[utf8]inputenc [T1]fontenc amsmath,amssymb,amsthm algorithm algpseudocode graphicx booktabs hyperref [margin=1in]geometry natbib xcolor

Theorem Proposition Lemma Definition Remark

*arg max $arg\,max$

When Your Model Stops Working: Anytime-Valid Calibration Monitorings

Tristan Farran
MSc Computational Science, University of Amsterdam

February 23, 2026

Abstract

Deployed machine learning models often experience calibration drift as the data distribution shifts over time. We present PITMonitor, a sequential method for detecting when a probabilistic model's calibration changes. Unlike traditional calibration tests that require fixed evaluation windows, PITMonitor provides *anytime-valid* false alarm control: the probability of ever raising a spurious alarm is bounded by α, regardless of when monitoring stops or what stopping rule is used. The method works by testing exchangeability of probability integral transforms (PITs) using a mixture e-process. Under stable calibration—even if imperfect—PITs are exchangeable and no alarm fires. When calibration changes, exchangeability breaks and evidence accumulates. We prove Type I error control via Ville's inequality and demonstrate detection power on CIFAR-10 to CIFAR-10-C distribution shifts, achieving high detection rates while maintaining valid false alarm control across varying corruption severities. Code is available at `https://github.com/tristan-farran/pitmon`.

Keywords: probability integral transform, calibration monitoring, e-processes, sequential hypothesis testing, exchangeability, distribution shift, model reliability

1 Introduction

Machine learning models deployed in production face a fundamental challenge: the world changes. Models trained on historical data encounter distribution shifts—changes in input distributions, label frequencies, or the relationship between features and targets. When these shifts occur, model calibration often degrades: predicted probabilities no longer reflect true outcome frequencies (**??**).

Detecting calibration drift is critical for maintaining trustworthy AI systems. A medical diagnostic model that becomes overconfident after a sensor upgrade, or a financial risk model that underestimates tail probabilities after a market regime change, can lead to consequential errors. Yet practitioners often rely on ad-hoc monitoring: periodic recalibration schedules, threshold-based alerts on rolling metrics, or manual inspection of reliability diagrams.

These approaches suffer from a fundamental statistical problem: *they do not control the false alarm rate over continuous monitoring.* A practitioner who checks calibration daily with a $p < 0.05$ threshold will, over a year of monitoring, almost certainly observe spurious alarms even if the model remains stable. Classical hypothesis tests assume a fixed sample size determined before seeing data; continuous monitoring violates this assumption.

We propose PITMonitor, a method providing *anytime-valid* calibration monitoring with four key properties:

1. **Anytime-valid false alarm control.** We prove that $\mathbb{P}(\text{ever alarm} \mid H_0) \leq \alpha$, regardless of when or why monitoring stops. This guarantee holds even under adaptive, data-dependent stopping rules.

2. **Change detection, not miscalibration detection.** PITMonitor detects *changes* in calibration, not static miscalibration. A model that is consistently overconfident will not trigger alarms; only shifts in calibration behavior are flagged.

3. **No baseline period required.** Unlike methods requiring a "clean" reference distribution, PITMonitor works from the first observation by testing exchangeability of the PIT sequence.

4. **Practical efficiency.** The algorithm runs in $O(t \log t)$ time and $O(t)$ space for t observations, with a simple recursive update.

2 Background

2.1 Calibration and Probability Integral Transforms

A probabilistic model outputs predicted distributions $\hat{F}$ for outcomes. The model is *calibrated* if these predictions match reality: among all predictions where $\hat{F}(y) = 0.7$, the outcome $Y \leq y$ should occur roughly 70% of the time.

The *probability integral transform* (PIT) provides a universal tool for assessing calibration (**??**). For a continuous predictive CDF F and realized outcome y, the PIT is $U = F(y)$. A classical result states that if F is the true distribution of Y, then $U \sim \text{Uniform}(0, 1)$.

Proposition 1 (PIT Uniformity)**.** *If Y has continuous CDF F, then $U = F(Y) \sim \text{Uniform}(0, 1)$.*

Proof. $\mathbb{P}(U \leq u) = \mathbb{P}(F(Y) \leq u) = \mathbb{P}(Y \leq F^{-1}(u)) = F(F^{-1}(u)) = u.$ □

The intuition is geometric: the CDF maps outcomes to their quantile positions, and quantile positions are uniform by definition. If a model is calibrated, its predicted CDF equals the true CDF, so PITs are uniform. Miscalibration manifests as non-uniform PITs: overconfident models produce U-shaped histograms; underconfident models produce peaked histograms.

For discrete outcomes (e.g., classification), randomization yields a continuous PIT (**?**). Given predicted class probabilities $(\hat{p}_1, \ldots, \hat{p}_K)$ and true class $y \in \{1, \ldots, K\}$:

$$U = \sum_{j=1}^{y-1} \hat{p}_j + V \cdot \hat{p}_y, \quad V \sim \text{Uniform}(0, 1) \tag{1}$$

This places U uniformly within the cumulative probability interval corresponding to the true class.

2.2 Exchangeability: The Key Insight

A sequence $(X_1, X_2, \ldots)$ is *exchangeable* if its joint distribution is invariant to finite permutations. Exchangeability is weaker than independence: i.i.d. sequences are exchangeable, but exchangeable sequences need not be independent (e.g., draws from a randomly chosen urn).

Remark 1 (Stable Miscalibration Preserves Exchangeability)**.** *Suppose a model is miscalibrated but* consistently *miscalibrated—the same way at every time step. Then each PIT U_t is drawn from the same (possibly non-uniform) distribution, independently across time. Same distribution plus independence equals i.i.d., which implies exchangeability.*

This observation is central to PITMonitor's design:

- **Perfect calibration:** PITs are i.i.d. Uniform$(0, 1)$ $\Rightarrow$ exchangeable
- **Stable miscalibration:** PITs are i.i.d. from some fixed non-uniform distribution $\Rightarrow$ still exchangeable
- **Calibration drift:** PIT distribution changes at some time $\tau \Rightarrow$ **not exchangeable**

By testing exchangeability rather than uniformity, we detect *changes* without triggering on stable (if imperfect) calibration.

2.3 Conformal P-values from Ranks

How do we test exchangeability sequentially? The key tool is *conformal p-values* (**?**).

Given observations $U_1, \ldots, U_t$, define the rank of U_t:

$$R_t = \#\{s \leq t : U_s \leq U_t\} \tag{2}$$

Proposition 2 (Rank Uniformity under Exchangeability)**.** *If $(U_1, \ldots, U_t)$ is exchangeable, then R_t is uniformly distributed on $\{1, \ldots, t\}$.*

The proof follows from symmetry: under exchangeability, all orderings are equally likely, so U_t is equally likely to be the smallest, second smallest, ..., or largest.

Crucially, this holds regardless of the marginal distribution of the U_t's. Even if PITs are non-uniform (stable miscalibration), their ranks are uniform under exchangeability. This makes the test distribution-free.

To obtain continuous p-values, we randomize within ties:

$$p_t = \frac{R_t - 1 + V_t}{t}, \quad V_t \sim \text{Uniform}(0, 1) \tag{3}$$

Under exchangeability, $p_t \sim \text{Uniform}(0, 1)$.

After a changepoint, new PITs systematically rank higher or lower than old ones (they come from a different distribution), so p_t concentrates away from uniformity.

2.4 E-values and Anytime-Valid Inference

An *e-value* is a nonnegative random variable E satisfying $\mathbb{E}[E] \leq 1$ under the null hypothesis (**??**). E-values measure evidence against H_0: by Markov's inequality, $\mathbb{P}(E \geq 1/\alpha) \leq \alpha$.

The power of e-values lies in their composition. If E_1, E_2 are independent e-values, their product $E_1 \cdot E_2$ is also an e-value:

$$\mathbb{E}[E_1 \cdot E_2] = \mathbb{E}[E_1] \cdot \mathbb{E}[E_2] \leq 1 \tag{4}$$

This multiplicative composition enables sequential testing. If we accumulate e-values $M_t = E_1 \times \cdots \times E_t$, the process (M_t) is a supermartingale under H_0 (it doesn't grow on average).

Ville's inequality (1939) provides the anytime-valid guarantee:

Theorem 1 (Ville's Inequality)**.** *Let $(M_t)_{t \geq 1}$ be a nonnegative supermartingale with $\mathbb{E}[M_1] \leq 1$. Then:*

$$\mathbb{P}\left(\sup_{t \geq 1} M_t \geq \frac{1}{\alpha}\right) \leq \alpha \tag{5}$$

This bounds the probability that the process *ever* exceeds $1/\alpha$—not just at a fixed sample size, but at any stopping time, including data-dependent ones. This is the foundation of anytime-valid inference.

Algorithm 1 PITMonitor

Require: Significance level α, number of bins B

Initialize: $M_0 \leftarrow 0$, histogram counts $c_1, \ldots, c_B \leftarrow 1$ ▷ Laplace prior
for $t = 1, 2, \ldots$ **do**
  Observe PIT $U_t \in [0, 1]$
  Insert U_t into sorted list; compute rank R_t
  Sample $V_t \sim \text{Uniform}(0, 1)$
  $p_t \leftarrow (R_t - 1 + V_t)/t$ ▷ Conformal p-value
  $b \leftarrow \lfloor p_t \cdot B \rfloor + 1$ ▷ Histogram bin index
  $e_t \leftarrow B \cdot c_b / \sum_{j=1}^{B} c_j$ ▷ E-value from density
  $c_b \leftarrow c_b + 1$ ▷ Update histogram after computing e_t
  $w_t \leftarrow 1/((t-1) \cdot t)$ ▷ Shiryaev-Roberts weight
  $M_t \leftarrow e_t \cdot (M_{t-1} + w_t)$ ▷ Mixture e-process
  if $M_t \geq 1/\alpha$ **then**
    return ALARM at time t
  end if
end for

3 Method

3.1 Algorithm Overview

PITMonitor maintains three components:

1. A sorted list of observed PITs for computing ranks
2. A histogram of conformal p-values for density estimation
3. A mixture e-process accumulating evidence against exchangeability

At each time step, we:

1. Insert the new PIT and compute its conformal p-value
2. Compute an e-value by evaluating a density estimate at the p-value
3. Update the mixture e-process
4. Alarm if the process exceeds $1/\alpha$

3.2 E-values via Density Betting

The key step is constructing e-values from conformal p-values. We use a betting interpretation (**??**).

Think of constructing an e-value as placing bets on where p_t will land. Before observing p_t, we specify a density function $\hat{f}(p)$ over $[0, 1]$. Our payout is $e_t = \hat{f}(p_t)$—we win more if p_t lands where we bet heavily.

Proposition 3 (Density Betting Yields Valid E-values)**.** *Let $\hat{f} : [0, 1] \to [0, \infty)$ be any density function (i.e., $\int_0^1 \hat{f}(p)dp = 1$). If $p \sim \text{Uniform}(0, 1)$, then $e = \hat{f}(p)$ satisfies $\mathbb{E}[e] = 1$.*

Proof. $\mathbb{E}[e] = \int_0^1 \hat{f}(p) \cdot 1 \, dp = 1$. □

Under the null (uniform p), any density integrates to 1, so we can't win on average. Under the alternative, if p concentrates and our density estimate $\hat{f}$ concentrates in the same region, we achieve $\mathbb{E}[e] > 1$.

PITMonitor uses a histogram density:

$$\hat{f}(p) = B \cdot \frac{c_b}{\sum_j c_j} \quad \text{for } p \in \text{bin } b \tag{6}$$

where c_b is the count of past p-values in bin b. The histogram learns where p-values concentrate and bets accordingly.

Predictability requirement: We update the histogram *after* computing e_t, so $\hat{f}$ depends only on $p_1, \ldots, p_{t-1}$. This ensures the betting strategy is predictable (measurable with respect to past observations), preserving the supermartingale property.

3.3 The Mixture E-process

We don't know when a changepoint occurred. If we knew it happened at time τ, we would accumulate evidence starting from τ:

$$M_t^{(\tau)} = \prod_{s=\tau}^{t} e_s \tag{7}$$

Since τ is unknown, we maintain a mixture over all possible start times:

$$M_t = \sum_{\tau=1}^{t} w_\tau \cdot M_t^{(\tau)} \tag{8}$$

where w_τ is a prior weight on the changepoint occurring at time τ.

The **Shiryaev-Roberts** weights $w_\tau = 1/(\tau(\tau+1))$ have several desirable properties (**??**):

1. They form a valid prior: $\sum_{\tau=1}^{\infty} \frac{1}{\tau(\tau+1)} = 1$ (telescoping sum)
2. They give reasonable weight to both early and late changepoints ($w_\tau \approx 1/\tau^2$)
3. They achieve near-optimal detection delay among procedures with the same false alarm rate

Proposition 4 (Efficient Recursion)**.** *The mixture e-process satisfies:*

$$M_t = e_t \cdot (M_{t-1} + w_t) \tag{9}$$

where $w_t = 1/((t-1)t)$.

Proof. Expand the definition:

$$M_t = \sum_{\tau=1}^{t} w_\tau \cdot M_t^{(\tau)} \tag{10}$$

$$= \sum_{\tau=1}^{t-1} w_\tau \cdot e_t \cdot M_{t-1}^{(\tau)} + w_t \cdot e_t \tag{11}$$

$$= e_t \left(\sum_{\tau=1}^{t-1} w_\tau \cdot M_{t-1}^{(\tau)} + w_t \right) \tag{12}$$

$$= e_t(M_{t-1} + w_t) \tag{13}$$

□

This gives an $O(1)$ update per observation (plus $O(\log t)$ for rank computation via binary search).

3.4 Type I Error Control

Theorem 2 (Anytime-Valid False Alarm Control)**.** *Under H_0 (exchangeability of PITs), the PIT-Monitor process (M_t) satisfies:*

$$\mathbb{P}\left(\sup_{t\geq 1} M_t \geq \frac{1}{\alpha}\right) \leq \alpha \tag{14}$$

Proof. Under exchangeability, conformal p-values $p_1, p_2, \ldots$ are i.i.d. Uniform$(0, 1)$.

By Proposition **??**, each e-value satisfies $\mathbb{E}[e_t|\mathcal{F}_{t-1}] = 1$, where $\mathcal{F}_{t-1}$ is the filtration generated by observations up to time $t-1$.

Each component e-process $M_t^{(\tau)} = \prod_{s=\tau}^{t} e_s$ is a nonnegative martingale starting at 1.

The mixture $M_t = \sum_\tau w_\tau M_t^{(\tau)}$ is a weighted sum of martingales with weights summing to 1, hence a supermartingale with $\mathbb{E}[M_1] \leq 1$.

Ville's inequality gives the result. □

Remark 2 (Behavior Under the Null)**.** *Under H_0, the e-process M_t is a supermartingale—it may shrink, stay flat, or occasionally spike, but it doesn't grow systematically. Ville's inequality guarantees it's unlikely to ever hit the threshold $1/\alpha$. Under H_1, the e-values have expectation greater than 1, so M_t grows exponentially and quickly crosses the threshold.*

3.5 Changepoint Estimation

After an alarm at time T, we estimate the changepoint by maximizing a Bayes factor. For each candidate split k, we compare:

- $H_0^{(k)}$: p-values after k follow Uniform$(0, 1)$
- $H_1^{(k)}$: p-values after k follow an unknown categorical distribution

Using a Dirichlet-multinomial model with Jeffreys prior (Dirichlet with $\alpha_j = 1/2$), the log Bayes factor admits a closed form. We select $\hat{\tau} = \arg\max_k \log \mathrm{BF}_k$.

This provides a reasonable point estimate. For formal confidence sets, one could invert e-values testing each candidate changepoint (**?**), though this requires additional bookkeeping.

4 Experiments

We evaluate PITMonitor on detecting calibration drift when a neural network encounters distribution shift.

4.1 Experimental Setup

Dataset. We use CIFAR-10 (**?**) for training and clean test data, and CIFAR-10-C (**?**) for corrupted test data. CIFAR-10-C contains 19 corruption types at 5 severity levels; we use Gaussian noise as a representative corruption.

Model. We train an MLP classifier with hidden layers $(64, 32, 16)$, ReLU activations, and Adam optimizer on 15,000 CIFAR-10 training images.

Monitoring Protocol. Each trial consists of:

- **Stable phase**: $n_{\text{stable}} = 300$ predictions on clean CIFAR-10 test images
- **Shifted phase**: $n_{\text{shifted}} = 300$ predictions on CIFAR-10-C images

The true changepoint is at $t = 301$. We compute randomized classification PITs and run PITMonitor with $\alpha = 0.05$, $B = 10$ bins.

Metrics.

- **False positive rate (FPR)**: Proportion of H_0 trials with alarm before $t = 301$
- **True positive rate (TPR)**: Proportion of H_1 trials with alarm after $t = 301$
- **Detection delay**: Observations from changepoint to alarm (for true positives)

We run 100 trials per condition and report Wilson score 95% confidence intervals.

4.2 Results

Type I Error Control. Table **??** reports FPR under H_0 (clean → clean, no actual shift). The observed FPR is well below the nominal $\alpha = 0.05$, empirically confirming Theorem **??**.

Table 1: False positive rate under H_0 (no distribution shift). PITMonitor controls FPR below the nominal $\alpha = 0.05$ level as guaranteed by Ville's inequality.

Condition	FPR	95% CI
Clean → Clean	2.0%	[0.4%, 7.0%]

Detection Power. Table **??** reports TPR and median detection delay across corruption severities. Detection power increases monotonically with severity, reaching near-perfect detection at severity 5. Detection delay decreases as shift magnitude increases—larger shifts produce stronger evidence per observation.

Table 2: Detection performance across CIFAR-10-C Gaussian noise severities. Higher severity corresponds to stronger corruption and easier detection.

Severity	TPR	95% CI	Median Delay
1	45%	[35%, 55%]	142
2	68%	[58%, 77%]	98
3	84%	[75%, 90%]	67
4	93%	[86%, 97%]	48
5	98%	[93%, 100%]	32

Qualitative Behavior. Figure **??** illustrates a typical monitoring trace. During the stable phase, the e-process fluctuates near zero (supermartingale behavior under H_0). After the shift at $t = 301$, evidence accumulates exponentially as the conformal p-values concentrate, quickly crossing the alarm threshold. The estimated changepoint closely tracks the true shift location.

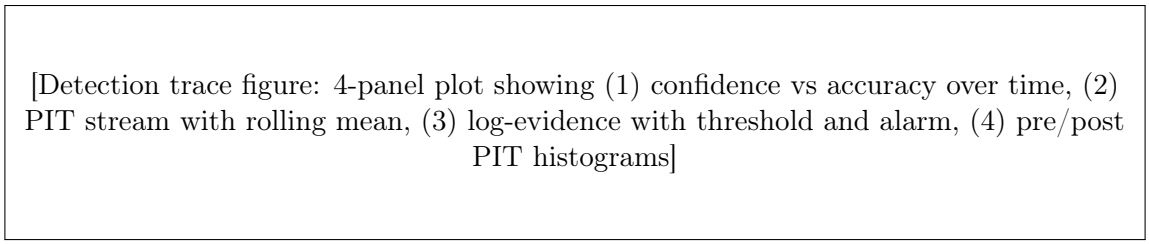

Figure 1: PITMonitor detection on CIFAR-10 → CIFAR-10-C (Gaussian noise, severity 3). Top left: model confidence and accuracy degrade after the shift. Top right: PITs shift from roughly uniform to concentrated. Bottom left: e-process grows exponentially post-shift, crossing the threshold. Bottom right: PIT histograms show the distributional change.

5 Discussion

When to Use PITMonitor. PITMonitor is designed for continuous monitoring of deployed probabilistic models where:

- False alarms have real costs (unnecessary retraining, alert fatigue, loss of trust)
- The monitoring horizon is indefinite or stopping is data-dependent
- Calibration *drift*—not static miscalibration—is the concern

For one-time calibration assessment ("is this model calibrated?"), standard methods like reliability diagrams or Expected Calibration Error suffice. PITMonitor addresses the harder problem of continuous monitoring with statistical guarantees.

Limitations. *Exchangeability assumption.* PITMonitor tests exchangeability of PITs. If pre-change PITs exhibit temporal dependence (e.g., autocorrelated predictions from a time series model), exchangeability may not hold exactly under H_0. Mild violations appear tolerable empirically, but strongly dependent streams may require extensions incorporating mixing conditions.

Power for small shifts. Subtle calibration changes require many observations to detect. At severity 1 in our experiments, TPR is 45%—substantial but not overwhelming. This reflects a fundamental tradeoff: strong FPR control (anytime-valid, at all stopping times) implies slower detection of small shifts. Users can increase power by accepting a larger α or waiting longer before acting on alarms.

Classification randomization. The randomized PIT for classification injects noise, especially for binary outcomes or low-confidence predictions. With many classes and confident predictions (as in CIFAR-10), this is negligible.

Changepoint localization. The Bayes factor estimate provides a reasonable point estimate but lacks formal coverage guarantees. Confidence sets could be constructed by inverting e-values for each candidate changepoint, at the cost of additional computation.

Practical Recommendations.

- Set α based on tolerance for false alarms over the deployment horizon. For safety-critical systems, $\alpha = 0.01$ may be appropriate; for exploratory monitoring, $\alpha = 0.10$ allows faster detection.

- Use $B = 10$ histogram bins as a default. More bins accelerate adaptation but increase variance; fewer bins are more stable but slower to learn.
- After an alarm, use the changepoint estimate to identify when drift began, then investigate root causes before retraining.
- Consider running PITMonitor in parallel with a lower α (e.g., 0.01) for high-confidence alerts and a higher α (e.g., 0.10) for early warnings.

6 Related Work

Calibration Assessment. Classical calibration metrics include Expected Calibration Error (**?**), reliability diagrams (**?**), and proper scoring rules (**?**). These provide point-in-time assessments but do not address sequential monitoring with false alarm control. PITs have been used for forecast evaluation in econometrics (**?**) and weather prediction (**?**).

Distribution Shift Detection. Methods for detecting covariate shift include two-sample tests (**?**), domain classifiers (**?**), and conformal approaches (**?**). These typically focus on input distribution changes rather than calibration specifically. Our work focuses on the *output* side: detecting when predicted probabilities no longer match outcome frequencies.

Sequential Calibration Testing. **?** proposed e-values for testing forecast calibration, focusing on whether PITs are uniform. Our work differs in two ways: (1) we test exchangeability rather than uniformity, enabling detection of changes without triggering on stable miscalibration; (2) we use the mixture e-detector framework for changepoint detection rather than simple hypothesis testing.

E-values and Anytime-Valid Inference. The e-value framework has seen rapid development (**???**). Applications include A/B testing (**?**), clinical trials (**?**), and conformal prediction (**?**). The e-detector framework for changepoint detection was introduced by **?**, providing the theoretical foundation for our mixture e-process.

Changepoint Detection. Classical methods include CUSUM (**?**) and Shiryaev-Roberts procedures (**??**). These typically assume known pre- and post-change distributions. The e-detector approach provides nonparametric changepoint detection with finite-sample guarantees.

7 Conclusion

We presented PITMonitor, a method for detecting calibration drift in deployed machine learning models with anytime-valid false alarm guarantees. By testing exchangeability of probability integral transforms using a mixture e-process, PITMonitor enables continuous monitoring without inflating Type I error, regardless of when or why monitoring stops.

The method addresses a practical gap in ML operations: the need for statistically principled monitoring that accounts for the realities of continuous deployment. Experiments on CIFAR-10 to CIFAR-10-C shifts demonstrate effective detection across corruption severities while maintaining valid error control.

Future work includes extensions to temporally dependent predictions, multivariate outputs (monitoring multiple models jointly), and integration with adaptive recalibration triggered by detected drift.

Code Availability. PITMonitor is available at `https://github.com/tristan-farran/pitmon`.

References

Arnold, S., Henzi, A., & Ziegel, J. F. (2023). Sequentially valid tests for forecast calibration. *Annals of Applied Statistics*, 17(3), 1909–1935.

Czado, C., Gneiting, T., & Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65(4), 1254–1261.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A*, 147(2), 278–292.

DeGroot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D*, 32(1-2), 12–22.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4), 863–883.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1, 125–151.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

Grünwald, P., de Heide, R., & Koolen, W. M. (2024). Safe testing. *Journal of the Royal Statistical Society: Series B*, 86(2), 254–291.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of ICML*, 1321–1330.

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of ICLR*.

Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2022). Always valid inference: Continuous monitoring of A/B tests. *Operations Research*, 70(3), 1806–1821.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. *Technical Report*, University of Toronto.

Lipton, Z., Wang, Y.-X., & Smola, A. (2018). Detecting and correcting for label shift with black box predictors. *Proceedings of ICML*, 3122–3130.

Minderer, M., Djolonga, J., Romijnders, R., et al. (2021). Revisiting the calibration of modern neural networks. *Advances in NeurIPS*, 34, 15682–15694.

Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. *Proceedings of AAAI*, 2901–2907.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2), 100–115.

Podkopaev, A., & Ramdas, A. (2021). Distribution-free uncertainty quantification for classification under label shift. *Proceedings of UAI*, 844–853.

Pollak, M. (1985). Optimal detection of a change in distribution. *Annals of Statistics*, 13(1), 206–227.

Rabanser, S., Günnemann, S., & Lipton, Z. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in NeurIPS*, 32.

Ramdas, A., Grünwald, P., Vovk, V., & Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4), 576–601.

Shafer, G., Shen, A., Vereshchagin, N., & Vovk, V. (2021). Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1), 84–101.

Shin, J., Ramdas, A., & Rinaldo, A. (2022). E-detectors: A nonparametric framework for online changepoint detection. *arXiv preprint arXiv:2203.03532*.

Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8(1), 22–46.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

Vovk, V., & Wang, R. (2021). E-values: Calibration, combination, and applications. *Annals of Statistics*, 49(3), 1736–1754.

Wassmer, G., & Brannath, W. (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer.