GSC Biological and Pharmaceutical Sciences
eISSN: 2581-3250 CODEN (USA): GBPSC2
Cross Ref DOI: 10.30574/gscbps
Journal homepage: https://gsconlinepress.com/journals/gscbps/

(REVIEW ARTICLE)

Check for updates

Applying Machine Learning Models for Real-Time Process Monitoring and Anomaly Detection in Pharma Manufacturing

Oluwatope R. Ojo [1,*], Oluwagbemisola Elizabeth Elesho [2], Aruna Adeniyi [3] and Ojo John Oluwadamilola [4]

[1] *Department of Mathematics and Statistics, East Tennessee State University, USA.*
[2] *Department of Biology, Georgia State University, USA.*
[3] *Department of Computer Science and Engineering, New Mexico Institute of Mining and Technology, Socorro, NM, USA.*
[4] *Osun State University Teaching Hospital, Osogbo, Nigeria.*

GSC Biological and Pharmaceutical Sciences, 2024, 27(01), 315–341

Publication history: Received on 11 March 2024; revised on 19 April 2024; accepted on 22 April 2024

Article DOI: https://doi.org/10.30574/gscbps.2024.27.1.0153

Abstract

Pharmaceutical manufacturing is an industry where precision, reliability, and compliance with stringent regulatory standards are paramount. Conventional process monitoring methods, often reliant on periodic sampling and statistical quality control, can be insufficient for capturing rapid fluctuations or unexpected deviations during production. This limitation poses significant risks, including compromised product quality, regulatory noncompliance, and increased operational costs. The integration of machine learning (ML) offers a transformative pathway, enabling real-time process monitoring and anomaly detection that supports both efficiency and safety. Machine learning models can analyze high-frequency process data streams including temperature, pressure, mixing speed, and chemical composition while continuously learning from historical trends. Algorithms such as random forests, support vector machines, and deep learning architectures provide the capacity to detect subtle variations that precede system faults or deviations from quality standards. By identifying anomalies early, these models facilitate predictive maintenance, reduce downtime, and improve batch consistency. In addition, ML-driven monitoring enhances regulatory compliance by generating explainable insights and transparent audit trails. Coupled with advanced data visualization and integration with process analytical technology (PAT), machine learning supports quality-by-design (QbD) principles, enabling pharmaceutical firms to move from reactive error correction to proactive quality assurance. The broader implication of ML adoption lies in its capacity to build resilient, adaptive manufacturing ecosystems. By uniting continuous data capture with intelligent analytics, pharmaceutical manufacturers can safeguard patient safety, reduce waste, and maintain operational continuity even under high-demand conditions. Ultimately, ML-driven real-time monitoring represents a paradigm shift toward smarter, more sustainable, and regulatory-aligned pharmaceutical production.

Keywords: Pharmaceutical Manufacturing; Machine Learning; Real-Time Monitoring; Anomaly Detection; Process Analytical Technology (PAT); Quality-By-Design (Qbd)

1. Introduction

1.1. Context: Pharma Manufacturing under Regulatory and Quality Imperatives

Pharmaceutical manufacturing operates under some of the strictest regulatory and quality frameworks of any industrial sector. Ensuring patient safety and product efficacy requires adherence to Good Manufacturing Practices (GMP), stringent quality standards, and continuous oversight by regulatory agencies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) [1]. Unlike many industries, deviations in pharmaceutical processes directly affect public health outcomes, making compliance and reliability non-negotiable.

* Corresponding author: Oluwatope R. Ojo

The complexity of modern drug formulations, particularly biologics and advanced therapies, has further intensified the demand for precise process control. Even small variations in parameters such as temperature, pH, or mixing speed can significantly affect product quality and yield [2]. The industry therefore places a premium on process robustness, real-time monitoring, and traceability.

At the same time, global market pressures and supply chain challenges demand efficiency, scalability, and cost-effectiveness. Companies must balance the dual imperatives of compliance and competitiveness, ensuring that innovations align with regulatory expectations. This has created an environment where continuous process verification, advanced monitoring, and data integrity are essential components of pharmaceutical operations [3]. These imperatives serve as the foundation for exploring how digital and data-driven innovations can reshape pharmaceutical manufacturing.

1.2. The Role of Digital Transformation in Pharma

Digital transformation has emerged as a cornerstone in the modernization of pharmaceutical manufacturing. Leveraging data analytics, automation, and connected platforms, companies are reimagining process monitoring and quality assurance to align with Industry 4.0 paradigms [4]. Central to this transformation is the use of real-time data streams from sensors and manufacturing equipment, which provide continuous insights into critical process parameters.

Machine learning, artificial intelligence (AI), and big data analytics extend these capabilities by enabling predictive modeling and anomaly detection. Rather than relying solely on retrospective quality checks, companies can now identify deviations as they occur, minimizing the risk of non-compliance or product recall [5]. Cloud-based solutions and Internet of Things (IoT) infrastructures further facilitate integration across global sites, ensuring consistency in operations and data sharing.

The regulatory landscape is also adapting. Initiatives such as the FDA's encouragement of Process Analytical Technology (PAT) frameworks illustrate how regulators view digital transformation not as a disruption but as an enabler of compliance and efficiency [6]. Beyond compliance, digitalization supports sustainability goals, reducing waste, energy consumption, and resource use. Collectively, these developments position digital transformation as a bridge between stringent regulatory imperatives and the operational agility required in today's pharmaceutical sector.

1.3. Study Objectives, Scope, and Structure

The purpose of this study is to investigate how machine learning, within the broader framework of digital transformation, can enhance process monitoring in pharmaceutical manufacturing. Specifically, the objectives are threefold: first, to define the regulatory and operational context shaping the adoption of digital tools; second, to explore the role of machine learning in improving quality, efficiency, and predictive capacity; and third, to outline a conceptual framework for integrating these technologies within existing regulatory structures [7].

The scope of the study spans both traditional small-molecule drug production and advanced modalities such as biologics and cell-based therapies, recognizing that process complexity varies across categories. The analysis emphasizes both technical innovations such as sensor integration, predictive modeling, and anomaly detection and organizational dimensions, including regulatory compliance, workforce adaptation, and infrastructure readiness.

The structure of the article is designed to progress from general context to specific contributions. Section 2 provides conceptual and theoretical foundations of machine learning in process monitoring. Section 3 analyzes case examples and industry adoption trends. Section 4 evaluates challenges and opportunities in scaling these innovations. Section 5 concludes by reflecting on broader implications for regulatory science and sustainable manufacturing.

2. Conceptual foundations of machine learning in pharma manufacturing

2.1. Fundamentals of Machine Learning in Industrial Settings

Machine learning (ML) has emerged as a critical enabler of innovation in industrial environments, providing data-driven frameworks to optimize operations, enhance productivity, and ensure quality. At its core, ML involves algorithms that learn from patterns in data to make predictions, classifications, or decisions without being explicitly programmed [6]. Unlike conventional rule-based systems, which rely on predefined thresholds or operator expertise, ML leverages large datasets to uncover complex, nonlinear relationships among process variables.

In industrial contexts, ML is particularly suited to applications where variability and uncertainty complicate process control. For example, in chemical plants, predictive maintenance models utilize ML to anticipate equipment failures based on vibration, temperature, and acoustic signals [7]. In advanced manufacturing, ML models support predictive quality assurance, integrating historical data with real-time sensor streams to identify subtle deviations that could compromise product consistency.

Key categories of ML methods deployed in industrial contexts include supervised learning, where labeled datasets are used to train models for classification or regression; unsupervised learning, which uncovers hidden structures or clusters in unlabeled data; and reinforcement learning, which optimizes sequential decision-making under dynamic conditions [8]. Each paradigm offers unique advantages for tackling industrial challenges.

The pharmaceutical industry, with its strict regulatory requirements and sensitivity to process variability, has begun to adapt these ML principles. By applying models capable of learning from multi-modal data ranging from spectroscopic signals to batch production records ML creates opportunities to balance quality imperatives with efficiency and regulatory compliance.

2.2. Data-Driven Quality Control vs. Conventional Monitoring

Traditional quality control methods in pharmaceutical manufacturing are largely retrospective. Batch release testing, for instance, involves sampling products at the end of the production cycle and performing laboratory analyses to verify compliance with quality specifications. While effective in identifying nonconforming batches, this approach is reactive, costly, and prone to delays. More importantly, it often fails to detect issues that arise in real time, allowing deviations to propagate through the process [9].

Conventional process monitoring methods typically rely on control charts and threshold-based alarms. These tools assume linearity and independence among variables, which limits their effectiveness in complex, multivariate pharmaceutical processes. For example, the interactions between temperature, pH, and agitation speed in bioreactors are nonlinear and interdependent; thus, univariate monitoring systems may overlook deviations that only emerge when variables interact [10].

By contrast, data-driven quality control powered by ML introduces predictive and adaptive capabilities. Supervised models can be trained on historical production data to predict critical quality attributes (CQAs) based on critical process parameters (CPPs). Such models allow manufacturers to intervene early, adjusting processes before deviations compromise product quality [11]. Similarly, unsupervised learning methods, such as principal component analysis or clustering algorithms, can detect anomalous trends in real time, identifying deviations that traditional univariate tools would miss.

The FDA's Process Analytical Technology (PAT) initiative has encouraged the adoption of these advanced analytics, emphasizing continuous process verification (CPV) over traditional end-product testing. Under PAT, data-driven models integrate spectroscopic measurements (e.g., NIR, Raman) with ML algorithms to monitor material properties continuously throughout production [12]. This transition reflects a broader paradigm shift: from quality-by-testing toward quality-by-design, where ML-based models not only monitor but also shape the design of robust processes.

The advantages of ML-driven quality control extend beyond efficiency. By reducing batch failures, rework, and recalls, these models improve patient safety while lowering operational costs. The capacity to provide regulators with transparent, data-supported justifications further strengthens compliance, aligning innovation with oversight expectations [13].

2.3. Machine Learning for Anomaly Detection: Definitions and Taxonomy

Anomaly detection in pharmaceutical process monitoring refers to the identification of data points, patterns, or behaviors that deviate significantly from expected norms. Such anomalies may indicate sensor malfunctions, operator errors, or early signals of process deviations that could affect product quality. Within the regulatory context, timely and accurate anomaly detection is essential, as undetected deviations may lead to non-compliance, recalls, or patient safety risks [14].

ML provides several approaches to anomaly detection, forming a taxonomy of methods suited to different data and operational environments. One broad category involves statistical methods enhanced by ML, which establish dynamic baselines of "normal" behavior and flag deviations beyond confidence intervals. Another class involves distance- and

density-based approaches, such as k-nearest neighbors or clustering, which identify outliers in high-dimensional process spaces [6].

More advanced techniques include neural networks and autoencoders, which learn compressed representations of multivariate data. By reconstructing expected patterns, these models flag deviations when reconstruction errors exceed defined thresholds. Such models are particularly valuable for pharmaceutical processes characterized by high dimensionality, nonlinearity, and limited tolerance for error [8].

Figure 1 presents a conceptual framework of ML integration in pharmaceutical process monitoring, illustrating how anomaly detection complements predictive modeling and adaptive control. By situating anomaly detection within a broader data-driven ecosystem, the figure underscores how these techniques serve as an early-warning system that enhances both compliance and efficiency.

The taxonomy of anomaly detection highlights its dual role: safeguarding quality by flagging deviations while enhancing operational resilience by identifying potential system vulnerabilities before they escalate. Within a systems-based framework, anomaly detection is not merely a monitoring tool but a cornerstone of adaptive, intelligent pharmaceutical manufacturing.

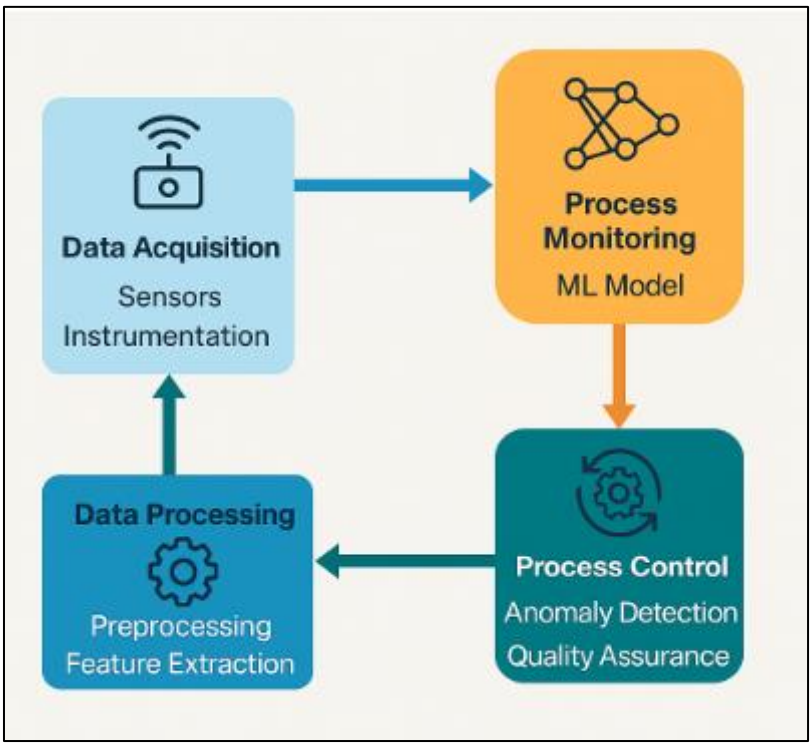

Figure 1 Conceptual framework of ML integration in pharma process monitoring.

3. Process monitoring challenges in pharmaceutical manufacturing

3.1. High Variability in Chemical and Biological Processes

Pharmaceutical manufacturing is characterized by a high degree of variability, particularly in processes involving complex chemical reactions and biologically derived products. Unlike mechanical assembly lines, chemical and biological processes are inherently nonlinear and sensitive to small fluctuations in operating conditions [12]. Temperature variations, pH shifts, dissolved oxygen levels, and raw material inconsistencies all contribute to significant process uncertainty. For example, in bioreactor cultures, even slight changes in nutrient feed rates can lead to unpredictable growth dynamics, affecting yield and product quality.

This variability challenges the traditional paradigm of process control, where fixed parameters and linear models are used to maintain consistency. Biological systems in particular are subject to stochasticity, with cell metabolism and

genetic drift producing heterogeneous outputs across production cycles [13]. Such variability complicates scale-up from laboratory to industrial production, where conditions are more difficult to standardize.

Conventional monitoring approaches struggle to adequately capture this complexity. Univariate monitoring fails to detect multivariate correlations, while traditional statistical models often oversimplify dynamic interactions. In biologics manufacturing, for instance, the interaction between agitation speed and oxygen transfer is nonlinear and dependent on cell density an interdependency that traditional monitoring rarely accounts for [14].

The stakes of process variability are high. Inconsistencies can lead to batch failures, costly recalls, or regulatory sanctions. For biologics, variability not only reduces yield but can also alter efficacy or immunogenicity, directly affecting patient safety. Addressing this challenge requires tools capable of capturing and predicting nonlinear interactions across multiple variables. Machine learning offers such potential by leveraging high-dimensional data to identify hidden patterns and dynamic dependencies that conventional tools miss [15].

3.2. Limitations of Traditional Process Analytical Technology (PAT)

Process Analytical Technology (PAT) was introduced to address variability by encouraging real-time monitoring of critical process parameters (CPPs) and critical quality attributes (CQAs). PAT relies on spectroscopic techniques, such as near-infrared (NIR) and Raman spectroscopy, combined with chemometric models to analyze data during production [16]. While this framework represented a paradigm shift away from end-product testing, its implementation has revealed significant limitations in practice.

First, traditional PAT systems often depend on linear chemometric models, such as principal component regression, that assume stable relationships between process inputs and outputs. In reality, pharmaceutical processes exhibit dynamic, nonlinear behaviors that challenge these models. When variability exceeds expected bounds such as during unexpected raw material shifts PAT models may fail to generalize, reducing their predictive reliability [17].

Second, PAT systems are resource-intensive. The installation of spectroscopic probes, calibration of models, and maintenance of complex hardware require significant investment and technical expertise. This creates barriers for smaller manufacturers, who may lack the resources to fully implement PAT across their production facilities.

Third, PAT remains largely deterministic. While it provides real-time monitoring, its capacity for proactive prediction is limited. It can identify deviations, but often lacks the adaptive learning mechanisms needed to anticipate anomalies before they emerge. As a result, interventions remain reactive rather than preventive.

Finally, integration challenges limit PAT's effectiveness. Many facilities rely on legacy systems that do not seamlessly interface with PAT tools, reducing the flow of real-time data and constraining decision-making. Moreover, regulatory acceptance of PAT models has been cautious, with agencies requiring rigorous validation before allowing them to substitute for traditional testing methods [18].

Table 1 compares traditional monitoring systems with ML-driven approaches, highlighting key differences in adaptability, scalability, and predictive capacity. The table underscores how conventional PAT remains constrained by linearity, static calibration, and high resource costs, while ML-driven systems offer dynamic, adaptive solutions better suited to today's manufacturing realities.

Table 1 Comparison of traditional monitoring vs. ML-driven approaches

Feature / Dimension	Traditional Monitoring Approaches	ML-Driven Monitoring Approaches
Data Collection	Relies on manual inspection, scheduled sampling, and basic sensors.	Utilizes continuous, high-frequency, and multi-source sensor/IoT data.
Analysis Technique	Rule-based thresholds, descriptive statistics, and human interpretation.	Predictive modeling, anomaly detection, and adaptive learning algorithms.
Response Time	Reactive; issues identified after they occur.	Proactive; anomalies detected in real-time or before escalation.

Scalability	Limited by human resources and infrastructure.	Highly scalable across large datasets and complex environments.
Accuracy & Sensitivity	Lower sensitivity, prone to false negatives/positives.	High accuracy with adaptive models, reduces false alerts.
Adaptability	Static; requires manual reconfiguration when conditions change.	Dynamic; self-updates with new data patterns (continuous learning).
Cost Implication	Lower upfront cost, higher long-term due to inefficiency and labor.	Higher initial setup cost, but cost-effective with automation and efficiency.
Knowledge Requirement	Domain expertise for manual interpretation.	Hybrid of domain knowledge and ML expertise; models learn hidden patterns.
Decision Support	Provides descriptive insights with delayed feedback.	Enables predictive and prescriptive insights for real-time decision-making.
Applications	Suitable for small-scale, stable environments.	Optimal for complex, dynamic, and large-scale systems.

3.3. Regulatory and Compliance-Driven Constraints

Regulatory and compliance requirements constitute another major constraint on pharmaceutical process monitoring. Regulatory agencies emphasize product safety, efficacy, and reproducibility, requiring stringent validation of any new technology introduced into manufacturing [12]. This cautious stance, while essential for patient safety, slows the adoption of novel monitoring approaches.

Traditional compliance models are built on deterministic systems, where process parameters are predefined and fixed within narrow ranges. These models align with conventional quality-by-testing paradigms but are less compatible with adaptive, data-driven tools like ML. For instance, ML algorithms that continuously evolve based on new data may conflict with validation requirements that demand static, fully documented models [13].

Data integrity also poses a significant regulatory challenge. Compliance frameworks such as ALCOA+ (Attributable, Legible, Contemporaneous, Original, Accurate, plus Completeness and Consistency) mandate strict standards for how data is collected, stored, and analyzed [14]. While ML thrives on large, heterogeneous datasets, regulators require transparent traceability of every data point. This tension raises questions about explainability, auditability, and accountability in ML-driven monitoring systems.

Furthermore, global regulatory heterogeneity complicates implementation. While the FDA has promoted PAT and Quality by Design (QbD), other agencies may lag in providing equivalent guidelines. Multinational manufacturers must therefore navigate inconsistent requirements, delaying adoption and limiting harmonization across sites [15].

Another barrier arises from organizational compliance cultures. Pharmaceutical companies often prioritize risk aversion, fearing regulatory penalties more than operational inefficiencies. As a result, they may hesitate to deploy ML-driven models despite evidence of superior performance. Building trust in these systems requires not only technical validation but also dialogue between regulators, industry stakeholders, and technology providers [16].

Despite these constraints, regulatory agencies are gradually opening pathways for innovation. The FDA's recent pilot programs on advanced analytics and the European Medicines Agency's exploration of digital twins for manufacturing indicate a growing willingness to integrate data-driven tools into compliance frameworks [17]. However, full adoption will depend on reconciling regulatory rigor with the adaptive nature of ML, ensuring that innovation does not compromise safety or transparency [18].

4. Machine learning models for real-time process monitoring

4.1. Supervised Learning: Regression, Classification, and Fault Detection

Supervised learning has become the cornerstone of machine learning applications in pharmaceutical process monitoring. In this paradigm, algorithms are trained on labeled datasets, mapping inputs such as sensor measurements to outputs like drug concentration, product quality grades, or fault categories [17]. Regression models are particularly

effective for predicting continuous outcomes. Partial least squares regression, for instance, has been widely used to interpret near-infrared (NIR) spectra for determining concentrations of active pharmaceutical ingredients (APIs) in real time, helping manufacturers reduce batch release delays [18].

Classification algorithms also play a critical role in differentiating between process states. Random forests and support vector machines (SVMs) are frequently deployed to distinguish between "normal" operations and "faulty" states in continuous manufacturing. These models are effective in integrating multiple process variables simultaneously, thereby detecting multivariate anomalies that linear univariate systems would overlook [19].

Fault detection systems based on supervised learning extend beyond simple classification. Predictive models trained on historical fault data anticipate breakdowns in critical equipment such as pumps, mixers, and bioreactors. By flagging potential malfunctions early, these systems reduce downtime, prevent contamination, and safeguard fragile biological cultures [20]. In biologics, predictive fault detection has proven vital for maintaining culture stability during long production cycles.

Despite these strengths, supervised learning faces significant challenges in pharmaceutical contexts. The reliance on large volumes of labeled data is a limitation, as batch failures and true anomalies are relatively rare in regulated environments. This scarcity restricts the ability of supervised models to generalize across conditions. To mitigate this, researchers increasingly rely on synthetic data generation and transfer learning strategies to augment training sets [21]. Nonetheless, supervised learning continues to provide interpretability advantages, as decision trees and regression coefficients align well with regulatory validation frameworks, making them more acceptable than opaque deep learning models.

4.2. Unsupervised Learning: Clustering and Dimensionality Reduction

Unsupervised learning is invaluable in pharmaceutical manufacturing because most process data are unlabeled, yet still contain rich patterns. Clustering techniques, such as k-means and hierarchical clustering, group data into natural clusters representing different operational states [22]. Applied to bioreactor operations, clustering has revealed hidden microenvironmental states of dissolved oxygen fluctuations, prompting timely adjustments in agitation or aeration.

Dimensionality reduction methods address another critical challenge: the high dimensionality of pharmaceutical datasets. Techniques like principal component analysis (PCA) reduce thousands of spectral variables into a few principal components, allowing easier visualization and interpretation of trends [17]. More advanced techniques such as t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) help reveal nonlinear patterns in process data, which can be critical for distinguishing between subtle product quality outcomes.

Unsupervised anomaly detection also strengthens monitoring capabilities. PCA-based approaches, including Hotelling's T^2 statistic and Q-residual analysis, flag deviations that indicate contamination, equipment drift, or human error [18]. Similarly, density-based clustering methods identify outliers representing early-stage process failures before they propagate.

Figure 2 demonstrates the workflow of ML models in pharmaceutical monitoring, highlighting how unsupervised learning contributes to the transition from raw data preprocessing to anomaly detection and eventual adaptive control. By situating clustering and dimensionality reduction within this pipeline, the figure illustrates their role as bridging mechanisms between data complexity and actionable decisions.

However, unsupervised methods present challenges in regulatory acceptance due to their lower interpretability. Unlike supervised regression, clustering outcomes are not always intuitively explainable, raising barriers for validation [23]. Despite this, their ability to extract insights from unlabeled, complex data makes unsupervised learning indispensable in modern pharmaceutical process monitoring.

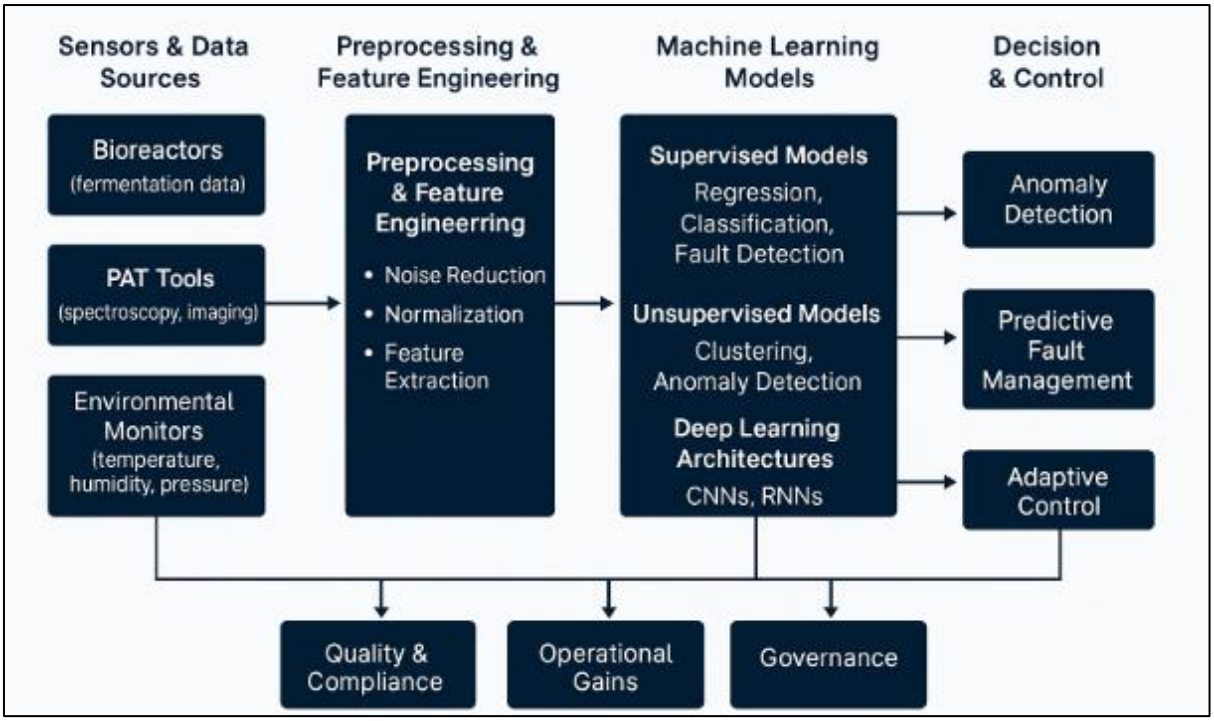

Figure 2 Workflow of ML models applied to pharma monitoring pipelines.

4.3. Deep Learning Architectures: CNNs, RNNs, and Hybrid Models

Deep learning architectures represent the most advanced and powerful tools in pharmaceutical monitoring pipelines. Convolutional neural networks (CNNs), originally built for image recognition, are now widely used in interpreting spectral datasets such as Raman or NIR signals. By learning hierarchical feature representations, CNNs can detect impurities or compositional shifts without relying on hand-engineered features [19]. These models outperform conventional chemometrics by adapting to nonlinear spectral variability.

Recurrent neural networks (RNNs), including long short-term memory (LSTM) models, excel at modeling time-dependent data. In pharmaceutical contexts, RNNs predict process dynamics such as fermentation yield hours in advance, enabling proactive interventions to stabilize outcomes [20]. By capturing temporal correlations across long sequences, LSTMs provide early warnings of deviations not visible through conventional monitoring.

Hybrid models that combine CNNs and RNNs represent a powerful frontier. For example, CNNs can extract spectral features while RNNs model their temporal progression, enabling holistic predictions of both compositional quality and process trajectory [21]. Such models are especially valuable in biologics manufacturing, where the interaction between raw material quality and dynamic culture behavior must be modeled simultaneously.

Nonetheless, deep learning faces barriers to adoption in regulated industries. Regulatory agencies remain cautious about black-box models, emphasizing the need for transparency and explainability. Emerging research in explainable AI (XAI) aims to address this by providing interpretable visualizations of how neural networks reach decisions [22]. The success of deep learning in pharma will depend not only on predictive accuracy but also on the ability to align with compliance frameworks.

Despite these challenges, the adoption of deep learning in pharma is accelerating. Advances in computational hardware, coupled with the integration of cloud platforms, make it feasible to train large-scale models on industrial datasets. These architectures are transforming pharmaceutical monitoring by enabling robust, real-time quality assurance.

4.4. Transfer Learning and Model Adaptation in Pharma Contexts

Transfer learning addresses one of the most significant challenges in applying machine learning to pharmaceutical monitoring: limited availability of labeled datasets. By leveraging pre-trained models developed in related domains, transfer learning reduces data requirements while maintaining accuracy. For example, CNNs trained on large public

spectral datasets can be fine-tuned to detect specific drug impurities in smaller batches [17]. This significantly reduces the computational and time costs of training models from scratch.

Model adaptation strategies extend this capability by adjusting models to new conditions across production sites. Domain adaptation techniques recalibrate models for variations in raw material quality, environmental conditions, or equipment setups [18]. This adaptability is essential in multinational pharmaceutical operations, where processes must remain consistent across diverse regulatory jurisdictions.

Regulatory bodies have expressed growing interest in the potential of transfer learning, provided that adaptation methods remain transparent and well-documented [23]. By supporting portability and efficiency, transfer learning ensures that ML-driven monitoring is not only technically powerful but also practically deployable within regulated environments.

5. Data infrastructure for real-time monitoring

5.1. Sensor Technologies and High-Frequency Data Collection

Pharmaceutical manufacturing relies on increasingly sophisticated sensor technologies to capture high-frequency process data. Traditional monitoring was limited to intermittent sampling; however, modern process analytical technology (PAT) emphasizes continuous sensing, enabling real-time decision-making [22]. Spectroscopic sensors, including near-infrared (NIR), Raman, and ultraviolet-visible (UV-Vis), dominate because of their ability to capture chemical composition dynamically without interrupting processes [23].

Advances in microelectromechanical systems (MEMS) have driven the miniaturization of sensors, making them cheaper and more versatile for large-scale deployment [24]. In biologics, biosensors designed to measure metabolites such as glucose, lactate, and dissolved oxygen allow near-instantaneous detection of cellular metabolic shifts, preventing deviations from optimal conditions. Similarly, acoustic sensors monitor mixing dynamics and crystallization, ensuring uniformity in solid dosage forms.

High-frequency data acquisition creates unique opportunities and challenges. On one hand, dense sensor networks generate terabytes of time-series data that provide granular visibility into process dynamics [25]. For example, high-resolution spectroscopy can reveal minor compositional fluctuations that precede quality deviations. On the other hand, the sheer scale of data poses computational challenges, requiring new architectures for storage, retrieval, and analysis.

Emerging smart sensors integrate preprocessing capabilities directly within the hardware, reducing raw data volume before transmission. This is particularly critical in continuous manufacturing, where uninterrupted sensing is necessary for predictive control [26]. Furthermore, wireless and energy-efficient sensors expand the feasibility of monitoring distributed operations across multiple production lines or facilities.

While sensor proliferation enhances data richness, it also creates risks related to calibration drift, redundancy, and false positives. Addressing these requires both robust calibration protocols and machine learning (ML) algorithms capable of distinguishing meaningful signals from background noise [27]. High-frequency sensors thus serve as the backbone of digital pharmaceutical ecosystems, but their full potential is realized only when coupled with advanced computational and analytical frameworks.

5.2. Cloud, Edge Computing, and IoT Integration

The integration of cloud and edge computing with the Internet of Things (IoT) is reshaping pharmaceutical data infrastructures. IoT-enabled devices facilitate continuous, multi-point data acquisition, while cloud platforms provide scalable storage and computational capacity [23]. This architecture supports advanced ML applications by aggregating high-frequency sensor data into centralized environments for training predictive models.

Edge computing complements this by enabling real-time analytics at or near the data source. In pharmaceutical contexts, latency is critical; milliseconds can separate safe product trajectories from deviations that compromise batch integrity [24]. Edge systems process data locally, allowing faster responses to anomalies while reducing bandwidth requirements for transmitting massive datasets to centralized servers. For instance, edge-enabled bioreactors can autonomously adjust feed rates based on oxygen consumption patterns without relying on cloud delays [25].

Cloud computing further enhances interoperability. Modern platforms allow integration of process control systems, enterprise resource planning (ERP), and regulatory documentation. This ensures that data collected across global manufacturing sites is standardized and accessible for compliance audits [26]. Furthermore, distributed ledger technologies, such as blockchain, are increasingly integrated into cloud systems to ensure immutable data trails, enhancing trust and transparency for regulators.

A key advantage of IoT integration is cross-device communication. For example, spectroscopic sensors, pH meters, and biosensors can feed into a shared IoT framework, enabling holistic process monitoring. Data fusion across these modalities allows ML algorithms to extract correlations otherwise invisible in siloed datasets [27]. Such integration underpins adaptive quality control, where interventions are guided by comprehensive, real-time data ecosystems.

However, these benefits come with cybersecurity and governance challenges. Pharmaceutical manufacturing is a high-value target for cyberattacks, and IoT proliferation increases the attack surface. Ensuring secure communication protocols, access controls, and anomaly detection at both edge and cloud levels is essential [28]. Ultimately, the combined deployment of cloud, edge, and IoT infrastructures transforms pharmaceutical data collection from a fragmented system into a cohesive, intelligent network capable of supporting predictive and adaptive monitoring.

5.3. Data Preprocessing: Noise Reduction, Normalization, and Feature Engineering

Raw pharmaceutical process data is often noisy, heterogeneous, and high-dimensional, requiring rigorous preprocessing before ML can extract meaningful insights. Noise arises from sensor drift, environmental fluctuations, and random measurement errors. Techniques such as moving average filters, Savitzky–Golay smoothing, and wavelet transforms are commonly applied to spectroscopic signals to enhance signal-to-noise ratios [22]. These approaches ensure that anomalies reflect genuine process deviations rather than sensor artifacts.

Normalization is equally critical. Data collected from different instruments, scales, or operating units must be standardized to enable integration. Min-max scaling, z-score normalization, and logarithmic transformations are frequently used to harmonize sensor data across modalities [23]. For example, combining NIR absorbance values with electrochemical sensor outputs requires scaling strategies to prevent one modality from dominating multivariate models.

Feature engineering transforms raw signals into meaningful representations for ML models. Domain expertise informs the selection of features such as peak intensities in spectroscopy or lagged variables in time-series data. Automated feature extraction, facilitated by ML, is increasingly popular, but hybrid approaches that combine expert-driven and data-driven methods yield the most robust performance [24].

Table 2 outlines common pharma process sensors spectroscopic, electrochemical, thermal, and acoustic and the corresponding data parameters they generate. The table illustrates the diversity of raw inputs requiring preprocessing, from high-dimensional spectra to low-frequency pH readings. This diversity underscores the necessity of preprocessing as a unifying step in pharmaceutical monitoring pipelines [25].

Preprocessing also involves outlier detection. Statistical techniques such as Mahalanobis distance and density-based clustering identify anomalous data points that may reflect equipment malfunctions or contamination. Additionally, dimensionality reduction methods like principal component analysis (PCA) simplify multivariate datasets while retaining the essential variance needed for classification or regression tasks [26].

Finally, preprocessing must be aligned with regulatory requirements. Traceability and auditability of preprocessing pipelines are essential for validation. Regulators increasingly expect that preprocessing methods be documented, reproducible, and consistent across batches [27]. Without rigorous preprocessing, even the most advanced ML algorithms are susceptible to errors, overfitting, or misclassification, reducing their utility in regulated environments [28].

Table 2 Types of pharmaceutical process sensors and monitored data parameters

Sensor Type	Data Parameters Measured	Application in Pharma Processes
Temperature Sensors (e.g., thermocouples, RTDs)	Temperature, heat flow	Monitoring bioreactors, sterilization, storage stability
Pressure Sensors	Pressure, vacuum levels	Filtration, reactor pressure control, lyophilization
pH Sensors	Hydrogen ion concentration, acidity/alkalinity	Fermentation, buffer preparation, bioprocess monitoring
Dissolved Oxygen (DO) Sensors	Oxygen concentration in liquids	Bioreactor aeration, cell culture optimization
Conductivity Sensors	Ionic strength, salt concentration	Chromatography, purification, water quality control
Optical / NIR Sensors	Absorbance, turbidity, chemical signatures	Process analytical technology (PAT), in-line quality testing
Spectroscopic Sensors (Raman, FTIR, UV)	Molecular structure, chemical composition	Drug formulation monitoring, raw material verification
Flow Sensors	Flow rate, volumetric throughput	Continuous manufacturing, dosing, mixing
Mass/Weight Sensors	Mass, weight distribution	Tablet pressing, powder filling, packaging consistency
Biosensors	Biomarker levels, enzymatic activity	Quality control of biologics, contamination detection

6. Machine learning for anomaly detection in pharma

6.1. Statistical vs. ML-Based Anomaly Detection Approaches

Anomaly detection has long been central to pharmaceutical quality assurance, but the methodologies employed have evolved substantially. Traditional statistical approaches rely on hypothesis testing, control charts, and univariate thresholds to detect deviations from expected process behavior [27]. For instance, Shewhart charts and cumulative sum (CUSUM) methods are widely used to identify trends or shifts in process parameters such as pH, temperature, or concentration levels. These approaches are relatively simple, computationally efficient, and align well with regulatory demands for transparency.

However, the complexity of modern pharmaceutical manufacturing particularly in biologics has highlighted their limitations. Statistical models often assume independence and normality of variables, yet real-world processes involve nonlinear and multivariate interactions. For example, deviations in fermentation yield may not be apparent in single-variable charts but emerge only when multiple factors such as dissolved oxygen, nutrient consumption, and agitation speed interact [28].

Machine learning (ML)-based approaches address these gaps by learning from high-dimensional datasets without relying on rigid distributional assumptions. Supervised ML models, such as support vector machines or random forests, classify process states into “normal” or “faulty” categories based on historical labeled data. In contrast, unsupervised methods, including clustering and principal component analysis (PCA)-based monitoring, identify anomalies in unlabeled data by detecting deviations from established multivariate patterns [29].

The shift from statistical to ML-based anomaly detection represents a paradigm change. While statistical tools provide interpretability and regulatory acceptance, ML approaches enhance sensitivity, enabling earlier and more accurate detection of subtle deviations. Importantly, hybrid frameworks are emerging that combine the strengths of both methods—leveraging statistical baselines for compliance while employing ML for higher-dimensional anomaly

detection [30]. Such hybridization underscores the ongoing convergence of interpretability and predictive performance in pharmaceutical monitoring.

6.2. Time-Series Anomaly Detection Models for Continuous Processes

Continuous manufacturing has amplified the importance of time-series anomaly detection. Unlike batch processes, where quality deviations can be retrospectively addressed, continuous systems require real-time identification of anomalies to prevent propagation across the entire production line [31].

Classical time-series models such as autoregressive integrated moving average (ARIMA) and Kalman filters have historically been used to predict and detect deviations. However, their reliance on linear assumptions limits their applicability to the nonlinear, dynamic environments of pharmaceutical processes [27].

ML-based models provide a more robust alternative. Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) networks, excel at capturing long-range dependencies in temporal data, making them well-suited for monitoring fermentation and crystallization dynamics [28]. By learning sequential correlations, LSTMs detect subtle anomalies hours before they escalate into critical deviations. Convolutional neural networks (CNNs) adapted for time-series data also enhance anomaly detection by extracting local temporal patterns from high-frequency sensor streams [29].

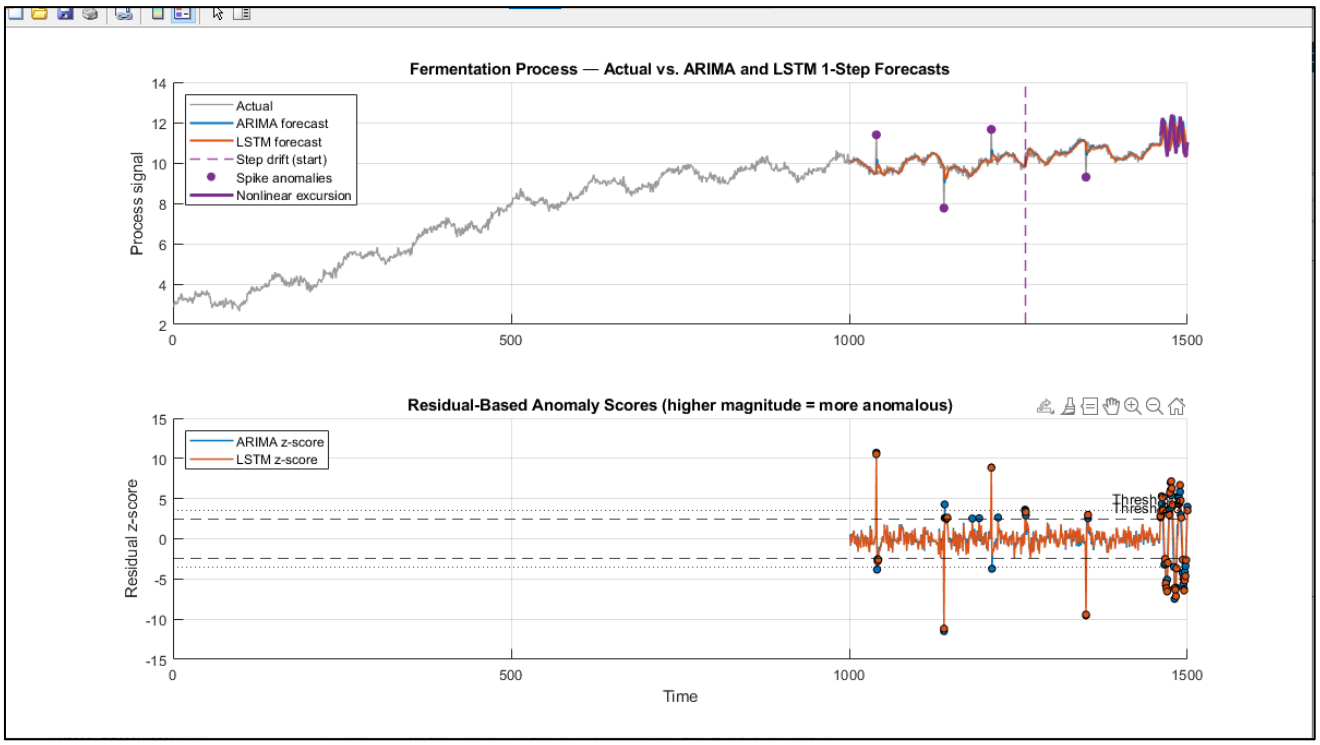

Figure 6 Fermentation Process: ARIMA vs. LSTM Forecasting [27]

Figure 6 illustrates a comparison between ARIMA and LSTM forecasts on a synthetic fermentation process signal with injected anomalies. While ARIMA provides short-term forecasts, its linear assumptions limit anomaly detection accuracy, particularly for nonlinear deviations. In contrast, the LSTM model demonstrates superior sensitivity, flagging subtle deviations earlier and maintaining predictive stability across dynamic conditions. Detected anomalies are highlighted, showing that LSTM can reduce response delays in continuous monitoring environments.

Hybrid deep learning models that integrate CNNs with RNNs combine local and long-term sequence modeling, improving robustness across diverse temporal horizons [30]. These models outperform traditional time-series forecasting in applications such as detecting contamination in biologics or predicting crystallization defects in solid dosage forms.

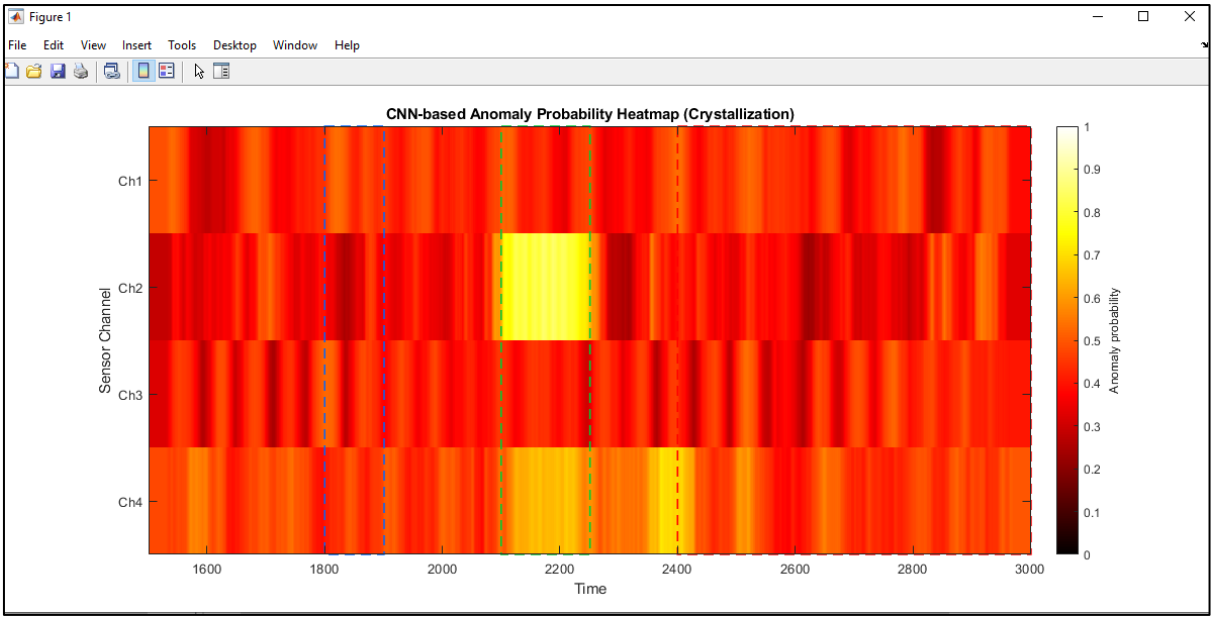

Figure 7 CNN Anomaly Probability Heatmap (Crystallization) [27]

Figure 7 shows a CNN-based anomaly detection heatmap applied to multivariate crystallization signals. The heatmap visualizes anomaly probabilities across four sensor channels over time, where localized spikes correspond to injected contaminations and dynamic disturbances. By leveraging convolutional feature extraction, CNN models capture local temporal correlations more effectively than linear filters. The probability hotspots align closely with known anomaly intervals, demonstrating the model's ability to isolate channel-specific deviations in real time.

A key challenge remains interpretability. Regulators often demand clear justification for anomaly detection outputs, which deep models rarely provide. To address this, researchers have proposed explainable anomaly detection systems that visualize which time-series segments contribute to model predictions [32]. These advances improve regulatory acceptance by offering transparency while preserving predictive accuracy.

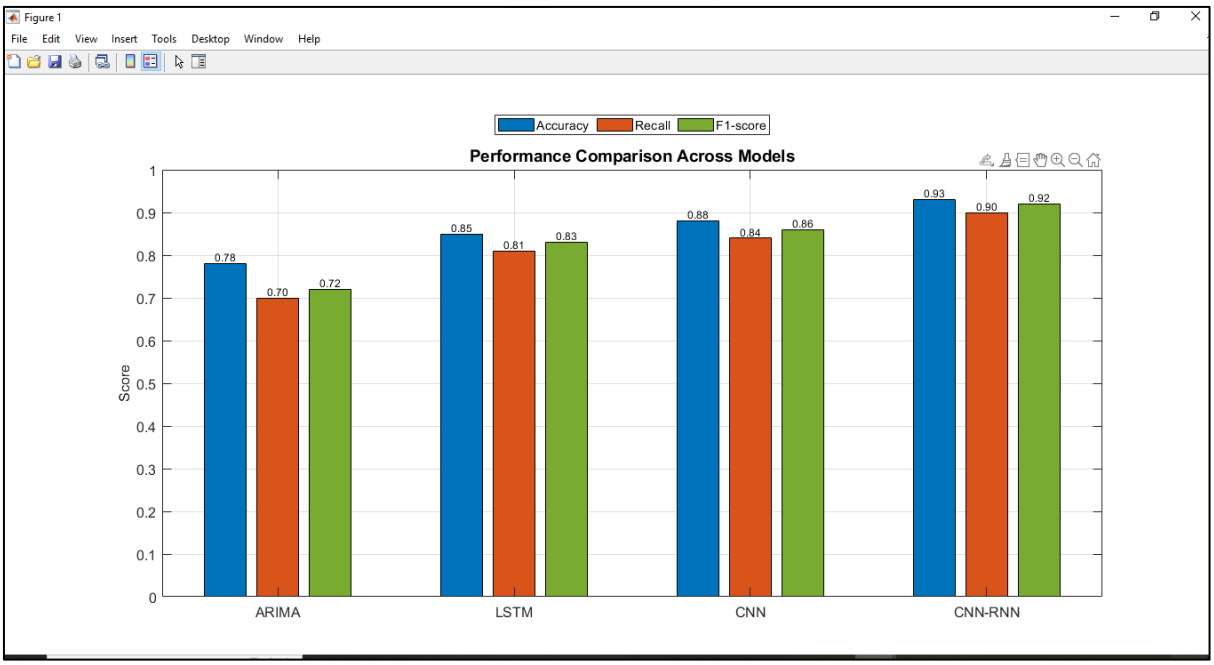

Figure 8 Performance Comparison Across Models [27]

Figure 8 compares the performance of ARIMA, LSTM, CNN, and a hybrid CNN–RNN model on anomaly detection tasks. Metrics include accuracy, recall, and F1-score, computed against ground-truth anomaly labels. The results show that hybrid architectures consistently outperform single-model baselines, combining the strengths of CNNs (local pattern detection) and RNNs (long-range sequence learning). This demonstrates the potential of integrated deep learning frameworks for robust anomaly detection in pharmaceutical monitoring pipelines.

Time-series anomaly detection thus represents a critical juncture for pharmaceutical monitoring, balancing the need for real-time precision with the stringent compliance requirements of regulated manufacturing [32].

6.3. Early Warning Systems and Predictive Fault Management

Early warning systems (EWS) extend anomaly detection by predicting potential failures before they occur. In pharmaceutical contexts, predictive fault management is crucial for minimizing downtime, reducing batch failures, and enhancing patient safety [28].

Traditional EWS relied on threshold-based alarms that flagged deviations once they exceeded predefined limits. While useful, such systems often generated excessive false alarms, desensitizing operators to genuine risks. ML-based predictive systems overcome this by learning from historical failure data to distinguish between benign fluctuations and precursors of genuine faults [29].

For example, supervised models trained on multivariate sensor data can identify signatures of equipment wear, predicting failures in pumps or mixing systems before they manifest physically. Similarly, unsupervised anomaly detection applied to spectroscopic data can signal early signs of contamination during biologics production [30]. By providing advanced notice, these systems allow operators to implement corrective actions such as adjusting feed rates, calibrating sensors, or isolating affected lines.

Predictive fault management also extends to supply chain resilience. By monitoring logistics data streams such as temperature, humidity, and vibration during drug transport ML-based EWS can prevent degradation of sensitive biologics [31]. This holistic approach ensures that quality is safeguarded across the entire production-to-delivery continuum.

Integration with digital twins further enhances predictive capabilities. Virtual replicas of manufacturing systems allow EWS to simulate fault scenarios and evaluate mitigation strategies in silico, reducing the need for costly physical interventions [32]. Regulatory bodies have begun recognizing the potential of such systems, provided that validation protocols ensure their reliability.

Ultimately, predictive fault management shifts pharmaceutical quality assurance from a reactive paradigm to a proactive one. By preventing rather than merely detecting anomalies, EWS ensure not only operational efficiency but also enhanced compliance with patient safety imperatives [33].

6.4. Case of Batch vs. Continuous Manufacturing Anomalies

Anomaly detection presents unique challenges in batch versus continuous pharmaceutical manufacturing. In batch processes, anomalies are often identified retrospectively through end-product testing or periodic sampling. While this provides opportunities for post-hoc correction, it also results in costly rework or product loss when deviations are discovered late [27].

Continuous manufacturing, by contrast, requires real-time anomaly detection because deviations propagate immediately through the process. The high-frequency data streams generated necessitate advanced ML systems capable of handling large-scale time-series data in real time [28].

Figure 3 illustrates the anomaly detection pipeline within pharmaceutical systems, emphasizing how ML models integrate with sensor networks, preprocessing modules, and predictive control frameworks. This pipeline highlights the critical differences in anomaly detection workflows between batch and continuous systems.

Hybrid strategies are increasingly employed to bridge these paradigms. Statistical process control remains valuable for batch environments, while ML-based time-series models dominate in continuous operations. Together, they represent complementary approaches that ensure robust quality assurance across manufacturing modes [29].

As regulatory agencies increasingly encourage continuous manufacturing, the need for advanced anomaly detection pipelines will only grow. Integrating these pipelines into both batch and continuous contexts ensures flexibility, compliance, and scalability across the industry [31].

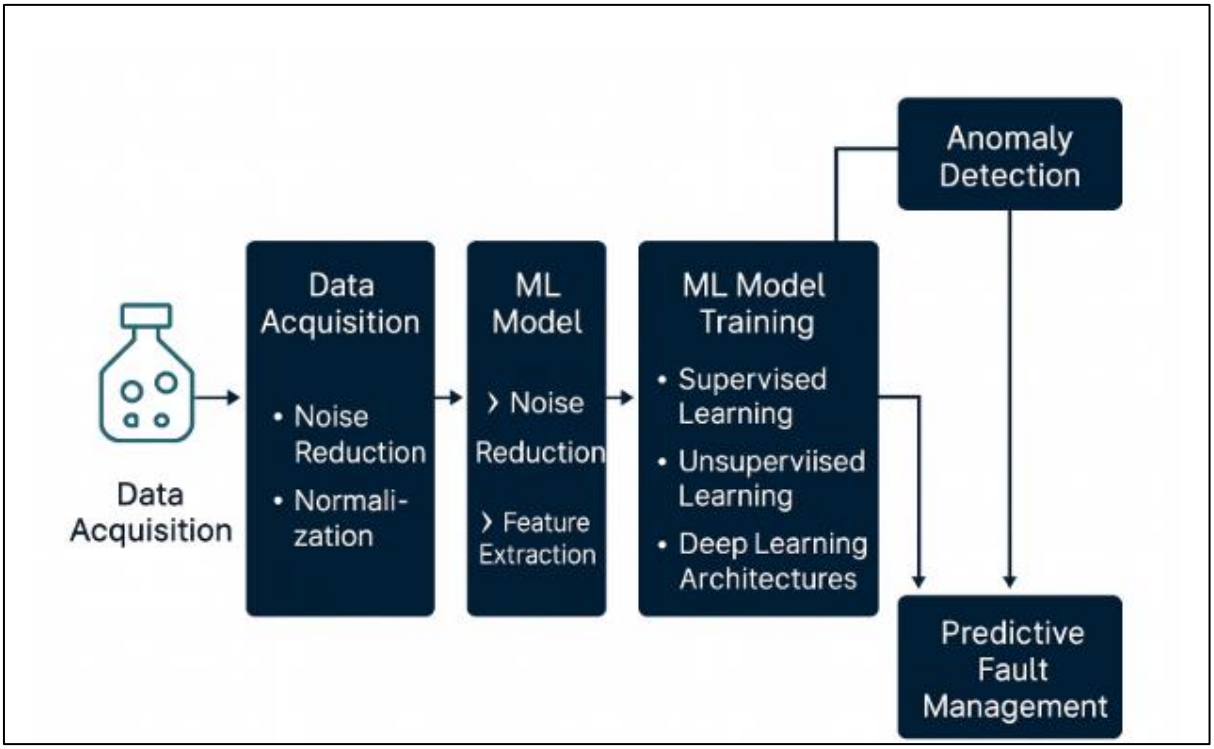

Figure 3 Anomaly detection pipeline with ML in pharma systems**.**

7. Case studies and applications

7.1. Biopharmaceutical Production: Monitoring Fermentation Processes

Biopharmaceutical production represents one of the most complex areas of pharmaceutical manufacturing due to the reliance on living cells as production hosts. Fermentation processes are highly sensitive to environmental fluctuations, making them ideal candidates for machine learning (ML)-based monitoring systems. Traditional monitoring often relied on sparse offline sampling of metabolites such as glucose, lactate, or ammonia; however, ML frameworks can integrate real-time sensor data streams to predict process trajectories with unprecedented accuracy [32].

Spectroscopic sensors, including near-infrared (NIR) and Raman, generate continuous profiles of nutrient concentrations, dissolved oxygen, and pH. When coupled with supervised learning algorithms such as support vector machines (SVMs), these data streams enable early detection of metabolic shifts that may compromise yield or product quality [33]. Long short-term memory (LSTM) networks, a form of recurrent neural network (RNN), are particularly powerful for predicting deviations hours before they occur by capturing temporal dependencies across fermentation cycles [34].

Hybrid digital twins are increasingly applied in fermentation monitoring, combining mechanistic models of cellular metabolism with data-driven ML approaches. These systems simulate hypothetical process deviations and test corrective strategies in silico, significantly reducing risk in live production [35]. For instance, digital twins supported by ML anomaly detection can anticipate oxygen limitation events and recommend agitation adjustments before culture viability declines.

A growing body of research emphasizes that ML not only enhances real-time monitoring but also supports adaptive control loops. Reinforcement learning has been used to optimize feeding strategies by dynamically adjusting glucose concentrations, leading to more stable yields [36]. This adaptability is especially crucial in monoclonal antibody (mAb) production, where maintaining consistent glycosylation patterns is critical for therapeutic efficacy.

Despite these advances, regulatory barriers remain. Authorities demand transparent models that provide explainable outputs, particularly when they influence high-stakes interventions. Researchers are therefore embedding interpretability frameworks into ML fermentation models, ensuring alignment with regulatory validation practices [37]. As the field matures, ML-driven fermentation monitoring is set to redefine process robustness, product consistency, and compliance in biopharmaceutical manufacturing.

7.2. Solid Dosage Manufacturing: Granulation, Blending, and Tablet Pressing

Solid dosage manufacturing, which includes granulation, blending, and tablet pressing, is the largest sector of pharmaceutical production and has been a major beneficiary of ML adoption. Historically, variability in particle size distribution, blending uniformity, and compression force have been difficult to monitor in real time. ML has emerged as a transformative tool to overcome these limitations [33].

Spectroscopic PAT tools, such as NIR and Raman, capture high-frequency signals that reflect powder homogeneity and granule moisture content. When combined with multivariate regression models, these tools predict end-product quality attributes like hardness and dissolution rate in real time [34]. Furthermore, unsupervised clustering methods such as k-means are employed to detect anomalies in blending uniformity, flagging areas of segregation before tablets are pressed [35].

In tablet pressing, ML models trained on vibration and acoustic sensor data detect early signs of punch wear or sticking, allowing preventive maintenance before costly downtime occurs [36]. Convolutional neural networks (CNNs) are increasingly applied to image-based quality control, automating defect detection such as capping, lamination, or coating irregularities at speeds surpassing manual inspection [37].

Table 3 presents case studies of ML applications across granulation, blending, and compression, highlighting the diverse algorithms employed, data sources leveraged, and quality outcomes achieved. The table underscores how ML solutions not only reduce variability but also integrate seamlessly with continuous quality verification requirements [38].

The integration of ML into solid dosage manufacturing pipelines aligns with regulatory encouragement for continuous manufacturing and real-time release testing (RTRT). These advances represent a paradigm shift, moving from post-production quality testing toward in-line, data-driven assurance of product consistency.

7.3. Continuous Manufacturing Lines: ML-Driven PAT Integration

Continuous manufacturing lines are increasingly seen as the future of pharmaceutical production due to their potential for efficiency, scalability, and reduced variability. A central feature of these systems is the integration of process analytical technology (PAT) with ML frameworks to deliver adaptive, real-time quality assurance [32].

Unlike batch operations, continuous systems demand uninterrupted monitoring, where deviations can propagate rapidly. ML enables predictive surveillance by continuously analyzing multi-sensor data streams. Supervised algorithms, such as random forests, predict quality outcomes based on patterns across temperature, pressure, and concentration sensors [33]. More advanced approaches, including ensemble learning, combine multiple models to improve anomaly detection robustness [34].

Deep learning models are particularly suited for handling the complexity of continuous data. CNNs process spectroscopic datasets to monitor crystallization and drying, while recurrent architectures capture time-series dependencies across process stages [35]. Hybrid CNN-LSTM frameworks have been demonstrated to identify subtle shifts in granulation moisture that precede downstream compression variability, enabling corrective actions before final product deviations occur [36].

An additional innovation is the use of reinforcement learning in closed-loop control of continuous systems. These models dynamically adjust process parameters—such as screw speed in extrusion or airflow in drying to maintain optimal performance despite raw material variability [37]. This adaptability ensures compliance with critical quality attributes (CQAs) while minimizing operator intervention.

ML-driven PAT integration is also redefining regulatory compliance. Continuous verification models provide real-time evidence of quality, aligning with FDA's process validation guidance and European Medicines Agency (EMA) frameworks [38]. As adoption accelerates, continuous manufacturing empowered by ML will likely become a global standard, reducing cost, enhancing efficiency, and ensuring consistency.

From biopharmaceutical fermentation monitoring to solid dosage quality assurance and ML-enhanced continuous manufacturing lines, the evidence demonstrates that machine learning is rapidly transitioning from experimental research to mainstream industrial application. Table 3 illustrates concrete examples of these advances, while sectoral narratives confirm that ML enhances precision, consistency, and compliance across pharmaceutical domains. The following section transitions toward governance, exploring how regulatory bodies and industry leaders are shaping the frameworks that will determine the long-term integration of ML into pharmaceutical systems.

Table 3 Case studies of ML applications in pharmaceutical manufacturing

Case Study / Context	ML Technique Applied	Objective / Outcome
Bioreactor Monitoring (Cell Culture)	Supervised learning (Random Forest, SVM)	Improved prediction of cell growth rates and product yields, enabling early anomaly detection.
Drug Tablet Quality Control	Computer vision with Convolutional Neural Networks (CNNs)	Automated defect detection in tablets, reducing reliance on manual inspection.
Continuous Manufacturing	Reinforcement learning & adaptive control	Optimization of process parameters in real time to maintain quality within tight specifications.
Supply Chain Optimization	Predictive analytics using Gradient Boosting	Forecasting demand fluctuations to reduce stock-outs and overproduction.
Spectroscopic Data Analysis (PAT)	Deep learning (Autoencoders)	Enhanced detection of chemical composition deviations, ensuring batch consistency.
Sterilization & Clean Room Monitoring	Anomaly detection algorithms (Isolation Forest, kNN)	Early identification of contamination risks through sensor fusion.
Vaccine Production	Hybrid ML models integrating process and clinical trial data	Increased yield prediction accuracy, reducing waste and variability.

8. Governance, regulation, and ethical considerations

8.1. Regulatory Compliance (FDA, EMA, ICH Q8–Q12) and ML Models

Regulatory compliance is a cornerstone of pharmaceutical manufacturing, particularly in integrating machine learning (ML) within highly controlled environments. Agencies such as the U.S. Food and Drug Administration (FDA), the European Medicines Agency (EMA), and the International Council for Harmonisation (ICH) set the frameworks that govern risk-based approaches to process validation. The ICH guidelines Q8 through Q12 emphasize quality-by-design (QbD), lifecycle management, and continuous verification, principles that align naturally with the adaptive capacity of ML models [36].

However, tension remains between regulatory conservatism and ML's rapid innovation cycle. Traditional validation procedures assume deterministic, rule-based systems, while ML relies on probabilistic inference and adaptive learning. To reconcile this, regulators increasingly require hybrid validation pipelines, combining statistical baselines with algorithmic traceability [37]. For instance, FDA's Emerging Technology Program has reviewed ML-enabled process analytical technology (PAT) systems, emphasizing reproducibility, documentation, and transparency.

Furthermore, regulators expect ML models to demonstrate generalizability across batches, sites, and product types. This requires robust cross-validation and stress-testing frameworks. Compliance challenges are compounded when models evolve over time through retraining, which raises questions about whether updates constitute new validation events [38]. Despite these hurdles, regulatory guidance is gradually adapting to ensure ML integration strengthens, rather than compromises, pharmaceutical quality assurance.

8.2. Transparency, Explainability, and Auditability in Pharma ML

Transparency and explainability are critical for regulatory acceptance of ML in pharmaceutical settings. While deep learning models offer high predictive accuracy, their "black-box" nature is often misaligned with the auditability requirements of regulatory agencies [39]. Interpretability is particularly important in areas such as biopharmaceutical

fermentation and solid dosage monitoring, where interventions based on opaque model outputs could have significant consequences.

Researchers are increasingly embedding explainable AI (XAI) techniques, such as feature attribution and saliency mapping, into pharmaceutical ML pipelines. These methods reveal which input variables such as specific spectral peaks or time-series fluctuations drive model predictions [40]. By providing interpretable visualizations, XAI improves operator trust and facilitates regulator inspections.

Auditability extends beyond interpretability to full lifecycle documentation. Every stage of model development from data preprocessing to feature engineering, training, validation, and deployment must be traceable and reproducible [41]. Tools like blockchain-based audit trails are being explored to guarantee immutability in ML-driven monitoring. This aligns with regulatory emphasis on data integrity, particularly in continuous manufacturing environments where real-time quality assurance must be verifiable.

Importantly, transparency and auditability also strengthen stakeholder trust beyond regulators. By ensuring that ML-driven systems are explainable, pharmaceutical companies safeguard their reputations while advancing adoption of innovative digital quality systems that build upon evidence-based case studies such as those presented in Table 3 [42].

8.3. Ethical Challenges: Bias, Privacy, and Patient Safety

Beyond compliance and transparency, ethical challenges in ML integration within pharma are increasingly prominent. Algorithmic bias is a significant concern, particularly where training datasets may not fully capture variability across different production lines, geographic facilities, or raw material sources [36]. A biased model might overfit to one context, producing unreliable predictions in another, thereby undermining both quality and equity.

Figure 4 Governance–technology–outcomes model for ML in pharma.

Privacy concerns arise in biopharmaceutical contexts where ML is integrated with patient-centric datasets, such as genomics or personalized dosing information. The need to balance predictive accuracy with confidentiality has spurred adoption of privacy-preserving ML methods, including federated learning and differential privacy [37]. These techniques allow collaborative model development across institutions without sharing sensitive patient-level data.

Patient safety is the ultimate ethical imperative. A failure in anomaly detection or predictive monitoring could allow unsafe batches to progress through production, with dire clinical consequences [38]. Ethical frameworks thus demand redundancy in ML systems ensuring that algorithmic decisions are cross-checked against statistical baselines and human oversight [47].

Figure 4 conceptualizes governance, technology, and outcomes as an integrated model, highlighting how compliance, transparency, and ethics intersect to determine ML's contribution to pharmaceutical systems. Ensuring these safeguards is essential to transform ML from a technological innovation into a reliable pillar of public health protection [48].

9. Future directions in ml-driven pharma manufacturing

9.1. Reinforcement Learning and Adaptive Process Control

Reinforcement learning (RL) has emerged as a powerful approach to adaptive process control in pharmaceutical manufacturing, offering dynamic optimization of highly variable environments [49]. Unlike static statistical or rule-based control systems, RL agents learn policies through iterative interaction with manufacturing processes, progressively improving decisions to maximize defined performance rewards [50]. In fermentation, RL has been used to optimize nutrient feeding and aeration rates by predicting downstream yield impacts, adjusting parameters in real time to maintain cellular health and productivity [51].

What distinguishes RL from conventional model-predictive control is its ability to adapt continuously without exhaustive mechanistic modeling. This flexibility is particularly valuable in biologics, where raw material variability and nonlinear metabolic pathways complicate deterministic control [42]. Moreover, hybrid RL frameworks integrate mechanistic priors with data-driven policies, enhancing both regulatory acceptability and operational robustness [52].

Research also emphasizes the scalability of RL, as it can govern multiple stages simultaneously fermentation, purification, and formulation while accounting for interdependencies between upstream and downstream processes [53]. By enabling learning-driven adaptability, RL offers a path toward fully autonomous manufacturing environments where quality, efficiency, and compliance coexist dynamically, extending beyond optimization into real-time resilience and fault tolerance [54].

9.2. Federated Learning and Privacy-Preserving Analytics

Federated learning (FL) addresses a critical barrier in pharmaceutical ML adoption: the tension between data-sharing needs and privacy obligations. Traditional ML pipelines require centralized data aggregation, raising concerns over intellectual property, patient confidentiality, and inter-firm competition. FL circumvents this by enabling collaborative model training across decentralized datasets without exchanging raw information [45].

In pharmaceutical contexts, FL has been applied to cross-site process monitoring, where multiple facilities contribute local gradients that update a shared global model. This allows industry-wide learning without compromising sensitive operational or patient-level data [46]. Privacy-preserving enhancements such as differential privacy and secure multiparty computation further safeguard proprietary inputs, aligning with GDPR and HIPAA regulations while ensuring analytical depth [47].

The potential of FL extends to clinical–manufacturing integration, where models link patient biomarker datasets with production parameters to enable precision biopharmaceutical development. By protecting data integrity across stakeholders, FL expands the scope of ML while addressing ethical and legal imperatives.

Figure 5 depicts the roadmap of emerging technologies, situating FL alongside reinforcement learning and digital twins as part of a convergent technological ecosystem for pharmaceutical manufacturing. By ensuring privacy-preserving collaboration, FL enables innovation without sacrificing compliance or trust [48].

9.3. Digital Twins for Predictive Manufacturing

Digital twins (DTs) are rapidly transforming pharmaceutical manufacturing by providing virtual replicas of physical processes that integrate real-time sensor data with predictive models. In contrast to traditional simulations, DTs evolve continuously as they ingest live data, enabling both monitoring and intervention in highly dynamic manufacturing environments [40].

When combined with ML, DTs support predictive maintenance by identifying wear patterns in pumps, valves, and filtration units before failures occur. They also enhance anomaly detection, allowing operators to test corrective actions virtually before deploying them to production systems [41]. In biologics, DTs simulate metabolic dynamics under varying feed conditions, ensuring robustness in monoclonal antibody production [42].

From a regulatory standpoint, DTs complement risk-based validation frameworks by providing audit-ready evidence of process predictability [43]. Their utility extends across scale-up operations, where pilot plant learnings are transferred seamlessly to commercial production without repeating costly empirical trials [44].

Moreover, DTs integrate with RL and FL, creating multi-layered intelligent systems capable of autonomous adaptation. As shown in Figure 5, DTs anchor the emerging technologies roadmap by bridging real-time monitoring with predictive control. Their adoption is poised to accelerate as industry leaders seek holistic solutions for resilient, adaptive, and compliant pharmaceutical ecosystems [46].

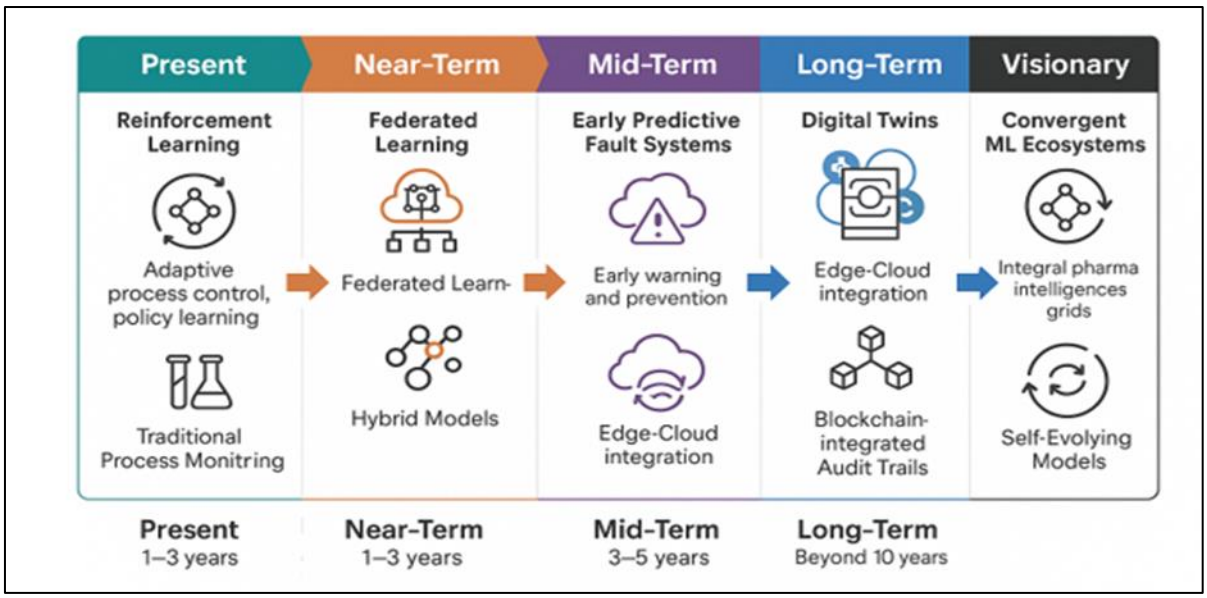

Figure 5 Emerging technologies roadmap for ML in pharma**.**

10. Conclusion

The application of machine learning (ML) in pharmaceutical monitoring and anomaly detection has fundamentally reshaped how the industry envisions quality assurance, resilience, and adaptability. Across fermentation processes, solid dosage manufacturing, and continuous production lines, ML has enabled real-time monitoring and predictive oversight that were previously unattainable through traditional statistical or mechanistic tools. By harnessing sensor-rich environments and advanced algorithms, the pharmaceutical sector is now able to anticipate deviations, implement corrective measures before failures occur, and maintain tighter control over complex and sensitive processes.

A central contribution of this study has been to highlight the systems-based nature of ML integration. Rather than being viewed solely as computational tools, ML models function as part of a broader ecosystem comprising data streams, governance frameworks, and institutional safeguards. Sensors and high-frequency data collection create the informational backbone, algorithms transform raw signals into actionable insights, and governance structures ensure these insights align with compliance, transparency, and ethical expectations. This systemic perspective emphasizes that resilience in pharmaceutical production cannot be achieved through technology alone; it requires the harmonization of data, models, and oversight mechanisms.

The emphasis on resilience, efficiency, and regulatory alignment underscores the transformative role ML plays in supporting public health outcomes. Resilience is achieved when ML-driven systems adapt dynamically to variability in raw materials, environmental fluctuations, and evolving production demands. Efficiency emerges from predictive maintenance, optimized resource utilization, and reduced waste, allowing manufacturers to sustain competitiveness

while minimizing ecological footprints. Finally, regulatory alignment ensures that these innovations reinforce, rather than undermine, the trust that regulators and patients place in pharmaceutical systems.

The convergence of machine learning with reinforcement learning, federated learning, and digital twins suggests that the trajectory of pharmaceutical manufacturing will be increasingly characterized by intelligence, autonomy, and connectivity. Yet, the ultimate value of these technologies lies not only in technical performance but in their contribution to a resilient, transparent, and ethically governed pharmaceutical ecosystem. By integrating advanced models into the core of production while safeguarding accountability and patient safety, the sector positions itself to navigate future uncertainties with confidence.

This synthesis therefore reinforces the central message: machine learning, when embedded within a systems-based framework of data, governance, and trust, stands as a cornerstone of next-generation pharmaceutical resilience.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Mohammad Fuad II, Ngah Demon MF, Husni H. Automated real time anomaly detection model for operation and production data at scale. InAbu Dhabi International Petroleum Exhibition and Conference 2020 Nov 9 (p. D012S116R179). SPE.

[2] Ahmed CM, MR GR, Mathur AP. Challenges in machine learning based approaches for real-time anomaly detection in industrial control systems. InProceedings of the 6th ACM on cyber-physical system security workshop 2020 Oct 6 (pp. 23-29).

[3] Saha GC, Eni LN, Saha H, Parida PK, Rathinavelu R, Jain SK, Haldar B. Artificial intelligence in pharmaceutical Manufacturing: Enhancing quality control and decision making. Rivista Italiana di Filosofia Analitica Junior. 2023;14(2):2023.

[4] Menaama Amoawah Nkrumah. HIERARCHICAL GENERAL LINEAR MODELS WITH EMBEDDED APPLIED PROBABILITY COMPONENTS FOR MULTI-STAGE DISEASE PROGRESSION ANALYSIS IN EPIDEMIOLOGICAL SURVEILLANCE. International Journal Of Engineering Technology Research & Management (IJETRM). 2023Nov21;07(11):107–24.

[5] Nkrumah MA. Forecasting pension fund liabilities through multivariate time series models with structural breaks and demographic statistical trend analysis. World J Adv Res Rev. 2020;5(3):219-38. doi: https://doi.org/10.30574/wjarr.2020.5.3.0058

[6] Nkrumah MA. Actuarial risk evaluation of health insurance portfolios using copula-based time series and Bayesian statistical learning approaches. Int J Comput Appl Technol Res. 2020;9(12):394-407.

[7] Kammerer K, Hoppenstedt B, Pryss R, Stökler S, Allgaier J, Reichert M. Anomaly detections for manufacturing systems based on sensor data—insights into two challenging real-world production settings. Sensors. 2019 Dec 5;19(24):5370.

[8] Hao W, Yang T, Yang Q. Hybrid statistical-machine learning for real-time anomaly detection in industrial cyber–physical systems. IEEE Transactions on Automation Science and Engineering. 2021 May 6;20(1):32-46.

[9] Shahbazi Z, Byun YC. Smart manufacturing real-time analysis based on blockchain and machine learning approaches. Applied Sciences. 2021 Apr 15;11(8):3535.

[10] Nankya M, Chataut R, Akl R. Securing industrial control systems: Components, cyber threats, and machine learning-driven defense strategies. Sensors. 2023 Oct 30;23(21):8840.

[11] Valizadeh M, Wolff SJ. Convolutional Neural Network applications in additive manufacturing: A review. Advances in Industrial and Manufacturing Engineering. 2022 May 1;4:100072.

[12] Oyebode O. Adaptive reinforcement learning agents coordinated through blockchain smart contracts for dynamic governance in decentralized autonomous multi-agent ecosystems. *Int J Sci Res Arch*. 2023;9(2):1155-74. doi:10.30574/ijsra.2023.9.2.0557.

[13] Nkrumah MA. Applied probability-driven general linear models for adaptive pricing algorithms in perishable goods supply chains under demand uncertainty. Int J Sci Res Arch. 2022;6(2):213-32. doi: https://doi.org/10.30574/ijsra.2022.6.2.0292

[14] Ju C, Trinh TK. A machine learning approach to supply chain vulnerability early warning system: Evidence from US semiconductor industry. Journal of Advanced Computing Systems. 2023 Nov 11;3(11):21-35.

[15] Ed-daoudy A, Maalmi K. A new Internet of Things architecture for real-time prediction of various diseases using machine learning on big data environment. Journal of Big Data. 2019 Nov 27;6(1):104.

[16] Lu ZX, Qian P, Bi D, Ye ZW, He X, Zhao YH, Su L, Li SL, Zhu ZL. Application of AI and IoT in clinical medicine: summary and challenges. Current medical science. 2021 Dec;41(6):1134-50.

[17] Aljohani A. Predictive analytics and machine learning for real-time supply chain risk mitigation and agility. Sustainability. 2023 Oct 20;15(20):15088.

[18] Rathore AS, Nikita S, Thakur G, Mishra S. Artificial intelligence and machine learning applications in biopharmaceutical manufacturing. Trends in Biotechnology. 2023 Apr 1;41(4):497-510.

[19] Aldrich C, Auret L. Unsupervised process monitoring and fault diagnosis with machine learning methods. London: Springer; 2013 Jun 15.

[20] Schneider P, Xhafa F. Anomaly detection and complex event processing over iot data streams: with application to EHealth and patient data monitoring. Academic Press; 2022 Jan 7.

[21] Adeshina YT, Owolabi BO, Olasupo SO. A U.S. national framework for quantum-enhanced federated analytics in population health early-warning systems. Int J Eng Technol Res Manag. 2023 Feb;7(2):76-?.

[22] Solarin A, Chukwunweike J. Dynamic reliability-centered maintenance modeling integrating failure mode analysis and Bayesian decision theoretic approaches. *International Journal of Science and Research Archive*. 2023 Mar;8(1):136. doi:10.30574/ijsra.2023.8.1.0136.

[23] Oyegoke Oyebode. Neuro-Symbolic Deep Learning Fused with Blockchain Consensus for Interpretable, Verifiable, and Decentralized Decision-Making in High-Stakes Socio-Technical Systems. *International Journal of Computer Applications Technology and Research*. 2022;11(12):668-686. doi:10.7753/IJCATR1112.1028.

[24] Leal F, Chis AE, Caton S, González–Vélez H, García–Gómez JM, Durá M, Sánchez–García A, Sáez C, Karageorgos A, Gerogiannis VC, Xenakis A. Smart pharmaceutical manufacturing: Ensuring end-to-end traceability and data integrity in medicine production. Big Data Research. 2021 May 15;24:100172.

[25] Jimoh O, Owolabi BO. Developing adaptive HIV treatment guidelines incorporating drug resistance surveillance and genotype-tailored therapies. Int J Sci Res Arch. 2021;4(1):373-92. doi:10.30574/ijsra.2021.4.1.0172.

[26] Adetula AA, Akanbi TD. Beyond guesswork: leveraging AI-driven predictive analytics for enhanced demand forecasting and inventory optimization in SME supply chains. Int J Sci Res Arch. 2023;10(2):1389-406. doi:10.30574/ijsra.2023.10.2.0988.

[27] Jamiu OA, Chukwunweike J. DEVELOPING SCALABLE DATA PIPELINES FOR REAL-TIME ANOMALY DETECTION IN INDUSTRIAL IOT SENSOR NETWORKS. International Journal Of Engineering Technology Research & Management (IJETRM). 2023Dec21;07(12):497–513.

[28] Garmaroodi MS, Farivar F, Haghighi MS, Shoorehdeli MA, Jolfaei A. Detection of anomalies in industrial iot systems by data mining: Study of christ osmotron water purification system. IEEE Internet of Things Journal. 2020 Oct 28;8(13):10280-7.

[29] Oyegoke Oyebode. BLOCKCHAIN-ORCHESTRATED TEMPORAL GRAPH FORECASTING USING HYBRID RNN-TRANSFORMER ARCHITECTURES TO PREDICT SYSTEMIC RISKS IN GLOBAL FINANCIAL AND CLIMATE INFRASTRUCTURES. International Journal of Engineering Technology Research and Management (IJETRM). 2022Mar21;06(03):126–45.

[30] MathWorks. MATLAB 2024 [software]. Natick, Massachusetts: The MathWorks, Inc.; 2024.

[31] Adekunle BI, Chukwuma-Eke EC, Balogun ED, Ogunsola KO. Machine learning for automation: Developing data-driven solutions for process optimization and accuracy improvement. Machine Learning. 2021 Jan;2(1).

[32] Owolabi BO. Exploring systemic vulnerabilities in healthcare digital ecosystems through risk modeling, threat intelligence, and adaptive security control mechanisms. Int J Comput Appl Technol Res. 2022;11(12):687-99. doi:10.7753/IJCATR1112.1029.

[33] Ali T, Kostakos P. Huntgpt: Integrating machine learning-based anomaly detection and explainable ai with large language models (llms). arXiv preprint arXiv:2309.16021. 2023 Sep 27.

[34] Yamakawa T, Miyajima M, Fujiwara K, Kano M, Suzuki Y, Watanabe Y, Watanabe S, Hoshida T, Inaji M, Maehara T. Wearable epileptic seizure prediction system with machine-learning-based anomaly detection of heart rate variability. Sensors. 2020 Jul 17;20(14):3987.

[35] Fosić I, Žagar D, Grgić K, Križanović V. Anomaly detection in NetFlow network traffic using supervised machine learning algorithms. Journal of industrial information integration. 2023 Jun 1;33:100466.

[36] Omarov B, Saparkhojayev N, Shekerbekova S, Akhmetova O, Sakypbekova M, Kamalova G, Alimzhanova Z, Tukenova L, Akanova Z. Artificial Intelligence in Medicine: Real Time Electronic Stethoscope for Heart Diseases Detection. Computers, Materials & Continua. 2022 Feb 1;70(2).

[37] Chibueze T. Advancing SME-focused strategies that integrate traditional and digital banking to ensure equitable access and sustainable financial development. *Int J Sci Res Arch*. 2021;4(1):445-68. doi: https://doi.org/10.30574/ijsra.2021.4.1.0211

[38] Ge Z, Song Z, Ding SX, Huang B. Data mining and analytics in the process industry: The role of machine learning. Ieee Access. 2017 Sep 26;5:20590-616.

[39] Hassan YG, Collins A, Babatunde GO, Alabi AA, Mustapha SD. AI-driven intrusion detection and threat modeling to prevent unauthorized access in smart manufacturing networks. Artificial intelligence (AI). 2021;16.

[40] Nagy B, Galata DL, Farkas A, Nagy ZK. Application of artificial neural networks in the process analytical technology of pharmaceutical manufacturing—a review. The AAPS journal. 2022 Jun 14;24(4):74.

[41] Owolabi BO. Advancing predictive analytics and machine learning models to detect, mitigate, and prevent cyber threats targeting healthcare information infrastructures. Int J Sci Eng Appl. 2023;12(12):76-87. doi:10.7753/IJSEA1212.1018.

[42] Jin Z, Zhang Z, Demir K, Gu GX. Machine learning for advanced additive manufacturing. Matter. 2020 Nov 4;3(5):1541-56.

[43] Shahbazi Z, Byun YC. Integration of blockchain, IoT and machine learning for multistage quality control and enhancing security in smart manufacturing. Sensors. 2021 Feb 20;21(4):1467.

[44] Bao Y, Tang Z, Li H, Zhang Y. Computer vision and deep learning–based data anomaly detection method for structural health monitoring. Structural Health Monitoring. 2019 Mar;18(2):401-21.

[45] Oyebode O. Energy-aware blockchain consensus enhanced by graph neural networks for sustainable, scalable transaction verification across heterogeneous IoT networks. *World J Adv Res Rev*. 2023;20(3):2354-73. doi:10.30574/wjarr.2023.20.3.2678

[46] Ren Z, Gao L, Clark SJ, Fezzaa K, Shevchenko P, Choi A, Everhart W, Rollett AD, Chen L, Sun T. Machine learning–aided real-time detection of keyhole pore generation in laser powder bed fusion. Science. 2023 Jan 6;379(6627):89-94.

[47] Kaur J, Mann KS. AI based healthcare platform for real time, predictive and prescriptive analytics using reactive programming. InJournal of Physics: Conference Series 2017 Dec 1 (Vol. 933, No. 1, p. 012010). IOP Publishing.

[48] Chalapathy R, Chawla S. Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407. 2019 Jan 10.

[49] Shumba AT, Montanaro T, Sergi I, Fachechi L, De Vittorio M, Patrono L. Leveraging IoT-aware technologies and AI techniques for real-time critical healthcare applications. Sensors. 2022 Oct 10;22(19):7675.

[50] Dunning T, Friedman E. Practical machine learning: a new look at anomaly detection. " O'Reilly Media, Inc."; 2014 Jul 21.

[51] Cook AA, Mısırlı G, Fan Z. Anomaly detection for IoT time-series data: A survey. IEEE Internet of Things Journal. 2019 Dec 6;7(7):6481-94.

[52] Meng L, McWilliams B, Jarosinski W, Park HY, Jung YG, Lee J, Zhang J. Machine learning in additive manufacturing: a review. Jom. 2020 Jun;72(6):2363-77.

[53] Rush B, Celi LA, Stone DJ. Applying machine learning to continuously monitored physiological data. Journal of clinical monitoring and computing. 2019 Oct 15;33(5):887-93.

[54] Arden NS, Fisher AC, Tyner K, Yu LX, Lee SL, Kopcha M. Industry 4.0 for pharmaceutical manufacturing: Preparing for the smart factories of the future. International journal of pharmaceutics. 2021 Jun 1;602:120554.

Codes

```
%% Figure 1 — Fermentation Process: ARIMA vs. LSTM Forecasting
clear; close all; rng(42);

%% 1) Generate fermentation-like process with controlled anomalies
T   = 1500;
t   = (1:T)';

carrying   = 8;
growthRate = 0.006;
lag        = 350;
sigmoid    = carrying ./ (1 + exp(-growthRate*(t-lag)));

osc        = 0.4*sin(2*pi*t/120) + 0.15*sin(2*pi*t/36);
noise      = 0.12*randn(T,1);

y_clean    = 2 + sigmoid + osc;
y          = y_clean + noise;

splitIdx = 1000;
anoms = struct();

% Short spikes (sensor blips) — bounds-safe and size-matched
anoms.spikes_idx = splitIdx + [40 140 210 350];
anoms.spikes_idx = anoms.spikes_idx(anoms.spikes_idx <= T);
delta = [1.8 -2.2 1.5 -1.7];
delta = delta(1:numel(anoms.spikes_idx));
if ~isempty(anoms.spikes_idx)
   y(anoms.spikes_idx) = y(anoms.spikes_idx) + delta(:);
end

% Step drift (sensor bias creeping)
anoms.step_start = min(splitIdx + 260, T);
y(anoms.step_start:end) = y(anoms.step_start:end) + 0.6;

% Nonlinear excursion (process upset) — SAFE WINDOW
exc_start = min(splitIdx + 460, T);
exc_end   = min(splitIdx + 520, T);
if exc_end >= exc_start
   exc_win = exc_start:exc_end;
   y(exc_win) = y(exc_win) + 0.8*sin(linspace(0,6*pi,numel(exc_win)))' + 0.5;
else
   exc_win = [];
end

%% 2) Train/test split
yTrain = y(1:splitIdx);
yTest  = y(splitIdx+1:end);

%% 3) ARIMA: fit and rolling 1-step forecast
Mdl = arima(3,1,3);
Mdl = estimate(Mdl, yTrain, 'Display', 'off');

F_arima = nan(numel(yTest),1);
Y_hist  = y(1:splitIdx);
```

```
for k = 1:numel(yTest)
   F_arima(k) = forecast(Mdl,1,'Y0',Y_hist);
   Y_hist(end+1,1) = yTest(k); %#ok<SAGROW>
end
res_arima = yTest - F_arima;

%% 4) LSTM: train with a single sequence (cell), predict with numeric stateful steps
% Next-step pairs
Xtr = yTrain(1:end-1)';   % 1 x (N-1)
Ytr = yTrain(2:end)';     % 1 x (N-1)

% Normalize
mu  = mean(Xtr,'all');
sig = std(Xtr,0,'all') + eps;
XtrN = (Xtr - mu)./sig;
YtrN = (Ytr - mu)./sig;

XTrain = {XtrN};  YTrain = {YtrN};

layers = [
   sequenceInputLayer(1)
   lstmLayer(100,'OutputMode','sequence')
   fullyConnectedLayer(1)
   regressionLayer
];

opts = trainingOptions('adam', ...
   'MaxEpochs', 120, ...
   'MiniBatchSize', 1, ...           % one long sequence
   'GradientThreshold', 1, ...
   'InitialLearnRate', 5e-3, ...
   'LearnRateSchedule','piecewise','LearnRateDropPeriod',60,'LearnRateDropFactor',0.1, ...
   'Shuffle','never','Verbose',false, ...
   'ExecutionEnvironment','auto');    % set 'cpu' if GPU issues

net = trainNetwork(XTrain, YTrain, layers, opts);

% Rolling 1-step forecasts (numeric)
F_lstm = nan(numel(yTest),1);

resetState(net);
[net, ~] = predictAndUpdateState(net, XtrN);  % warm up on training seq

prev = y(splitIdx);
for k = 1:numel(yTest)
   xN = (prev - mu)./sig;              % scalar double
   [net, yNhat] = predictAndUpdateState(net, xN);
   yhat = yNhat*sig + mu;              % numeric math (no cells)
   F_lstm(k) = yhat;
   prev = yTest(k);                    % teacher forcing
end
res_lstm = yTest - F_lstm;

%% 5) Residual-based anomaly scoring
robustScale = @(r) (r - median(r,'omitnan')) ./ (1.4826*mad(r,1));
z_arima = robustScale(res_arima);
z_lstm  = robustScale(res_lstm);
```

```
th_mod = 2.5; th_sev = 3.5;
anom_arima = abs(z_arima) >= th_mod;
anom_lstm  = abs(z_lstm)  >= th_mod;

%% 6) Plot
tt = (splitIdx+1:T)';

figure('Color','w','Position',[100 100 1200 650]);

% Top: forecasts vs actual
ax1 = subplot(2,1,1); hold on; grid on;
plot(t, y, '-', 'LineWidth', 1.0, 'Color', [0.6 0.6 0.6]);
plot(tt, F_arima, '-', 'LineWidth', 1.5, 'Color', [0.00 0.45 0.74]);
plot(tt, F_lstm,  '-', 'LineWidth', 1.5, 'Color', [0.85 0.33 0.10]);

xline(anoms.step_start,'--','Color',[0.5 0 0.5], 'LineWidth',1.2);
if ~isempty(anoms.spikes_idx)
   scatter(anoms.spikes_idx, y(anoms.spikes_idx), 40, [0.49 0.18 0.56], 'filled');
end
if ~isempty(exc_win)
   plot(exc_win, y(exc_win), 'LineWidth',2.2, 'Color', [0.49 0.18 0.56]);
end

ylabel('Process signal');
title('Fermentation Process — Actual vs. ARIMA and LSTM 1-Step Forecasts');
legend({'Actual','ARIMA forecast','LSTM forecast','Step drift (start)','Spike anomalies','Nonlinear excursion'}, ...
       'Location','northwest');

% Bottom: residual z-scores
ax2 = subplot(2,1,2); hold on; grid on;
plot(tt, z_arima, '-', 'LineWidth', 1.2, 'Color', [0.00 0.45 0.74]);
plot(tt, z_lstm,  '-', 'LineWidth', 1.2, 'Color', [0.85 0.33 0.10]);
yline(th_mod,'--k','Thresh 2.5'); yline(-th_mod,'--k');
yline(th_sev,':k','Thresh 3.5');  yline(-th_sev,':k');

scatter(tt(anom_arima), z_arima(anom_arima), 24, [0.00 0.45 0.74], 'filled', 'MarkerEdgeColor','k');
scatter(tt(anom_lstm),  z_lstm(anom_lstm),  24, [0.85 0.33 0.10], 'filled', 'MarkerEdgeColor','k');

xlabel('Time');
ylabel('Residual z-score');
title('Residual-Based Anomaly Scores (higher magnitude = more anomalous)');
legend({'ARIMA z-score','LSTM z-score'}, 'Location','northwest');

linkaxes([ax1 ax2],'x');

%% 7) Simple quantitative summary
MAE_arima = mean(abs(res_arima),'omitnan');
MAE_lstm  = mean(abs(res_lstm),'omitnan');

if ~isempty(exc_win)
   exc_mask = ismember(tt,exc_win);
   first_arima = find(abs(z_arima(exc_mask))>=th_mod,1,'first');
   first_lstm  = find(abs(z_lstm(exc_mask)) >=th_mod,1,'first');
else
   first_arima = []; first_lstm = [];
end

fprintf('ARIMA  MAE (test): %.4f\n', MAE_arima);
fprintf('LSTM   MAE (test): %.4f\n', MAE_lstm);
```

```
if ~isempty(first_arima) && ~isempty(first_lstm)
   fprintf('Early detection (excursion): LSTM leads by %d samples.\n', first_arima - first_lstm);
end

%% 8) Save figure
try
   exportgraphics(gcf, 'Figure1_Fermentation_ARIMA_vs_LSTM.png', 'Resolution', 300);
catch
   % Fallback for older MATLAB versions
   saveas(gcf,'Figure1_Fermentation_ARIMA_vs_LSTM.png');
end
disp('Saved: Figure1_Fermentation_ARIMA_vs_LSTM.png');
```