical medical settings remains challenging. Despite these advancements significantly improving diagnostic accuracy and patient care, integrating them into clinical practice remains a challenge.

A major limitation in the clinical deployment of these approaches is the lack of integration of uncertainty estimates in Deep Learning (DL) models, as relying solely on maximum-likelihood estimates and their probabilities may not ensure optimal diagnostic reliability. For example, if a DL model provides a prediction with high uncertainty or low confidence, it could abstain from decision-making, thus enabling the pathologist to intervene in such instances.

Some advancements have been made in integrating uncertainty quantification (UQ) into DL workflows for computational pathology [12, 35, 34], utilizing probabilistic techniques such as Monte Carlo (MC) dropout, ensemble methods, and test-time augmentation (TTA), which are inherently computationally intensive and have limited real-world deployment.

An alternative approach to uncertainty quantification (UQ) is Conformal Prediction (CP), which delivers actionable uncertainty estimates by generating prediction sets that are likely to contain the true outcome with a user-specified probability. This computationally efficient method is particularly helpful in medical applications, where it is crucial not only to identify the most probable diagnosis but also to confirm or exclude potentially harmful diagnoses. For example, even if the most likely diagnosis is a common throat infection, CP methods help rule out serious conditions like Tuberculosis (TB), COPD, or Lung cancer by including all possible outcomes in the prediction sets.

Extensive works in CP have been concentrated on improving prediction set size and coverage guarantees, which ensure that the true outcome is included within the predicted set with a specified probability. However, these metrics are insufficient for integrating CP methods within computational pathology workflows. For clinical deployment, it is vital to effectively control the compositionality of conformal prediction sets to ensure that the included classes align with the needs of pathologists.

The compositionality of the prediction sets is primarily dependent on the probability distribution across different classes, which is influenced by the training method used for the underlying DL

model. Therefore, it is essential to provide a training approach that enables some control over the compositionality of the prediction sets and develop metrics for their evaluation tailored for clinical pathology workflows. Working upon these, the main contributions of this work are:

1. We introduce a novel training approach using hinge loss for underlying models in CP methods, offering effective control over the compositionality of the prediction sets and enhancing the integration of CP methods into pathological workflows.

2. We provide three application specific metrics dealing with the compositionality of conformal prediction sets, specifically designed to align CP methods with the needs of pathologists.

## 2 Related Works

### 2.1 Conformal Prediction (CP) methods

Conformal prediction (CP), is built on the foundational work by Vovk et al. [55], which has been explored in both regression [39] and classification settings [42, 1]. Most CP methods use a split approach [25] with a held-out calibration set, though some alternatives employ cross-validation [54]. Additional works have been focused on achieving and improving conditional coverage [53, 11, 40], obtaining smaller confidence sets [4, 48, 14], addressing issues related to out-of-distribution settings [33, 5], & domain adaptation [15].

### 2.2 CP methods for Medical Settings

CP methods have been utilized across various healthcare domains, including MRI [28], CT scans [2, 61], X-rays [22], and other clinical areas [29, 63].

## 3 Conformal Prediction & Procedures

Conformal prediction is a statistical framework that generates prediction sets containing ground truth labels with a desired coverage guarantee. It is distribution-free, rigorous in statistics, easy to integrate with any ML model, and agnostic to the underlying prediction problem. 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

A trained model is used to quantify its uncertainty based on a holdout set known as the calibration set, which is not used during the model's training phase. 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 real-valued score function is defined for each input-output pair $(x, y)$ in the calibration set, for measuring its similarity to training examples. 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 a quantile of the sorted conformity scores from the calibration set. Specifically, $\hat{q}$ is the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of these scores, where $n$ is the number of examples in the calibration set, $\alpha$ is the error rate (e.g., $\alpha = 0.05$ for a $95\%$ confidence level), and $\lceil \cdot \rceil$ denotes the ceiling function. $\hat{q}$ serves as a threshold to ensure that the true class is included in the prediction set with a $1 - \alpha$ confidence level during testing.

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 performance of any CP method is generally evaluated using the following metrics:

1. Set Size: It corresponds to $|\textit{Prediction Set}|$, which denotes the cardinality of the prediction set or the number of elements it contains.

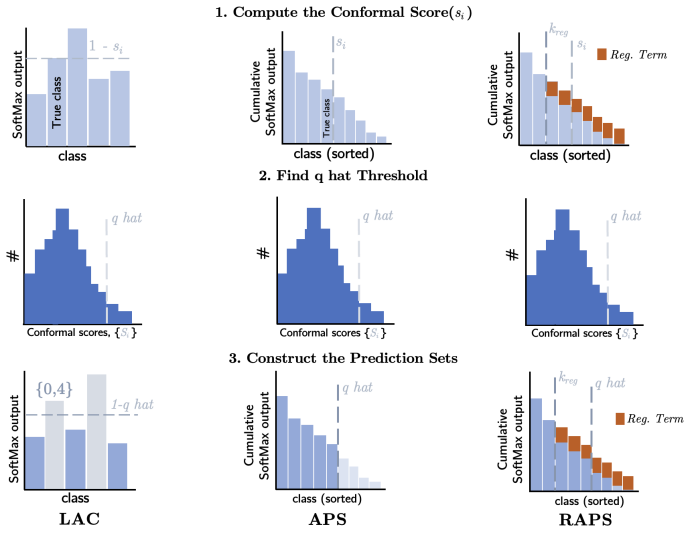


Figure 2: Overview of the working of LAC, APS, and RAPS conformal prediction methods

2. 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-$\alpha$, where $\alpha$ is the error rate. For example, with 95% coverage ($\alpha$=0.05), the method ensures that the true label is included in the prediction set for at least 95% of test data points.

### 3.1 LAC

LAC [42] 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 $1 - \hat{q}$.

### 3.2 Adaptive Prediction Sets (APS)

APS [40] 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 $\hat{q}$.

### 3.3 Regularized Adaptive Prediction Sets (RAPS)

RAPS [1] 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 $\lambda$ is added for classes beyond a randomized threshold ($k_{\text{reg}}$). 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 $\hat{q}$.

# 4 Loss Functions for Classification

In this section, we explain in detail two commonly used loss functions for multi-class classification: Cross Entropy loss and Hinge loss.

## 4.1 Cross Entropy (CE) Loss

Cross Entropy (CE) is a loss function that measures a classification model's performance by quantifying the difference between true and predicted class distributions, as shown in eq 1:

$$\mathcal{L} = -\sum_{i=1}^{K} y_i \log p_i \tag{1}$$

where $K$ is the number of total classes, $y_i$ is the true probability for class $i$ & $p_i$ is the predicted probability for class $i$.

CE loss operates on the softmax scores of different classes. It directly influences the probability distribution by concentrating a high probability mass on the true class while reducing it for the incorrect classes, without any regard to the relative ordering of the probabilities assigned to the incorrect classes.

Considering a three-class classification problem with Class A as the true class, the model initially predicts probabilities as [0.33, 0.32, 0.35] for classes A, B, and C respectively. As training progresses with cross-entropy loss, the probabilities might evolve as follows:

After some training: [0.5, 0.3, 0.2]
After further training: [0.7, 0.2, 0.1]

Cross-entropy loss focuses on directing a higher probability mass on the true class (A) and lowering it for the incorrect classes (B and C), but it does not regulate the relative probability ordering between incorrect classes.

## 4.2 Hinge Loss

Hinge loss is a loss function that guarantees the correct class is predicted with a specified margin of confidence. It achieves this by summing the excess margins of the correct class over the incorrect classes, adjusted by a margin parameter. For a true class label $y$ and score vector $f$, it is computed as given by eq 2:

$$\mathcal{L} = -\sum_{i \neq y} \max(0, f_i - f_y + \Delta) \tag{2}$$

where $f_i$ is the predicted score for class $i$, $f_y$ is the predicted score for the true class $y$ & $\Delta$ is the margin parameter which specifies the minimum difference required between the true class score and the scores of other classes.

Hinge loss operates on logits (raw score) and aims to maximize the score difference between true class and incorrect class based on the specified margin. It enforces the margin constraints between the correct class and other classes by penalizing the scores that fail to meet these margins and adjusting the logits to ensure the true class has a higher score relative to others, with varying degrees of emphasis based on the margin requirements. This margin control property affects score adjustments and subsequently impacts probability mass distribution.

Considering a three-class classification problem with Class A as the true class, where the initial logits are [87.2, 87.9, 87.5] and the corresponding SoftMax outputs are [0.28, 0.42, 0.30] for Classes A, B, and C, respectively. Our goal is to predict the correct class A while also ensuring that the probability pattern follows: prob(A) $>$ prob(C) $>$ prob(B). To achieve this, we set the margins as follows:

Margin 1 (between A and B): 1.0
Margin 2 (between A and C): 0.5

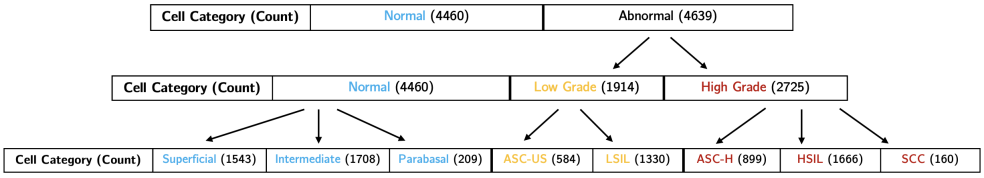


Figure 3: Detailed class-wise distribution of the CRIC dataset for 8-way classification task

After further training with hinge loss, the logits might update to [90, 86, 88], resulting in probabilities [0.86, 0.02, 0.12] for classes A, B, and C, respectively. The specified margins impose penalties to ensure greater separation between classes A and B compared to the separation between classes A and C.

### 4.3 Hinge Loss for Controlling Set Composition

In the previous sub-sections, we discussed how Hinge Loss affects the probability distribution by enforcing margin constraints between the correct and incorrect classes. This margin control property adjusts the scores and the probability mass distribution accordingly. It can be leveraged for effective control over the composition of conformal prediction sets by tuning the margin parameter. In this section, we provide a detailed discussion of how Hinge Loss can be integrated with the three CP methods to gain effective control over the composition of conformal prediction sets.

1. **LAC**: LAC generates prediction sets by including classes with probabilities above a specified threshold. By setting appropriate margins for hinge loss based on requirements, the distribution of probability mass can be adjusted through model training. This impacts the inclusion or exclusion of certain classes by making certain classes more or less likely to exceed the threshold. Therefore, integrating Hinge Loss with LAC can precisely adjust the model's probability outputs to fulfill specific prediction set requirements, effectively controlling the set composition.

2. **APS & RAPS**: Adaptive Prediction Sets (APS) & Regularized Adaptive Prediction Sets (RAPS) form prediction sets by including the ordered probabilities of all classes until their cumulative probability mass exceeds a predefined threshold. By setting appropriate margins for hinge loss based on requirements, the model can be trained to adjust the probabilities of certain classes, pushing them to either end of the ordered probabilities and impacting their inclusion or exclusion in the prediction set. Therefore, using Hinge Loss with APS/RAPS fine-tunes the model's probabilities to alter the ranking of certain classes in the ordered probabilities and effectively control the composition of prediction sets.

## 5 Methodology

In this section, we provide details about the dataset and the experimental setup for the multi-class classification task.

### 5.1 Dataset Used

The utilized dataset is from the Center for Recognition and Inspection of Cells (CRIC) Searchable Image Database [38], which was classified by three cytopathologists using the Bethesda System [44]. The dataset comprises 400 images at 1376 × 1020 pixels and 150 dpi from Pap smear tests, with each cell annotated by spatial coordinates. These cells were originally categorized into 6 classes: NILM, ASC-US, LSIL, ASC-H, HSIL, and SCC. For this study, our pathologist further divided the NILM class into Superficial, Intermediate, and Parabasal classes based on the Bethesda System and categorized all classes into three risk levels categories. The detailed class-wise distribution of the dataset is shown in Figure 3.

### 5.2 Experimental Setup

The dataset is split into 70% training/validation and 30% calibration/test sets, with 5-fold cross-validation applied to the training and validation set.We trained several underlying models for CP

| Coverage | LAC | APS | RAPS |
|---|---|---|---|
| *90%* | $1.78_{\pm 0.17}$ | $4.36_{\pm 0.26}$ | $2.12_{\pm 0.09}$ |
| *95%* | $2.4_{\pm 0.27}$ | $5.25_{\pm 0.32}$ | $2.73_{\pm 0.17}$ |
| *99%* | $3.78_{\pm 0.50}$ | $6.45_{\pm 0.31}$ | $4.28_{\pm 0.75}$ |

Table 1: Comparison of average set size for various CP methods across coverages.

| Model | AUC | Accuracy |
|---|---|---|
| **DieT** | $\mathbf{78.24}_{\pm 0.49}$ | $\mathbf{71.28}_{\pm 0.17}$ |
| ViT | $77.35_{\pm 0.39}$ | $70.76_{\pm 0.21}$ |
| ResNet 152 | $75.66_{\pm 0.35}$ | $68.29_{\pm 0.19}$ |
| DenseNet | $73.55_{\pm 0.31}$ | $63.29_{\pm 0.22}$ |
| ResNet 50 | $73.04_{\pm 0.41}$ | $68.90_{\pm 0.15}$ |
| RegNet | $72.27_{\pm 0.56}$ | $65.16_{\pm 0.39}$ |
| Xception | $70.80_{\pm 0.35}$ | $60.95_{\pm 0.17}$ |
| EfficientNet | $70.79_{\pm 0.49}$ | $60.93_{\pm 0.30}$ |

Table 2: Comparison of the AUC and accuracy for various underlying models.

methods, including DieT [52], ViT [13], ResNet152 [16], DenseNet [18], ResNet50 [16], RegNet [59], Xception [9], and EfficientNet [49] for 8-way classification on the CRIC dataset and evaluated their performance using AUC and accuracy. Each model was trained for 100 epochs with early stopping, using patience of 20 epochs and validation loss as the stopping criterion. The Adam optimizer was used with a learning rate of 0.0002, a weight decay of 5e-3, and a batch size of 512. DieT, which achieved the highest AUC and accuracy across the 5 folds, was chosen as the underlying model for our CP methods for all our experiments. We trained the models using both cross-entropy and hinge loss functions. For hinge loss, we set the low margin to 1 and the high margin to 5 based on the requirements of our experiments. We also performed a grid search for hyperparameter tuning ($k_{\text{reg}}$, $\lambda$) for the RAPS method. For all experiments, we evaluated the desired metrics for each CP method across three coverage levels (i.e. 90%, 95% & 99%) and conducted 100 trials to obtain robust and reliable results.

## 6 Experiments

In this section, we provide detailed information on our three experiments and the associated metrics used to evaluate CP methods.

### 6.1 Risk-Based Controlled Set Sizes

The pathological diagnostic workflow is crucial for precise cancer diagnosis and treatment. The WHO estimates a sharp increase in cancer cases, from 20 million in 2022 to 35 million by 2050 [7]. However, pathologists constitute only 0.8% of physicians globally [32], which is insufficient to meet this rising demand. This results in significant disparities in pathologist-to-population ratios and imposes a heavy workload on pathologists, involving samples with varying risk levels and requiring specialized expertise. This increased workload and diverse cases can affect the quality and efficiency of diagnostics.

To tackle this issue, pathologists need to manage their workflow efficiently to allocate time to each case based on its risk level to ensure accurate diagnoses. This necessitates conformal methods that adjust prediction set sizes based on risk levels to align with pathological workflows.

In our study, we aim to establish a descending order of set sizes based on the associated risk level of each diagnosis, where High Grade cells have the largest set size, Low-Grade cells have a moderate set size, and Normal cells have the smallest set size. We track the average set size of prediction sets across varying risk levels to achieve the order of set sizes as mentioned in eq 3.

$$\textit{C(High Grade)} > \textit{C(Low Grade)} > \textit{C(Normal)} \tag{3}$$

where $C(A)$ denotes the set size or cardinality of A.

Thus, a large set size for a particular diagnosis can signal to the pathologist that the specimen is likely in the High-Grade category, which requires more time and expertise due to its higher risk. In contrast, smaller set sizes suggest that the specimen is in the Normal or Low-Grade category, involving lower risk and allowing for faster processing. This risk-based approach to controlling set size will help streamline workflow management and improve diagnostic accuracy, ultimately leading to better patient outcomes.

### 6.2 Avoiding Classes from Various Risk Levels

Pathological diagnostic workflow often handles a variety of conditions, including common low-risk cases, borderline cases, and rare high-risk conditions. The accurate differentiation between these varied risk levels of cases is essential for providing explicit diagnoses and delivering effective treatments for each case. However, this need for accurate differentiation of cases can be challenging and heighten the cognitive workload for pathologists, potentially leading to a higher likelihood of misdiagnosis and greater patient anxiety.

Although CP methods used in medical settings provide reliable results, they often lack control over the compositionality of prediction sets, resulting in prediction sets that include conditions from various risk levels. This poses a significant challenge when integrating CP methods into pathological diagnosis workflows. For example, if a pathologist encounters a prediction set with conditions of varying risk levels, it can lead to diagnostic ambiguity and increased patient anxiety due to additional tests.

Focusing on this problem, our experiment aims to enhance conformal predictors by providing effective control over the compositionality of prediction sets to avoid the inclusion of classes with different risk levels such as Normal, Low Grade, and High Grade. We track the reduction of prediction sets with overlapping classes across various risk levels using eq 4:

$$\textit{Overlapping Set \%} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}! \, , y_i \in \mathcal{C}(x_i), \tag{4}$$

where $N_{\text{overlap}}$ is the total number of cases with overlapping classes in their confidence sets & $N_{\text{total}}$ is the total number of test cases.

This approach will minimize diagnostic ambiguity for pathologists, leading to increased diagnostic efficiency and improved patient outcomes through limited unnecessary testing.

### 6.3 Avoiding Confusing Classes

Pathologists frequently encounter the challenge of distinguishing between diseases with similar characteristics that require distinct treatments. For example, distinguishing between the two most common non-melanoma skin cancers, Basal Cell Carcinoma(BCC) and Squamous Cell Carcinoma(SCC), is challenging due to overlapping clinical and histological features [41].

Providing specialized expertise for such critical and ambiguous cases is pivotal for accurate diagnoses and effective treatments. However, these confusing cancer cell types often hinder reaching a consensus on its pathological interpretation, leading to misdiagnosis and impacting patient turnaround times.

CP methods aim to include the true class within a user-defined coverage, but this approach is inadequate for pathological workflows. There is an inherent need for CP methods to avoid including confusing classes along with the true class in the prediction sets. This shall reduce diagnostic confusion and assist pathologists in making accurate diagnoses.

Our experiment aims to reduce the occurrence of prediction sets with the confusing cancer cell pairs listed below, which pathologists often misidentify due to similar cytological features [20]:

1. Atypical Squamous Cells, cannot exclude HSIL (ASC-H) and High-Grade Squamous Intraepithelial Lesions (HSIL)
2. Atypical Squamous Cells of Undetermined Significance (ASC-US) and Low-Grade Squamous Intraepithelial Lesions (LSIL)

| | LAC | | APS | | RAPS | |
|---|---|---|---|---|---|---|
| | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE |
| **Covr.= 90%** | | | | | | |
| *Normal* | $1.73_{\pm 0.15}$ | $1.75_{\pm 0.17}$ | $3.93_{\pm 0.19}$ | $3.58_{\pm 0.22}$ | $1.99_{\pm 0.20}$ | $1.84_{\pm 0.26}$ |
| *Low Grade* | $2.23_{\pm 0.26}$ | $2.41_{\pm 0.2}$ | $5.66_{\pm 0.24}$ | $4.68_{\pm 0.29}$ | $2.75_{\pm 0.36}$ | $2.53_{\pm 0.29}$ |
| *High Grade* | $1.68_{\pm 0.15}$ | $2.68_{\pm 0.29}$ | $4.15_{\pm 0.22}$ | $5.08_{\pm 0.27}$ | $1.98_{\pm 0.23}$ | $2.82_{\pm 0.45}$ |
| **Covr.= 95%** | | | | | | |
| *Normal* | $2.17_{\pm 0.19}$ | $2.29_{\pm 0.21}$ | $4.27_{\pm 0.33}$ | $4.64_{\pm 0.96}$ | $3.87_{\pm 0.58}$ | $2.4_{\pm 0.28}$ |
| *Low Grade* | $2.95_{\pm 0.32}$ | $3.04_{\pm 0.26}$ | $6.60_{\pm 0.36}$ | $5.77_{\pm 0.75}$ | $5.50_{\pm 0.70}$ | $3.18_{\pm 0.33}$ |
| *High Grade* | $2.15_{\pm 0.21}$ | $3.75_{\pm 0.29}$ | $5.04_{\pm 0.39}$ | $6.08_{\pm 0.66}$ | $2.53_{\pm 0.18}$ | $3.69_{\pm 0.39}$ |
| **Covr.= 99%** | | | | | | |
| *Normal* | $3.35_{\pm 0.48}$ | $3.45_{\pm 0.46}$ | $6.06_{\pm 0.45}$ | $7.72_{\pm 0.47}$ | $3.87_{\pm 0.58}$ | $3.55_{\pm 0.36}$ |
| *Low Grade* | $4.74_{\pm 0.70}$ | $4.49_{\pm 0.61}$ | $7.46_{\pm 0.2}$ | $7.81_{\pm 0.3}$ | $5.50_{\pm 0.7}$ | $4.67_{\pm 0.48}$ |
| *High Grade* | $3.42_{\pm 0.53}$ | $4.92_{\pm 0.55}$ | $6.49_{\pm 0.45}$ | $7.92_{\pm 0.55}$ | $4.06_{\pm 0.63}$ | $5.05_{\pm 0.43}$ |

Table 3: Comparision of average set size for cross-entropy and hinge loss used for CP methods across various coverages.

We track the reduction of prediction sets with confusing cell class pairs across various risk levels using eq 5:

$$\textit{Confusing Set \%} = \frac{N_{\text{confusing}}}{N_{\text{total}}} \times 100 \tag{5}$$

where $N_{\text{confusing}}$ is the total number of cases with confusing class pairs in their confidence sets & $N_{\text{total}}$ is the total number of test cases.

This approach will reduce patient turnaround time, leading to improved diagnostic workflow and overall patient care.

## 7 Results & Discussion

In this section, we present the results for each of our experiments, along with their corresponding analyses and findings.

### 7.1 Risk-Based Controlled Set Sizes

In Table 3, we track the average prediction set sizes of Normal, Low Grade, and High Grade categories for various CP methods across three coverage levels, with the underlying model trained using cross-entropy and hinge loss.

Table 3 demonstrates that CE loss with any CP method does not produce set sizes ordered by associated risk levels at any coverage setting. In contrast, hinge loss used with any CP method across all coverage settings consistently achieves the desired order set sizes, where High-Grade cells have the largest set sizes, Low-Grade cells have moderate set sizes, and Normal cells have the smallest sets.

The desired set size order is achieved by adjusting the margin control property of Hinge Loss for each category as follows:

1. *Normal Grade*: When the true class (e.g., Superficial cell, Intermediate cell, Parabasal cell) is in the Normal category, we apply a high margin to the other 7 incorrect classes. This high margin tunes the probability of the true class to be substantially higher than all incorrect ones, limiting the inclusion of incorrect classes and resulting in the smallest set sizes.

2. *Low Grade (LG)*: When the true class (e.g., ASC-US, LSIL) belongs to the Low-Grade category, we ease the margin constraints compared to the Normal category. We use a high margin for any 5 incorrect classes and a lower margin for the remaining 2 incorrect classes. This adjustment allocates a higher probability mass to the true class and the two incorrect classes corresponding to lower margins

| | LAC | | APS | | RAPS | |
|---|---|---|---|---|---|---|
| | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE |
| **Coverage** | | | | | | |
| *90%* | $25.40_{\pm 2.39}$ | $17.15_{\pm 1.24}$ | $69.28_{\pm 3.68}$ | $37.03_{\pm 5.53}$ | $31.25_{\pm 3.18}$ | $19.06_{\pm 1.16}$ |
| *95%* | $33.88_{\pm 4.08}$ | $25.65_{\pm 5.80}$ | $80.66_{\pm 4.19}$ | $57.95_{\pm 5.5}$ | $41.19_{\pm 3.81}$ | $27.17_{\pm 0.60}$ |
| *99%* | $57.76_{\pm 8.98}$ | $53.69_{\pm 11.8}$ | $92.22_{\pm 3.19}$ | $87.58_{\pm 7.84}$ | $64.56_{\pm 6.48}$ | $60.81_{\pm 5.30}$ |

Table 4: Comparision of *Overlapping Set %* for cross-entropy and hinge loss used for CP methods across various coverages.

| | LAC | | APS | | RAPS | |
|---|---|---|---|---|---|---|
| | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE |
| **Coverage** | | | | | | |
| *90%* | $19.47_{\pm 2.91}$ | $8.89_{\pm 1.64}$ | $42.2_{\pm 1.11}$ | $23.25_{\pm 1.60}$ | $23.54_{\pm 2.66}$ | $11.0_{\pm 0.79}$ |
| *95%* | $27.72_{\pm 3.18}$ | $14.37_{\pm 2.56}$ | $45.41_{\pm 0.84}$ | $28.75_{\pm 2.55}$ | $31.59_{\pm 3.60}$ | $15.65_{\pm 1.67}$ |
| *99%* | $38.78_{\pm 3.03}$ | $27.20_{\pm 4.86}$ | $48.05_{\pm 0.72}$ | $43.35_{\pm 5.75}$ | $41.08_{\pm 3.58}$ | $20.87_{\pm 1.83}$ |

Table 5: Comparision of *Confusing Set %* for Cross Entropy and Hinge loss used for CP methods across various coverages.

relative to the remaining 5 incorrect classes with high margins, resulting in a larger prediction set size than the Normal group but smaller than the High-Grade group.

3. *High Grade (HG)*: For the High-Grade category (e.g., ASC-H, HSIL, SCC), we further ease the margin constraints. In this case, we apply a high margin to any 4 incorrect classes, while the remaining 3 incorrect classes receive a lower margin. This adjustment allocates a higher probability mass to the true class and the 3 incorrect classes corresponding to lower margins relative to the remaining 4 incorrect classes with high margins, resulting in the largest prediction set size among the three categories.

This strategical utilization of hinge loss margins across different risk categories allows for effective control over the compositionality of prediction sets, resulting in the desired order of set sizes based on associated risk levels.

### 7.2 Avoiding Classes from Various Risk Levels

In Table 4, we track the *Overlapping Set %* for various CP methods across 3 coverage guarantees, with the underlying model trained using cross-entropy and hinge loss.

Table 4 demonstrates that the *Overlapping Set %* for Hinge Loss is consistently lower than for CE Loss across all CP methods and coverage settings. The difference of *Overlapping Set %* ranges from 5% to 46% across various CP methods and coverage settings. This reduction in *Overlapping Set %* is achieved by adjusting the margin control property of Hinge Loss for all groups as detailed below:

For a true class in a specific group, we apply low margins for all classes in that category and high margins for all other classes. This approach concentrates a higher probability mass on the classes belonging to the correct category, thereby reducing the inclusion of classes from all incorrect categories. For example, if the true class belongs to the Normal category, we assign low margins to all Normal classes and high margins to Low-Grade and High-Grade classes. This approach concentrates a greater probability mass on the Normal classes, ultimately reducing the likelihood of including classes from the Low Grade & High-Grade categories in the prediction set.

This effective usage of the Hinge loss margin results in a reduced *Overlapping Set %* compared to CE loss.

### 7.3 Avoiding Confusing Classes

In Table 5, we track the *Confusing Set %* for various CP methods across 3 coverage guarantees, with the underlying model trained using cross-entropy and hinge loss.

Table 5 demonstrates that the *Confusing Set %* for Hinge Loss is consistently lower than that for CE Loss across all CP methods and coverage settings. The difference of *Overlapping Set %* ranges from

9% and 54% across various CP methods and coverage settings. This reduction in *Confusing Set %* is achieved by adjusting the margin control property of Hinge Loss for all groups as detailed below:

For a true class belonging to a confused class pair, we apply a high margin to its corresponding confused class and low margins for all other classes. This approach reduces the probability mass assigned to the corresponding confusing class, thereby lowering the likelihood of its inclusion in the prediction set. For example, if the true class is ASC-H, we assign a high margin to its corresponding confused class i.e., HSIL and low margins for all other classes. This approach reduces the likelihood of including HSIL in the prediction set.

This effective usage of the Hinge Loss margins results in a reduced *Confusing Set %* compared to CE loss.

## 8 Conclusion

In this study, we propose a novel training method using Hinge loss for underlying models in CP methods, which enables effective control over the compositionality of prediction sets. We evaluate this method using three application-specific metrics vital for deploying and aligning CP methods with real world pathological workflows. Our results indicate that the Hinge loss approach consistently surpasses the traditional Cross Entropy loss method across all three metrics for the CRIC dataset, demonstrating effective control over set compositionality. We believe that this control over the composition of prediction sets is crucial for better alignment with the needs of pathologists, leading to smoother integration into clinical workflows and enhancing overall patient care.

## References


[1] Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

[2] Anastasios Nikolas Angelopoulos, Stuart R Pomerantz, Synho Do, Stephen Bates, Christopher P Bridge, Daniel C Elton, Michael H Lev, R Gilberto Gonzalez, Michael I Jordan, and Jitendra Malik. Conformal triage for medical imaging ai deployment. *medRxiv*, pages 2024–02, 2024.

[3] Cagla Deniz Bahadir, Mohamed Omar, Jacob Rosenthal, Luigi Marchionni, Benjamin Liechty, David J Pisapia, and Mert R Sabuncu. Artificial intelligence applications in histopathology. *Nature Reviews Electrical Engineering*, 1(2):93–108, 2024.

[4] Anthony Bellotti. Optimized conformal classification using gradient descent approximation. *arXiv preprint arXiv:2105.11255*, 2021.

[5] Thomas Bonnier and Benjamin Bosch. Engineering uncertainty representations to monitor distribution shifts. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

[6] Lei Cao, Jinying Yang, Zhiwei Rong, Lulu Li, Bairong Xia, Chong You, Ge Lou, Lei Jiang, Chun Du, Hongxue Meng, et al. A novel attention-guided convolutional network for the detection of abnormal cervical cells in cervical cancer screening. *Medical image analysis*, 73:102197, 2021.

[7] Wei Cao, Kang Qin, Feng Li, and Wanqing Chen. Comparative study of cancer profiles between 2020 and 2022 using global cancer statistics (globocan). *Journal of the National Cancer Center*, 2024.

[8] Shenghua Cheng, Sibo Liu, Jingya Yu, Gong Rao, Yuwei Xiao, Wei Han, Wenjie Zhu, Xiaohua Lv, Ning Li, Jing Cai, et al. Robust whole slide image analysis for cervical cancer screening using deep learning. *Nature communications*, 12(1):5639, 2021.

[9] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[10] Miao Cui and David Y Zhang. Artificial intelligence and computational pathology. *Laboratory Investigation*, 101(4):412–422, 2021.

[11] Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] James M Dolezal, Andrew Srisuwananukorn, Dmitry Karpeyev, Siddhi Ramesh, Sara Kochanny, Brittany Cody, Aaron S Mansfield, Sagar Rakshit, Radhika Bansal, Melanie C Bois, et al. Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. *Nature communications*, 13(1):6572, 2022.

[13] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. Training uncertainty-aware classifiers with conformalized deep learning. *Advances in Neural Information Processing Systems*, 35:22380–22395, 2022.

[15] Clara Fannjiang, Stephen Bates, Anastasios N Angelopoulos, Jennifer Listgarten, and Michael I Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022.

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

[17] Mahdi S Hosseini, Babak Ehteshami Bejnordi, Vincent Quoc-Huy Trinh, Lyndon Chan, Danial Hasan, Xingwen Li, Stephen Yang, Taehyo Kim, Haochen Zhang, Theodore Wu, et al. Computational pathology: a survey review and the way forward. *Journal of Pathology Informatics*, page 100357, 2024.

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[19] Dongyao Jia, Jialin Zhou, and Chuanwang Zhang. Detection of cervical cells based on improved ssd network. *Multimedia Tools and Applications*, 81(10):13371–13387, 2022.

[20] Jayasree Kattoor and Meherbano M Kamal. The gray zone squamous lesions: Asc-us/asc-h. *CytoJournal*, 19, 2022.

[21] Nfn Kiran, FNU Sapna, FNU Kiran, Deepak Kumar, FNU Raja, Sheena Shiwlani, Antonella Paladini, FNU Sonam, Ahmed Bendari, Raja Sandeep Perkash, et al. Digital pathology: transforming diagnosis in the digital age. *Cureus*, 15(9), 2023.

[22] Bhawesh Kumar, Anil Palepu, Rudraksh Tuwani, and Andrew Beam. Towards reliable zero shot classification in self-supervised models with conformal prediction. *arXiv preprint arXiv:2210.15805*, 2022.

[23] Rohit Kundu and Soham Chattopadhyay. Deep features selection through genetic algorithm for cervical pre-cancerous cell classification. *Multimedia Tools and Applications*, 82(9):13431–13452, 2023.

[24] Sunil R Lakhani, Ian O Ellis, Stuart Schnitt, Puay Hoon Tan, and Marc van de Vijver. Who classification of tumours of the breast. *Volume 4. IARC WHO Classification of Tumours*, 2012.

[25] Jing Lei, Alessandro Rinaldo, and Larry Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74:29–43, 2015.

[26] Wanli Liu, Chen Li, Ning Xu, Tao Jiang, Md Mamunur Rahaman, Hongzan Sun, Xiangchen Wu, Weiming Hu, Haoyuan Chen, Changhao Sun, et al. Cvm-cervix: A hybrid cervical pap-smear image classification framework using cnn, visual transformer and multilayer perceptron. *Pattern Recognition*, 130:108829, 2022.

[27] David N Louis, Georg K Gerber, Jason M Baron, Lyn Bry, Anand S Dighe, Gad Getz, John M Higgins, Frank C Kuo, William J Lane, James S Michaelson, et al. Computational pathology: an emerging definition. *Archives of pathology & laboratory medicine*, 138(9):1133–1138, 2014.

[28] Charles Lu, Anastasios N Angelopoulos, and Stuart Pomerantz. 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*, pages 545–554. Springer, 2022.

[29] Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12008–12016, 2022.

[30] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[31] Ankur Manna, Rohit Kundu, Dmitrii Kaplun, Aleksandr Sinitca, and Ram Sarkar. A fuzzy rank-based ensemble of cnn models for classification of cervical cytology. *Scientific Reports*, 11(1):14538, 2021.

[32] Bruno Märkl, Laszló Füzesi, Ralf Huss, Svenja Bauer, and Tina Schaller. Number of pathologists in germany: comparison with european countries, usa, and canada. *Virchows Archiv*, 478:335–341, 2021.

[33] Paul Novello, Joseba Dalmau, and Léo Andeol. Out-of-distribution detection should use conformal prediction (and vice-versa?). *arXiv preprint arXiv:2403.11532*, 2024.

[34] Shubham Ojha and Aditya Narendra. Uncertainty quantification in dl models for cervical cytology. In *Medical Imaging with Deep Learning*, 2024.

[35] Milda Pocevičiūtė, Gabriel Eilertsen, Sofia Jarkman, and Claes Lundström. Generalisation effects of predictive uncertainty estimation in deep learning for digital pathology. *Scientific Reports*, 12(1):8329, 2022.

[36] Ioannis Prassas, Blaise Clarke, Timothy Youssef, Juliana Phlamon, Lampros Dimitrakopoulos, Andrew Rofaeil, and George M Yousef. Computational pathology: an evolving concept. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 62, 2024.

[37] Md Mamunur Rahaman, Chen Li, Yudong Yao, Frank Kulwa, Xiangchen Wu, Xiaoyan Li, and Qian Wang. Deepcervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques. *Computers in Biology and Medicine*, 136:104649, 2021.

[38] Mariana T Rezende, Raniere Silva, Fagner de O Bernardo, Alessandra HG Tobias, Paulo HC Oliveira, Tales M Machado, Caio S Costa, Fatima NS Medeiros, Daniela M Ushizima, Claudia M Carneiro, et al. Cric searchable image database as a public platform for conventional pap smear cytology data. *Scientific data*, 8(1):151, 2021.

[39] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[40] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

[41] Tea Hyung Ryu, Heesang Kye, Jae Eun Choi, Hyo Hyun Ahn, Young Chul Kye, and Soo Hong Seo. Features causing confusion between basal cell carcinoma and squamous cell carcinoma in clinical diagnosis. *Annals of dermatology*, 30(1):64–70, 2018.

[42] Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

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

[44] D Solomon, RL Aamodt, K Anderson, P Ashton, L Bergner, M Bodo, DN Collins, V Conley, IS Cox, YS Erozan, et al. The 1988 bethesda system for reporting cervical/vaginal cytologic diagnoses. developed and approved at the national cancer institute workshop, bethesda, maryland, usa, december 12-13, 1988. *Acta cytologica*, 33(5):567–575, 1989.

[45] Andrew H Song, Guillaume Jaume, Drew FK Williamson, Ming Y Lu, Anurag Vaidya, Tiffany R Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12):930–949, 2023.

[46] Youyi Song, Ee-Leng Tan, Xudong Jiang, Jie-Zhi Cheng, Dong Ni, Siping Chen, Baiying Lei, and Tianfu Wang. Accurate cervical cell segmentation from overlapping clumps in pap smear images. *IEEE transactions on medical imaging*, 36(1):288–300, 2016.

[47] Fabio A. Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016.

[48] David Stutz, Krishnamurthy Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. *CoRR*, abs/2110.09192, 2021.

[49] Mingxing Tan. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[50] Afaf Tareef, Yang Song, Heng Huang, Yue Wang, Dagan Feng, Mei Chen, and Weidong Cai. Optimizing the cervix cytological examination based on deep learning and dynamic shape modeling. *Neurocomputing*, 248:28–40, 2017.

[51] Nishant Thakur, Mohammad Rizwan Alam, Jamshid Abdul-Ghafar, and Yosep Chong. Recent application of artificial intelligence in non-gynecological cancer cytopathology: a systematic review. *Cancers*, 14(14):3529, 2022.

[52] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[53] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

[54] Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28, 2015.

[55] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[56] Tao Wan, Shusong Xu, Chen Sang, Yulan Jin, and Zengchang Qin. Accurate segmentation of overlapping cells in cervical cytology with deep convolutional neural networks. *Neurocomputing*, 365:157–170, 2019.

[57] Cholatip Wiratkapun, Pawat Piyapan, Panuwat Lertsithichai, and Noppadol Larbcharoensub. Fibroadenoma versus phyllodes tumor: distinguishing factors in patients diagnosed with fibroepithelial lesions after a core needle biopsy. *Diagnostic and Interventional Radiology*, 20(1):27, 2014.

[58] Chuanyun Xu, Mengwei Li, Gang Li, Yang Zhang, Chengjie Sun, and Nanlan Bai. Cervical cell/clumps detection in cytology images using transfer learning. *Diagnostics*, 12(10):2477, 2022.

[59] Jing Xu, Yu Pan, Xinglin Pan, Steven Hoi, Zhang Yi, and Zenglin Xu. Regnet: Self-regulated network for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):9562–9567, 2022.

[60] Xiaoli Yan and Zhongsi Zhang. Hsdet: A representative sampling based object detector in cervical cancer cell images. In *Bio-Inspired Computing: Theories and Applications: 15th International Conference, BIC-TA 2020, Qingdao, China, October 23-25, 2020, Revised Selected Papers 15*, pages 406–418. Springer, 2021.

[61] Xianghao Zhan, Zhan Wang, Meng Yang, Zhiyuan Luo, You Wang, and Guang Li. An electronic nose-based assistive diagnostic prototype for lung cancer detection with conformal prediction. *Measurement*, 158:107588, 2020.

[62] Han Zhang, Hongqing Zhu, and Xiaofeng Ling. Polar coordinate sampling-based segmentation of overlapping cervical cells using attention u-net and random walk. *Neurocomputing*, 383:212–223, 2020.

[63] Yizhe Zhang, Shuo Wang, Yejia Zhang, and Danny Z Chen. 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*, pages 12–21. Springer, 2023.

# A Supplemental Material

This supplementary material contains additional qualitative results and other details for the Hinge Loss training method applied to the BreakHis dataset.

# B Dataset

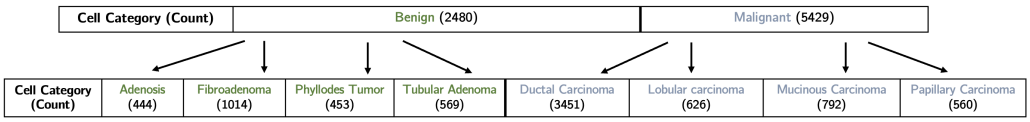


Figure 4: Detailed class-wise distribution of the BreakHist dataset for 8-way classification task

The dataset used is the Breast Cancer Histopathological Image Classification (BreakHis) dataset [47] comprising 7,909 microscopic images of breast tumour tissue from 82 patients. These images were obtained using the SOB method, also known as partial mastectomy or excisional biopsy at a resolution of 700x460 pixels across different magnification levels (40x, 100x, 200x, and 400x). The dataset contains four histological distinct types of benign breast tumours: Adenosis (A), Fibroadenoma (F), Phyllodes Tumour (PT), and Tubular Adenona (TA); and four malignant tumours (breast cancer): Ductal Carcinoma (DC), Lobular Carcinoma (LC), Mucinous Carcinoma (MC) and Papillary Carcinoma (PC). The detailed class-wise distribution of the dataset is shown below in Figure 4.

| **Coverage** | **LAC** | **APS** | **RAPS** |
|---|---|---|---|
| *90%* | $1.15_{\pm 0.04}$ | $4.66_{\pm 0.14}$ | $1.40_{\pm 0.01}$ |
| *95%* | $1.39_{\pm 0.07}$ | $5.88_{\pm 0.11}$ | $1.66_{\pm 0.08}$ |
| *99%* | $2.55_{\pm 0.40}$ | $6.85_{\pm 0.25}$ | $2.93_{\pm 0.28}$ |

Table 6: Comparison of average set size for various CP methods across coverages.

# C Experiments

In this section, we provide detailed information on 3 experiments and their corresponding metrics for evaluating CP methods.

## C.1 Risk-Based Controlled Set Sizes

In our experiment, we aim to establish a descending order of set sizes based on the associated risk level of each diagnosis with the Malignant cells category having a larger set size than the Benign cells category. We track the average set size of prediction sets across varying risk levels to achieve the order of set sizes as mentioned in eq 6:

$$C(Malignant) > C(Benign) \tag{6}$$

where $C(A)$ denotes the set size or cardinality of A.

Therefore, a large set size for a certain diagnosis can serve as a signal to the pathologist that the specimen likely falls under the Malignant category, requiring more time and expertise due to its higher associated risk. In contrast, smaller set sizes suggest the specimen is in the Benign category, presenting lower risk and enabling faster processing. This risk-based controlled set size approach will streamline workflow management and enhance diagnostic accuracy, leading to improved patient outcomes.

## C.2 Avoiding Classes from Various Risk Levels

Our experiment aims to enhance conformal predictors by providing effective control over the compositionally of prediction sets to avoid the inclusion of classes with varied risk categories namely

| Model | AUC | Accuracy |
|---|---|---|
| **DieT** | $\mathbf{89.68}_{\pm 0.59}$ | $\mathbf{84.53}_{\pm 0.19}$ |
| ViT | $88.64_{\pm 0.27}$ | $83.45_{\pm 0.18}$ |
| ResNet 152 | $87.34_{\pm 0.23}$ | $82.25_{\pm 0.29}$ |
| DenseNet | $85.64_{\pm 0.48}$ | $78.87_{\pm 0.44}$ |
| ResNet 50 | $84.98_{\pm 0.64}$ | $78.68_{\pm 0.59}$ |
| RegNet | $83.43_{\pm 0.57}$ | $77.23_{\pm 0.33}$ |
| Xception | $81.15_{\pm 0.17}$ | $72.34_{\pm 0.67}$ |
| EfficientNet | $80.89_{\pm 0.29}$ | $71.87_{\pm 0.27}$ |

Table 7: Comparison of the AUC and accuracy for various underlying models.

Benign and Malignant. We track the reduction of prediction sets with overlapping classes across various risk categories using eq 7:

$$\textit{Overlapping Set \%} = \frac{N_{\text{overlap}}}{N_{\text{total}}} \times 100 \tag{7}$$

where $N_{\text{overlap}}$ is the total number of cases with overlapping classes in their confidence sets & $N_{\text{total}}$ is the total number of test cases.

This method will reduce diagnostic uncertainty for pathologists, enhancing diagnostic efficiency and improving patient outcomes by minimizing unnecessary testing.

### C.3 Avoiding Confusing Classes

This experiment aims to reduce the occurrence of prediction sets with the confusing cancer cell pairs listed below, which pathologists often misidentify due to similar histological features [57, 24] :

1. Ductal Carcinoma (DC) and Lobular Carcinoma (LC)
2. Fibroadenoma (F) and Phyllodes Tumor (PT)

We track the reduction of prediction sets with confusing cell class pairs across various risk levels using eq 8:

$$\textit{Confusing Set \%} = \frac{N_{\text{confusing}}}{N_{\text{total}}} \times 100 \tag{8}$$

where $N_{\text{confusing}}$ is the total number of cases with confusing class pairs in their confidence sets & $N_{\text{total}}$ is the total number of test cases.

This approach will shorten patient turnaround times, improving diagnostic workflows and enhancing overall patient care.

## D Results & Discussion

In this section, we present the results for each of our experiments, along with their corresponding analyses and findings.

### D.1 Risk-Based Controlled Set Sizes

In Table 8, we track the average prediction set sizes of Benign and Malignant categories for various CP methods across three coverage levels, with the underlying model trained using cross-entropy and hinge loss.

Table 8 shows that Hinge Loss, when used with any CP method across all coverage settings, consistently produces the desired set size order, with Malignant cells having larger set sizes than Benign cells. In contrast, CE loss with any CP method fails to achieve risk-ordered set sizes at any coverage setting. The desired set size order is achieved by adjusting the margin control property of Hinge Loss for each category as follows:

| | LAC | | APS | | RAPS | |
|---|---|---|---|---|---|---|
| | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE |
| **Covr.= 90%** | | | | | | |
| *Benign* | $1.19_{\pm 0.04}$ | $1.03_{\pm 0.01}$ | $5.057_{\pm 0.16}$ | $3.90_{\pm 0.26}$ | $1.47_{\pm 0.005}$ | $1.13_{\pm 0.01}$ |
| *Malignant* | $1.17_{\pm 0.03}$ | $1.17_{\pm 0.05}$ | $4.55_{\pm 0.19}$ | $7.82_{\pm 0.05}$ | $1.37_{\pm 0.02}$ | $1.86_{\pm 0.07}$ |
| **Covr.= 95%** | | | | | | |
| *Benign* | $1.44_{\pm 0.08}$ | $1.11_{\pm 0.02}$ | $6.12_{\pm 0.14}$ | $7.45_{\pm 0.97}$ | $1.74_{\pm 0.07}$ | $1.25_{\pm 0.04}$ |
| *Malignant* | $1.35_{\pm 0.06}$ | $1.61_{\pm 0.10}$ | $5.70_{\pm 0.17}$ | $7.99_{\pm 0.004}$ | $1.60_{\pm 0.05}$ | $2.46_{\pm 0.19}$ |
| **Covr.= 99%** | | | | | | |
| *Benign* | $3.10_{\pm 0.24}$ | $1.82_{\pm 0.35}$ | $7.07_{\pm 0.14}$ | $7.92_{\pm 0.009}$ | $3.27_{\pm 0.27}$ | $2.03_{\pm 0.21}$ |
| *Malignant* | $2.68_{\pm 0.22}$ | $4.98_{\pm 1.05}$ | $6.78_{\pm 0.15}$ | $7.99_{\pm 0.001}$ | $2.84_{\pm 0.27}$ | $5.37_{\pm 0.33}$ |

Table 8: Comparision of average set size for cross-entropy and hinge loss used for CP methods across various coverages.

| | LAC | | APS | | RAPS | |
|---|---|---|---|---|---|---|
| | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE |
| **Coverage** | | | | | | |
| *90%* | $5.33_{\pm 0.64}$ | $2.93_{\pm 0.23}$ | $65.91_{\pm 2.29}$ | $6.91_{\pm 0.44}$ | $10.37_{\pm 0.71}$ | $3.68_{\pm 0.24}$ |
| *95%* | $10.12_{\pm 1.02}$ | $3.28_{\pm 0.28}$ | $78.65_{\pm 1.34}$ | $7.94_{\pm 0.80}$ | $14.62_{\pm 1.22}$ | $4.50_{\pm 0.32}$ |
| *99%* | $32.02_{\pm 4.42}$ | $5.80_{\pm 0.83}$ | $90.1_{\pm 2.66}$ | $15.27_{\pm 3.28}$ | $37.42_{\pm 7.52}$ | $6.92_{\pm 0.83}$ |

Table 9: Comparison of *Overlapping Set %* for cross-entropy and hinge loss used for CP methods across various coverages.

1. *Benign cells*: When the true class (e.g., Adenosis, Fibroadenoma, Phyllodes Tumour & Tubular Adenona) belongs to the Benign cells category, we apply a high margin to the 7 incorrect classes. This high margin tunes the probability of the true class to be substantially higher than all incorrect ones, limiting the inclusion of incorrect classes and resulting in the smallest set sizes.

2. *Malignant cells*: For the Malignant cells category (e.g., Ductal Carcinoma, Lobular Carcinoma, Mucinous Carcinoma & Papillary Carcinoma), we ease the margin constraints. In this case, we apply a low margin to all 7 incorrect classes. This adjustment allocates a uniform probability mass distribution across the 7 incorrect classes, resulting in a larger prediction set size than normal.

This strategical utilization of hinge loss margins across different risk categories allows for effective control over the compositionality of prediction sets, resulting in the desired order of set sizes based on associated risk levels.

### D.2 Avoiding Classes from Various Risk Levels

In Table 9, we track the *Overlapping Set %* for various CP methods across coverage guarantees, with the underlying model trained using cross entropy and hinge loss.

Table 9 demonstrates that the *Overlapping Set %* for Hinge Loss is consistently lower than for CE Loss across all CP methods and coverage settings. The difference of *Overlapping Set %* ranges from 5% to 46% across various CP methods and coverage settings. This reduction in *Overlapping Set %* is achieved by adjusting the margin control property of Hinge Loss for all groups as detailed below:

For a true class in a specific group, we apply low margins for all classes in that category and high margins for all other classes. This approach concentrates a higher probability mass on the classes belonging to the correct category, thereby reducing the inclusion of classes from all incorrect categories. For example, if the true class belongs to the Benign category, we assign low margins to all Benign cell classes and high margins to Malignant classes. This approach concentrates a greater probability mass on the Benign classes, ultimately reducing the likelihood of including classes from the Malignant category in the prediction set.

This effective usage of the Hinge Loss margins results in a reduced *Overlapping Set %* compared to CE loss.

| | LAC | | APS | | RAPS | |
|---|---|---|---|---|---|---|
| | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE |
| **Coverage** | | | | | | |
| *90%* | $10.48_{\pm 0.46}$ | $7.07_{\pm 0.64}$ | $56.45_{\pm 1.07}$ | $20.27_{\pm 1.32}$ | $15.61_{\pm 1.06}$ | $10.02_{\pm 0.38}$ |
| *95%* | $15.76_{\pm 1.19}$ | $11.17_{\pm 0.80}$ | $61.36_{\pm 1.02}$ | $25.82_{\pm 2.06}$ | $21.36_{\pm 0.87}$ | $12.30_{\pm 0.93}$ |
| *99%* | $40.86_{\pm 4.59}$ | $24.24_{\pm 3.49}$ | $67.79_{\pm 1.23}$ | $44.20_{\pm 5.18}$ | $41.07_{\pm 3.41}$ | $21.89_{\pm 1.67}$ |

Table 10: Comparision of *Confusing Set %* for cross-entropy and hinge loss used for CP methods across various coverages .

| | LAC | | APS | | RAPS | |
|---|---|---|---|---|---|---|
| | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE |
| **Covr.= 90%** | | | | | | |
| *Normal* | 0 | 2.26 | 0 | 0 | 0 | 0 |
| *Low Grade* | 7.55 | 1.10 | 0 | 0 | 0 | 0 |
| *High Grade* | 0.40 | 0 | 0 | 0 | 0 | 0 |
| **Covr.= 95%** | | | | | | |
| *Normal* | 0 | 0.81 | 0 | 0 | 0 | 0 |
| *Low Grade* | 4.27 | 0 | 0 | 0 | 0 | 0 |
| *High Grade* | 1.31 | 0 | 0 | 0 | 0 | 0 |
| **Covr.= 99%** | | | | | | |
| *Normal* | 0 | 1.52 | 0 | 0 | 0 | 0 |
| *Low Grade* | 2.16 | 0.71 | 0 | 0 | 0 | 0 |
| *High Grade* | 0.58 | 0 | 0 | 0 | 0 | 0 |

Table 11: Comparision of average class coverage gap (CovGap) for cross-entropy and hinge loss used for CP methods across various coverages for CRIC dataset w.r.t Risk-Based Controlled Set Sizes.

### D.3 Avoiding Confusing Classes

In Table 10, we track the *Confusing Set %* for various CP methods across 3 coverage guarantees, with the underlying model trained using cross-entropy and hinge loss.

Table 10 demonstrates that the *Confusing Set %* for Hinge Loss is consistently lower than that for CE Loss across all CP methods and coverage settings. The difference of *Overlapping Set %* ranges from 9% to 54% across various CP methods and coverage settings. This reduction in *Confusing Set %* is achieved by adjusting the margin control property of Hinge loss for all groups as detailed below:

For a true class belonging to a confused class pair, we apply a high margin to its corresponding confused class and low margins for all other classes. This approach reduces the probability mass assigned to the corresponding confusing class, thereby lowering the likelihood of its inclusion in the prediction set. For example, if the true class is Fibroadenoma, we assign a high margin to its corresponding confused class i.e., Phyllodes Tumour and low margins for all other classes. This approach reduces the likelihood of including Phyllodes Tumour in the prediction set.

This effective usage of the Hinge Loss margins results in a reduced *Overlapping Set %* compared to CE loss.

## E Loss Function and Coverage Tradeoff

This section provides additional results regarding the effect of training with different loss functions on the empirical coverage of the prediction sets for all three experiments across both datasets. For $I_{\text{test}}$ be a test set and $|I_{\text{test}}|$ size of test set, then the empirical coverage $\hat{c}$ is defined in Eq. 9:

$$\hat{c} = \frac{1}{|I_{\text{test}}|} \sum_{i \in I_{\text{test}}} \delta[y_i \in C(x_i)] \tag{9}$$

| | LAC | | APS | | RAPS | |
|---|---|---|---|---|---|---|
| | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE |
| **Covr.= 90%** | | | | | | |
| *Benign* | 0.71 | 1.87 | 0 | 0 | 0 | 0 |
| *Malignant* | 1.22 | 0.44 | 0 | 0 | 0 | 0 |
| **Covr.= 95%** | | | | | | |
| *Benign* | 0.26 | 1.96 | 0 | 0 | 0 | 0 |
| *Malignant* | 0.91 | 0 | 0 | 0 | 0 | 0 |
| **Covr.= 99%** | | | | | | |
| *Benign* | 0.30 | 1.8 | 0 | 0 | 0 | 0 |
| *Malignant* | 1.9 | 0 | 0 | 0 | 0 | 0 |

Table 12: Comparision of average class coverage gap (CovGap) for cross-entropy and hinge loss used for CP methods across various coverages for BreakHist dataset w.r.t Risk-Based Controlled Set Sizes.

| | LAC | | APS | | RAPS | |
|---|---|---|---|---|---|---|
| | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE |
| **Coverage** | | | | | | |
| *90%* | 0 | 0.71 | 0 | 0 | 0 | 0 |
| *95%* | 0 | 0.41 | 0 | 0 | 0 | 0 |
| *99%* | 0.50 | 0.63 | 0 | 0 | 0 | 0 |

Table 13: Comparison of average class coverage gap (CovGap) for cross-entropy and hinge loss used for CP methods across various coverages for CRIC dataset w.r.t Avoiding Classes from various classes.

where $\delta$ denotes an indicator function that is 1 when its argument is true and 0 otherwise. We examine the empirical coverage tradeoff between the loss functions by utilizing the average class coverage gap (CovGap) metric, which is outlined in Eq. 10:

$$\text{CovGap} = 100 \times \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \max\left(0, (1-\alpha) - \hat{c}_k\right) \tag{10}$$

where $\hat{c}_k$ is the empirical class-conditional coverage of class $k$.

### E.1 Risk-Based Controlled Set Sizes

In Table 11 & Table 12 , we track the average class coverage gap (CovGap) for various CP methods across 3 coverage guarantees, with the underlying model trained using cross-entropy and hinge loss on the CRIC and BreakHist datasets w.r.t to the first experiment.

Table 11 & Table 12 demonstrates that the average class coverage gap (CovGap) for Hinge loss and Cross Entropy consistently achieves empirical coverage evident from across all risk categories for APS and RAPS on CRIC and BreakHist dataset. In the case of LAC, Cross-Entropy Loss results in a random distribution of CovGap, whereas Hinge Loss leads to an increased CovGap which is consistent with the desired order of set sizes according to associated risk levels.

### E.2 Avoiding Classes from Various Risk Levels

In Table 13 & Table 14, we track the average class coverage gap (CovGap) for various CP methods across 3 coverage guarantees, with the underlying model trained using cross-entropy and hinge loss on both datasets w.r.t to the second experiment.

Table 13 & Table 14 demonstrates that the average class coverage gap (CovGap) for Hinge loss is consistently higher than that for CE Loss in all CP methods and coverage settings across both datasets. This trend can be attributed to the slight decrease in AUC (76.14 for CRIC dataset & 88.42 for BreakHist) and accuracy (70.01 for CRIC dataset & 82.64 for BreakHist) when the DieT model when trained to explicitly avoid classes based on varying risk levels, compared to the baseline.

| | LAC | | APS | | RAPS | |
|---|---|---|---|---|---|---|
| | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE |
| **Coverage** | | | | | | |
| *90%* | 0.07 | 0.60 | 0 | 0 | 0 | 0 |
| *95%* | 0.18 | 0.73 | 0 | 0 | 0 | 0 |
| *99%* | 0.29 | 0.64 | 0 | 0 | 0 | 0 |

Table 14: Comparison of average class coverage gap (CovGap) for cross-entropy and hinge loss used for CP methods across various coverages for BreakHist dataset w.r.t Avoiding Classes from various classes.

| | LAC | | APS | | RAPS | |
|---|---|---|---|---|---|---|
| | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE |
| **Coverage** | | | | | | |
| *90%* | 0 | 0.26 | 0 | 0 | 0 | 0 |
| *95%* | 0 | 0.1 | 0 | 0 | 0 | 0 |
| *99%* | 0.55 | 0.84 | 0 | 0 | 0 | 0 |

Table 15: Comparison of average class coverage gap (CovGap) for cross-entropy and hinge loss used for CP methods across various coverages for CRIC dataset w.r.t Avoiding Confusing Classes.

| | LAC | | APS | | RAPS | |
|---|---|---|---|---|---|---|
| | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE | CROSS ENTROPY | HINGE |
| **Coverage** | | | | | | |
| *90%* | 0.22 | 0.94 | 0 | 0 | 0 | 0 |
| *95%* | 0.01 | 0.07 | 0 | 0 | 0 | 0 |
| *99%* | 0.41 | 0.93 | 0 | 0 | 0 | 0 |

Table 16: Comparison of average class coverage gap (CovGap) for cross-entropy and hinge loss used for CP methods across various coverages for BreakHist dataset w.r.t Avoiding Confusing Classes.

### E.3 Avoiding Confusing Classes

In Table 15 & Table 16, we track the average class coverage gap (CovGap) for various CP methods across 3 coverage guarantees, with the underlying model trained using cross-entropy and hinge loss on both datasets w.r.t to the second experiment.

Table 15 & Table 16 demonstrates that the average class coverage gap (CovGap) for Hinge loss is consistently higher than that for CE Loss in all CP methods and coverage settings across both datasets. Similar to the above experiment,this trend can also be associated to the slight decrease in AUC (76.34 on CRIC dataset & 87.38 on BreakHist) and accuracy (70.02 on CRIC dataset & 82.12 on BreakHist) when the DieT model when trained to explicitly avoid confusing pairs of classes, compared to the baseline.