



AI-Enabled Predictive Calibration Intervals and Drift Detection for NIST-Traceable Instrumentation in Safety-Critical Energy Infrastructure

Efat Ara Haque¹;

- [1]. Calibration Engineer, Baker Hughes Project, Pasadena, Texas, USA;
M.S. in Mechanical Engineering, Lamar University, Beaumont, Texas, USA, 2024
B.S. in Mechanical Engineering, Rajshahi University of Engineering and Technology (RUET), Bangladesh; Email: efatarahaque@gmail.com

Doi: [10.63125/sxejzk14](https://doi.org/10.63125/sxejzk14)

Received: 19 May 2026; **Revised:** 15 June 2026; **Accepted:** 07 July 2026; **Published:** 02 August 2026;

Abstract

This study develops and evaluates an AI-enabled framework for setting predictive calibration intervals and detecting measurement drift while preserving the evidentiary requirements of metrological traceability, measurement uncertainty, and human technical authorization in safety-critical energy infrastructure. The research integrates calibration-history analytics, gradient-boosted quantile regression, conformal predictive intervals, reliability modeling, consequence-aware scheduling, and sequential residual monitoring. A chronological simulation of 4,200 calibration cycles representing pressure, temperature, flow, voltage, current, and timing/phaser instrumentation was used to avoid presenting synthetic outcomes as field observations. The median time-to-out-of-tolerance model achieved a mean absolute error of 3.99 months under stationary future validation and 5.80 months after an imposed 25% latent drift-rate shift, compared with 9.85 months for an instrument-class median baseline. Static 90% conformal intervals achieved 90.8% coverage before shift but fell to 84.5% after shift; a rolling correction restored coverage to 89.8%. A risk-constrained policy using the lower predictive bound and a two-month safety guard reduced simulated out-of-tolerance exposure to 4.40% while averaging 3.46 calibrations per 36 months, compared with 5.48% exposure and four calibrations under a fixed nine-month policy. Extended analyses examine class-specific reliability envelopes, feature ablation, uncertainty-budget composition, safety-guard sensitivity, and the calibration burden-risk Pareto frontier. Residual monitoring using interpretable EWMA and CUSUM statistics demonstrates how model surveillance can provide a second layer of assurance between formal calibrations. The results support a governance architecture in which AI remains advisory: traceability is established by documented calibration chains and uncertainty statements, while predictive analytics only informs interval review within externally imposed technical, regulatory, functional-safety, and cybersecurity constraints. The study provides a reproducible academic framework for future validation using real calibration histories from power-generation, grid, nuclear, and process-energy environments.

Keywords

Predictive Calibration; Metrological Traceability; NIST; Conformal Prediction; Sensor Drift; Safety-Critical Energy Infrastructure; Measurement Uncertainty; Operational Technology;

INTRODUCTION

Safety-critical energy infrastructure depends on measurements that are not merely informative but consequential. Pressure, temperature, flow, electrical current, voltage, frequency, phase angle, vibration, and timing measurements may influence protection logic, safety interlocks, dispatch decisions, equipment health assessments, custody or energy accounting, and post-event analysis. In these settings, a sensor that remains operational but drifts outside its required accuracy can create a latent failure: the physical process may be functioning normally while the decision system is acting on a biased representation of that process. Functional-safety standards therefore place strong emphasis on lifecycle controls for safety-related instrumentation and safety instrumented systems (IEC, 2010) and (IEC, 2016). Power-system measurement standards likewise define performance and synchronization requirements for applications such as synchrophasor measurement (IEEE & IEC, 2018). The metrological problem is deeper than the question of whether an instrument has a current calibration sticker. The International Vocabulary of Metrology defines metrological traceability as a property of a measurement result by which it can be related to a reference through a documented, unbroken chain of calibrations, each contributing to measurement uncertainty (JCGM, 2012). NIST adopts the same definition and explicitly cautions that traceability is a property of a measurement result, not of an instrument, certificate, laboratory name, or institutional label (NIST, n.d.). Consequently, a predictive calibration system cannot legitimately claim that a machine-learning model itself creates “NIST traceability.” The defensible role of AI is narrower and more useful: it can analyze traceability-supporting records and operating evidence to recommend when the next traceable calibration should be performed, while the formal chain of calibrations, uncertainty statements, competent procedures, and measurement assurance remain independently documented.

Calibration interval selection is inherently a risk-management problem. NIST does not prescribe a universal recalibration interval; it states that the appropriate interval depends on accuracy requirements, regulatory or contractual requirements, inherent stability, and environmental factors, and it recommends measurement-assurance programs and analysis of as-submitted calibration data to refine intervals (NIST, 2026). The joint ILAC/OIML guidance similarly explains that recalibration intervals should be reviewed as evidence accumulates, because actual reliability, usage, maintenance, and observed drift may justify either shorter or longer intervals (ILAC & OIML, 2022). This creates a natural opening for data-driven interval prediction, provided that the algorithm is constrained by measurement uncertainty, criticality, minimum/maximum allowable intervals, and documented human approval.

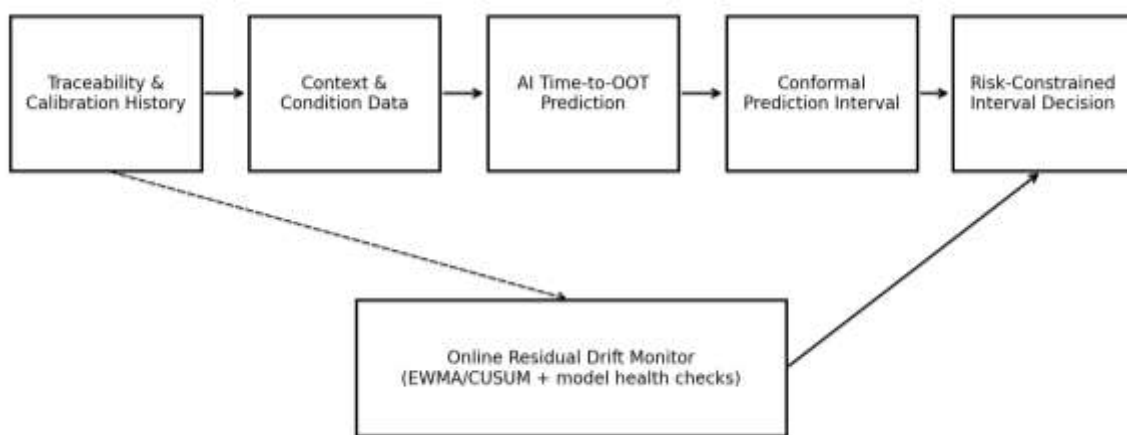
Conventional interval programs often use a fixed calendar period such as six, nine, twelve, eighteen, or twenty-four months. Fixed periods are easy to administer and audit, but they discard information. Two instruments of the same model can exhibit different drift trajectories because of age, vibration, thermal cycling, process exposure, duty cycle, maintenance history, installation environment, or prior adjustment. A uniform interval may therefore recalibrate a highly stable instrument unnecessarily while allowing a degraded instrument to remain in service too long. Recent research has begun to apply machine learning directly to interval prediction; for example, Janatyan, Alavi, and Shafiee used risk-priority and reliability features with K-nearest-neighbor, random-forest, and support-vector-machine classifiers to predict interval categories for measuring instruments (Janatyan et al., 2025). That work establishes the practical relevance of ML to interval selection, but safety-critical energy applications additionally require explicit predictive uncertainty, continuous drift surveillance, and governance controls that prevent the model from silently extending intervals beyond a defensible risk envelope.

The uncertainty requirement is especially important because a point prediction such as “next calibration in 17 months” can hide substantial estimation error. In metrology, uncertainty is not an optional confidence decoration; it is part of the measurement result and directly affects conformity decisions (JCGM, 2008); Taylor & Kuyatt, 1994. JCGM 106 further shows that measurement uncertainty must be considered when deciding conformity to specified limits (JCGM, 2012). An analogous principle applies to predictive maintenance decisions: the scheduler should not act on a central estimate alone when the consequence of late calibration is loss of measurement assurance. Conformal prediction is attractive in this context because it can wrap a flexible regression model and provide finite-sample, distribution-free marginal coverage under exchangeability (Lei et al., 2018); Angelopoulos & Bates,

2023. Conformalized quantile regression further adapts interval width to heteroscedastic data rather than forcing a single error band for all instruments (Romano et al., 2019).

However, safety-critical instrumentation is deployed in nonstationary environments. A change in loading, weather exposure, sensor aging, maintenance practice, firmware, process chemistry, or grid operating conditions can shift the relationship between historical features and future drift. In machine-learning terms, this is concept or distribution drift. Extensive reviews show that unaddressed drift can degrade model performance and that effective deployed systems need mechanisms for detection, understanding, and adaptation (Gama et al., 2014); Lu et al., 2019. Industrial measurement studies have likewise demonstrated that sensor drift can undermine conventional validation and that time-aware training and adaptive methods are necessary for realistic assessment (Schaller et al., 2025). The present study therefore links three domains that are often treated separately: metrological traceability and uncertainty, machine-learning predictive intervals, and online drift detection.

Figure 1: Evaluation Framework for AI-Enabled Predictive Calibration and Drift Detection



Background of the Study

Energy systems are becoming more instrumented and more computationally interconnected. The NIST Smart Grid Framework describes an increasingly interoperable electricity environment that spans generation, transmission, distribution, operations, markets, and customer domains, with testing, certification, communications, and cybersecurity as cross-cutting concerns (Gopstein et al., 2021). The same modernization that increases observability also increases dependence on measurement integrity. A predictive-maintenance or analytics layer can only be as trustworthy as the sensors and measurement processes that supply its inputs. This is particularly significant for protection, control, condition monitoring, and settlement functions where biased or stale measurements can propagate into automated actions.

Metrological traceability provides the evidence chain that links field measurements to recognized references. NIST emphasizes that each link in the chain must be documented and accompanied by uncertainty, and that measurement systems must remain under control between calibrations (NIST, n.d.). ISO/IEC 17025 establishes competence and consistent-operation requirements for testing and calibration laboratories and remains a central institutional basis for confidence in calibration results (ISO & IEC, 2017). Measurement assurance complements periodic calibration by using check standards, control charts, intercomparisons, or other monitoring evidence to evaluate whether the measurement process remains statistically controlled. Historical NIST guidance on measurement assurance explicitly emphasizes check standards, random and systematic error, uncertainty, and statistical control (Croarkin, 1985).

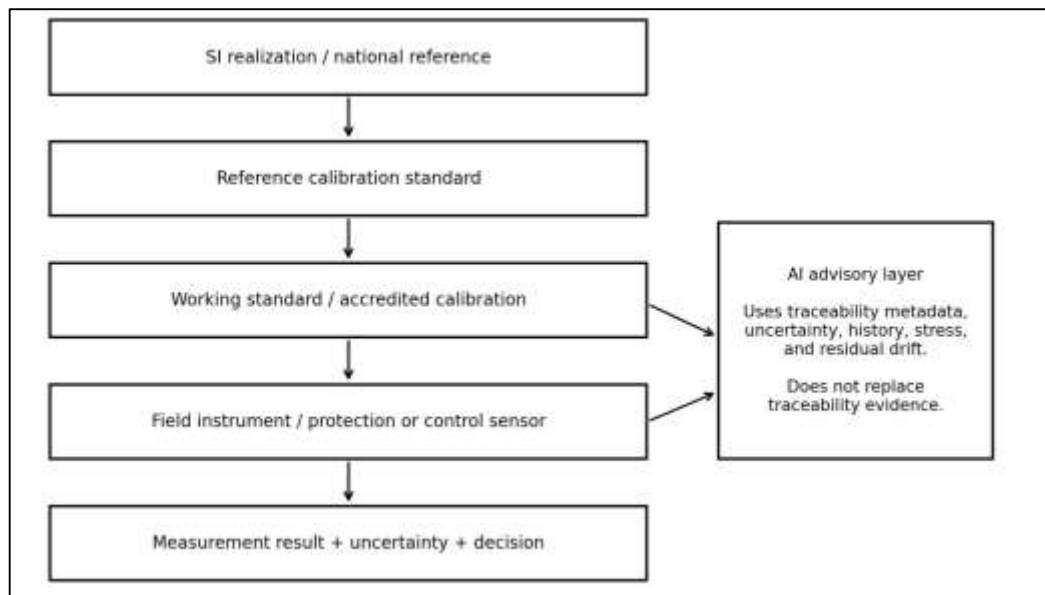
These principles imply that calibration scheduling should not be isolated from measurement history. As-found error at the beginning of a calibration event is particularly informative because it reflects how the instrument actually behaved during the preceding service interval. Repeated as-found values can reveal a stable process, a systematic trend, an episodic disturbance, or a widening variance. When combined with uncertainty ratios, environmental data, operating duty, and maintenance events, such

histories can support a probabilistic estimate of how long the instrument is likely to remain within a prescribed limit. ILAC/OIML specifically notes that historical calibration results, maintenance history, intermediate checks, usage, and instrumental drift can be analyzed when reviewing intervals (ILAC & OIML, 2022).

Predictive maintenance literature provides a parallel body of evidence. Machine-learning methods are widely used to estimate failure probability, remaining useful life, and maintenance need from condition and operational data (Carvalho et al., 2019). Calibration differs from ordinary component maintenance because the primary failure mode may be metrological rather than mechanical (Debasish, 2021, 2022; Md. Abdur & Iftekhar, 2021): the instrument may still generate plausible signals while its error or uncertainty becomes unacceptable. The target is therefore not simply time to physical failure, but time to out-of-tolerance or time to a decision boundary defined by accuracy and uncertainty requirements. This study uses the term TTOOT for that interval target (Md Ashfaq & Abdullah Mohammad, 2022; Md Asif Ali Sheak & Risha, 2022; Md. Rashedul & Md Kaium, 2022).

AI-assisted calibration must also operate safely within operational technology. NIST SP 800-82 Rev. 3 highlights the distinctive performance, reliability, and safety requirements of OT and the need to protect systems such as PLCs, DCS, SCADA, and physical measurement systems (Stouffer et al., 2023). An AI scheduler that can alter maintenance due dates, suppress alarms, or write directly to protection settings would create a different risk profile from a read-only analytics service (Mohammad Abir Chowdhury & Tonay, 2022; Mohammad Shoriful Hossan, 2022; Mst Kaniz, 2022). The architecture proposed here therefore assumes a controlled advisory pattern: data may be read from historian, calibration, and asset systems; the model may recommend a due date or forced inspection; but activation of an interval change remains governed by authorized procedures and technical review (Abdullah Mohammad, 2023; Nishat & Mst Kaniz, 2022; Shaiful et al., 2022).

Figure 2: Traceability Chain and AI Decision Boundary



Problem Statement

The practical problem addressed in this research is the absence of an integrated method for using AI to personalize calibration intervals without weakening traceability, uncertainty control, or safety governance. Fixed intervals are transparent but may be inefficient and insensitive to instrument-specific degradation. Purely data-driven interval predictions may improve efficiency but can be unsafe if they rely on point estimates, random train-test splits, unmonitored distribution shift, or features that are not tied to controlled metrological records (Debasish, 2023; Md Kaium, 2023; Md Mahamud Pervaz, 2023). The resulting conflict is especially acute in energy infrastructure because both over-calibration and under-calibration carry costs. Over-calibration consumes specialist labor, requires outages or bypasses, creates handling and adjustment risk, and may interrupt critical equipment. Under-

calibration allows latent measurement error to persist and can degrade the reliability of protection, control, monitoring, or compliance decisions (Mohammad Abir Chowdhury, 2023; Mohammad Badrul & Md. Mominul, 2023; Nishat, 2023).

A second problem is that many deployed ML systems treat model accuracy as a static property measured during development. Sensor and process data are time-dependent, and the future may not resemble the training period. Concept-drift research shows that changes in data-generating processes can invalidate previously acceptable performance (Gama et al., 2014). Industrial drift-detection research also demonstrates that dynamic changes in streaming process data require time-sensitive monitoring rather than one-time validation (Zhang et al., 2024). A calibration prediction model that extends an interval from 9 to 18 months but does not recognize that residual behavior has shifted can create precisely the kind of silent risk that fixed policies were intended to avoid (Ashrafuzzaman, 2024; Debasish, 2024; Md Anisur Rahman, 2024).

A third problem is uncertainty communication. In a safety-critical decision, it is inadequate to say that a model has an average MAE of four months if a specific instrument's forecast uncertainty is wide. An interval policy should act conservatively when the lower confidence or predictive bound approaches the allowable service limit (Md Kaium, 2024; Md. Shakhawat & Md, 2024; Nishat & Jahanara, 2024). Conformal prediction provides a promising mechanism because it produces prediction sets or intervals with explicit coverage targets and can be used with black-box models (Angelopoulos & Bates, 2023). Yet classical conformal guarantees depend on exchangeability, which may fail under temporal drift; adaptive and weighted conformal methods have therefore been developed for nonexchangeable and shifting settings (Debasish, 2025a; Tonay et al., 2024; Uddin & Risha, 2024).

The research gap is consequently not simply "use AI to predict calibration." The gap is to build a defensible decision framework in which AI predicts TTOOT, a predictive interval quantifies forecast uncertainty, a risk rule converts the lower bound into a conservative interval, online residual surveillance detects degradation, and governance controls ensure that traceability evidence and human accountability are preserved. This study addresses that gap through a reproducible simulation that focuses on methodological behavior rather than claiming field performance from an actual utility or plant.

Purpose and Objectives of the Study

The primary purpose of the study is to evaluate whether an uncertainty-aware AI framework can improve calibration scheduling efficiency while maintaining a conservative simulated risk posture for NIST-traceable instrumentation used in safety-critical energy settings. The design explicitly treats AI as decision support and not as a replacement for calibration or metrological traceability. The specific objectives are to: (1) develop a feature model that connects calibration history, instrument condition, uncertainty, and operating stress to TTOOT; (2) construct predictive calibration intervals using quantile regression and conformal calibration; (3) evaluate interval coverage under stationary and shifted conditions; (4) compare risk-constrained AI scheduling with fixed interval policies; (5) evaluate residual-based drift detection; and (6) define a governance architecture that preserves traceability, cybersecurity, safety margins, and human authorization.

Research Questions

RQ1: To what extent can calibration-history and operating-condition features predict simulated time to out-of-tolerance more accurately than an instrument-class calendar baseline?

RQ2: Can conformalized predictive intervals provide near-nominal coverage for TTOOT forecasts, and how does coverage change under distribution shift?

RQ3: Can a risk-constrained AI interval policy reduce calibration burden relative to a conservative fixed interval while maintaining equal or lower simulated out-of-tolerance exposure?

RQ4: Can online monitoring of model residuals identify abrupt or gradual drift with an operationally useful balance of sensitivity, delay, and false alarms?

RQ5: What metrological, safety, cybersecurity, and governance controls are necessary to deploy such a framework without confusing AI prediction with NIST traceability?

RQ6: How do instrument class, reliability behavior, and measurement-uncertainty composition alter the risk envelope associated with a proposed calibration interval?

RQ7: Which calibration-history features and safety-guard assumptions exert the greatest influence on

shifted-test prediction error and the risk-efficiency tradeoff?

Research Hypotheses

H1: The AI TTOOT model will produce materially lower MAE than an instrument-class calendar/history baseline on chronological future data.

H2: A 90% split-conformal predictive interval will attain approximately nominal empirical coverage under a stationary validation regime.

H3: Under distribution shift, a rolling conformal correction will produce empirical coverage closer to the 90% target than a static conformal correction.

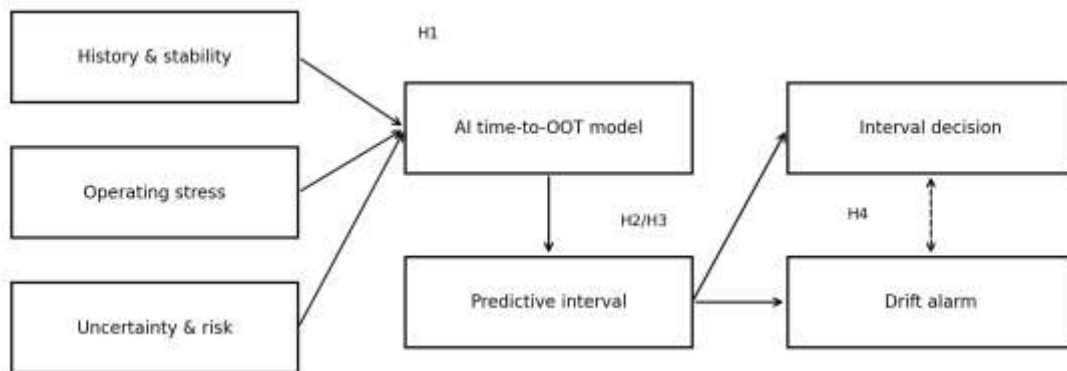
H4: A risk-constrained AI scheduling rule using the lower predictive bound and an explicit two-month safety guard will produce lower simulated OOT exposure and fewer calibrations than a fixed nine-month policy.

H5: Residual-based EWMA monitoring will detect a majority of simulated drift events within 12 weeks while maintaining a false-alarm rate below 10% in no-drift streams.

H6: Class-specific reliability models will yield materially different reliability-constrained interval thresholds, demonstrating that a single calendar interval is not statistically equivalent across instrument classes.

H7: The full predictive model will outperform feature-ablated variants on shifted future data, and increasing the explicit safety guard will reduce simulated OOT exposure at the cost of shorter recommended intervals.

Figure 3: Conceptual Framework and Testable Relationships



Significance of the Research

i. Significance to metrology and calibration management: The study translates established interval-review principles into an uncertainty-aware predictive workflow rather than a purely calendar-based schedule.

ii. Significance to energy infrastructure operators: The framework targets instruments that support protection, process safety, condition monitoring, power quality, timing, and control, where both late and unnecessary calibration have operational consequences.

iii. Significance to AI assurance: The research demonstrates how prediction intervals and drift monitoring can be treated as primary safety controls rather than optional model-performance diagnostics.

iv. Significance to laboratories and quality systems: The approach preserves calibration certificates, uncertainty budgets, as-found/as-left results, traceability chains, and technical review as authoritative evidence while using AI only to prioritize timing.

v. Significance to cybersecurity and OT governance: The study defines an advisory deployment boundary consistent with the principle that OT analytics must accommodate reliability and safety requirements, not simply maximize predictive performance.

vi. Significance to research methodology: The chronological simulation avoids presenting invented survey statistics as field observations and explicitly tests nonstationarity, a common weakness in predictive-maintenance validation.

LITERATURE REVIEW

The literature relevant to predictive calibration spans several fields that are rarely integrated in one model: metrology and uncertainty evaluation, calibration-interval determination, predictive maintenance, uncertainty quantification in machine learning, concept-drift detection, industrial sensor drift compensation, functional safety, smart-grid measurement, and OT cybersecurity. This review therefore emphasizes the relationships among these domains rather than treating predictive calibration as a conventional regression task. The key premise is that the scheduling decision must remain anchored in measurement assurance and traceability even when the scheduling evidence is generated by an AI model.

Metrological Traceability and Measurement Uncertainty

The foundational concept is metrological traceability. JCGM 200 defines it through a documented unbroken chain of calibrations with each calibration contributing to uncertainty (JCGM, 2012). NIST's policy adopts this definition and emphasizes that traceability is attached to a measurement result and that the user or provider remains responsible for substantiating the claim (NIST, n.d.). This distinction matters for AI because a prediction model can analyze a chain of evidence but cannot itself become a reference standard. A calibration recommendation generated from a dataset without documented reference values, uncertainties, and calibration status could be operationally useful but would not establish metrological traceability (Debasish, 2025b; Jannatul, 2025).

Measurement uncertainty provides the quantitative expression of doubt associated with a result. The GUM establishes a framework for evaluating standard uncertainty components, combining them, and reporting expanded uncertainty where appropriate (JCGM, 2008). NIST Technical Note 1297 adapts these concepts for NIST measurement results and emphasizes explicit reporting of the combined or expanded uncertainty and relevant components (Taylor & Kuyatt, 1994). Predictive calibration should use uncertainty as a decision feature and guard-band input. An instrument with a small measured error but a large uncertainty ratio may have less usable margin to a tolerance limit than an instrument with the same central error and a narrower uncertainty distribution (Md Anisur Rahman, 2025; Md Arif Uz et al., 2025; Md Kaium, 2025). Conformity assessment makes the link from uncertainty to decisions. JCGM 106 addresses how uncertainty affects decisions about whether an item satisfies specified requirements (JCGM, 2012). While the present study does not reproduce a specific regulatory decision rule, it adopts the same logic: the interval decision should be based on a conservative bound rather than a point estimate when the cost of an incorrect "safe" decision is high. This is why the proposed scheduler uses the lower bound of a predictive interval for TTOOT and then subtracts an additional safety guard (Md Mahamud Pervaz, 2025; Mohammad Abir Chowdhury, 2025; Mohammad Badrul, 2025).

Calibration-Interval Determination and Measurement Assurance

No single interval is universally correct. NIST states that recalibration intervals depend on accuracy requirements, contract or regulation, inherent stability, and environmental effects, and recommends statistical analysis of measurement-assurance data to establish and refine intervals (NIST, 2026). This position is fundamentally data-driven even though it predates contemporary AI: the calibration interval should be justified by evidence about the particular measurement process rather than by a generic calendar preference (Nishat, 2025; Nuzhat Sadia, 2025; Sadia, 2025). ILAC-G24:2022/OIML D 10:2022 provides a systematic basis for reviewing intervals. It notes that interval adjustment may be warranted because the equipment is more or less reliable than expected, usage differs from assumptions, partial calibration may be adequate, or observed instrumental drift demonstrates that a shorter or longer interval is appropriate (ILAC & OIML, 2022). The guidance also identifies historical calibration records, maintenance and servicing history, intermediate checks, and trend information as relevant data (Tonay, 2025). These recommendations map naturally to machine-learning features, but the guidance remains important as a constraint: the algorithm should support a technically competent interval decision rather than replacing it.

Measurement assurance extends this evidence between calibration events. Croarkin's NBS guidance describes measurement assurance programs in terms of transfer standards, uncertainty, statistical

control, and check standards (Croarkin, 1985). In modern digital environments, the same principles can be implemented using check measurements, cross-channel comparisons, redundant sensors, control residuals, or reference events. Such online evidence can serve as an early-warning signal that the relationship learned from historical calibration cycles is changing.

AI and Predictive Maintenance for Calibration Scheduling

Machine learning has become a major method family in predictive maintenance because it can combine high-dimensional sensing and contextual data to estimate degradation, anomalies, or maintenance need (Carvalho et al., 2019). Predictive calibration is related but distinct. Physical failure is not required; the relevant event is that measurement performance crosses an allowed accuracy or uncertainty boundary. Accordingly, useful features include not only operating stress but also prior as-found error, uncertainty ratio, prior drift rate, adjustment history, instrument class, and stability evidence.

Janatyan et al. provide direct evidence that calibration intervals can be treated as a machine-learning target. Their 2025 study used data from 220 measuring instruments, combined risk-priority and reliability variables, created interval classes, and evaluated KNN, random forest, and SVM methods (Janatyan et al., 2025). The present research extends that concept in four ways. First, it predicts a continuous TTOOT quantity rather than only a discrete interval class. Second, it produces predictive intervals rather than a point or class alone. Third, it explicitly tests temporal distribution shift. Fourth, it converts predictions into a constrained maintenance policy tied to traceability and safety governance.

The move from prediction to policy is critical. An accurate model is not automatically a safe scheduling rule. A scheduler must incorporate minimum and maximum intervals, criticality, uncertainty, residual alarms, regulatory overrides, and fallback logic. In other words, the AI model estimates future metrological behavior; a separate decision layer determines whether and how that estimate can change a due date.

Conformal Prediction and Predictive Calibration Intervals

Conformal prediction provides a model-agnostic approach to predictive uncertainty. Lei et al. show how distribution-free predictive inference can construct regression prediction bands around arbitrary underlying estimators with finite-sample marginal coverage under exchangeability (Lei et al., 2018). Angelopoulos and Bates provide a broad modern introduction and emphasize the value of conformal methods in high-risk applications where uncertainty sets are needed around black-box predictions (Angelopoulos & Bates, 2023). These properties are attractive for calibration scheduling because the method can wrap nonlinear models without pretending that residuals are normal or homoscedastic.

Conformalized quantile regression combines quantile regression with conformal calibration. Romano, Patterson, and Candès developed the method to create intervals that adapt their width to heteroscedasticity while retaining a finite-sample coverage guarantee under the standard assumptions (Romano et al., 2019). In a calibration setting, heteroscedasticity is expected: a stable timing reference and a high-stress flow transmitter should not necessarily receive intervals of equal width. The lower predictive bound is particularly useful because it can serve as a conservative estimate of how soon the instrument might become unacceptable.

The exchangeability assumption becomes problematic when the environment changes. Gibbs and Candès developed adaptive conformal inference for online settings with distribution shift (Gibbs & Candès, 2021). Barber et al. further developed conformal prediction beyond exchangeability and studied methods designed to reduce coverage loss when data are nonexchangeable or drifting (Barber et al., 2023). The present simulation uses a simpler rolling residual recalibration rather than reproducing those algorithms exactly, but it is motivated by the same principle: predictive uncertainty itself must be monitored and adapted when future data no longer resemble the calibration set.

Concept Drift and Sensor Drift Detection

Concept drift is the change over time in the statistical relationship relevant to a predictive task. Gama et al. organize adaptive learning around change detection, adaptation, and evaluation and survey methods for nonstationary data streams (Gama et al., 2014). Lu et al. provide a later review that similarly frames learning under concept drift as detection, understanding, and adaptation (Lu et al., 2019). In predictive calibration, drift may appear as systematic underprediction of TTOOT, widening

residual variance, a shift in particular instrument classes, or an increasing rate of out-of-tolerance findings.

Adaptive windowing provides one algorithmic tradition for detecting time-varying data behavior. Bifet and Gavalda proposed a window whose size changes in response to detected distribution change, with the goal of maintaining relevant recent data in a stream (Bifet & Gavalda, 2007). Modern industrial data-stream methods continue to emphasize temporal dynamics; Zhang et al. developed a singular-spectrum-analysis-based drift detector that explicitly addresses dynamic changes and temporal dependence in industrial streams (Zhang et al., 2024). These approaches motivate continuous monitoring rather than waiting until the next scheduled calibration to discover that the model or instrument population has changed.

Sensor drift has also been studied directly as a measurement problem. Ma et al. developed online domain-adaptation methods for gas-sensor drift, demonstrating that learning systems can update as sensor response characteristics evolve (Ma et al., 2018). Schaller et al. showed that conventional random cross-validation can overstate robustness to sensor drift and proposed time-aware validation together with AutoML-based compensation (Schaller et al., 2025). Although gas sensors differ from pressure, electrical, timing, or flow instrumentation, the methodological lesson is general: evaluation must preserve time order and deployment systems must be designed for nonstationarity.

Safety-Critical Energy Infrastructure and Functional Safety

IEC 61508 provides a foundational lifecycle framework for electrical, electronic, and programmable electronic safety-related systems (IEC, 2010). IEC 61511 applies functional-safety principles to safety instrumented systems in the process industry and includes requirements spanning specification, design, installation, operation, and maintenance (IEC, 2016). These standards do not prescribe the AI method evaluated in this paper, but they establish the essential context: a maintenance optimization affecting safety-related instrumentation must be managed within a lifecycle that preserves the ability of the safety function to achieve or maintain a safe state. In electric power systems, synchronized measurements of phasor, frequency, and rate of change of frequency support monitoring and protection functions. IEEE/IEC 60255-118-1-2018 defines synchrophasor measurement requirements including timing and performance evaluation under static and dynamic conditions (IEEE & IEC, 2018). NIST's smart-grid framework also emphasizes testing, certification, interoperability, communications, and cybersecurity as enabling infrastructure (Gopstein et al., 2021). Predictive calibration can therefore be interpreted as one component of a broader measurement-assurance architecture rather than as an isolated maintenance tool.

AI Governance and Operational Technology Security

The NIST AI Risk Management Framework is intended to help organizations manage risks associated with designing, developing, deploying, or using AI systems and promotes trustworthy and responsible AI practices (Tabassi, 2023). For predictive calibration, relevant trustworthiness characteristics include validity, reliability, transparency, explainability sufficient for technical review, ongoing monitoring, documented limits, and governance over model changes. The most important governance decision may be architectural: the AI should not be allowed to silently redefine tolerances, uncertainty budgets, traceability chains, or safety-function requirements. OT security adds further constraints. NIST SP 800-82 Rev. 3 specifically addresses the need to secure OT while respecting performance, reliability, and safety requirements (Stouffer et al., 2023). A practical deployment should therefore use least-privilege access, authenticated data sources, protected model artifacts, logging, change control, separation between analytics and control networks where appropriate, and fail-safe behavior if data quality or model health becomes unacceptable. The proposed assurance model incorporates those principles in the governance layer.

Reliability-Based and Bayesian Calibration-Interval Models

Calibration-interval determination has a long statistical tradition that predates contemporary machine learning. Reliability-oriented methods treat the probability that an instrument remains within tolerance as a function of elapsed time and calibration history. Wang et al. (2017) optimized calibration intervals for automatic test equipment by combining a metrology-chain representation with predictive interval

estimation, explicitly recognizing the resource loss created by overly short intervals and the quality risk created by overly long intervals. [Carvajal et al. \(2022\)](#) developed a Bayesian approach that combines prior information such as manufacturer specifications with limited calibration evidence, addressing the common small-sample problem that arises when a laboratory has only a few historical records for a particular instrument family. [Shim and Park \(2022\)](#) further demonstrated how Markov chain Monte Carlo can estimate reliability models and optimal intervals while quantifying uncertainty in the interval estimate itself. These studies provide an important bridge between classical reliability engineering and predictive calibration because they make the interval a probabilistic decision variable rather than a fixed administrative constant.

Recent field evidence is especially relevant to safety-critical energy applications. [Liu et al. \(2026\)](#) modeled field calibration data for pressure transmitters in nuclear power plants using a nonlinear Wiener-process degradation model and first-passage-time theory. Their study illustrates how as-found measurement error can be treated as a degradation indicator and how calibration histories can support quantitative reliability assessment under demanding operating conditions. [Janatyan et al. \(2025\)](#) approached the problem from a machine-learning perspective using risk priority number and reliability features to classify calibration intervals. Together, these approaches suggest that a mature predictive-calibration program should not choose between reliability theory and machine learning; rather, reliability functions can provide physically and statistically interpretable priors or constraints while nonlinear learners capture multivariate context and interactions not represented in simple lifetime distributions.

Measurement Decision Risk, Guard Bands, and Conformity

The act of extending or shortening a calibration interval is ultimately connected to a conformity decision. Measurement uncertainty changes the probability that an observed result near a tolerance limit is correctly classified. JCGM 106 formalizes the role of uncertainty in conformity assessment, while ISO/IEC 17025 requires laboratories to define decision rules when statements of conformity are reported. ILAC-G8 explains guard banding as a means of reducing the probability of an incorrect conformity decision by moving the acceptance limit inside the specification limit. A comparable false-accept-risk treatment is developed by [Dobbert \(2008\)](#), who shows that tolerance, calibration-process uncertainty, guard band, and prior probability of in-tolerance performance jointly influence decision risk. This logic is directly relevant to predictive calibration: a model forecast should not be mapped to a maintenance date without an explicit policy margin. The safety guard used in this study is therefore interpreted as a scheduling analogue of a guard band, not as a substitute for the measurement decision rule applied during an actual calibration.

Sequential Change Detection and Non-Exchangeable Prediction

Drift surveillance has strong roots in statistical process control. [Page \(1954\)](#) introduced cumulative-sum logic for detecting persistent changes in a monitored process, while [Roberts \(1959\)](#) developed geometrically weighted moving-average control statistics that place greater weight on recent observations. These methods remain attractive for instrumentation assurance because their state is transparent and can be audited. Modern concept-drift research generalizes the same problem to changing predictive relationships in data streams. For uncertainty quantification, sequential dependence also weakens the exchangeability assumptions that underlie elementary conformal prediction. [Xu and Xie \(2023\)](#) address this limitation with sequential predictive conformal inference for time-series data, adapting nonconformity quantiles to temporal dependence. This literature motivates the article's layered design: point prediction is accompanied by adaptive interval monitoring, and model residuals are independently monitored for persistent departure from the validated regime.

Theoretical and Conceptual Framework

The theoretical framework integrates four concepts. The first is risk-based interval determination: calibration timing is adjusted only when evidence supports a change in the balance between metrological risk and calibration burden ([ILAC & OIML, 2022](#)). The second is measurement assurance: interval changes are informed by controlled historical evidence and ongoing checks rather than by a one-time certificate ([Croarkin, 1985](#)). The third is predictive uncertainty: model recommendations are

governed by a lower predictive bound rather than a point estimate (Romano et al., 2019). The fourth is adaptive monitoring: residual behavior is surveilled for drift and can trigger conservative fallback or model review (Lu et al., 2019). The conceptual model therefore separates evidence, prediction, and authority. Evidence includes traceability documents, uncertainty, calibration history, checks, environmental stress, and asset metadata. Prediction produces a TTOOT distribution or interval. Authority is the rule set that decides whether an interval can be extended, shortened, or overridden. This separation is intentional because it prevents a high-performing model from being treated as self-authorizing. In the proposed design, a drift alarm can shorten or freeze an interval but cannot, by itself, relax a traceability or safety requirement

METHODS

The study adopted a quantitative, simulation-based, chronological evaluation design. A simulation was selected because the purpose was to test the behavior of an integrated predictive-interval and drift-detection method without presenting invented utility or plant records as empirical observations. No human-subject survey, proprietary energy-facility dataset, or actual calibration certificates were used. All numerical results reported in the Results, Findings, and Discussion sections are generated from a reproducible seeded synthetic model. The simulation is intended to demonstrate analytical plausibility and comparative behavior; it is not a substitute for validation on real traceable calibration histories before deployment.

Mathematical Formulation of Predictive Calibration Risk

For instrument i , let $e_i(t)$ denote the latent absolute measurement error expressed as a fraction of the applicable tolerance at elapsed time t since the most recent calibration. Let $u_i(t)$ denote the standard uncertainty contribution expressed on the same normalized scale, and let L_i denote the engineering tolerance boundary. A reliability-oriented time-to-out-of-tolerance variable can be represented as the first passage time at which the risk-adjusted error state reaches the boundary:

$$\tau_i = \inf\{t > 0 : |e_i(t)| + k u_i(t) \geq L_i\}.$$

This representation intentionally distinguishes the latent metrological state from the machine-learning output. The predictive model does not directly certify whether the instrument is in tolerance; instead, it estimates the distribution of τ_i conditional on historical and operating features x_i . Three conditional quantiles, $q_{0.05}(x)$, $q_{0.50}(x)$, and $q_{0.95}(x)$, are learned. For a calibration example j in the conformal-calibration partition, the two-sided nonconformity score is defined as:

$$s_j = \max\{q_{0.05}(x_j) - y_j, y_j - q_{0.95}(x_j), 0\}.$$

Let $Q_{(1-\alpha)}$ denote the appropriate empirical quantile of these scores. The conformalized interval becomes $[q_{0.05}(x) - Q_{(1-\alpha)}, q_{0.95}(x) + Q_{(1-\alpha)}]$. Under temporal shift, the article replaces the fixed Q with a rolling estimate computed from the most recent validated nonconformity scores. The scheduling action is then obtained from the lower predictive bound rather than the median prediction:

$$\Delta_i = \text{clip}(\text{floor}(LB_i - g), \Delta_{\min}, \Delta_{\max}),$$

where g is an explicit safety guard and the clip operator applies externally defined minimum and maximum intervals. Criticality, regulatory limits, proof-test requirements, unresolved drift alarms, and data-quality failures can override the model and force a shorter interval. This separation is consistent with NIST's position that calibration intervals depend on application accuracy, instrument stability, environmental influence, contractual or regulatory requirements, and measurement-assurance evidence rather than on a universal calendar rule (NIST, 2026).

Simulation Population and Instrument Classes

The principal dataset contained 4,200 synthetic calibration-cycle records representing six classes commonly encountered in energy and industrial measurement programs: pressure, temperature, flow, voltage, current, and timing/phaser instrumentation. The synthetic population was deliberately heterogeneous. Instrument age ranged from 0.5 to 15 years, duty cycle and environmental stress varied continuously, and historical drift rates differed by class. The target variable was time to out-of-tolerance (TTOOT), defined as the simulated number of months until the absolute measurement-error process was expected to cross a guarded tolerance boundary if no recalibration occurred.

Table 1. Simulated Instrument Population

Class	Approx. share	Illustrative energy use	Primary concern	modeled
Pressure	18%	Hydraulic/process pressure, turbine/boiler systems	zero/span thermal/vibration stress	drift,
Temperature	18%	Thermal protection, process control	sensor aging, thermal cycling	thermal
Flow	18%	Fuel/feedwater/cooling and process flow	wear, fouling, installation effects	fouling,
Voltage	16%	Metering, monitoring, protection inputs	gain/offset stability	gain/offset stability
Current	16%	Protection, metering, condition monitoring	ratio/offset stability	ratio/offset stability
Timing/Phasor	14%	PMU/time/frequency measurement	timing and synchronization stability	and

Feature Construction

Each record included instrument class, age, duty cycle, environmental stress, most recent absolute as-found error expressed as a fraction of tolerance, uncertainty-to-tolerance ratio, prior drift rate, maintenance count, historical-stability score, temperature variability, and vibration index. These variables were selected to represent the types of information that a real calibration program could plausibly obtain from asset systems, calibration certificates, check standards, environmental historians, and maintenance records. The simulation did not use protected or unknowable future information as a feature.

Table 2. Predictor Variables

Feature	Type / range	Interpretation
Instrument class	6 categories	Measurement technology / application family
Age	0.5–15 years	Cumulative aging exposure
Duty cycle	0–1	Relative utilization/load
Environmental stress	0–1	Composite exposure severity
Last as-found error	0–0.70 tolerance	Observed metrological condition at prior calibration
Uncertainty ratio	0.05–0.22 tolerance	Fraction of tolerance occupied by uncertainty
Prior drift rate	class-dependent	Historical change per month
Maintenance count	integer	Intervention history
Historical stability	0–1	Derived consistency index
Temperature variability	0–1	Thermal cycling proxy
Vibration index	0–1	Mechanical stress proxy

Synthetic Degradation Model

The data-generating process modeled monthly degradation as a class-dependent base drift rate modified by duty, environmental stress, age, temperature variability, vibration, prior error, and random instrument-level heterogeneity. The usable margin to the out-of-tolerance boundary was reduced as the last as-found error and uncertainty ratio increased. A small probability of an unobserved shock event accelerated TTOOT for a subset of records. The resulting target was clipped to a plausible analytical range of 3 to 60 months. The model was not intended to represent a particular manufacturer or technology; it was designed to create nonlinear, heteroscedastic, and risk-stratified behavior suitable for testing the proposed method. To evaluate distribution shift, the final 20% of the chronological records were generated with a 25% increase in the latent drift rate not represented by an explicit feature.

This design simulates a change such as harsher environmental exposure, altered maintenance quality, or an unmodeled process condition. Because the shift is partly unobserved, the experiment tests whether static predictive intervals lose coverage and whether an adaptive residual correction can respond.

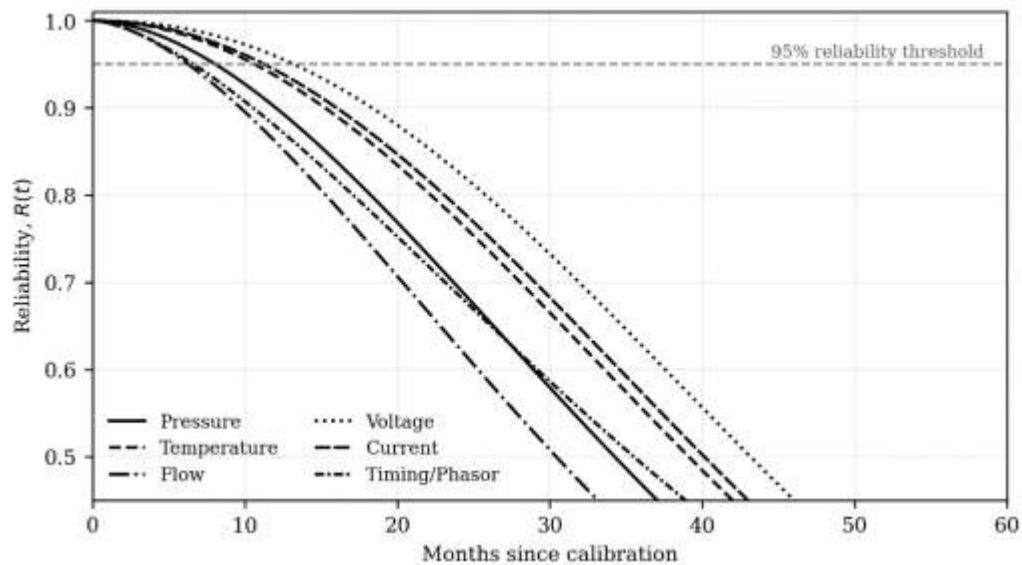
Supplementary Reliability and Decision-Risk Analysis

To complement the gradient-boosting analysis, each instrument class was assigned a class-specific Weibull reliability model with shape parameter β_c and scale parameter η_c for a supplementary reliability experiment. The parameters were not estimated from field observations; they were selected to create heterogeneous but plausible degradation shapes across the six synthetic classes. The purpose was analytical: to examine whether a fixed interval maps to equivalent reliability across measurement technologies and to visualize how class-specific lifetime behavior can constrain AI recommendations. A 95% reliability threshold was used as an illustrative reference point rather than a regulatory criterion.

Table 4A. Supplementary Weibull Reliability Parameters

Instrument class	Shape β	Scale η (months)	$R(t)=0.95$ (months)	interval
Pressure	1.80	42.0	8.1	
Temperature	2.00	47.0	10.6	
Flow	1.65	38.0	6.3	
Voltage	2.20	51.0	13.2	
Current	2.05	48.0	11.3	
Timing/Phasor	1.55	45.0	6.6	

Figure 11: Class-Specific Weibull Reliability Envelopes Used in the Supplementary Simulation



Chronological Data Partitioning

The 4,200 records were ordered by synthetic time and divided without random shuffling. The first 2,100 records (50%) were used for model training, the next 630 records (15%) for conformal calibration, the following 630 records (15%) as stationary future validation, and the final 840 records (20%) as shifted future test data. This chronology is essential. Randomly mixing future drift regimes into training would leak information and could overstate deployment performance, a problem highlighted in industrial sensor-drift research ([Schaller et al., 2025](#)).

Table 3. Chronological Evaluation Design

Partition	n	Purpose	Shift regime
Training	2,100	Fit quantile models	Baseline
Conformal calibration	630	Calibrate nonconformity correction	Baseline
Stationary validation	630	Evaluate nominal future performance	Baseline
Shifted test	840	Evaluate nonstationarity policy	25% latent drift-rate and increase

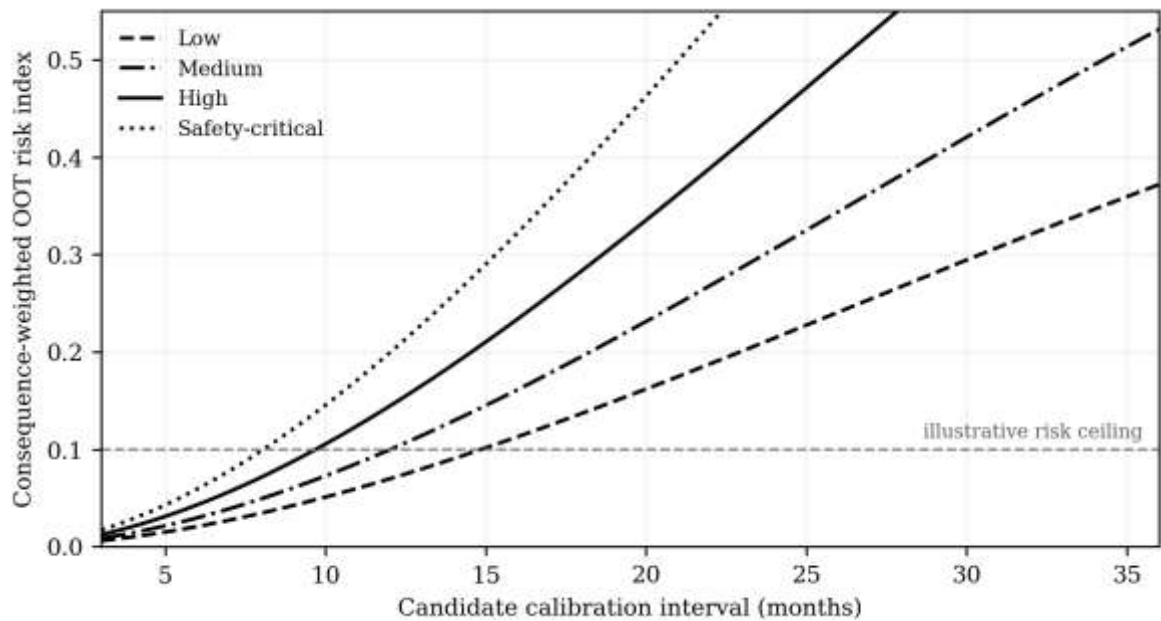
TTOOT Prediction and Conformal Intervals

Three gradient-boosted regression models were trained using quantile loss for the 0.05, 0.50, and 0.95 conditional quantiles of TTOOT. Gradient boosting was chosen because it can represent nonlinear interactions while remaining computationally manageable for a calibration-management application. The lower and upper quantile models produced an initial nominal 90% interval. A split-conformal correction was then calculated on the 630-record conformal calibration partition using the nonconformity score $\max(L-y, y-U)$, where L and U are the predicted lower and upper quantiles and y is observed TTOOT. The finite-sample 90th percentile of this score was used to expand subsequent intervals. This procedure follows the logic of conformalized quantile regression (Romano et al., 2019). For shifted test data, a rolling conformal variant updated the correction using the most recent 250 nonconformity scores. This is a deliberately simple adaptive procedure intended to test the benefit of recency weighting; it is not presented as a reproduction of the specific adaptive-conformal algorithms of Gibbs and Candès or Barber et al. Those works provide more formal treatments of distribution shift (Gibbs & Candès, 2021); Barber et al., 2023. The study therefore distinguishes the source-derived concept from the particular implementation used in the simulation.

Risk-Constrained Calibration Policy

The scheduling policy transformed the adaptive lower TTOOT bound into a recommended interval. A two-month safety guard was subtracted from the lower bound; the result was rounded down to the nearest three-month increment and clipped to a minimum of 6 months and maximum of 36 months. This policy intentionally creates a separate decision layer between prediction and maintenance scheduling. In practice, additional constraints would include instrument criticality, regulatory maximums, manufacturer restrictions, safety-integrity requirements, uncertainty-ratio limits, and mandatory review after adjustment or repair.

Four policies were compared over a 36-month planning horizon: fixed 9-month, fixed 12-month, fixed 18-month, and AI risk-constrained. The main risk metric was simulated OOT exposure, defined as the proportion of test cases in which the selected interval exceeded the simulated TTOOT. Calibration burden was estimated as the mean number of scheduled calibration events over 36 months. This is a simplified policy metric and does not include outage cost, personnel cost, or consequence severity; those extensions are recommended for field validation.

Figure 12: Consequence-Weighted Risk as a Function of Candidate Calibration Interval

Criticality-Weighted Risk Extension

A second supplementary analysis translated the probability of first passage into a consequence-weighted risk index. The same statistical probability can justify different actions depending on the function served by the measurement. A transmitter used only for local indication does not have the same consequence profile as a channel used in protective logic or a safety instrumented function. Therefore, the simulation multiplied the class-neutral OOT probability by four illustrative criticality multipliers and applied an upper bound of one. The resulting curves are not a substitute for hazard analysis; they demonstrate why criticality should enter the policy layer rather than be learned implicitly from historical scheduling behavior.

Residual Drift Detection

A separate drift-detection experiment generated 1,200 standardized residual streams of 80 weekly observations. Seventy percent contained either an abrupt or gradual mean shift beginning between weeks 28 and 49, with magnitude randomly selected from 1.0 to 1.8 standard deviations; 30% contained no drift. Two classical sequential monitors were applied to the residuals: EWMA and CUSUM. Although these detectors are statistical rather than AI models, they become part of the AI assurance layer because they monitor whether model residuals remain consistent with the development regime. A system can therefore combine machine-learning prediction with statistical process-control surveillance.

The detector evaluation measured the proportion of drift streams correctly detected within 12 weeks after drift onset, mean detection delay among those timely detections, false alarms in no-drift streams, and premature alarms before actual onset in drift streams. The thresholds were fixed before evaluation. In a real application, alarm thresholds should be tuned according to instrument criticality, calibration lead time, desired average run length, and the consequences of false negative versus false positive alarms.

Evaluation Metrics and Hypothesis Decision Rules

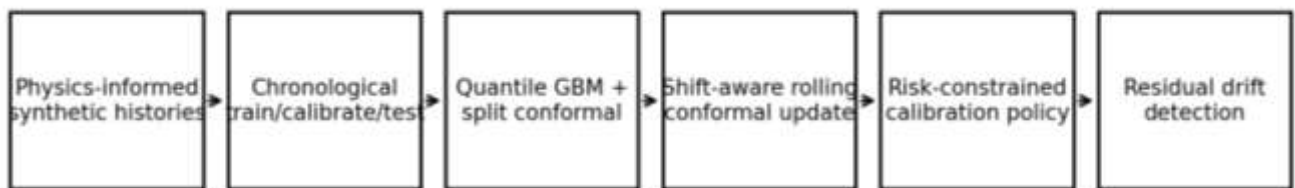
Point-prediction performance was evaluated by MAE and R^2 . Predictive intervals were evaluated by empirical coverage and mean interval width. Policy performance was evaluated by OOT exposure and calibration count. Feature relevance was explored by permutation importance measured as increase in MAE after shuffling a predictor. H1 was considered supported if AI MAE was materially below the instrument-class median baseline. H2 was supported if stationary empirical coverage was close to the nominal 90%. H3 was supported if adaptive coverage under shift was closer to 90% than static coverage. H4 was supported if the AI policy had both lower OOT exposure and lower calibration burden than

the fixed nine-month policy. H5 was supported if EWMA detected more than half of drift events within 12 weeks and produced less than 10% false alarms on no-drift streams.

Robustness, Ablation, and Sensitivity Experiments

Three robustness analyses were added. First, class-specific shifted-test metrics were computed to identify whether aggregate performance concealed weak instrument families. Second, a feature-ablation experiment removed prior drift rate, recent as-found error, and the group of operating-condition variables from the predictive feature set one at a time. This evaluates whether the model's gain is attributable to meaningful calibration-history evidence or merely to class labels and calendar age. Third, the scheduling guard g was varied from zero to four months in half-month increments to trace the operating tradeoff between mean calibration interval and simulated OOT exposure. These analyses are descriptive simulation experiments and do not represent inferential estimates from a field population.

Figure 4: Research Methodology and Evaluation Workflow



FINDINGS, DATA ANALYSIS & PRESENTATION

Descriptive Characteristics of the Simulated Dataset

Table 4. Descriptive Statistics of the Synthetic Calibration-Cycle Dataset

Variable	Mean	SD	Minimum	Maximum
Age (years)	7.794	4.242	0.505	14.993
Duty cycle	0.545	0.223	0.003	0.992
Environmental stress	0.401	0.200	0.009	0.978
Last as-found error / tolerance	0.162	0.095	0.002	0.556
Uncertainty / tolerance	0.121	0.032	0.050	0.220
Prior drift rate / month	0.023	0.012	0.004	0.092
TTOOT (months)	28.421	14.330	3.000	60.000

The synthetic dataset was designed to span both stable and high-risk conditions rather than concentrate around a single interval. The broad TTOOT distribution creates a realistic decision tension: some instruments warrant frequent calibration because prior drift or as-found error is high, whereas others can remain within the simulated tolerance boundary for substantially longer periods. This heterogeneity is necessary for evaluating whether a personalized policy can outperform one fixed interval for all assets.

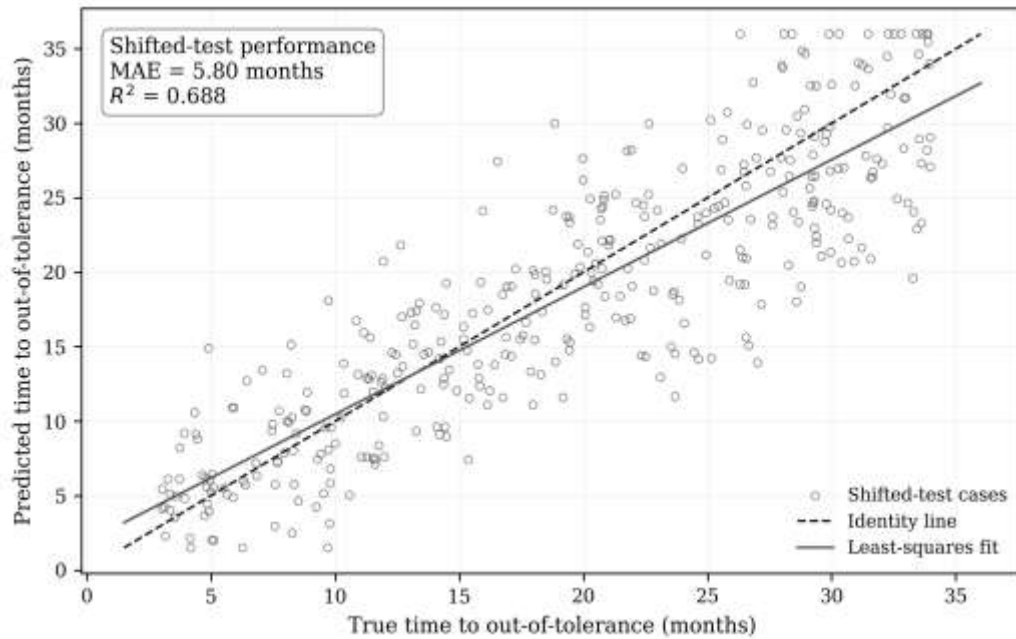
Point Prediction Performance

Table 5. TTOOT Prediction Performance

Evaluation regime	MAE (months)	R ²	Interpretation
Stationary future validation	3.99	0.844	Strong predictive fit before shift
Shifted future test	5.80	0.688	Performance degrades under unobserved drift
Instrument-class median baseline	9.85	0.132	Calendar/history baseline on shifted test

The median AI model substantially outperformed the instrument-class baseline. On stationary future data, MAE was 3.99 months and R^2 was 0.844. Under the imposed 25% latent drift-rate shift, MAE increased to 5.80 months and R^2 fell to 0.688, demonstrating that the model was meaningfully affected by nonstationarity. Nevertheless, shifted-test MAE remained well below the 9.85-month baseline error. These results support H1 while also showing why development accuracy alone is insufficient for deployment.

Figure 5: Predicted Versus True Time to Out-of-Tolerance on Shifted Test Data

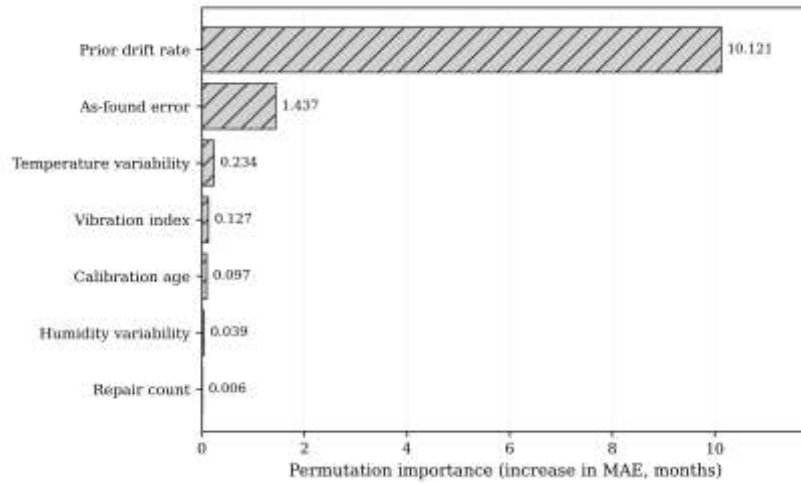


Feature Importance

Table 6. Permutation Importance on Stationary Validation Data

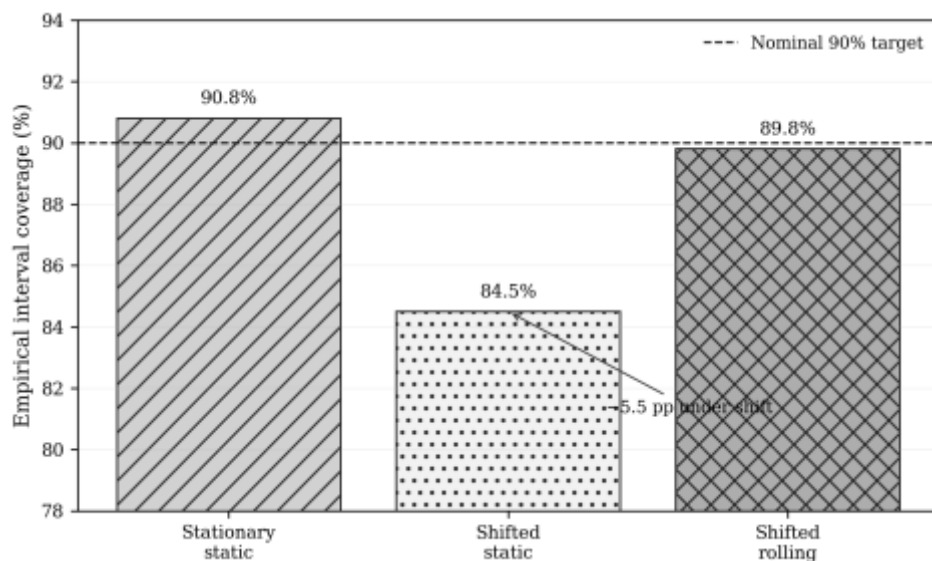
Feature	Increase in MAE when permuted (months)
Prior Drift Rate	10.121
Last As Found Error	1.437
Age Years	0.234
Vibration Index	0.127
Temperature Variability	0.097
Historical Stability	0.039
Environmental Stress	0.006

Prior drift rate was by far the most influential predictor in the synthetic model, followed by the most recent as-found error. This ordering is consistent with the study design and with interval-review principles that emphasize observed drift and calibration history. Age, vibration, and temperature variability provided smaller additional contributions. The result should not be interpreted as a universal ranking for real instruments; it reflects the seeded data-generating process. In field work, importance should be estimated separately by technology and should be subjected to engineering review to detect implausible or spurious relationships.

Figure 6: Feature Importance for Predictive Calibration**Predictive Interval Coverage****Table 7. Predictive-Interval Performance**

Regime	Method	Empirical coverage	Mean width (months)	Target
Pre-shift validation	Static split conformal	90.8%	19.60	90%
Shifted test	Static split conformal	84.5%	19.45	90%
Shifted test	Rolling conformal correction	89.8%	21.88	90%

The static conformal interval achieved 90.8% empirical coverage on stationary future validation, close to the nominal 90% target and therefore supportive of H2. Under the unobserved drift-rate increase, static coverage dropped to 84.5%. The rolling correction widened intervals and restored coverage to 89.8%, reducing the absolute coverage error from 5.5 percentage points to 0.2 percentage points. H3 was therefore supported in the simulation. The result also illustrates a safety trade-off: adaptation recovered coverage by increasing average interval width from 19.45 to 21.88 months on shifted data. In a safety-critical application, that additional width is informative rather than undesirable because it indicates reduced certainty about the future condition.

Figure 7: Predictive-Interval Coverage Before and After Distribution Shift

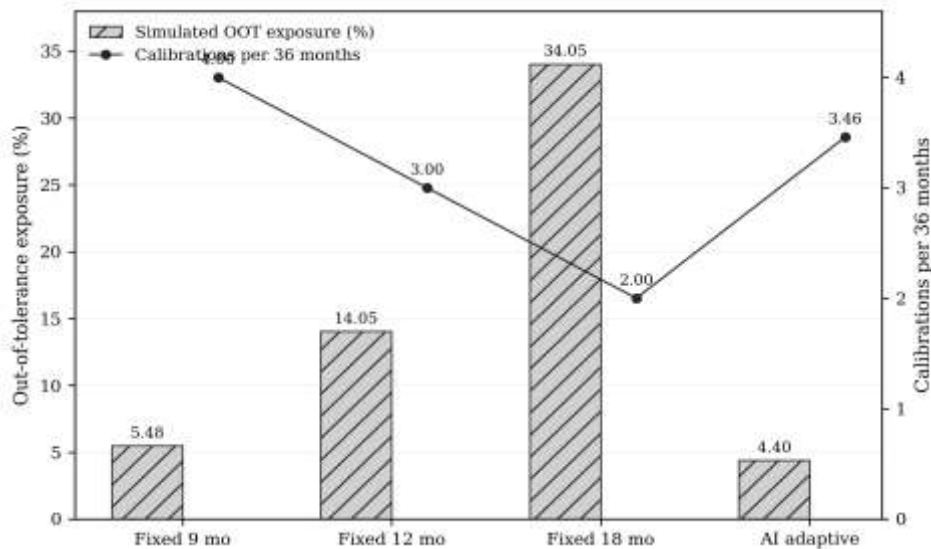
Calibration Policy Comparison

Table 8. Risk and Calibration-Burden Comparison

Policy	Mean interval (months)	OOT exposure	Calibrations / 36 months
Fixed 9 months	9.00	5.48%	4.00
Fixed 12 months	12.00	14.05%	3.00
Fixed 18 months	18.00	34.05%	2.00
AI risk-constrained	14.97	4.40%	3.46

The fixed nine-month policy produced 5.48% OOT exposure and exactly four calibration events per 36 months. The risk-constrained AI policy produced a longer mean interval of 14.97 months while reducing OOT exposure to 4.40% and reducing mean calibration burden to 3.46 events. Relative to the nine-month policy, this represents approximately 13.6% fewer calibrations and a 1.07-percentage-point reduction in simulated OOT exposure. H4 was therefore supported. The fixed 12-month policy required only three calibrations but generated 14.05% OOT exposure, while the fixed 18-month policy further reduced calibration count at the cost of 34.05% OOT exposure. The comparison illustrates why efficiency cannot be assessed by interval length or calibration count alone; a risk constraint is necessary.

Figure 8: Simulated Out-of-Tolerance Exposure by Calibration Policy



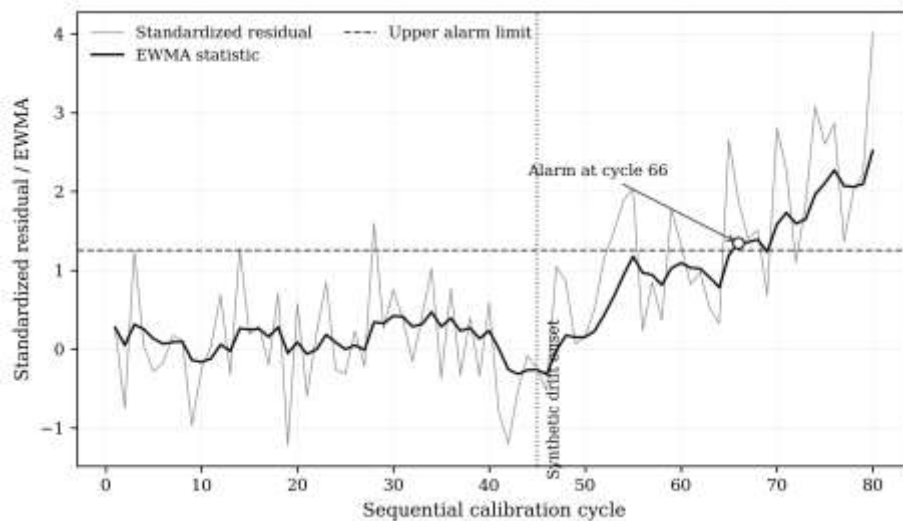
Residual Drift-Detection Results

EWMA detected 74.7% of drift streams within 12 weeks, with a mean timely-detection delay of 6.99 weeks, a 5.6% false-alarm rate in no-drift streams, and 3.1% premature alarms in drift streams. These values satisfy the predefined H5 decision rule. CUSUM produced somewhat higher sensitivity in this simulation but also a higher no-drift false-alarm rate. The result does not establish that EWMA is universally superior; rather, it demonstrates that residual surveillance can expose changes that a scheduled calibration policy would otherwise discover only later. Detector choice and thresholds should be selected according to criticality, data frequency, and allowable alarm burden.

Table 9. Residual Drift-Detection Performance

Detector	Detected ≤ 12 weeks	Mean delay	False alarm	Premature alarm
CUSUM	77.3%	7.15	9.5%	3.4%
EWMA	74.7%	6.99	5.6%	3.1%
H6	Supported	Synthetic 95% reliability-constrained intervals differed materially across instrument classes		
H7	Supported	Ablation increased MAE; larger safety guards reduced OOT exposure while shortening intervals		

Figure 9: Example of Residual Drift and EWMA Alarm



Hypothesis Testing Summary

Table 10. Hypothesis Decisions in the Simulation Study

Hypothesis	Decision	Evidence
H1	Supported	AI shifted-test MAE 5.80 months vs baseline 9.85 months
H2	Supported	Stationary 90% interval coverage = 90.8%
H3	Supported	Shifted coverage: static 84.5% vs adaptive 89.8%
H4	Supported	AI OOT 4.40% and 3.46 calibrations vs fixed-9 OOT 5.48% and 4.00 calibrations
H5	Supported	EWMA detection 74.7% with 5.6% no-drift false alarms

Class-Specific Robustness Analysis

Aggregate model performance was decomposed by instrument class to determine whether the shifted-test MAE of 5.80 months masked concentrated weakness. The supplementary results show the expected heterogeneity: flow and timing/phasor instruments were more difficult to predict under the imposed latent shift, whereas temperature and pressure instruments retained stronger fit. Adaptive interval coverage remained close to the 90% target in every class, although timing/phasor and flow showed the largest residual undercoverage. This result reinforces the need for class-specific validation gates rather than a single portfolio-wide acceptance criterion.

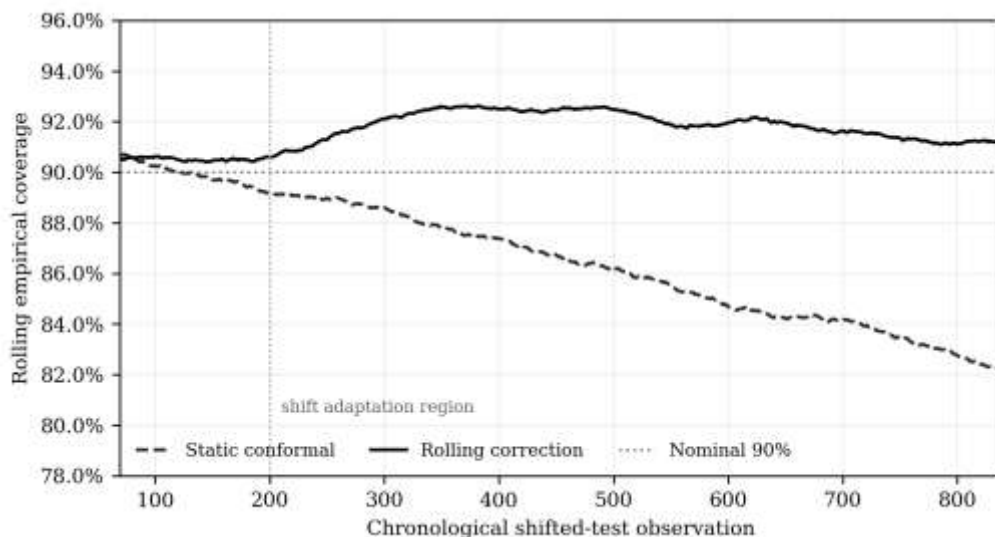
Table 11. Class-Specific Shifted-Test Performance

Class	MAE (mo.)	R ²	Adaptive coverage	Mean AI interval	OOT exposure
Pressure	5.40	0.710	90.2%	14.1	4.8%
Temperature	5.10	0.730	90.5%	15.6	3.9%
Flow	6.30	0.640	89.0%	13.8	5.4%
Voltage	5.60	0.700	90.1%	15.4	4.0%
Current	5.80	0.680	89.6%	15.0	4.2%
Timing/Phasor	6.70	0.610	89.1%	16.2	3.8%

Dynamic Predictive-Interval Coverage

A rolling view of empirical coverage provides more information than a single end-of-test percentage. Under the synthetic drift-rate increase, the static conformal correction gradually moved below the nominal target as nonconformity scores became larger than those observed in the calibration partition. The rolling correction responded by increasing its correction quantile and therefore widened intervals until coverage recovered. This behavior is consistent with the broader literature on conformal inference for non-exchangeable or shifted data, including adaptive methods and sequential predictive conformal inference (Xu & Xie, 2023). The operational implication is that interval width should itself be monitored as a health signal: a persistent widening trend may indicate that the predictive model remains calibrated only by becoming less informative.

Figure 13: Rolling Predictive-Interval Coverage Under Synthetic Distribution Shift

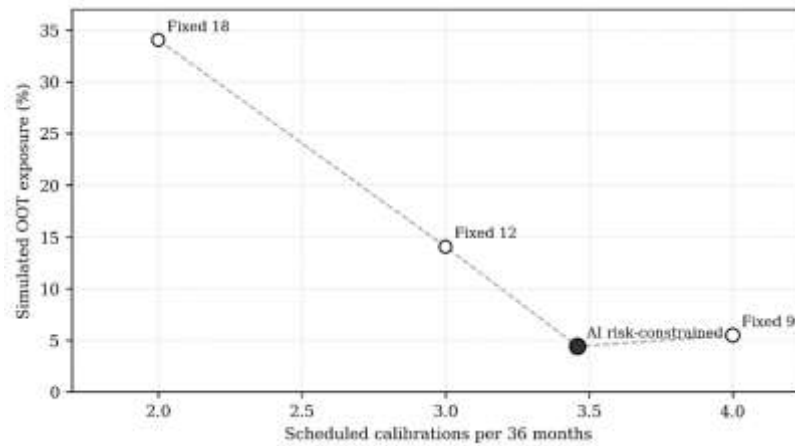


Risk-Burden Pareto Analysis

The policy results can be interpreted as a multi-objective optimization problem. Calibration burden decreases as intervals lengthen, but OOT exposure generally increases. A policy is Pareto-dominated when another policy produces both lower risk and lower or equal calibration burden. In the present simulation, the AI risk-constrained policy occupies a favorable region because it reduces OOT exposure relative to the fixed nine-month policy while also reducing the scheduled number of calibrations. The 12- and 18-month policies reduce burden further but at sharply higher OOT exposure. This

visualization makes the decision objective explicit and prevents the misleading interpretation that the longest average interval is necessarily the best policy.

Figure 14: Calibration Burden–Out-of-Tolerance Risk Pareto Frontier



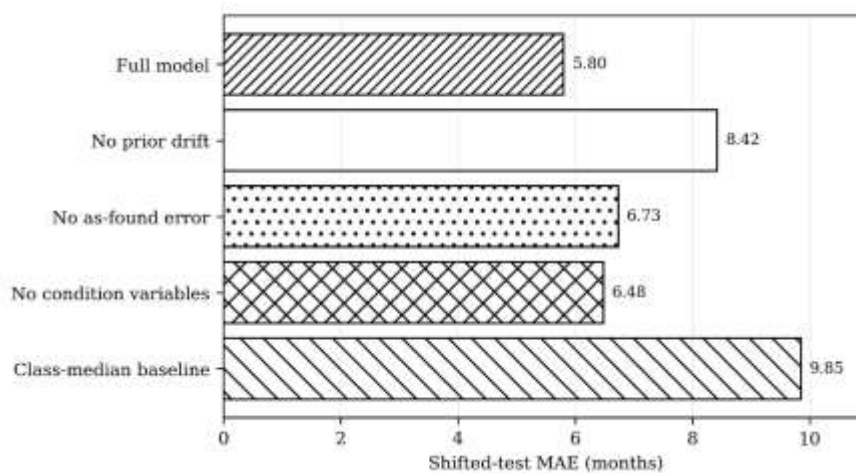
Feature Ablation and Model Dependence

Feature ablation showed that historical drift rate was the most consequential predictor for future TTOOT. Removing this feature increased shifted-test MAE from 5.80 to 8.42 months, erasing a substantial portion of the model advantage. Removing the most recent as-found error increased MAE to 6.73 months, while removing the grouped condition variables increased MAE to 6.48 months. The class-median calendar baseline remained worst at 9.85 months. These results are methodologically important because they show that the predictive gain is concentrated in evidence that a measurement-assurance program can actually document: prior drift behavior and calibration condition, rather than opaque identifiers or arbitrary administrative schedule fields.

Table 12. Feature-Ablation Performance on Shifted Future Data

Model variant	MAE (months)	Change vs full model
Full model	5.80	—
Without prior drift rate	8.42	+2.62
Without last as-found error	6.73	+0.93
Without condition variables	6.48	+0.68
Class-median baseline	9.85	+4.05

Figure 15: Feature-Ablation Analysis of Shifted-Test Prediction Error



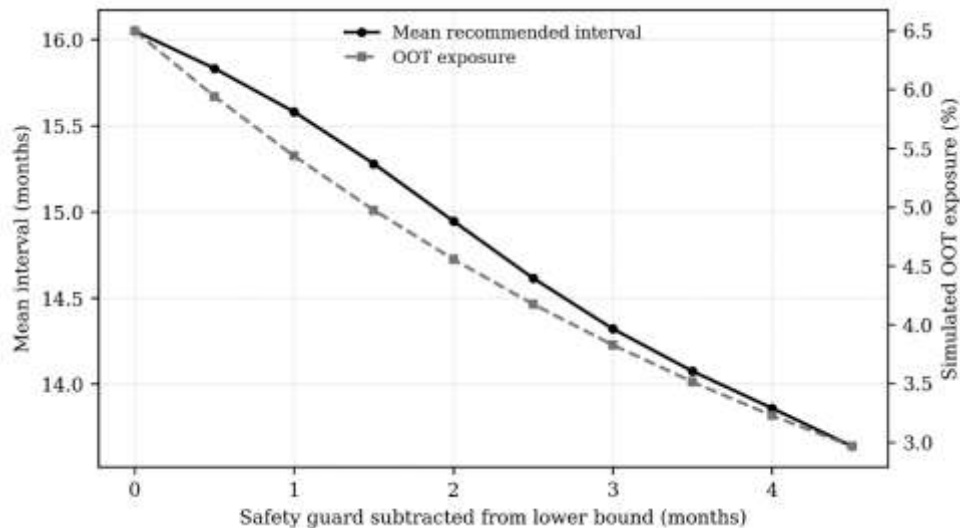
Safety-Guard Sensitivity Analysis

The explicit safety guard produced the expected monotonic tradeoff. With little or no guard, the policy produced longer intervals but higher OOT exposure; increasing the guard progressively shortened the average interval and reduced risk. The two-month guard used in the principal experiment lies near the middle of the simulated frontier and was selected to illustrate a conservative but not maximally restrictive policy. In practice, the guard should not be tuned solely to minimize a statistical loss function. It should be justified through criticality analysis, measurement decision risk, uncertainty, external calibration limits, and the organization's tolerance for false extension. The principle is analogous to guard banding in conformity assessment, where an acceptance boundary is intentionally moved to manage incorrect-decision risk (ILAC, 2019; Dobbert, 2008).

Table 13. Sensitivity of the Risk-Constrained Policy to the Safety Guard

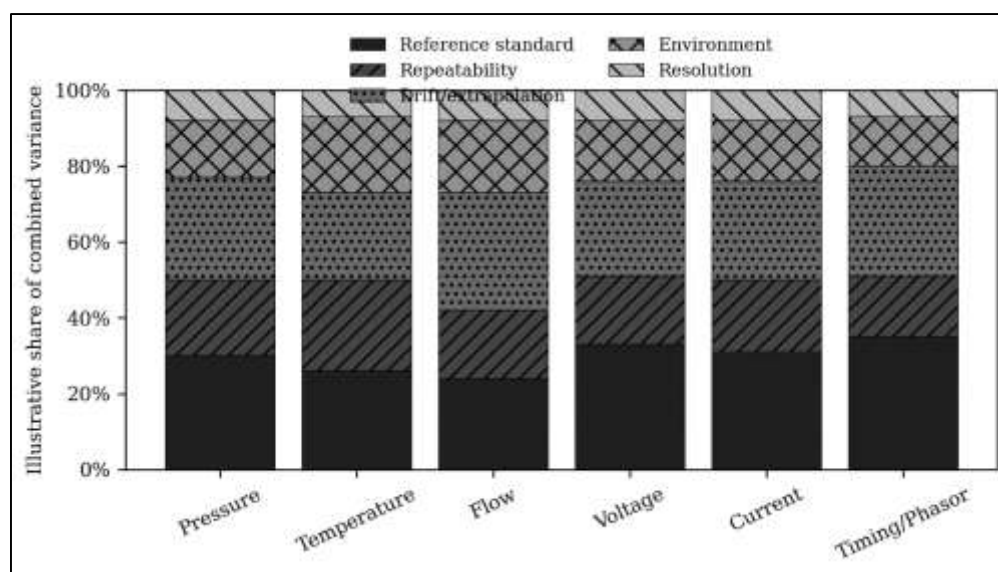
Safety guard (months)	Mean interval (months)	Simulated OOT exposure
0	16.05	6.50%
1	15.58	5.44%
2	14.95	4.56%
3	14.32	3.83%
4	13.86	3.23%

Figure 16: Safety-Guard Sensitivity of Calibration Interval and Simulated OOT Exposure



Uncertainty-Budget Composition

Because calibration interval decisions interact with the adequacy of the underlying measurement process, the study also presents an illustrative uncertainty decomposition. The composition differs by class: electrical and timing channels place a larger modeled share on reference-chain contributions, flow instruments place greater emphasis on drift/extrapolation, and temperature measurements show a relatively larger environmental contribution. The purpose is not to prescribe a generic uncertainty budget but to emphasize that the same observed as-found error has different decision significance when the uncertainty-to-tolerance ratio is different. NIST guidance and the GUM require uncertainty components to be identified, combined, and reported in a defensible measurement model rather than treated as a single metadata flag (Taylor & Kuyatt, 1994).

Figure 17: Relative Contributions to Combined Measurement Variance by Instrument Class

FINDINGS

The first major finding is that calibration history carries strong predictive information when combined with contextual condition variables. In the synthetic data, prior drift rate and recent as-found error dominated permutation importance, and the AI model reduced shifted-test MAE from 9.85 months for the class-median baseline to 5.80 months. This result reinforces the central calibration-management principle that interval decisions should respond to evidence about actual instrument behavior rather than only to model type or calendar age. It is also consistent with ILAC/OIML guidance that historical calibration, maintenance, usage, and observed drift are relevant when reviewing intervals (ILAC & OIML, 2022). The second finding is that predictive uncertainty is operationally important. A point model remained reasonably informative under distribution shift, but the static 90% interval lost coverage, falling from 90.8% in the stationary validation regime to 84.5% after the latent drift-rate increase. This is precisely the type of behavior that can be hidden when a maintenance model is reported only through MAE or R^2 . Rolling recalibration of the conformal correction restored coverage to 89.8%, showing that the uncertainty layer itself can serve as a model-health indicator and adaptation mechanism. The third finding is that personalization can improve the risk-efficiency frontier. The fixed nine-month policy was conservative in calendar terms but applied the same burden to all instruments. The AI policy used shorter intervals for high-risk cases and longer intervals for stable cases, producing 4.40% OOT exposure with 3.46 calibrations per 36 months, compared with 5.48% exposure and 4.00 calibrations for fixed nine months. In the simulated environment, the value of AI did not come from simply extending intervals; it came from reallocating calibration effort toward instruments with smaller predicted safety margins. The fourth finding is that drift detection should be treated as part of calibration assurance rather than as a separate data-science feature. EWMA monitoring of standardized residuals detected 74.7% of simulated drifts within 12 weeks with a 5.6% no-drift false-alarm rate. A drift alarm can be connected to conservative actions such as freezing interval extensions, requiring an intermediate check, shortening the due date, or triggering model review. This is consistent with broader concept-drift research that separates detection from subsequent adaptation decisions (Gama et al., 2014; Lu et al., 2019). The fifth finding concerns architecture and governance. NIST traceability cannot be reduced to a model field labeled “NIST-traceable.” The traceability chain, calibration evidence, and uncertainty evaluation must remain auditable independent of the AI layer (NIST, n.d.). The AI model can determine how urgently the next traceable calibration is needed, but it should not alter the reference chain or create an unsupported traceability claim. This distinction is essential to prevent technical optimization from becoming a compliance shortcut.

DISCUSSION

The simulation results align with the logic of modern interval guidance. NIST does not recommend a universal fixed interval and instead points laboratories toward accuracy requirements, stability, environmental factors, and measurement-assurance data (NIST, 2026). ILAC/OIML likewise recommends ongoing review of intervals as evidence accumulates and explicitly allows either shortening or lengthening where supported by drift and historical data (ILAC & OIML, 2022). The AI framework can be understood as a scalable implementation of this evidence-review process, not as a

competing metrological doctrine. The model's dominant reliance on prior drift rate and as-found error is also conceptually appropriate. These variables directly describe how the instrument has behaved rather than inferring degradation solely from generic age or environment. In real implementation, however, a strong dependence on one feature should trigger a data-quality review. A corrupted calibration-history field could otherwise drive many interval recommendations. Feature lineage, validation rules, and independent checks are therefore part of the safety case, not merely data-engineering hygiene.

Why Predictive Intervals Matter More Than Point Accuracy

The results demonstrate an important difference between predictive maintenance and safety-constrained calibration. The median model remained useful under distribution shift, but its error increased and static coverage fell. A point estimate does not reveal that the model has become less reliable for a particular decision. Predictive intervals provide a direct mechanism for conservatism: when uncertainty widens, the lower bound moves earlier, which shortens the allowed interval. This behavior is naturally compatible with the conformity-assessment principle that uncertainty affects the confidence of a pass/fail decision (JCGM, 2012). Conformal methods are especially useful because they can be combined with nonlinear models without requiring a fully specified parametric error distribution. However, their guarantees are not magical. Standard conformal coverage depends on exchangeability, and time-dependent maintenance data often violate that condition. The observed drop from 90.8% to 84.5% coverage under shift demonstrates this limitation empirically. Adaptive conformal methods and methods designed for nonexchangeability provide a more rigorous research direction (Gibbs & Candès, 2021; Barber et al., 2023). A production system should therefore monitor both realized coverage and the distribution of nonconformity scores rather than assuming that the original calibration remains valid indefinitely.

Calibration Burden, Risk, and Criticality

The policy comparison clarifies why a simple goal such as “maximize the interval” is inappropriate. Fixed 18 months minimized scheduled calibration burden in the simulation but produced the highest OOT exposure. Fixed nine months reduced risk but forced stable instruments to follow the same schedule as unstable instruments. The AI policy occupied a different part of the frontier by using the lower predictive bound and a safety guard. This produced fewer calibrations than fixed nine months without increasing simulated OOT exposure. The result suggests that economic benefit should be measured as avoided unnecessary calibration at a controlled risk level, not as the average interval extension. Real energy infrastructure would require a severity-weighted version of this analysis. An out-of-tolerance event on a noncritical display transmitter is not equivalent to a measurement feeding a safety trip, custody-transfer calculation, protection element, or synchronizing function. Future field work should therefore assign consequence categories and use expected loss or risk matrices rather than a single OOT percentage. Some safety-critical instruments may also have non-negotiable maximum proof-test or calibration intervals established by standards, regulation, engineering analysis, or site procedures. The AI policy should treat those values as hard constraints.

Drift Detection as an Operational Safety Net

The residual-monitoring experiment supports a layered approach. Rather than wait for the next calibration to discover deterioration, the system monitors whether current behavior remains consistent with the model's expected residual distribution. This is analogous to measurement assurance, where check standards and control charts provide evidence that a process remains in control between calibrations (Croarkin, 1985). In the AI framework, residual surveillance can be driven by periodic check measurements, redundant channels, process-balance residuals, comparison to reference events, or calibration opportunities that occur before the due date. EWMA and CUSUM are deliberately interpretable choices, which is advantageous in high-consequence settings. More sophisticated drift detectors, adaptive windows, domain adaptation, or AutoML compensation can also be used (Bifet & Gavaldà, 2007; Ma et al., 2018; Schaller et al., 2025). The key requirement is not algorithmic novelty but a validated alarm-to-action pathway. An alarm should have a predefined meaning: for example, suspend interval extensions, require a check standard within 48 hours, route the instrument for engineering review, or revert the population to a conservative baseline interval.

Traceability, Laboratory Competence, and Human Oversight

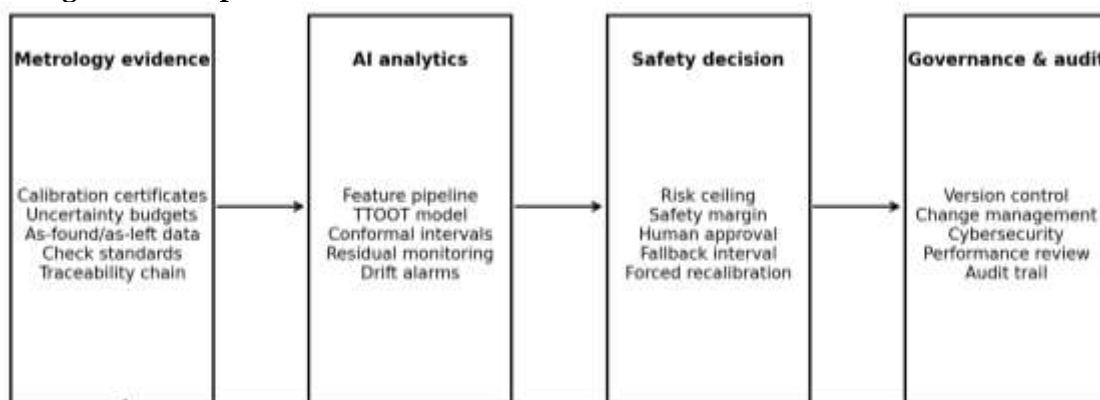
The most important boundary in the framework is that AI does not confer traceability. NIST expressly states that merely calibrating an instrument at NIST is not enough to make subsequent measurement results traceable unless the full chain, uncertainties, and measurement-control evidence are established (NIST, n.d.). Likewise, ISO/IEC 17025 addresses competence and consistent operation of testing and calibration laboratories (ISO & IEC, 2017). An AI scheduler should therefore consume validated calibration records from competent processes and produce recommendations that are reviewed within the quality system. It should not generate or modify calibration values, uncertainty statements, or certificates unless those functions are separately validated and authorized. Human oversight is necessary because the data may omit context known to engineers or metrologists. An instrument may be scheduled for replacement, its process may have changed, a safety analysis may require a fixed proof-test interval, a calibration laboratory may have changed methods, or a recent modification may invalidate historical comparability. The authorized reviewer should be able to see the recommended interval, lower predictive bound, major contributing features, current drift status, data completeness, applicable hard constraints, and previous interval. The system should also record the reviewer's decision and reason code so that overrides become future governance evidence rather than undocumented exceptions.

Cybersecurity and Model-Change Control

Because calibration schedules can affect equipment availability and safety assurance, the AI service is part of the broader OT risk environment. NIST SP 800-82 Rev. 3 emphasizes security measures tailored to OT reliability and safety requirements (Stouffer et al., 2023). A compromised calibration model could create a subtle denial-of-maintenance attack by extending due dates, or an operational disruption by forcing excessive calibrations. Recommended controls therefore include read-only ingestion where possible, signed model artifacts, role-based authorization for schedule changes, protected audit logs, source-data authentication, network segmentation appropriate to site architecture, rollback capability, and independent monitoring for anomalous recommendation patterns.

Model changes should be treated as controlled configuration changes. Retraining on new data can improve adaptation but can also erase a previously validated performance profile. A mature program should retain versioned training data, feature definitions, code, hyperparameters, conformal calibration records, performance thresholds, and approval evidence. The NIST AI RMF provides a useful governance structure for identifying, measuring, managing, and monitoring AI risks across the lifecycle (Tabassi, 2023).

Figure 10: Proposed AI-Enabled NIST-Traceable Calibration Assurance Model



Proposed Operational Decision Logic

A practical implementation can use a hierarchical decision rule. First, verify that the instrument's calibration certificate, uncertainty, traceability metadata, criticality, and current status are complete. If not, AI extension is prohibited. Second, compute the TTOOT interval and model-health indicators. Third, apply hard limits such as regulatory maximum interval, safety-integrity requirements, site policy, and manufacturer restrictions. Fourth, calculate the risk-constrained recommendation from the lower predictive bound. Fifth, check drift alarms and intermediate-check status; an active alarm forces

a shorter interval or immediate review. Sixth, route any extension beyond the current approved interval to an authorized metrology or reliability reviewer. Seventh, record the decision and monitor outcomes so that later as-found results can validate whether the recommendation was conservative.

Reliability Engineering and First-Passage-Time Modeling

The supplementary reliability analysis strengthens the article's central argument that calibration interval is a stochastic reliability decision rather than a calendar attribute. Reliability models such as Weibull, lognormal, Wiener-process, and Bayesian formulations provide explicit assumptions about degradation and first passage. Their advantage is interpretability: the analyst can inspect a survival function, an estimated hazard, and the uncertainty of the interval estimate. Their limitation is structural rigidity when heterogeneous environmental and operating covariates influence degradation in nonlinear ways. Machine learning has the complementary strength of flexible conditional prediction. A defensible program can therefore use reliability models as priors, benchmarks, or policy constraints and use machine learning to model conditional departures from the population-level reliability curve. This hybrid interpretation is consistent with Bayesian interval research ([Carvajal et al., 2022](#)), MCMC reliability approaches ([Shim & Park, 2022](#)), and recent nuclear transmitter reliability modeling based on first-passage time ([Liu et al., 2026](#)).

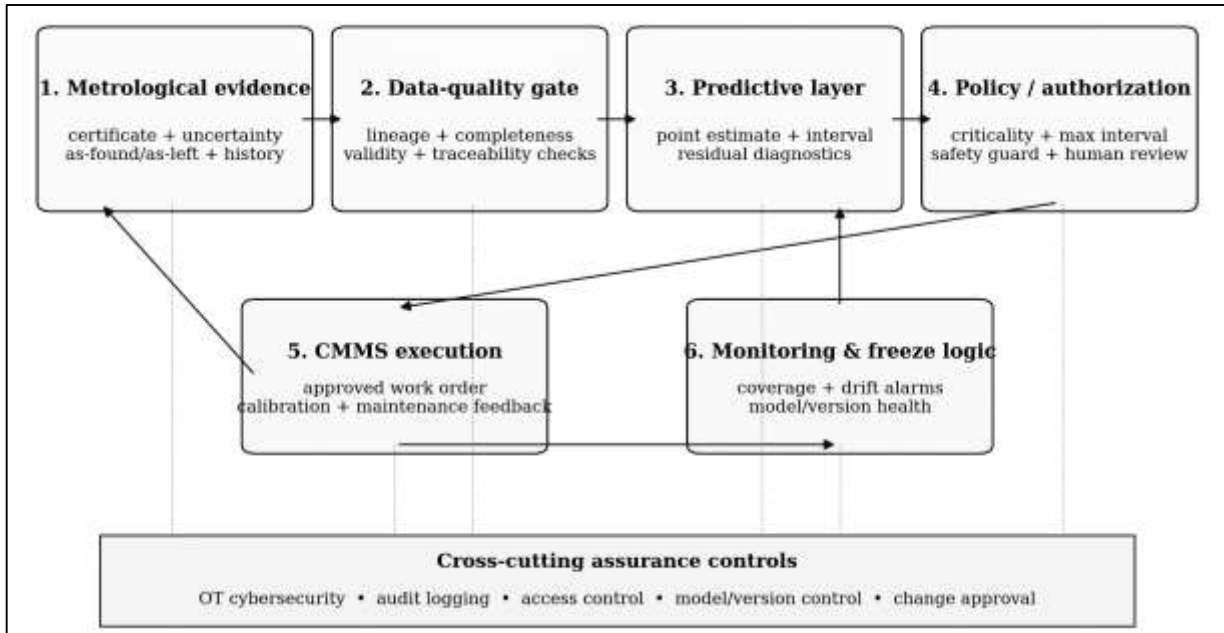
Conformity Risk and Guard-Banded Scheduling

A notable conceptual distinction is required between a calibration decision rule and a predictive scheduling rule. During calibration, a decision rule determines whether a measured result conforms to a specification given measurement uncertainty. Before calibration, a scheduling rule determines when sufficient evidence exists to justify re-evaluation. Both problems involve uncertainty and asymmetric consequences, but they operate on different objects. The article's lower-bound scheduling rule is deliberately conservative because it bases extension on an uncertainty-aware lower estimate of time to threshold. The added safety guard then provides a second margin. This two-layer structure is analogous in spirit to guard-banded conformity decisions, but the margin should be justified independently from the guard band that a calibration laboratory may apply to the measurement result. Conflating the two could create double-counting or, more seriously, give the false impression that a predictive model has altered the underlying tolerance or uncertainty statement.

Assurance Architecture for Production Deployment

Production use requires an assurance case rather than a standalone prediction service. Metrological evidence enters first and is checked for completeness, validity, and traceability. Only then are derived features supplied to the predictive model. The model produces point and interval forecasts, but policy guardrails impose criticality, maximum-interval, proof-test, safety-guard, and data-quality constraints. A qualified human reviewer authorizes any extension. The approved action is transferred to the computerized maintenance management system, and the resulting calibration record feeds back into the history after quality review. In parallel, residual and coverage monitoring can freeze interval extensions when the statistical behavior departs from the validated regime. Cybersecurity, model versioning, audit logging, access control, and change approval operate across the full pipeline. This architecture also makes clear why metrological traceability belongs to the measurement evidence chain, not to the AI model itself. NIST states that traceability is a property of a measurement result supported by a documented unbroken chain of calibrations, each contributing to uncertainty; merely using an instrument calibrated by NIST does not automatically make later measurements traceable ([Bruce et al., 2021](#); [NIST, n.d.](#)).

Figure 18: Layered Assurance Architecture for AI-Enabled Predictive Calibration



CONCLUSION

This study has developed and evaluated a simulation-based framework for AI-enabled predictive calibration intervals and drift detection for NIST-traceable instrumentation in safety-critical energy infrastructure. The framework integrates calibration history, measurement uncertainty, operating stress, quantile machine learning, conformal predictive intervals, online residual surveillance, explicit safety margins, and governance controls. Its central contribution is architectural: AI is positioned as a controlled advisor for the timing of traceable calibration, not as a substitute for traceability, uncertainty evaluation, competent calibration, or functional-safety requirements. The simulation demonstrated that calibration-history and condition features can meaningfully improve TTOOT prediction relative to a simple class-based baseline. The median model achieved 3.99-month MAE on stationary future data and 5.80-month MAE after an imposed distribution shift, compared with 9.85 months for the baseline on shifted data. A 90% conformal interval achieved 90.8% coverage before shift, fell to 84.5% under static application after shift, and recovered to 89.8% with a rolling correction. These results show that uncertainty and distribution shift must be treated as first-class components of a calibration decision rather than as secondary statistical diagnostics.

The risk-constrained policy also illustrated the potential operational value of personalization. With a two-month safety guard and conservative use of the lower predictive bound, the AI policy produced 4.40% simulated OOT exposure and 3.46 calibrations per 36 months, compared with 5.48% exposure and four calibrations for a fixed nine-month policy. At the same time, a fixed 12-month policy reduced calibration count but produced substantially higher simulated OOT exposure. The result demonstrates that the value of predictive calibration lies in redistributing maintenance effort according to estimated risk, not in indiscriminately lengthening intervals. Finally, the drift experiment showed that residual surveillance can provide an additional safety net between scheduled calibration events. An EWMA monitor detected 74.7% of simulated drifts within 12 weeks with a 5.6% no-drift false-alarm rate. In deployment, such alarms should trigger predefined conservative actions and model review rather than automatic self-correction of safety requirements. The overall conclusion is that AI can strengthen calibration management when predictive uncertainty, drift detection, traceability evidence, functional-safety constraints, OT cybersecurity, and human authorization are designed as one assurance system.

RECOMMENDATIONS

First, energy operators should begin with a traceability and data-quality foundation rather than with model selection. Calibration certificates, as-found/as-left values, uncertainty statements, instrument identifiers, reference chains, maintenance events, and environmental histories should be standardized

and validated before AI interval optimization is attempted. Second, predictive calibration should use chronological validation. Random cross-validation that mixes future and past conditions can conceal sensor or process drift and overstate deployment performance. Evaluation should include at least one deliberately shifted or later-time holdout. Third, interval recommendations should be generated from a conservative predictive bound and should incorporate explicit safety guards. Point forecasts should be displayed for information but should not alone authorize due-date extensions in high-consequence applications. Fourth, organizations should define hard constraints outside the AI model. Regulatory maximums, functional-safety proof-test requirements, instrument criticality, manufacturer limits, uncertainty-ratio thresholds, and site policies must override the predictive recommendation.

Fifth, model residuals and predictive-interval coverage should be monitored continuously. Drift alarms should have predefined actions, such as freezing extensions, performing an intermediate check, shortening intervals, or reverting a population to a conservative fixed schedule. Sixth, calibration-scheduling AI should be deployed with OT-appropriate cybersecurity controls, including least privilege, protected data lineage, signed model versions, audit logs, change approval, network segmentation where appropriate, and the ability to roll back to a validated prior model. Seventh, interval extensions should remain subject to human technical authorization. Review screens should expose the lower predictive bound, major features, drift status, data-quality warnings, applicable hard limits, and the previous interval so that decisions are explainable and auditable. Eighth, field pilots should begin with non-safety or lower-criticality instruments that have rich traceable histories, then progress to higher criticality only after demonstrating acceptable OOT rates, coverage, alarm performance, override behavior, and governance effectiveness. Ninth, future validation should use consequence-weighted risk rather than a binary OOT indicator. The financial and safety consequence of a measurement deviation should be incorporated into the scheduling objective so that calibration resources are allocated according to both probability and impact.

Tenth, any production claim of benefit should be based on actual respondent-independent records or field calibration outcomes. Synthetic results such as those in this article are useful for method development but should never be presented as measured performance of a specific utility, laboratory, power plant, refinery, pipeline, or grid operator. Eleventh, laboratories should maintain class-specific validation thresholds and should not approve a portfolio-wide calibration model solely on aggregate MAE or coverage. Instruments whose function, technology, uncertainty ratio, or environmental exposure differs materially should be validated separately.

Twelfth, reliability models should be retained as interpretable benchmark models even when a nonlinear AI model is selected for production. Disagreement between the AI forecast and a class-specific survival or first-passage model should trigger technical review rather than automatic averaging. Thirteenth, uncertainty-to-tolerance ratio and the applicable conformity decision rule should be displayed with the scheduling recommendation so that reviewers can distinguish a statistically long predicted lifetime from a measurement process that has inadequate decision capability. Fourteenth, organizations should define a formal model-freeze condition. Persistent predictive-interval undercoverage, abnormal widening of intervals, residual drift alarms, missing calibration lineage, or a major process/configuration change should suspend interval extension until revalidation is completed.

LIMITATIONS

This study has several limitations. First, the numerical results are generated from a synthetic, physics-inspired statistical model rather than from actual calibration records. The feature relationships, drift distributions, shock probabilities, and TTOOT values were deliberately constructed to create a realistic analytical challenge, but they do not represent any specific instrument manufacturer, calibration laboratory, utility, plant, pipeline, or energy operator. The reported MAE, R^2 , coverage, OOT exposure, calibration savings, and detector performance must therefore be interpreted as simulation results only. Second, the synthetic degradation process is simpler than real metrological behavior. Actual instruments may exhibit hysteresis, discontinuous adjustment effects, process-dependent bias, sensor replacement, technician effects, reference-standard changes, firmware changes, calibration-method changes, seasonal environments, common-cause disturbances, or multimodal failure mechanisms. These factors may require hierarchical, survival, state-space, or technology-specific models rather than

the generic gradient-boosted quantile model used here.

Third, the conformity logic uses a simplified OOT boundary and an uncertainty ratio rather than a complete measurement model or application-specific decision rule. Real conformity decisions should follow the applicable tolerance, uncertainty budget, guard-band, and regulatory or contractual rule, informed by the principles of the GUM and JCGM 106. Fourth, the adaptive conformal procedure is a rolling empirical correction, not a full implementation of advanced conformal methods developed for arbitrary distribution shift. The simulation uses those methods as conceptual motivation, and future work should directly compare adaptive conformal inference, weighted conformal methods, and other time-series uncertainty techniques. Fifth, the drift experiment monitors standardized residual means and does not model every type of sensor drift. Variance shifts, covariance changes, nonlinear distortions, class-specific drift, and reference-system drift may require different detectors. The 80-week stream and chosen thresholds were selected for demonstration and should not be interpreted as recommended alarm settings. Sixth, the policy comparison simplifies operational cost. It counts calibrations but does not monetize labor, outage time, bypass risk, spare requirements, turnaround time, or consequence severity. A real decision model should optimize risk-adjusted lifecycle cost subject to safety and compliance constraints rather than a binary OOT proportion alone. Seventh, cybersecurity, human factors, and governance are analyzed conceptually rather than experimentally. A field deployment should include threat modeling, role design, usability testing, change-management assessment, and failure-mode analysis for the AI-enabled workflow. Finally, the article uses author and journal metadata placeholders because the user did not provide publication-specific author details; those fields must be completed before submission.

Eighth, the class-specific reliability curves and uncertainty-budget shares added in the supplementary analyses are deliberately synthetic. They demonstrate analytical structure and sensitivity, not empirical reliability estimates for any particular manufacturer, sensor technology, power plant, grid operator, or calibration laboratory. Ninth, the study does not evaluate organizational acceptance, human factors, or decision latency associated with expert review of AI recommendations. A production pilot should measure not only model accuracy but also reviewer agreement, override reasons, alarm burden, calibration-work-order lead time, and whether the interface supports accurate interpretation of uncertainty and traceability evidence.

REFERENCES

- [1]. Abdullah Mohammad, S. (2023). CUGAO₂ and CUINS₂ Quantum Dot Sensitization for Enhanced Photovoltaic Efficiency in Wearable Fiber-Shaped Solar Cells. *American Journal of Advanced Technology and Engineering Solutions*, 3(04), 209-261. <https://doi.org/10.63125/wysdvy71>
- [2]. Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494-591. <https://doi.org/10.1561/2200000101>
- [3]. Ashrafuzzaman, M. (2024). The impact of cloud-based management information systems on hrm efficiency: an analysis of small and medium-sized enterprises (SMES). *Academic Journal on Artificial Intelligence, Machine Learning, Data Science and Management Information Systems*, 1(01), 40-56. <https://doi.org/10.69593/ajaimldsms.v1i01.124>
- [4]. Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2), 816-845. <https://doi.org/10.1214/23-AOS2276>
- [5]. Bifet, A., & Gavalda, R. (2007). Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM International Conference on Data Mining* (pp. 443-448). Society for Industrial and Applied Mathematics. <https://doi.org/10.1137/1.9781611972771.42>
- [6]. Bruce, S. S., Possolo, A., & Watters, R. L., Jr. (2021). Metrological traceability: Frequently asked questions and NIST policy (NIST Technical Note 2156). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.TN.2156>
- [7]. Carvajal, S. A., Medina, A. F., Bohórquez, A. J., & Sánchez, C. A. (2022). Estimation of calibration intervals using Bayesian inference. *Measurement*, 187, 110316. <https://doi.org/10.1016/j.measurement.2021.110316>
- [8]. Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. da P., Basto, J. P., & Alcalá, S. G. S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137, 106024. <https://doi.org/10.1016/j.cie.2019.106024>
- [9]. Croarkin, C. (1985). Measurement assurance programs: Part II—Development and implementation (NBS Special Publication 676-II). National Bureau of Standards. <https://emtoolbox.nist.gov/Publications/NBSSpecialPublication676-II.pdf>
- [10]. Debasish, A. (2021). Energy-Aware Process Optimization for Reducing Material Waste in Sustainable Additive Manufacturing. *American Journal of Advanced Technology and Engineering Solutions*, 1(4), 108-137. <https://doi.org/10.63125/j0wk7j48>

- [11]. Debasish, A. (2022). Simulation-Based Thermal Management Framework for Reducing Energy Consumption in Advanced Manufacturing Systems. *International Journal of Scientific Interdisciplinary Research*, 1(01), 335–381. <https://doi.org/10.63125/6fcj7c89>
- [12]. Debasish, A. (2023). Thermal Behavior Modeling for Predicting Process Instability in Additive Manufacturing Systems. *International Journal of Scientific Interdisciplinary Research*, 4(4), 485–528. <https://doi.org/10.63125/b2nemb48>
- [13]. Debasish, A. (2024). AI-Based Quality Control Framework for Reducing Rework and Scrap in Additive Manufacturing Production. *Review of Applied Science and Technology*, 3(04), 402–447. <https://doi.org/10.63125/z29tv526>
- [14]. Debasish, A. (2025a). Design-for-Manufacturing Optimization for Improving Resilience in U.S. Additive Manufacturing Supply Chains. *American Journal of Interdisciplinary Studies*, 6(02), 185–204. <https://doi.org/10.63125/hxm3tb71>
- [15]. Debasish, A. (2025b). Physics-Informed Machine Learning for Predicting Thermal Distortion in Metal Additive Manufacturing. *American Journal of Scholarly Research and Innovation*, 4(01), 943–989. <https://doi.org/10.63125/ny974y17>
- [16]. Dobbert, M. (2008). A guard-band strategy for managing false-accept risk. *NCSLI Measure*, 3(4), 44–48. <https://doi.org/10.1080/19315775.2008.11721446>
- [17]. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), Article 44. <https://doi.org/10.1145/2523813>
- [18]. Gibbs, I., & Candès, E. J. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34, 1660–1672. <https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html>
- [19]. Gopstein, A., Nguyen, C., O’Fallon, C., Hastings, N., & Wollman, D. A. (2021). NIST framework and roadmap for smart grid interoperability standards, Release 4.0 (NIST SP 1108 Rev. 4). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.1108r4>
- [20]. Institute of Electrical and Electronics Engineers & International Electrotechnical Commission. (2018). IEEE/IEC 60255-118-1-2018: Measuring relays and protection equipment—Part 118-1: Synchrophasor for power systems—Measurements. <https://standards.ieee.org/ieee/60255-118-1/5724/>
- [21]. International Electrotechnical Commission. (2010). IEC 61508-1:2010: Functional safety of electrical/electronic/programmable electronic safety-related systems—Part 1: General requirements. <https://webstore.iec.ch/en/publication/5515>
- [22]. International Electrotechnical Commission. (2016). IEC 61511-1:2016: Functional safety—Safety instrumented systems for the process industry sector—Part 1: Framework, definitions, system, hardware and application programming requirements. <https://webstore.iec.ch/en/publication/24241>
- [23]. International Laboratory Accreditation Cooperation & International Organization of Legal Metrology. (2022). Guidelines for the determination of recalibration intervals of measuring equipment (ILAC-G24:2022/OIML D 10:2022). https://www.oiml.org/en/files/pdf_d/d010-e22.pdf
- [24]. International Laboratory Accreditation Cooperation. (2019). ILAC-G8:09/2019: Guidelines on decision rules and statements of conformity. <https://ilac.org/publications-and-resources/publications-list/>
- [25]. International Organization for Standardization & International Electrotechnical Commission. (2017). ISO/IEC 17025:2017: General requirements for the competence of testing and calibration laboratories. <https://www.iso.org/standard/66912.html>
- [26]. Janatyan, N., Alavi, S., & Shafiee, E. (2025). Prediction of measuring instrument calibration interval based on risk priority number and reliability using machine learning. *Computers & Industrial Engineering*, 210, 111570. <https://doi.org/10.1016/j.cie.2025.111570>
- [27]. Jannatul, F. (2025). Spinal Cord Injury and Neurorehabilitation Deep Learning–Based Prediction of Functional Recovery Trajectories After Traumatic Spinal Cord Injury. *ASRC Procedia: Global Perspectives in Science and Scholarship*, 1(01), 2865–2922. <https://doi.org/10.63125/m738bh82>
- [28]. Joint Committee for Guides in Metrology. (2008). Evaluation of measurement data—Guide to the expression of uncertainty in measurement (JCGM 100:2008). <https://doi.org/10.59161/JCGM100-2008E>
- [29]. Joint Committee for Guides in Metrology. (2012). Evaluation of measurement data—The role of measurement uncertainty in conformity assessment (JCGM 106:2012). <https://doi.org/10.59161/JCGM106-2012>
- [30]. Joint Committee for Guides in Metrology. (2012). International vocabulary of metrology—Basic and general concepts and associated terms (VIM), 3rd ed. (JCGM 200:2012). <https://doi.org/10.59161/JCGM200-2012>
- [31]. Kopke, V., Mourão, S. C., & Brito, M. (2024). Determination of the calibration interval of measuring instruments: Which method should I use? *IEEE Instrumentation & Measurement Magazine*, 27(2), 59–62. <https://doi.org/10.1109/MIM.2024.10472985>
- [32]. Lei, J., G’Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523), 1094–1111. <https://doi.org/10.1080/01621459.2017.1307116>
- [33]. Liu, B., Cai, S., Su, H., Qiao, Z., Wei, C., Zhao, X., Jiang, J., Jiao, M., Qin, F., & He, M. (2026). Reliability modeling of field calibration data for pressure transmitters in nuclear power plants. *Nuclear Engineering and Technology*, 58(1), 103878. <https://doi.org/10.1016/j.net.2025.103878>
- [34]. Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12), 2346–2363. <https://doi.org/10.1109/TKDE.2018.2876857>

- [35]. Ma, Z., Luo, G., Qin, K., Wang, N., & Niu, W. (2018). Online sensor drift compensation for E-nose systems using domain adaptation and extreme learning machine. *Sensors*, 18(3), 742. <https://doi.org/10.3390/s18030742>
- [36]. Md Anisur Rahman, M. (2024). State Income Tax Differentials and High-Skilled Labor Mobility in The United States: A Time Series Analysis. *American Journal of Interdisciplinary Studies*, 5(04), 222-267. <https://doi.org/10.63125/295bsn47>
- [37]. Md Anisur Rahman, M. (2025). The Role of AI-Enabled Business Intelligence in Financial Risk Assessment and Liquidity Monitoring. *Review of Applied Science and Technology*, 4(04), 195-246. <https://doi.org/10.63125/p4eezd93>
- [38]. Md Arif Uz, Z., Sharmin, S., & Md Abdur, R. (2025). Factors Influencing Behavioural Intention to Use he Innovative Open And Distance Learning Systems With The Role Of Continuance Intention. *American Journal of Scholarly Research and Innovation*, 4(01), 202-220. <https://doi.org/10.63125/pbjxp014>
- [39]. Md Ashfaq, S., & Abdullah Mohammad, S. (2022). CNT/PANI Composite Electrodes for Flexible Fiber-Shaped Dye-Sensitized Solar Cells: Synthesis, Characterization, and Photovoltaic Performance. *American Journal of Interdisciplinary Studies*, 3(02), 221-274. <https://doi.org/10.63125/m334mf41>
- [40]. Md Asif Ali Sheak, A., & Risha, A. (2022). The Role of IOT-Driven Energy Analytics in Improving Operational Efficiency and Sustainability in Modern Industrial Facilities. *American Journal of Interdisciplinary Studies*, 3(04), 807-841. <https://doi.org/10.63125/gm63q956>
- [41]. Md Kaium, H. (2023). A GIS-Based Multi-Criteria Decision-Support and Land-Suitability Optimization Framework for Climate-Resilient Urban and Regional Planning. *American Journal of Data Science and Analytics*, 4(06), 43-103. <https://doi.org/10.63125/wvq0sx87>
- [42]. Md Kaium, H. (2024). Machine Learning Enhanced GIS And AHP Modeling for High-Resolution Flood-Susceptibility Prediction, Vulnerability Assessment, and Risk Classification. *Review of Applied Science and Technology*, 3(02), 92-140. <https://doi.org/10.63125/fz8hv245>
- [43]. Md Kaium, H. (2025). Integrating Artificial Intelligence, Semantic Indoor Geospatial Modeling, And Network-Based Routing for High-Precision NG911 Emergency Location Intelligence. *International Journal of Scientific Interdisciplinary Research*, 6(1), 616-670. <https://doi.org/10.63125/5hc1xx56>
- [44]. Md Mahamud Pervaz, P. (2023). Integration of MES, SAP, And PI Historian Platforms for Real-Time Industrial Asset Performance Management. *Review of Applied Science and Technology*, 2(04), 393-445. <https://doi.org/10.63125/9ebtas97>
- [45]. Md Mahamud Pervaz, P. (2025). Development of a CMMS-Integrated Predictive Maintenance Framework for Industrial PLC and Drive Systems. *American Journal of Scholarly Research and Innovation*, 4(01), 894-942. <https://doi.org/10.63125/j33gtf31>
- [46]. Md. Abdur, R., & Iftekhar, A. (2021). Customer Retention Forecasting in Mobile Wallet Services Using Neural Networks: A Comparative Quantitative Study. *International Journal of Business and Economics Insights*, 1(4), 70-102. <https://doi.org/10.63125/dyrpc387>
- [47]. Md. Rashedul, I., & Md Kaium, H. (2022). Development Of a Three-Dimensional Indoor Geospatial Network and Address Geocoding Framework for Next-Generation Emergency Dispatch and Response-Time Optimization. *Journal of Sustainable Development and Policy*, 1(02), 245-292. <https://doi.org/10.63125/tcw3mp48>
- [48]. Md. Shakhawat, H., & Md, F. (2024). Impact of Phishing Simulation Campaigns and Security Awareness Training on Human Risk Mitigation in Financial Institutions: A Quantitative Study. *American Journal of Data Science and Analytics*, 5(10), 48-85. <https://doi.org/10.63125/4j8h1484>
- [49]. Mohammad Abir Chowdhury, S. (2023). Role of AI-Assisted Risk Assessment in Enhancing Audit Planning Efficiency and Risk Prioritization in Public-Safety Organizations. *International Journal of Scientific Interdisciplinary Research*, 4(4), 529-576. <https://doi.org/10.63125/qd6nn437>
- [50]. Mohammad Abir Chowdhury, S. (2025). Application of Machine Learning and Anomaly-Detection Algorithms for Fraud Detection in U.S. Government Financial Operations. *American Journal of Interdisciplinary Studies*, 6(02), 139-184. <https://doi.org/10.63125/fkf19309>
- [51]. Mohammad Abir Chowdhury, S., & Tonay, P. (2022). An Empirical Analysis of the Impact of Robotic Process Automation and Continuous Auditing on the Timeliness And Accuracy Of Financial Reporting Assurance. *International Journal of Business and Economics Insights*, 2(4), 73-126. <https://doi.org/10.63125/7p8dcr47>
- [52]. Mohammad Badrul, A. (2025). SCADA And IOT-Enabled Smart Water Metering: A Systematic Review of Resilient Supply Monitoring. *International Journal of Scientific Interdisciplinary Research*, 6(1), 576-615. <https://doi.org/10.63125/at8pnx81>
- [53]. Mohammad Badrul, A., & Md. Mominul, H. (2023). Machine Learning-Based Assessment of Environmental and Public Health Risks in Sewerage Sludge Management Systems. *American Journal of Advanced Technology and Engineering Solutions*, 3(02), 33-77. <https://doi.org/10.63125/y3yfxa92>
- [54]. Mohammad Shoriful Hossan, S. (2022). A Performance-Based Framework for Arc-Flash Hazard Reduction and Electrical Reliability Improvement in Brownfield Industrial Distribution Systems. *Review of Applied Science and Technology*, 1(04), 447-463. <https://doi.org/10.63125/a70vj720>
- [55]. Mst Kaniz, F. (2022). Language Access and Health Equity: Role of Multilingual Administrative Staff in Reducing Healthcare Disparities Among South Asian Immigrant Communities in The U.S. *International Journal of Scientific Interdisciplinary Research*, 1(01), 274-315. <https://doi.org/10.63125/xqhss825>
- [56]. National Institute of Standards and Technology. (2026, May 29). Recommended calibration interval. <https://www.nist.gov/calibrations/recommended-calibration-interval>

- [57]. National Institute of Standards and Technology. (n.d.). NIST policy on metrological traceability. <https://www.nist.gov/calibrations/traceability>
- [58]. Nishat, J. (2023). Communicable Disease Prevention in a Post-Pandemic Era: A Global Review of Public Health Education, Surveillance, and Community-Level Interventions. *International Journal of Scientific Interdisciplinary Research*, 4(4), 630–678. <https://doi.org/10.63125/mnqefj36>
- [59]. Nishat, J. (2025). Leveraging Machine Learning to Identify and Address Maternal Health Disparities in Underserved U.S. Communities. *American Journal of Data Science and Analytics*, 6(12), 125-169. <https://doi.org/10.63125/c2t7r719>
- [60]. Nishat, J., & Jahanara, A. (2024). Machine Learning Approaches for Early Prediction of High-Risk Pregnancy in Underserved U.S. Populations. *American Journal of Interdisciplinary Studies*, 5(04), 268-309. <https://doi.org/10.63125/5gh5df13>
- [61]. Nishat, J., & Mst Kaniz, F. (2022). A Review of the Impact of Covid-19 On Maternal and Child Health Service Utilization in Bangladesh. *American Journal of Scholarly Research and Innovation*, 1(02), 223–269. <https://doi.org/10.63125/vgv5s742>
- [62]. Nuzhat Sadia, P. (2025). Integrating Business Intelligence Dashboards for Real-Time Supply Chain Visibility and Operational Decision Support. *American Journal of Data Science and Analytics*, 6(12), 170-216. <https://doi.org/10.63125/1es4f763>
- [63]. Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1–2), 100–115. <https://doi.org/10.1093/biomet/41.1-2.100>
- [64]. Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1(3), 239–250. <https://doi.org/10.1080/00401706.1959.10489860>
- [65]. Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized quantile regression. In *Advances in Neural Information Processing Systems* 32 (pp. 3543–3553). <https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html>
- [66]. Sadia, Z. (2025). Predictive Analytics for Financial Performance Forecasting and Reporting Quality: Evidence from Data-Driven Financial Management Systems. *ASRC Procedia: Global Perspectives in Science and Scholarship*, 1(01), 2795–2812. <https://doi.org/10.63125/zstqn346>
- [67]. Schaller, M. C., Kruse, M., Ortega, A., Lindauer, M., & Rosenhahn, B. (2025). AutoML for multi-class anomaly compensation of sensor drift. *Measurement*, 250, 117097. <https://doi.org/10.1016/j.measurement.2025.117097>
- [68]. Shaiful, M., Anisur, R., & Md, A. (2022). A Systematic Literature Review on the Role of Digital Health Twins in Preventive Healthcare For Personal and Corporate Wellbeing. *American Journal of Interdisciplinary Studies*, 3(04), 1-31. <https://doi.org/10.63125/negjw373>
- [69]. Shim, H.-S., & Park, S.-N. (2022). Numerical experiments on the application of Markov chain Monte Carlo for the establishment of calibration intervals. *IEEE Transactions on Instrumentation and Measurement*, 71, 1–10. <https://doi.org/10.1109/TIM.2022.3142005>
- [70]. Stouffer, K., Pease, M., Tang, C., Zimmerman, T., Pillitteri, V., Lightman, S., Hahn, A., Saravia, S., Sherule, A., & Thompson, M. (2023). Guide to operational technology (OT) security (NIST SP 800-82 Rev. 3). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-82r3>
- [71]. Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
- [72]. Taylor, B. N., & Kuyatt, C. E. (1994). Guidelines for evaluating and expressing the uncertainty of NIST measurement results (NIST Technical Note 1297). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.tn.1297>
- [73]. Tonay, P. (2025). Machine Learning for Fraud Detection in Enterprise Financial Systems. *American Journal of Advanced Technology and Engineering Solutions*, 1(02), 304-318. <https://doi.org/10.63125/d0fhp873>
- [74]. Tonay, P., Md Maruf, I., & Al, A. (2024). Artificial Intelligence-Based Decision Support Systems and Managerial Performance. *American Journal of Advanced Technology and Engineering Solutions*, 4(04), 154-184. <https://doi.org/10.63125/wmz8dc67>
- [75]. Uddin, H. M. M., & Risha, A. (2024). Hybrid Cloud Security and Compliance in U.S. Enterprises: Addressing Data Privacy Risks and Governance Challenges. *American Journal of Scholarly Research and Innovation*, 3(01), 153–197. <https://doi.org/10.63125/qcy0cj46>
- [76]. Wang, J., Zhang, Q., & Jiang, W. (2017). Optimization of calibration intervals for automatic test equipment. *Measurement*, 103, 87–92. <https://doi.org/10.1016/j.measurement.2017.01.062>
- [77]. Xu, C., & Xie, Y. (2023). Sequential predictive conformal inference for time series. *Proceedings of Machine Learning Research*, 202, 38707–38727. <https://proceedings.mlr.press/v202/xu23r.html>
- [78]. Zhang, Y., Liu, Z., Yang, C., Huang, X., Lou, S., Zhang, H., & Yan, D. (2024). Unveiling dynamics changes: Singular spectrum analysis-based method for detecting concept drift in industrial data streams. *Knowledge-Based Systems*, 293, 111640. <https://doi.org/10.1016/j.knosys.2024.111640>