Distribution-Free Predictive Inference for Classification and Time Series

1 Introduction

1.1 Limitations of Traditional Prediction Methods

Traditional statistical forecasting is often reported as a single point estimate, and its uncertainty quantification is frequently fragile. Obtaining a valid prediction interval usually relies on strong assumptions—such as correct model specification, homoscedastic and approximately normal errors, and asymptotic approximations.

However, in many real-world applications, these conditions fail, so the resulting intervals can be poorly calibrated. This problem is even worse in the modern machine learning, where the most powerful models used are black-box (e.g. random forests, neural networks, etc) and they do not provide a parametric error distribution, making it impossible to achieve reliable prediction intervals using traditional statistical tools.

1.2 Conformal Prediction and How It Works

Under this circumstance, a new prediction framework known as *conformal prediction* emerges. This method is model-agnostic, distribution-free, and capable of providing reliable prediction intervals. The only assumption required is the exchangeability of the data.

To construct a conformal prediction interval with target coverage $1 - \alpha$, the following steps are required:

1. **Data Splitting:** Divide the dataset into a training set and a calibration set.
2. **Model Training:** Fit any predictive model on the training set to obtain a prediction function $\hat{f}(x)$.
3. **Calibration Score:** Define a score function to measure the discrepancy between the true value and the prediction. The larger the discrepancy, the larger the score. Compute the nonconformity scores on the calibration set.

   For regression models, the score function is typically defined as the absolute residual:

   $$s(x_i, y_i) = r_i = |y_i - \hat{f}(x_i)|, \quad i = 1, \dots, m. \tag{1}$$

   For classification tasks, a common choice is:

   $$s(x_i, y_i) = 1 - p_{\hat{f}}(y_i \mid x_i), \quad i = 1, \dots, m. \tag{2}$$

4. **Quantile Computation:** Compute the $(1 - \alpha)$ quantile of the calibration scores:

   $$\hat{q} = \text{Quantile}_{\lceil (n+1)(1-\alpha) \rceil / n} \left(s_1, \dots, s_n \right). \tag{3}$$

5. **Prediction Set Construction:** For a new input x_{new}, construct the conformal prediction set:

   $$\mathcal{T}(x_{\text{new}}) = \left\{ y : s(x_{\text{new}}, y) \leq \hat{q} \right\}. \tag{4}$$

1.3 Theoretical Foundation: Exchangeability

The validity of conformal prediction relies on the assumption of *exchangeability.* If the data points are i.i.d. or simply exchangeable, then the residuals observed on the calibration set have the same distribution as the residual associated with a new, unseen observation. Consequently, the quantile computed from calibration residuals provides a valid error bound for future predictions.

This leads to strong theoretical guarantees: conformal prediction is distribution-free, model-agnostic, and valid in finite samples, making it a robust and practical method for uncertainty quantification in forecasting tasks.

2 Distribution-Free Risk Control for Multi-Label Classification

2.1 Motivation: From Coverage to Risk Constraints

Classical conformal prediction (CP) is most often introduced through a *coverage* guarantee: prediction sets are constructed so that the probability of containing the true label is at least $1-\alpha$ in finite samples (under exchangeability). While this is appealing, coverage alone does not directly capture the operational risks that matter in many decision-making pipelines. In multi-label medical screening, for example, a method that returns overly large sets can trivially achieve high coverage but may overwhelm downstream review and increase false alarms.

Distribution-Free Risk Control (DFRC) generalizes conformal ideas by replacing "coverage" with an *arbitrary bounded risk* $R(\lambda)$, and aims to select a parameter λ such that

$$P(R(\lambda) \leq \alpha) \geq 1-\delta,$$

where α is a target risk tolerance and δ is an allowable failure probability. This framework accommodates a wide range of objectives beyond coverage, including false discovery rate (FDR), false negative rate (FNR), intersection-over-union losses, and fairness-related risks. In contrast to standard CP (which typically targets marginal coverage directly), DFRC emphasizes a probabilistic statement about *risk control* that holds in finite samples and remains distribution-free.

2.2 Methodology: Learn-Then-Test (LTT) for FDR Control

We instantiate DFRC for multi-label classification by thresholding the output probabilities of an upstream neural classifier. Given an input image x and predicted class probabilities $\hat{p}_k(x)$ for labels $k \in \{1,\ldots,K\}$, we define a family of prediction sets indexed by a threshold λ:

$$T_\lambda(x) = \{k : \hat{p}_k(x) > \lambda\}.$$

To control false discoveries, we measure set quality using a false-discovery-proportion (FDP) style loss. For a ground-truth label set Y and a predicted set T, the loss is

$$L_{\text{FDP}}(T,Y) = \begin{cases} 1 - \frac{|T\cap Y|}{|T|}, & |T| > 0, \\ 0, & |T| = 0. \end{cases}$$

The corresponding DFRC risk target is the expectation of this bounded loss:

$$R_{\text{FDR}}(\lambda) = E[L_{\text{FDP}}(T_\lambda(X), Y)]\,,$$

which plays the role of an "expected FDP" and is aligned with the notion of FDR in this setting.

To select λ with statistical safety guarantees, we use the Learn-Then-Test (LTT) procedure. Concretely, we evaluate a grid Λ of candidate thresholds and, on a calibration split, compute the

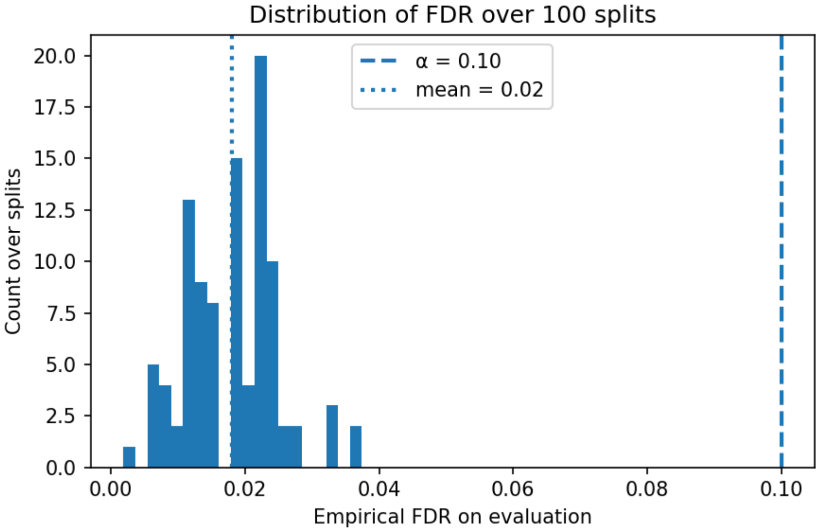

Figure 1: Distribution of empirical FDR over 100 calibration/evaluation splits.

empirical risk $\hat{R}(\lambda)$. A Hoeffding-type bound yields a conservative p-value for testing whether $R(\lambda) \leq \alpha$. Because many thresholds are tested, we apply a Bonferroni correction over the grid to control family-wise error across Λ. The output of LTT is a *safe set* of thresholds that pass the corrected test; we then select the least conservative threshold among those deemed safe (in our implementation, the smallest valid λ, which tends to maximize set size subject to risk control).

2.3 Experimental Setup

We validate DFRC on top of an upstream NIH Chest X-ray multi-label classifier (ResNet-50 backbone, trained with focal loss). Our subset contains 7,550 images across 15 labels. On this data, the base model achieves an average ROC–AUC of about 0.739, and we use its sigmoid probability outputs to form the DFRC prediction sets $T_\lambda(x)$.

For DFRC evaluation, we follow a repeated split protocol on a held-out pool. Each run randomly partitions the pool into calibration and evaluation halves; in our case this corresponds to 534 calibration images and 534 evaluation images per split. We report results over multiple reshuffles (e.g., 100 splits) to check that the empirical violation frequency is consistent with the theoretical δ level. We use $\alpha = 0.10$, $\delta = 0.10$, and a uniform grid of $|\Lambda| = 101$ thresholds in $[0, 1]$, with Bonferroni adjustment across the grid.

Empirical stability and the role of Bonferroni correction. Figure 1 reports the empirical FDR on the held-out evaluation subset across 100 independent calibration/evaluation reshuffles.

The distribution is tightly concentrated near zero: most splits lie in the $[0.01, 0.03]$ range, and the empirical mean is about 0.02 (vertical dotted line). Crucially, we do not observe any split exceeding the target level $\alpha = 0.10$ (vertical dashed line), which is consistent with the DFRC/LTT objective that violations should occur in at most a $\delta = 0.10$ fraction of random splits. Because our risk is defined via the false discovery proportion, $\mathrm{FDP}(T, Y) = 1 - \frac{|T \cap Y|}{|T|}$ (with the standard convention when $|T| = 0$), this guarantee has a direct precision-style interpretation: with probability at least $1 - \delta = 0.90$ over the calibration randomness and the LTT selection step,

$$E\left[\frac{|T_{\lambda(X)\cap Y|}}{|T_{\lambda(X)|}}\right] \geq 1 - \alpha = 0.90.$$

In practice, the histogram suggests substantially stronger behavior on this dataset (mean FDR ≈ 0.02, i.e., precision close to 0.98 on average), but this is an empirical observation rather than a worst-case claim.

A key driver of the observed stability is the Bonferroni adjustment applied over the discretized λ-grid (multiple tests over candidate thresholds). By controlling the family-wise error rate, Bonferroni reduces the chance of selecting an overly permissive λ due to calibration noise, at the cost of increased conservatism. In our runs, this manifests as a relatively high selected threshold (approximately $\lambda^{\approx 0.75}$), which shrinks the prediction sets $T_\lambda(x) = \{k : \hat{p}_k(x) > \lambda\}$ and further suppresses false discoveries—potentially trading off sensitivity/recall for rare pathologies.

Conclusion. This experiment demonstrates how DFRC extends conformal principles from coverage guarantees to broader risk-control objectives. Using LTT with Bonferroni adjustment, we selected $\lambda^{\approx 0.75}$ for multi-label prediction sets and achieved the desired reliability behavior: FDR above 10% occurs in at most 10% of reshuffles, consistent with $(\alpha, \delta) = (0.10, 0.10)$. The result provides an interpretable operational guarantee—precision at least 90% with probability at least 90%—which is well-suited to domains where false positives are more damaging than missed detections.

3 Conformal Prediction for Time Series

3.1 Problem Formulation: Non-Exchangeability

As delineated previously, the theoretical validity of conformal prediction relies on the assumption of data exchangeability. This assumption is inherently violated in time series forecasting, where observations exhibit serial correlation and heteroscedasticity, commonly manifested as volatility clustering. Standard approaches that approximate the marginal distribution of residuals, such as the industry baseline Ensemble Bootstrap Prediction Intervals (EnbPI), often fail to capture these temporal dynamics. By relying on a static empirical distribution of past errors, EnbPI produces conservative, "step-like" prediction intervals that adapt slowly to regime shifts.

3.2 Methodology: Sequential Predictive Conformal Inference (SPCI)

To address non-exchangeability, we implement the Sequential Predictive Conformal Inference (SPCI) framework, which reformulates interval construction as a supervised learning problem. Unlike EnbPI, SPCI employs Conditional Quantile Regression, implemented via a Quantile Random Forest (QRF), to estimate the conditional distribution of the next-step residual given the history of past errors. This design enables dynamic re-estimation of interval widths at each time step. By leveraging the feedback structure inherent in sequential prediction, SPCI achieves context-aware uncertainty quantification that adapts to evolving temporal regimes.

3.3 Experimental Setup

We evaluate the efficacy of SPCI against the EnbPI baseline across three datasets representing distinct residual structures: Australian Electricity (high volatility), Electricity Transformer (periodic and structured), and Air Quality (stochastic noise). All experiments enforce a strict validity constraint of 90% target coverage, i.e., $1 - \alpha = 0.9$. Predictive efficiency, measured by the average prediction interval width, serves as the primary performance metric.

3.4 Empirical Results and Discussion

The empirical results are summarized in Table 1. Overall, SPCI demonstrates superior efficiency in regimes characterized by strong temporal dependence, but exhibits limitations in predominantly stochastic environments.

Table 1: Coverage and Efficiency Comparison Between EnbPI and SPCI

Dataset	Metric	EnbPI (Baseline)	SPCI (Ours)
Australian Electricity	Coverage	91.00%	93.32%
	Avg. Width	0.32	0.22
Transformer Data	Coverage	88.65%	90.00%
	Avg. Width	0.88	0.81
Air Quality	Coverage	92.77%	84.94%
	Avg. Width	5.41	5.43

In the Australian Electricity Market, SPCI achieves a 31% reduction in average interval width (0.22 vs. 0.32) while maintaining higher empirical coverage (93.32%) compared to the baseline. This result confirms the model's ability to exploit volatility clustering for sharper uncertainty estimation. In contrast, SPCI fails to satisfy the validity constraint on the Air Quality dataset, achieving only 84.94% coverage.

Conclusion The divergent performance across datasets can be attributed to differences in residual structure. SPCI relies on predictable temporal dependence to learn the conditional distribution of residuals. However, the Air Quality residuals exhibit independent but non-identically distributed (i.n.i.d.) behavior with weak serial correlation. In the absence of exploitable temporal structure, the supervised learning component struggles to generalize, resulting in under-coverage. Consequently, while SPCI substantially improves efficiency in structured time series, careful validation is required in stochastic regimes to avoid violations of coverage guarantees.

4 Literature

1. Distribution-Free Predictive Inference for Regression, J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, 2017

2. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, A. N. Angelopoulos and S. Bates, 2022

3. Sequential Predictive Conformal Inference for Time Series, by Chen Xu, Yao Xie, 2023.