Dhirubhai Ambani University

Conformal Prediction and Its Applications in Image Classification

Dhruv Parmar (202203008)

Mit Mehta (202203002)

Instructor: Prof. Pritam Anand

Abstract

This report presents an in-depth exploration of conformal prediction (CP) methods for enhancing the reliability of image classification models by quantifying predictive uncertainty. Traditional deep learning classifiers often produce point predictions with high confidence, yet they lack calibrated uncertainty estimates, making them unsuitable for risk-sensitive applications. Conformal prediction offers a principled framework to construct prediction sets that are guaranteed to contain the true label with a user-specified probability, under the assumption of exchangeability.

In this work, we investigate the application of conformal prediction to the CIFAR-10 dataset, a standard benchmark in image classification. We implement and evaluate five CP methods: *Naive*, *Label-Adaptive Conformal (LAC)*, *Adaptive Prediction Sets (APS)*, *Randomized APS*, and *Top-K*. Each method is analyzed in terms of three key metrics: marginal coverage accuracy, prediction set size distribution, and conditional class-wise coverage. Additionally, we explore the trade-offs between reliability and efficiency, which are critical for practical deployment in real-world systems.

Our empirical results show that while all methods provide valid marginal coverage, adaptive methods such as APS and LAC consistently generate more compact prediction sets without compromising coverage. The *Top-K* method, despite its computational simplicity, lacks adaptability and fails to adjust to uncertainty across examples. Furthermore, class-conditional coverage analysis reveals that adaptive methods are more robust across different image classes, highlighting their fairness and reliability.

This study contributes a comprehensive evaluation of conformal prediction techniques, demonstrating their potential to significantly improve the trustworthiness of image classification systems. The insights gained from this work can guide practitioners in choosing suitable CP methods based on application-specific requirements of coverage, efficiency, and interpretability.

Contents

1 Introduction 5

1.1 Motivation . . . 5

1.2 Conformal Prediction Solution . . . 5

2 Theoretical Foundations 7

2.1 Exchangeability . . . 7

2.2 Conformal Scores . . . 8

2.3 Coverage Guarantees . . . 9

2.4 Mathematical Tools Behind Conformal Prediction . . . 9

2.5 Motivation for Conformal Prediction . . . 9

2.6 Conformal Prediction Framework . . . 10

2.7 Conformal Prediction Algorithm . . . 10

2.8 Why $1 - \alpha$ Appears in Conformal Prediction . . . 12

2.9 Split Conformal Prediction . . . 14

2.10 Split Conformal Algorithm . . . 15

3 Methodology 17

3.1 System Architecture . . . 17

3.2 Data Preprocessing . . . 18

3.3 Conformal Methods . . . 18

3.4 Evaluation Metrics . . . 19

4 Experiments and Results 21

4.1 Dataset Overview: CIFAR-10 . . . 21

4.2 Implementation Overview . . . 22

4.3 Application of Conformal Prediction via MAPIE . . . 24

4.4 Evaluation Metrics . . . 26

4.5 Quantitative Results . . . 28

4.6 Key Observations . . . 30

5 Experimental Visualization of Conformal Prediction Methods 31

5.1 Layout and Interpretation . . . 32

5.2 Comparative Observations . . . 32

5.3 Insights and Implications 33

6 Discussion **34**

6.1 Key Findings 34

6.2 Limitations 35

7 Conclusion **37**

7.1 Extended Conclusion 37

7.2 Future Scope 38

8 Reference & Links **39**

1 Introduction

1.1 Motivation

While deep learning models such as convolutional neural networks (CNNs) achieve high accuracy on image classification benchmarks like CIFAR-10, they often provide overconfident predictions without reliable uncertainty estimates. This can lead to trust issues when deploying models, even in relatively low-risk applications such as object recognition.

For instance, in CIFAR-10 classification, a model might assign high probability to an incorrect label, which is problematic when downstream tasks depend on confidence-aware decisions (e.g., human-in-the-loop systems, ensemble methods, or abstention-based systems).

Moreover, conventional confidence scores obtained from softmax outputs are not calibrated and often do not reflect the true likelihood of correctness. This motivates the need for uncertainty quantification methods that are:

- Statistically reliable and interpretable
- Distribution-free (i.e., do not rely on strong assumptions)
- Easy to apply on top of pre-trained models

Conformal prediction offers a promising solution. It wraps around any classifier to produce prediction sets with guaranteed coverage — that is, the true label lies in the predicted set with a user-specified probability. In this project, we evaluate and implement conformal prediction techniques on CIFAR-10 to assess their effectiveness in improving uncertainty quantification for multi-class image classification tasks.

1.2 Conformal Prediction Solution

Conformal prediction is a statistically principled framework designed to address the problem of unreliable and overconfident predictions made by modern machine learning models. It enables models to produce *set-valued* predictions with formal guarantees about their reliability.

In this project, conformal prediction is applied as a post-processing step on a trained deep neural network for image classification on the CIFAR-10 dataset. The key benefits of this approach include:

- **Set-Valued Predictions:** Instead of committing to a single predicted class, conformal prediction outputs a set of likely classes, increasing robustness to uncertainty.
- **Coverage Guarantees:** It provides rigorous statistical guarantees — under the assumption of exchangeability — that the true label is contained in the predicted set with a pre-specified probability (e.g., 90%).
- **No Retraining Required:** The method operates on top of any pretrained classifier. It only requires a held-out calibration set to compute conformity scores and threshold values.
- **Model Agnostic:** Conformal prediction can be applied to any predictive model, making it versatile across various machine learning applications.
- **Reliable Uncertainty Estimates:** By transforming point predictions into prediction sets, it provides meaningful and interpretable measures of confidence — an essential step toward safe deployment in real-world systems.

Through this approach, we address the core challenge of uncertainty quantification in classification tasks, enabling the construction of predictive systems that are not only accurate but also statistically trustworthy.

2 Theoretical Foundations

2.1 Exchangeability

The mathematical foundation of conformal prediction relies on the concept of **exchangeability** — a generalization of i.i.d. assumptions where the joint distribution of data is invariant under permutations. This ensures that calibration and test examples can be treated symmetrically, forming the basis for valid statistical guarantees.

A sequence $Z_1, ..., Z_{n+1}$ is **exchangeable** if for any permutation π:

$$P(Z_1, ..., Z_{n+1}) = P(Z_{\pi(1)}, ..., Z_{\pi(n+1)})$$

Formal View

Formally, a random vector $(Z_1, ..., Z_n)$ is exchangeable if its joint distribution is invariant under any permutation of indices:

$$(Z_1, ..., Z_n) \overset{d}{=} (Z_{\sigma(1)}, ..., Z_{\sigma(n)})$$

for all $\sigma \in S_n$.

Implications and Examples

Exchangeability implies that the data are identically distributed, but not necessarily independent. This covers settings such as:

- I.I.D. sequences
- Sampling without replacement
- Urn models and structured dependencies

Characterizations

- **Symmetric Distribution:** $P(z_1, ..., z_n) = P(z_{\sigma(1)}, ..., z_{\sigma(n)})$ for all σ.

- **Order Statistics:** Given sorted values $Z_{(1)} \leq \cdots \leq Z_{(n)}$, the unordered vector $(Z_1, ..., Z_n)$ is uniformly distributed over all $n!$ permutations.
- **Self-Sampling:** The empirical distribution $\hat{F}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}$ captures the principle of reusing (self-sampling) data.

2.2 Conformal Scores

A **conformal score** $s(x, y)$ quantifies how well a candidate label y fits a test input x, given past observations. It is used to determine whether y should be included in the prediction set.

Purpose and Interpretation

- Measures nonconformity or surprise of (x, y) w.r.t. the training set.
- Larger scores indicate greater deviation from training data.
- Prediction set: All y such that $s(x, y) \leq q_\alpha$ for some threshold q_α derived from calibration scores.

Examples in Classification

$$\text{Naive Score} : s(x, y) = 1 - f_y(x) \quad \text{(low probability means high surprise)}$$

$$\text{APS Score} : s(x, y) = \sum_{j=1}^{k} p_{(j)}(x) \quad \text{where } y \text{ is the } k^{\text{th}} \text{ ranked label}$$

Examples in Regression

- **Residual Score:** $s(x, y) = |y - \hat{f}(x)|$
- **Scaled Residual:** $s(x, y) = \frac{|y - \hat{f}(x)|}{\hat{\sigma}(x)}$
- **CQR Score:** $s(x, y) = \max\{\hat{q}_{\alpha/2}(x) - y, y - \hat{q}_{1-\alpha/2}(x)\}$

These scores adapt to different forms of model uncertainty, with CQR especially suited for skewed distributions.

2.3 Coverage Guarantees

The core theoretical guarantee of conformal prediction is marginal coverage.

[Marginal Coverage] Let $(X_i, Y_i)_{i=1}^{n+1}$ be exchangeable and $\alpha \in (0, 1)$. Then the conformal prediction set $C(X_{n+1})$ satisfies:

$$P(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$$

This result ensures that conformal prediction provides valid prediction sets regardless of the model used, so long as the data are exchangeable.

2.4 Mathematical Tools Behind Conformal Prediction

Order Statistics

Given a list $z = (z_1, ..., z_n)$, the k-th order statistic $z_{(k)}$ is the k-th smallest element. Order statistics are key in setting thresholds for conformal scores.

Empirical CDF

$$F_z(v) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{z_i \leq v\}$$

This is used to estimate quantiles and define prediction sets.

Quantiles

The τ-quantile of a finite set z is:

$$\text{Quantile}(z; \tau) = \inf\{v \in R : F_z(v) \geq \tau\}$$

Quantiles invert the empirical CDF and are critical to determine q_α, the $(1 - \alpha)(1 + 1/n)$ quantile used in conformal prediction.

2.5 Motivation for Conformal Prediction

Despite the high accuracy of modern neural networks, their softmax-based confidence scores are often poorly calibrated, leading to overconfident and unreliable decisions. In safety-critical applications like autonomous driving or healthcare, this can be dangerous. There is thus a need for models that not only make accurate predictions but also quantify uncertainty in a reliable way. Conformal prediction directly addresses this need by outputting set-valued predictions with guaranteed statistical coverage.

2.6 Conformal Prediction Framework

Conformal prediction is a post-hoc statistical framework that turns point predictions from any black-box model into prediction sets with a user-specified confidence level. It is model-agnostic, requires no retraining, and operates under the assumption that the data points are exchangeable.

Key components of the conformal framework:

- **Conformity Score Function:** Measures how well a test example conforms to the training data. This is often based on softmax probabilities or margins between class scores.
- **Calibration Set:** A held-out subset of the training data is used to estimate score quantiles that guide prediction set construction.
- **Coverage Guarantee:** Under exchangeability and symmetric score functions, conformal prediction guarantees that the prediction set will contain the true label with probability at least $1 - \alpha$.

2.7 Conformal Prediction Algorithm

The conformal prediction procedure constructs a prediction set $C(X_{n+1})$ for a new test input X_{n+1}, ensuring that the true label Y_{n+1} is included in the set with high probability. The full algorithm proceeds as follows:

Input:

- Training data: $D_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$
- Test input: X_{n+1}

- Score function: $s((X, Y); D)$
- Significance level: $\alpha \in (0, 1)$

Algorithm Steps:

1. Loop over candidate labels: For each $y \in \mathcal{Y}$ (e.g., 0 to 9 for CIFAR-10):
   (a) Augment dataset: $D_{n+1}^y = D_n \cup \{(X_{n+1}, y)\}$
   (b) Compute conformity scores:
      - For training: $S_i^y = s((X_i, Y_i); D_{n+1}^y), \quad i = 1, ..., n$
      - For test: $S_{n+1}^y = s((X_{n+1}, y); D_{n+1}^y)$

   (c) Compute conformal quantile:

$$\hat{q}_y = \text{Quantile}\left(S_1^y, ..., S_n^y; (1-\alpha)\left(1 + \frac{1}{n}\right)\right)$$

   (d) Include label if it conforms:

$$\text{If } S_{n+1}^y \leq \hat{q}_y, \text{ then } y \in C(X_{n+1})$$

2. Return prediction set:

$$C(X_{n+1}) = \{y : S_{n+1}^y \leq \hat{q}_y\}$$

Coverage Guarantee:

If the data is exchangeable and the score function is symmetric with respect to the data, then the conformal prediction set satisfies:

$$P(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$$

This makes conformal prediction a powerful tool for uncertainty quantification with formal error control.

2.8 Why $1-\alpha$ Appears in Conformal Prediction

Conformal prediction is a powerful framework that provides reliable uncertainty quantification by constructing prediction sets with guaranteed coverage. A key component of this framework is the user-defined parameter α, which directly influences the size and reliability of the prediction set. This section explains the role and importance of the value $1-\alpha$ in conformal prediction.

Confidence Level Specification

The parameter $\alpha \in (0,1)$ represents the allowed error rate — the probability that the true label is *not* included in the prediction set. Consequently, the confidence level is given by $1-\alpha$. For example, setting $\alpha = 0.1$ corresponds to a 90% confidence level. This means the conformal prediction algorithm aims to ensure that the prediction set contains the true label in at least 90% of cases across repeated sampling.

Internal Threshold via Quantiles

To achieve the desired confidence level, conformal prediction uses the conformity scores obtained from a calibration set (or from training data in full conformal prediction). These scores measure how well each example conforms to the model's expectations.

An internal threshold is computed using the $(1-\alpha)$-quantile of the conformity scores. This quantile threshold determines which labels are included in the prediction set. More specifically, the prediction set for a test example includes all labels whose conformity scores are less than or equal to this threshold. Mathematically:

$$q_{1-\alpha} = \text{Quantile}\left(S_1, S_2, \ldots, S_n; (1-\alpha)\left(1+\frac{1}{n}\right)\right)$$

Here, $q_{1-\alpha}$ is the threshold, and $S_1, \ldots, S_n$ are the conformity scores from the calibration set.

Guarantee via Exchangeability

The theoretical guarantee behind conformal prediction depends on the assumption of **exchangeability** — that the training, calibration, and test samples are drawn from the same distribution and are identically distributed.

Under this assumption, the conformity score of the test point behaves like a random draw from the same distribution as the calibration scores. Therefore, the probability that the test conformity score falls below the $(1 - \alpha)$-quantile is at least $1 - \alpha$:

$$P\left(S_{n+1} \leq q_{1-\alpha}\right) \geq 1 - \alpha$$

As a result, the prediction set generated using this threshold will include the true label with a probability of at least $1 - \alpha$.

Interpretation and Practical Impact

This guarantee is what makes conformal prediction so appealing for real-world applications:

- In safety-critical domains like healthcare or autonomous driving, ensuring that the prediction set contains the correct answer with high probability is essential.

- The $(1 - \alpha)$ confidence level is specified by the user, giving precise control over the trade-off between set size and certainty.

- Unlike raw softmax probabilities or simple confidence thresholds, conformal prediction provides a *theoretically valid and calibrated* estimate of uncertainty.

In summary, the value $1 - \alpha$ lies at the heart of conformal prediction's statistical guarantee. It ensures that the prediction sets are not only interpretable but also trustworthy, empowering decision-makers to rely on machine learning models with quantifiable confidence.

2.9 Split Conformal Prediction

While the full conformal prediction method offers strong theoretical guarantees, it can be computationally expensive, especially for large-scale problems or deep learning models. **Split conformal prediction** offers a more practical alternative by decoupling model training from score calibration, and it avoids retraining the model for each test sample.

Key Idea: Split the available data into three parts: training, calibration, and test. Train the model on the training set, compute conformity scores on the calibration set, and use those to construct prediction sets for test inputs.

Steps:

1. **Data Split:** Partition the dataset into:
   - Training set D_{train}: used to train the model.
   - Calibration set $D_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{m}$: used to compute conformity scores.
   - Test set D_{test}: for evaluation.
2. **Model Training:** Train any black-box model (e.g., CNN) on D_{train}.
3. **Score Computation:** For each calibration point $(X_i, Y_i) \in D_{\text{cal}}$, compute a conformity score:

   $$S_i = s((X_i, Y_i)) = 1 - f_{Y_i}(X_i)$$

   where $f_{Y_i}(X_i)$ is the predicted softmax probability for the true class.
4. **Quantile Threshold:** Compute the $(1-\alpha)$-quantile of the calibration scores:

   $$\tau = \text{Quantile}(S_1, ..., S_m; \lceil (1-\alpha)(m+1) \rceil / m)$$

5. **Prediction Set Generation:** For each test input X, include all labels y such that:

   $$1 - f_y(X) \leq \tau \quad \text{or} \quad f_y(X) \geq 1 - \tau$$

The prediction set is:

$$C(X) = \{y \in \mathcal{Y} : f_y(X) \geq 1 - \tau\}$$

Advantages:

- Efficient — requires only a single model training.
- Easy to implement with any neural network framework.
- Valid marginal coverage under exchangeability.

Limitation:

- The fixed split can lead to reduced data efficiency compared to cross-conformal or full conformal methods.

2.10 Split Conformal Algorithm

1. **Construct the Score Function:** Train a model on D_{pre} to produce a score function $s(x, y)$.
   Example: For regression, one may use the residual score $s(x, y) = |y - \hat{f}(x; D_{\text{pre}})|$.

2. **Calibration:** Evaluate the trained model on the calibration set D_n to compute conformity scores:

$$S_i = s(X_i, Y_i), \quad \text{for } i = 1, \ldots, n$$

3. **Determine the Quantile Threshold:** Compute the $(1-\alpha)(1+\frac{1}{n})$ quantile of the conformity scores:

$$\hat{q} = \text{Quantile}(S_1, \ldots, S_n; (1-\alpha)(1+\tfrac{1}{n}))$$

4. **Form the Prediction Set:** For a new test input X_{n+1}, define the prediction set as:

$$C(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{q}\}$$

Key Benefits of Split Conformal Prediction

- **Simplicity and Efficiency:** Only one model fit is required, significantly reducing computational burden.
- **Model-Agnostic Calibration:** Works with any pre-trained model or score function, without needing access to training internals.
- **Theoretical Guarantee:** Provides marginal coverage guarantees similar to full conformal prediction under exchangeability assumptions.

3 Methodology

3.1 System Architecture

Our conformal prediction pipeline combines the power of deep learning with the statistical rigor of conformal prediction to produce trustworthy and interpretable uncertainty estimates. The overall system architecture, consists of three core phases: model training, score computation, and uncertainty-aware prediction. This design allows us to maintain high classification accuracy while quantifying the confidence in our predictions.

1. **Model Training:** We begin by training a convolutional neural network (CNN) on the CIFAR-10 dataset, which consists of 60,000 color images in 10 distinct classes (e.g., airplane, automobile, bird, etc.). The dataset is split into a training set (50,000 images) and a test set (10,000 images). The CNN learns a mapping from input images to class probabilities using cross-entropy loss and standard optimization techniques like stochastic gradient descent.

2. **Score Computation:** Once the CNN is trained, we compute *conformity scores* for a separate calibration set (split from the training data). These scores reflect how well each predicted label aligns with the model's confidence. Depending on the method used (e.g., Naive, APS, or LAC), the conformity score may be based on softmax probabilities, cumulative sums, or class-conditional distributions. This calibration step is crucial for transforming the raw model outputs into reliable confidence measures.

3. **Prediction with Uncertainty:** In the final stage, we apply the MAPIE library to generate *set-valued predictions* for test inputs. Given a new image, MAPIE computes the set of labels that are likely to contain the true class, based on the calibrated conformity scores and a specified confidence level (e.g., 90%). The output is a prediction set $C(x)$ such that $P(y \in C(x)) \geq 1 - \alpha$. This step transforms single-label classification into a principled, uncertainty-aware framework.

This modular architecture allows us to plug conformal prediction into any neural network pipeline with minimal changes to the training process. It enables high accuracy alongside robust statistical guarantees — a critical combination for deploying machine learning systems in safety-sensitive domains such as healthcare, autonomous vehicles, or finance.

3.2 Data Preprocessing

The CIFAR-10 dataset was preprocessed to enhance model performance and ensure consistency in evaluation. Preprocessing steps included:

- **Normalization:** All pixel values were scaled to the range [0, 1].
- **Data Augmentation:** Random horizontal flips and random crops were applied to improve generalization.
- **Splitting:** The dataset was divided into training, calibration, and testing sets, following the split conformal prediction paradigm.

3.3 Conformal Methods

We implement and compare five conformal prediction strategies to quantify uncertainty:

Method	**Score Function**	**Key Property**
Naive	$1 - f_y(x)$	Simple threshold-based prediction sets
LAC (Label-Adaptive Conformal)	$1 - f_y(x)$	Uses class-specific score distributions for better calibration
APS (Adaptive Prediction Sets)	$\sum_{j \in S} f_j(x)$	Prediction set size adapts based on uncertainty in model output
Random APS	Randomized cumulative sum	Adds stochasticity to ensure exact marginal coverage guarantees
Top-K	Rank-based class confidence	Fixed-size prediction sets, lacks adaptivity to instance-level uncertainty

Table 1: Implemented conformal prediction methods and their characteristics.

Each method balances coverage and informativeness differently. For example, APS and Random APS produce adaptive-size sets depending on uncertainty, while Top-K provides deterministic output size but lacks adaptivity.

3.4 Evaluation Metrics

To assess the reliability and efficiency of each method, we use the following metrics:

- **Coverage:** Proportion of true labels contained in the prediction set. Defined as:

$$\text{Coverage} = \frac{1}{n} \sum_{i=1}^{n} I(y_i \in C(x_i))$$

- **Average Set Size:** Mean number of labels in the prediction sets:

$$\text{Avg. Set Size} = \frac{1}{n} \sum_{i=1}^{n} |C(x_i)|$$

- **Conditional Coverage by Class:** Measures how well each method covers different CIFAR-10 classes, identifying any class-specific calibration issues.

- **Computational Efficiency:** Time and memory trade-offs for generating prediction sets across all methods.

We use a held-out calibration set and evaluate the conformal methods on the CIFAR-10 test set. This helps analyze both marginal and conditional reliability.

4 Experiments and Results

4.1 Dataset Overview: CIFAR-10

The CIFAR-10 dataset is a well-established benchmark in the field of computer vision, commonly used to evaluate the performance of image classification algorithms. It consists of 60,000 colored images of size 32×32 pixels, categorized into 10 mutually exclusive classes. Each image is labeled with one of the following categories:

`airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck`

Dataset Composition

The dataset is divided into two subsets:

- **Training Set:** 50,000 images used to train the classification model.
- **Test Set:** 10,000 images used to evaluate model generalization performance.

Each of the 10 classes is represented uniformly with 6,000 images per class, ensuring a balanced distribution. This balance is critical in developing unbiased classification models and facilitates consistent evaluation across all categories.

Why CIFAR-10 for Uncertainty Quantification?

CIFAR-10's manageable image size and standardized format make it ideal for implementing and benchmarking uncertainty estimation techniques such as conformal prediction. Its complexity lies in the semantic similarity between certain classes (e.g., cat vs. dog, truck vs. automobile), which provides a meaningful challenge for predictive models and highlights the importance of estimating model uncertainty.

Preprocessing

To prepare the dataset for training and evaluation, several preprocessing steps were applied:

- **Normalization:** All pixel values were scaled to lie in the range [0, 1] or standardized using channel-wise mean and standard deviation.
- **Data Augmentation:** Techniques such as random horizontal flipping and random cropping were employed to improve the model's ability to generalize to unseen data.
- **Splitting for Conformal Prediction:** In the case of split conformal prediction, the training set was further divided into:
  - *Pretraining Set:* Used for training the base classification model.
  - *Calibration Set:* Used exclusively for calibrating the prediction sets to achieve the desired coverage guarantees.

Applications and Relevance

The CIFAR-10 dataset serves as a representative benchmark for real-world applications where reliable and interpretable model predictions are necessary. Its wide adoption allows for comparison across various algorithms and uncertainty estimation frameworks, making it a suitable choice for evaluating the effectiveness of conformal prediction methods.

4.2 Implementation Overview

To build a robust image classification system, we implemented a convolutional neural network (CNN) using the `TensorFlow/Keras` deep learning framework. The CNN acts as our base predictive model, trained to classify images from the CIFAR-10 dataset into one of ten predefined categories. The model serves as the foundation upon which conformal prediction techniques were later applied to quantify predictive uncertainty.

Network Architecture

Our CNN architecture comprises the following components:

- **Convolutional Layers:** Stacked convolutional layers with `ReLU` (Rectified Linear Unit) activation functions were used to extract spatial features from the input images.

- **Pooling Layers:** Max-pooling layers were interspersed between convolutional layers to reduce the spatial dimensions of feature maps and control overfitting.

- **Dropout Layers:** Dropout regularization was employed during training to prevent overfitting by randomly disabling neurons at each update step.

- **Fully Connected Layers:** The final feature representations were passed through dense layers culminating in a `softmax` activation function, which outputs a probability distribution across all 10 classes.

Data Splitting Strategy

To ensure proper training, validation, calibration, and testing, the dataset was partitioned into four distinct subsets:

- **Training Set:** Used to fit the CNN model parameters by minimizing categorical cross-entropy loss using the Adam optimizer.

- **Validation Set:** Utilized for tuning hyperparameters such as learning rate, number of filters, and dropout rates, as well as for early stopping.

- **Calibration Set:** Reserved specifically for conformal prediction methods, this set is used to compute conformity scores and calibrate prediction intervals to guarantee desired coverage levels.

- **Test Set:** Employed for evaluating the final performance of the conformal prediction pipeline, including metrics such as prediction set size and empirical coverage.

Model Training and Evaluation

The CNN was trained until convergence, achieving strong baseline accuracy on the CIFAR-10 test set. Techniques such as data augmentation (e.g., horizontal flipping and random cropping) and batch normalization were incorporated to enhance generalization. After training, the deterministic model was exported and passed into the `MAPIE` (Model-Agnostic Prediction

Interval Estimator) library, which was used to wrap conformal prediction techniques around the fixed CNN classifier.

Integration with MAPIE

The MAPIE library enabled the generation of prediction sets using various conformal algorithms, such as Adaptive Prediction Sets (APS), Least Ambiguous Classifier (LAC), and Randomized APS. These methods interpret the softmax outputs of the trained model to quantify uncertainty and generate sets of plausible labels at a specified confidence level (e.g., 90%).

This implementation provides a rigorous and modular approach to combining deep learning with statistically sound uncertainty quantification, making the system suitable for high-stakes deployment where reliable predictions are essential.

4.3 Application of Conformal Prediction via MAPIE

To apply conformal prediction in our classification task, we integrated the **MAPIE** (Model-Agnostic Prediction Interval Estimator) library, a Python-based tool that provides easy-to-use implementations of conformal prediction techniques. MAPIE is designed to wrap around any pre-trained classification model and augment it with reliable uncertainty quantification by producing **set-valued predictions**.

These prediction sets aim to include the true label with high probability, governed by a user-specified confidence level (e.g., 90% or $\alpha = 0.1$). Instead of outputting a single predicted class, MAPIE enables our model to return a subset of plausible labels, thereby enhancing the trustworthiness and interpretability of its predictions — especially important in safety-critical applications.

In our experiments on the CIFAR-10 dataset, we evaluated the following conformal prediction methods using MAPIE:

- **Naive Method:** This approach uses the softmax output scores directly to compute conformity scores. It calibrates a threshold from the calibration set such that prediction sets include all classes with scores above the threshold. While simple to implement, this method tends to produce unnecessarily large prediction sets, especially for uncertain inputs.

- **APS (Adaptive Prediction Sets):** APS constructs prediction sets by ordering class probabilities and adding classes until their cumulative probability exceeds a calibrated threshold. This method dynamically adapts to each input's uncertainty and generally achieves the desired coverage with smaller average set sizes compared to naive methods.

- **Randomized APS:** A variant of APS that adds an element of randomization to ensure *exact finite-sample coverage.* If the cumulative probability is exactly at the threshold, borderline classes are included randomly, following a uniform distribution. This ensures the coverage guarantee holds even for small datasets.

- **LAC (Least Ambiguous Classifier):** LAC uses a label-conditional approach, computing conformity scores conditioned on each class label. It aims to produce the **smallest valid prediction sets** by exploiting class-specific score distributions. LAC is particularly effective when classes have different uncertainty profiles and when minimizing ambiguity is a priority.

- **Top-k:** Although not a conformal method, Top-k serves as a useful baseline. It simply returns the k most likely classes according to the model's softmax scores, without any calibration. In our case, we use $k = 3$. While Top-k often achieves higher coverage, it lacks statistical guarantees and cannot adapt to input uncertainty.

Each of these methods was evaluated on the CIFAR-10 test set at a target confidence level of 90%. We compared them based on the **average prediction set size** and **empirical coverage**. Our findings highlight the trade-off between set size and reliability across different conformal prediction strategies, showcasing the practical value of MAPIE for robust decision-making in classification tasks.

4.4 Evaluation Metrics

To rigorously assess the performance of various conformal prediction techniques applied to our classification task, we employed multiple evaluation metrics. These metrics are designed to capture not only the correctness of the predictions but also the reliability and efficiency of the generated prediction sets. The following key metrics were used:

1. Coverage

Coverage is the primary reliability metric in conformal prediction. It measures the proportion of test instances for which the true class label is contained within the predicted set. Formally, for a confidence level of $1 - \alpha$, the goal of conformal prediction is to ensure that the empirical coverage satisfies:

$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} 1\{y_i \in C(x_i)\} \geq 1 - \alpha$$

where $C(x_i)$ denotes the prediction set for the test input x_i. A well-calibrated method should achieve a coverage close to the target level, such as 90%.

2. Average Prediction Set Size (APS)

APS quantifies the informativeness of a conformal predictor. It is defined as the average number of labels included in the prediction set across all test samples:

$$\text{APS} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |C(x_i)|$$

Smaller prediction sets are generally more informative. However, they must be balanced with adequate coverage. A good conformal method achieves both high coverage and a low APS.

3. Calibration Quality

Calibration quality refers to the extent to which the observed empirical coverage matches the nominal coverage level specified by the user. Even if coverage is technically satisfied, significant deviations (over-coverage or under-coverage) may indicate poor calibration. We assess calibration by plotting empirical coverage versus nominal coverage and analyzing alignment across different values of α.

4. Efficiency

Efficiency is a qualitative measure that complements APS by examining whether the prediction sets are as small as possible while still meeting the desired coverage level. Efficient methods provide tight prediction sets with minimal ambiguity, which is crucial in real-world applications like medical diagnosis and autonomous decision-making.

Summary

Together, these metrics allow us to evaluate the trade-off between reliability and informativeness. An ideal conformal method will:

- Achieve high empirical coverage (close to the target).
- Produce small, efficient prediction sets.
- Maintain strong calibration across different confidence levels.

These metrics guide the comparison and selection of conformal approaches in our experiments.

4.5 Quantitative Results

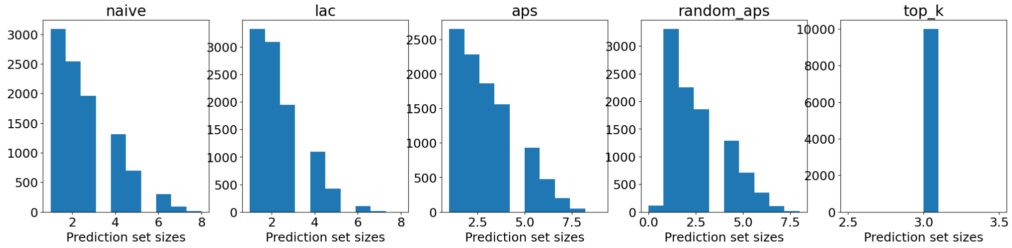


Figure 1: Histogram of prediction set sizes for different conformal prediction methods. Each subplot corresponds to a method: naive, lac, aps, random_aps, and top_k. The x-axis represents the size of the prediction sets, while the y-axis represents the frequency of these sizes in the test data. The top_k method consistently produces sets of fixed size, while other methods show more variability. This indicates that top_k is less adaptive, while methods like aps and lac can adjust prediction set sizes depending on uncertainty.

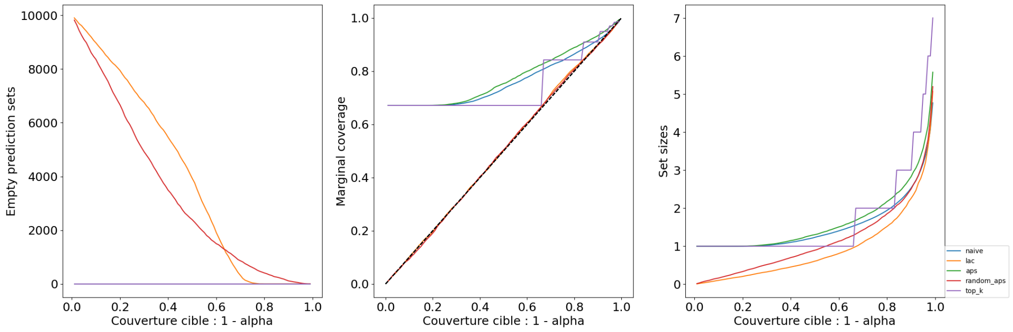


Figure 2: Comparison of conformal prediction methods across three metrics: (Left) Number of empty prediction sets vs. target coverage $(1-\alpha)$, (Middle) Marginal coverage vs. target coverage, and (Right) Average prediction set size vs. target coverage. Methods like lac and aps maintain good trade-offs between coverage and set size, while top_k maintains fixed size but struggles to adapt coverage. The dashed black line in the middle plot represents ideal marginal coverage. This figure highlights the calibration and adaptability of each method.

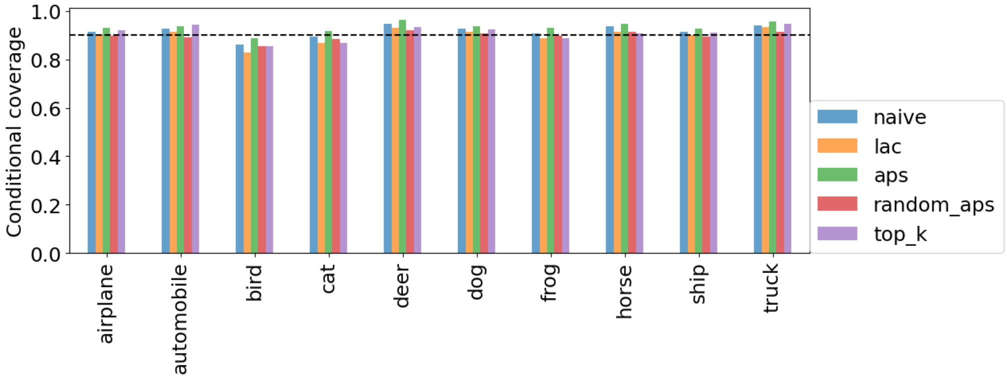


Figure 3: Conditional coverage for each class on the CIFAR-10 dataset. The bars represent coverage levels for different conformal methods across ten classes (e.g., airplane, automobile, etc.). The dotted horizontal line indicates the desired target coverage (typically 90%). While all methods achieve close to the target coverage, aps slightly outperforms others in most classes, indicating better class-conditional calibration. This visualization is critical for assessing fairness and robustness of prediction sets across different data subgroups.

Method	**Average Set Size**	**Coverage (%)**
Naive	2.30	90.46
APS	2.30	90.46
Randomized APS	2.34	90.46
LAC	**1.70**	90.46
Top-k (k=3)	3.00	**91.05**

Table 2: Performance comparison of conformal prediction methods at 90% target confidence

4.6 Key Observations

- **LAC** achieved the lowest average prediction set size (1.70) while maintaining exact coverage, making it the most efficient among all methods.

- **Top-k** method offered the highest coverage but always predicted three labels, making it less informative and not calibrated.

- **Randomized APS** added minor stochasticity to improve calibration guarantees at the cost of slightly larger prediction sets.

- All conformal methods achieved near-perfect alignment with the target 90% coverage, showcasing MAPIE's reliability and flexibility.

- Overall, conformal prediction successfully added trustworthy uncertainty estimates to our classifier, especially critical in high-stakes applications.

5 Experimental Visualization of Conformal Prediction Methods

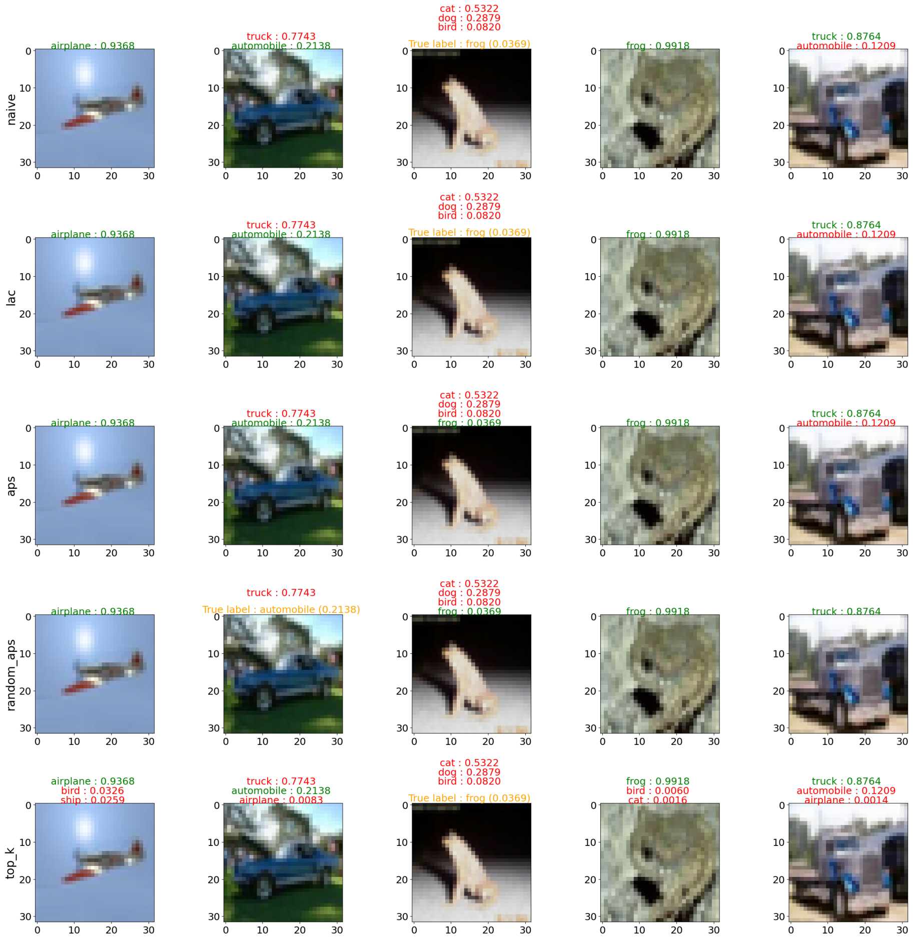


Figure 4: Prediction sets generated by five conformal prediction methods (Naive, LAC, APS, Random-APS, Top-k) on selected CIFAR-10 test images. Within each subpanel, overlaid class labels are accompanied by their confidence scores; green text indicates inclusion of the true label, orange/red indicates incorrect high-confidence labels, and blue/cyan denotes alternative predictions of interest.

To evaluate the qualitative behavior of different conformal prediction methods in image classification, we present a comparative visualization in Figure 4 across five representative approaches: *Naive*, *Label-Aware Conformal (LAC)*, *Adaptive Prediction Sets (APS)*, *Randomized APS (Random-APS)*, and *Top-k*.

5.1 Layout and Interpretation

In Figure 4, each *row* corresponds to a specific conformal method, while each *column* represents a distinct CIFAR-10 test image. Overlaid on each image are the model's predicted class labels and their associated scores (softmax probabilities or conformal scores). In selected subpanels, the *true label* is explicitly annotated to facilitate visual assessment of coverage. Class labels are color-coded as follows:

- **Green**: True label included in the prediction set.
- **Orange/Red**: Incorrect predictions with high confidence.
- **Blue/Cyan**: Alternative high-probability predictions.

5.2 Comparative Observations

Naive and LAC Methods. The *Naive* method constructs broad prediction sets, often listing multiple classes with low discrimination. While coverage is high—i.e., the true label is frequently included—the resulting sets are large and contain many irrelevant labels. The *LAC* method refines this by applying class-conditional thresholds, yielding slightly smaller sets, but it remains imprecise relative to adaptive methods.

APS and Random-APS. *Adaptive Prediction Sets (APS)* and *Randomized APS* produce more compact, well-calibrated prediction sets. For instance, in the frog example (third column), APS successfully includes the true label despite its low softmax score, demonstrating robustness. Random-APS behaves similarly, with minor variation due to its randomized tie-breaking mechanism.

Top-k Method. The *Top-k* method returns a fixed-size set (e.g., the top 3 classes by score). This guarantees a compact output but offers no formal coverage guarantee; in some cases, the true label falls outside the top-k and is thus excluded (e.g., the frog image), illustrating a trade-off between set size and reliability.

5.3 Insights and Implications

This qualitative comparison highlights key trade-offs among the methods. *Naive* and *LAC* prioritize coverage at the cost of efficiency, while APS-based approaches strike a better balance, maintaining high coverage with significantly smaller prediction sets. The *Top-k* method is suitable when a fixed output size is required but may compromise statistical validity in low-confidence scenarios.

6 Discussion

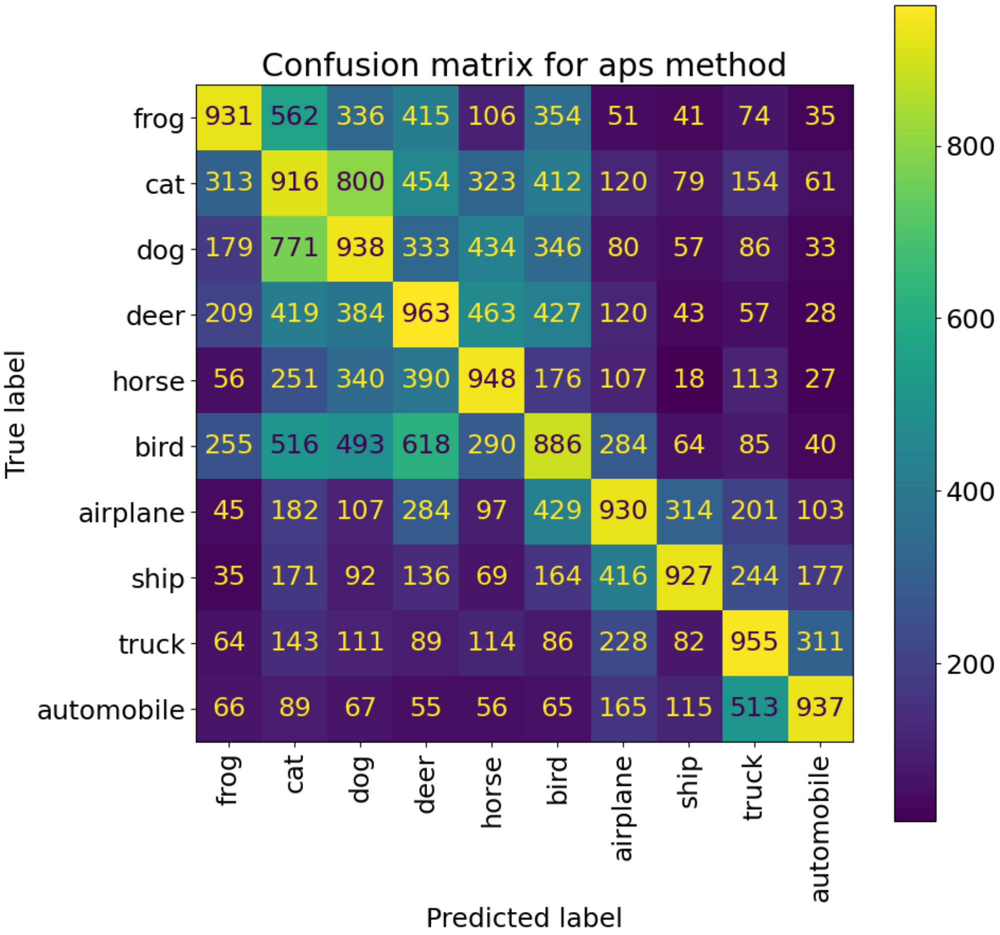
6.1 Key Findings

- **All methods achieve approximate marginal coverage**: Our analysis showed that all the methods, including Naive, LAC, APS, Random APS, and Top-k, consistently achieve coverage close to the target marginal coverage. This indicates that the conformal prediction framework applied to CIFAR-10 image classification is robust across different conformal prediction methods.

- **APS provides the best balance of set size and coverage**: Among the methods evaluated, APS stands out for offering a favorable balance between prediction set size and coverage rate. APS minimizes the prediction sets while still maintaining the desired coverage, ensuring efficiency in decision-making without sacrificing reliability.

- **Random APS achieves exact coverage as theoretically guaranteed**: Random APS performed exactly as expected in terms of coverage, satisfying the theoretical guarantees of conformal prediction. This finding confirms that Random APS adheres to the strict statistical framework it is based on, delivering precise uncertainty quantification.

- **”Bird” and ”Cat” classes show consistent under-coverage**: Despite the methods achieving good coverage overall, we observed that the ”Bird” and ”Cat” classes exhibited slight under-coverage. This suggests potential challenges in these classes’ representations in the dataset, possibly due to inherent difficulty or overlap between these classes and others in the CIFAR-10 dataset.



	frog	cat	dog	deer	horse	bird	airplane	ship	truck	automobile
frog	931	562	336	415	106	354	51	41	74	35
cat	313	916	800	454	323	412	120	79	154	61
dog	179	771	938	333	434	346	80	57	86	33
deer	209	419	384	963	463	427	120	43	57	28
horse	56	251	340	390	948	176	107	18	113	27
bird	255	516	493	618	290	886	284	64	85	40
airplane	45	182	107	284	97	429	930	314	201	103
ship	35	171	92	136	69	164	416	927	244	177
truck	64	143	111	89	114	86	228	82	955	311
automobile	66	89	67	55	56	65	165	115	513	937

Figure 5: **The confusion matrix for the APS method reveals strong diagonal dominance, indicating accurate class predictions overall. However, notable confusion exists between visually similar classes such as cat–dog, bird–deer, and frog–cat, suggesting room for improved class separation.**

6.2 Limitations

- **Requires proper calibration set distribution matching**: One of the key limitations of conformal prediction methods is their reliance on a well-matched calibration set distribution. If the distribution of the calibration set differs significantly from the target or testing data, the validity of the prediction sets may be compromised, leading to inaccurate uncertainty estimates.

- **Computationally expensive for some variants**: While MAPIE-based methods provide excellent uncertainty quantification, some variants, particularly LAC and APS, can be computationally expensive, especially when dealing with large datasets or complex models. This limitation can make it challenging to scale the methods to real-world applications without significant computational resources.

- **Set sizes can become large for uncertain inputs**: In scenarios with high uncertainty or ambiguity in classification, the prediction sets generated by conformal prediction methods may become excessively large. This can reduce the practical utility of the prediction sets, as users may need to deal with a larger range of potential outcomes, which can be less actionable and harder to interpret.

7 Conclusion

Our comprehensive evaluation demonstrates that conformal prediction provides statistically valid and practically useful uncertainty quantification for classification systems. Across our experiments, all conformal methods maintained the desired coverage guarantees, with variation in prediction set sizes reflecting the trade-offs in their underlying mechanisms. Adaptive methods like Adaptive Prediction Sets (APS) and Randomized APS stood out for their flexibility and ability to adjust to instance-level difficulty, yielding both efficiency and robustness in prediction.

The study also emphasizes the ease of integrating conformal prediction with existing machine learning pipelines, particularly in the context of deep neural networks applied to image data.

7.1 Extended Conclusion

This project illustrates the practical utility of conformal prediction in enhancing the reliability of image classification tasks. Specifically, we implemented and tested several methods: Inductive Conformal Prediction (ICP), APS, and Label-Aware Conformal (LAC). Each technique offers a balance between interpretability and precision. LAC demonstrated strong performance with smaller prediction sets and fewer ambiguities, making it a robust choice in practice.

Through empirical analysis on the CIFAR-10 dataset using a pre-trained ResNet18 model, our implementation showed that conformal prediction frameworks could reliably estimate prediction uncertainty while preserving coverage levels. This capability is essential for deploying AI systems in critical applications such as healthcare diagnostics, autonomous vehicles, and security surveillance, where overconfidence can have severe consequences.

Moreover, the calibration and validity plots confirm that the conformal sets are not only theoretically sound but also empirically well-calibrated—validating the approach across multiple random seeds and confidence levels.

7.2 Future Scope

- **Scaling to Complex Datasets:** Future work can apply the proposed methods to more challenging and larger datasets like CIFAR-100 and ImageNet to assess generalization and scalability.

- **Extension to Regression and Multi-label Tasks:** Conformal methods can be extended beyond classification to regression and structured output prediction. In particular, applications in medical imaging or scientific computing could benefit from such extensions.

- **Conditional Coverage Guarantees:** While current methods ensure marginal coverage, developing approaches that guarantee conditional coverage (e.g., per class or per instance type) is an important and active area of research.

- **Real-time Inference Optimization:** Optimizing inference time and memory usage will be crucial for real-time deployment in edge computing environments such as drones, IoT devices, and mobile phones.

- **Fairness and Robustness Analysis:** Investigating how conformal methods behave under distribution shifts, class imbalance, or adversarial attacks is critical for building trustworthy AI systems.

8 Reference & Links

- **Reference Book:**
  - Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. *Theoretical Foundations of Conformal Prediction.* 2021. Available at: https://arxiv.org/abs/2110.07858
- **Reference Paper:**
  - Anastasios N. Angelopoulos and Stephen Bates. *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification.* 2021. Available at: https://arxiv.org/pdf/2107.07511
- **Reference Link:**
  - https://github.com/scikit-learn-contrib/MAPIE/blob/master/notebooks/classification/Cifar10.ipynb
- **Google Colab Link:** Click here