Data Augmentation and Conformal Prediction

by

Helen Lu

S.B. in Computer Science and Engineering
Massachusetts Institute of Technology (2022)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Authored by:	Helen Lu Department of Electrical Engineering and Computer Science May 12, 2023
Certified by:	John Guttag Professor Thesis Supervisor
Certified by:	Divya Shanmugam Doctoral Candidate Thesis Supervisor
Accepted by:	Katrina LaCurts Chair, Master of Engineering Thesis Committee

Data Augmentation and Conformal Prediction

by

Helen Lu

Submitted to the Department of Electrical Engineering and Computer Science
on May 12, 2023, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Conformal prediction is a popular line of research in uncertainty quantification. Conformal predictors output sets of predictions accompanied by a guarantee that the set contains the true label. Conformal prediction is particularly promising because it makes no distributional assumptions and requires only a black-box classifier to produce sets with this type of guarantee. Unfortunately, existing conformal predictions can produce uninformatively large prediction sets for certain examples, which limits their applications to real-world contexts. In this thesis, we explore the impact of data augmentation, a popular computer vision technique, on the performance of conformal predictors. In particular, we present multiple ways of combining data augmentation with conformal prediction by introducing five methods of test-time-augmentation-enhanced conformal prediction (TTA-CP). We find that certain TTA-CP methods can improve upon the size and stability of prediction sets created by traditional conformal prediction. Using ImageNet and Fitzpatrick 17k, two datasets differing in size, complexity, and balance, we reveal dataset-dependent decisions that are key to improving performance in conformal prediction.

Thesis Supervisor: John Guttag
Title: Professor

Thesis Supervisor: Divya Shanmugam
Title: Doctoral Candidate

Acknowledgments

First, I would like to thank John for all of his support and mentorship over the past two years. Under his patient guidance, I have grown as a researcher, thinker, and communicator. His dedication to helping his students reach their full potential shines through, and I could not have asked for a more rewarding M.Eng experience.

I would also like to thank Divya for taking a chance on me as a UROP in 2021. She has been an incredible mentor and role model through and through, patiently teaching me how to become a better researcher, writer, and person. Her genuine passion, intellectual curiosity, and superhuman-like abilities to balance a million things with the utmost calm and humility have been a source of inspiration over the past two years. I am tremendously grateful to have had the opportunity to learn from her.

I'd also like to thank my labmates - Harini, Aniruddh, Jose, Katie L., Emily, Hallee, Katie M. Marianne, Andrew, and Victor for welcoming me into the lab. Thank you for all the laughter, baked goods, stimulating conversations, running recommendations, and wonderful memories.

This year would not have been possible without my roommates and friends by my side. Thank you to Meagan for letting me distract you from your own thesis with late night chats and being my Hayden buddy; to Emelie for letting me crash her morning routine and inspiring me with positivity in even the most stressful of times; and to Jess for supplying humorous content for us to laugh ourselves silly to and showing us the light of gluten-free baked goods. Thank you to Caroline, Raina, Albert, Madison, Claire, and Meghana for being my family-away-from-home. And of course, a special shoutout to Pod.

Finally, I'd like to thank my family–Mom, Dad, and Katie for your unwavering love and support. Thank you for picking up my phone calls and FaceTimes, in times of panic and in times of celebration.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The deployment of machine learning models in high-stakes applications (including healthcare, finance, and sustainability) requires faithful estimates of a model's confidence in a particular prediction. Research on uncertainty quantification in machine learning focuses on this problem and aims to develop models that "know what they do not know." A naive approach is to train a probabilistic classifier and use the outputted probability. For instance, consider a dermatology classifier meant to distinguish between skin images of atopic dermatitis (i.e., eczema) and images of other skin conditions. If the classifier outputs a probability of 0.8 that atopic dermatitis is depicted in an image, one is tempted to interpret this as the model being 80% confident that the image is of atopic dermatitis. Unfortunately, neural networks are known, however, to produce overconfident predictions that cannot be interpreted as true probabilities [15] [34] [29]. Such classifiers also lack comprehensive statistical guarantees on prediction accuracy [38]. Without guarantees, it is difficult to ensure the reliability of a classifier's predictions in important applications.

Conformal prediction (CP) is an emerging area of research that provides these types of guarantees about the correctness of a model's output. CP methods do so by reframing the problem: instead of returning a single prediction, conformal prediction algorithms output a set of predicted classes, accompanied by a probabilistic guarantee for how often the set contains the true class. This set is termed the **prediction set**, and this probabilistic guarantee is called the **coverage guarantee**. The cover-

Figure 1-1: **Example of a prediction set created by the APS conformal predictor with a coverage guarantee of 90% (i.e., $\alpha = 0.1$) on an image labeled "saxophone" from the ImageNet validation set.** Conformal predictor did not use randomization nor allowed empty prediction sets. 1000 images were used at calibration

['accordion', 'bassoon', 'cornet', 'flute', 'French horn', 'microphone', 'oboe', 'saxophone', 'stage', 'trombone']

age guarantee is defined as $1 - \alpha$, where α is a user-provided acceptable error rate. Figure 1-1 depicts an example of a prediction set. To output these prediction sets, the conformal prediction pipeline involves two stages: calibration and inference. At calibration, a threshold is determined based on the classifier's predicted probabilities on each sample. This threshold is then applied at inference to determine which classes will be included in a sample's prediction set. Large prediction sets can be interpreted as a form of uncertainty; when the model is uncertain, the prediction set contains many classes in order to adhere to the pre-specified guarantee.

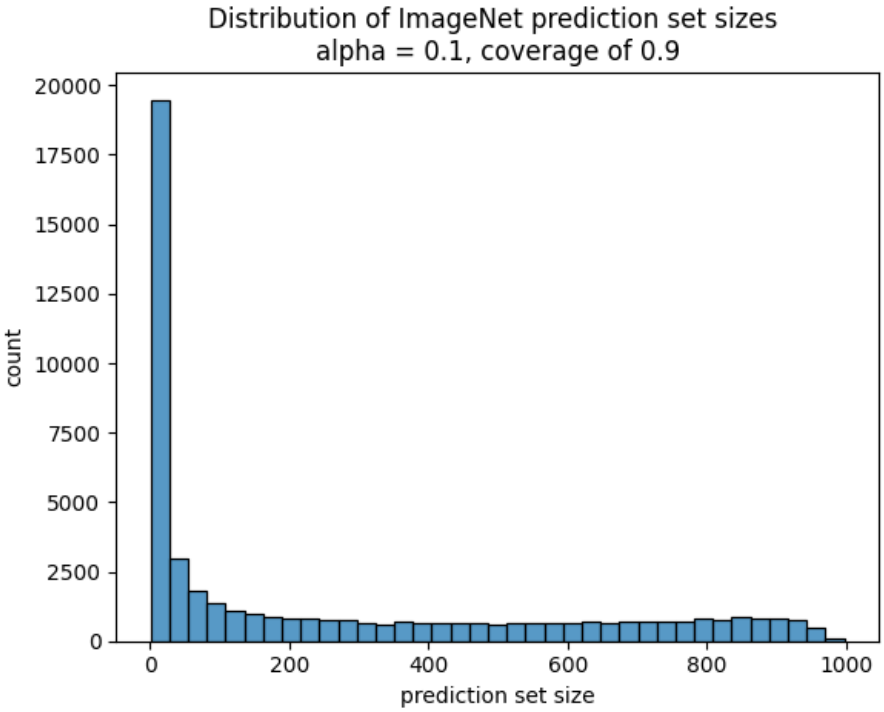
How might this be used in practice? Consider, for instance, clinical diagnosis. When predicting the true diagnosis on an unseen patient presenting with chest pain, a traditional classification model might yield a diagnosis of heartburn with 90% prob-

ability. More useful, however, is a set of predictions that rules in other potential diagnoses, such as cardiac arrest or pulmonary embolism, for a physician to consider. Knowing that the true diagnosis is contained in the set with 90% probability, a physician could use this set of predictions to order additional diagnostic tests–such as an upper gastrointestinal endoscopy, an electrocardiogram, and a chest X-ray in this example–to perform a differential diagnosis and confirm the true diagnosis. CP allows us to realize the second scenario. By creating prediction sets tailored to a user-defined confidence level, CP provides guarantees not found typically in other machine learning classifiers. CP is also more robust than other uncertainty quantification methods, which also make assumptions to achieve guarantees. CP is model-agnostic, making no assumptions about the structure or training procedure of the underlying classifier. CP is also distribution-free. It does not require knowledge about the probability distribution of the data and assumes only that the train, calibration, and test sets are independent and identically distributed.

There is a catch: CP methods achieve high coverage guarantees at the cost of larger prediction set sizes. This presents a significant challenge in practice because a prediction set with many classes is less informative than a prediction set with few classes. We can always achieve high coverage by naively outputting prediction sets that contain all possible classes. Although the true class is guaranteed to be in the prediction set, this outcome is not useful because it does not provide any information about what the true class could be. Existing CP approaches today, while more advanced than this naive approach, run into a similar issue. Figure 1-2 displays the distribution of prediction set size when running the adaptive prediction sets (APS) CP algorithm developed by Angelopoulos et. al. [1] on the ImageNet validation set. Achieving a coverage guarantee of 90% results in 48.59% of prediction sets containing 100 or more of the 1000 possible classes. While variants of APS exist, we explore the interaction between data augmentation and the basic principles of conformal prediction. Further work can explore how data augmentation interacts with other modifications (e.g., regularization [1]).

While current CP approaches are able to achieve high coverage guarantees, they

Figure 1-2: **Prediction sets created with APS conformal prediction on ImageNet validation set are large, with a mean size of 214.04 and a median size of 41.0 when** $\alpha = 0.1$**.** ImageNet validation set contains 1000 unique classes. Prediction sets created without randomization and without empty prediction sets. 1000 samples used for calibration.



often do so at the cost of prediction sets that are large and thus lacking in meaningful insight into the true class. Smaller prediction sets provide more information about the true class, but are challenging to attain while meeting the coverage guarantee.

The difficulty of achieving smaller prediction sets is, in part, due to the limited availability of data across the conformal prediction pipeline, during calibration and inference. **In this work, we explore how data augmentation, a common and successful technique in machine learning for generating additional data points, can be used in conformal prediction to achieve smaller prediction sets and thus improve the real-world applicability of conformal prediction. We examine the relative benefits of data augmentation at different stages of the conformal prediction pipeline.** Our main contributions are:

- We present five test-time-augmentation-enhanced-conformal prediction (TTA-CP) methods and apply them to two datasets differing in size, distribution, and difficulty: ImageNet [**?**] and Fitzpatrick 17k-8 (a modification of the Fitzpatrick 17k dataset [12]). We show that the stage (calibration or prediction set creation), augmentation policy, and aggregation function (when augmenting at prediction set creation) affect achieved coverage. **We find that, conditional on prediction set size,**

- We study how randomization and allowing empty prediction sets affected achieved coverage on Fitzpatrick 17k-8. We explain why randomization should not be used when applying a conformal predictor to a poor-performing classifier, and why empty prediction sets should not be allowed.

Chapter 2 reviews related work on conformal prediction and data augmentation. In Chapter 3, we present our methods of applying data augmentation to conformal prediction and outline our experimental approach to understanding the effects of these methods. Chapter 4 presents our findings on the overall effect of data augmentation in reducing the size of generated prediction sets. Chapter 5 examines these effects at a class-specific level to determine the utility of data augmentation in conformal prediction on underrepresented classes. In Chapter 6, we pivot to understanding how data augmentation can improve efficiency of calibration. We summarize our findings and present areas for further exploration in Chapter 7.

Chapter 2

Related Works

2.1 Conformal Prediction

Vovk et al. [42, 41] created the framework of conformal prediction (at the time, called "transductive confidence machine (TCM)"), in which outputted prediction sets were accompanied by a probabilistic guarantee (i.e., coverage guarantee) that the outputted set contains the true label. This framework required only that data used at train and test time be independent and identically distributed. They showed empirically that this framework maintained the coverage guarantee in an online setting, using each sample of the test set as part of the learning process to improve calibration. By design, TCM is "statistically efficient". It can achieve the coverage guarantee with a limited amount of labeled training data and an unlabeled test dataset. This statistical efficiency, however, comes at the cost of computational inefficiency. TCM achieves coverage guarantee by calibrating on the labeled training data and then performing a nearest neighbor search for each sample in the test set. The iterations of nearest neighbor search are computationally inefficient, especially with large training sets. Today, TCM is known as full conformal prediction or transductive conformal prediction.

To address the computational inefficiency of full conformal prediction, Papadopoulos et al. [23] introduced split conformal prediction. Split conformal prediction operates in an offline setting but relies on the same assumptions as full conformal pre-

diction (i.e., independent and identically distributed train and test sets) and outputs prediction sets that meet the coverage guarantee. Split conformal prediction is more efficient, however, because it performs calibration on a smaller subset of the train data (i.e., calibration set). This splitting allows the remaining train data to be used to train a classifier that can generate predictions on the test set, eliminating the need to run nearest neighbor search for each sample in the test set. Papadoupolos et al. showed empirically that, at the cost of slightly larger prediction set sizes, split conformal prediction maintains the coverage guarantee while being more computationally efficient than full conformal prediction [23]. For this reason, split conformal prediction is much more suited for practical deployment. Today, most research in conformal prediction uses some version of split conformal prediction [1, 2, 11, 27, 28, 7, 40].

While we require computational efficiency for conformal predictors to be fit for practical deployment, we also require informative prediction set outputs to consider these predictors useful. One metric for how "informative" a prediction set is prediction set size (i.e. a prediction set with 3 classes is more informatiove than a prediction set with 10 classes). Angelopoulos et al. [1, 2] propose two split conformal predictors, Adaptive Prediction Sets (APS) and Regularized Adaptive Prediction Sets (RAPS), that use prediction set size as a proxy for classification difficulty. Larger prediction sets outputted from APS and RAPS imply a more challenging sample to classify, whereas smaller predictive sets imply an easier sample. In our work, we use APS and describe notation and implementation in more detail in 3.1.

One line of research in conformal prediction is centered around the coverage guarantee. The conformal prediction approaches we covered earlier guarantee coverage over the entire dataset. There may exist contexts, however, in which we want to achieve coverage guarantee for each class. Derhacobian et al. [11] introduced a conformal predictor, CCAPS, that achieves class-conditional coverage guarantee. Another modification to the original coverage guarantee focuses on conditional coverage. Angelopoulos et al. [1, 2] introduce the concept of conditional coverage, where coverage is guaranteed for every input. This definition of coverage is much stronger than the marginal coverage guarantee of conformal prediction and often impossible to achieve

[?]. Romano et al. [27, 28], Cauchois et al. [7], and Tibshirani et al. [40] present conformal predictors that attempt to approximately meet conditional coverage guarantee.

Conformal prediction is not limited to classification. Recent work has explored conformal methods in quantile regression [27] and outlier detection [6, 19, 14, 13]. Angelopoulos et al. also present adaptations of conformal prediction to achieve guarantees on risk control metrics other than accuracy, such as false discovery rate [4, 3]. In our work, we focus on conformal prediction in classification.

2.2 Data Augmentation

Data augmentation–the expanding of a dataset by generating new transformed data points from existing ones–is a technique commonly used in image classification. Traditionally used at training time, data augmentation can prevent overfitting and improve generalization [25, 31, 44]. Recent work has also begun exploring data augmentation at test-time. By aggregating predictions across transformed versions of a test input, test-time augmentation (TTA) has been shown to improve model robustness [26, 35, 10], accuracy [18, 37, 32, 22, 30], and uncertainty estimates [22, 33, 5, 43]. TTA has become popular because these benefits can be realized using off-the-shelf libraries [9, 24], without requiring resource-intensive retraining or additional data. TTA in the context of conformal prediction, however, has yet to be considered rigorously. In our work, we investigate if data augmentation may confer similar benefits for conformal predictors as we've seen with other image classifiers.

Current questions of interest in data augmentation research center around design. What is the best way to aggregate predictions generated from augmented and original test samples? How should we decide which augmentations should be used? Further, data augmentation Advances in these directions of research are complementary to the work I present here and can be readily combined with our work.

Chapter 3

Experimental Setup

In this chapter, we describe our test-time augmented conformal prediction methods and the experiments we ran to evaluate their performance. We outline the datasets, models, and augmentations that we use in our work.

3.1 Conformal Prediction Algorithms and Notation

In our work, we focus on the adaptive prediction sets (APS) conformal predictor presented by Angelopoulos et al. [1, 2]. We considered using the regularized adaptive prediction sets (RAPS) algorithm introduced by Angelopoulos et al., which produced smaller prediction sets than APS on ImageNet without sacrificing coverage [1]. However, we chose to use APS for two reasons: simplicity and widespread adoption. APS requires minimal finetuning, with only one hyperparameter: the size of the calibration set. RAPS, on the other hand, requires finetuning two additional hyperparameters in order to achieve smaller prediction sets than APS [1]. Moreover, other conformal predictors, including RAPS [1], are built on the APS algorithm [11, 45]. These predictors would also benefit from improvements to the performance of APS.

We borrow terminology and notation from Angelopoulos et al. [1, 2]. Formally, there are three inputs to the APS algorithm:

- A pre-trained black-box classifier $f : X \to [0, 1]^K$ that maps inputs to a vector of class probabilities, where X is the input domain and K is the number of classes.

(Note: while APS can be applied to regression problems with continuous output, we focus on discrete classification.)

- A user-specified error rate $\alpha \in [0, 1]$ that defines the coverage guarantee, which ensures that the correct class exists in the outputted prediction set with probability $1 - \alpha$.

- A calibration set C of n_C held-out inputs $X_C = x_1, \ldots, x_n$ and their associated labels $Y_C = y_1, \ldots, y_n$, where $y_i \in 1, ..., K$. We assume this set is independent of and identically distributed with respect to data drawn from the the train and test domain.

The APS procedure improves upon conformal prediction. In addition to maintaining coverage guarantee, APS enforces that prediction set sizes reveal information about instance-wise model uncertainty. The adaptiveness of prediction sets is not explicitly guaranteed with other conformal predictors. To achieve adaptive prediction sets, APS operates in the following manner. First, the classifier is used to obtain class probabilities across all C classes for each sample in the calibration set. Next, a score function is defined and applied to each sample of the calibration set. The score function can be any function that ranks inputs from lowest to highest magnitude of model error (i.e., larger scores encode worse agreement between an input x and a label y). In our work, we use the score function $s(x, y) = \sum_{j=1}^{k} \hat{f}(x)_{\pi_j(x)}$, where $y = \pi_k(x)$ and $pi_k(x)$ is the permutation of $1, ..., K$ that sorts the class probabilities in descending order. In other words, for each sample, the score is the minimum running sum of class probabilities sorted in descending order that includes the true label of the sample. To concretize our explanation, we will use a running example in which we apply a classifier to an input, and it outputs the probability that the input is in one of three classes (i.e., $K = 3$). Imagine our input x_i has a true label $y_i = 3$ and the classifier outputs probabilities $[0.25, 0.4, 0.35]$ corresponding to classes $[1, 2, 3]$. Sorted in descending order, the class probabilities are $[0.4, 0.35, 0.25]$ and correspond to permutation $\pi_k(x_i) = [2, 3, 1]$. We sum the class probabilities up to and including the true class 3 to obtain the score $s(x_i, y_i) = 0.75$.

Once the score of each calibration sample is calculated, we find the $\frac{(n_C+1)(1-\alpha)}{n_C}$ quantile q. Here, q is roughly the $(1-\alpha)$ quantile of the scores (with a small adjustment of $\frac{(n_C+1)}{n_C}$ to account for the sample size of the calibration set). In other words, α of the scores lie above q.

The quantile q is then used to generate prediction sets at test-time. Given a test set T of n_T held-out inputs X_T and associated labels $Y_T = y_1, \ldots, y_t$ where $y_i \in 1, ..., K$, we apply f to X_T and obtain class probabilities for each sample. For each sample, we again sort the class probabilities in descending order and calculate the running sum. Each class whose probability is included in the greatest running sum less than quantile q is included as part of the output prediction set for the test sample. Returning to our earlier example of a three-class setting, imagine we determined q to be 0.85 at calibration and we have a test sample x_j for which classifier f returns class probabilities of $[0.25, 0.4, 0.35]$ for classes 1, 2, and 3 respectively. Sorted in descending order, the class probabilities $[0.4, 0.35, 0.25]$ correspond to a running sum vector of $[0.4, 0.75, 1.0]$ and a $\pi_k(x_i) = [2, 3, 1]$. We find the greatest running sum less than or equal to quantile q to be 0.75, and thus include only classes 2 and 3 in our prediction set for sample x_j. Following this procedure, the prediction sets created by APS achieve an average coverage that maintains the promised coverage guarantee while communicating instance-wide uncertainty via size.

In our work, we use a variation of the APS procedure: randomized APS. Randomized APS achieves smaller prediction sets than those generated by the APS procedure by adjusting calibration sample scores with randomization. For each sample, we begin by sorting the class probabilities and calculating the running sum. If we were to follow the APS calibration procedure without randomization, the score for a calibration sample is the minimum running sum that includes the true label of the sample. With randomization, however, we set the score to be a random value between a and b. a is the maximum running sum excluding the true class (or 0, if the model correctly classified the calibration sample) and b is the minimum running sum including the true class. Recalling our three-class setting at calibration, classifier f outputted class probabilities $[0.25, 0.4, 0.35]$ corresponding to classes $[1, 2, 3]$ for our calibration

sample x_i belonging to class 3. Sorted in descending order, the class probabilities $[0.4, 0.35, 0.25]$ corresponded to permutation $\pi_k(x_i) = [2, 3, 1]$. As in APS, we find the cumulative sum vector of $[0.4, 0.75, 1.0]$. Here, randomized APS deviates from APS. Whereas APS defined the score $s(x_i, y_i)$ to be 0.75 (the minimum cumulative sum of sorted class probabilities that includes the true class), randomized APS defines the score to be a random value between 0.4 (the maximum cumulative sum of sorted class probabilities up until the true class) and 0.75. Empirically, research has found that introducing randomization in this way on ImageNet produces dramatically smaller prediction sets while maintaining the coverage guarantee [1, 28].

APS and Randomized APS include a flag for controlling whether or not the algorithm is allowed to create empty prediction sets. We evaluate the effect of allowing and not allowing empty prediction sets in Section 4.1

3.2 Test-Time Data Augmentation

In the following experiments, we investigate the effects of data augmentation at each stage of the (randomized) APS conformal prediction pipeline: calibration and prediction set formation. We denote data augmentation at calibration as a form of test-time augmentation because the augmentation occurs after a model has been trained and made its predictions. We add two inputs to the conformal prediction pipeline:

- Augmentation policy A which consists of M augmentation functions $a_m : X \to X$, and always includes an identity transform. As a result, augmentation policy $A(x_i)$ maps sample x_i to a set of inputs: the original image and $a_m(x_i)$ for each of the remaining $M - 1$ augmentations.

- Aggregation function g which maps a set of prediction sets to a single prediction set.

We consider five plausible test-time-augmentation-enhanced conformal prediction (TTA-CP) methods. We investigate the impact of data augmentation at each stage alone, and then in combination.

- **TTA-Cal**: In TTA-Cal, we augment only the calibration set C. We apply augmentation policy A, consisting of M augmentation functions, to each of the n_C samples in C, obtaining a new set C' of Mn samples. We use C' as the new calibration set for determining quantile threshold q in the conformal prediction pipeline. The test set remains unaugmented.

- **TTA-Set-Intersection**: In TTA-Set-Intersection, we augment only the test set T. We used the unaugmented calibration set C to determine the quantile threshold q. We apply augmentation policy A, consisting of M augmentation functions, to each of the n_T samples in T, such that each sample in X_T is now mapped to M augmented samples. An intermediate prediction set is created for each of the Mt samples in T'. For TTA-Set-Intersection, we define the aggregation function g to be the set intersection of the M intermediate prediction sets created for each original test sample, and apply g to obtain our final n_T prediction sets.

- **TTA-Set-Majority**: In TTA-Set-Majority, we again augment only the test set T. We follow the same procedure as outlined in TTA-Set-Intersection, except that we define aggregation function g to contain only classes $1, \ldots, K$ that appear in more than half of the intermediate prediction sets.

- **TTA-Cal-Set-Intersection**: In TTA-Cal-Set-Intersection, both the calibration and test sets are augmented. We generate an augmented calibration set C' from applying augmentation policy A, to each sample in C and determine the quantile threshold q from C'. We then use q to form intermediate prediction sets on each of the samples in the augmented prediction set T', and form the final prediction sets by applying set intersection as defined by g.

- **TTA-Cal-Set-Majority**: In TTA-Cal-Set-Majority, both the calibration and test sets are augmented. We follow the same procedure as defined in TTA-Cal-Set-Intersection, except that we define g to aggregate the intermediate prediction sets by applying the set majority procedure outlined in TTA-Set-Majority.

The augmentation policies A we apply consist of a subset of augmentations seen at train-time. Of course, this augmentation policy is not the only policy that could be considered. We leave this exploration to future work, and mean our augmentation policy to represent the salient case of augmentation policies drawn from train-time augmentations (as outlined in Section 3.3.2).

We compare the performance achieved by these TTA-CP methods against those of the standard and randomized APS procedures, which we designate as **Vanilla-CP** and **Randomized-Vanilla-CP**.

3.3 Experimental Setup

We conduct three experiments to understand how test-time augmentation affects the performance of conformal prediction. First, we compare the five TTA-CP methods on achieved coverage and average prediction set size, the two primary metrics for conformal prediction performance. We describe these results in Chapter 4. In our second experiment, we disaggregate these results to be class-specific, analyzing how TTA-CP methods affect performance on underrepresented classes. The findings of this experiment are presented in Chapter 5. Finally, we narrow in on TTA-Cal in Chapter 6 and explore its potential in improving conformal prediction in contexts with limited labeled data available for calibration.

3.3.1 Datasets

We report our findings on two datasets: ImageNet and Fitzpatrick 17k. We chose these two datasets for their differences in size, class balance, and difficulty.

The ImageNet dataset, a widely-used dataset for image classification in computer vision, consists of over 1.2 million training images and 50,000 validation images across 1000 classes. Unlike the training set, the validation set is balanced, with 50 images per class. We reserve 50% of the validation set (25,000 images) for calibration, and generate prediction sets on the other half.

The Fitzpatrick 17k dataset is a dermatology dataset put forth by Groh. et al.

[12]. Each of the 16,577 clinical images is annotated for one of 114 dermatological conditions. These 114 labels are part of a hierarchy of skin conditions, which groups them into 9 medium-level labels and 3 high-level labels. Groh et. al. annotated each image with a score from the Fitzpatrick six-point scale [12], which is a widely-used proxy for skin tone. The dataset is imbalanced at the class-level and skin-tone-level. We chose to focus on the nine-class level of classification. With the three-class level of classification, prediction sets would contain at most 3 classes, limiting our study of the adaptivity and reduction of prediction set sizes. On the other hand, the 114-class level of classification has few examples per class and low accuracy with state-of-the-art models [12], limiting our ability to test for sensitivity to calibration set size in Chapter 6.

For the nine-way classification problem, the training data is skewed towards the "inflammatory" class, as shown in Figure 3-1. We use a common resampling technique of removing the majority class for improving classification accuracy among minority classes to mitigate against potential prediction bias [39, 17, 16, 8]. After removing images labeled "inflammatory" from the Fitzpatrick dataset, we reduce the task to an eight-way classification problem. We subsequently refer to this dataset as Fitzpatrick 17k-8. We reserved 10% of the dataset for testing, of which 50% was used for calibration and 50% for evaluating performance.

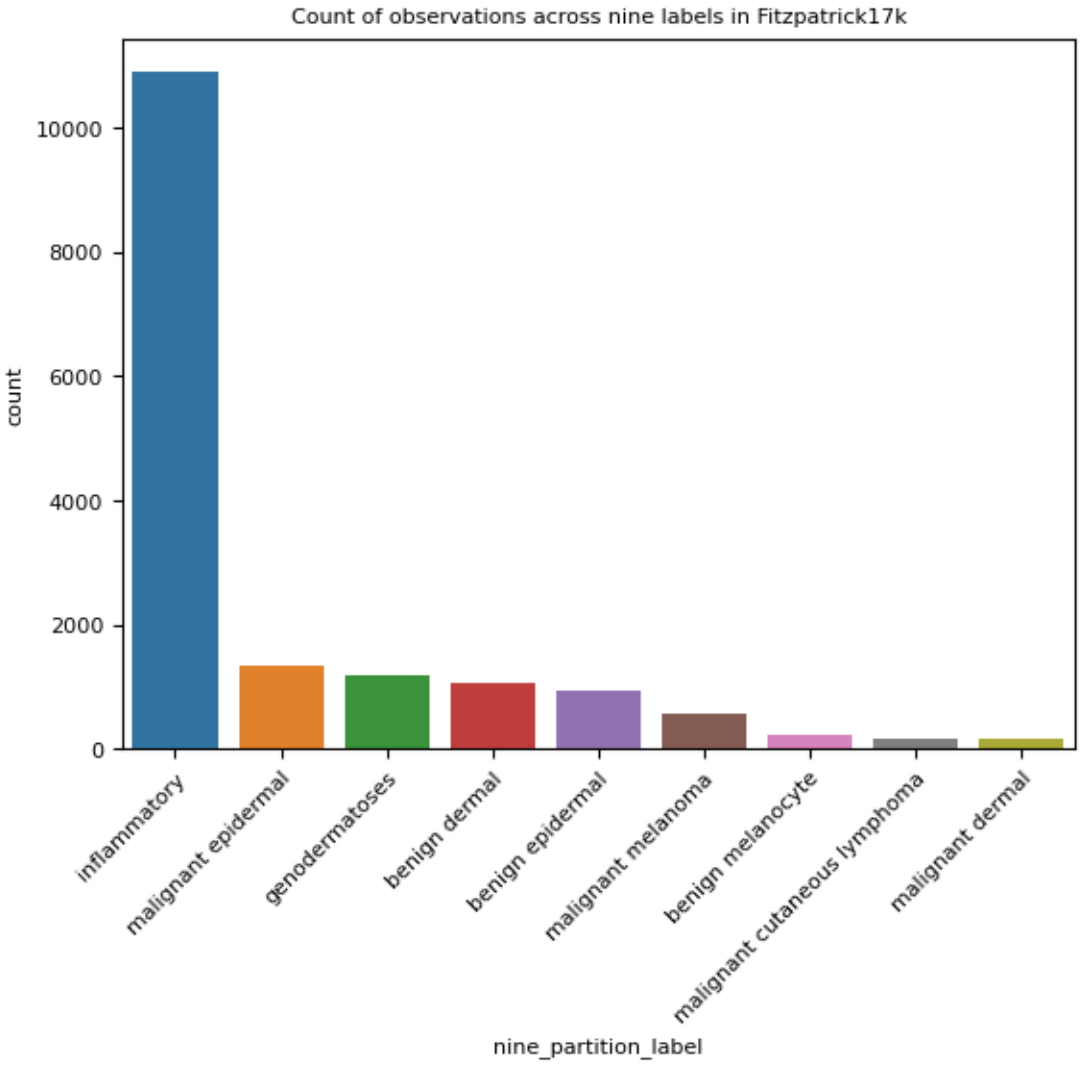
3.3.2 Model Architecture and Performance

We use a ResNet-152 classifier when predicting on the ImageNet dataset and a VGG-16 classifier when predicting on the Fitzpatrick 17k-8 dataset. We normalize images prior to training and inference, in line with past work.

In all ImageNet experiments, we use the PyTorch ResNet-152 classifier pretrained on ImageNet. This classifier applies random crop, random horizontal flip, and resize augmentations at training time, and achieves a top-1 classification accuracy of 82.28% on the validation set.

For Fitzpatrick 17k-8 experiments, we trained our own VGG-16 classifier to predict one of the eight classes. We used a 60/20/20 split of the dataset for training,

Figure 3-1: Distribution of observations across nine classes in the Fitzpatrick 17k dataset created by Groh. et al. [12]



validation, and testing, and applied five augmentations at training time: color jitter, random crop, random rotation, random horizontal flip, and center crop augmentations. The classifier achieves a top-1 accuracy of 56.06% on the training set and 48.06% on the test set.

We chose random crop and horizontal flip–the intersection of train-time augmentations used by both classifiers–in addition to the identity augmentation function to form our augmentation policy.

Chapter 4

Impact of TTA-CP on Coverage and Prediction Set Size

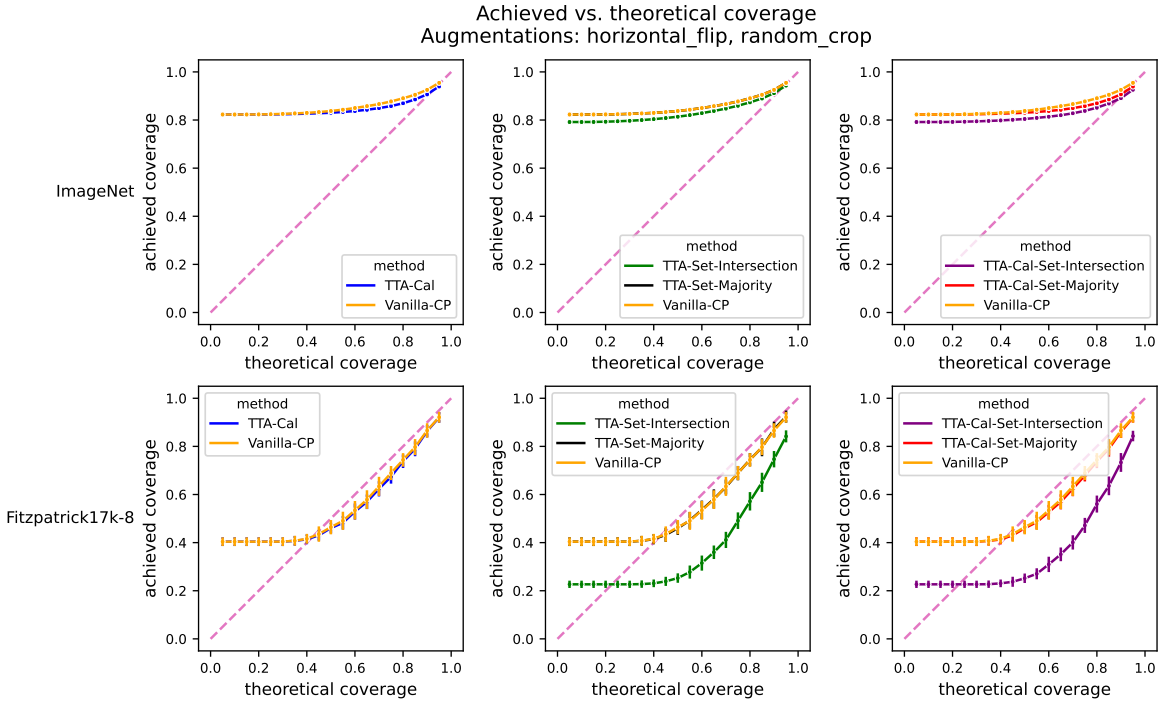
In this chapter, we present our findings on how TTA-CP methods compare to existing approaches in conformal prediction.

4.1 Achieved vs. Theoretical Coverage

We begin by examining the effect of test-time augmentation on the primary objective of conformal prediction: the probability that the true label appears in the outputted prediction set. It is important to note, however, that a trade-off can occur between maintaining the coverage guarantee and reducing prediction set size. We explore this tradeoff in Section 4.2.

In Figure 4-1, we compare the coverage achieved by the five TTA-CP methods proposed in Chapter 3 (TTA-Cal, TTA-Set-Intersection, TTA-Set-Majority, TTA-Cal-Set-Intersection, and TTA-Cal-Set-Majority) with the theoretical coverage guaranteed for both ImageNet and Fitzpatrick 17k-8. Augmentation at calibration time alone (TTA-Cal, in blue) achieves the same coverage as that of Vanilla-CP for all levels of α. Coverage achieved when augmenting at the prediction set creation phase, however, is more complex. TTA-Set-Majority (in black) matches the coverage achieved by Vanilla-CP. Meanwhile, TTA-Set-Intersection (in green) causes a drop in achieved

Figure 4-1: **Augmentation at calibration alone matches Vanilla-CP in achieved coverage. The performance of augmentation at prediction set creation (with and without augmentation at calibration) depends on the aggregation function used.** Achieved vs. theoretical coverage is compared across ImageNet and Fitzpatrick17k-8 datasets for all TTA-CP methods and Vanilla-CP. Achieved coverage results are averaged across 10 runs. Randomization is applied to all TTA-CP methods and to Vanilla-CP (in yellow). TTA-Cal (in blue), TTA-Set-Majority (in black), and TTA-Cal-Set-Majority (in red) trendlines lie under Vanilla-CP line for both datasets.



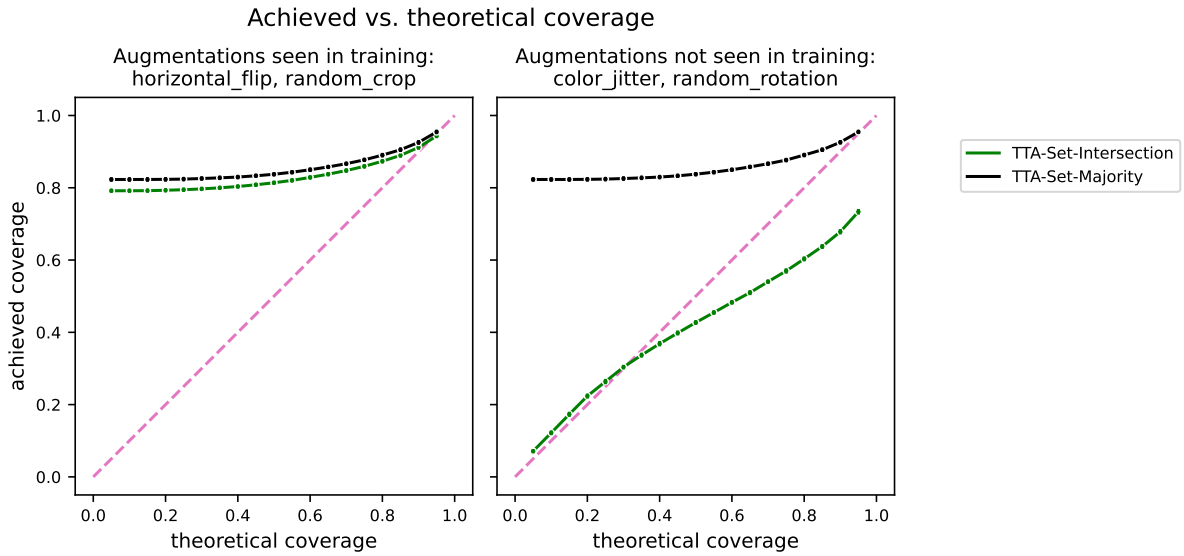
coverage from Vanilla-CP. These results indicate that the dominant factor determining achieved coverage of an augmented conformal predictor are the aggregation strategies used during prediction set creation. TTA-Set-Intersection uses a stricter aggregation function g than TTA-Set-Majority. For the true class to appear in the prediction set outputted by TTA-Set-Intersection, the class must appear in every prediction set outputted after applying the augmentation policy. TTA-Set-Intersection achieves lower coverage when there is disagreement among the prediction sets created for each augmented sample than when there is strong agreement.

The augmentation policy is one factor contributing to the degree of agreement across augmented prediction sets. As seen in Figure 4-2, TTA-Set-Intersection produces an achieved coverage that approximates that of Vanilla-CP and TTA-Set-Majority with augmentation policy of horizontal flip and random crop. These two augmentations are seen at train-time for ImageNet. When we switch the augmentation policy to instead include augmentations not seen in training for ImageNet (i.e. color jitter and random rotation), TTA-Set-Intersection produces a coverage much worse than that of TTA-Set-Majority. Intuitively, augmentations seen in training time will create more overlapping prediction sets than those not seen in training.

Applying test-time augmentation to conformal prediction at prediction set creation has a larger impact on coverage than at calibration. In Figure 4-1, TTA-Cal-Set-Majority (in red) achieves the same coverage as TTA-Set-Majority, and TTA-Cal-Set-Intersection (in purple) performs similarly to TTA-Set-Intersection. Introduction of augmentation during calibration time has only a marginal effect on improving achieved coverage. The negative effects of TTA-Set-Intersection, however, drive TTA-Cal-Set-Intersection to achieve lower coverage than Vanilla-CP.

The relative effects of TTA-CP are consistent across both ImageNet and Fitzpatrick 17k-8 (Figure 4-1). However, it might be surprising that with Fitzpatrick 17k-8, all methods–including Randomized-Vanilla-CP–fail to achieve the coverage guarantee at $\alpha <= 0.6$. We explore this behavior in the following section.

Figure 4-2: **TTA-Set-Intersection is more affected by augmentation policy that TTA-Set-Majority, even producing coverage below guaranteed levels when augmentation policy uses augmentations not see at train time.** Comparison of how achieved coverage (with respect to theoretical coverage) changes for TTA-Set-Intersection and TTA-Set-Majority between augmentations seen and not-seen in training. Coverage is reported for ImageNet prediction sets. Achieved coverage results are averaged across 10 runs. Randomization is applied to TTA-Set-Intersection and TTA-Set-Majority.



Randomization and Empty Prediction Sets in APS. Performing conformal prediction relies on many choices that can have a large impact on the achieved coverage and mean prediction set size of a conformal predictor. Two of these choices relate to the existence of empty sets (allowed vs. not allowed) and the choice of threshold (randomized vs. non-randomized). Using Fitzpatrick 17k-8, we analyze the decision to allow empty prediction sets and use randomization. We report results on Fitzpatrick 17k-8, since ImageNet prediction sets become prohibitively large without randomization.

Prior work in conformal prediction [1] allows empty prediction sets to ensure an exact coverage guarantee. However, in a machine learning classification setting, empty prediction sets artificially handicap conformal predictors when the coverage guarantee is less than the model accuracy. If a classifier, for example, has a top-1 accuracy of 95% but the desired theoretical coverage is 49% (i.e., $\alpha = 0.1$), empty prediction sets are needed to achieve an achieved coverage of exactly 90%, capping the potential achieved coverage. We empirically show this effect on achieved coverage in Figure 4-3, where prohibiting empty prediction sets improves achieved coverage for all levels of theoretical coverage on Fitzpatrick 17k-8. For this reason, it is almost always preferable to output sets of at least size one, including the class with the highest predicted probability in every outputted set. In all subsequent experiments, we prohibit the creation of empty prediction sets, ensuring that prediction sets.

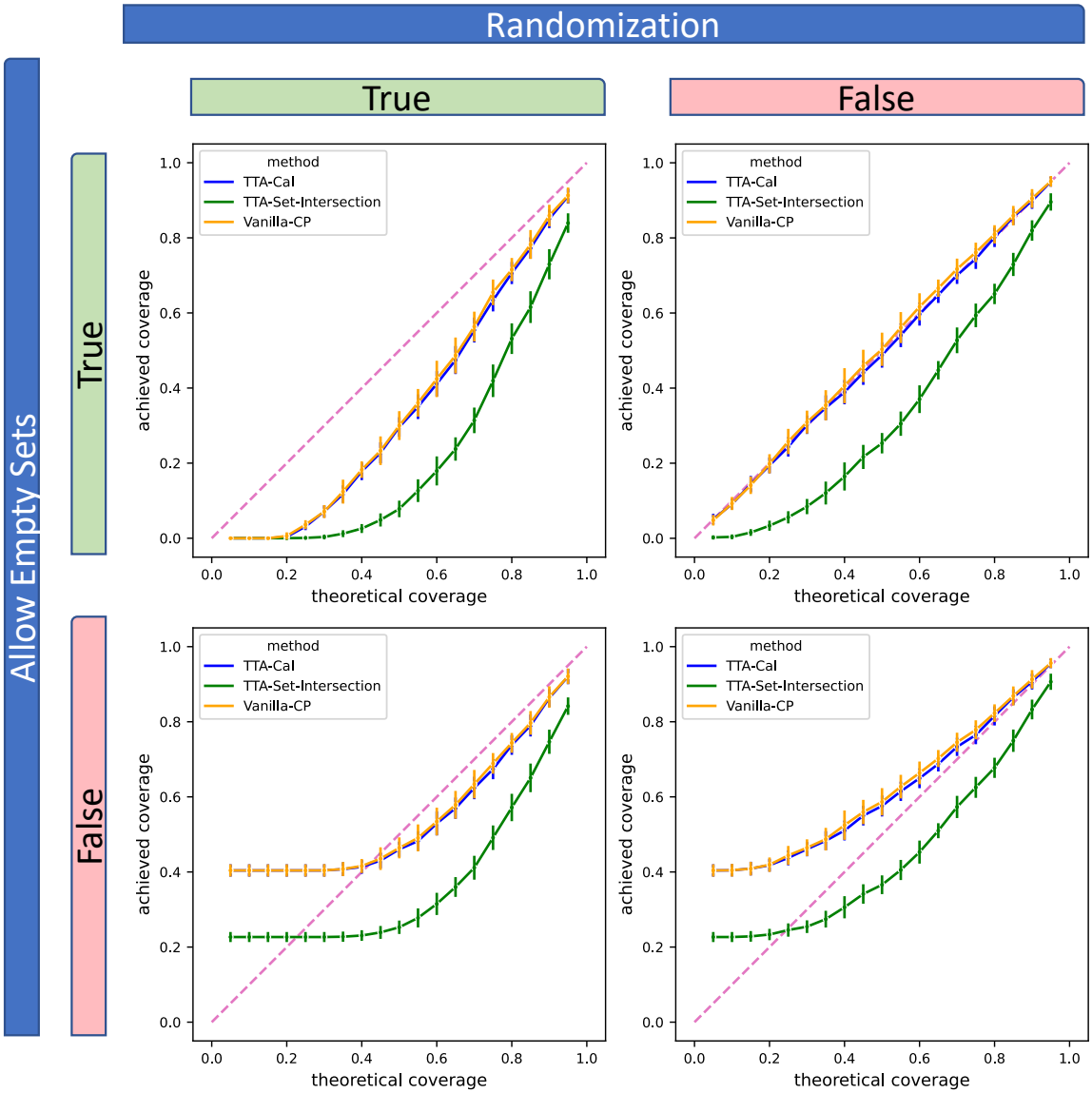
Angelopoulos et al. [1] use randomization to achieve, on average, smaller prediction sets with ImageNet while maintaining the coverage guarantee. As described in Section 3.1, randomization is the adjustment of calibration scores to be a random value between the running sum of sorted class probabilities before including the true label and the running sum of sorted class probabilities after including the true label. We find, however, that randomization breaks the coverage guarantee with Fitzpatrick 17k-8. As shown in Figure 4-3, randomization worsens achieved coverage for all levels of theoretical coverage. This drop off can largely be attributed to examples for which the classifier is incorrect. By definition, when a classifier correctly predicts the class for a test sample, the prediction/true class will always be included in the

prediction set. A drop off in coverage with randomization, therefore, must stem from the creation of prediction sets in which the classifier incorrectly predicts the class.

Let's illustrate this by revisiting the example presented in Section 3.1. We work in a three-class setting (i.e., $K = 3$) in which the classifier f is applied to a calibration sample x_i and outputs class probabilities $[0.25, 0.4, 0.35]$ corresponding to classes $[1, 2, 3]$. Sorted in descending order, the class probabilities are $[0.4, 0.35, 0.25]$ and the cumulative sum array is $[0.4, 0.75, 1.0]$, corresponding to permutation $\pi_k(x_i) = [2, 3, 1]$. Suppose the true class for sample x_i is $y_i = 2$. Here, the classifier correctly classifies sample x_i, and randomized APS will set the score $s(x_i, y_i)$ to be a random value between 0 and 0.4, the minimum cumulative sum of class probabilities in descending order that includes the true class. Now suppose the true class for sample x_i is not $y_i = 2$, but $y_i = 3$. In this case, classifier incorrectly classifies sample x_i. With randomized APS, the score $s(x_i, y_i)$ is defined to be a random value between 0.4, the maximum cumulative sum of descending class probabilities excluding the true label, and 0.75, the minimum cumulative sum of descending class probabilities including the true label. With a sufficiently-accurate classifier, this process of randomization would reduce scores of calibration samples, consequently lowering the quantile threshold and reducing the size of prediction sets at test time while maintaining coverage guarantee. Angelopoulos et al. [1] makes this observation with ImageNet. We do not, however, observe this phenomenon with Fitzpatrick 17k-8.

We hypothesize that the poor performance of the Fitzpatrick 17k-8 classifier contributes to the dropoff in coverage with randomization that we do not observe with ImageNet. The Fitzpatrick 17k-8 classifier (with top-1 accuracy of 48.06% and top-2 accuracy of 63.13% on the test set) is more frequently incorrect than the ImageNet classifier (with top-1 accuracy of 82.28% and top-2 accuracy of 91.03% on the validation set). We conclude that randomization is not wise with classifiers that achieve low accuracy, even if those classifiers perform significantly better than random classifiers. For this reason, we switch randomization off for all Fitzpatrick 17k-8 experiments. We keep randomization on for ImageNet, since prediction sets become prohibitively large without randomization.

Figure 4-3: **Using randomization and allowing empty prediction sets consistently worsen achieved coverage on Fitzpatrick 17k-8 for all methods.** Effects of applying randomization and accommodating empty prediction sets on achieved coverage for Vanilla-CP, TTA-Cal, and TTA-Set-Intersection methods shown for Fitpatrick 17k-8. Theoretical coverage is defined as $1 - \alpha$. Achieved coverage results are averaged across 10 runs. Augmentations used in TTA-Cal and TTA-Set-Intersection are horizontal flip and random crop.



4.2 Prediction Set Size vs. Achieved Coverage

Having examined how the proposed TTA-CP methods affect achieved coverage, we now study the tradeoff between achieved coverage and prediction set size. We compare this tradeoff among Vanilla-CP, TTA-Cal, and TTA-Set-Intersection in ImageNet (Figure 4-4) and Fitzpatrick 17k-8 (Figure 4-5).

For ImageNet, we observe that TTA-Cal (in blue) falls roughly along the Vanilla-CP trend (in yellow) (Figure 4-4). This indicates that the tradeoff in achieved coverage and prediction set size for TTA-Cal is the same as that of Vanilla-CP.

For TTA-Set-Intersection, the tradeoff in achieved coverage and prediction set size relative to that of Vanilla-CP depends on α. At lower α values (i.e. higher theoretical coverage), TTA-Set-Intersection (in green) lies along the Vanilla-CP trendline. At higher α values (i.e. lower theoretical coverage), however, TTA-Set-Intersection produces lower actual coverage than Vanilla-CP when conditioned on prediction set size.

For Fitzpatrick 17k-8 (Figure 4-5), TTA-Cal exhibits a different tradeoff than Vanilla-CP, as demonstrated by a slight left translation of TTA-Cal with respect to Vanilla-CP (Figure 4-5). This indicates that, conditional on prediction set size, TTA-Cal achieves a higher coverage than Vanilla-CP. We hypothesize this phenomenon occurs because calibration plays a greater role in conformal prediction on Fitzpatrick 17k-8 than it does in ImageNet. There are many factors that could increase the importance of calibration for Fitzpatrick 17k-8 conformal prediction. These two tasks vary in classification difficulty, randomization, and number of samples. In particular, Fitzpatrick 17k-8 has a smaller calibration set and a worse classifier than ImageNet. These two factors could contribute to the importance of calibration in conformal prediction, explaining why TTA-Cal improves the tradeoff between achieved coverage and prediction set size compared to Vanilla-CP.

On the other hand, it is hard to reason if TTA-Set-Intersection improves the tradeoff of achieved coverage and prediction set size. TTA-Set-Intersection (in green) departs from the tradeoff of Vanilla-CP. We again observe a translation to the left

of the Vanilla-CP tradeoff, indicating that, conditional on prediction set size, TTA-Set-Intersection achieves a higher coverage. We also note, however, that TTA-Set-Intersection produces an overall worse coverage than Vanilla-CP.

4.3 Summary

In this chapter, we determine that TTA-Cal, TTA-Set-Majority, and TTA-Cal-Set-Majority maintain the coverage guarantee, while the achieved coverage produced by TTA-Set-Intersection is dependent on augmentation policy. TTA-Set-Intersection can produce a better tradeoff between set size and achieved coverage, but TTA-Cal produces negligible differences across both datasets. There are a number of reasons this could be true - the datasets, models, or augmentations we look at - and future work should examine how our results depend on these factors. We build intuition as to why classifier performance causes randomization to worsen coverage for Fitzpatrick 17k-8, and provide evidence supporting the theoretical expectation that allowing empty prediction sets consistently worsens achieved coverage.

Figure 4-4: **For ImageNet, TTA-Cal preserves the tradeoff made by Vanilla-CP between achieved coverage and prediction set size. TTA-Set-Intesection does the same at low *alpha* (i.e. high theoretical coverage), but produces worse coverage conditional on prediction set size at higher α** Achieved coverage results were averaged across 10 runs. Average prediction set size is calculated as the median of mean prediction set sizes for 10 runs. Augmentations used in TTA-Cal and TTA-Set-Intersection are horizontal flip and random crop. α values plotted range from 0.05 to 0.95, at increments of 0.05.

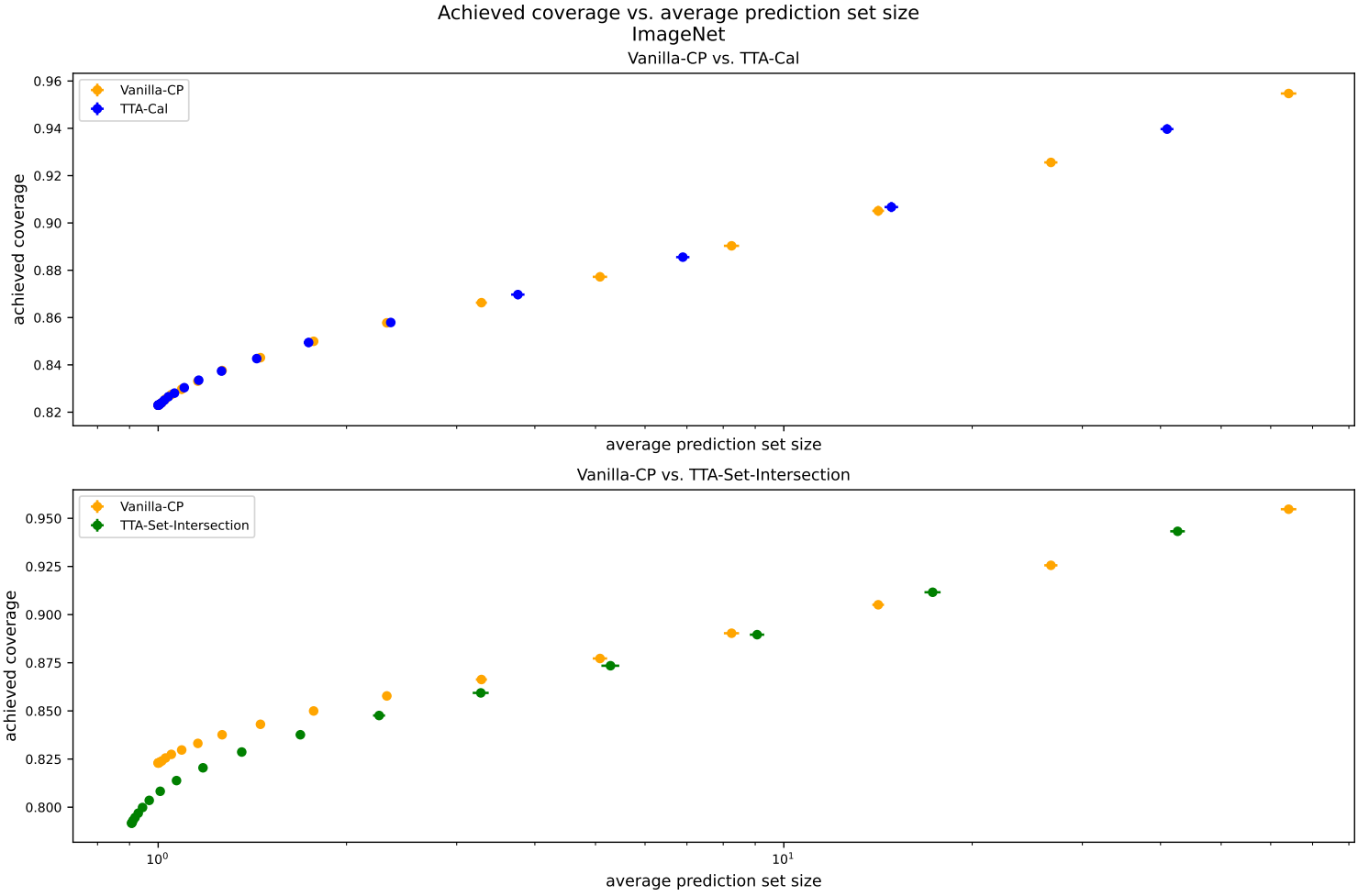
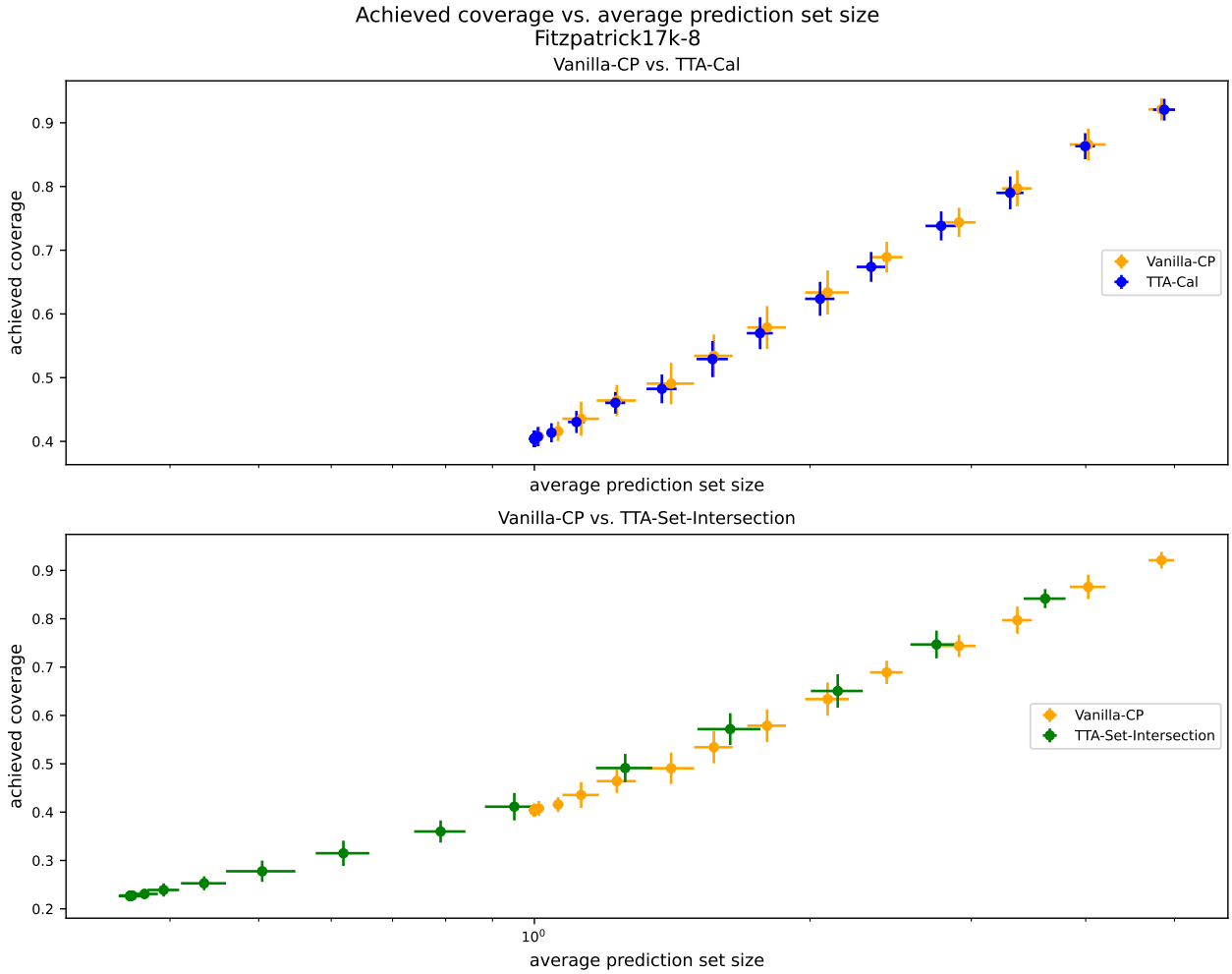


Figure 4-5: **For Fitzpatrick 17k-8, TTA-Cal produces a slightly more favorable tradeoff than Vanilla-CP between achieved coverage and prediction set size. Conditional on prediction set size, TTA-Set-Intersection achieves higher coverage than Vanilla-CP.** Achieved coverage results were averaged across 10 runs. Average prediction set size is calculated as the median of mean prediction set sizes for 10 runs. Augmentations used in TTA-Cal and TTA-Set-Intersection are horizontal flip and random crop. α values plotted range from 0.05 to 0.95, at increments of 0.05.



Chapter 5

Class-Level Effects of TTA-CP on Coverage and Prediction Set Size

In this chapter, we analyze class-specific performance of each conformal prediction method we study. We focus on the Fitzpatrick 17k-8 dataset to understand how class imbalance and class accuracy affects the performance of TTA-CP methods.

5.1 Fitzpatrick 17k-8 Class-Level Analysis

For Fitzpatrick 17k-8, we consider two class-related characteristics: prevalence in training data and model accuracy. In Chapter 3, we described the class imbalance found in the original Fitzpatrick dataset [12], motivating our decision to remove the "inflammatory" class to create the Fitzpatrick 17k-8 dataset. Even without the "inflammatory" class, Fitzpatrick 17k-8 demonstrates class imbalance (Figure 5-1). The classifier's performance on each class also differs greatly – from 81.08% on malignant cutaneous lymphoma to 31.46% on benign dermal (Table 5.1).

We find no notable difference in method performance trends when we break performance down by class, as shown in Figure 5-2. We hypothesize that the classifier's performance on augmented images is why we observe similarity between how TTA-CP methods affect prediction set size at a class-level and at the aggregate-level. Performance of TTA-CP methods depends primarily on the classifier's accuracy on

Figure 5-1: **Fitzpatrick 17k-8 dataset is imbalanced across the eight represented classes.** Count of samples across each of the eight classes in the entire Fitzpatrick 17k-8 dataset is shown in descending order. 60% of the samples were used for the train set, 20% for the validation set, and 20% for the test set.

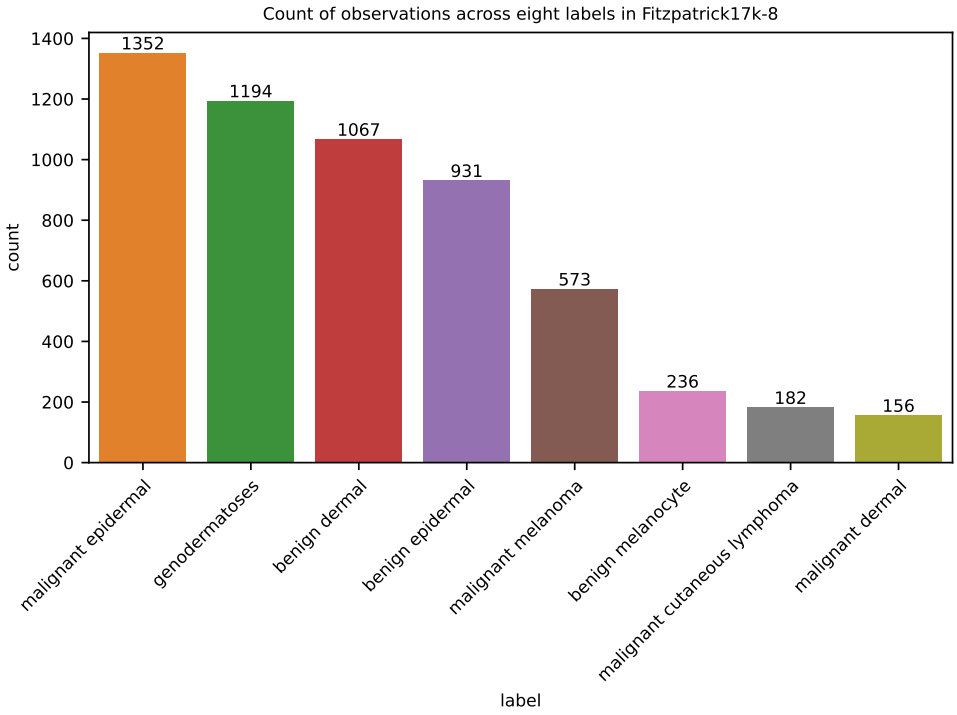
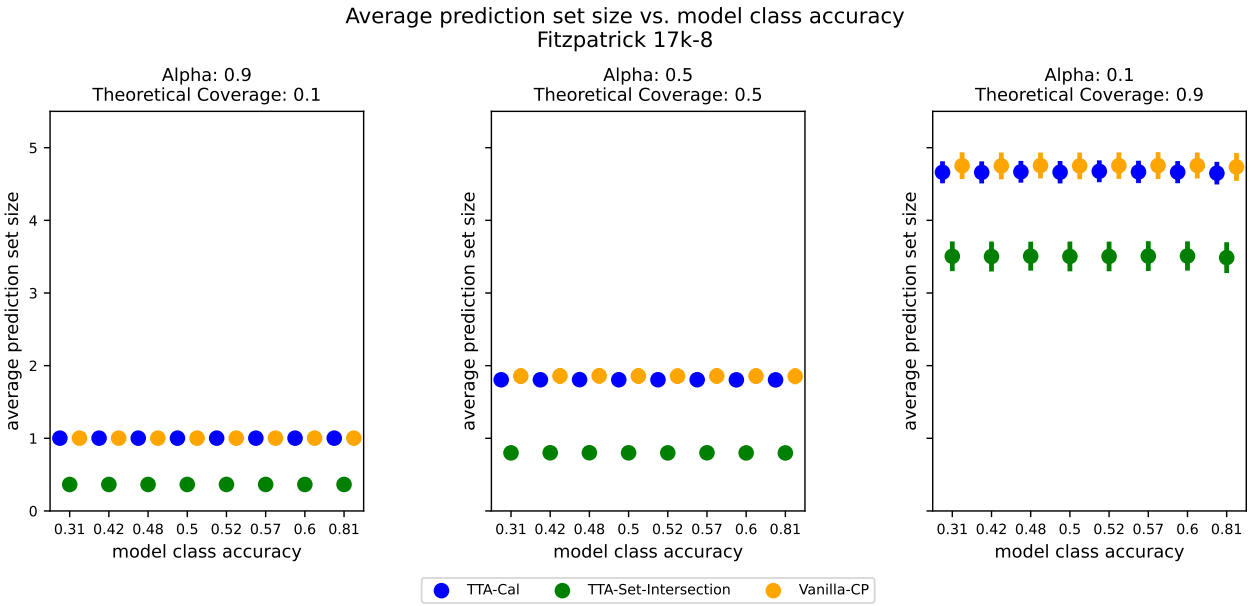


Table 5.1: **Fitzpatrick 17k-8 classifier achieves different class-specific accuracy across the eight classes.** Class-specific model accuracy ranges from 31.46% (benign dermal) to 81.08% (malignant cutaneous lymphoma).

label	accuracy
malignant cutaneous lymphoma	0.810811
benign melanocyte	0.595745
malignant epidermal	0.566667
malignant dermal	0.516129
malignant melanoma	0.504348
genodermatoses	0.481172
benign epidermal	0.424731
benign dermal	0.314554

Figure 5-2: **TTA-Set-Intersection reduces prediction set size uniformly across classes, regardless of model class accuracy. TTA-Cal produces a marginal reduction in prediction set size across all classes.** Average prediction set size calculated over 10 runs.



augmented images.

5.2 Summary

We find that the performance of TTA-CP methods at the class-level and dataset-level are consistent on the Fitzpatrick 17k-8 dataset. We recommend replicating this study on datasets with better-performing classifiers, such as ImageNet, to determine whether this pattern is an artifact of classifier accuracy.

Chapter 6

Applications of TTA-Cal in Decreasing Variance of Achieved Coverage

In this chapter, we investigate the performance of TTA-Cal in the presence of small calibration sets, which are known to increase the variance of coverage achieved by conformal prediction [2].

As defined in conformal prediction, the coverage guarantee ensures that the true class appears in the prediction set for a particular example with probability $1 - \alpha$. This guarantee is conditional on the calibration set. In other words, if we were to run conformal prediction multiple times using a different calibration set (drawn from the same distribution) each time, the coverage achieved on the same test sets for each run would on average satisfy the coverage guarantee. If $\alpha = 0.1$, two different runs might yield a coverage of 0.88 and a coverage of 0.92. On average, the runs meet the coverage guarantee, but produce different achieved coverage values while doing so. This variance in achieved coverage results from different thresholds being calculated from the calibration sets. Since the threshold used for creating prediction sets at test time relies on the samples seen at calibration, calibration sets composed of different samples will result in different threshold values.

The size of a calibration set will affect variance of achieved coverage, since we use

the finite calibration set to approximate a "true" threshold q for a particular α. For this reason, Angelopoulos et al. [2] investigate how calibration set size affects stability of achieved coverage and conclude that a calibration set of size 1000 is sufficient for most purposes. In our work, however, the TTA-CP method of TTA-Cal increases the size of the calibration set by treating each augmented version of an original calibration sample as a new calibration sample. Hence, we set out to understand if TTA-Cal can increase stability (i.e., reduce variance) in achieved coverage across subsamples of the same original calibration set.

In Chapters 4 and 5, we split the original test set 50/50 for calibration/test. To determine how modifications to the calibration set size affect achieved coverage on the test set, we resampled (with replacement) the calibration set of one 50/50 calibration/test split to create smaller calibration sets. We perform this resampling on on both ImageNet and Fitzpatrick 17k-8. For ImageNet, we examined calibration sets of size 5,000 and 10,000. Since the ImageNet validation set is uniformly distributed across 1,000 classes, this translates into approximately 5 calibration samples per class and 10 calibration samples per class, respectively. For Fitzpatrick 17k-8, we examined calibration sets of size 20, 50, 100, and 500. Since α ranges from 0 to 1, we select a low, intermediary, and high value of α (0.1, 0.5, and 0.9, respectively) and compare the mean and variance in achieved coverage by TTA-Cal and Vanilla-CP under these settings.

First, we examine ImageNet. Table 6.1 compares the mean and variance of coverage achieved by TTA-Cal and Vanilla-CP on ImageNet across 10 subsamples. We find that TTA-Cal causes a statistically significant decrease in achieved coverage for all considered calibration set sizes for α values of 0.1, 0.3, 0.5, and 0.7. At $\alpha = 0.9$ (i.e. coverage guarantee is 0.1), however, TTA-Cal and Vanilla-CP consistently achieve the same coverage regardless of calibration set size. We also observe that on ImageNet, TTA-Cal produces a statistically significant reduction in variance only when $\alpha = 0.5$ and the calibration set is of size 5000. At all other values of α and calibration set sizes considered for ImageNet, there is no statistically significant change in variance of achieved coverage.

Table 6.1: **On ImageNet, relative to Vanilla-CP TTA-Cal produces a statistically significant reduction in achieved coverage for high levels of coverage (i.e., $\alpha < 0.9$) across all calibration set sizes. For most combinations of calibration set size and α, TTA-Cal does not improve variance of achieved coverage.** 10 subsamples of the original calibration set are used to calculate mean and standard deviation. Statistical significance was calculated using a pairwise t-test for mean and Levnene's test for variance.

	calibration set size	**5000.0**		**10000.0**		**Original**	
		mean	**std**	**mean**	**std**	**mean**	**std**
α	**method**						
0.1	**TTA-Cal**	0.907****	0.001	0.907****	0.001	0.907****	0.001
	Vanilla-CP	0.926	0.001	0.926	0.001	0.926	0.000
0.3	**TTA-Cal**	0.851****	0.001	0.851****	0.000	0.851****	0.000
	Vanilla-CP	0.868	0.001	0.867	0.001	0.867	0.000
0.5	**TTA-Cal**	0.832****	0.000***	0.832****	0.000	0.832****	0.000
	Vanilla-CP	0.839	0.001	0.839	0.000	0.839	0.000
0.7	**TTA-Cal**	0.825****	0.000	0.825****	0.000	0.825****	0.000
	Vanilla-CP	0.827	0.000	0.827	0.000	0.827	0.000
0.9	**TTA-Cal**	0.825	0.000	0.825	0.000	0.825	0.000
	Vanilla-CP	0.825	0.000	0.825	0.000	0.825	0.000

Table 6.2: **On Fitzpatrick 17k-8, relative to Vanilla-CP TTA-Cal produces a statistically significant reduction in achieved coverage for high levels of coverage (i.e., $\alpha < 0.7$) across all calibration set sizes. For non-extreme values of α (i.e. 0.3, 0.5 and 0.7), TTA-Cal produces statistically-significant decreases in achieved coverage variance across all calibraion set sizes.** 100 subsamples of the original calibration set are used to calculate mean and standard deviation. Statistical significance was calculated using a pairwise t-test for mean and Levnene's test for variance.

	calibration set size	**20.0**		**50.0**		**100.0**		**500.0**		**Original**	
		mean	**std**	**mean**	**std**	**mean**	**std**	**mean**	**std**	**mean**	**std**
α	**method**										
0.1	**TTA-Cal**	0.933****	0.041*	0.927****	0.030	0.925****	0.022	0.924****	0.010*	0.900	0.0
	Vanilla-CP	0.965	0.031	0.942	0.027	0.939	0.020	0.933	0.009	0.916	0.0
0.3	**TTA-Cal**	0.757****	0.072*	0.739****	0.046*	0.738****	0.032	0.736****	0.018*	0.719	0.0
	Vanilla-CP	0.812	0.082	0.772	0.065	0.759	0.041	0.756	0.015	0.735	0.0
0.5	**TTA-Cal**	0.590****	0.064*	0.575****	0.042***	0.568****	0.027****	0.563****	0.013***	0.566	0.0
	Vanilla-CP	0.639	0.089	0.602	0.061	0.589	0.042	0.580	0.023	0.582	0.0
0.7	**TTA-Cal**	0.463****	0.039****	0.450***	0.021****	0.446	0.011***	0.444	0.003***	0.443	0.0
	Vanilla-CP	0.497	0.069	0.463	0.041	0.448	0.023	0.444	0.005	0.443	0.0
0.9	**TTA-Cal**	0.395****	0.010****	0.391***	0.001**	0.391*	0.001*	0.390	0.001	0.385	0.0
	Vanilla-CP	0.409	0.025	0.394	0.008	0.391	0.002	0.390	0.001	0.385	0.0

We conduct the same comparison for Fitzpatrick 17k-8. Table 6.2 compares the mean and variance of coverage achieved by TTA-Cal and Vanilla-CP on Fitzpatrick 17k-8. We find that, for α values of 0.3, 0.5, and 0.9, TTA-Cal frequently reduces (or matches) the variance in achieved coverage produced by Vanilla-CP. This trend is consistent across most calibration set sizes considered for Fitzpatrick 17k-8.

For a constant calibration set size, Vanilla-CP yields the lowest variance in achieved coverage at extreme values of α (i.e., 0.1 and 0.9). Variance is highest at $\alpha = 0.3$ at each calibration set size. Based on this observation, we hypothesize that TTA-Cal produces the most extreme reductions in variance at intermediary levels of α. Intuitively, TTA-Cal will not produce significant changes to prediction sets (and thus, to achieved coverage) at low or high values of α. For instance, when $\alpha = 0.9$, the coverage guarantee of 0.1 yields small prediction sets, many of which will be of size 1 because of our decision to not allow empty prediction sets. With this naturally tighter achieved coverage, there is little improvement that TTA-Cal can provide. However, the reason behind the insignificant reduction in variance at $\alpha = 0.1$ is less clear. Future work should examine the mechanisms affecting this observation.

We also find that the variance in achieved coverage is consistently higher for Fitzpatrick 17k-8 compared to ImageNet. Several factors may contribute to this variance, such as the larger calibration set sizes used for ImageNet; the smaller, more imbalanced nature of the Fitzpatrick 17k-8 dataset; and the use of randomization for ImageNet (but not Fitzpatrick 17k-8). Since Fitzpatrick 17k-8 is small and imbalanced, there are fewer representative samples per class in the calibration set. ImageNet calibration sets, on the other hand, have 5 images per class in our smallest setting, providing more information over which to form the quantile threshold used to create prediction sets.

6.1 Summary

From these results, we conclude that, when the calibration set is small, TTA-Cal may be useful in reducing the variance of achieved coverage at α values between 0.1 (high

coverage) and 0.9 (low coverage). We find that TTA-Cal provides stability while still preserving the coverage guarantee, despite a statistically significant reduction in achieved coverage. These findings imply that in settings where available data for calibration is limited, TTA-Cal could be used to produce levels of coverage variance closer to that produced by Vanilla-CP on larger calibration sets.

Chapter 7

Discussion and Future Work

In this thesis, we presented five methods of test-time augmentation-enhanced conformal prediction (TTA-CP) and investigated their performance on two classification tasks. In Chapter 4, we studied how test-time augmentation at the calibration and prediction set creation stages of APS conformal prediction affect achieved coverage and prediction set size. We found that there are many factors contributing to the performance of TTA-CP, such as task difficulty, dataset size, and augmentation policy. Since aggregate metrics can hide important class-specific behavior, we conducted a class-specifc analysis of TTA-CP in Chapter 5. We found no notable differences from the aggregate trends established in Chapter 4. In Chapter 6 of this work, we considered a different metric for conformal prediction performance: variance in achieved coverage. We found that, at α values between 0.1 (high coverage) and 0.9 (low coverage) and with reduced calibration set sizes, augmentation at the calibration stage (TTA-Cal) proves most useful in reducing the variance of achieved coverage. These results hold implications for using conformal predictors on datasets where obtaining a large calibration set is impossible.

Our learnings contribute to the conformal prediction field of research an understanding of how TTA might be used with CP and opens up avenues for future research. In our work, we found preliminary results indicating augmentation policy design to be a key factor in TTA-CP performance. A key decision is how to design a successful test-time augmentation policy; how many augmentations do optimal policies use?

How do the optimal augmentations relate to those used during training? Additionally, future work should explore designing optimal aggregation functions for test-time augmentation at the prediction set creation stage. Set-Intersection and Set-Majority, the two aggregation functions we used, are strongly affected by augmentations that are not label-preserving and create disagreement among prediction sets. An aggregation function that might mitigate this effect could be one that is learned and weighted based on class invariances, as studied by Shanmugam et al. [30]. With these changes to augmentation policy and aggregation function, the results may be different from what we present in Chapters 5 and 6. It is also possible that varying the choice of augmentation policy and aggregation function would reveal different behavior from what we describe for class-specific performance and reduction in achieved coverage variance.

Additional work includes exploring if other benefits from data augmentation translate to TTA-CP. Research has shown TTA to be useful in natural language classification too [21], and future work should consider how our results generalize to conformal prediction tasks in that domain. Data augmentation at train time is known to improve robustness in neural networks [26, 10, 36] and address covariate shift [20], an area of active conformal prediction research.

While in our work we focus on using conformal prediction to perform classification with certainty guarantees, conformal predictors have also been extended to other tasks, such as quantile regression [27], outlier detection [6, 19, 14, 13], and risk control metrics beyond accuracy [4, 3]. Future work includes the generalization of TTA-CP methods to these conformal prediction variants.

Bibliography

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

[2] Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021.

[3] Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control, 2022.

[4] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control, 2023.

[5] Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2018.

[6] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values, 2022.

[7] Maxime Cauchois, Suyash Gupta, and John Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction, 2020.

[8] Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6(1):1–6, jun 2004.

[9] Francois Chollet et al. Keras, 2015.

[10] Jeremy M Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing, 2019.

[11] Alex Derhacobian, John Guibas, Lin Tzy Li, and Bharath Namboothiry. Adaptive prediction sets with class conditional coverage. 2021.

[12] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828, 2021.

[13] Leying Guan. Conformal prediction with localization, 2020.

[14] Leying Guan and Robert Tibshirani. Prediction and outlier detection in classification problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):524–546, 2022.

[15] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. arXiv, 2017.

[16] Sotiris Kotsiantis, D. Kanellopoulos, and P. Pintelas. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30:25–36, 11 2005.

[17] Sotiris Kotsiantis and P. Pintelas. Mixture of expert agents for handling imbalanced data sets. *Annals of Mathematics, Computing Teleinformatics*, 1:46–55, 01 2004.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[19] Rikard Laxhammar and Göran Falkman. Conformal prediction for distribution-independent anomaly detection in streaming vessel data. In *Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*, StreamKDD '10, page 47–55, New York, NY, USA, 2010. Association for Computing Machinery.

[20] Ziquan Liu, Yi Xu, Yuanhong Xu, Qi Qian, Hao Li, Rong Jin, Xiangyang Ji, and Antoni B. Chan. An empirical study on distribution shift robustness from the perspective of pre-training and data augmentation, 2022.

[21] Helen Lu, Divya Shanmugam, Harini Suresh, and John Guttag. Improved text classification via test-time augmentation, 2022.

[22] Kazuhisa Matsunaga, Akira Hamada, Akane Minagawa, and Hiroshi Koga. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble, 2017.

[23] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. *Lecture Notes in Computer Science*, 2430:345–356, 2002.

[24] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Workshop on Autodiff*, 2017.

[25] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning, 2017.

[26] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection, 2018.

[27] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression, 2019.

[28] Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage, 2020.

[29] Tom Schaul, Sixin Zhang, and Yann LeCun. No more pesky learning rates, 2012.

[30] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. 2020.

[31] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60, 2019.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[33] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection, 2018.

[34] Liangchen Song, Jialian Wu, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. Handling difficult labels for multi-label image classification via uncertainty distillation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 2410–2419, New York, NY, USA, 2021. Association for Computing Machinery.

[35] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples, 2017.

[36] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples, 2018.

[37] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[38] Mahsa Taheri, Fang Xie, and Johannes Lederer. Statistical guarantees for regularized neural networks, 2020.

[39] Qi Tian, Kun Kuang, Kelu Jiang, Fei Wu, and Yisen Wang. Analysis and applications of class-wise robustness in adversarial training, 2021.

[40] Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candes, and Aaditya Ramdas. Conformal prediction under covariate shift, 2020.

[41] Volodya Vovk. On-line confidence machines are well-calibrated. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 187–196, 2002.

[42] Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 444–453, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[43] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, apr 2019.

[44] Suorong Yang, Weikang Xiao, Mengcheng Zhang, Suhan Guo, Jian Zhao, and Furao Shen. Image data augmentation for deep learning: A survey, 2022.

[45] Margaux Zaffran, Aymeric Dieuleveut, Olivier Féron, Yannig Goude, and Julie Josse. Adaptive conformal predictions for time series, 2022.