WHEN DRIFT DETECTORS CRY WOLF FALSE ALARM RATES IN CONTINUOUS ML MONITORING

Raj Shekhar Singh
Indian Institute of Technology, Roorkee
Roorkee, Uttarakhand, India
raj_ss@ece.iitr.ac.in

ABSTRACT

Drift detection is a core component of production machine learning monitoring systems, where detectors are used to compare incoming data with a reference distribution and trigger alerts when changes occur. However, these detectors are often evaluated in research settings that emphasize detection accuracy under synthetic shifts, while overlooking false alarms under continuous monitoring. In production environments, models are monitored repeatedly over time and across many features, and even small false positive rates can accumulate into frequent alerts, leading to alarm fatigue.

We empirically analyze false positive behavior across five commonly used drift detectors: PSI, KS, MMD, LSDD, and adversarial validation. Consistent with existing literature, PSI exhibits strong sensitivity to batch size, producing frequent false alarms at small sample sizes; however, we further observe that its behavior stabilizes and improves substantially once batch sizes exceed approximately 200 samples. In contrast, KS, MMD, and LSDD display persistent fluctuations across batch sizes, while remaining comparatively more reliable than PSI in low-data regimes. Applying a Bonferroni correction reduces false positive rates, but often at the cost of reduced true positive sensitivity, reinforcing the well-known stability–sensitivity trade-off in drift detection.

This work provides a systematic comparison of false positive behavior across multiple drift detectors under continuous monitoring conditions. We identify trade-offs across detector families and provide practical guidelines for selecting and calibrating drift detectors in production ML systems.

1 INTRODUCTION

Machine learning systems in production operate in environments that change over time. Shifts in user behavior, upstream pipelines, or data sources can alter the input distribution, leading to data drift and degraded model performance (Gama et al., 2014). To address this, production systems deploy drift detectors (Quionero-Candela et al., 2009; Lu et al., 2018) that compare incoming data with a reference distribution and trigger alerts when differences are detected. However, most detectors are evaluated in research settings (Rabanser et al., 2019) using synthetic shifts and single-shot tests, with emphasis on detection accuracy. In production, monitoring is continuous, models rely on many features, and alerts are generated repeatedly over time (Montiel et al., 2018). Under these conditions, even detectors with nominal false positive rates can produce frequent alerts, leading to alarm fatigue, where engineers begin to ignore or disable monitoring signals (Sculley et al., 2015). Despite its practical importance, this issue is rarely studied systematically. In this work, we aim to quantify how often drift detectors "cry wolf" under realistic continuous monitoring and understand the trade-offs between sensitivity and stability.

We conduct controlled experiments across five widely used detectors—PSI, KS, MMD, LSDD, and adversarial validation—using the Adult Income dataset under both no-drift and gradual-drift scenarios (Gama et al., 2014; Rabanser et al., 2019). We simulate 30-day monitoring cycles across multiple

batch sizes and drift magnitudes, and evaluate false alarm rates, true positive rates, and time to detection (Bifet & Gavaldà, 2018; Lu et al., 2018). Our results confirm that PSI is highly sensitive to small batch sizes, producing frequent false alarms below approximately 200 samples (Sidhu, 2021; Baesens et al., 2015), while its false positive behavior stabilizes noticeably at larger batch sizes. In contrast, statistical detectors exhibit more consistent performance in low-data regimes but show moderate and persistent false alarm fluctuations across batch sizes under standard thresholds (Benjamini & Hochberg, 1995; Montiel et al., 2018). We further observe that applying a Bonferroni correction substantially reduces false alarms, consistent with prior work, but often at the expense of detection sensitivity, highlighting a clear stability–sensitivity trade-off across methods (Dunn, 1961; Rabanser et al., 2019). These findings provide practical deployment guidelines and highlight configuration choices that materially affect monitoring reliability in production ML systems (Sculley et al., 2015).

2 Related Work

2.1 Drift Detection Methods

A wide range of drift detection methods have been proposed for monitoring distributional changes (Gama et al., 2014; Lu et al., 2018). Classical statistical tests such as the Kolmogorov–Smirnov (KS) test (Kolmogorov, 1933) and Population Stability Index (PSI) (Lewis, 1994) compare univariate feature distributions and are widely used in credit scoring and tabular ML pipelines (Baesens et al., 2015; Sidhu, 2021). Kernel-based methods extend this to multivariate settings by measuring distances in reproducing kernel Hilbert spaces, with Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) and Least-Squares Density Difference (LSDD) (Sugiyama et al., 2012) as representative examples. Adversarial detectors (Lopez-Paz & Oquab, 2016), where a classifier distinguishes reference from current data, have also gained popularity in production systems (Lipton et al., 2018). These statistical, kernel, and adversarial families form the core toolkit for modern ML monitoring (Rabanser et al., 2019).

2.2 Continuous Monitoring and Multiple Testing

Most drift detection research evaluates detectors on benchmark datasets or synthetic shifts, focusing on detection accuracy and time-to-detection at single change points (Bifet & Gavaldà, 2018; Montiel et al., 2018). However, production monitoring involves continuous testing across many features and time windows (Sculley et al., 2015). Under such conditions, even detectors with nominal false positive rates can generate frequent alerts due to random variation—a problem that has received limited empirical attention in the drift detection literature (Rabanser et al., 2019). The issue of repeated hypothesis testing is well studied in statistics under multiple testing frameworks (Benjamini & Hochberg, 1995). Methods like Bonferroni correction adjust significance thresholds to control family-wise error rates, though at the cost of reduced statistical power (Dunn, 1961). While standard in genomics and A/B testing, such corrections are rarely incorporated into drift detection pipelines, where detectors typically use default thresholds designed for single-shot experiments rather than continuous monitoring (Fanaee-T & Gama, 2014).

3 Experiments

3.1 Experimental Setup

We evaluate five drift detection methods on the Adult Income dataset (Dua & Graff, 2019), a standard tabular benchmark with mixed numerical and categorical features (Kohavi, 1996). A continuous monitoring scenario spanning 30 days is constructed using a fixed reference dataset and daily batches representing production data (Sculley et al., 2015). In the no-drift setting, daily batches are sampled from the reference distribution. In the drift setting, controlled shifts are injected into a single feature to simulate gradual distributional changes (Gama et al., 2014).

3.2 Detectors and Methods

Five detection methods spanning different methodological families are considered: Population Stability Index (PSI), a histogram-based univariate metric widely used in industry; Kolmogorov–Smirnov (KS) test, a classical non-parametric statistical test; two kernel-based multivariate detectors, Maximum Mean Discrepancy (MMD) and Least-Squares Density Difference (LSDD); and adversarial validation, where a classifier distinguishes reference from current data. Detector behavior is examined under two scenarios: (1) batch-size sweep from 50 to 500 samples (increments of 50), and (2) drift-span experiments with shift magnitudes of 5, 10, 15, and 20 years injected into the age feature. For statistical detectors, both standard and Bonferroni-corrected thresholds are evaluated to analyze multiple-testing correction effects. Each configuration is repeated across five random seeds.

3.3 Evaluation Metrics

Performance is evaluated using three operational metrics: (1) false positive days—the number of days (out of 30) with alarms in the no-drift scenario (Basseville & Nikiforov, 1993), (2) true positive rate (TPR) (Gama et al., 2004)—the fraction of drift-period days with correct detection, and (3) time to detection (TTD)—days between drift onset and first detection (Bifet & Gavaldà, 2018). Results are reported as mean and standard deviation across seeds.

Full experimental details are provided in Appendix A.

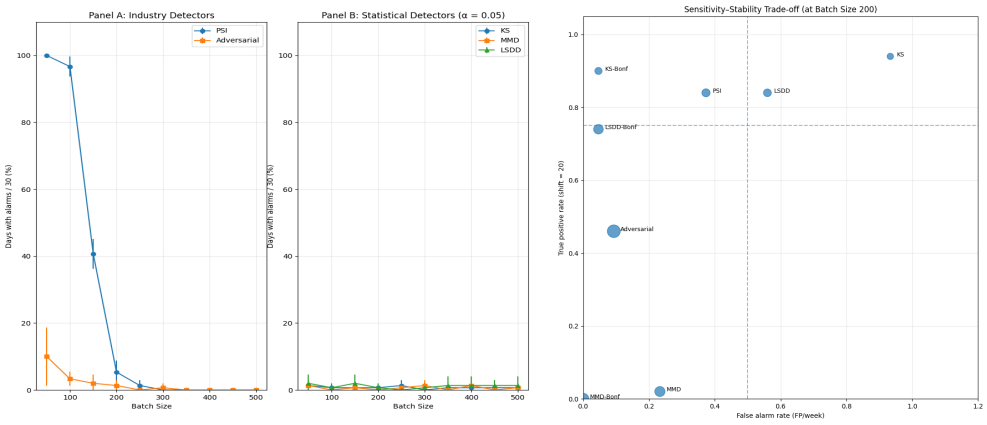

Figure 1: False alarm rates as a function of batch size. Panel A: Industry-style detectors (PSI and adversarial validation). PSI exhibits extreme false alarm rates at small batch sizes, approaching alarms on nearly all days below 200 samples, followed by a sharp stability transition. Panel B: Statistical detectors (KS, MMD, LSDD) at $\alpha = 0.05$. These methods maintain relatively low false alarm rates across batch sizes, indicating greater stability under continuous monitoring. Figure 2: Sensitivity–stability trade-off at batch size 200. Detectors are plotted by false alarm rate (x-axis) and true positive rate for a strong drift (y-axis), showing the trade-off between detection sensitivity and operational stability.

3.4 Results

Across all detectors, continuous monitoring reveals that false alarms are not a theoretical concern but a practical operational issue. In the no-drift setting, several detectors triggered alerts on multiple days despite the absence of any distributional change, confirming the presence of a "cry wolf" effect under realistic monitoring conditions.

PSI exhibited the most pronounced batch-size dependence, with alarm rates approaching daily triggers when fewer than approximately 200 samples were available, followed by a sharp reduction and stabilization as batch sizes increased. By comparison, KS and LSDD behaved more robustly in low-data settings but continued to show non-trivial false alarm variability across the full range of batch sizes. Adversarial validation demonstrated comparatively stable behavior, though with limited

Table 1: Summary of Detector Behavior and Practical Implications under Continuous Monitoring (Interpretation is based on experiments as well as literature review)

DETECTOR	KEY OBSERVATION	INTERPRETATION	PRACTICAL IMPLICATION
PSI	Extremely high false alarm rates at small batch sizes (≈30/30 days at batch 50–100), followed by a sharp drop after ∼200 samples and almost zero false rate for high batch size	PSI is highly sensitive to sampling noise when estimated from small batches due to its histogram-based formulation (Khademi et al., 2023; Yurdakul & Naranjo, 2018)	PSI would be desirable if batch sizes are sufficiently large (¿200 samples); otherwise, it may produce near-constant alarms
KS	Consistently low false alarm rates and strong detection performance as drift magnitude increases (TPR up to 0.90)	The KS test provides a favorable balance between stability and sensitivity under standard thresholds (Tonguz & Taschin, 2025)	KS serves as a reliable default detector for tabular monitoring with small to moderate batch sizes
LSDD	Moderate false alarm rates and lower sensitivity than KS for small shifts	Kernel density difference methods appear more conservative, trading sensitivity for stability under simple univariate drift (Bu et al., 2017)	Suitable in settings where fewer false alarms are preferred, at the cost of slower or weaker detection
MMD	Very low false alarm rates across batch sizes, but near-zero detection rates under the tested drift scenarios	Under the chosen kernel and drift configuration, MMD exhibited low sensitivity to the injected univariate drift (Zhou, 2025; Schrab et al., 2023)	Kernel-based multivariate tests may require careful kernel selection or tuning; default settings may not detect simple localized drift
Adversarial	False alarms decrease rapidly with increasing batch size, but detection occurs only for larger shifts (TPR up to 0.46)	Classifier-based detectors appear conservative, primarily detecting strongly separable distribution changes (Palli et al., 2022)	Appropriate when minimizing false alarms is critical, though subtle drift may go undetected
Bonferroni (correction)	Substantial reduction in false alarms across statistical detectors, typically from multi-day alarms to near-zero, accompanied by a drop in detection rates	Stricter significance thresholds reduce Type I errors (false alarms) but also lower sensitivity to small or gradual drift (Liu et al., 2022)	Effective when operational stability is prioritized; however, practitioners must account for reduced sensitivity and potential delays in drift detection

responsiveness to smaller distributional shifts. The introduction of a Bonferroni correction consistently suppressed false alarms for statistical detectors, but this came with a reduction in sensitivity, emphasizing the inherent stability–sensitivity trade-off across detectors and configurations.

More exploration on each detector is given in Table 1. Detailed quantitative results for all settings are provided in Appendix B.

4 Conclusion

This work examined the behavior of commonly used drift detectors under realistic continuous monitoring conditions. While such detectors are typically evaluated on benchmark shifts with an emphasis on detection accuracy, the experiments show that their behavior can differ substantially in production-like settings. Across five detectors and a range of batch sizes, several methods produced frequent alarms even in the absence of drift, revealing a practical "cry wolf" problem. In particular,

PSI was highly sensitive to small batch sizes, producing near-constant alarms below roughly 200 samples.

We also analyzed the trade-off between false alarms and detection sensitivity. Statistical detectors such as KS and LSDD achieved strong true positive rates but still generated non-trivial false alarm frequencies under standard thresholds. Applying a simple Bonferroni correction substantially reduced false positives, though at the cost of reduced sensitivity for some detectors, highlighting a clear sensitivity–stability trade-off across methods and configurations.

Overall, these results suggest that drift detector behavior in continuous monitoring scenarios requires careful calibration, and that standard research settings may not fully reflect production realities. Future work can extend this analysis to additional datasets, more complex drift patterns, alternative detector configurations, and longer monitoring horizons to further validate the generality of these findings.

REFERENCES

Bart Baesens et al. *Credit Risk Analytics*. Oxford University Press, 2015.

Michèle Basseville and Igor V. Nikiforov. Detection of abrupt changes: Theory and application. *Prentice Hall*, 1993.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate. *Journal of the Royal Statistical Society*, 1995.

Albert Bifet and Ricard Gavaldà. *Machine learning for data streams*. MIT Press, 2018.

Xia Bu et al. An incremental change detection test based on density difference estimation. *Machine Learning (preprint)*, 2017.

Dheeru Dua and Casey Graff. Uci machine learning repository: Adult data set. `https://archive.ics.uci.edu/ml/datasets/Adult`, 2019. Accessed: 2026-02-11.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 1961.

Hadi Fanaee-T and João Gama. A survey on concept drift adaptation. *ACM Computing Surveys*, 2014.

João Gama, Pedro Medas, Gladys Castillo, and Pedro Rodrigues. Learning with drift detection. In *Brazilian Symposium on Artificial Intelligence*, pp. 286–295, 2004.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 2014.

Arthur Gretton et al. A kernel two-sample test. *Journal of Machine Learning Research*, 2012.

Aria Khademi, Michael Hopka, and Devesh Upadhyay. Model monitoring and robustness of in-use machine learning models: Quantifying data distribution shifts using population stability index. *arXiv preprint arXiv:2302.00775*, 2023.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers. *KDD*, 1996.

A. N. Kolmogorov. On the empirical determination of a distribution law. *Giornale dell'Istituto Italiano degli Attuari*, 1933.

Edward M. Lewis. *An Introduction to Credit Scoring*. Athena Press, 1994.

Zachary Lipton et al. Detecting and correcting for label shift with black box predictors. In *ICML*, 2018.

Anjin Liu, Jie Lu, Yiliao Song, Junyu Xuan, and Guangquan Zhang. Concept drift detection delay index. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4585–4597, 2022.

D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

Jie Lu, Anjin Liu, Feng Dong, Feng Gu, João Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

Jacob Montiel et al. Adaptive drift detection methods for data streams. *Machine Learning*, 2018.

A.S. Palli, J. Jaafar, H.M. Gomes, M.A. Hashmani, and A.R. Gilal. An experimental analysis of drift detection methods on multi-class imbalanced data streams. *Applied Sciences*, 12(22):11688, 2022.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil Lawrence. *Dataset shift in machine learning*. MIT Press, 2009.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *NeurIPS*, 2019.

Antonin Schrab et al. Mmd aggregated two-sample test. In *Proceedings of the 2023 Conference on Learning Theory*, 2023.

D. Sculley et al. Hidden technical debt in machine learning systems. In *NeurIPS*, 2015.

Harpreet Sidhu. Population stability index: A practical guide. *Towards Data Science*, 2021.

Masashi Sugiyama et al. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Ozan K. Tonguz and Federico Taschin. Using kolmogorov-smirnov distance for measuring distribution shift in machine learning. *arXiv preprint arXiv:2510.15996*, 2025.

Bilal Yurdakul and Joshua Naranjo. Statistical properties of the population stability index. *Journal of Risk Model Validation*, 14(4):89–100, 2018.

L Peng C Lei Zhou, X Tian. Dual: Learning diverse kernels for aggregated two-sample and independence tests. *arXiv preprint arXiv:2510.11140*, 2025.

A Appendix A: Experimental Details

A.1 Dataset Preprocessing

Dataset: Adult Income

Source: UCI Machine Learning Repository

Missing values: removed

Categorical encoding: label encoding

Features used: all except target

Final dataset size: $\sim$30k samples

Number of features: 14

A.2 Monitoring Protocol

Monitoring duration: 30 days

Daily execution frequency: once per day

Reference dataset: fixed training distribution

Daily batch sampling: random sampling from dataset

No-drift experiment

- Daily batches drawn from reference distribution
- Used for false positive evaluation

Drift-span experiment

- Drift injected into age feature
- Drift begins after day 30
- Gradual drift over 10 days

A.3 Experimental Variables

Batch size sweep

Batch sizes evaluated:

$$50, 100, 150, 200, 250, 300, 350, 400, 450, 500$$

Drift span shifts

Injected shifts in age feature:

$$5, 10, 15, 20 \text{ years}$$

Random seeds

All experiments repeated with:

$$\text{Seeds: } 0, 1, 2, 3, 4$$

Results reported as:

- Mean
- Standard deviation

A.4 Detector Configurations

PSI

- Binning: 10 equal-frequency bins
- Threshold: 0.2 (standard industry setting)

KS Test

- Test: two-sample KS test
- Standard threshold: $\alpha = 0.05$
- Bonferroni-corrected threshold:

$$\alpha = \frac{0.05}{\text{number\_of\_features}}$$
$$\approx 0.007 \text{ for 14 features}$$

MMD

- Kernel: Gaussian (default)
- Reference subset size: 2000 samples
- Permutations: 100
- Standard threshold: $\alpha = 0.05$
- Bonferroni threshold: $\alpha \approx 0.007$

LSDD

- Kernel: Gaussian (default)
- Reference subset size: 2000 samples
- Permutations: 100
- Standard threshold: $\alpha = 0.05$
- Bonferroni threshold: $\alpha \approx 0.007$

Adversarial Detector

- Model: logistic regression
- Train/test split: reference vs current batch
- Metric: ROC-AUC
- Alarm threshold: AUC ≥ 0.6

A.5 Evaluation Metrics

False positive days

Definition: Number of days with alarm in no-drift setting

Range: 0 to 30 days

True positive rate (TPR)

Definition: Fraction of drift-period days with alarm

Computed only during drift window

Time to detection (TTD)

Definition: Days between drift start and first detection

Reported as: Mean $\pm$ standard deviation across seeds

A.6 Implementation Details

Programming language: Python

Libraries:

- NumPy
- Pandas
- scikit-learn
- alibi-detect
- evidently

Hardware:

- CPU-based experiments

Random number generator:

- NumPy default RNG

B Complete Experimental Tables

Table 2 and 3 — False positives vs. batch size (no-drift scenario)

False positives are measured as the number of days with alarms out of 30, reported as mean $\pm$ standard deviation across 5 seeds.

Table 2: False positives vs. batch size (50–200).

Detector	Variant	50	100	150	200
Adversarial	Standard	3.00±2.61	1.00±0.63	0.60±0.80	0.40±0.49
KS	Standard	0.40±0.49	0.20±0.40	0.20±0.40	0.20±0.40
KS	Bonferroni	0.40±0.49	0.20±0.40	0.20±0.40	0.20±0.40
LSDD	Standard	0.60±0.80	0.20±0.40	0.60±0.80	0.20±0.40
LSDD	Bonferroni	0.60±0.80	0.20±0.40	0.60±0.80	0.20±0.40
MMD	Standard	0.40±0.49	0.00±0.00	0.20±0.40	0.00±0.00
MMD	Bonferroni	0.40±0.49	0.00±0.00	0.20±0.40	0.00±0.00
PSI	Standard	30.00±0.00	29.00±0.89	12.20±1.33	1.60±1.02

Table 3: False positives vs. batch size (250–500).

Detector	Variant	250	300	350	400	450	500
Adversarial	Standard	0.00±0.00	0.20±0.40	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
KS	Standard	0.40±0.49	0.00±0.00	0.20±0.40	0.20±0.40	0.20±0.40	0.20±0.40
KS	Bonferroni	0.40±0.49	0.00±0.00	0.20±0.40	0.20±0.40	0.20±0.40	0.20±0.40
LSDD	Standard	0.00±0.00	0.20±0.40	0.40±0.80	0.40±0.80	0.40±0.49	0.40±0.80
LSDD	Bonferroni	0.00±0.00	0.20±0.40	0.40±0.80	0.40±0.80	0.40±0.49	0.40±0.80
MMD	Standard	0.20±0.40	0.40±0.49	0.00±0.00	0.40±0.80	0.00±0.00	0.20±0.40
MMD	Bonferroni	0.20±0.40	0.40±0.49	0.00±0.00	0.40±0.80	0.00±0.00	0.20±0.40
PSI	Standard	0.40±0.49	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00

TABLE 4 — TRUE POSITIVE PERFORMANCE VS. DRIFT MAGNITUDE

Metrics are reported as mean ± standard deviation across 5 seeds.

Table 4: True Positive Performance vs drift magnitude

Detector	Variant	Shift	TPR	FP	TTD (days)
Adversarial	Standard	5	0.00±0.00	0.00±0.00	N/A
Adversarial	Standard	10	0.00±0.00	0.00±0.00	N/A
Adversarial	Standard	15	0.28±0.04	0.00±0.00	8.20±0.40
Adversarial	Standard	20	0.46±0.05	0.00±0.00	6.40±0.49
KS	Standard	5	0.30±0.00	0.20±0.40	7.40±1.20
KS	Standard	10	0.68±0.04	0.20±0.40	4.20±0.40
KS	Standard	15	0.86±0.05	0.20±0.40	2.20±0.40
KS	Standard	20	0.90±0.06	0.20±0.40	2.00±0.63
LSDD	Standard	5	0.08±0.07	0.40±0.49	7.00±0.00
LSDD	Standard	10	0.48±0.07	0.40±0.49	5.80±1.17
LSDD	Standard	15	0.64±0.08	0.40±0.49	4.40±0.80
LSDD	Standard	20	0.74±0.05	0.20±0.40	3.60±0.49
MMD	Standard	5	0.00±0.00	0.40±0.49	N/A
MMD	Standard	10	0.00±0.00	0.40±0.49	N/A
MMD	Standard	15	0.00±0.00	0.20±0.40	N/A
MMD	Standard	20	0.00±0.00	0.00±0.00	N/A
PSI	Standard	5	0.40±0.06	1.60±1.02	5.20±2.14
PSI	Standard	10	0.66±0.05	1.60±1.02	3.60±0.80
PSI	Standard	15	0.74±0.05	1.60±1.02	3.20±0.98
PSI	Standard	20	0.84±0.05	1.60±1.02	2.60±0.49

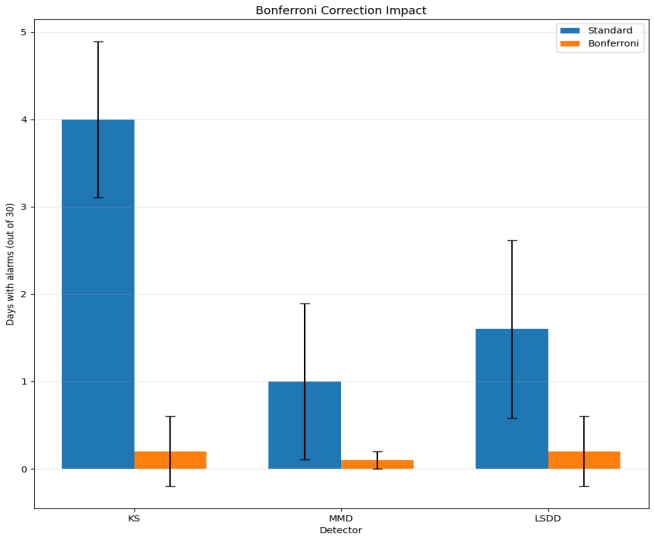

Figure 2: Effect of Bonferroni correction on false alarms.