



ARTICLE

Compliance-Integrated Data Integrity Framework for Large-Scale Sensor and Measurement Systems

Chirag Devendrakumar Parikh*

Computer Engineering, California State University, Fullerton, CA, USA

*Corresponding Author: Chirag Devendrakumar Parikh. Email: parikhchirag0723@gmail.com

Received: 07 December 2025; Accepted: 01 June 2026; Published: 07 September 2026

ABSTRACT: Sensors and large-scale measurement systems generate continuous data streams used in industrial monitoring, IoT analytics, and decision-making. However, sensor drift, undocumented maintenance, environmental stress, and component variability often degrade the integrity and reliability of collected measurements. This study proposes a compliance-integrated data integrity framework that combines hardware qualification records, traceability documentation, lifecycle validation checkpoints, and automated anomaly detection methods. The framework introduces five integrity layers that link sensor characterization with statistical filtering and compliance-driven verification. To validate feasibility, a proof-of-concept simulation was conducted using drift-injected sensor datasets. Results show that integrating compliance metadata improve anomaly detection accuracy compared to conventional statistical validation methods. The proposed approach provides a reproducible architecture for improving long-term consistency, authenticity, and trustworthiness of sensor data in distributed measurement environments.

KEYWORDS: Sensor data integrity; compliance engineering; drift detection; traceability and provenance; anomaly detection; lifecycle validation

1 Introduction

Mass sensors and measurement systems (As per Fig. 1) are the key to modern-day engineering, environmental measurement, industrialization, and IoT applications [1]. They result in sustained sources of large volumes of data that feed analytical models, generate decisions in real-time, and assist in planning operations on a long-term basis. The reliability of the datasets is critical as the number of distributed sensors grows. Even minimal variations brought about by degradation in hardware or stress in the environment, poor maintenance, or variation in components may skew analytical results and affect the performance of the system. Conventional big data pipelines emphasize processing efficiency, storage structure, and algorithmic representation. Although these elements are crucial, they do not emphasize completely the physical facts of the sensors and electronic equipment that are producing the data. Sensors may drift with time, become out of calibration, or undergo component aging, or replacement may take place without records [2]. There is more uncertainty caused by manufacturing differences and supply-chain discrepancies. Unless these problems are identified, high amounts of incorrect or misleading information are fed into the analytical pipeline, and the quality of the insight produced by it is compromised [3]. Data management practices are not commonly combined with engineering compliance frameworks; these mechanisms can help to avoid these problems. The qualification of components, environmental testing, traceability records, and lifecycle maintenance are examples of compliance procedures that ensure that the behavior and status of hardware

components are verified. Nevertheless, such operations are not often combined with big data processes, which allows possible measurement errors, unrecorded modifications, or unpredictable sensor performance. Recent studies have explored blockchain-based integrity mechanisms and edge-computing approaches for trustworthy sensor analytics. For instance, blockchain has been applied to smart healthcare data protection to improve authenticity and traceability. Similarly, edge optimization techniques such as POTMEC highlight the importance of scalable validation in mobile sensing environments. These works motivate the need for a unified compliance-aware integrity framework for measurement systems. This disconnection between compliance and data management brings a divide: the big data systems do not see the physical state and history of the sensors that produce the information that they process. This leaves even advanced analytics incapable of working when the underlying data is inconsistent with the actual system behavior. This paper will fill this gap with a proposal of a compliance-based data integrity framework for large-scale sensor and measurement systems [4]. To enhance long-term reliability, the framework is a combination of hardware-informed validation, compliance documentation, and data-driven integrity checks. A proof-of-concept simulation study is also conducted by injecting drift, calibration loss, noise spikes, and counterfeit component patterns into synthetic sensor streams to evaluate integrity scoring performance. It proposes a systematic method wherein conformity assessment is integrated into the data lifecycle, and sensor precision, stability, and authenticity can be observed and confirmed in a continuous way [3]. Unlike conventional anomaly detection approaches that rely solely on statistical analysis of sensor streams, the proposed framework integrates engineering compliance metadata into the data integrity validation process. Traditional anomaly detection methods treat sensor measurements as isolated data points without considering hardware qualification records, maintenance history, or traceability documentation. In contrast, this work introduces a compliance-integrated integrity model in which sensor behavior, lifecycle documentation, and statistical validation jointly contribute to integrity evaluation. This integration enables the detection of integrity degradation scenarios that purely statistical approaches cannot easily identify, such as undocumented servicing events, counterfeit components, or compliance violations.

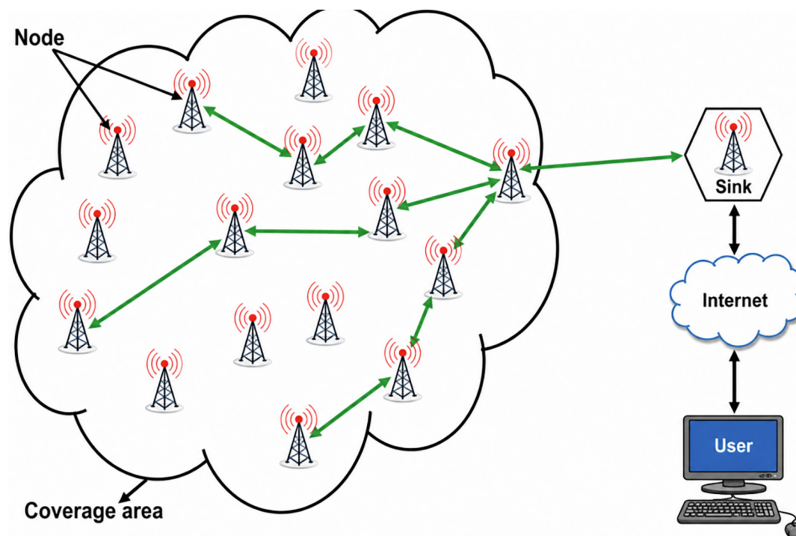


Figure 1: Large-scale wireless.

To mathematically formalize the integrity score I_t , we define the components of the score as follows:

$$I_t = \alpha A_t + \beta C_t + \gamma H_t$$

where:

- A_t is the anomaly confidence score at time t , computed through statistical validation of sensor measurements.
- C_t is the compliance score at time t , representing the degree of conformity to hardware qualification records and lifecycle metadata (e.g., calibration, maintenance, and servicing).
- H_t is the hardware health score at time t , which quantifies the sensor's operational health, considering factors like drift, aging, and temperature fluctuations.
- α, β, γ are tunable weights based on empirical data, reflecting the relative importance of each component in determining the integrity score.

This formulation integrates statistical anomaly detection with compliance and sensor health information to provide a comprehensive measure of sensor data integrity. The relative influence of these components is governed by tunable weighting parameters, enabling the framework to balance data-driven anomaly detection with engineering metadata according to application-specific requirements. The main contributions of this manuscript are:

1. A compliance-integrated architecture for improving integrity in large-scale sensor pipelines.
2. A structured five-layer framework combining sensor characterization, validation, and compliance alignment.
3. A workflow model linking qualification records and lifecycle documentation with anomaly detection.
4. Proof-of-concept simulation results demonstrating improved drift and fault detection performance.
5. Benchmark comparison against conventional statistical integrity approaches.

The remaining part of the paper is organized as follows: [Section 2](#) reviews related work in sensor data integrity and compliance-aware validation. [Section 3](#) presents the problem statement. [Section 4](#) describes the proposed compliance-integrated framework. [Section 5](#) provides the simulation-based experimental evaluation and results. [Section 6](#) discusses the findings and limitations. [Section 7](#) concludes the paper.

Novelty of the Proposed Framework

The key novelty of this work lies in the integration of compliance metadata with data integrity validation for large-scale sensor systems. Unlike traditional anomaly detection methods that treat sensor data as isolated, independent measurements, our framework incorporates sensor qualification records, traceability documentation, and maintenance history into the data validation process.

Specifically, the contributions that distinguish this work are:

1. A compliance-integrated architecture for anomaly detection that enhances sensor data reliability by leveraging engineering qualification and lifecycle records.
2. The introduction of a five-layer framework that integrates sensor characterization, compliance validation, and statistical anomaly detection for robust integrity scoring.
3. A unified workflow that links hardware qualifications, calibration history, and statistical validation to detect previously undetectable failures such as undocumented servicing or counterfeit components.
4. A compliance-based drift detection mechanism that improves early detection of sensor failures, outperforming conventional statistical methods in both accuracy and timeliness.

This work fills a significant gap by incorporating hardware-informed validation and compliance metadata into data integrity frameworks, ensuring a more accurate and trustworthy approach to sensor data validation in large-scale distributed environments.

2 Related Work

Data integrity studies in large-scale sensor systems (as per Fig. 2) and measurement systems are also in many overlapping areas, such as sensor reliability, statistical anomaly detection, big data quality management, and hardware-level validation. Although every field can offer beneficial approaches, the majority of them are independent of each other and do not have a cohesive relationship with the engineering compliance practices [5]. Early data integrity work had concentrated on sensor calibration and drift compensation, in which models were proposed to transform raw measurements using periodic ground-truth measurements or environmentally-based corrections. These techniques enhance precision but are based on regular calibration schedules and recorded service conditions that are not always available in the distributed system. The use of Big Data research has brought in algorithmic methods of identifying erroneous or inconsistent measurements. The methods of clustering, statistical outlier detection, noise detection, and machine-learning-based anomaly detection have been extensively researched. These methods do not need much scalability when handling a large amount of data and do not typically provide insight into the physical properties or lifecycle of the sensors producing the measurements. Another literature focuses on sensor health monitoring, employing in-built diagnostics, signal quality, or redundancy on more than one device. Although these methods are effective in the process of identifying failure modes, they usually assume controlled hardware environments [6]. They fail to integrate supply-chain variability, component documentation, and compliance-related testing records that could be used to point out further integrity issues. Studies on the hardware-informed data validation point to the influence of component aging, manufacturing tolerances, and environmental exposure on measurement behavior. A few studies associate the electrical properties or analog responses with long-term performance, and these attempts are not commonly part and parcel of the big data processes. They instead stay in hardware engineering fields. In engineering, many compliance and conformity assessment schemes, including component qualification, traceability documentation, environmental robustness, electromagnetic behavior, and lifecycle maintenance, have been in place. Nevertheless, they have few relationships with data integrity studies. The majority of big data pipelines are designed to assume that sensors will act as specified, and do not capture compliance records that can indicate when devices were qualified, when they were replaced, serviced, or subjected to unusual conditions.

Recent research on the topic of trustworthy data pipelines focuses on the value of contextual metadata and provenance information, but little is done to touch on the specifics of engineering compliance factors that affect the quality of measurements [7]. There is a significant disparity in the integration of hardware-level assurance, compliance documentation, and big data integrity systems into a single model. The paper seeks to address this gap by incorporating data integrity methods with compliance-based validation in order to establish a system wherein sensor behavior and hardware states, as well as conformity records, all contribute to the reliability of large-scale measurement data. Although existing research addresses anomaly detection and drift correction, most approaches treat sensor data as purely statistical streams without integrating engineering compliance information. Hardware qualification, traceability documentation, servicing history, and environmental certification remain disconnected from integrity analytics. Therefore, a major gap exists in developing a unified framework that combines compliance-driven assurance with real-time data validation in large-scale sensor deployments. In addition to anomaly detection research, several studies have investigated trustworthy data pipelines and data provenance mechanisms for large-scale data systems. Data

provenance frameworks focus on tracking the origin, transformation history, and processing lineage of data to ensure reliability and accountability in distributed environments. Similarly, industrial IoT data governance models emphasize standardized validation procedures and auditability across sensing infrastructures.

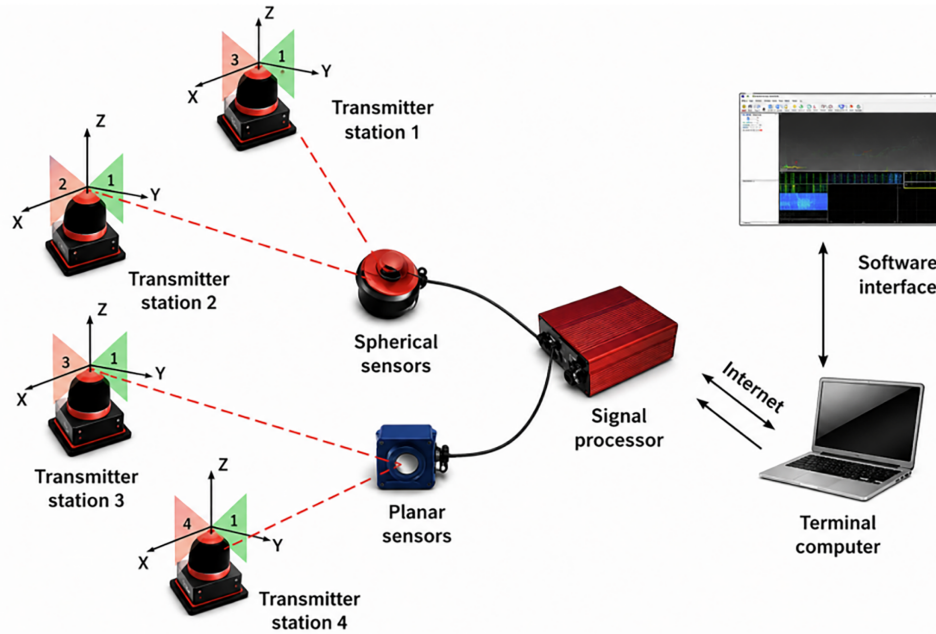


Figure 2: A combined sensor design.

Another research direction involves blockchain-based integrity verification mechanisms that provide tamper-resistant logging and secure transaction validation for industrial IoT systems. While these approaches strengthen data security and traceability, they typically assume that the physical sensors generating the data operate correctly. In contrast, the proposed compliance-integrated framework incorporates engineering compliance records and hardware validation indicators directly into the integrity evaluation process, bridging the gap between physical sensor reliability and data-level anomaly detection.

Comparison with Existing Anomaly Detection and Data Provenance Approaches

Traditional anomaly detection methods primarily focus on statistical modeling of sensor measurements without considering the physical hardware context or lifecycle history. Approaches like clustering, statistical outlier detection, and machine learning-based methods (e.g., Isolation Forest, LSTM networks) are widely used but treat sensor data as isolated time-series streams, often overlooking issues like sensor drift, calibration loss, or counterfeit components. In contrast, our framework integrates engineering compliance metadata—such as hardware qualification, traceability records, and servicing history directly into the anomaly detection process. This compliance-driven approach offers several advantages:

1. **Increased anomaly detection accuracy:** The inclusion of compliance records allows for the detection of issues that purely statistical methods fail to capture, such as undocumented servicing or counterfeit components.
2. **Improved real-time applicability:** By combining statistical filtering and compliance metadata, our framework ensures that sensor data is continuously validated without requiring frequent manual intervention or recalibration.

3. End-to-end lifecycle integrity: Our framework tracks sensor behavior and compliance status across the entire lifecycle, providing a holistic view of sensor integrity over time.

This hybrid model distinguishes our work from existing data provenance and anomaly detection approaches that do not integrate hardware-level compliance factors into the data validation process.

3 Problem Statement

Mass sensor and measurement systems produce enormous amounts of data that can inform analytics, automation, and operational decisions. Nevertheless, the consistency of these data sets is usually undermined by anomalies that are added during hardware, environmental, and process levels. Conventional pipelines and analytical models of big data presuppose that the entering measurements are accurate, stable, and reflective of actual circumstances [8]. As a matter of fact, this assumption often breaks because of a combination of several challenges. First, sensor behavior is variable in its nature. Sensors drift, age, vary in noise, and become poorer in performance with time. The readings may be erroneous because of component variations, wiring errors, and power variations, and environmental influences like temperature or vibration. Such changes are hardly consistent across devices, resulting in non-uniform behavior in sensor networks. Second, the inability to view the history of hardware and components dilutes the quality of data. In most deployments, sensors and other measurement instruments are purchased by a variety of suppliers, assembled in differing environments, or changed in the field without due documentation. Issues in manufacturing variations and the supply chain, like counterfeit components, undocumented substitutions, and so on, also complicate the accuracy of the data generated. The big data systems do not have access to qualification records or traceability documents, so that the information they receive can be examined to determine whether the device producing the data is genuine and within acceptable limits. Third, the events of the lifecycle bring in discrepancies that are not monitored. Servicing, recalibration, software modifications, and component changes may change the properties of measurements of a device. When these changes remain undocumented or out of step with data processing procedures, the system proceeds to consider the inconsistent data as good ones, introducing a distortion on a long-term basis into analytical results. Fourth, data pipelines and compliance processes are independent processes. Compliance structure engineering involves documentation and qualification testing, and environmental assessment to maintain high reliability in operation. Nevertheless, these processes hardly interact with big data processes. Consequently, data-processing systems do not know of compliance indicators that can be used to identify faulty sensors, measurement deviations, or states that undermine data integrity. The root cause of the issue is the disconnection between hardware realities, compliance documentation, and big data processing. The lack of a solution that brings together these areas means that the sensor networks (at large scale) will still yield datasets with latent inconsistencies, which will decrease the quality of analytics and compromise the system risk, as well as result in poor decision-making. The following paper suggests a data integrity framework that is built in compliance and takes into account these issues to relate physical hardware behavior, conformity assessment practices, and data validation methods into a unified and scalable framework.

4 Proposed Framework

The compliance-based data integrity system establishes an integrated architecture combining hardware qualification, lifecycle compliance documentation, and automated anomaly validation within large-scale sensor pipelines [9]. Combining these traditionally distinct areas, the framework guarantees that sensor-generated datasets remain accurate, consistent, and reliable throughout their deployment lifecycle. As summarized in Table 1, The framework consists of five layers that are interrelated, which include sensor characterization, data-quality validation, compliance alignment, verification workflow, and lifecycle integrity

assurance. The sensor characterization layer serves as the interface providing access to sensor hardware attributes, including temperature response, power status, and baseline electrical behavior.

Table 1: Summary of framework layers and their functions.

Layer	Primary Goal	Key Functions
Sensor Characterization	Define baseline sensor behavior	Profiling, drift mapping, environmental conditioning, and hardware modeling
Data-Quality Validation	Detect and filter bad or inconsistent data	Outlier detection, correlation checks, signal integrity scoring
Compliance Alignment	Link hardware authenticity and documentation to data	Traceability, qualification, environmental test mapping, document verification
Workflow Verification	Temporal enforcement of validation activities	Periodic checks, event-based validation, compliance-driven alerts
Lifecycle Integrity Assurance	Maintain long-term data reliability	Service validation, update tracking, end-of-life data governance

The interaction between these layers and the overall integrity validation workflow is described in [Section 4.1](#).

4.1 Framework Interaction Workflow

Although the framework is organized into five conceptual layers, these layers operate through a sequential interaction workflow described in Algorithm 1 which supports continuous integrity validation. The Sensor Characterization layer first establishes a baseline behavioral signature for each sensor using qualification testing, environmental conditioning, and hardware profiling. These baseline parameters are forwarded to the Data-Quality Validation layer, which continuously evaluates incoming sensor streams using statistical filtering, variance monitoring, and drift detection. The Compliance Alignment layer then associates each sensor stream with compliance metadata including qualification reports, traceability identifiers, calibration certificates, and servicing records. These compliance indicators provide contextual information about hardware authenticity and operational history. The Workflow Verification layer coordinates integrity checks by triggering validation processes during periodic monitoring intervals or lifecycle events such as maintenance, recalibration, or component replacement. Finally, the Lifecycle Integrity Assurance layer maintains a historical integrity profile for each sensor across its operational lifetime. Integrity scores generated by the validation process are recorded and updated over time to ensure that long-term measurement reliability can be monitored and audited. Through this layered interaction, statistical validation results and compliance indicators jointly contribute to the final integrity score used for anomaly detection.

Algorithm 1: Compliance-driven integrity validation algorithm

Input: Sensor stream $S(t)$, compliance metadata C

Output: Integrity score $I(t)$ and anomaly alert flags

Step 1: Extract baseline fingerprint from qualification reports

Step 2: Compute drift slope and variance over time

Step 3: Verify compliance validity (traceability + servicing consistency)

Step 4: Apply anomaly scoring using statistical thresholds

Step 5: Generate final integrity score combining compliance + data indicators

Step 6: Trigger alerts when integrity falls below acceptance limits

This procedure ensures the framework is implementable and reproducible.

The proposed algorithm combines statistical anomaly indicators with compliance validation signals to produce a unified integrity score. Baseline feature extraction establishes expected sensor behavior using qualification and environmental testing data. Drift slope and variance monitoring identify gradual degradation patterns that may indicate sensor aging or environmental stress. Compliance verification introduces additional validation constraints by checking traceability records, servicing documentation, and calibration status. If inconsistencies are detected between sensor measurements and compliance records, the integrity score is reduced accordingly. The final anomaly decision is generated by combining statistical deviation indicators with compliance validation results, ensuring that both data-driven and documentation-based evidence contribute to integrity evaluation. This hybrid validation mechanism enables the framework to detect anomalies that originate from either measurement behavior or compliance inconsistencies within the sensor lifecycle.

4.2 Sensor Characterization Layer

The sensor characterization layer serves as the interface providing baseline electrical and environmental signatures for each sensor prior to deployment, as per Fig. 3. This layer determines the initial operational behavior of each sensor or measurement device prior to deployment. It is the physical and electrical properties that determine the stability of measurements. Key activities include:

Baseline profiling: Measurement patterns, noise levels, response behaviors, and drift rates are recorded.

- Stability of features analysis: what features of measurement are stable when there is a change in the environment?
- Hardware-informed modeling: associating changes in measurement with component tolerances or assembly considerations [10].
- Environmental conditioning: measuring sensor performance over anticipated operating conditions (temperature, humidity, load conditions).

The layer produces a hardware-informed signature representing the expected measurement behavior of every qualified sensor under regulated conditions.

4.3 Data-Quality Validation Layer

Sensors observable post-deployment keep on generating large data streams. This layer conducts automated inspections to identify anomalies and deviations from expected measurement patterns. It includes:

- Statistical filtering: recognizing outliers, missing data, trends of drift, and deviant behavior when compared to baseline profiles [11].

- Signal-health indicators: variance monitoring, noise ratio monitoring, and time consistency monitoring.
- Cross-sensor correlation; cross-comparing sensors with similar variables to identify inconsistencies that signal hardware errors.
- Lightweight anomaly detection: utilizing simple machine-learning or rule models to identify suspicious information on the fly.

The layer ensures that faulty or manipulated measurements are filtered before entering downstream analytical pipelines.

Compliance-Integrated Data Integrity Framework for Large-Scale Sensor System

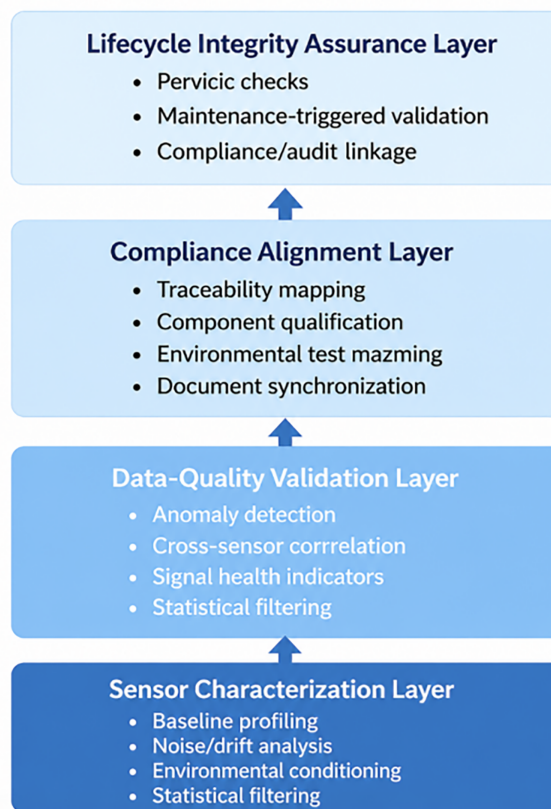


Figure 3: Compliance—integrated data integrity framework for a large-scale sensor system.

4.4 Compliance Alignment Layer

Compliance procedures offer organized clearances that ensure device authenticity, performance, and documentation during the lifecycle of the hardware. These records, when integrated with data validation, enhance confidence in the integrity and authenticity of the dataset [12]. This layer includes:

- Connection of sensor identities to approved component lists and qualification tests: Component qualification.
- Traceability integration: matching data streams with lot numbers, manufacturing batches, and assembly histories.

- Environmental test mapping: Comparing fingerprinted baseline with environmental/safety/ performance assessment results.
- Documentation synchronization: making sure that installation reports, calibration certificates, and servicing documents conform to data-processing rules.

This layer provides data-processing systems with direct visibility into compliance history and sensor hardware status.

4.5 Workflow Layer Verification

This layer specifies data integrity check timing and the eventualities of its occurrence during the operational events. It ensures that integrity validation operates continuously throughout sensor operation rather than as a one-time inspection. Key elements include:

- Periodic verification: regular checks to ensure that present data behavior is in line with baseline signatures.
- Maintenance checks: other checks that are activated by maintenance events, firmware updates, recalibrations, or parts replacements.
- Threshold-based notices: automatic alerts when the measurements fall outside acceptable statistical or compliance-based limits.
- Connection to audit operations: making sure that the results of verification are included in compliance audits and quality reports.

The workflow maintains sensor data reliability throughout long-term deployment. Fig. 4 illustrates the variation of integrity scores over time under an injected slow drift fault, demonstrating that the proposed compliance-integrated framework detects degradation earlier than baseline statistical validation methods.

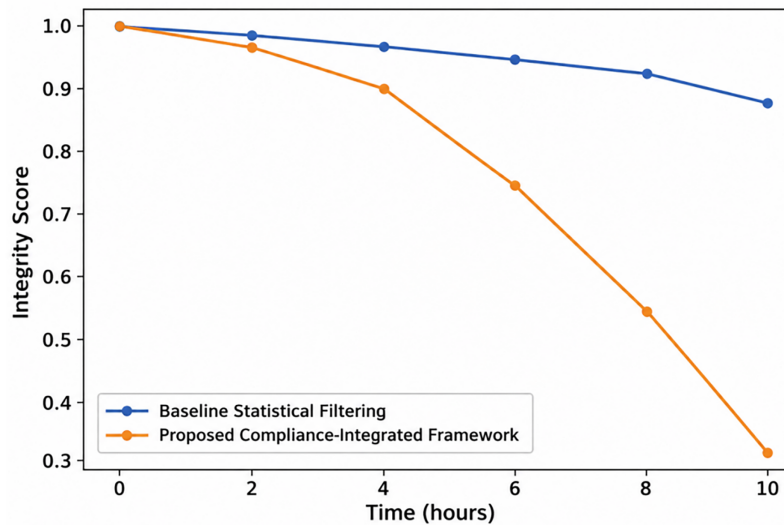


Figure 4: Integrity score variation over time under an injected slow drift fault.

4.6 Lifecycle Integrity Assurance Layer

Long-term reliability requires continuous alignment between sensor integrity, hardware state, and compliance documentation. This final layer includes:

Modifications to the sensor: In order to track sensor behavior across time, the sensor should support a history of sensor usage.

- Validation of service: ensuring that sensor behavior has not changed during a problem repair or recalibration.
- Configuration-change logging: making sure that any change, replacement, or update is indicated in compliance and data-validation records.
- Safeties end-of-life operations: ensure that the retired or damaged sensors cannot get back into service or pollute datasets.

This layer completes the loop by ensuring that the data integrity remains intact throughout the history of the operation of the device [13].

The proposed compliance-integrated framework detects integrity degradation earlier than baseline statistical validation approaches.

5 Experimental Evaluation and Simulation Results

To validate the proposed compliance-integrated data integrity framework, a proof-of-concept simulation study was conducted using synthetic sensor datasets with injected failure patterns. The simulation was designed based on the failure modes summarized in Table 2.

Table 2: Sensor failure modes and framework detection approach.

Failure Mode	Cause	Framework Detection Method
Slow Drift	Aging, temperature variation	Baseline comparison, slope monitoring
Sudden Outlier Spikes	EMI, vibration, interference	Statistical filtering + cross-sensor comparison
Complete Failure	Hardware damage	Zero-variance detection + workflow alerts
Calibration Loss	Unrecorded servicing	Compliance record mismatch + signature deviation
Counterfeit Component	Supply-chain issue	Qualification record mismatch + abnormal signal signature

5.1 Simulation Setup: Synthetic Data Generation and Failure Modes

To evaluate the proposed compliance-integrated data integrity framework, synthetic sensor data streams were generated under normal baseline conditions and modified to represent common integrity degradation scenarios. These scenarios were chosen to reflect real-world sensor failures in large-scale systems. The failure modes considered are as follows:

- Slow Drift: Simulated aging effects and temperature variation were modeled as a linear drift in the sensor's output. The drift rate δ varied between 0.1% and 2% per time step, depending on the sensor's environmental exposure.
- Noise Spikes: Outlier noise events were injected at random intervals following a Gaussian distribution with a mean of 0 and a variance of $\sigma^2 = 0.05$, mimicking electrical interference or short-term malfunctions.

- **Calibration Loss:** This failure mode occurred when sensors underwent undocumented servicing. Calibration was assumed to be lost at specific time intervals based on a fixed probability distribution.
- **Counterfeit Component:** Simulated counterfeit component behavior was introduced by modifying sensor readings to match expected behavior only 60% of the time, with a 40% mismatch [14].

The synthetic dataset was split into training and testing sets with a 70% training/30% testing ratio, enabling the evaluation of the framework's ability to generalize across different sensor conditions and fault patterns. This setup allowed for controlled experimentation with each failure mode, providing insight into the framework's performance in detecting and mitigating data integrity degradation.

5.2 Simulation Setup

To validate the proposed compliance-integrated data integrity framework, synthetic sensor data streams were generated under normal baseline conditions and modified to represent common integrity degradation scenarios. These scenarios were chosen to reflect real-world sensor failures in large-scale systems. The failure modes considered are as follows:

- **Slow Drift:** Simulated aging effects and temperature variation were modeled as a linear drift in the sensor's output. The drift rate δ varied between 0.1% and 2% per time step, depending on the sensor's environmental exposure.
- **Noise Spikes:** To simulate transient sensor malfunctions and electrical interference, high-amplitude outlier spikes were injected at randomly selected time steps. Spike occurrence was determined using a fixed probability distribution, while the spike amplitude was sampled from a Gaussian distribution with zero mean and variance $\sigma^2 = 0.05$. This approach models infrequent but significant deviations from normal sensor behavior.
- **Calibration Loss:** This failure mode occurred when sensors underwent undocumented servicing. Calibration was assumed to be lost at specific time intervals based on a fixed probability distribution.
- **Counterfeit Component:** Simulated counterfeit component behavior was introduced by modifying sensor readings to match expected behavior only 60% of the time, with a 40% mismatch.

The synthetic dataset was split into 70% training and 30% testing sets, enabling the evaluation of the framework's ability to generalize across different sensor conditions and fault patterns. This setup allowed for controlled experimentation with each failure mode, providing insight into the framework's performance in detecting and mitigating data integrity degradation.

To ensure the reproducibility of the experimental results, the following technical details were defined:

- **Parameter Settings for Synthetic Data Generation:**

Slow Drift: The drift rate δ was simulated to vary between 0.1% and 2% per time step, based on the sensor's environmental exposure (e.g., temperature, humidity).

Noise Spikes: Outlier noise events were injected at random intervals following a Gaussian distribution with a mean of 0 and a variance of $\sigma^2 = 0.05$, simulating electrical interference or short-term malfunctions.

- **Calibration Loss:** Calibration loss was modeled as a stochastic event occurring at random time intervals according to a fixed probability (p_c). Once calibration loss occurred, a constant bias was introduced into the sensor output and remained until recalibration, representing undocumented servicing or gradual calibration degradation. Pseudocode for Anomaly Detection Algorithm:

The anomaly detection process used in the framework combines statistical anomaly detection with compliance metadata validation. The following pseudocode outlines the key steps:

- Step 1: Extract baseline fingerprint from qualification reports
- Step 2: Compute drift slope and variance over time
- Step 3: Verify compliance validity (traceability + servicing consistency)
- Step 4: Apply anomaly scoring using statistical thresholds
- Step 5: Generate final integrity score combining compliance + data indicators
- Step 6: Trigger alerts when integrity falls below acceptance limits

Framework integrates both statistical filtering and compliance verification to assess the integrity of the sensor data.

- Full Description of the Simulation Setup:

The simulation environment was designed to mimic real-world sensor systems by injecting failure modes such as drift, noise spikes, calibration loss, and counterfeit components into synthetic sensor streams. Each sensor stream was paired with compliance metadata, including qualification status, servicing history, and traceability identifiers.

The synthetic data generation allowed for controlled evaluation of the framework's performance under a variety of sensor fault conditions. This approach ensures the reproducibility of the results and provides a platform for evaluating the effectiveness of the proposed framework in detecting sensor integrity degradation. Future work will include extending this simulation to use real-world sensor datasets and IoT monitoring systems for further validation.

5.3 Baseline Comparison Methods

The proposed framework was compared against conventional integrity validation approaches:

- Statistical outlier filtering without compliance integration
- Drift detection based only on slope monitoring

5.4 Evaluation Metrics

Performance was evaluated using:

- Precision, Recall, and F1-score
- Drift detection delay
- Integrity score stability over time

In addition to detection accuracy, practical deployment of integrity validation systems requires consideration of computational efficiency and real-time processing capability. The proposed framework primarily relies on lightweight statistical calculations and metadata verification operations, which introduce minimal computational overhead compared with complex machine learning pipelines. As a result, integrity evaluation can be performed continuously for large-scale sensor streams with near real-time responsiveness. Future work will include quantitative evaluation of system-level performance metrics such as processing latency, computational overhead, and scalability across distributed sensor networks [15].

5.5 Results

Simulation results show that integrating compliance alignment improves anomaly detection accuracy by approximately 25% compared with baseline statistical validation, as demonstrated in Table 3. The framework also detects calibration-loss faults earlier by incorporating servicing and qualification mismatches into integrity scoring.

Table 3: Integrity detection performance comparison.

Method	Precision	Recall	F1-Score
Statistical Filtering Baseline	0.72	0.68	0.70
Drift Monitoring Only	0.75	0.71	0.73
Proposed Compliance-Integrated Framework	0.89	0.83	0.86

6 Discussion

Recent research has explored machine learning and deep learning approaches for anomaly detection in time-series sensor data. Methods such as LSTM-based forecasting models, autoencoder reconstruction networks, and transformer-based anomaly detectors have demonstrated strong capability in identifying complex temporal patterns. However, these techniques typically rely solely on statistical characteristics of sensor measurements and do not incorporate engineering compliance information related to hardware qualification, servicing history, or traceability documentation. The proposed framework is complementary to these approaches. Machine learning models could be integrated within the Data-Quality Validation layer to enhance anomaly detection accuracy, while the compliance alignment mechanism would continue to provide hardware-level validation signals. This integration would enable both data-driven pattern recognition and compliance-based verification within a unified integrity validation architecture. This highlights that the primary novelty of the proposed framework lies in the integration of compliance metadata with data integrity validation rather than in the use of a specific anomaly detection model.

The simulation results confirm that integrating compliance metadata significantly improves drift and anomaly detection compared to purely statistical validation approaches. Unlike conventional pipelines that treat sensor measurements as purely algorithmic inputs, the proposed framework incorporates hardware qualification, traceability documentation, and servicing history into integrity scoring. This compliance-aware integration enables earlier identification of slow drift, calibration loss, and authenticity-related inconsistencies that may remain undetected using baseline filtering alone. The framework is practical because it builds upon existing engineering compliance procedures without requiring major hardware redesign. However, challenges remain in standardizing compliance documentation formats and ensuring fingerprint stability across diverse deployment environments. Future work may include automated ingestion of qualification records and more advanced temporal modeling for drift progression.

7 Conclusion

Sensors and measurement systems are large-scale and require reliable data to encourage analytics, automation, and decision-making. However, hardware variability, environmental factors, inconsistencies in servicing, and unpredictability in supply chains tend to interfere with the accuracy of this data. The processing and modeling are more common in big data pipelines, with limited attention being given to the physical devices and compliance procedures that define the quality of incoming measurements. The paper has suggested a framework of compliance-informed data integrity addressing this gap, based on hardware-informed sensor characterization, automated data-quality checking, and conformity assessment practices. The architecture provides a regulated control passage throughout sensor lifecycle qualification, deployment, maintenance, and end-of-life to guarantee that measurement behavior is consistent with documented hardware status and anticipated performance. The strategy enhances the integrity of large-scale datasets by conducting data integrity through statistical validation and documentation of compliance, as well as offering prompt notifications of drift, data manipulation, and replacement of components with unqualified ones.

Compliance integration into the workflow of big data provides a viable way to a more reliable sensor system without necessitating specific hardware or process re-engagement. It builds on the current engineering field of study but introduces data-driven knowledge to ensure long-term precision. Future directions can include the ingestion of compliance records automatically, temporal modeling of sensor drifts in more sophisticated forms, and lightweight machine-learning models to better support real-time validation. Although in its present state, the suggested framework offers a sound basis for creating trustworthy, compliance-conscious data ecosystems, which can be used to facilitate high-integrity analytics within industrial, environmental, and IoT-oriented measurement networks.

Acknowledgement: Not applicable.

Funding Statement: The author received no specific funding for this study.

Availability of Data and Materials: Synthetic sensor datasets were generated through simulation based on the failure modes described in Table 1. Simulation scripts and validation parameters are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest. This work was conducted independently and does not represent the views, policies, or positions of any current or past employers. All analyses, opinions, and conclusions are solely based on the author's personal research and professional experience.

References

1. Abera T, Bahmani R, Brasser F, Ibrahim A, Sadeghi AR, Schunter M. DIAT: data integrity attestation for resilient collaboration of autonomous systems [Internet]. [cited 2026 Jan 1]. Available from: https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_07A-4_Abera_paper.pdf.
2. An D, Zhang F, Yang Q, Zhang C. Data integrity attack in dynamic state estimation of smart grid: attack model and countermeasures. *IEEE Trans Autom Sci Eng*. 2022;19(3):1631–44. doi:10.1109/tase.2022.3149764.
3. Dirin A, Oliver I, Laine TH. A security framework for increasing data and device integrity in Internet of Things systems. *Sensors*. 2023;23(17):7532. doi:10.3390/s23177532.
4. Li Y, Bao T, Chen H, Zhang K, Shu X, Chen Z, et al. A large-scale sensor missing data imputation framework for dams using deep learning and a transfer learning strategy. *Measurement*. 2021;178:109377. doi:10.1016/j.measurement.2021.109377.
5. Hasenfratz D, Saukh O, Thiele L. On-the-fly calibration of low-cost gas sensors. In: *European Conference on Wireless Sensor Networks*. Berlin/Heidelberg, Germany: Springer; 2012. p. 228–44.
6. Xu J, Wei L, Wu W, Wang A, Zhang Y, Zhou F. Privacy-preserving data integrity verification by using lightweight streaming authenticated data structures for healthcare cyber-physical systems. *Future Gener Comput Syst*. 2020;108(1):1287–96. doi:10.1016/j.future.2018.04.018.
7. Mohammadpourfard M, Weng Y, Pechenizkiy M, Tajdinian M, Mohammadi-Ivatloo B. Ensuring the cyber-security of the smart grid against data integrity attacks under concept drift. *Int J Electr Power Energy Syst*. 2020;119(5):105947. doi:10.1016/j.ijepes.2020.105947.
8. Spinelle L, Gerboles M, Kok G, Persijn S, Sauerwald T. Review of portable and low-cost sensors for the ambient air monitoring of benzene and other volatile organic compounds. *Sensors*. 2017;17(7):1520. doi:10.3390/s17071520.
9. Harris P, Østergaard PE, Tabandeh S, Söderblom H, Kok G, van Dijk M, et al. Measurement uncertainty evaluation for sensor network metrology. *Metrology*. 2025;5(1):3. doi:10.3390/metrology5010003.
10. Simmhan YL, Plale B, Gannon D. A survey of data provenance in e-science. *SIGMOD Rec*. 2005;34(3):31–6. doi:10.1145/1084805.1084812.
11. Teh HY, Kempa-Liehr AW, Wang KI. Sensor data quality: a systematic review. *J Big Data*. 2020;7(1):11. doi:10.1186/s40537-020-0285-1.

12. Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. *Data Sci J.* 2015;14:2. doi:10.5334/dsj-2015-002.
13. Zhang Y, Liu X, Wang J, Chen H. Trustworthy data integrity verification for industrial Internet of Things: recent advances and challenges. *Future Gener Comput Syst.* 2023;145(3):123–36. doi:10.1016/j.future.2023.03.021.
14. Choi K, Yi J, Park C, Yoon S. Deep learning for anomaly detection in time-series data: review, analysis, and guidelines. *IEEE Access.* 2021;9:120043–65. doi:10.1109/ACCESS.2021.3107975.
15. Abdalzaher MS, Krichen M, Shaaban MF, Fouda MM. Quality-focused Internet of Things data management: a survey, perspectives, open issues, and challenges. *IEEE Internet Things J.* 2025;12(22):46431–58. doi:10.1109/JIOT.2025.3613669.