**Journal of Information Systems Engineering and Management**
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

# Systematic Literature Review of Continuous Validation and Improvement Methods for AI Systems

Bavani Raja Pandian [1]*, Kamsiah Mohamed [2]
[1] *Tunku Abdul Rahman University of Management and Technology (TAR UMT), Jalan Genting Kelang, Setapak, 53300 Kuala Lumpur, Malaysia. Email: bavanirp@tarc.edu.my, Orchid Id: 0009-0000-9870-1988*
[2] *Universiti Selangor, Jalan Timur Tambahan, 45600 Batang Berjuntai, Selangor, Malaysia. Email: kamsh@unisel.edu.my*
**Corresponding Author: Bavani Raja Pandian*

**ARTICLE INFO**

Received:01 Sept 2025
Revised:07 Oct 2025
Accepted:18 Oct 2025

**ABSTRACT**

This systematic literature review aims to systematically analyze current methods of continuous validation and improvement in AI systems. It investigates techniques, application domains, scope of implementation, and real-world challenges to identify prevailing trends, limitations, and opportunities for enhancing adaptive, reliable, and ethically responsible AI deployment This systematic review employed a structured search across IEEE Xplore, Scopus, Web of Science, and ACM Digital Library (2018-2025), using defined keywords. Studies were screened through title, abstract, and full-text review, applying inclusion/exclusion criteria. Quality was assessed using context, design, validity, rigor, and relevance dimensions. A total of 51 studies were analyzed. Cross-validation (25.49%) and online retraining (23.53%) were the most used validation methods. Improvement efforts centered on iterative model refinement (21.57%) and integration with feedback and workflows (15.69% each). Most studies focused on smart manufacturing and robotics (each 27.45%), with healthcare (13.73%) and environmental systems (11.76%) trailing behind. Industry-focused deployments dominated (19.61%), while only 9.80% addressed cross-regional implementations. Key challenges included model drift (13.73%), generalizability issues (11.76%), and ethical concerns (11.76%). Real-time feedback mechanisms, regulatory alignment, and interpretability remained under-addressed, signaling critical gaps in sustainable and trustworthy AI development. While technical validation and improvement strategies are maturing, gaps persist in real-world adaptability, ethical integration, and socio-technical feedback loops. Bridging these gaps will require collaborative, context-aware, and regulatory-informed AI systems capable of maintaining performance and trust across diverse, evolving environments.

**Keywords:** Continuous Validation, Continuous Improvement, AI Systems, Model Drift, Real-World Deployment.

## INTRODUCTION

One of the most transformational and revolutionary technologies of our time is Artificial Intelligence (AI). The advent of AI technology has transformed the role of robots from mere tools to active participants in solving problems, driving innovation, and making decisions, owing to their ability to emulate human intelligence and decision-making processes. The development of AI has been accelerated by breakthroughs in machine learning, deep learning, neural networks, and natural language processing. Consequently, AI now permeates almost every aspect of human life, from the personal electronic devices which we now carry to the sophisticated systems which control industries and economies [1]. Over the last few years, AI systems have been experiencing rapid development,

**Journal of Information Systems Engineering and Management**
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

which has resulted in their integration in a variety of industries. These include self-driving cars, diagnostics in healthcare, natural language processing, and finance. Given this pervasive impact, the dependability and trustworthiness of AI systems has become a critical concern. AI systems are designed to tackle sophisticated problems, and as such, they must be able to perform correct, ethical, and effective functioning. Validation processes are critical in confirming that AI models perform as expected and protect against errors and biases [2].

AI-based systems are commonly constructed using various Machine Learning (ML) approaches, which depend on vast amounts of data to train models. With time, the models are refined based on new information and numerous factors. Consequently, performance or behavioral drifting may occur in AI systems due to the data they are exposed to, the environment they are in, or changes to the internal parameters. This phenomenon known as "model drift" highlights the need for systematic validation and continuous iteration to sustain the AI systems' efficiency. It is for these reasons that there is an accelerated focus on the continuous validation and enhancement of AI models [3]. Through continuous validation, an iterative feedback loop is established, which ensures that the AI system is capable of retaining the desired performance metrics during the course of its life-cycle. Such systems are geared towards the dynamic monitoring, assessment, and iteration of AI systems to combat the degradation of performance due to model drift and shifting data patterns. Unlike traditional validation approaches, which occur during the development phase or post-deployment, continuous validation allows for real-time adaptability to shifting contexts [4,5].

The expansion of AI technologies into sensitive sectors like healthcare, finance, and criminal justice amplifies the need for effective, constant validation processes. These systems can lead to serious consequences such as misdiagnoses, unjust legal determinations, financial errors, and system malfunctions. The rising need for precision and real-time adaptability in AI systems is what drives the need for new validation techniques that incorporate dynamic feedback systems to allow models self-improvement and adjustment beyond static testing [6,7]. The drive for this systematic literature review is to analyze and amalgamate the different methods for active validation and enhancement of AI systems. This review focuses on integrated studies, frameworks, and techniques to showcase the most prevalent gaps in research alongside the most effective solutions, challenges, and emerging trends within AI validation. As with any high-stakes technology, the importance of AI validation is paramount. AI validation is the mission-critical process of confirming that an AI system employs the expected processes and operations with the frameworks of compliance to the determined benchmarks of precision, bias, security, and resilience. Approximately, every AI model in the lifecycle requires incessant validation at all stages to ensure the systems are functioning to standards. Appropriate validation can avert catastrophic failures, improve system dependability, and build trust in AI technologies [8].

Validation, in the case of AI, refers to checking the model against a number of scenarios, datasets, and metrics to ascertain whether the model can effectively generalize to new data. Traditionally, validation has placed greater emphasis on offline evaluation, which refers to the model validation using the historical data prior to the model being put to use. This technique has its drawbacks, however, for the systems that work in ever changing data and condition contexts. Offline validation fails to consider the instantaneous shifts that happen post-deployment, such as user behavior, environmental interactions, or other systems interdependencies, and interactions[9]. Given the rate at which AI technologies are developing, there is an urgent interdisciplinary demand for continuous validation frameworks that ensure an AI system's integrity, reliability, and ethical accountability in real-world interactions. The models of autonomous vehicles, for instance, need to validate them against necessity of adapting to ever-changing and unpredictable road, weather, and traffic conditions. AI in healthcare also requires continuous evaluation of model predictions in the context of new data and shifting trends in medicine,

**Journal of Information Systems Engineering and Management**
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

demography, or disease patterns. The lack of continuous assessment may cause AI systems to become obsolete, which can result in risks, or in the worst-case scenarios, harm, in the vital decisions that they make [10,11].

These challenges can be solved using continuous methods validation, which incorporates the real-time monitoring, testing, and adjustment of AI systems. With these systems, it is possible to intervene early with the corrections required to prevent failures and ensure optimal outcomes. Moreover, AI systems can be adequately trained to learn and adjust to new data as it is continuously validated, meaning they can be effective in novel situations [12]. Aside from system performance, continuous validation is important in managing the dangers of bias and fairness in AI. It is possible to imbalance social justice issues in trained AI systems since they rely on datasets that are often diverse, which can contain latent biases forged from social inequality. Without active intervention models, these systems will unfairly and unjustly discriminate algorithmically. Through continuous validation, biased behavior can be monitored and discovered, thus allowing fairness in AI and enabling the “right to biased behavior” systems to be set in place [13]. Continuous validation is equally important to the ethical side of fairness, as AI systems are often employed in areas with a critical need for public safety. Sectors such as healthcare, finance, and defense rely on AI systems to deliver correct predictions all the time, as even slight deviations from the truth can have catastrophic repercussions. For instance, an AI system applied in medical imaging might incorrectly classify a tumor, resulting in a misdiagnosis and a lapse in care that could severely compromise the patient's health. Continuous validation endeavors to identify and rectify such errors as they occur, ensuring that the AI systems in use will repeatedly and accurately provide safe and dependable results [14]. Trust and acceptance of AI technologies are maintained through public validation. AI systems operate within a framework of increasingly critical societal expectations, therefore, their accountability and transparency is of utmost importance. Validation is useful in confirming AI systems meet performance benchmarks. AI developers that undertake validation protocols demonstrate a willing commitment to public acceptance, trust, and the adoption of Artificial Intelligence technologies, thereby enhancing public confidence across sectors [15]. Validation in the case of AI systems is of critical concern as trust, ethics, safety, and fairness are non-negotiable in the responsible governance of these technologies. With the increasing pervasiveness of AI in various sectors, the need for extensive and continuous validation procedures will increase. This literature review will provide the scope of validation procedures to support researchers, practitioners, and policy makers in their work and ensure AI systems are trustworthy across all applications.

## METHODS AND METHODOLOGY

**Research Objective:**

- To assess the categories and prevalence of continuous validation practices and their application across different AI domains.
- To assess the application and incorporation of the methods of adaptive learning, automation, feedback loops, and self-improvement systems.
- To examine the contextual scope, sectoral focus, and geographical distribution of AI implementations in real-world settings.
- To identify key limitations, ethical concerns, and technical gaps that hinder the sustainable and trustworthy deployment of AI systems

**Research Questions:**

- What continuous validation methods are most commonly used in AI systems, and how do they vary across sectors?

- How are continuous improvement strategies such as feedback loops, AutoML, or manual refinement being applied to maintain or enhance AI performance?
- In which domains and regions are continuous validation and improvement practices most frequently implemented, and where are the major gaps?
- What are the critical barriers (e.g., drift, bias, legal concerns) affecting long-term AI system reliability and scalability, as reported in existing studies?

**Search Strategy:**

The search strategy for this systematic literature review (SLR) on continuous validation and improvement methods for AI systems involved a comprehensive search across multiple academic databases, including IEEE Xplore, Scopus, Web of Science (WOS), and ACM Digital Library. The search was conducted using a combination of keywords such as *(Artificial Intelligence OR Machine Learning OR "autonomous system" OR "Deep Learning" OR "self-adaptive system" OR "autonomous vehicle" OR "self-driving car" OR "autonomous car" OR "self-driving vehicle" OR robot OR "intelligent system") AND (validation OR verification OR testing OR "continuous validation" OR "model monitoring" OR "performance evaluation" OR "v-model") AND (user OR customer OR industrial OR industry OR "Healthcare AI").* The review focused on studies published between 2018 and 2025 to ensure the inclusion of recent advancements in the field. The search aimed to capture a wide range of literature, covering theoretical frameworks to empirical investigations, focusing on those relating to application-based evaluation, performance assessment, and adaptive learning frameworks. After deduplication, relevant papers were filtered based on their titles and abstracts, after which inclusion and exclusion criteria were applied to identify papers that emphasized core themes of validation and continuous improvement of AI systems.

**Inclusion and Exclusion Criteria:**

**i). Inclusion:**

This systematic literature review (SLR) incorporated literature spanning from 2018 to 2025 that centered on the continuous validation and improvement of frameworks pertaining to artificial intelligence (AI) systems. Only peer-reviewed periodicals, conference proceedings, white papers, and research reports were included. The review focuses on primary literature describing performance evaluation, validation, and techniques aimed at the continuous improvement of AI models in diverse areas like natural language processing, computer vision, robotics, and reinforcement learning. Furthermore, the studies must document implementation frameworks that substantially improve the scalability, reliability, and robustness of AI systems operating in dynamic environments, especially those based on supervised, unsupervised, and reinforcement learning models.

**ii). Exclusion:**

The exclusion criteria were based on whether the studies focused on the continuous improvement processes of the AI systems, specifically on the practical application of the concepts, as well as on the systems' real-world validation and improvement application. Additionally, publications preceding the year 2018 were irrelevant to modern practices and did not meet the criterion. Papers that discussed autonomous AI systems or specialized within narrow fields that lacked generalizable applications were disregarded. Moreover, the review excluded studies that focused solely on static validation or improvement methodologies, controlled testing environments, and those written in languages other than English to ensure cohesiveness and comprehensibility.

**Journal of Information Systems Engineering and Management**
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

**Study Selection and Screening:**

The study selection and screening process for this SLR is presented in Figure 1. The search was conducted across four major academic databases: Web of Science (WOS) (155 records), Scopus (7,027 records), IEEE Xplore (67,194 records), and ACM Digital Library (150 records), with a focus on publications from 2018 to 2025, resulting in a total of 57,294 records. Duplicate records (96) were removed, leaving 57,294 records for further screening. During the screening phase, 15,046 records were evaluated based on titles and abstracts. Following the application of exclusion criteria, 13,320 records were discarded for being irrelevant, 1,726 records were excluded for focusing solely on AI and autonomous systems, and 174 records were removed for discussing only AI system testing. After assessing eligibility, 68 records were deemed suitable for further evaluation, ultimately leading to the inclusion of 51 [16–66] studies in the review, all of which contribute valuable insights into the continuous validation and improvement methods for AI systems.

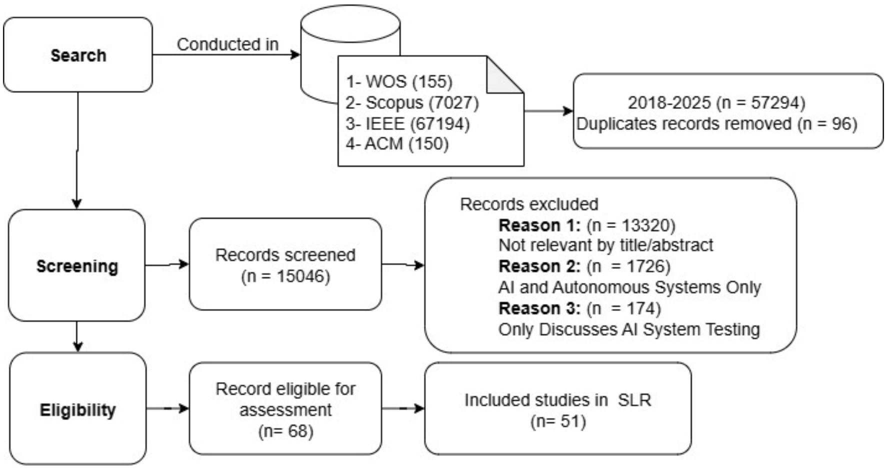


**Figure 1.** Search and selection process

**Data Extraction and Synthesis:**

A structured framework was applied to extract key attributes from each included study, including Author(s), Title, Year of publication, continuous validation techniques, improvement strategies, AI application domains, implementation scope, and reported challenges. This information was systematically organized to facilitate cross-study comparison and thematic synthesis. The approach enabled the identification of common practices, emerging trends, and research gaps in how AI systems are continuously validated and improved across various real-world contexts.

**Quality Assessment:**

To assess the quality of the final papers, we adopted the rigour and relevance framework proposed by Ivarsson and Gorschek (2011) [67], which evaluates research quality along two orthogonal dimensions. Relevance refers to the realism of the study setting, such as the use of industrial contexts, real-world data, and practitioner involvement. Instead of applying relevance strictly as defined in the original framework, we assessed it based on contextual realism observed during the study selection process, which closely aligns with the framework's intent. As relevance was implicitly considered during inclusion, only papers with moderate to high relevance (scores 2-4 on a 4-point scale) were retained. Rigour was assessed by summing three core concerns: context description, study design, and validity discussion, each rated on a 0-1 scale, yielding a maximum score of 3. Among the 51 included studies, 11 (21.6%) were categorized as high rigour (rigour score ≥ 2.0), and 24 (47.1%) exhibited high relevance (relevance score ≥ 4.0). Only one study Kim et al. 2023 [63] achieved the maximum rigour score of 3.0.

**Journal of Information Systems Engineering and Management**
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

These results highlight that while nearly half of the included studies are highly relevant to the practical context, fewer demonstrate strong methodological rigour. This finding underscores a need for more robust research designs to support evidence-based practices in this domain.

**Table 1.** Quality assessment

| Study | Context | Design | Validity | Rigour | Relevance |
|---|---|---|---|---|---|
| Cruz-Benito et al. 2017 [16] | 1 | 1 | 0 | 2 | 4 |
| Alvares et al. 2020 [17] | 1 | 1 | 0 | 2 | 4 |
| Skrede et al. 2024 [18] | 0.5 | 1 | 0 | 2 | 4 |
| Liu et al. 2023 [19] | 1 | 0.5 | 1 | 1.5 | 4 |
| Kaminwar et al. 2023 [20] | 1 | 0 | 1 | 2 | 4 |
| Nam et al. 2022 [21] | 1 | 1 | 1 | 1 | 4 |
| Semjon et al. 2024 [22] | 1 | 1 | 0 | 2 | 5 |
| Yoon et al. 2024 [23] | 1 | 1 | 1 | 3 | 4 |
| Wan et al. 2024 [24] | 1 | 1 | 0 | 2 | 3 |
| Higgins et al. 2023 [25] | 1 | 1 | 1 | 1 | 3 |
| Soltan et al. 2023 [26] | 1 | 1 | 0 | 2 | 4 |
| Hussain et al. 2024 [27] | 0.5 | 0.5 | 0 | 1 | 4 |
| Khaliq et al. 2022 [28] | 0.5 | 1 | 0 | 1.5 | 5 |
| Biro et al. 2020 [29] | 0.5 | 0.5 | 0 | 1 | 4 |
| Kuts et al. 2022 [30] | 1 | 0.5 | 1 | 1.5 | 3 |
| Dashti et al. 2023 [31] | 1 | 1 | 0 | 2 | 3 |
| Ketcham et al. 2025 [32] | 1 | 1 | 0.5 | 1.5 | 5 |
| Chen et al. 2025 [33] | 1 | 0.5 | 1 | 1.5 | 4 |
| Leong et al. 2021 [34] | 1 | 0.5 | 0.5 | 2 | 4 |
| Yoo et al. 2020 [35] | 1 | 1 | 0.5 | 1.5 | 3 |
| Sukarti et al. 2025 [36] | 0.5 | 0.5 | 0 | 1 | 3 |
| Ramos-Rojas et al. 2024 [37] | 0.5 | 1 | 0.5 | 1 | 4 |
| Ozkan et al. 2023 [38] | 1 | 1 | 1 | 2 | 5 |
| Gartziandia et al. 2022 [39] | 1 | 1 | 0.5 | 1.5 | 3 |
| Hislop et al. 2021 [40] | 0.5 | 1 | 0.5 | 1 | 3 |
| Davis et al. 2022 [41] | 1 | 0.5 | 1 | 1.5 | 4 |
| Freire-Obregón et al. 2021 [42] | 1 | 0.5 | 0.5 | 1 | 5 |
| Arriba-Perez et al. 2024 [43] | 1 | 1 | 0 | 1 | 3 |
| Wang et al. 2025 [44] | 0.5 | 1 | 1 | 2.5 | 5 |

**Journal of Information Systems Engineering and Management**
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

| Study | Context | Design | Validity | Rigour | Relevance |
|---|---|---|---|---|---|
| Urrea et al. 2021 [45] | 1 | 0.5 | 0 | 1.5 | 3 |
| Khan et al. 2025 [46] | 0.5 | 1 | 1 | 1.5 | 4 |
| Tang et al. 2020 [47] | 0.5 | 0.5 | 0.5 | 1.5 | 3 |
| Widodo et al. 2023 [48] | 0.5 | 1 | 0.5 | 1 | 5 |
| Sekaran et al. 2023 [49] | 1 | 1 | 0.5 | 1.5 | 4 |
| XUE et al. 2024 [50] | 1 | 1 | 1 | 2 | 5 |
| Capy et al. 2022 [51] | 0.5 | 0.5 | 1 | 1 | 4 |
| Moreira et al. 2023 [52] | 0.5 | 1 | 1 | 1.5 | 3 |
| Płaczek et al. 2018 [53] | 1 | 0.5 | 1 | 1.5 | 3 |
| Ye et al. 2021 [54] | 1 | 1 | 0.5 | 1.5 | 5 |
| Kong et al. 2025 [55] | 1 | 0.5 | 0 | 1 | 4 |
| Zhou et al. 2022 [56] | 0.5 | 1 | 1 | 1.5 | 3 |
| Bairagi 2022 [57] | 1 | 1 | 0 | 1 | 5 |
| Wang et al. 2024 [58] | 1 | 0.5 | 1 | 1.5 | 4 |
| Pinto et al. 2025 [59] | 1 | 1 | 0 | 1 | 3 |
| Yang et al. 2023 [60] | 1 | 1 | 0.5 | 1.5 | 4 |
| Nguyen et al. 2024 [61] | 0.5 | 1 | 1 | 1.5 | 5 |
| Chen et al. 2024 [62] | 1 | 1 | 0.5 | 1.5 | 4 |
| Kim et al. 2023 [63] | 1 | 1 | 0 | 1 | 4 |
| Lwakatare et al. 2021 [64] | 1 | 0.5 | 1 | 1.5 | 4 |
| Vitui et al. 2024 [65] | 0.5 | 1 | 0.5 | 1 | 4 |
| Wang et al. 2024 [66] | 1 | 1 | 1 | 2 | 5 |

## RESULTS

Figure 2 illustrates the distribution of the 51 included studies across five core themes. Most studies addressed Continuous Validation Methods and Improvement Techniques, followed by AI System Focus. Fewer studies focused on Scope of Implementation and Real-World Utility, highlighting a research gap in deployment challenges and contextual applicability.

**Journal of Information Systems Engineering and Management**
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

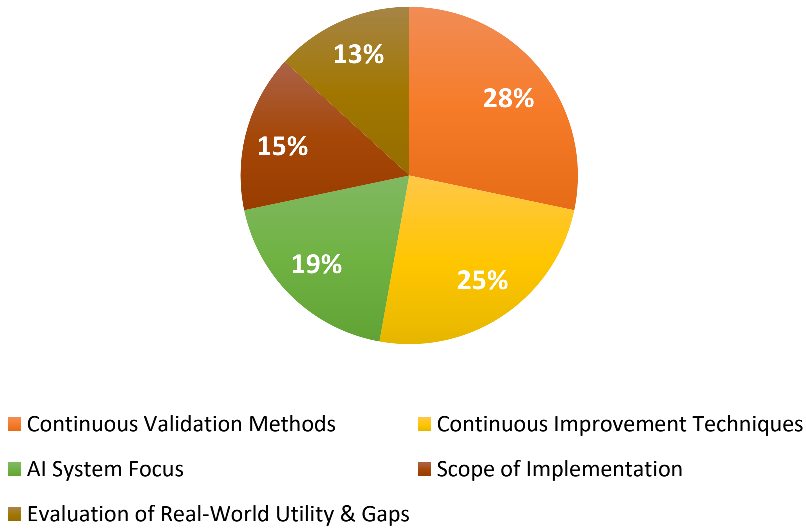


**Figure 2.** Distribution of 51 included studies across five core themes: Continuous Validation Methods, Continuous Improvement Techniques, AI System Focus, Scope of Implementation, and Evaluation of Real-World Utility and Gaps

Table 2 presents the distribution of continuous validation techniques identified across 51 studies in the systematic review. Cross-validation methods, including K-fold and stratified variants, were the most commonly employed (25.49%), followed closely by online retraining or incremental learning approaches (23.53%). Performance monitoring dashboards appeared in 15.69% of studies, while drift detection mechanisms, targeting model or concept drift, were used in 13.73%. Ensemble or hybrid validation strategies accounted for 11.76%, and A/B testing was applied in 9.80% of cases. These findings highlight a diverse yet focused landscape of validation techniques aimed at ensuring reliability and adaptability of AI systems in dynamic settings

**Table 2.** Continuous Validation Techniques Used in Included Studies

| Continuous Validation Technique | No. of Studies (n = 51) | % of Studies | Studies References |
|---|---|---|---|
| Cross-Validation (K-fold, Stratified) | 13 | 25.49 | [16,18,20,21,23,26,28,30,32,37,42,50,63] |
| Online Retraining / Incremental Learning | 12 | 23.53 | [17,19,22,24,25,27,29,33,36,40,45,48] |
| Drift Detection (Model/Concept Drift) | 7 | 13.73 | [31,41,49,51,55,60,66] |
| Performance Monitoring Dashboards | 8 | 15.69 | [34,38,43,44,52,54,58,62] |
| A/B Testing | 5 | 9.80 | [35,46,57,59,61] |
| Ensemble/Hybrid Validation Approaches | 6 | 11.76 | [39,47,53,56,64,65] |

Table 3 presents a quantitative breakdown of continuous improvement techniques across 51 studies. Iterative model refinement was the most prevalent method, employed in 11 studies (21.57%), emphasizing the role of ongoing tuning and updates. Integration with real-world feedback and industry workflow integration were each utilized in 8 studies (15.69%), reflecting the growing emphasis on practical applicability and user-informed development. Adaptive learning or meta-learning and Auto ML or optimization algorithms were each reported in 5 studies (9.80%), indicating a moderate focus on automation and dynamic learning enhancements. Community or open-source contributions and proactive error correction mechanisms appeared in 4 studies each (7.84%), showcasing collaborative and preventative strategies. Longitudinal deployment monitoring was noted in 6 studies (11.76%), underlining the need for sustained post-deployment evaluation. These findings highlight a diversified approach to continuous improvement, with over 60% of studies incorporating at least one feedback or refinement mechanism to maintain and boost AI performance over time.

**Table 3.** Continuous Improvement Techniques Identified Across Included Studies

| Continuous Improvement Techniques | No. of Studies (n = 51) | % of Studies | Studies References |
|---|---|---|---|
| Iterative Model Refinement | 11 | 21.57 | [16,18,22,25,27,29,31,33,36,40,43] |
| Integration with Real-World Feedback | 8 | 15.69 | [17,19,24,26,32,38,41,47] |
| Industry Workflow Integration | 8 | 15.69 | [20,21,28,35,37,39,44,50] |
| Adaptive Learning / Meta-Learning | 5 | 9.80 | [23,30,34,42,51] |
| Use of AutoML or Optimization Algorithms | 5 | 9.80 | [45,46,48,49,52] |
| Community or Open-Source Contributions | 4 | 7.84 | [53–56] |
| Proactive Error Correction Mechanisms | 4 | 7.84 | [57–60] |
| Longitudinal Deployment Monitoring | 6 | 11.76 | [61–66] |

Table 4 outlines the distribution of AI system domains, techniques, and data types across the 51 included studies. Smart manufacturing and industrial AI and robotics (industrial, medical, service) were the leading application domains, each represented in 14 studies (27.45%), using techniques like machine learning (RF, SVM, CNN), digital twins, and trajectory tracking with data from sensors, quality inspections, and force-feedback systems. Healthcare and biomedicine accounted for 7 studies (13.73%), applying CNNs, NLP, and diagnostic tools to patient records and clinical data. Environmental and energy systems were explored in 6 studies (11.76%), leveraging deep learning models like CNN, LSTM, and XGBoost on climate, SCADA, and smart meter data. UI/UX and human-computer interaction comprised 4 studies (7.84%), employing explainable AI and object detection on interface and interaction data. Regulatory and life sciences (5.88%), cyber-physical systems, agriculture, and decision modeling domains (each 1.96%) used specialized models with domain-specific datasets. This reflects a broad yet industrial-heavy focus in AI system applications.

**Table 4.** AI System Domains, Techniques, and Data Types Across Included Studies

| AI Application Domain | No. of Studies (n = 51) | % of Studies | Common AI Techniques Used | Typical Data Types | Studies References |
|---|---|---|---|---|---|
| Smart Manufacturing & Industrial AI | 14 | 27.45 | ML (RF, SVM, CNN), Digital Twin, Fault Detection | Sensor data, operational data, quality inspection | [17,20,23,26,33,38,40,46,51,56,57,61,63,66] |
| Robotics (Industrial, Medical, Service) | 14 | 27.45 | Robot control, calibration, trajectory tracking | Positional, force-torque, sensor feedback | [18,22,24,29,30,35,37,41,43,47,53,58,64,65] |

**Journal of Information Systems Engineering and Management**
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

| AI Application Domain | No. of Studies (n = 51) | % of Studies | Common AI Techniques Used | Typical Data Types | Studies References |
|---|---|---|---|---|---|
| Healthcare & Biomedicine | 7 | 13.73 | ML (CNN, ANN), NLP, Conversational AI, Diagnostic tools | Patient records, mammograms, speech/text data | [21,32,34,42,52,59,62] |
| Environmental & Energy Systems | 6 | 11.76 | CNN, LSTM, Hybrid DL, XGBoost, AI-based anomaly detection | Climate data, SCADA, smart meters, grid signals | [19,27,31,36,55,60] |
| UI/UX & HCI | 4 | 7.84 | Object detection, Explainable AI, Deep Learning (CNN) | UI screenshots, interaction logs, test cases | [16,28,48,49] |
| Regulatory & Life Sciences | 3 | 5.88 | AutoML, Medical validation, AI in pharma/devices | Simulation data, clinical records, reports | [25,44,45] |
| Cyber-Physical Systems & Simulation | 1 | 1.96 | Regression, Simulated AI systems, Twin models | Simulated data, synthetic logs | [39] |
| Agriculture | 1 | 1.96 | AI-based environmental control | Sensor data (temp, humidity, soil moisture) | [50] |
| Decision Models & Theoretical Systems | 1 | 1.96 | Fuzzy logic, AI-based decision matrix evaluation | Benchmark datasets, synthetic input | [54] |

Table 5 provides an overview of the geographical scope, target audiences, contextual factors, and application areas addressed in the 51 included studies. Industry-focused implementations were the most common, appearing in 10 studies (19.61%), primarily covering smart factories, automation, and manufacturing control systems. Localized or regional implementations were reported in 8 studies (15.69%), often focused on country-specific deployments in China, the EU, India, and the USA. Healthcare-specific deployments appeared in 7 studies (13.73%), targeting hospitals and diagnostic centers. Academic and research settings were represented in 6 studies (11.76%), leveraging simulated environments or curated datasets for model evaluation. Another 6 studies (11.76%) emphasized ethical and socio-technical factors, including data privacy, bias mitigation, and regulatory constraints. Global or cross-regional implementations and human-AI interaction contexts were each found in 5 studies (9.80%), while cyber-physical systems and agricultural/environmental applications each appeared in 2 studies (3.92%), demonstrating a relatively diverse but industry-heavy implementation landscape

**Journal of Information Systems Engineering and Management**
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

**Table 5.** Geographical Scope, Target Audience, Contextual Factors, and Application Areas Across Included Studies

| Scope of Implementation | No. of Studies (n = 51) | % of Studies | Common Elements | Studies References |
|---|---|---|---|---|
| Global or Cross-Regional Scope | 5 | 9.80 | Multi-continent deployment, international data sources, collaborative AI validation | [16,24,34,44,48] |
| Industry-Focused Implementation | 10 | 19.61 | Manufacturing, automation, process control, smart factories | [19,21,25,37,39,41,42,49,50,59] |
| Healthcare-Specific Deployment | 7 | 13.73 | Hospitals, diagnostic centers, biomedical testing facilities | [17,22,33,45,52,53,66] |
| Academic & Research Settings | 6 | 11.76 | Model evaluation via academic datasets, simulated environments | [23,27,30,40,47,64] |
| Localized/Regional Implementations | 8 | 15.69 | Country-specific AI models (e.g., China, EU, India, USA) | [20,29,32,43,51,55,57, 63] |
| Human-Interaction & UI Contexts | 5 | 9.80 | AI in user interfaces, testing human-AI interaction | [18,26,28,46,60] |
| Ethical & Socio-Technical Emphasis | 6 | 11.76 | Data privacy, bias, regulatory constraints, human oversight | [31,35,38,54,56,62] |
| Cyber-Physical Systems | 2 | 3.92 | Smart infrastructure, embedded systems | [36,58] |
| Agricultural & Environmental Focus | 2 | 3.92 | Climate models, energy grid, resource monitoring | [61,65] |

Table 6 summarizes the key challenges, limitations, ethical concerns, and practical gaps identified across the 51 included studies. The most frequently reported issue was model drift and data distribution shifts, appearing in 7 studies (13.73%), highlighting concerns about performance degradation over time and the need for adaptive retraining strategies. Generalizability and dataset bias were identified in 6 studies (11.76%), often due to narrow or unrepresentative training data. Similarly, ethical and legal concerns including privacy, fairness, and explainability were discussed in 6 studies (11.76%). Computational and resource constraints and poor human-AI interpretability were each noted in 5 studies (9.80%), reflecting barriers to scalability and real-world usability. Less frequently reported but still significant were challenges related to scalability (5.88%), regulatory guidance (5.88%), security and adversarial risks (5.88%), and inconsistent recommendations (5.88%). A smaller subset of studies highlighted gaps in evaluation metrics (3.92%), continuous feedback loops (5.88%), and the over-reliance on retrospective data (3.92%), pointing to the need for more robust, forward-looking validation approaches.

**Journal of Information Systems Engineering and Management**
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

**Table 6.** Challenges, Limitations, Ethical Factors, and Practical Gaps in Included Studies

| Evaluation of Real-World Utility and Gaps | No. of Studies (n = 51) | % of Studies | Common Elements | Studies References |
|---|---|---|---|---|
| Model Drift & Data Distribution Shifts | 7 | 13.73 | Performance decay over time, retraining triggers, concept drift mitigation | [16,22,28,36,39,46,64] |
| Generalizability & Dataset Bias | 6 | 11.76 | Limited external validation, non-representative training data | [17,24,27,38,41,51] |
| Computational & Resource Constraints | 5 | 9.80 | High compute costs, model scalability, energy demands | [20,33,42,52,62] |
| Ethical and Legal Concerns | 6 | 11.76 | Data privacy, fairness, accountability frameworks, explainability | [21,25,29,47,53,56] |
| Poor Human-AI Interpretability | 5 | 9.80 | Lack of explainability, end-user mistrust, usability in real settings | [18,32,40,50,58] |
| Scalability to Industry or Population Level | 3 | 5.88 | Deployment at national or enterprise level hindered by cost, access, or integration barriers | [26,34,44] |
| Lack of Regulatory or Ethical Guidance | 3 | 5.88 | AI deployment without policy standards | [23,35,48] |
| Limited Evaluation Metrics | 2 | 3.92 | Overuse of accuracy/F1 without robustness, fairness, or uncertainty quantification | [19,30] |
| Security & Adversarial Risks | 3 | 5.88 | Vulnerability to model poisoning, adversarial examples | [31,49,54] |
| Inconsistent Recommendations | 3 | 5.88 | Studies provided vague or conflicting deployment advice | [37,43,57] |
| Under-addressed Socio-Technical Dimensions | 3 | 5.88 | User involvement, team workflows, contextual settings missing from validation design | [45,55,63] |
| No Clear Strategy for Continuous Feedback | 3 | 5.88 | Feedback loops lacking or poorly implemented | [59–61] |
| Over-reliance on Retrospective Data | 2 | 3.92 | Absence of prospective evaluation or field deployment | [65,66] |

## DISCUSSION

This review highlights that AI systems predominantly rely on internal mechanisms to maintain performance, with cross-validation (25.49%) and online retraining (23.53%) being the most frequently employed techniques. These methods demonstrate a continued preference for data-driven fine-tuning within controlled settings. However, the relatively limited use of drift detection (13.73%) and A/B testing (9.80%) points to a persistent gap in real-time performance evaluation. This imbalance suggests

**Journal of Information Systems Engineering and Management**
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

that while internal validation is well-developed, operational and external validation remains underemphasized. Addressing this requires the integration of drift detection tools and hybrid evaluation frameworks to bolster system robustness under dynamic conditions. These findings align with recent studies advocating for advanced cross-validation strategies that ensure robustness across diverse datasets [68]. Despite widespread recognition of its importance, drift detection continues to face challenges in deployment due to complexity, as documented in systematic reviews emphasizing gaps in operational validation under real-world constraints [69]. Since 2017, there has been growing advocacy for hybrid validation systems combining internal checks with external, real-time monitoring, particularly in policy and governance literature [70]. In terms of continuous improvement strategies, AI development remains heavily reliant on manual iterative model refinement (21.57%), underscoring a developer-centric evolution model. Nonetheless, increasing incorporation of real-world feedback and workflow integration (15.69% each) reflects rising industry engagement and responsiveness to operational contexts. The limited use of AutoML and adaptive learning (9.80% each), however, reveals missed opportunities for scalable self-improving systems. Additionally, community-driven improvements and proactive error correction remain peripheral (7.84%). These trends mirror observations from Microsoft Research, which noted that developers and data scientists still favor human-in-the-loop workflows due to their control, interpretability, and trust benefits [71]. Ahmed and Lin (2021) similarly observed that manual model tuning and data pipeline adjustments are persistent sources of technical debt in AI system maintenance [72]. These patterns demonstrate that despite advances in automation technologies, the field continues to rely heavily on expert oversight. Other studies from the late 2010s and early 2020s also emphasized expert-driven tuning as central to managing drift and data variability in deployed systems [73].

Current study shows a heavy concentration of continuous validation and improvement research in industrial AI domains, particularly smart manufacturing and robotics (27.45% each), driven by structured data and well-defined outcomes. The study further reveals a domain imbalance in the application of continuous validation and improvement strategies. Industrial sectors particularly smart manufacturing and robotics dominate the landscape (27.45% each), benefitting from structured data and clearly defined performance goals. Conversely, complex and sensitive domains like healthcare (13.73%) and environmental systems (11.76%) are underrepresented, largely due to data heterogeneity and ethical concerns. Minimal representation in agriculture, regulatory science, and cyber-physical systems further signals an uneven focus across sectors. The readiness of manufacturing environments for continuous AI adaptation can be attributed to high-frequency sensor data and structured workflows, as noted in Industry 4.0 frameworks [74]. Similarly, intelligent robotics excels in structured industrial contexts that support iterative validation and learning cycles [75]. Broader industrial transformation under Industry 5.0 further emphasizes ethical deployment, human-machine collaboration, and sustainable innovation [76]. A recent review also confirmed that AI integration with Industry 4.0 technologies such as CPS, IoT, and big data improves predictive maintenance and process optimization, although system integration remains a challenge [77]. Implementation scope across studies shows fragmentation. Nearly one-fifth (19.61%) of studies target industry-specific use cases, and 15.69% focus on regional applications. While healthcare (13.73%) and academic settings (11.76%) often demonstrate high theoretical rigor, their scalability outside controlled environments is limited. Only a small proportion (9.80%) of studies reflect global or cross-contextual implementations. Furthermore, socio-technical and human-centric dimensions are underexplored. This disconnects between experimental development and real-world adaptability emphasizes the need for context-aware frameworks that integrate regulatory diversity, cultural variability, and stakeholder engagement. According to Baytech Consulting's 2025 report, AI integration in enterprises surged in 2024, but regional disparities and policy environments continue to shape deployment outcomes [78]. A recent Science Direct review also points to ongoing fragmentation, calling for more holistic frameworks that account for socio-technical complexity [79]. Meanwhile, recent advances in autonomous AI agents demonstrate increasing technical capabilities in planning and reasoning, yet also underscore the importance of balancing performance gains with contextual sensitivity [80].

Critical barriers to sustainable AI performance remain. Model drift (13.73%) emerged as the most frequently cited challenge, followed by dataset bias and ethical or legal issues (11.76% each). These concerns represent foundational vulnerabilities that can erode trust and degrade performance if unaddressed. Additionally, limited interpretability and high computational demands (10%) hinder deployment in sensitive domains. Less frequently discussed but equally vital challenges include weak feedback loops, adversarial risks, and immature evaluation metrics. The low reliance on prospective, continuous assessment methods (3.92%) further illustrates a reactive approach to validation. Building adaptive and trustworthy AI systems will require both institutional innovation and technical advances. For example, Kirichenko et al. (2023) warned about feedback loops that reinforce model bias, urging the development of adaptive monitoring mechanisms [81]. Papernot et al. (2022) highlighted the need for adversarial robustness and continuous training to counter evolving threats [82]. Similarly, drift detection remains essential, as recent findings confirm reliability issues in systems lacking ongoing performance checks [83] The 2025 Artificial Intelligence Index Report also identifies the interpretability gap and high resource consumption as major deployment hurdles, noting that while hardware costs are declining, training large models remains costly and centralized [84]. In sum, advancing AI systems toward sustainable, trustworthy, and context-aware performance will require a deliberate shift from isolated technical optimization to holistic frameworks that integrate continuous validation, adaptive improvement, domain diversity, and socio-technical resilience.

## CONCLUSION

This systematic literature review analyzed 51 studies to map the landscape of continuous validation and improvement strategies for AI systems. Findings indicate a robust foundation in technical validation methods particularly cross-validation and online retraining yet reveal a lack of comprehensive integration of real-world monitoring tools like drift detection and A/B testing. Improvement strategies are varied, but predominantly manual, with iterative refinement, feedback incorporation, and workflow alignment taking precedence. Automation-oriented approaches, such as Auto ML and adaptive learning, are underutilized, limiting scalability. The application landscape is heavily skewed toward industrial domains, particularly manufacturing and robotics, with healthcare and environmental applications emerging but less prevalent. In terms of scope, most studies target localized or sector-specific implementations, with few addressing global or cross-context interoperability. Alarmingly, critical dimensions such as ethical considerations, regulatory frameworks, interpretability, and user engagement remain inconsistently addressed. Furthermore, key challenges including model drift, data bias, and limited evaluation metrics suggest that AI systems are not yet fully equipped for adaptive, context-sensitive deployment in dynamic environments. Future research should prioritize cross-sector integration, longitudinal deployment monitoring, and inclusive validation frameworks that incorporate socio-technical factors. A shift toward collaborative, adaptive, and ethically anchored AI development will be essential for real-world sustainability and trust.

**Acknowledgement:**

I would like to express my heartfelt gratitude to Tunku Abdul Rahman University of Management and Technology (TAR UMT) for providing me with the opportunity to publish this journal, along with their invaluable support and generous funding. My deepest appreciation also goes to my supervisor, Dr. Kamsiah Mohamed from UNISEL, for her continuous guidance, encouragement, and insightful feedback throughout the research and writing process.

e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

**Funding Statement:**

This research was supported by Tunku Abdul Rahman University of Management and Technology (TAR UMT) through institutional funding.

**Data Availability:**

No new data were created or analyzed in this study.

**Conflict of Interest:**

None.

## REFERENCES

[1] How artificial intelligence is transforming the world n.d. https://www.brookings.edu/articles/how-artificial-intelligence-is-transforming-the-world/ (accessed June 14, 2025).

[2] Ibegbulam C.M, Olowonubi, J.A, Fatounde, S.A, Oyegunwa, O.A. ARTIFICIAL INTELLIGENCE IN THE ERA OF 4IR: DRIVERS, CHALLENGES AND OPPORTUNITIES. Engineering Science & Technology Journal 2023;4:473–88. https://doi.org/10.51594/ESTJ.V4I6.668.

[3] Naveen Kodakandla. Data drift detection and mitigation: A comprehensive MLOps approach for real-time systems. International Journal of Science and Research Archive 2024;12:3127–39. https://doi.org/10.30574/ IJSRA.2024.12.1.0724.

[4] Gallab M, Di Nardo M. New innovation, sustainability, and resilience challenges in the X.0 era. Appl Syst Innov 2023;6:39. https://doi.org/10.3390/asi6020039.

[5] Sinha S, Lee YM. Challenges with developing and deploying AI models and applications in industrial systems. Discover Artificial Intelligence 2024;4:1–19. https://doi.org/10.1007/S44163-024-00151-2/FIGURES/3.

[6] Ferrara E. GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. J Comput Soc Sci 2024;7:549–69. https://doi.org/10.1007/s42001-024-00250-1.

[7] Ferrara E. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci 2024, Vol 6, Page 3 2023;6:3. https://doi.org/10.3390/SCI6010003.

[8] AI Model Validation: Ensuring Accuracy and Reliability n.d. https://focalx.ai/ai/ai-model-validation/ (accessed June 14, 2025).

[9] Ogrizović M, Drašković D, Bojić D. Quality assurance strategies for machine learning applications in big data analytics: an overview. J Big Data 2024;11:1–48. https://doi.org/10.1186/S40537-024-01028-Y/TABLES/6.

[10] Aldoseri A, Al-Khalifa KN, Hamouda AM. Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. Applied Sciences (Switzerland) 2023;13. https://doi.org/10.3390/ APP13127082.

[11] Radanliev P. AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development. Applied Artificial Intelligence 2025;39. https://doi.org/10.1080/08839514.2025.2463722.

[12] Guan H, Bates D, Zhou L. Keeping Medical AI Healthy: A Review of Detection and Correction Methods for System Degradation 2025.

[13] Benneh Mensah G. Artificial Intelligence and Ethics: A Comprehensive Review of Bias Mitigation, Transparency, and Accountability in AI Systems. 2023. https://doi.org/10.13140/RG.2.2.23381.19685/1.

[14] Tilala MH, Chenchala PK, Choppadandi A, Kaur J, Naguri S, Saoji R, et al. Ethical Considerations in the Use of Artificial Intelligence and Machine Learning in Health Care: A Comprehensive Review. Cureus 2024;16: e62443. https://doi.org/10.7759/CUREUS.62443.
[15] Ayodeji A. Artificial Intelligence in Enhancing Regulatory Compliance and Risk Management. 2024. https://doi.org/10.13140/RG.2.2.20915.44326.
[16] Cruz-Benito J, Vazquez-Ingelmo A, Sanchez-Prieto JC, Theron R, Garcia-Penalvo FJ, Martin-Gonzalez M. Enabling Adaptability in Web Forms Based on User Characteristics Detection Through A/B Testing and Machine Learning. IEEE Access 2017;6:2251–65. https://doi.org/10.1109/ACCESS.2017.2782678.
[17] Alvares AJ, Rodriguez E, Riano Jaimes CI, Toquica JS, Ferreira JCE. STEP-NC Architectures for Industrial Robotic Machining: Review, Implementation and Validation. IEEE Access 2020;8:152592–610. https://doi.org/10.1109/ACCESS.2020.3017561.
[18] Skrede A, Dalen AF, Hellevik AI, Stavdahl O, Bye RT. Quantifying Passive Biomechanical Stability Using an Industrial Robot: Development and Experimental Validation of a Task Space Motion Framework. IEEE Access 2024;12:29397–413. https://doi.org/10.1109/ACCESS.2024.3368390.
[19] Liu S, Zhang J, Wang K, Wu X, Chen W, Liang S, et al. Structural indicator synergy for mitigating extreme urban heat island effects in industrial city: Simulation and verification based on machine learning. Ecol Indic 2023;157:111216. https://doi.org/10.1016/J.ECOLIND.2023.111216.
[20] Kaminwar SR, Goschenhofer J, Thomas J, Thon I, Bischl B. Structured Verification of Machine Learning Models in Industrial Settings. Big Data 2023;11:181–98. https://doi.org/10.1089/BIG.2021.0112/ASSET/ IMAGES/BIG.2021.0112_FIGURE3.JPG.
[21] Nam KH, Kim DY, Kim DH, Lee JH, Lee J Il, Kim MJ, et al. Conversational Artificial Intelligence for Spinal Pain Questionnaire: Validation and User Satisfaction. Neurospine 2022;19:348. https://doi.org/10.14245/NS. 2143080.540.
[22] Semjon J, Varga J, Pilat P, Marcinko P, Janos R. VERIFICATION OF THE PROPERTIES OF AN INDUSTRIAL ROBOT FOR THE NUCLEAR INDUSTRY. MM Science Journal 2024;2024:7788–94. https://doi.org/10. 17973/MMSJ.2024_11_2024056.
[23] Yoon H, Kim HJ, Kim S. Validation and Implementation of Customer Classification System using Machine Learning. Measurement 2024;22:131–40. https://doi.org/10.1080/15366367.2023.2246111;JOURNAL: JOURNAL:HMES20;WGROUP:STRING:PUBLICATION.
[24] Wan L, Liu K, Mengash HA, Alruwais N, Duhayyim M Al, Venkatachalam K. Deep learning-based photoplethysmography biometric authentication for continuous user verification. Appl Soft Comput 2024; 156:111461. https://doi.org/10.1016/J.ASOC.2024.111461.
[25] Higgins DC, Johner C. Validation of Artificial Intelligence Containing Products Across the Regulated Healthcare Industries. Ther Innov Regul Sci 2023;57:797–809. https://doi.org/10.1007/S43441-023-00530-4/FIGURES/3.
[26] Soltan H, Janada K, Omar M. FAQT-2: A customer-oriented method for MCDM with statistical verification applied to industrial robot selection. Expert Syst Appl 2023;226:120106. https://doi.org/10.1016/J.ESWA. 2023.120106.
[27] Hussain A, Goljandin D, Podgursky V, Rüstü Yörük C, Sergejev F, Kübarsepp J, et al. Industrial sustainable Fabrication, SEM Characterization, mechanical Testing, ANOVA analysis of PP-PETF recycled Composites: Artificial intelligence and deep learning studies for nuclear shielding

**Journal of Information Systems Engineering and Management**
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

applications. Eur Polym J 2024; 213:113082. https://doi.org/10.1016/J.EURPOLYMJ.2024.113082.
[28] Khaliq Z, Farooq SU, Khan DA. A deep learning-based automated framework for functional User Interface testing. Inf Softw Technol 2022;150:106969. https://doi.org/10.1016/J.INFSOF.2022.106969.
[29] Biro I, Kinnell P. Performance Evaluation of a Robot-Mounted Interferometer for an Industrial Environment. Sensors 2020, Vol 20, Page 257 2020;20:257. https://doi.org/10.3390/S20010257.
[30] Kuts V, Marvel JA, Aksu M, Pizzagalli SL, Sarkans M, Bondarenko Y, et al. Digital Twin as Industrial Robots Manipulation Validation Tool. Robotics 2022, Vol 11, Page 113 2022;11:113. https://doi.org/10.3390/ ROBOTICS11050113.
[31] Dashti A, Raji M, Riasat Harami H, Zhou JL, Asghari M. Biochar performance evaluation for heavy metals removal from industrial wastewater based on machine learning: Application for environmental protection. Sep Purif Technol 2023;312:123399. https://doi.org/10.1016/J.SEPPUR.2023.123399.
[32] Ketcham M, Boonyopakorn P, Ganokratanaa T. Resolution-Aware Deep Learning with Feature Space Optimization for Reliable Identity Verification in Electronic Know Your Customer Processes. Mathematics 2025, Vol 13, Page 1726 2025;13:1726. https://doi.org/10.3390/MATH13111726.
[33] Chen YT, Liu CS, Lee BK. Trajectory positioning error compensation and verification for six-axis industrial robot. International Journal of Advanced Manufacturing Technology 2025;137:5205–12. https://doi.org/10. 1007/S00170-025-15486-Y/FIGURES/8.
[34] Leong PH, Goh OS, Kumar YJ, Sam YH, Fong CW. The Evaluation of User Experience Testing for Retrieval-based Model and Deep Learning Conversational Agent. International Journal of Advanced Computer Science and Applications 2021;12:216–21. https://doi.org/10.14569/IJACSA.2021.0120429.
[35] Yoo S, Jang J, Lee JK, Park JW. Optimal Robot Motion for Accelerated Life Testing of a 6-DoF Industrial Robot. Applied Sciences 2020, Vol 10, Page 7459 2020;10:7459. https://doi.org/10.3390/APP10217459.
[36] Sukarti S, Sulaima MF, Abdul Kadir AF, Zulkafli NI, Othman ML, Hanak DP. Enhancing energy savings verification in industrial settings using deep learning and anomaly detection within the IPMVP framework. Energy Build 2025;327:115096. https://doi.org/10.1016/J.ENBUILD.2024.115096.
[37] Ramos-Rojas J, Castano JA, Fernández PR, Carballeira J, Pérez-Martín E, Lora-Millan JS, et al. Design and Validation of an Ambulatory User Support Gait Rehabilitation Robot: NIMBLE. Actuators 2024, Vol 13, Page 348 2024;13:348. https://doi.org/10.3390/ACT13090348.
[38] Ozkan M, Demirci Z, Aslan Ö, Yazıcı A. Safety Verification of Multiple Industrial Robot Manipulators with Path Conflicts Using Model Checking. Machines 2023, Vol 11, Page 282 2023;11:282. https://doi.org/10.3390/ MACHINES11020282.
[39] Gartziandia A, Arrieta A, Ayerdi J, Illarramendi M, Agirre A, Sagardui G, et al. Machine learning-based test oracles for performance testing of cyber-physical systems: An industrial case study on elevators dispatching algorithms. Journal of Software: Evolution and Process 2022;34. https://doi.org/10.1002/SMR.2465.
[40] Hislop J, Isaksson M, McCormick J, Hensman C. Validation of 3-Space Wireless Inertial Measurement Units Using an Industrial Robot. Sensors 2021, Vol 21, Page 6858 2021;21:6858.

https://doi.org/10.3390/
S21206858.
[41] Davis C, Aid G. Machine learning-assisted industrial symbiosis: Testing the ability of word vectors to estimate similarity for material substitutions. J Ind Ecol 2022;26:27–43. https://doi.org/10.1111/JIEC.13245;
WGROUP:STRING:PUBLICATION.
[42] Freire-Obregón D, Rosales-Santana K, Marín-Reyes PA, Penate-Sanchez A, Lorenzo-Navarro J, Castrillón-Santana M. Improving user verification in human-robot interaction from audio or image inputs through sample quality assessment. Pattern Recognit Lett 2021;149:179–84. https://doi.org/10.1016/J.PATREC.2021.06.014.
[43] De Arriba-Perez F, Garcia-Mendez S, Otero-Mosquera J, Gonzalez-Castano FJ, Gil-Castineira F. Automatic Generation of Insights From Workers' Actions in Industrial Workflows With Explainable Machine Learning: A Proposed Architecture With Validation. IEEE Industrial Electronics Magazine 2024;18:17–29. https://doi.org/
10.1109/MIE.2023.3284203.
[44] Wang D, Jin J, Shi G, Bao J, Wang Z, Li S, et al. ADMET evaluation in drug discovery: 21. Application and industrial validation of machine learning algorithms for Caco-2 permeability prediction. J Cheminform 2025;17:1–14. https://doi.org/10.1186/S13321-025-00947-Z/FIGURES/6.
[45] Urrea C, Pascal J. Design and validation of a dynamic parameter identification model for industrial manipulator robots. Archive of Applied Mechanics 2021;91:1981–2007. https://doi.org/10.1007/S00419-020-01865-2/METRICS.
[46] Khan AH, Su H, Cao X, Pham DT, Ji Z, Packianather M, et al. BOMO-RNN: a novel neural network controller for industrial robots with experimental validation. Int J Syst Sci 2025. https://doi.org/10.1080/00207721.
2025.2482871;PAGE:STRING:ARTICLE/CHAPTER.
[47] Tang S, Liu G, Lin Z, Li X, Pan M. A Test Procedure Optimization Method for an Industrial Robot Servo System on an Integrated Testing Platform. Complexity 2020;2020:7906019. https://doi.org/10.1155/2020/7906019.
[48] Widodo AP, Wibowo A, Kurniawan K. Enhancing Software User Interface Testing Through Few Shot Deep Learning: A Novel Approach for Automated Accuracy and Usability Evaluation. International Journal of Advanced Computer Science and Applications 2023;14:578–85. https://doi.org/10.14569/IJACSA.2023.
0141260.
[49] Sekaran JF, Hildas R. Implementation of an Industrial Automation Remote Lab (IARL) and validation using a deep learning approach during the COVID pandemic. Computer Applications in Engineering Education 2023;31:1593–607. https://doi.org/10.1002/CAE.22663;REQUESTEDJOURNAL:JOURNAL:10990542;
PAGE:STRING:ARTICLE/CHAPTER.
[50] XUE F, LIU Q, ZHANG T, LIU S, CHENG J, MA C. Automatic Construction of Graphical User Interfaces Semantic Models Using Robots for Mobile Application Testing. Studies in Informatics and Control 2024;33:73–90. https://doi.org/10.24846/V33I1Y202407.
[51] Capy S, Rincon L, Coronado E, Hagane S, Yamaguchi S, Leve V, et al. Expanding the Frontiers of Industrial Robots beyond Factories: Design and in the Wild Validation. Machines 2022, Vol 10, Page 1179 2022;10:1179. https://doi.org/10.3390/MACHINES10121179.
[52] Moreira PA, Fernandes RM, Avila LV, Bastos LDSL, Martins VWB. Artificial Intelligence and Industry 4.0? Validation of Challenges Considering the Context of an Emerging Economy Country Using Cronbach's Alpha and the Lawshe Method. Eng 2023, Vol 4, Pages 2336-2351

2023;4:2336–51. https://doi.org/10.3390/ENG4030133.
[53] Płaczek M, Piszczek Ł. Testing of an industrial robot's accuracy and repeatability in off and online environment. Eksploatacja i Niezawodność – Maintenance and Reliability 2018;20:455–64. https://doi.org/10.17531/EIN.2018.3.15.
[54] Ye J, Du S, Yong R. Similarity Measures between Intuitionistic Fuzzy Credibility Sets and Their Multicriteria Decision-Making Method for the Performance Evaluation of Industrial Robots. Math Probl Eng 2021; 2021:6630898. https://doi.org/10.1155/2021/6630898.
[55] Kong D, Hong Y, Yang Y, Gu T, Fu Y, Ye Y, et al. A parametric, control-integrated and machine learning-enhanced modeling method of demand-side HVAC systems in industrial buildings: A practical validation study. Appl Energy 2025;379:124971. https://doi.org/10.1016/J.APENERGY.2024.124971.
[56] Zhou Z, Yang X, Wang H, Zhang X. Coupled dynamic modeling and experimental validation of a collaborative industrial mobile manipulator with human-robot interaction. Mech Mach Theory 2022;176:105025. https://doi.org/10.1016/J.MECHMACHTHEORY.2022.105025.
[57] Bairagi B. Technique of Accurate Ranking Order (TARO): A novel multi criteria analysis approach in performance evaluation of industrial robots for material handling. Decision Science Letters 2022;11:563–89. https://doi.org/10.5267/J.DSL.2022.5.001.
[58] Wang YC, Papastathopoulos A. Cross-segment validation of customer support for AI-based service robots at luxury, fine-dining, casual, and quick-service restaurants. International Journal of Contemporary Hospitality Management 2024;36:1744–65. https://doi.org/10.1108/IJCHM-11-2022-1448.
[59] Pinto A, Lemos L, Carvalho C, Santos J, Menezes P, Nomura T. Translation, Adaptation, and Validation in Portuguese of an Acceptance Scale for Human–Robot Interaction in an Industrial Context. Hum Behav Emerg Technol 2025;2025:8816379. https://doi.org/10.1155/HBE2/8816379.
[60] Yang X, Luan F, Zhang J, Zhang Z. Testing for quadratic impact of industrial robots on environmental performance and reaction to green technology and environmental cost. Environmental Science and Pollution Research 2023;30:92782–800. https://doi.org/10.1007/S11356-023-28864-4/METRICS.
[61] Nguyen VT, Do CD, Dang TV, Bui TL, Tan PX. A comprehensive RGB-D dataset for 6D pose estimation for industrial robots pick and place: Creation and real-world validation. Results in Engineering 2024;24:103459. https://doi.org/10.1016/J.RINENG.2024.103459.
[62] Chen T, Chen J, Or CK. Heuristic evaluation and end-user testing of a machine learning-based lower-limb exercise training system for management of knee pain in individuals aged 55 years or over. Int J Ind Ergon 2024;102:103607. https://doi.org/10.1016/J.ERGON.2024.103607.
[63] Kim DJ, Locke S, Chen THP, Toma A, Sporea S, Weinkam L, et al. Challenges in Adopting Artificial Intelligence Based User Input Verification Framework in Reporting Software Systems. Proceedings - International Conference on Software Engineering 2023:99–109. https://doi.org/10.1109/ICSE-SEIP58684.2023.00014.
[64] Lwakatare LE, Range E, Crnkovic I, Bosch J. On the experiences of adopting automated data validation in an industrial machine learning project. Proceedings - International Conference on Software Engineering 2021:248–57. https://doi.org/10.1109/ICSE-SEIP52600.2021.00034.
[65] Vitui A, Chen TH. MLOLET - Machine Learning Optimized Load and Endurance Testing: An industrial experience report. Proceedings - 2024 39th ACM/IEEE International Conference on Automated Software Engineering, ASE 2024 2024:1956–66.

https://doi.org/10.1145/3691620.3695258;SUBPAGE:STRING:ABSTRACT;CSUBTYPE:STRING:CONFERENCE.
[66] Wang H, Wang Y, Ming X. Motion control strategies for industrial robots and their performance evaluation. ACM International Conference Proceeding Series 2024:50–3. https://doi.org/10.1145/3689299.3689309;SUBPAGE:STRING:ABSTRACT;CSUBTYPE:STRING:CONFERENCE.
[67] Ivarsson M, Gorschek T. A method for evaluating rigor and industrial relevance of technology evaluations. Empir Softw Eng 2011;16:365–95. https://doi.org/10.1007/S10664-010-9146-4/METRICS.
[68] Hartung T, Kleinstreuer N. Challenges and opportunities for validation of AI-based new approach methods. ALTEX - Alternatives to Animal Experimentation 2025;42:3–21. https://doi.org/10.14573/ALTEX.2412291.
[69] Myllyaho L, Raatikainen M, Männistö T, Mikkonen T, Nurminen JK. Systematic literature review of validation methods for AI systems. Journal of Systems and Software 2021;181:111050. https://doi.org/10.1016/J.JSS.2021.111050.
[70] Feldman DK-D, Benoliel D. Beyond Explainability: The Case for AI Validation 2025.
[71] Alenezi M, Akour M. AI-Driven Innovations in Software Engineering: A Review of Current Practices and Future Directions. Applied Sciences 2025, Vol 15, Page 1344 2025;15:1344. https://doi.org/10.3390/APP15031344.
[72] Tang Y, Khatchadourian R, Bagherzadeh M, Singh R, Stewart A, Raja A. An empirical study of refactorings and technical debt in machine learning systems. Proceedings - International Conference on Software Engineering 2021:238–50. https://doi.org/10.1109/ICSE43902.2021.00033.
[73] Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review 2022 56:4 2022;56:3005–54. https://doi.org/10.1007/S10462-022-10246-W.
[74] Varriale V, Cammarano A, Michelino F, Caputo M. Critical analysis of the impact of artificial intelligence integration with cutting-edge technologies for production systems. J Intell Manuf 2025;36:61–93. https://doi.org/10.1007/S10845-023-02244-8.
[75] Licardo JT, Domjan M, Orehovački T. Intelligent Robotics—A Systematic Review of Emerging Technologies and Trends. Electronics 2024, Vol 13, Page 542 2024;13:542. https://doi.org/10.3390/ELECTRONICS13030542.
[76] Dewangan S, Kumar S. AI and big data analytics revolutionizing industry 5.0: Unlocking the power of smart manufacturing and beyond. Big Data Analytics Techniques for Market Intelligence 2024:398–410. https://doi.org/10.4018/979-8-3693-0413-6.CH016.
[77] Ahmmed MS, Isanaka SP, Liou F. Promoting Synergies to Improve Manufacturing Efficiency in Industrial Material Processing: A Systematic Review of Industry 4.0 and AI. Machines 2024, Vol 12, Page 681 2024;12:681. https://doi.org/10.3390/MACHINES12100681.
[78] The State of Artificial Intelligence in 2025 n.d. https://www.baytechconsulting.com/blog/the-state-of-artificial-intelligence-in-2025 (accessed July 28, 2025).
[79] Rashid A Bin, Kausik MAK. AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. Hybrid Advances 2024;7:100277. https://doi.org/10.1016/J.HYBADV.2024.100277.
[80] AI Agents in 2025: Expectations vs. Reality | IBM n.d. https://www.ibm.com/think/insights/ai-agents-2025-expectations-vs-reality (accessed July 28, 2025).
[81] Humeniuk D, Khomh F, Antoniol G. AmbieGen: A Search-based Framework for Autonomous Systems Testing 2023.

**Journal of Information Systems Engineering and Management**
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

[82] Nedic A. Distributed Gradient Methods for Convex Machine Learning Problems in Networks: Distributed Optimization. IEEE Signal Process Mag 2020;37:92–101. https://doi.org/10.1109/MSP.2020.2975210.
[83] Houben S, Abrecht S, Akila M, Bär A, Brockherde F, Feifel P, et al. Inspect, Understand, Overcome: A Survey of Practical Methods for AI Safety. Deep Neural Networks and Data for Automated Driving: Robustness, Uncertainty Quantification, and Insights Towards Safety 2022:3–78. https://doi.org/10.1007/978-3-031-01233-4_1.
[84] Maslej N, Fattorini L, Perrault R, Gil Y, Parli V, Kariuki N, et al. Artificial Intelligence Index Report 2025 2025.