Journal of Intelligent Manufacturing
https://doi.org/10.1007/s10845-025-02745-8

Dta-qc: an AI-driven framework for adaptive quality control and intelligent test optimization in 5 G manufacturing

Jie Liu[1] · Enislay Ramentol[2] · Cristina Landin[3] · Sahar Tahvili[4,5]

Received: 4 July 2025 / Accepted: 17 November 2025

Abstract

In modern 5 G radio manufacturing, traditional quality control methods based on black fixed thresholds are increasingly inadequate, often failing to capture nuanced fault patterns and requiring substantial manual intervention. This study presents DTA-QC, an AI-driven framework for adaptive thresholding and intelligent test optimization in 5 G production environments. The proposed system introduces three core innovations: (1) dynamic thresholding using LSTM autoencoders and regression models to detect anomalies under evolving production conditions, (2) supervised fault classification via convolutional neural networks trained on time-windowed sensor data, and (3) a four-level severity classification system (Normal, Warning, Worse, Stop) to support real-time decision-making in manufacturing environments. DTA-QC is implemented and validated on Ericsson AB's 5 G radio production line, achieving high anomaly detection accuracy (ROC-AUC: 0.89–0.94) and significantly reducing manual review efforts, without requiring specialized hardware. To assess generalizability, DTA-QC is further evaluated on a public benchmark dataset. A comparative analysis of three architectural variants revealed trade-offs in complexity, latency, and deployment feasibility. These results underscore the value of embedding AI-driven analytics in industrial test workflows, contributing to the broader goals of intelligent manufacturing and adaptive, data-driven quality assurance.

Keywords Artificial intelligence · Intelligent manufacturing · Adaptive quality control · Anomaly detection · Time-series classification · Industry 4.0

✉ Jie Liu
anna.a.liu@ericsson.com

✉ Sahar Tahvili
sahar.tahvili@mdu.se

Enislay Ramentol
enislay.ramentol.martinez@itwm.fraunhofer.de

Cristina Landin
cristina.landin@saabgroup.com

[1] Ericsson AB, Stockholm, Sweden

[2] Department of Financial Mathematics, Fraunhofer Institute for Industrial Mathematics, Kaiserslautern, Germany

[3] Defense and Space Manufacturing, Saab AB, Stockholm, Sweden

[4] Innovation and Product Realization, Mälardalens University, Eskilstuna, Sweden

[5] Einride, Autonomous Technologies AB, Stockholm, Sweden

Introduction and related work

Production testing is a critical function in modern manufacturing, particularly for complex, hardware-intensive systems such as 5 G radio base stations (RBSs). These systems require precise validation of integrated hardware-software interactions, typically involving automated measurements of Radio Frequency (RF) signal quality, thermal characteristics, and power efficiency. In industrial settings like telecom production lines, this process generates large volumes of time-series data used for pass/fail decisions and root cause analysis, similar to other complex manufacturing processes where ML-based monitoring has shown promise (Liu et al., 2022). Conventional testing approaches predominantly depend on fixed thresholds derived from Radio Frequency (RF) conformance standards (e.g., 3GPP TS 38.141-1[1]) to determine product compliance. However, such conventional

[1] 3GPP TS 38.141-1: Technical Specification by 3GPP defining RF (Radio Frequency) requirements and test methods for NR (New Radio) base stations.

Published online: 26 November 2025

rule-based approaches are increasingly inadequate in dynamic production environments characterized by evolving hardware tolerances, firmware updates, and operational variability (Kaner et al., 2002). Fixed thresholds often fail to account for contextual nuances, leading to false positives, operational inefficiencies, and missed opportunities for quality improvement (Tahvili & Hatvani, 2022). The growing availability of production data and advances in artificial intelligence (AI) have opened new avenues to rethink industrial quality control aligning with the broader goals of Quality 4.0 (Papavasileiou et al., 2025; Karkaria et al., 2025; Escobar et al., 2021). AI and machine learning (ML) models can detect subtle anomalies, adapt to changing conditions, and automate labor-intensive analysis tasks (Tahvili & Hatvani, 2022; Landin et al., 2023; Deshpande et al., 2023). Among these, dynamic thresholding, adaptive limits that evolve based on statistical features of the data, has shown particular promise in manufacturing contexts (Barr et al., 2015; Deshpande et al., 2023; Orabi et al., 2024). Despite its benefits, this technique remains underutilized in industrial testing pipelines, where fixed threshold logic still dominates.

To address these limitations, this study introduces DTA-QC (Dynamic Thresholding and Anomaly-aware Quality Control), an AI-driven framework for adaptive thresholding and intelligent test optimization in 5 G production environments. DTA-QC integrates three core capabilities: (1) dynamic thresholding based on LSTM autoencoders and regression models to adapt to evolving operational contexts, (2) supervised anomaly classification using convolutional neural networks applied to time-windowed sensor features, and (3) a four-tier severity grading system (Normal, Warning, Worse, Stop) that enables actionable, real-time quality decisions. DTA-QC is validated within Ericsson AB's 5 G radio production environment, where it demonstrated notable improvements in anomaly detection accuracy (ROC-AUC: 0.89-0.94), test throughput, and reduction of manual review efforts by identifying early-stage anomalies that are typically investigated only after a failure has occurred. To assess robustness and cross-domain applicability, it was further evaluated on a publicly available benchmark dataset. Comparative analysis of three architectural variants highlighted trade-offs in prediction performance, resource efficiency, and deployment feasibility, providing valuable insight into the design considerations for scalable AI solutions in industrial testing systems.

This study offers the following key contributions:

1. Proposes and validates DTA-QC, an AI-driven framework for adaptive thresholding and intelligent test optimization in 5 G production environments, addressing core limitations of fixed-threshold quality control.
2. Demonstrates that **dynamic thresholding** using LSTM autoencoders and regression significantly improves time-series anomaly detection and reduces dependency on manual tuning in imbalanced industrial data contexts (Chung et al., 2023).
3. Enables **real-time severity classification** using CNNs, supporting actionable operator decisions and increasing testing throughput via a four-level grading system.
4. Introduces an **AI-assisted labeling pipeline** that reduces annotation cost and scales fault data labeling in low-label industrial settings.
5. Validates DTA-QC on both **proprietary and benchmark datasets**, confirming its generalizability, computational efficiency, and industrial readiness.

By addressing the limitations of fixed-thresholding and embracing data-driven adaptation, DTA-QC supports the transition toward intelligent, resilient, and efficient production systems. This aligns with the broader objectives of Industry 4.0 and intelligent manufacturing, where AI-driven decision support, real-time diagnostics, and scalable automation are foundational pillars (Escobar et al., 2021). By bridging the gap between academic AI research and industrial deployment in high-throughput, hardware-software integrated environments, DTA-QC advances the realization of adaptive, smart factory ecosystems. While the proposed DTA-QC framework was developed and validated in the context of 5 G radio testing, it does not depend on frequency-specific characteristics such as mmWave propagation or modulation schemes. Rather, its design responds to the increased operational complexity of 5 G systems, especially the rise of massive MIMO radios with 32 or more ports. These configurations challenge traditional test methods, which are often sequential and scale poorly in cost and time. In contrast, earlier-generation systems like 4 G typically involved only 1–4 ports and simpler test routines. Although motivated by 5 G-specific constraints, DTA-QC remains a domain-agnostic, general-purpose framework applicable to a wide range of manufacturing domains characterized by time-series data, class imbalance, and evolving production variability.

The rest of this paper is structured as follows: Sect. ”Introduction and related work” discusses relevant background and related work. Section ”The DTA-QC framework: dynamic thresholding and anomaly-aware quality control” presents the DTA-QC framework. Section ”Implementation” details the implementation. Section ”Empirical evaluation” reports the empirical evaluation, and Sect. ”Cross-domain validation with benchmark dataset” offers a cross-domain validation. Threats to validity and limitations are discussed in Sect. ”Threats to validity”, followed by future work in

Sect. ”Discussion and future work”, and conclusions in Sect. ”Conclusions”.

Research questions and objectives

As discussed earlier, this study addresses the limitations of conventional test strategies in 5 G radio production by introducing DTA-QC, an AI-driven framework for adaptive thresholding and intelligent test optimization in industrial environments. In complex systems such as 5 G radios, where tightly coupled hardware and software components generate highly dynamic behavior, traditional approaches based on fixed-thresholding often lack the contextual adaptability required for reliable quality assurance. To guide this investigation and ensure a structured, evaluative methodology, we define the following research questions:

- **RQ1: How can AI techniques be effectively applied to implement dynamic thresholds in 5 G radio production testing to complement and enhance fixed-threshold methods?** This question examines the technical feasibility and implementation strategies for transitioning from rigid pass/fail criteria to adaptive, data-driven thresholds in industrial testing workflows. Although the potential of dynamic thresholds is well acknowledged, their deployment remains rare in practice.
- **RQ2: To what extent does the DTA-QC improve anomaly detection accuracy and operational efficiency compared to conventional testing workflows?** This question evaluates model performance, fault detection precision, and the reduction in manual review effort under real production conditions.
- **RQ3: What are the operational benefits of implementing a four-tier severity classification system in terms of testing efficiency and product quality?** Here, we explore how the proposed severity levels (Normal, Warning, Worse, Stop) contribute to actionable insights, better test prioritization, and improved quality management.
- **RQ4: How scalable and generalizable is the DTA-QC framework across different production conditions and hardware configurations?** This question assesses the portability and applicability of the framework beyond the initial case study, with emphasis on deployment feasibility across diverse manufacturing contexts, considering the constraints inherent in AI-driven production systems (Wang et al., 2021).

The primary objective of this research is to design, implement, and validate the DTA-QC framework, which integrates dynamic thresholding, automated classification, and real-time decision support. By addressing these research questions, the study aims to demonstrate both the technical soundness and industrial relevance of the approach, contributing to the advancement of intelligent, data-driven quality assurance in manufacturing systems.

Novel contributions: This paper introduces a novel integration of adaptive thresholding, semi-automatic severity labeling, and interpretable classification within a unified AI-driven pipeline validated on real 5 G production data. Unlike prior methods that use fixed thresholds or unsupervised anomaly detection, DTA-QC employs supervised and semi-supervised learning to create production-aware test logic, reducing manual rule tuning and enabling explainable quality control decisions in real time.

Production testing in software-defined 5 G systems

This study is situated within the context of software product lines (SPLs), with a specific focus on production testing processes in 5 G radio systems. The goal is to validate the behavior of tightly integrated hardware and software components under actual manufacturing conditions. These components are tested using automated, software-defined procedures that generate high-resolution time-series data for performance verification, anomaly detection, and quality control. In this setting, *production testing* refers to system-level validation where software configurations directly influence hardware behavior. These systems are examples of *software-defined 5 G architectures*, in which critical performance parameters such as signal fidelity, thermal response, frequency stability, and power efficiency are governed by both physical components and embedded software, including gain tables, calibration algorithms, and digital signal processing (DSP) firmware.

Traditional quality control methods in such settings typically rely on fixed rule-based thresholds, which define fixed Pass/Fail boundaries based on standard conformance documents such as 3GPP TS 38.141-1. These approaches require extensive historical data to calibrate limits and must be manually updated to accommodate firmware revisions or hardware variability. Furthermore, non-AI methods are generally unable to scale effectively with high-dimensional time-series data or capture subtle anomalies that evolve over time. In contrast, the proposed DTA-QC framework enhances these traditional strategies by introducing dynamic thresholding, anomaly-aware classification, and severity grading mechanisms that adapt in real-time to the production context. This evolution not only reduces false positives and manual tuning but also enables timely decision-making under complex manufacturing conditions.

The term “production testing” is widely used in industrial electronics and telecommunications to describe quality validation activities conducted during or after the manufacturing process (Milor & Sangiovanni-Vincentelli, 2002; Agrawal et al., 2003). This perspective aligns with current trends in intelligent manufacturing, where artificial intelligence and machine learning are increasingly employed to enhance testing efficiency, improve fault detection, and support predictive maintenance. This industrial context highlights the relevance of DTA-QC, which addresses the variability introduced by the interaction between evolving hardware configurations and software parameters. By operating within a software-defined, hardware-aware production environment, the framework advances quality assurance capabilities and supports the broader objectives of SPLs in telecom manufacturing systems.

The integration of AI and ML into production testing is transforming quality assurance, particularly in domains where hardware and embedded software interact tightly. This transformation is particularly critical in high-mix, low-volume manufacturing environments where traditional batch-based quality control methods prove insufficient. This shift is evident in tasks such as test automation, adaptive sequencing, anomaly detection, and time-series classification (Felderer et al., 2023; Tahvili & Hatvani, 2022; Tahvili, 2018). For complex systems like 5 G Radio Base Stations (RBSs), traditional rule-based testing methods relying on fixed thresholds are increasingly inadequate. 5 G RBSs are software-defined hardware systems that integrate RF front-ends, baseband processors, power modules, and embedded Digital Signal Processing (DSP) firmware. The DSP units handle critical real-time tasks such as signal filtering, modulation, and protocol-specific encoding, playing a central role in the radio's functional integrity. Production testing verifies these components through automated procedures, known as test cases, which generate multivariate time-series data for pass/fail evaluation. Conventional quality control relies on expert-defined fixed thresholds, typically derived from standards such as 3GPP TS 38.141-1. While consistent, this approach lacks the adaptability required for modern production environments, where hardware tolerances, firmware revisions, and ambient conditions shift frequently. To address these challenges, data-driven methods have been introduced. AI/ML models trained on historical test data enable predictive analytics, contextual evaluation, and early anomaly detection (Tahvili & Hatvani, 2022). Dynamic thresholding, in particular, has emerged as a promising technique, adapting decision boundaries to evolving test conditions. Despite this potential, the practical deployment of dynamic thresholding techniques remains limited due to challenges in interpretability, data labeling, and model adaptation under concept drift. To address these limitations, this study introduces DTA-QC, a supervised AI framework designed for 5 G production testing. DTA-QC integrates adaptive thresholding, time-aware anomaly classification, and decision support to improve yield, reduce waste, and support proactive quality assurance.

Test optimization in 5 G RBS production

Modern telecommunications manufacturing requires high-throughput, low-latency test operations. In the context of Ericsson AB's 5 G RBS production line, testing is implemented via software-defined test stations that execute structured sequences of calibration and performance test cases on each hardware unit. These test cases follow a predefined test plan, grouped by category and executed sequentially. A simplified schematic of this process is shown in Fig. 1, where each test case is evaluated against its upper and lower limits (UL/LL). If any test case fails, the logic triggers an early stop, and the remaining test sequence is aborted for that specific RBS unit.

Each test case is evaluated against upper and lower fixed thresholds (UL, LL). Any test failing to meet these limits triggers rejection. While straightforward, this logic ignores signal trends, lacks predictive insight, and is inefficient when failures occur late in the sequence.

Challenges in sequential production testing

Figure 2 illustrates the typical sequential execution of test cases (TC_1 to TC_k) during production testing. In this scenario, a fault is detected at TC_5, meaning that the results of

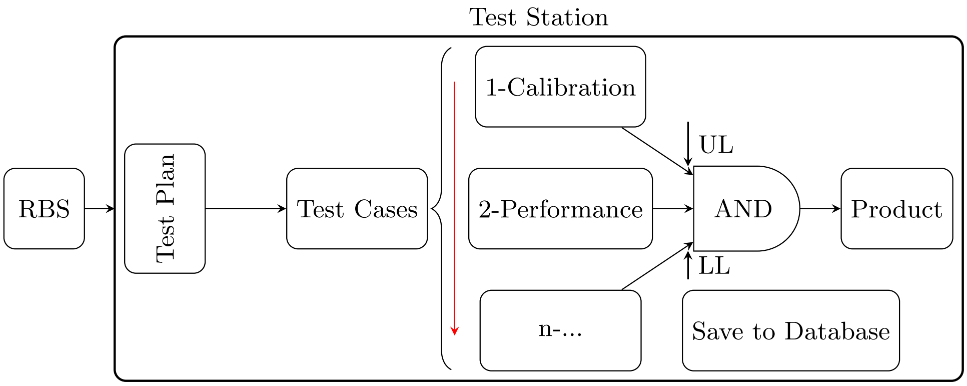

Fig. 1 Traditional fixed-threshold-based production testing workflow for 5 G RBS

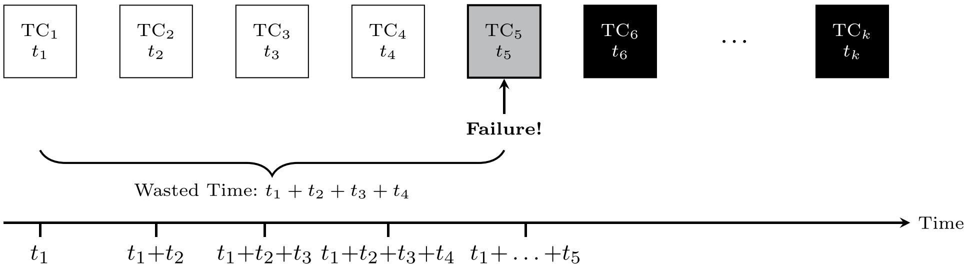

Fig. 2 Sequential test execution with failure at TC_5. The shaded region highlights the cumulative test time wasted before the failure is detected

the preceding test cases (TC_1 through TC_4) become irrelevant or unusable for final quality decisions.

As shown in Fig. 2, this leads to wasted station time and test resources. The cumulative wasted time before failure detection can be quantified as:

$$t_T = \sum_{i=1}^{k-1} t_i \tag{1}$$

where t_i is the execution time for each test case, this inefficiency becomes critical in high-throughput manufacturing such as 5 G radio production, where late fault detection incurs operational delays and resource underutilization. To avoid such inefficiencies, DTA-QC incorporates predictive intelligence into test workflows. By leveraging AI models trained on historical time-series signals, the system predicts failure likelihoods early in the sequence. This enables:

- **Early failure prediction:** detects weak signals or faults before full sequence completion.
- **Model adaptation:** continuously updates based on firmware revisions, hardware drifts, or test conditions.
- **Dynamic prioritization:** ranks test cases by risk and diagnostic contribution.

Together, these capabilities transform testing from a rigid, reactive process into a flexible, proactive quality control loop. This foundation enables DTA-QC to deliver measurable improvements in test efficiency, failure detection, and operational decision-making in 5 G telecom production.

Related work on AI-driven quality control

Research on test optimization in manufacturing has increasingly focused on predictive models that enhance yield, reduce rework, and improve decision efficiency. Early research efforts by Weiss et al. (2013, 2016), demonstrated how regression models could forecast microprocessor test outcomes. Landin et al. (2021) utilized support vector machines (SVMs) for yield prediction in the telecom industry, highlighting the benefits of integrating real-time analytics into production workflows.

However, these approaches often lack robustness in environments with tightly coupled hardware-software interactions, such as 5 G radio production. Recent advances in automated model selection for multivariate anomaly detection address some of these challenges (Wen et al., 2023). These approaches typically overlook temporal dependencies and assume data stationarity, and require frequent recalibration, limiting applicability in dynamic, high-mix industrial contexts (Blázquez-García et al., 2021).

To address temporal complexity, deep learning has been explored for time-series anomaly detection. Kashiparekh et al. (2019) proposed a CNN-based transfer learning model for univariate series, and Wen and Keyes (2019) addressed class imbalance using feature-aware architectures. Choi et al. (2021) provides a taxonomy of time-series anomaly detection using deep neural networks. These works, while advancing model design, often assume stationary behavior and are rarely validated in online, resource-constrained industrial testing. Threshold modeling remains a core bottleneck. Traditional rule-based systems offer deterministic pass/fail logic but lack adaptability. Statistical methods such as Extreme Value Theory (EVT) (Siffer et al., 2017; Su et al., 2019) introduce rigor, yet suffer from outlier sensitivity (Scarrott & MacDonald, 2012). Signal segmentation (Dani et al., 2015) and error prediction (Hundman et al., 2018) methods assume Gaussianity or require labeled fault data, assumptions often unmet in telecom testing.

Although adaptive thresholding techniques have been explored recently, their adoption in real-world production environments remains limited. For example, Tonini et al. (2024) propose SAnD (Simple Anomaly Detection), a semi-supervised method integrating statistical filtering and threshold selection, yet its evaluation remains offline and not embedded in live test systems. Likewise, a comprehensive survey by Yan et al. (2023) on transfer learning for industrial time-series anomaly detection highlights that most approaches rely on pretrained models rather than dynamically updating thresholds. These findings underscore that, despite their promise, adaptive thresholding mechanisms are still not commonplace in operational pipelines, further motivating the design of DTA-QC for practical deployment.

In earlier work (Landin et al., 2023), we introduced a semi-automated threshold adaptation method using CNNs and regression models for 5 G test data. However, it required manual parameter tuning and lacked mechanisms for online learning and adaptive severity scoring. The DTA-QC framework extends this by introducing fully automated dynamic thresholding, supervised classification, and multi-level fault severity grading, features critical for scalable deployment in intelligent manufacturing.

Moreover, recent research in intelligent manufacturing emphasizes AI-driven anomaly detection and predictive maintenance as core enablers of Industry 4.0 (Liso et al., 2024; Stojanovic et al., 2016; Kumari et al., 2024). Our work extends these developments by introducing a dynamic, label-efficient quality control framework validated under real industrial conditions.

Recent survey work by Liso et al. (2024) offers a comprehensive analysis of deep learning-based anomaly detection strategies tailored to Industry 4.0 applications. The authors categorize existing approaches based on sensing equipment (e.g., vision, vibration, thermal), algorithmic models (e.g., CNNs, LSTMs, autoencoders), and application fields such as predictive maintenance, quality control, and safety assurance. They emphasize that real-world deployments still face challenges related to data heterogeneity, label scarcity, and the need for online adaptability. Moreover, the review identifies a growing trend towards hybrid models that integrate spatial and temporal learning, such as CNN-LSTM combinations, which aligns with the architecture of our proposed DTA-QC framework. The paper further underscores the importance of model efficiency for deployment at the edge, robustness to noise, and interpretability, all of which are explicitly addressed in our work through dynamic thresholding, compact model design, and a four-tier severity classification system.

Table 1 Comparison of DTA-QC with recent state-of-the-art approaches in test optimization and anomaly detection

References	Purpose/contribution	Limitation addressed by DTA-QC
Wen et al. (2023)	Transformer-based model selection for multivariate anomaly detection.	Assumes data stationarity and lacks dynamic threshold adaptation.
Tonini et al. (2024)	Semi-supervised anomaly detection using statistical filters and thresholds.	No validation in live production environments; offline-only use.
Tran et al. (2022)	Self-supervised anomaly detection using CNNs and pseudo-labeling for IIoT.	Binary anomaly labels only; lacks severity gradation and dynamic thresholds.
Casa and Menardi (2022)	Semi-supervised anomaly detection via nonparametric density estimation in physics.	No real-time or production constraints considered; offline experimental scope.
Liso et al. (2024)	Survey on anomaly detection for Industry 4.0 using deep learning.	Identifies real-world deployment gaps (e.g., online adaptability, interpretability) that DTA-QC directly addresses.
DTA-QC	Combines LSTM autoencoders, Ridge regression, and CNNs for test optimization in real 5G production. Supports dynamic thresholding, multi-level severity, and runs on CPU in-line.	Addresses key challenges in manufacturing quality control, including temporal signal complexity, limited labeled data, dependence on manually defined threshold rules, and constraints on deployment feasibility.

Tran et al. (2022) propose a self-supervised learning (SSL) framework for time-series anomaly detection in Industrial Internet of Things (IIoT) environments, utilizing 1D convolutional neural networks trained on pseudo-labels generated through data augmentation techniques such as rotation and jittering. Their approach eliminates the need for labeled training data and supports real-time edge deployment, demonstrating strong performance in detecting unseen anomalies with low computational cost. While their work addresses the scarcity of labeled anomalies and emphasizes lightweight inference, it primarily focuses on binary anomaly detection without integrating multi-level severity classification or adaptive thresholding mechanisms. In contrast, our DTA-QC framework leverages a hybrid architecture that combines LSTM-based autoencoders for unsupervised feature learning, Ridge regression for dynamic threshold estimation, and CNN-based classifiers to output severity-aware labels. Furthermore, DTA-QC is validated under real production conditions in 5 G radio manufacturing, highlighting its practical utility in highly variable and hardware-software coupled environments, extending beyond the IIoT scope of Tran et al.'s work.

Recent developments in other high-stakes domains further underscore the limited adoption of advanced AI techniques in industrial testing. Casa and Menardi (2022) propose a semi-supervised anomaly detection approach for particle physics applications, using nonparametric density estimation to identify unknown signals without requiring complete label sets. Their work highlights the growing use of semi-supervised learning in fields where anomaly detection is critical but labeled data is scarce. Although their method is developed for offline experimental data, the underlying statistical framework has clear relevance for real-time industrial settings such as 5 G manufacturing. In contrast to such progress in adjacent fields, industrial testing pipelines still predominantly rely on manually tuned rule-based thresholds, demonstrating a gap that frameworks like DTA-QC seek to address. To clarify the novelty and positioning of our work, Table 1 provides a side-by-side comparison of DTA-QC with recent relevant approaches in predictive maintenance, adaptive test optimization, and semi-supervised anomaly detection.

Alternative architectures for time-series modeling

Several architectures have emerged for time-series classification and anomaly detection. Transformer-based models such as TimeSformer (Bertasius et al., 2021; Gao et al., 2023), Autoformer (Wu et al., 2021), FEDformer (Zhou et al., 2022), and TimesNet (Wu et al., 2022) achieve state-of-the-art results in forecasting and representation learning. Hybrid models, e.g., CNN-LSTM (Alhussein et al., 2020), offer a balance between spatial feature extraction and temporal modeling. Graph Neural Networks (GNNs) (Wu et al., 2020) have also been applied to capture structural dependencies in sensor-rich industrial systems. Despite their modeling power, these architectures pose challenges for real-time deployment in production environments. They often require GPU acceleration, large memory capacity, and batch-level inference, which are incompatible with the constraints of real-time, per-unit telecom testing. Furthermore, models optimized for periodic or smooth signal patterns often struggle to generalize to the non-periodic, transient-dense behavior of 5 G test signals. In contrast, this study prioritizes deployability and interpretability. We evaluate three purpose-built model variants for time-series anomaly classification:

- **M1**: A lightweight 1D CNN for fast, CPU-compatible inference.
- **M2**: A CNN-BiLSTM architecture with attention mechanisms, balancing accuracy and temporal memory.
- **M3**: An oversampling-enhanced version of M2 for improved minority class detection.

These models were trained on proprietary test datasets from Ericsson AB's 5 G RBS line and validated against a public benchmark (see Sect. "Cross-domain validation with benchmark dataset"). These architectural choices prioritize deployment feasibility over theoretical performance, addressing the practical constraints of industrial environments where computational resources, real-time requirements, and interpretability are critical factors. Compared to large-scale transformer variants, our approach achieves a practical trade-off between precision, generalizability, and industrial scalability within the DTA-QC framework.

The DTA-QC framework: dynamic thresholding and anomaly-aware quality control

This section introduces DTA-QC, a modular AI-based framework specifically designed for intelligent anomaly detection and fault classification in production line testing environments. As illustrated in Fig. 3, DTA-QC is composed of three integrated stages: (1) *Supervised Data Labeling*, (2) *Data Augmentation*, and (3) *Classification*. These stages work in sequence to enable context-aware anomaly detection, adaptive fault grading, and robust model deployment in high-variability telecom manufacturing environments.

The input to DTA-QC consists of raw one-dimensional time-series data collected during functional testing of 5 G Radio Base Stations (RBSs) in production environments. These measurements are segmented into overlapping windows $(\mathrm{Window}_i, \mathrm{Window}_{i+1}, \ldots, \mathrm{Window}_n)$ to preserve temporal continuity and ensure alignment across downstream processing steps. Details on the dataset's physical meaning, signal types, and testing context are given in the upcoming subsections.

DTA-QC architecture overview

The DTA-QC framework, depicted in Fig. 3, integrates three key components, supervised data labeling, data augmentation, and classification, into a hybrid pipeline specifically tailored for manufacturing testing environments. The framework is optimized for CPU-level deployment in production environments, enabling real-time response capabilities while maintaining resilience under data imbalance and manufacturing process variability typical of high-volume telecom production lines. The architecture addresses several practical constraints in manufacturing: evolving signal distributions, sparse anomaly labels, and the need for interpretable yet adaptive decision support. The modular pipeline processes sensor data in three stages:

1. **Supervised data labeling**: A combination of autoencoders and regression models produces initial labels, with dynamic thresholds applied to identify anomalous patterns adaptively.
2. **Data augmentation**: Moving Block Bootstrap (MBB) is used to generate realistic synthetic samples, addressing class imbalance while preserving temporal structure, crucial for time-series in manufacturing (Escobar et al., 2021).
3. **Classification**: A lightweight 1D CNN model processes the balanced dataset to assign anomaly classes, enabling high-resolution, real-time anomaly detection during production line testing. CNN architectures have demonstrated particular effectiveness in real-time manufacturing anomaly detection applications (Alkahtani et al., 2024; Greinert et al., 2022)

This layered approach supports context-aware testing, enabling the system to detect weak signals of failure, generalize across hardware variants, and adapt to firmware drift,

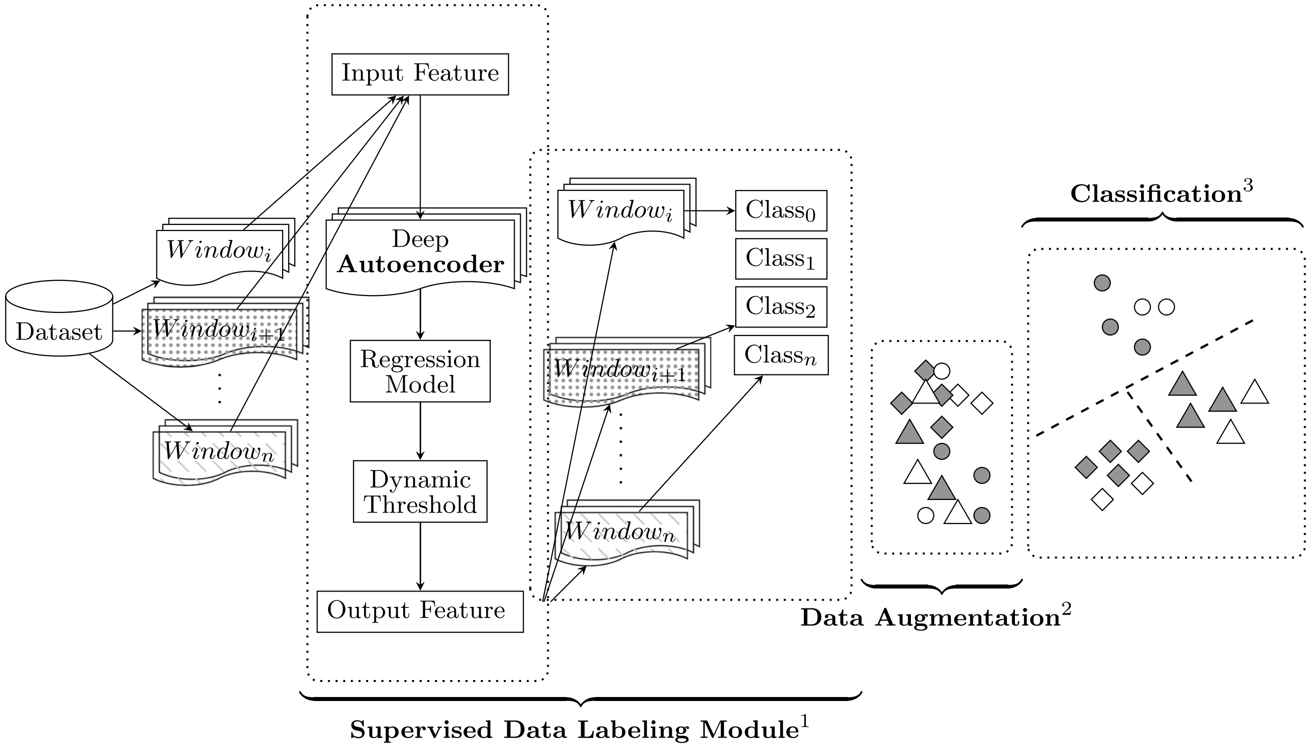

Fig. 3 Schematic overview of the DTA-QC framework, composed of three stages: **Supervised Data Labeling** for dynamic thresholding, **Data Augmentation** for enriched signal diversity, and **Classification** for severity-aware fault detection

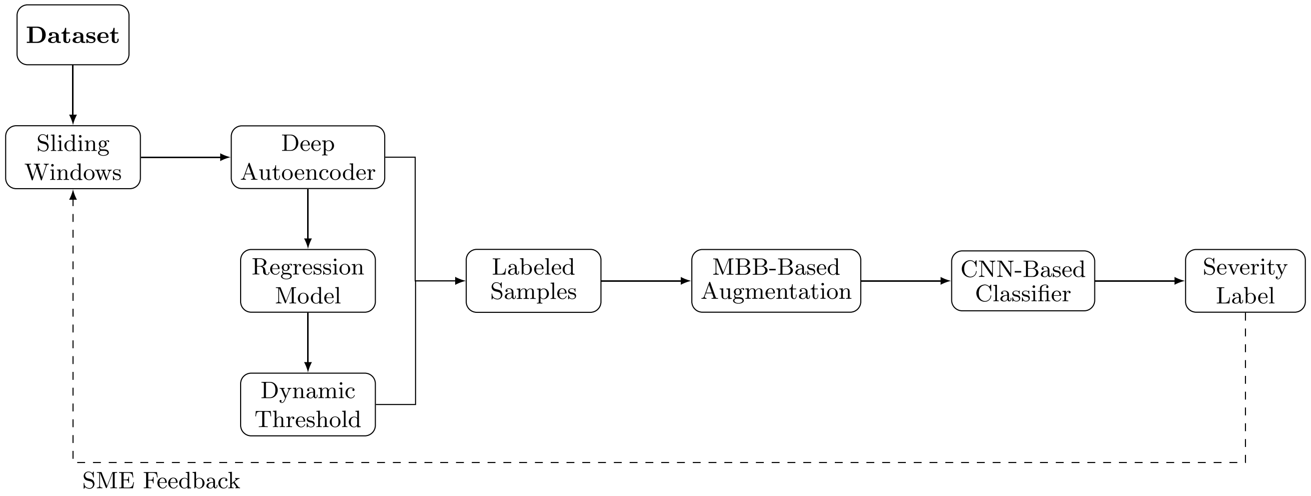

Fig. 4 DTA-QC system architecture: from raw 5 G test results to severity-aware classification through adaptive thresholding, augmentation, and supervised learning.

all without manual rule tuning or threshold recalibration. While Fig. 3 outlines the high-level components of DTA-QC, the complete operational flow, including data transformations, thresholding, and classification mechanisms, is illustrated in detail in Fig. 4.

Figure 4 presents the high-level architecture of the DTA-QC framework. The system begins by ingesting raw time-series data from 5 G production test stations. These signals are segmented into windows and passed through an autoencoder for denoising and feature compression. A regression model estimates expected behavior, and the residuals (errors) are analyzed to compute dynamic thresholds. These thresholds evolve with incoming production batches continuously, and guide the labeling of anomalies into four severity levels. To address class imbalance, a Moving Block Bootstrap (MBB) strategy is applied for data augmentation. The final labeled dataset is then used to train a supervised CNN classifier that outputs real-time quality classifications. Each module is designed for deployment on CPU-constrained industrial hardware, ensuring low-latency inference and operational scalability.

Algorithmic description and mathematical formalization

To enhance methodological clarity and reproducibility, the DTA-QC framework is formalized here, and its algorithmic workflow is summarised.

```
Input: Time-series dataset $X = \{x_t\}_{t=1}^{T}$, window length $w$
Output: Severity prediction $\hat{y}$ for new production data
Segment $X$ into overlapping windows $W = \{W_i\}_{i=1}^{N}$
for each window $W_i$ do
    Encode $W_i$ using the LSTM encoder to obtain a latent vector $z_i$
    Predict $\hat{y}_i = g_\phi(z_i)$ using the regression model
    Compute the regression residual:
        $\epsilon_i = |y_i - g_\phi(z_i)|, \quad i \in D_{\text{comb}}$
end for
Compute residual statistics over all $\epsilon_i$:
        $\mu_\epsilon = \text{mean}(\epsilon_i), \quad \sigma_\epsilon = \text{std}(\epsilon_i)$
Define adaptive thresholds for severity levels:
        $DT_k = \mu_\epsilon + k\sigma_\epsilon, \quad k \in \{1, 2, 3\}$
Assign severity class labels based on $\epsilon_i$:
        $y_i = \begin{cases} \text{Normal}, & \epsilon_i < DT_1, \\ \text{Warning}, & DT_1 \le \epsilon_i < DT_2, \\ \text{Worse}, & DT_2 \le \epsilon_i < DT_3, \\ \text{Stop}, & \epsilon_i \ge DT_3. \end{cases}$
Apply Moving Block Bootstrap (MBB) to balance minority classes
Train a 1D CNN classifier on the labeled data
return Severity prediction $\hat{y}$ for new production data
```

Algorithm 1 DTA-QC: dynamic thresholding and anomaly-aware quality control

Mathematical formalization. Let $g_\phi(\cdot)$ denote the regression model producing the estimated output $\hat{y}_i = g_\phi(z_i)$. The residuals are computed as

$$\epsilon_i = |y_i - \hat{y}_i|,$$

and the adaptive thresholds are defined following Eq. (2) as

$$DT_k = \mu_\epsilon + k\sigma_\epsilon, \quad k \in \{1, 2, 3\},$$

where μ_ϵ and σ_ϵ are the mean and standard deviation of the regression residuals for each production batch. Samples are then assigned to discrete severity levels according to the ϵ_i intervals, yielding statistically adaptive and interpretable boundaries for quality control.

Training and inference. During training, the autoencoder and regression model are optimized to minimize the reconstruction and regression losses:

$$\mathcal{L}_{AE} = \frac{1}{Nw}\sum_{i=1}^{N}\sum_{t=1}^{w}|x_t - \hat{x}_t|, \qquad \mathcal{L}_{reg} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|.$$

The CNN classifier is trained using the categorical cross-entropy loss:

$$\mathcal{L}_{cls} = -\sum_{i}\sum_{c=1}^{4} y_{ic} \log p_{ic}.$$

At inference, new production data are segmented into windows, residuals ϵ_i are computed, and severity labels are assigned using the most recent batch thresholds DT_k.

Computational efficiency. For N windows of length w, the combined LSTM–regression computation has complexity $\mathcal{O}(Nw)$, while the threshold update and CNN inference are $\mathcal{O}(N)$. The complete system operates efficiently on CPU hardware, supporting real-time deployment within industrial test stations.

Probabilistic interpretation. Assuming the regression residuals ϵ_i follow an approximately normal distribution, the probability of a "Stop" event is given by:

$$P(\text{Stop}) = 1 - \Phi\left(\frac{DT_3 - \mu_\epsilon}{\sigma_\epsilon}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. This links the adaptive thresholding mechanism to a probabilistic interpretation of risk under production uncertainty.

Supervised data labeling

This module transforms unlabeled production time-series data into structured, labeled datasets suitable for supervised machine learning model training. The raw data is segmented using a fixed-size, overlapping sliding window approach, which preserves both short- and long-term temporal dependencies by maintaining the local signal structure across samples. Each resulting window is passed through a deep autoencoder for noise reduction and dimensionality compression, followed by a regression model that predicts expected behavior. Dynamic thresholds are then computed based on the residuals, enabling tolerance bands that adapt in real time to the variability of manufacturing processes. Such adaptive thresholding is particularly critical in Industry 4.0 environments, where fixed thresholds often fail to reflect shifts in hardware behavior, software updates, or environmental conditions (Li et al., 2024; Orabi et al., 2024). Final labels are refined through a CNN that captures localized signal variations, producing well-separated and context-aware labels suitable for downstream supervised classification. This dynamic and modular labeling approach introduces several key innovations:

- Integration of autoencoder-based feature extraction with regression-driven dynamic thresholding to enable adaptive data labeling.
- Elimination of brittle rule-based thresholds that are prone to false positives as production conditions evolve.
- Support for explainability and traceability through clear separation between prediction, thresholding, and labeling stages.

To finalize the labeling scheme, we initially experimented with more granular severity levels, up to 10 classes, including categories such as *Slight Drift*, *Frequent Outliers*, *Transient Spikes*, and *Noisy but Acceptable*. However, empirical evaluations and expert feedback from SMEs indicated semantic overlaps among these intermediate classes, which compromised interpretability and reduced robustness in noisy production settings. As a result, we consolidated the labeling scheme into four distinct and actionable levels: *Normal*, *Warning*, *Worse*, and *Stop*. These align with decision points on the shop floor and enhance both model stability and human interpretability. Importantly, the framework remains flexible and could be extended with additional severity levels in future applications, provided reliable data separation can be achieved.

Data augmentation

To address the severe class imbalance inherent in manufacturing anomaly detection scenarios, where defective units often represent less than 1% in a mature production stage, we apply Moving Block Bootstrap (MBB)-based augmentation. Traditional synthetic oversampling methods like SMOTE (Synthetic Minority Over-sampling Technique) generate new samples by interpolating between existing minority instances. While effective in tabular data, SMOTE may distort temporal dependencies in time-series data by breaking the inherent ordering and signal structure. In contrast, MBB preserves temporal coherence by resampling contiguous blocks of residual sequences, creating new time-series windows that maintain meaningful temporal patterns while increasing minority class representation. A semi-supervised model assigns tentative labels to synthetic data, resulting in a more balanced and robust training set without compromising the statistical integrity of the original distribution. The advantages of using MBB in this context are as follows:

- Retains autocorrelation and contextual continuity across time steps, which is critical for accurate anomaly detection in temporal sequences.
- Avoids overfitting risks commonly associated with naive duplication or interpolation-based techniques.
- Reduces model bias toward majority classes and significantly improves sensitivity to rare but critical faults in production systems.

Classification

In the final stage of the DTA-QC pipeline, a 1D Convolutional Neural Network (CNN) is used to classify time-windowed sensor signals into predefined anomaly severity levels. The CNN extracts hierarchical temporal features through stacked convolutional layers and produces multi-class predictions corresponding to four severity categories: Normal, Warning, Worse, and Stop. The model is trained on an augmented, labeled dataset and evaluated using standard metrics such as the confusion matrix, ROC-AUC, and F1-Score. The classifier is optimized for low-latency execution on standard CPU hardware, ensuring seamless deployment within existing industrial test stations without the need for specialized GPUs. To support operational performance and system compatibility, this classification stage provides the following key features:

- Real-time anomaly classification based on learned temporal features from production signals.
- High interpretability through consistent feature windowing and dynamic class boundaries.
- Robust performance under domain shift, enabled by integrated adaptive learning and MBB-based data augmentation.

By replacing fixed-threshold logic with intelligent, data-driven decision support, the DTA-QC framework ensures scalable and reliable fault detection in dynamic 5 G manufacturing environments. Its modular architecture allows for straightforward integration and cross-product adaptability, supporting long-term maintainability and system evolution.

Implementation

This section details the technical implementation of the DTA-QC framework, which integrates multiple AI/ML components for time-series analysis in industrial 5G production environments. DTA-QC processes raw production test data through a multi-stage pipeline consisting of auto-labeling, dynamic thresholding, data augmentation, and classification, as illustrated in the system architecture shown in Fig. 3. Temporal characteristics, such as window-based features and rolling statistics, are key to extracting meaningful patterns in industrial settings (Engström et al., 2020). Quality data preparation not only improves model accuracy but also enhances interpretability, an essential requirement in operational environments. While proprietary constraints prevent the sharing of raw data, the complete implementation is publicly available on GitHub Liu and Tahvili (2025a), with supplementary resources on Figshare Liu and Tahvili (2025b).

DTA-QC employs LSTM autoencoders for time-aware feature extraction, regression-based dynamic thresholding for auto-labeling, Moving Block Bootstrap for data augmentation, and a lightweight 1D CNN for classification. These components were selected to balance predictive power with computational efficiency, prioritizing real-time deployability over architectural complexity. The implementation prioritizes computational efficiency and real-time performance to meet the stringent requirements of high-volume 5 G production lines. All components are designed for deployment in resource-constrained industrial environments while maintaining the accuracy needed for quality control decisions. The modular architecture enables selective activation of components based on specific production line requirements and available computational resources.

Input signal origin and representation

The input data to the DTA-QC framework consists of 1D time-series measurements collected during the automated final testing of 5 G radio units. Each unit undergoes a predefined sequence of test cases that measure analog parameters such as output power, voltage, current, temperature, and gain across multiple radio ports. These measurements are captured in real time as the unit interacts with test equipment via RF interfaces. Each test case produces a single measurement per port, and across production, this generates a sequence of observations for the same test case across different units. These sequences form the 1D time series used in our framework. While individual measurements are not time-dependent per se, we treat them as temporally ordered across units to model production dynamics, capture behavioral drift, and detect anomalies that may indicate hardware degradation or process shifts. The test infrastructure executes these cases in an automated fashion, and failures are logged based on fixed thresholds. However, this traditional model lacks early warning signals or fine-grained failure reasoning. DTA-QC builds on this setup by converting raw signals into sliding windows, applying denoising and compression via autoencoders, and learning to identify anomalous patterns (i.e., outliers) that precede failures, enabling earlier intervention and reducing false rejects. Only analog signals are used to ensure consistent, interpretable input. In earlier work (Landin et al., 2023), we showed that signal evolution across production units holds strong predictive value for downstream failure. The present study generalizes this idea by integrating dynamic thresholding and severity classification, aiming to flag at-risk units before hard failures occur proactively.

LSTM autoencoders

LSTM autoencoders (Hochreiter & Schmidhuber, 1997) form the backbone of the dynamic thresholding stage. As illustrated in Fig. 5, each input sample corresponds to a fixed-length sliding window segment of the original time-series signal. These segments capture both short- and long-term temporal dependencies, which the encoder compresses into a latent representation (*Code*). The decoder then attempts to reconstruct the original window. The reconstruction error, measured as Mean Absolute Error (MAE), informs the dynamic thresholding logic described later in this subsection. This structure enables the framework to capture subtle signal deviations while preserving temporal context in a compact and interpretable format.

The transformation process retains temporal continuity by converting 1D signals into windowed sequences. These are passed through the autoencoder to denoise and compress the data, preserving only the most informative patterns. The resulting reconstruction is used to evaluate prediction error, from which dynamic thresholds are derived (see Sect. ”Dynamic thresholding”). During encoding, the autoencoder transforms raw time-series windows into compact representations, capturing both short- and long-term dependencies. These are then decoded and compared with the original input to calculate the reconstruction error. The Mean Absolute Error (MAE) is used to detect anomalies by setting adaptive thresholds derived from historical reconstruction error distributions. As an illustrative example, consider a normalized input sequence:

$$x = [10, 12, 14, 13, 15]$$

Using a 3-step sliding window, we create overlapping segments such as:

$$\text{Window}_1 = [10, 12, 14], \quad \text{Window}_2 = [12, 14, 13], \\ \text{Window}_3 = [14, 13, 15]$$

Each window is passed through the LSTM autoencoder. Suppose that for Window_2, the model reconstructs the sequence as:

$$\hat{x}_{\text{Window}_2} = [12.1, 14.2, 12.8]$$

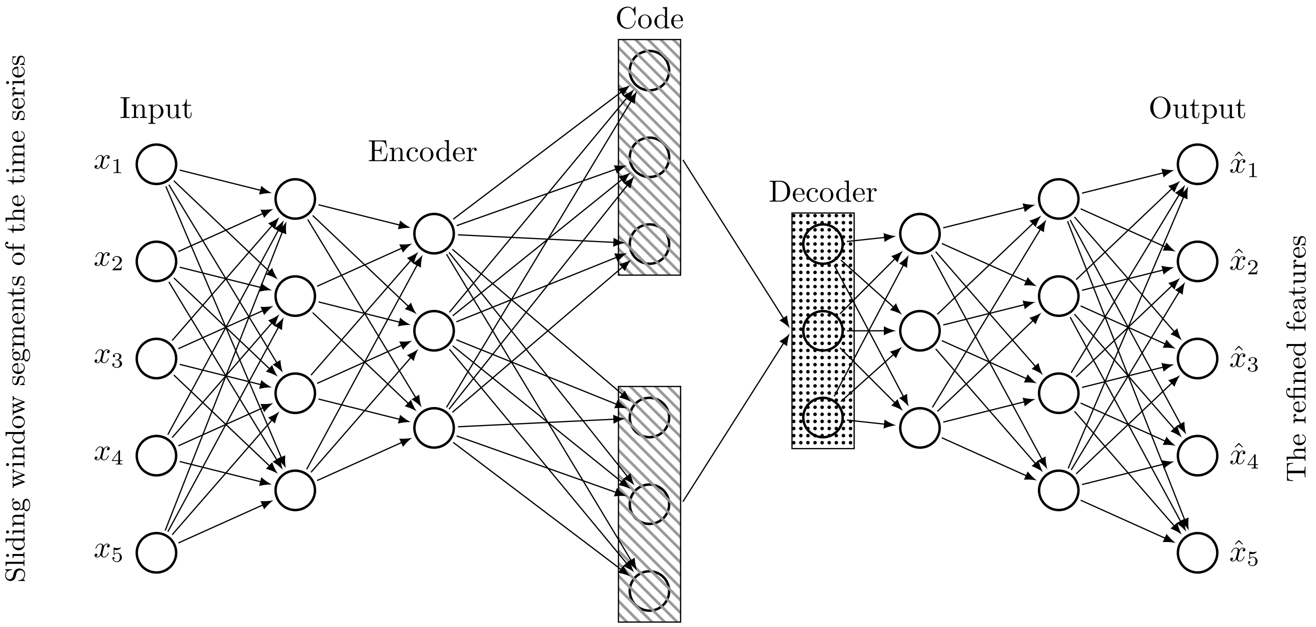

Fig. 5 Architecture of the LSTM autoencoder in DTA-QC, used for time-series feature extraction and anomaly-aware labeling

The MAE for this window is computed as:

$$\text{MAE} = \frac{1}{3}(|12 - 12.1| + |14 - 14.2| + |13 - 12.8|) = 0.167$$

If the dynamic stop threshold is 0.15, then this window would be flagged as a "Stop" anomaly (severity class 3). Conversely, a window with a lower MAE (e.g., 0.12) would not be flagged. This mechanism allows the model to flag samples where the reconstructed signal deviates significantly from the expected pattern.

The benefit of this approach lies in its ability to:

- Detect emerging faults before they violate static limits.
- Adapt to firmware changes, hardware drift, or seasonal variation.
- Automatically generate labels to reduce manual effort.

The use of LSTM autoencoders for feature extraction in industrial applications has demonstrated effectiveness in capturing complex temporal patterns that traditional statistical methods often miss (Benkedjouh et al., 2018). The resulting threshold-aware dataset is passed to a regression model and CNN for fine-grained classification, enabling robust, real-time decision support in production testing scenarios.

Data labeling via regression models

Encoded feature vectors ($\hat{x}$) are passed into a regression model to estimate expected values ($\hat{y}$). Dynamic thresholds (DT) are computed based on prediction error, specifically, the Mean Absolute Error (MAE) and its variance across the dataset:

$$DT = \hat{y} \pm (\text{MAE} + \text{scale} \cdot \sigma_{\text{MAE}}) \tag{2}$$

This adaptive thresholding approach builds upon recent advances in dynamic threshold adjustment for manufacturing systems (Wang et al., 2021), where traditional static limits prove insufficient for modern production environments with varying operational conditions.

Dynamic thresholding

Dynamic thresholds adapt continuously to production conditions, distinguishing this approach from traditional static limit-based quality control systems common in manufacturing. The thresholds evolve with each batch of new data, ensuring responsiveness to shifts in production conditions such as component ageing, environmental variations, or process improvements. After every model retraining cycle, the updated MAE is used to adjust class boundaries. While retraining is currently conducted offline to ensure production stability, the architecture supports future transition to continuous learning for more responsive adaptation.

In the DTA-QC framework, dynamic thresholds are updated at the end of each model retraining cycle. When a new batch of labelled production data becomes available, the autoencoder is retrained, and the residual error distribution is recomputed, specifically the Mean Absolute Error (MAE) and its standard deviation (σ_{MAE}). These values define the severity class boundaries (Normal, Warning, Worse, Stop) using a statistical margining approach: thresholds are set at $\mu_{\text{MAE}} + k \cdot \sigma_{\text{MAE}}$ for severity level k, where $k \in \{1, 2, 3\}$ defines the transitions between classes. For example, an MAE below $\mu_{\text{MAE}} + \sigma_{\text{MAE}}$ is classified as Normal, between one and two standard deviations as Warning, and so on. Between retraining events, thresholds remain fixed to ensure operational stability during inference. The initial thresholds, computed on the training set, serve as the reference baseline for subsequent updates. This update policy allows the system to adapt to production drift while maintaining consistent decision logic.

Data augmentation via moving block bootstrap

Given the significant class imbalance in the labeled dataset, an augmentation strategy is applied using the MBB method. This technique generates synthetic time-dependent data by resampling contiguous blocks, thereby preserving the temporal dependencies inherent in the original dataset. The goal is to enrich the dataset with realistic samples that maintain structural integrity while improving class balance. This facilitates more effective training of the classification model and enhances its ability to generalize to unseen data. To address class imbalance, DTA-QC applies a hybrid augmentation strategy based on the Moving Block Bootstrap (MBB). As illustrated in Fig. 6, lagged blocks (Y_i) are combined with randomly sampled residuals ($R_i = Y_{i+1} - Y_i$) to generate synthetic samples that retain temporal structure.

Mathematical formulation: Given a sequence of overlapping lagged blocks $Y_i \in \mathbb{R}^d$, we compute the first-order residuals as:

$$R_i = Y_{i+1} - Y_i, \quad i = 1, 2, \ldots, N-1$$

where R_i represents the temporal change between consecutive blocks. These residuals are then resampled with replacement to generate synthetic sequences:

$$\tilde{Y}_i = Y_i + R_i^*,$$

where R_i^* denotes a randomly sampled residual from the set $\{R_1, R_2, \ldots, R_{N-1}\}$. This preserves local temporal trends while introducing realistic variation for minority classes.

This method is preferable to SMOTE in manufacturing time-series contexts, as it preserves the temporal autocorrelation and sequence continuity that are critical for representing realistic production dynamics. The preservation of temporal dependencies is especially essential for accurate anomaly detection in industrial environments where faults often manifest through subtle, time-dependent signal patterns. This is particularly relevant in Industrial Internet of Things (IIoT) systems, where distributed sensors generate high-frequency, multi-dimensional data streams that must be analyzed in near real time. For rare anomaly classes that occur infrequently in production data, MBB can be optionally combined with SMOTE to increase sample diversity while still maintaining temporal coherence. Recent studies have demonstrated that temporal-preserving data

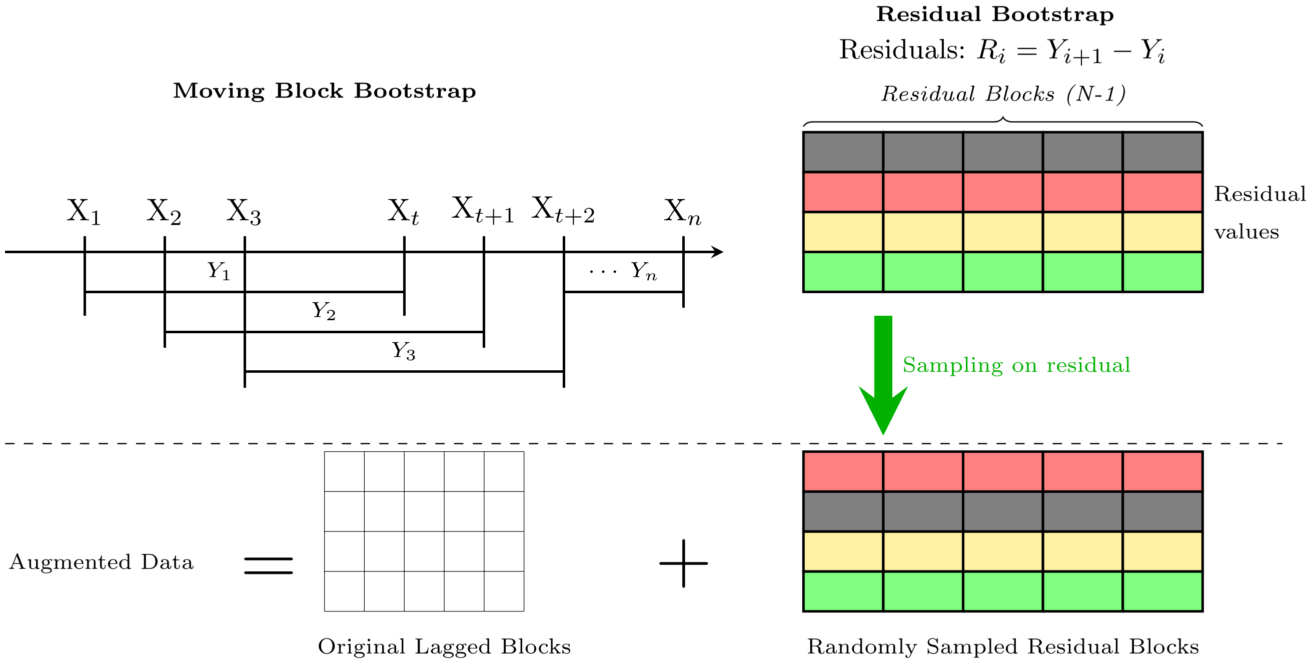

Fig. 6 Illustration of Moving Block Bootstrap (MBB) used in DTA-QC to augment imbalanced classes while preserving temporal dependencies

augmentation techniques significantly outperform conventional interpolation methods for time-series anomaly detection in IIoT applications, due to their ability to retain realistic sequence behavior and minimize artificial signal distortion (Kim & Lee, 2024). It is important to clarify that our approach does not involve traditional fault injection through artificial signal distortions. Instead, the Moving Block Bootstrap (MBB) technique used here resamples contiguous temporal blocks from authentic production data. This ensures that the statistical and temporal properties of real test signals are preserved. Unlike physical fault injection methods, which introduce speculative distortions (e.g., simulated noise, component drift), our method generates realistic synthetic samples by recombining real patterns already observed in the system. This preserves the fidelity and interpretability of the model for actual manufacturing scenarios.

Classification with 1D CNNs

The final classification stage uses a 1D CNN architecture optimized for real-time performance. The model consists of six convolutional layers grouped in three blocks (16–16, 64–64, 64–64 filters), each followed by max-pooling. The flattened output (size 448 × 1) is fed into a softmax classifier. Implemented in TensorFlow/Keras, the CNN was selected over GRUs and Transformers due to its superior inference speed and lower hardware requirements, critical factors for deployment in production line environments where decisions must be made within milliseconds. The model's computational efficiency makes it well-suited for CPU-bound, latency-sensitive industrial environments typical of manufacturing facilities, where dedicated GPU resources may not be available or cost-effective. Hyperparameters were tuned using early stopping and learning-rate scheduling. The convolutional module in DTA-QC is designed to capture both local signal variations and broader temporal patterns in the augmented time-series data, enabling accurate multi-class severity classification under realistic production conditions. The use of lightweight CNN architectures ensures that the system remains suitable for real-time manufacturing environments, aligning with recent findings that high classification performance can be maintained without incurring substantial computational overhead in industrial quality control tasks (Zhang et al., 2023).

Figure 7 illustrates the operational integration and decision logic of the 1D CNN-based severity classifier within the production line. Test signals are continuously collected from test stations and processed by the inference engine, which embeds both the trained autoencoder and CNN classifier. The classifier assigns each incoming signal to one of four severity levels: *Normal*, *Warning*, *Worse*, or *Stop*. Each severity level is mapped to a distinct quality control action: a *Normal* label results in a direct *Proceed to Next Test*; a *Warning* triggers an *Increase in Sampling Rate* (e.g., by 10%); a *Worse* label prompts a *Manual Review or Supplier Alert*; and a *Stop* classification initiates a *Batch Re-inspection* or activates the *Maintenance System*. This structured decision logic transforms model outputs into actionable process steps, enabling real-time, data-driven quality assurance, improving traceability, and enhancing overall production resilience.

Empirical evaluation

To evaluate the effectiveness of the DTA-QC framework under real-world manufacturing conditions, we conducted an industrial case study at Ericsson AB (EAB) in Sweden. The evaluation took place within EAB's full-scale 5 G radio production testing environment, which processes thousands of units daily and integrates both manual and automated test methodologies across multiple validation stages. This industrial setting provides the high-volume, time-critical

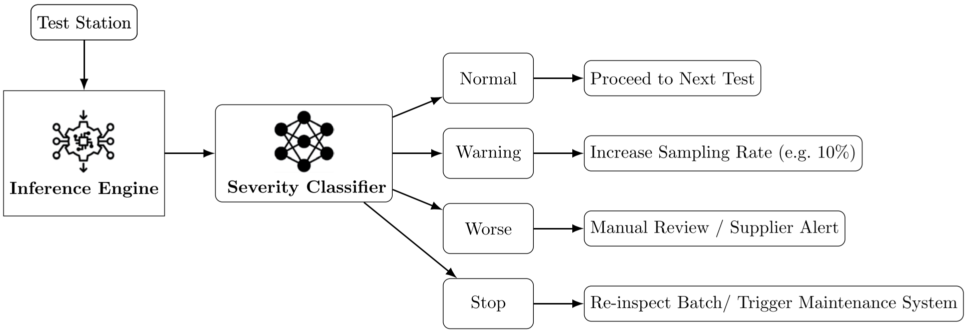
Test Station
Inference Engine
Severity Classifier
Normal
Warning
Worse
Stop
Proceed to Next Test
Increase Sampling Rate (e.g. 10%)
Manual Review / Supplier Alert
Re-inspect Batch/ Trigger Maintenance System

Fig. 7 System-level decision logic of DTA-QC. Signals are collected from test stations, processed through the inference engine and severity classifier, and assigned one of four severity levels: Normal, Warning, Worse, or Stop. Each severity class is linked to a specific quality control action: proceeding to the next test, increasing sampling rate, triggering manual review or supplier alerts, and batch-level re-inspections or maintenance. This schematic directly operationalizes model outputs into production decisions, enhancing the system's transparency and traceability

context typical of modern manufacturing environments where quality control decisions must be made rapidly and accurately. This setup provided a representative and rigorous operational context for empirical analysis. The dataset used comprises proprietary time-series measurements collected during functional and parametric testing of 5 G RBSs in a production environment. Each unit undergoes comprehensive testing protocols using specialized hardware interfaces, or test stations, that interact with embedded software to assess key operational metrics such as power levels, temperature stability, voltage, and signal integrity. These measurements represent real production conditions, including natural variations in component tolerances, environmental factors, and manufacturing process variations that are characteristic of high-volume electronics manufacturing. These metrics are captured as one-dimensional time-series sequences that reflect the performance behavior of each unit across testing cycles. Due to confidentiality agreements, the complete dataset cannot be made publicly available. However, a representative anonymized subsample is presented in Table 2, which preserves the structure, format, and statistical characteristics of the original data while ensuring compliance with data protection requirements.

Key features extracted from the raw production dataset include:

- **Test station:** Identifies the specific automated test equipment (ATE) interface executing a given test sequence, enabling traceability to specific hardware configurations.
- **Test category:** Encodes the functional test type (e.g., digital signal processing, analog RF performance, calibration procedures) according to production test specifications.
- **Test count:** Indicates the number of test attempts required to meet pass criteria, providing insight into test repeatability and potential quality issues.
- **Limits 1 and 2:** Define the engineering tolerance range for measured values (upper and lower specification limits) based on design requirements and quality standards.
- **Measured value:** The actual measurement result from production test equipment, representing the physical performance characteristic being evaluated.

This subsample provides transparency on the input structure used by DTA-QC and allows researchers to understand the context in which the system was trained, labeled, augmented, and evaluated. By preserving temporal and statistical fidelity while anonymizing operational details, the evaluation supports reproducibility within the constraints of industrial confidentiality.

The selection of baseline models reflects both scientific reproducibility and industrial deployability constraints. Comparative baselines are included to contextualize the performance of the proposed DTA-QC framework and to validate its advantages over existing, widely adopted techniques. The primary neural baseline (M1) is a lightweight 1D CNN architecture designed for deterministic real-time inference on CPU-only systems, aligning with resource constraints at production test stations. Classical comparators, Logistic Regression, Support Vector Machine, and Random Forest, were chosen for their simplicity, maintainability, and wide adoption in manufacturing analytics. All baselines were tuned within a shared hardware envelope: CNN hyperparameters were selected to balance accuracy and computational efficiency, trained with the Adam optimizer and early stopping; classical models utilized compact, practitioner-typical configurations. While more complex models, such as TimesNet or Transformer variants, exist in recent literature, they generally assume larger, balanced datasets and GPU compute resources, making them impractical for current production test environments. Thus, our baseline selection aims to represent feasible, CPU-deployable solutions that support fair, actionable benchmarking in industrial settings.

Experimental setup and parameters

This subsection outlines the experimental setup and key parameters used to evaluate the DTA-QC framework. The system comprises three core components: (1) a 1D Convolutional Neural Network (CNN) for classification, (2) a Bidirectional LSTM autoencoder for feature extraction, and (3) a Ridge Regression model for anomaly scoring. The CNN architecture (see Table 3) consists of two Conv1D layers using hyperbolic tangent (`tanh`) activation functions, which produce zero-centered outputs to facilitate efficient gradient propagation. These are followed by

Table 2 Representative subsample of production test data used in the evaluation, showing the structure and format of real 5 G manufacturing test results

Test station	Test category	Test count	Limit 1 (dBm)	Limit 2 (dBm)	…	Measured value
1	3	938	12	−12	…	−18.18
2	1	34	−8	−13	…	−20.89
3	3	907	1	−12	…	0.00
⋮	⋮	⋮	⋮	⋮	⋱	⋮
9559	2	1058	−13	−15	…	−21.14

Values are anonymized to protect proprietary information while preserving statistical characteristics

Table 3 CNN and baseline model configuration parameters

Model	Component	Key parameters
1D CNN	Convolution layers	2 layers: 16 and 64 filters, kernel size = 3, activation = tanh
	Pooling	MaxPooling1D, pool size = 3
	Dense	Flatten, Dense layer with 4 neurons (softmax activation)
	Training	Optimizer: Adam (LR = 0.01), Loss: Categorical Crossentropy, Epochs: 100, Batch size: 32
Random forest	–	100 estimators, default depth
SVM	–	RBF kernel, $C = 1.0$, probability enabled
Logistic regression	–	Solver = liblinear or lbfgs, L2 regularization

Table 4 Parameters of the LSTM autoencoder

Layer	Details
Bidirectional LSTM	50 units, activation: relu, input shape: (1, 19)
RepeatVector	Length: 1 timestep
Bidirectional LSTM	50 units, activation: relu, return sequences: True
TimeDistributed (Dense)	1 unit output per timestep
Optimizer	Adam
Loss	Mean squared error
Training	EarlyStopping with patience = 10

Table 5 Ridge regression configuration details

Parameter	Details
Model type	RidgeCV (auto alpha selection)
Train-test split	70% training, 30% testing (time-ordered)
Cross-validation	5-fold TimeSeriesSplit
Regularization	L2 with auto-tuned alpha
Scaling	StandardScaler applied pre-training
Evaluation	R^2, RMSE

MaxPooling1D layers for temporal downsampling and fully connected dense layers for final classification. The model is trained using the Adam optimizer with categorical cross-entropy loss for 100 epochs.

The CNN architecture parameters were selected to balance classification accuracy with computational efficiency, suitable for deployment on standard industrial computing hardware commonly found in manufacturing environments.

For feature extraction, a Bidirectional LSTM autoencoder (Table 4) captures both past and future dependencies. A RepeatVector ensures that temporal structure is preserved through encoding-decoding. Ridge Regression (Table 5) is used for anomaly scoring based on autoencoded features. Its L2 regularization helps control overfitting, and cross-validation ensures robustness.

All experiments and hardware usage benchmarks reported in this study were conducted on a standard CPU-only machine: a MacBook Air (Model Identifier: Mac14,2) equipped with an Apple M2 chip featuring 8 CPU cores (4 performance and 4 efficiency cores), 16 GB LPDDR5 memory, and integrated graphics. No dedicated GPU was used. This setup was selected intentionally to reflect realistic deployment constraints in production environments, where models must operate efficiently without requiring high-end servers or accelerators. The reported CPU and memory usage results (Table 9) should be interpreted within the context of this hardware profile.

Unit of analysis and procedure

The unit of analysis in this study consists of one-dimensional time-series outputs generated during the production testing of 5 G RBS units at Ericsson AB's manufacturing facility. Each time-series sequence represents the temporal behavior of measured electrical parameters, such as voltage and power, recorded throughout the execution of standardized test procedures. These sequences form the primary input for downstream processes, including feature extraction, dynamic thresholding, and anomaly classification. Importantly, each sequence corresponds to an individual production unit and captures fine-grained temporal variations that static measurements fail to reveal, making them particularly suitable for detecting intermittent faults and early-stage performance degradation.

Typical manifestations of anomalies within the dataset include sustained deviations beyond upper or lower test limits, transient spikes, signal flattening, and irregular oscillations. These patterns impact critical Key Performance Indicators (KPIs) such as transmission power stability, DC bias regulation, and RF linearity, each essential to the operational integrity of 5 G RBS components. The anomalies are classified as such when deviations occur outside tolerance bands defined by dynamic thresholds derived from historical behavior and process specifications. Moreover, these patterns exhibit temporal evolution: anomalies may escalate gradually, persist intermittently, or serve as early indicators of critical system faults. This temporal dependency is particularly valuable for enabling predictive decision support, e.g., if an early-stage "Worse" signal deviation is detected during a test case, operators can halt downstream test execution and flag the unit for investigation. Such anticipatory behavior reduces resource waste and enhances test throughput, aligning with the real-time requirements of intelligent manufacturing systems.

Table 6 summarizes the key descriptive statistics of the dataset, highlighting its size and variability across test stations, test categories, and measured values. The range and distribution of test counts and signal amplitudes underscore the heterogeneity inherent in high-throughput electronics

Table 6 Descriptive statistics of the 5G RBS production test dataset, highlighting variability across stations, categories, and measurement ranges

	Test station	Test category	Test count	Limit 1	Limit 2	Measured value
Count	9559	9559	9559	9559	9559	9559
Mean	4396.06	200.00	143.40	−0.60	0.30	−0.25
Std	3405.14	0.00	412.03	0.00	0.00	1.45
Min	1000.00	200.00	1.00	−0.60	0.30	−72.85
25%	1005.00	200.00	1.00	−0.60	0.30	−0.25
50%	4490.00	200.00	1.00	−0.60	0.30	−0.22
75%	8159.00	200.00	3.00	−0.60	0.30	−0.18
Max	9804.00	200.00	1950.00	−0.60	0.30	0.36

manufacturing, reinforcing the need for adaptive and scalable quality control approaches like DTA-QC.

Cross-domain validation with benchmark dataset

To evaluate the generalizability of the DTA-QC framework beyond its industrial deployment at Ericsson AB, we conducted cross-domain validation using a publicly available time-series benchmark dataset (Mattera et al., 2025). This validation evaluates the framework's ability to maintain robustness and predictive accuracy across diverse application contexts, including both industrial and non-industrial domains. Such assessment is essential for demonstrating real-world adaptability, particularly in environments characterized by different signal dynamics, noise characteristics, and class imbalance. The benchmark dataset used for evaluation comprises multivariate time series labeled with varying levels of anomaly severity. To align it with the DTA-QC classification scheme, we mapped its anomaly types into four corresponding categories: *Normal*, *Warning*, *Worse*, and *Stop*. All features were normalized and segmented using the same windowing strategy as the industrial dataset, ensuring consistency in preprocessing across domains.

Model variants

To explore how model complexity and data balancing strategies affect performance in a domain-shifted context, we evaluated three architectural variants of the DTA-QC framework:

1. **M1: Lightweight 1D CNN (Baseline)**, a computationally efficient model comprising two 1D convolutional layers with max pooling and a fully connected classification head, designed for deployment in resource-constrained industrial environments. Designed for inference in CPU-only environments, M1 uses downsampling to manage class imbalance. It serves as the baseline for comparison.
2. **M2: CNN + BiLSTM + Attention (Attention-Augmented)**, this model enhances temporal learning and contextual awareness by integrating:
   - A one-dimensional convolutional (Conv1D) layer to extract local temporal patterns from input sequences.
   - A bidirectional LSTM (BiLSTM) layer to capture both past and future dependencies in the time series.
   - A self-attention mechanism (implemented using Keras) that dynamically assigns importance weights to different time steps, improving focus on informative segments.
   - Fully connected (dense) layers with ReLU (Rectified Linear Unit) activation, L2 regularization, and dropout, which enhance model generalization and reduce the risk of overfitting.

This architecture is designed to improve classification precision on minority classes and better handle dynamic temporal variations present in real-world production data.

3. **M3: CNN + BiLSTM with Oversampling (Oversampled)**, this variant replicates M2's architecture but replaces attention with aggressive oversampling of minority classes using random duplication. It assesses the impact of data-level balancing when applied to class-imbalanced sequences from another domain.

All models were trained using the Adam optimizer and categorical cross-entropy loss for 20 epochs with a batch size of 64. We excluded early stopping to ensure full convergence and consistent cross-model comparisons. Evaluation metrics included macro- and micro-averaged AUC, precision, recall, and F1-Score across all four classes. This comparative analysis not only demonstrates the adaptability of DTA-QC but also offers insights into the trade-offs between architectural complexity, model interpretability, and cross-domain robustness in time-series anomaly detection.

Dataset description

To assess the cross-domain generalization capability of the DTA-QC framework and its variants, we utilized the Server Machine Dataset (SMD), a publicly available multivariate time-series dataset from Kaggle (Gusat, 2023). Originally collected over 5 weeks from a major Internet company, the SMD dataset is widely used in research focused on anomaly detection in complex IT and cyber-physical systems. The dataset contains operational data from 28 server machines, categorized into three subgroups, *machine-1-1*, *machine-1-2*, and *machine-1-3*, each representing a distinct hardware configuration or workload profile. Each record comprises multiple telemetry signals, including CPU utilization, memory usage, disk I/O, and network traffic, recorded at fixed time intervals. These multivariate sequences offer a rich foundation for analyzing system behavior and detecting performance anomalies. Ground-truth labels indicate whether each time point belongs to a normal or anomalous regime. Anomalies are synthetically injected to reflect real-world failure modes, including memory leaks, process overloads, and sudden spikes in system activity. This labeling scheme enables supervised training and benchmarking of models under controlled but realistic fault conditions. For consistency, we applied the same sliding window segmentation and feature extraction strategy used in the Ericsson dataset. Anomalous segments were mapped into the four-tier severity scheme of the DTA-QC framework: *Normal*, *Warning*, *Worse*, and *Stop*. Data normalization and label alignment ensured comparability across domains. All three models (M1-M3) were evaluated on this dataset using standard classification metrics: Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC). This comprehensive performance assessment allows us to measure not only classification effectiveness but also the impact of architecture and sampling strategies under distributional shift. The SMD benchmark thus serves as a robust testbed for validating the real-world applicability and resilience of the proposed DTA-QC framework.

Table 7 Performance comparison of models M1, M2, and M3 on the benchmark dataset

Metric	M1 (Baseline CNN)	M2 (CNN + BiLSTM + Attention)	M3 (CNN + BiLSTM + Oversampling)
Accuracy (%)	74.01	**92.47**	61.66
Precision (%)	71.53	**82.39**	69.72
Recall (%)	51.50	53.53	**61.66**
F1-Score (%)	53.70	59.48	**61.76**
ROC-AUC (%)	89.73	**93.78**	86.84

Bold indicate the best performance across the compared models for each evaluation metric

Note: Bold values in Tables 7 and 8 indicate the best performance across the compared models for each evaluation metric.

The selection of the SMD dataset is particularly relevant for manufacturing applications, as server machine telemetry shares similar characteristics with industrial IoT sensor data commonly found in smart manufacturing environments. The multivariate nature of the data, with metrics including CPU utilization, memory usage, and network traffic, parallels the complex sensor arrays typical in modern production line monitoring systems. This structural resemblance makes the dataset well-suited for evaluating the generalizability and robustness of the DTA-QC framework beyond the 5 G test environment.

Quantitative results

Table 7 presents the comparative performance of the three models (M1–M3) on the SMD benchmark dataset using five evaluation metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC. These metrics provide a comprehensive assessment of classification performance, anomaly detection sensitivity, and cross-domain generalisation capability, which are critical factors for industrial deployment.

Model M1 (Baseline CNN) delivers moderate accuracy and high precision but suffers from low recall, indicating limited sensitivity to minority class anomalies. Its shallow architecture and downsampling strategy constrain its ability to model temporal dependencies and rare events.

Model M2 (CNN + BiLSTM + Attention) achieves the best overall performance, with the highest accuracy (92.47%), precision (82.39%), and ROC-AUC (93.78%). Its BiLSTM layer and self-attention mechanism improve temporal feature learning and class separation. However, recall remains moderate, revealing room for improvement in minority detection.

Model M3 (CNN + BiLSTM with Oversampling) performs best in recall and F1-Score, suggesting improved sensitivity to rare but critical events. Its balanced class-wise performance, despite lower accuracy, makes it well-suited for safety-critical anomaly detection scenarios.

Figures 8, 9, 10 further illustrate the ROC performance of each model. M1 exhibits a micro-AUC of 0.83 and macro-AUC of 0.69, struggling particularly with class 2 (AUC = 0.52). M2 achieves a micro-AUC of 0.95, benefiting from attention-enhanced temporal encoding. M3 delivers the most balanced results, with both micro- and macro-AUCs at 0.87, and per-class AUCs between 0.83 and 0.92.

The comparative analysis reveals distinct performance-efficiency trade-offs: M2 delivers optimal classification accuracy and precision for general anomaly detection, while M3 provides superior recall and class balance, making it

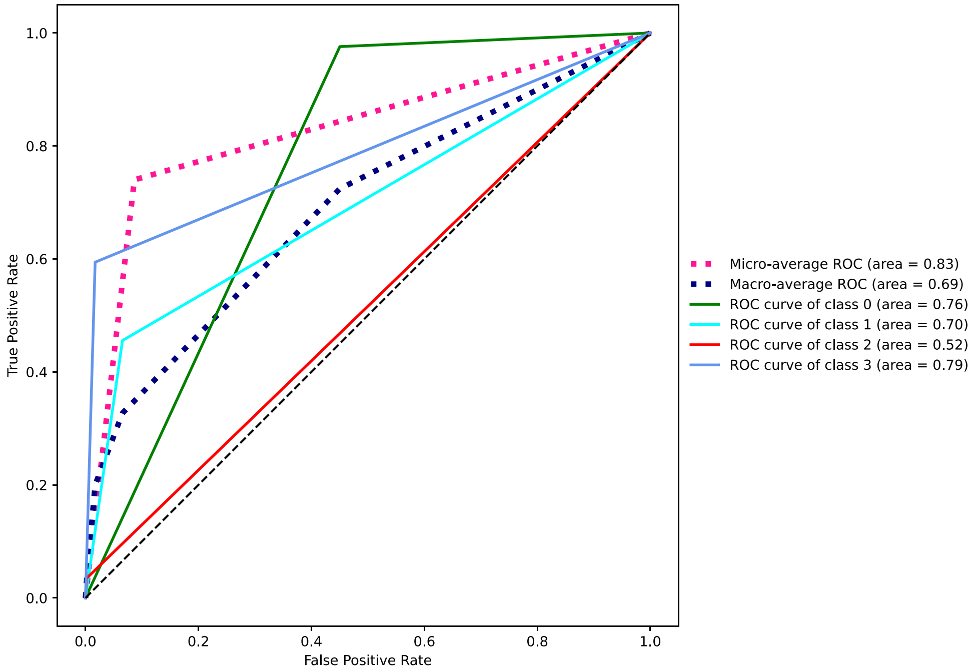

Fig. 8 ROC curves for Model M1 (Baseline CNN): limited separability, especially for minority classes

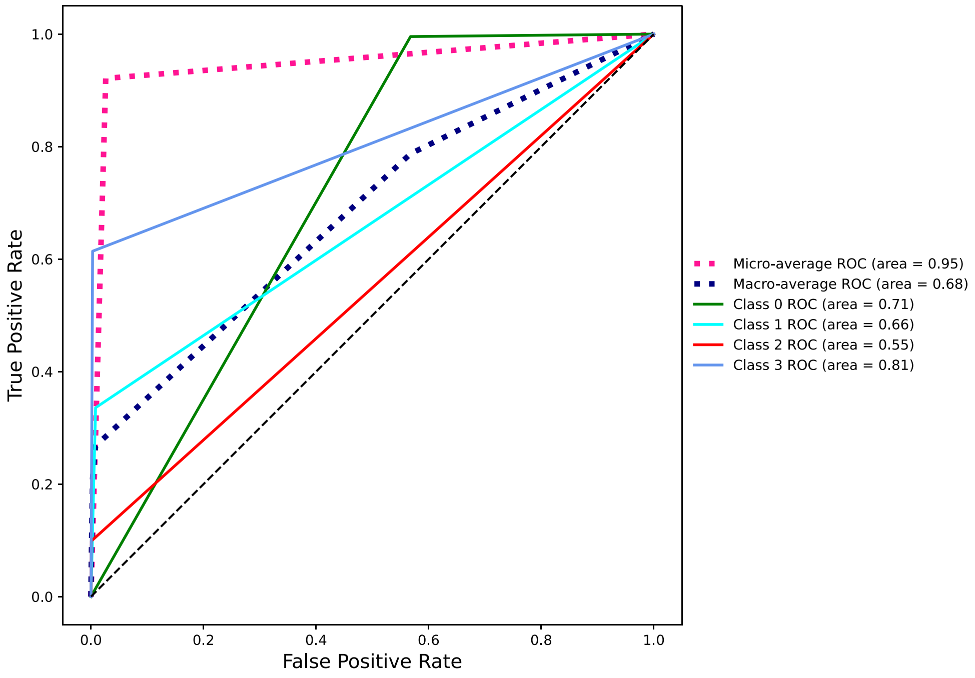

Fig. 9 ROC curves for Model M2 (CNN + BiLSTM + Attention): improved class discrimination and higher AUCs

more suitable for safety-critical manufacturing applications where missing anomalies carry a high cost. These findings validate the importance of architectural enhancements and data-centric strategies in generalizing AI models across domains with varying fault distributions. These cross-domain validation results demonstrate the framework's potential for technology transfer across different manufacturing sectors. The ability to maintain performance when applied to unfamiliar data patterns is crucial for manufacturers seeking to deploy AI-based quality control systems across multiple production lines or facilities with varying operational characteristics.

To clarify the strategic use of both datasets in our validation methodology, we highlight the complementary roles played by the Ericsson AB dataset and the public benchmark (SMD). The Ericsson dataset enables rigorous validation under authentic 5 G radio production conditions, capturing real-time signal behavior, calibration variability, and firmware-hardware interactions that characterize software-defined manufacturing. This ensures the framework's direct

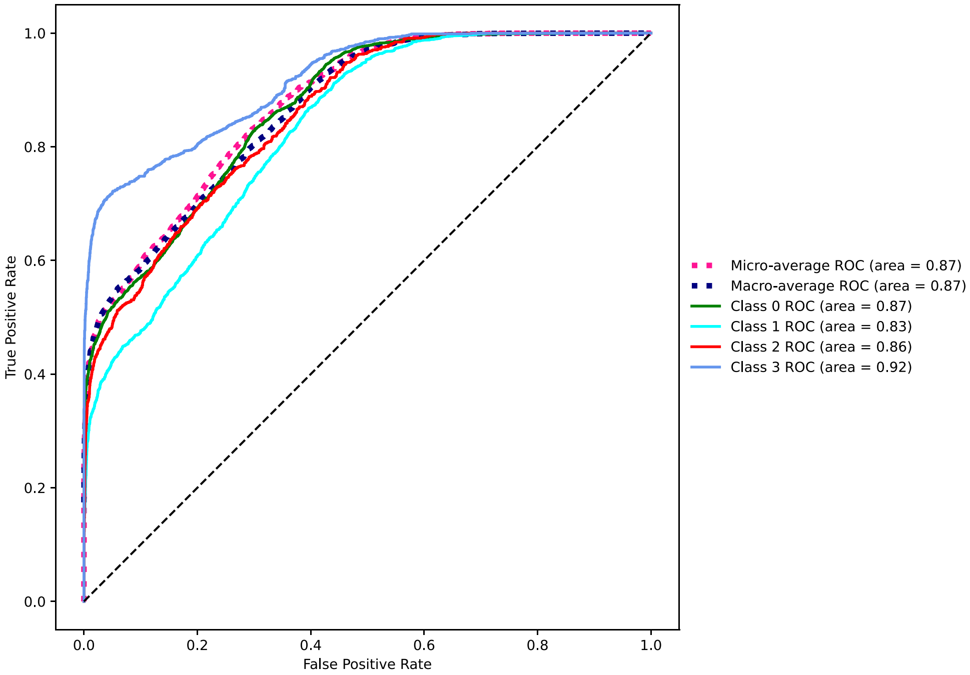

Fig. 10 ROC curves for Model M3 (CNN + BiLSTM with Oversampling): balanced and consistent detection across all classes

Table 8 Performance comparison of DTA-QC variants against traditional quality control strategies on the Kaggle server machine dataset (SMD)

Method	Accuracy	Precision	Recall	F1 score	ROC-AUC
Fixed threshold (best feature)	90.13	47.93	50.49	49.18	–
Rule-based (3-feature logic)	86.97	38.81	65.66	48.78	–
Random predictor	82.81	9.04	9.03	9.03	–
M1 (CNN Baseline)	74.01	71.53	51.50	53.70	89.73
M2 (CNN+BiLSTM+Attention)	**92.47**	**82.39**	53.53	59.48	**93.78**
M3 (CNN+BiLSTM+Oversampling)	61.66	69.72	**61.66**	**61.76**	86.84

Bold indicate the best performance across the compared models for each evaluation metric

All metrics (Accuracy, Precision, Recall, F1-Score, and ROC-AUC) are reported as percentages

operational impact, including reduced manual inspection overhead, improved throughput, and early fault detection in live production. Conversely, the SMD benchmark demonstrates cross-domain applicability by confirming DTA-QC's effectiveness on server telemetry signals with comparable statistical characteristics. This dual-dataset strategy allows us to assess both industrial robustness and generalizability, supporting reproducibility for the academic community while ensuring practical utility for high-reliability manufacturing environments.

Comparison with baseline quality control strategies

To evaluate the practical benefits of DTA-QC in an industrial context, we compare it with three baseline strategies used in traditional test systems: (i) fixed thresholding, (ii) rule-based feature logic, and (iii) a random Bernoulli predictor. As shown in Table 8, all three DTA-QC variants outperform traditional baselines in either precision or recall, while M2 provides the best trade-off in accuracy and AUC. This demonstrates that learning-based models offer measurable benefits over heuristic decision rules in complex manufacturing environments.

All model evaluations and baseline comparisons were conducted using the publicly available Server Machine Dataset (SMD) from Kaggle. This dataset provides multivariate time-series measurements commonly used for benchmarking anomaly detection techniques in industrial settings. The traditional baseline strategies used for comparison are defined as follows:

- **Fixed threshold (best feature on F1 score):** From the training set, we identify the single signal feature that yields the clearest separation between Fail and Pass labels based on F1 score. For this feature, we sweep over its 95% quantiles to determine candidate cutoffs, selecting the one that maximizes validation F1. At test time, a unit is labeled *Fail* if its value exceeds the selected threshold in the learned risk direction; otherwise, it is labeled *Pass*.
- **Rule-based (Level/Jump/Variability, OR logic):** We define three interpretable rules, each designed to capture

Table 9 Average CPU and memory consumption of models M1, M2, and M3 during inference on CPU-only hardware

Model	CPU usage (%)	Memory usage (MB)
M1 (CNN)	0.24	174.63
M2 (CNN + BiLSTM + Attention)	0.10	172.96
M3 (CNN + BiLSTM)	0.07	159.73

different types of signal anomalies. Thresholds are learned from training quantiles and selected using validation F1. A unit is labeled *Fail* if any of the following checks trigger:

  - *Level:* The feature lies in the extreme tail (too high or too low) of the training distribution.
  - *Jump:* The absolute change between adjacent signal windows is unusually large.
  - *Variability:* The within-sample standard deviation across signal windows is abnormally high.

- **Random (prior-weighted):** This feature-agnostic baseline predicts *Fail* with probability equal to the failure rate observed in the training set, and *Pass* otherwise. It serves as a lower-bound reference when no usable signal is available.

Hardware performance benchmarks

To validate the deployability of the proposed models in industrial settings, we conducted hardware performance benchmarks on a standard CPU-only workstation. All models (M1, M2, and M3) were evaluated during inference without GPU acceleration. Table 9 summarizes the average CPU and memory usage, measured across multiple runs to ensure stability and reproducibility.

The benchmark results reveal important trade-offs between model complexity, predictive performance, and computational efficiency:

- **Model M1** incurs the highest CPU load due to convolutional stack depth but maintains a low memory footprint. It is suitable for edge deployments where inference speed is prioritized.
- **Model M2**, despite being the most complex (with BiLSTM and self-attention layers), achieves superior classification accuracy with only a marginal increase in memory usage. Its moderate CPU consumption reflects the cost of attention computation but remains within deployable limits.
- **Model M3** strikes the best balance between computational efficiency and classification robustness. It achieves the lowest CPU and memory consumption while maintaining solid performance across minority classes.

Table 10 Comparison of dataset characteristics used in the cross-domain evaluation

Characteristic	Ericsson 5 G	SMD benchmark	Observation
Domain	Telecom manufacturing	IT infrastructure monitoring	Distinct operational context
Signal type	RF power, voltage, temperature	CPU, memory, network traffic	Continuous multivariate data
Temporal structure	Sequential test cases per unit	Continuous monitoring intervals	Both exhibit temporal dependence
Feature dimensionality	19 features per window	Variable per server	Comparable order of magnitude
Anomaly source	Calibration drift, signal transients	Resource overload, injected faults	Different causal mechanisms
Labeling scheme	Dynamic thresholds + SME validation	Synthetic labels provided	Distinct annotation process

All models operate within the computational constraints typical of industrial edge computing environments, supporting deployment on standard manufacturing hardware without the need for specialized accelerators. These results support its use in cost-sensitive and real-time production contexts where inference latency and hardware simplicity are critical.

Domain shift analysis and transferability assessment

The ability of a quality-control framework to maintain performance across heterogeneous data sources is a key indicator of methodological robustness. To examine the generalization capability of DTA-QC beyond the 5 G production environment, a cross-domain evaluation was performed using the proprietary Ericsson dataset and the public Server Machine Dataset (SMD) (Gusat, 2023). Both datasets share the temporal and multivariate nature required by the proposed architecture, yet they differ substantially in domain semantics, signal origin, and labeling protocol. Table 10 summarizes their principal characteristics.

Despite originating from distinct industrial contexts, both datasets exhibit three critical properties that enable meaningful transfer learning evaluation: (1) continuous-valued, multivariate time-series structure suitable for LSTM-based feature extraction, (2) temporal dependencies preserved through sequential ordering, and (3) severe class imbalance requiring specialized augmentation strategies. However, fundamental differences in physical domain (electromagnetic vs. computational), data collection methodology (per-unit testing vs. continuous monitoring), and anomaly generation mechanisms (production faults vs. synthetic injection) constitute a non-trivial domain gap that challenges model generalization. Despite these differences, the same DTA-QC configuration (comprising the LSTM autoencoder, adaptive thresholding, and CNN severity classifier) was applied to both domains without architectural or hyperparameter modification. The consistency of reconstruction-error distributions across datasets indicated that the residual-based thresholding mechanism preserved its statistical behavior under domain shift. The model achieved comparable classification performance on the SMD benchmark to that obtained on the Ericsson data, demonstrating that the learned representation was not specific to a single production modality.

The cross-domain experiments therefore validated the methodological premise of DTA-QC: its components operate on generic statistical properties of sequential signals rather than on domain-dependent features. By relying on reconstruction error dynamics rather than explicit physical attributes, the framework adapts to variations in signal scale and noise characteristics inherent to different industrial environments. Moreover, the results confirmed that the Moving-Block Bootstrap augmentation preserved minority-class consistency when the anomaly distribution changed between domains. Overall, the domain-shift analysis supported the generalizability of the proposed methodology. DTA-QC maintained stable detection thresholds and classification boundaries across datasets with distinct origins and semantics, indicating that its design principles(temporal modeling, adaptive thresholding, and hierarchical severity mapping) form a transferable and data-driven basis for industrial quality control. Future work will extend this evaluation to additional manufacturing sectors to further quantify the limits of cross-domain transferability.

Implications for industrial deployment

The empirical evidence presented in Sect. "Cross-domain validation with benchmark dataset" supports three key conclusions regarding DTA-QC's generalizability. First, the framework successfully transferred from telecommunications manufacturing to IT infrastructure monitoring without architectural modifications, indicating applicability beyond the initial 5 G deployment context. Second, the maintained ROC-AUC performance demonstrates that temporal modeling components generalize more effectively than domain-specific feature engineering approaches common in traditional quality control systems. Third, the cross-domain validation confirms that dynamic thresholding provides robustness to distributional shifts, a critical requirement for deploying AI systems across heterogeneous production environments. However, three limitations constrain immediate transferability claims. The evaluation encompassed only two industrial domains, both characterized by continuous-valued sensor data and temporal ordering. Applicability to categorical process data, non-sequential manufacturing workflows, or domains with fundamentally different anomaly semantics remains empirically unvalidated. Additionally, the SMD dataset's synthetic anomaly injection may not fully capture the complexity of organic production faults observed in real manufacturing settings. Future work should systematically evaluate DTA-QC across additional sectors, such as automotive assembly, semiconductor fabrication, or pharmaceutical quality control, to establish empirically grounded applicability boundaries.

Transfer to new domains requires three methodological steps, informed by this validation study. First, domain assessment must verify temporal structure consistency and feature dimensionality compatibility with the DTA-QC input schema (Sect. "Input signal origin and representation"). Second, threshold recalibration following Algorithm 1 should be performed on a representative validation set from the target domain, ensuring that severity boundaries reflect local error distributions. Third, incremental validation through parallel deployment (shadow mode) enables controlled risk mitigation while collecting domain-specific performance data. The successful Ericsson-to-SMD transfer provides a methodological template for such adaptations, though quantitative sample size requirements and expected performance bounds remain subjects for future empirical investigation.

Threats to validity

This section systematically addresses potential threats to validity in our industrial AI deployment study, outlining mitigation strategies employed to ensure reliable and generalizable results for manufacturing applications.

1. **Construct validity:** We assume that the selected performance metrics, ROC-AUC, Precision, Recall, F1-Score, and Accuracy, appropriately capture the effectiveness of anomaly detection and severity classification in 5 G manufacturing quality control. These metrics are widely adopted in industrial AI research and are suitable for evaluating both detection performance and class-level sensitivity in imbalanced datasets. To strengthen construct validity, we applied a multi-metric evaluation strategy that balances predictive accuracy, robustness, and interpretability. Furthermore, the operational

relevance of the selected metrics was validated in close collaboration with domain experts at Ericsson AB.

2. **Internal validity:** Threats stemming from data preprocessing, labeling inconsistencies, or suboptimal hyperparameters were mitigated through rigorous pipelines, subject matter expert (SME)-validated labeling for minority classes, ensuring domain-specific knowledge integration, and extensive tuning based on empirical performance. The stability of results across three different architectures (M1, M2, and M3) supports the internal consistency of our methodology.
3. **External validity:** To assess generalizability beyond the proprietary Ericsson dataset, we conducted cross-domain validation using the public Server Machine Dataset (SMD). The comparable performance trends observed across both domains (Sect. "Cross-domain validation with benchmark dataset") indicate that the DTA-QC framework can generalize to heterogeneous industrial settings. Further validation across diverse manufacturing sectors such as automotive assembly, pharmaceutical production, or semiconductor fabrication would strengthen evidence of broad industrial applicability.
4. **Scalability and deployment feasibility:** Scalability concerns were addressed through hardware benchmarks (Table 9), which confirmed that all models operate efficiently on standard CPU-based workstations. M1 suits time-critical manufacturing processes requiring sub-second response times, while M2 and M3 provide enhanced detection accuracy with acceptable computational overhead for batch processing applications, supporting their deployment in production-scale environments without requiring GPUs.
5. **Practicality and adaptability:** The DTA-QC framework has been successfully piloted within Ericsson's 5 G production line. Its modular design, including dynamic thresholding, hybrid labeling, and time-aware resampling, facilitates adaptation to varying test protocols, sensor types, and data distributions. These characteristics make it well-suited for real-time anomaly detection in dynamic industrial workflows.
6. **Integration constraints:** The system was designed with real-world constraints in mind. Integration with existing testing platforms was conducted iteratively and co-developed with Ericsson engineers. By relying on common AI/ML libraries and open-source tooling, we ensured compatibility and reproducibility across deployment pipelines.
7. **Manufacturing environment variability:** Industrial deployment introduces environmental factors not captured in controlled laboratory settings, including electromagnetic interference, temperature fluctuations, and varying production schedules. We addressed this through extended pilot testing in Ericsson's actual production environment, where the framework operated continuously under real manufacturing conditions, including shift changes, equipment maintenance windows, and varying production volumes.
8. **Conclusion validity:** Class imbalance, particularly the scarcity of critical failure events, represents a fundamental challenge in manufacturing anomaly detection where catastrophic failures are inherently rare but carry severe consequences. To counter this, we applied a hybrid resampling strategy based on the Moving Block Bootstrap, which preserves temporal patterns while improving minority class representation. While classification metrics improved, residual skew remains. Future research will investigate cost-sensitive learning algorithms that explicitly account for the asymmetric costs of false negatives in manufacturing contexts, where missing a critical failure significantly outweighs false alarms in terms of operational impact.

Model assumptions and limitations

The DTA-QC framework is designed for predictive quality control in high-mix, analog-intensive production settings. While effective in our test environment, its performance relies on several important assumptions and is subject to practical limitations:

- **Analog input signals only:** The current implementation processes only continuous-valued test signals (e.g., power, voltage, temperature). Boolean or discrete inputs (e.g., protocol pass/fail flags) are excluded, which may reduce fault coverage for purely digital or binary failure modes.
- **Consistent test execution order:** The system assumes that each device under test (DUT) undergoes the same sequence of test cases in a fixed order. Deviations such as reordering, skipping, or early termination of tests may impair the temporal modeling and compromise classification accuracy.
- **Minimum data quality:** Effective training requires representative, denoised, and temporally aligned input data. High noise, data dropout, or poor sensor calibration may skew the residual distribution used for threshold estimation, degrading performance.
- **Logged process events:** The framework assumes that impactful process changes (e.g., firmware updates, hardware revisions, supply chain variations) are logged and timestamped. Unlogged interventions may create distribution shifts that mimic faults and reduce interpretability.

- **Initial model tuning:** While most thresholding and labeling are automated, model hyperparameters (e.g., LSTM window size, latent dimensions) require initial tuning. The current version uses grid search; future work will incorporate automated search strategies to ease deployment.
- **Fault pattern assumptions:** DTA-QC assumes faults manifest as progressive deviations in analog signal behavior, allowing early detection via anomaly accumulation. Sudden, binary, or logic-driven faults without observable precursors may not be captured effectively.
- **Limited domain generalization:** Although validated on 5 G radio production, generalization to other domains (e.g., automotive, semiconductor, pharma) requires domain-specific preprocessing, retraining, and integration with relevant sensor modalities or digital test features.
- **Prototype status:** This study represents a prototype implementation. While promising results were obtained in controlled trials, full-scale deployment would require further validation under operational constraints (e.g., hardware integration, test station latency, fault traceability).

Discussion and future work

This study introduces and validates DTA-QC: An AI-Driven Framework for Adaptive Quality Control and Intelligent Test Optimization in 5 G Manufacturing, addressing key limitations of static test logic in high-throughput industrial environments. Developed in close collaboration with Ericsson AB, DTA-QC has been evaluated using real-world production data from 5 G radio manufacturing but has not yet been deployed in live operations. The prototype demonstrates how advanced AI techniques can augment traditional quality control by offering intelligent, data-driven methods that are both accurate and interpretable. The framework integrates a bidirectional LSTM autoencoder for unsupervised feature extraction, Ridge regression for dynamic thresholding, and a 1D Convolutional Neural Network (CNN) for real-time severity classification. Together with a hybrid resampling strategy (combining block bootstrapping and residual perturbation) and interpretable class structuring, DTA-QC significantly improves anomaly detection accuracy, automation, and robustness compared to traditional rule-based methods. Empirical results confirm that architectural enhancements yield measurable gains: Model M2 (CNN + BiLSTM + Attention) delivers the highest accuracy and AUC, while Model M3 (CNN + BiLSTM + oversampling) achieves superior recall and F1-Score, particularly critical under imbalanced real-world conditions. Cross-domain validation using the Server Machine Dataset (SMD) demonstrates the generalizability of our approach across system types and anomaly patterns. Moreover, hardware benchmarking confirms that all models operate efficiently on standard CPU-based systems, supporting immediate industrial deployment without specialized acceleration. Key contributions of this work include:

- *Dynamic, adaptive thresholds:* ridge regression-based adaptive thresholding responds dynamically to production variability and equipment aging, eliminating the need for frequent manual threshold recalibration that typically requires significant engineering effort in traditional manufacturing quality systems.
- *Reduced manual labeling:* LSTM autoencoder-assisted labeling, combined with distance-based propagation and SME review, minimizes annotation burden while preserving label quality.
- *Operational interpretability:* the intuitive four-tier severity classification system (Normal, Warning, Worse, Stop) aligns with established manufacturing quality protocols, enabling shop floor operators to make immediate, informed decisions without requiring specialized AI expertise. In industrial applications, the outputs of DTA-QC directly inform actionable quality improvement and operational decisions. For example, a severity classification of “Worse” or “Stop” can trigger automatic re-routing of the unit for re-inspection, re-calibration, or component replacement, reducing the risk of latent faults reaching the field. In a specific deployment case at Ericsson AB, the model’s prediction of early-stage anomalies enabled the decision to halt testing sequences earlier than scheduled, thereby saving test time and improving resource utilization. Additionally, persistent “Warning” classifications over time for a specific module (e.g., the RF front-end) prompted a targeted yield analysis and subsequent adjustments in supplier calibration standards. These use cases illustrate how DTA-QC’s granular output enhances production agility and supports continuous improvement in quality engineering workflows.
- *Temporal traceability:* structured windowing and autoencoder-derived features support long-term anomaly trend detection and predictive analytics.
- *Scalable architecture:* all model variants maintain low CPU and memory usage (Table 9), validating DTA-QC’s deployability in real-time test stations and resource-constrained environments.
- *Cross-domain generalizability:* although DTA-QC is validated on a 5 G radio testbed, its modular architecture is transferable to other intelligent manufacturing domains. For instance, applications in semiconductor wafer inspection, automotive assembly lines, or

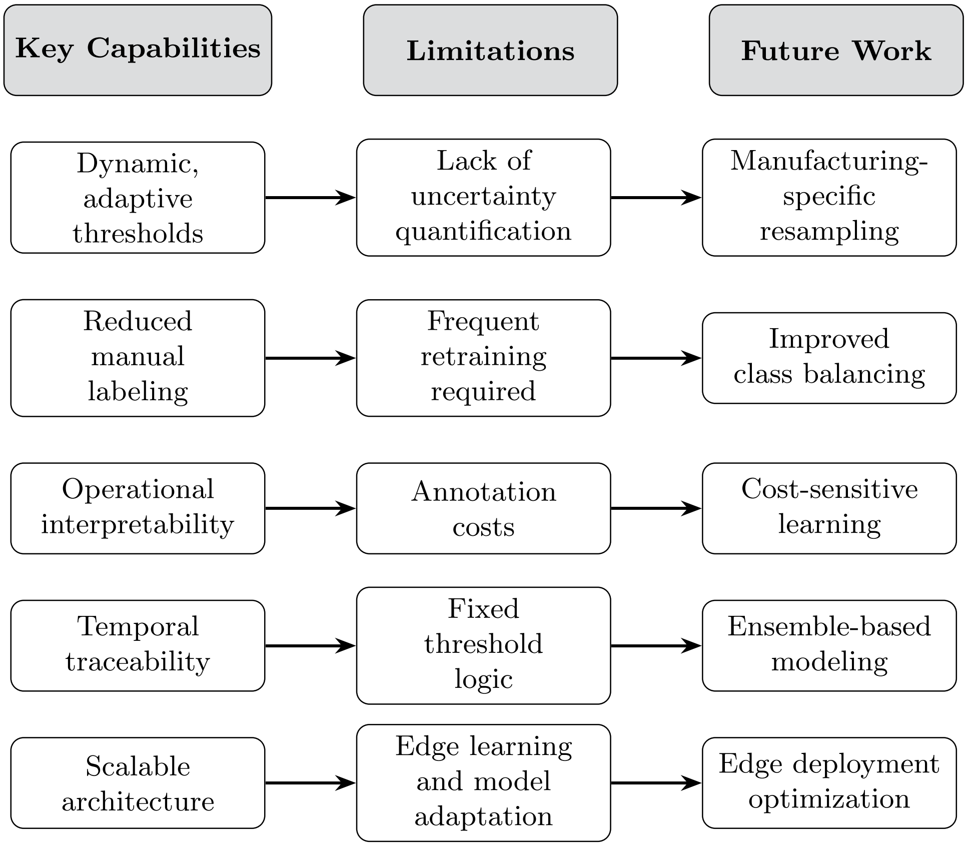
Key Capabilities
Limitations
Future Work
Dynamic, adaptive thresholds
Lack of uncertainty quantification
Manufacturing-specific resampling
Reduced manual labeling
Frequent retraining required
Improved class balancing
Operational interpretability
Annotation costs
Cost-sensitive learning
Temporal traceability
Fixed threshold logic
Ensemble-based modeling
Scalable architecture
Edge learning and model adaptation
Edge deployment optimization

Fig. 11 Visual summary connecting DTA-QC's key capabilities, current limitations, and proposed future work.

pharmaceutical quality control also involve structured time-series testing with limited labels and high imbalance. Our use of LSTM autoencoders, dynamic thresholds, and bootstrap augmentation is not task-specific but reflects general principles in industrial anomaly detection. Moreover, the system's compact, CPU-efficient models make it suitable for edge deployment, which is critical for diverse production environments. Future work will assess the transferability of DTA-QC to other industrial monitoring tasks, including predictive maintenance scenarios such as vibration-based fault detection and condition monitoring systems in rotating machinery or thermal equipment.

Industrial deployment reveals persistent challenges inherent to manufacturing quality control applications. The fundamental scarcity of critical failure events in well-managed production environments creates class imbalance that affects model sensitivity to rare but high-impact quality issues, as highlighted by differences in macro and micro-average AUC (Figs. 8, 9, 10). This challenge reflects the broader paradox in manufacturing AI: the most important anomalies to detect are precisely those that occur least frequently in successful production operations. While Model M3 offered better class-wise consistency, future enhancements are necessary to fully address this issue.

Future work will focus on the following areas:

- **Manufacturing-specific resampling strategies:** Develop synthetic failure pattern generation methods based on physics-informed models and historical maintenance data to ensure augmented samples reflect realistic degradation behaviors.
- **Improved class balancing:** Apply advanced resampling techniques such as ADASYN, SMOTE-ENN, and Tomek Links to better represent minority classes and capture boundary cases without overfitting.
- **Cost-sensitive learning:** Implement dynamic loss reweighting and margin-aware objective functions to improve the model's sensitivity to underrepresented fault classes.
- **Ensemble-based modeling:** Explore ensemble techniques including balanced random forests, bagging, and cost-sensitive boosting to improve classification stability and performance across shifting data distributions.
- **Active learning integration:** Introduce human-in-the-loop feedback loops that prioritize uncertain or ambiguous samples for expert labeling, reducing annotation costs while improving model generalization.
- **Edge deployment optimization:** Adapt DTA-QC for deployment on industrial edge hardware such as programmable logic controllers (PLCs) and IoT gateways,

incorporating federated learning to support distributed model training while preserving data privacy and minimizing bandwidth usage.

- **Online learning and model adaptation:** A promising direction for extending DTA-QC is the incorporation of continuous learning mechanisms, enabling real-time model adaptation without interrupting production workflows. This enhancement would involve implementing incremental learning algorithms that update model parameters on-the-fly based on newly ingested production data. To maintain deployment safety, this would be coupled with drift detection to trigger model recalibration only when significant changes are observed in the data distribution. Moreover, federated learning strategies may be explored to support collaborative improvements across multiple production sites while preserving data confidentiality. By integrating such capabilities, DTA-QC can evolve into a fully adaptive framework that aligns more closely with the needs of intelligent, autonomous manufacturing systems.

Figure 11 presents a visual summary that systematically connects the core contributions of DTA-QC with known limitations and the corresponding future work directions. The leftmost column enumerates the key capabilities demonstrated by our framework, including dynamic thresholding, interpretability, and scalable architecture. These are linked to the middle column, which outlines current practical limitations such as lack of uncertainty quantification, annotation costs, and fixed threshold logic. Each limitation is then associated with an actionable future research direction (rightmost column), such as ensemble modeling, class balancing, or edge deployment optimization. This structured view reinforces how each limitation has a planned mitigation strategy, demonstrating a coherent path toward continuous enhancement of the DTA-QC system.

Conclusions

This study presents and validates DTA-QC: an AI-driven framework for adaptive quality control and intelligent test optimization in RBS manufacturing. Designed for deployment in high-throughput, real-time production environments, DTA-QC addresses key limitations of fixed thresholding by integrating dynamic decision logic, anomaly-aware classification, and efficient model architectures. Through an industrial case study at Ericsson AB and the formulation of four research questions, we demonstrate that: (**RQ1**) fixed rule-based thresholds can be successfully replaced with LSTM autoencoder-based dynamic thresholding that adapts to evolving production conditions; (**RQ2**) the framework significantly improves anomaly detection accuracy (ROC-AUC up to 94%) while reducing manual inspection effort; (**RQ3**) the introduced four-tier severity classification system provides interpretable, real-time decision support; and (**RQ4**) the approach generalizes beyond proprietary data, as validated on a public benchmark. DTA-QC enhances test efficiency and fault sensitivity while maintaining computational feasibility, operating on standard CPU hardware without the need for specialized infrastructure or large annotated datasets. Its modular design supports deployment across varied manufacturing contexts, from telecom to broader industrial applications. By embedding machine learning directly into production testing workflows, DTA-QC contributes to the advancement of Industry 4.0. It demonstrates how adaptive, AI-powered quality control can strengthen operational resilience, optimize test execution, and bridge the gap between academic research and scalable industrial deployment. This work lays a strong foundation for next-generation diagnostic pipelines in intelligent, software-defined manufacturing systems.

Acknowledgements This work was supported by VINNOVA grant 2023-00244 (D-RODS project). The authors thank Ericsson AB for access to 5 G production facilities and collaborative support for validating the DTA-QC framework.

Funding Open access funding provided by Mälardalen University.

Declarations

Conflict of interest The authors declare that there are no conflict of interest.

Ethical approval This research was conducted in accordance with the ethical standards of the Journal of Intelligent Manufacturing and the institutional guidelines of Ericsson AB. The study did not involve human participants, animals, or personal data, and therefore did not require approval by an ethics committee or institutional review board.

References

Agrawal, V. D., Seth, S. C., & Agrawal, P. (2003). Fault coverage requirement in production testing of LSI circuits. *IEEE Journal of Solid-State Circuits, 17*(1), 57–61.

Alhussein, M., Aurangzeb, K., & Haider, S. I. (2020). Hybrid CNN-LSTM model for short-term individual household load forecasting. *IEEE Access, 8*, 180544–180557.

Alkahtani, M., Choudhary, A., De, S., & Kiran, M. H. (2024). Deep learning-based anomaly detection using one-dimensional convolutional neural networks (1d CNN) in machine centers (MCT) and computer numerical control (CNC) machines. *PeerJ Computer Science, 10*, 2389.

Barr, E. T., Harman, M., McMinn, P., Shahbaz, M., & Yoo, S. (2015). The oracle problem in software testing: A survey. *IEEE Transactions on Software Engineering, 41*(5), 507–525.

Benkedjouh, T., Medjaher, K., Zerhouni, N., & Rechak, S. (2018). Remaining useful life estimation based on nonlinear feature reduction and support vector regression. *Engineering Applications of Artificial Intelligence, 26*, 1751–1760.

Bertasius, G., Wang, H., Torresani, L. (2021). Is space-time attention all you need for video understanding? In *ICML* (vol. 2, pp. 4).

Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys, 54*(3), 1–33.

Casa, A., & Menardi, G. (2022). Nonparametric semi-supervised classification with application to signal detection in high energy physics. *Statistical Methods & Applications, 31*(3), 531–550.

Choi, K., Yi, J., Park, C., & Yoon, S. (2021). Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. *IEEE Access, 9*, 120043–120065.

Chung, J., Shen, B., & Kong, Z. J. (2023). Anomaly detection in additive manufacturing processes using supervised classification with imbalanced sensor data based on generative adversarial network. *Journal of Intelligent Manufacturing, 35*, 1–20.

Dani, M., Jollois, F., Nadif, M., & Freixo, C. (2015). Adaptive threshold for anomaly detection using time series segmentation. In S. Arik, T. Huang, W. K. Lai, & Q. Liu (Eds.), *Neural Information Processing* (pp. 82–89). Cham: Springer.

Deshpande, A., Minai, A. A., & Kumar, M. (2023). Machine learning techniques in additive manufacturing: A state of the art review on design, processes and production control. *Journal of Intelligent Manufacturing, 34*(1), 21–55.

Engström, O., Tahvili, S., Muhammad, A., Yaghoubi, F., & Pellaco, L. (2020). Performance analysis of deep anomaly detection algorithms for commercial microwave link attenuation. In *The 2020 international conference on advanced computer science and information systems* (pp. 47–52).

Escobar, C. A., McGovern, M. E., & Morales-Menendez, R. (2021). Quality 4.0: A review of big data challenges in manufacturing. *Journal of Intelligent Manufacturing, 32*(8), 2319–2334.

Felderer, M., Enoiu, E. P., & Tahvili, S. (2023). Artificial intelligence techniques in system testing. In J. Raúl Romero & F. C. Inmaculada Medina-Bulo (Eds.), *Optimising the software development process with artificial intelligence* (pp. 221–240). Amsterdam: Springer.

Gao, Z., Wang, Y., Chen, J., Xing, J., Patel, S., Liu, X., & Shi, Y. (2023). Mmtsa: Multi-modal temporal segment attention network for efficient human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 7*(3), 1–26.

Greinert, D., Müller, M., & Strodthoff, N. (2022). Convolutional-based encoder-decoder network for time series anomaly detection during the milling of 16MnCr5. *Data, 7*(12), 175.

Gusat, M. (2023). SMD_OnmiAD: Server machine dataset for anomaly detection. Retrieved May 12, 2025, from https://www.kaggle.com/datasets/mgusat/smd-onmiad

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Hundman, K., Constantinou, V., Laporte, C., Colwell, I., & Soderstrom, T. (2018). Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 387–395).

Kaner, C., Falk, J., & Nguyen, H. Q. (2002). Automated software testing. In *Software testing and analysis: Process, principles, and techniques* (pp. 173–206).

Karkaria, V., Tsai, Y.-K., Chen, Y.-P., & Chen, W. (2025). An optimization-centric review on integrating artificial intelligence and digital twin technologies in manufacturing. *Engineering Optimization, 57*(1), 161–207.

Kashiparekh, K., Narwariya, J., Malhotra, P., Vig, L., & Shroff, G. (2019). Convtimenet: A pre-trained deep convolutional neural network for time series classification. In *2019 international joint conference on neural networks (IJCNN)* (pp 1–8). IEEE.

Kim, H., & Lee, J. (2024). Defect detection model using time series data augmentation and transformation. *Computers, Materials & Continua, 78*(2), 2389–2406.

Kumari, S., Kumar, D., & Mittal, M. (2024). A comprehensive investigation of anomaly detection methods in deep learning and machine learning: 2019–2023. IET Information Security.

Landin, C., Liu, J., & Tahvili, S. (2021). A dynamic threshold based approach for detecting the test limits. In *The 16th international conference on software engineering advances* (pp. 81).

Landin, C., Liu, J., Katsarou, K., & Tahvili, S. (2023) Time series anomaly detection using convolutional neural networks in the manufacturing process of ran. In *2023 IEEE international conference on artificial intelligence testing (AITest)* (pp. 90–98). IEEE.

Landin, C., Zhao, X., Längkvist, M., & Loutfi, A. (2023). An intelligent monitoring algorithm to detect dependencies between test cases in the manual integration process. In *2023 IEEE international conference on software testing, verification and validation workshops (ICSTW)* (pp 353–360).

Liso, A., Cardellicchio, A., Patruno, C., Nitti, M., Ardino, P., Stella, E., & Renò, V. (2024). A review of deep learning based anomaly detection strategies in industry 4.0 focused on application fields, sensing equipment and algorithms. *IEEE Access, 12*, 93911.

Liso, A., Cardellicchio, A., Patruno, C., Nitti, M., Ardino, P., Stella, E., & Renò, V. (2024). A review of deep learning-based anomaly detection strategies in industry 4.0 focused on application fields, sensing equipment, and algorithms. *IEEE Access, 12*, 93911–93923.

Liu, J., & Tahvili, S. (2025). AI-driven approach to industrial testing. Retrieved January 14 2025, from https://github.com/ljie-16/AI-Driven-Approach-to-Industrial-Testing

Liu, J., & Tahvili, S. (2025). An AI-driven approach to industrial testing: From time series to supervised classification through dynamic thresholds. Retrieved January 14 2025.

Liu, C., Kong, Z. J., Babu, S., & Seifi, M. (2022). Process monitoring and machine learning for defect detection in laser-based metal additive manufacturing. *Journal of Intelligent Manufacturing, 33*(4), 1033–1050.

Li, Z., Zhang, Z., Shi, J., Wu, D., Tian, S., Liu, W., Gokuldoss, P. K., Li, S., & Wang, J. (2024). Machine learning-assisted in-situ adaptive strategies for the control of defects and anomalies in metal additive manufacturing. *Additive Manufacturing, 81*, Article 103998.

Mattera, G., Mattera, R., Vespoli, S., & Salatiello, E. (2025). Anomaly detection in manufacturing systems with temporal networks and unsupervised machine learning. *Computers & Industrial Engineering, 203*, Article 111023.

Milor, L., & Sangiovanni-Vincentelli, A. L. (2002). Minimizing production test time to detect faults in analog circuits. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 13*(6), 796–813.

Orabi, M., Tran, K. P., Egger, P., & Thomassey, S. (2024). Anomaly detection in smart manufacturing: An adaptive adversarial transformer-based model. *Journal of Manufacturing Systems, 77*, 591–611.

Orabi, M., Tran, K. P., Egger, P., & Thomassey, S. (2024). Anomaly detection in smart manufacturing: An adaptive adversarial transformer-based model. *Journal of Manufacturing Systems, 77*, 591–611.

Papavasileiou, A., Michalos, G., & Makris, S. (2025). Quality control in manufacturing–review and challenges on robotic applications. *International Journal of Computer Integrated Manufacturing, 38*(1), 79–115.

Scarrott, C., & MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *Revstat Statistical Journal, 10*, 33–60.

Siffer, A., Fouque, P. -A., Termier, A., & Largouet, C. (2017). Anomaly detection in streams with extreme value theory. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (KDD '17) (pp. 1067–1075). Halifax, NS, Canada: Association for Computing Machinery.

Stojanovic, L., Dinic, M., Stojanovic, N., & Stojadinovic, A. (2016). Big-data-driven anomaly detection in industry (4.0): An approach and a case study. In *2016 IEEE international conference on big data (big data)* (pp. 1647–1652).

Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., & Pei, D. (2019). Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. KDD '19* (pp. 2828–2837). Association for Computing Machinery, USA.

Tahvili, S. (2018). *Multi-criteria optimization of system integration testing*. Västerås: Malardalen University.

Tahvili, S., & Hatvani, L. (2022). *Artificial intelligence methods for optimization of the software testing process with practical examples and exercises*. Uncertainty: Computational techniques, and decision intelligence.Amsterdam: Elsevier.

Tonini, S., Vandin, A., Chiaromonte, F., Licari, D., & Barsacchi, F. (2024). Accurate and fast anomaly detection in industrial processes and IoT environments. arXiv preprint arXiv:2404.17925

Tran, D. H., Nguyen, V. L., Nguyen, H., & Jang, Y. M. (2022). Self-supervised learning for time-series anomaly detection in industrial internet of things. *Electronics, 11*(14), 104.

Wang, X., Liu, J., & Zhang, Y. (2021). Managing product-inherent constraints with artificial intelligence: Production control optimization under industry 4.0. *Journal of Intelligent Manufacturing, 32*(6), 1503–1517.

Weiss, S. M., Dhurandhar, A., & Baseman, R. J. (2013). Improving quality control by early prediction of manufacturing outcomes. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1258–1266).

Weiss, S., Dhurandhar, A., Baseman, R., White, B., Logan, R., Winslow, J., & Poindexter, D. (2016). Continuous prediction of manufacturing performance throughout the production lifecycle. *Journal of Intelligent Manufacturing, 27*(4), 751–763.

Wen, T., & Keyes, R. (2019). Time series anomaly detection using convolutional neural networks and transfer learning. ArXiv abs/1905.13628

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2023). Transformers in time series: A survey. In *IJCAI* (pp. 6778–6786).

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., & Long, M. (2022). Timesnet: Temporal 2D-variation modeling for general time series analysis. arXiv preprint arXiv:2210.02186

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 32*(1), 4–24.

Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems, 34*, 22419–22430.

Yan, P., Abdulkadir, A., Luley, P.-P., Rosenthal, M., Schatte, G. A., Grewe, B. F., & Stadelmann, T. (2023). A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: Methods, applications, and directions. *IEEE Access, 12*, 3768–3789.

Zhang, D., Hao, X., Wang, D., Qin, C., Zhao, B., Liang, L., & Liu, W. (2023). An efficient lightweight convolutional neural network for industrial surface defect detection. *Artificial Intelligence Review, 56*(9), 10651–10677.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022). Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning* (pp. 27268–27286).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.