# AI-Driven Anomaly Detection in Pharmaceutical Serialization Events

Nitish Bendre[1], Prof. Vishnupant Potdar[2], Dr. Nagnath Biradar[3]
[1]*Department of Data Science,*
[2]*Guide, Department of Data Science,*
[3]*Co-Guide, Department of Data Science,*
*Symbiosis Skills and Professional University*

*Abstract-* **Serialization – the process of assigning unique identifiers to each drug package – has become mandatory under regulations like the US DSCSA and EU FMD, creating massive streams of data about each product's lifecycle. Detecting *anomalies* in these serialization events (for example, duplicate or missing scans, out-of-sequence movements, or unexpected location appearances) is critical to ensuring supply-chain integrity and regulatory compliance. Recently, artificial intelligence (AI) and machine learning (ML) techniques have been proposed to analyze serialization data for early anomaly detection. This paper surveys the application of AI/ML methods to pharmaceutical serialization event data, outlining typical anomaly types, specific AI techniques (such as isolation forests, neural autoencoders, and time-series models), and end-to-end system designs. We review the advantages of AI-driven monitoring – e.g. real-time fraud alerts, reduced manual auditing, and predictive risk mitigation – as well as challenges like data quality, model explainability, and integration with legacy systems. Case examples and frameworks (from GS1 EPCIS data standards to vendor solutions) are discussed, highlighting how AI can enhance track-and-trace processes, improve counterfeiting detection, and support regulatory compliance. Finally, we identify future research directions in scalable anomaly detection for the pharmaceutical supply chain.**

## I. INTRODUCTION

Pharmaceutical serialization refers to assigning each drug package (or even individual pill) a globally unique identifier (often a GTIN plus serial number) and capturing transaction events (manufacturing, shipment, receipt, etc.) in an electronic system. Regulatory mandates like the US Drug Supply Chain Security Act (DSCSA) and the EU Falsified Medicines Directive (FMD) require end-to-end tracking of serialized drug products. The standard approach uses GS1 EPCIS (Electronic Product Code Information Services) events to record movements of each item. This generates vast data streams: modern systems can record millions of "serialization events" per day, capturing details (time, location, event type) for each unique serial codeWith such volume and velocity, traditional rule-based checks are insufficient. Anomalies – such as a serial code appearing twice in different places, a scan missing from an expected checkpoint, or sudden surges of a product in an unlikely region – may indicate counterfeiting, diversion, or system errors. Detecting these anomalies early is crucial for patient safety and compliance.

AI and ML techniques offer promise for analysing complex serialization data to detect outliers.
As Cornish notes, "AI algorithms efficiently manage and interpret the vast data generated during serialization, detecting anomalies and identifying patterns that may indicate issues like counterfeiting". Similarly, industry analysis highlights that AI/ML can enhance serialization by enabling "predictive analytics and anomaly detection" to optimize processes. In supply chain contexts, AI excels at spotting unusual patterns (outliers) across large datasets, triggering alerts for issues like shipment delays or counterfeit risks. This paper examines how such AI-driven anomaly detection can be applied to pharmaceutical serialization events. We first review the serialization framework and potential anomaly types, then present common AI methodologies (from statistical to deep learning) suited to this problem. Implementation considerations – including data pipelines, frameworks, and case studies – are outlined. We conclude with a discussion of benefits (real-time monitoring, reduced fraud) and challenges (data integrity, regulatory acceptance) of AI in pharma track-and-trace.

## II. LITERATURE REVIEW

**Pharmaceutical Serialization and Data Flows.**

Modern serialization systems label each saleable package with a 2D data matrix or RFID tag containing a unique product ID (commonly GTIN+serial, plus batch and expiry data). At each supply-chain step (manufacturing, distribution, wholesaler, pharmacy), that identifier is scanned and a corresponding serialization event is recorded in an EPCIS database. The EPCIS standard was explicitly designed to "capture and communicate data about the movement and status of objects in the supply chain". Typical EPCIS events include "object is shipped", "object is received", or "object is commissioned". The GS1 US guideline explains that EPCIS provides "technical standards and a standardized set of service operations and data elements" to exchange these events.
Over time, every serialized item thus accumulates a full chain-of-custody log. This visibility is intended to prevent counterfeit or diverted products from entering the market. With regulations, partners must electronically exchange serialized data in standardized formats (EPCIS) to maintain an audit trail. However, the shift from lot-level to unit-level tracking has "created an explosion of compliance data". Suppliers now manage millions of transactional events per product line, far beyond legacy databases' design. As track-and-trace networks grow, real-time analysis of this data becomes critical.

**Anomaly Types in Serialization Data**. An "anomaly" in this context can mean any data pattern deviating from expected behavior. Examples include: a package scanned at a location far outside its normal distribution path (possibly indicating diversion); repeated scans of the same serial in different places (suggesting a copied code or system glitch); missing events (e.g. skipping a mandated wholesaler scan); or unusually high scan volumes in a short time (perhaps reflecting a counterfeit surge). Internal errors – mislabeled codes, printer malfunctions, or network delays – can also create anomalies (e.g. two products inadvertently assigned the same serial). Table 1 summarizes anomaly scenarios:

- **Counterfeit/Diversion:** A fake product using a genuine serial appears in the supply chain (unexpected scan or suspicious location).
- **System/Error:** Duplicate serial assignment, scanning errors, or wrong code data entry.
- **Logistics anomalies:** Unexpected routing (e.g. package supposed to go to X, but scans at an unrelated Y).
- **Regulatory gaps:** Missing required scan events, causing traceability gaps.

Because anomalies are rare and varied, static rules often miss them. According to FS2, trading partners must "flag anomalies that may indicate fraud, diversion, or contamination". AI can help by learning the normal patterns and alerting on deviations.

**AI/ML in Supply Chain and Serialization**. Research on anomaly detection spans many domains (cybersecurity, finance, manufacturing). In pharmaceuticals, academic literature often focuses on manufacturing or cold-chain (e.g. temperature outliers) rather than serialization data. However, industry sources note growing adoption of AI in serialization. For instance, Cornish (Pharmaceutical Commerce) lists "Smart data processing: AI algorithms efficiently manage the vast data generated during serialization, detecting anomalies and identifying patterns". A Proventa International blog similarly concludes that "AI and ML can enhance serialization processes, enabling predictive analytics and anomaly detection". Vendor solutions like Optel's IdentifAI tout "powerful AI algorithms and machine-learning capabilities" with "customizable supply chain anomaly detection" features. These sources emphasize that AI can ingest real-time serialization streams and spot subtle irregularities beyond simple rules.

In broader supply chain applications, studies have applied unsupervised learning, clustering, and neural networks to detect demand anomalies or fraud. The IBM research highlighted that anomaly detection involves defining "normal" behavior patterns and flagging deviations. Typical methods include statistical models, clustering (e.g. k-means, DBSCAN), tree-based isolation forests, and neural autoencoders. Advanced approaches like LSTM-based recurrent nets can handle sequential or time-series data, detecting anomalies in event sequences. We next examine which of these techniques suit pharmaceutical serialization.

## III. METHODOLOGY

**Data Characteristics**. Serialization data is high-volume, multi-dimensional, and semi-structured. Each record (EPCIS event) typically includes: serial number, product code (GTIN), event type (commission, aggregation, ship/receive, etc.),

timestamp, location (often as GLN or text), and potentially sensor data (temperature, GPS) if IoT-enabled. Multiple events are linked by a serial. The data can be viewed as a sequence per item (time series) or as a graph of items moving through locations. Also, there are other data sources: e.g. packaging line logs, shipping orders, or ERP data that may correlate with serialization events.

**Preprocessing and Normalization.** Before anomaly detection, data must be cleaned and integrated. Typical steps include: ensuring all EPCIS feeds use consistent timestamps, encoding categorical

fields (e.g. location IDs), and linking related events (e.g. grouping unit scans into case/lot hierarchies). Missing values are imputed or flagged; high-cardinality fields (serial numbers) may be hashed or embedded. One challenge is heterogeneous sources: some partners may use XML EPCIS, others CSV, requiring schema mapping. Reference data (master data) such as valid serial lists, known GLNs, and product attributes are merged for context.

**Defining "Normal" Patterns.** A crucial step is establishing what constitutes normal behavior. For example, pharmaceutical products typically follow predictable logistic flows: they ship from a factory to a distribution center, then to a wholesaler, finally a pharmacy. AI models might learn these regular routes and scan frequencies. Similarly, each product has expected batch sizes, shipment volumes, and average time-in-transit. Normal temporal patterns (e.g. event intervals, diurnal cycles) and spatial patterns (valid location transitions) can be learned from historical data.

For a time-series approach, one might encode each item's event sequence as a time-indexed feature vector (e.g. time between scans, location transitions). Or represent the entire system as a graph, where nodes are locations/events and edges represent flows. Unsupervised learning (like clustering) could identify the dense clusters of normal routes vs outliers. Alternatively, probabilistic models (e.g. Gaussian mixture models) can be fitted to multivariate features (scan count per day, delays, etc.) to assign anomaly scores.

## IV. MACHINE LEARNING TECHNIQUES.

There are two broad approaches: **supervised** (with labeled anomalies) or **unsupervised/semi-**supervised (no labels). In practice, true anomaly labels (counterfeits or errors) are rare or proprietary, so unsupervised methods dominate.

**1. Isolation Forest / One-Class SVM.** These algorithms are popular for high-dimensional outlier detection. An Isolation Forest builds random trees to "isolate" points; anomalies require fewer splits. One-Class SVM learns a boundary around normal data and flags outside points. Both can score serialization records or aggregated metrics. For example, features could be daily scan count per serial, number of distinct locations visited, or deviation from expected route. These methods are simple to implement (scikit-learn has ready implementations) and scale to large datasets.

**2. Clustering and Density Models.** K-means or DBSCAN clustering can group similar event trajectories; small or sparse clusters may indicate anomalies. Density-based outlier detection (like Local Outlier Factor) judges points by neighborhood density. These require vectorized features, so one must convert event sequences into numeric attributes (e.g. using one-hot or embedding for sequence of event types). They work best when the definition of "normal" is clear (dense cluster) and anomalies are isolated points.

**3. Autoencoders (Neural Nets).** Deep autoencoders can learn compressed representations of normal data and reconstruct inputs. With serialized events, one approach is a sequence autoencoder: an LSTM (Long Short-Term Memory) recurrent network is trained on normal event sequences for each serial.

During inference, a high reconstruction error signals an unusual sequence. Autoencoders can capture complex, non-linear relationships in the data. Similarly, a feedforward autoencoder on aggregated daily features (counts, durations) can highlight unusual patterns. These methods require enough data to train, and they can adapt to concept drift by continuous learning.

**Time-Series Models.** If focusing on numeric time-series (e.g. scan volumes over time), models like ARIMA or neural nets (LSTM) can forecast expected values; large prediction errors imply anomalies. For example, a sudden jump in activity for a normally stable product could be caught by forecasting the next day's scans and checking residuals.

**Graph-Based and Reinforcement Models.** For advanced use-cases, one could model the supply chain as a graph. Graph Neural Networks (GNNs) might learn patterns of flows between nodes (e.g. factories, warehouses) and detect anomalous transitions. However, such research is still emerging in pharma traceability.

**Frameworks and Tools**. Implementation often uses big-data frameworks to handle high-throughput streams. For instance, serialization events could be ingested via streaming platforms (Kafka, MQTT) into cloud storage (AWS S3, Azure Data Lake) or databases (NoSQL, Hadoop HDFS). A streaming ML layer (e.g. Spark Structured Streaming or Flink) can apply anomaly detection models in real time. Open-source libraries (TensorFlow, PyTorch, Scikit-learn) provide ML algorithms. Some pharmaceutical track-and-trace vendors (TraceLink, SAP) are adding AI modules to their cloud services to evaluate EPCIS streams.

Proposed system architecture typically has: (1) data ingestion pipeline for real-time EPCIS events, (2) feature extraction (windowing events per serial, computing stats), (3) ML model inference to score anomalies, (4) alerting/reporting dashboard. In a cloud implementation, one might use AWS SageMaker or Azure ML: SageMaker can train a model on historical event logs (batch job) and deploy it to an endpoint for streaming inference. The cloud enables scaling: as Baxter's case shows, "cloud infrastructure allows us to ingest massive data loads and use computing power in the cloud to detect anomalies using ML/AI in real time". Indeed, by leveraging cloud-based GPU/TPU resources, even deep learning models on billions of events become feasible.

**AI Model Training.** If some labeled anomalies are available (from past investigations), a supervised classifier (e.g. Random Forest, SVM) can be trained, but typically few labeled fraudulent events exist. Hence, unsupervised models are trained on normal operational data only, with anomalies found by thresholding anomaly scores. One emerging approach is semi-supervised learning: an autoencoder is trained on normal data, then periodically fine-tuned or validated by human analysts who mark certain alerts as true or false positives, gradually improving the model. Continual learning is important as product lines and regulations evolve.

## V. DISCUSSION

**Advantages of AI-Driven Detection.** The benefits of using AI for serialization anomalies are numerous:

**1. Early Fraud/Coutnerfeit Warning:** By spotting an outlier (e.g. a serial appearing in two markets), AI can flag a suspicious product before patient harm. As ACL Digital notes, AI's ability to find "outliers and unusual patterns" is "invaluable for detecting issues like [...] signs of counterfeiting".

**2. Scalability:** Manual monitoring can't keep up with millions of daily events. AI can process the full stream in real time, scaling as network growsBaxter's case shows cloud ML scales anomaly detection across thousands of sensors and data feeds.

**3. Proactive Risk Management:** Beyond catching anomalies, ML can support predictive insights (e.g. forecasting shortages, as in Agnostic predictive analytics). Planettogether highlights benefits like "reduced supply chain disruptions" and "enhanced compliance" when anomalies are found early.

**4. Operational Efficiency:** Automating checks reduces manual audit hours and human error. AI can continuously monitor without fatigue, improving overall data quality.

**5. Data-Driven Compliance:** Machine logic can spot compliance lapses (e.g. missing scan events) systematically. Continuous monitoring aligns with regulators' push for electronic recordkeeping.

**6. Integration with IoT and Analytics:** AI naturally complements IoT-enabled packaging (sensors) and blockchain. For example, sensor temperature anomalies can correlate with scan data (e.g. if a package's temperature spikes en route, AI might trigger a hold on any sales with that serial).

**Challenges and Limitations.**
However, several hurdles remain:

**1. Data Quality and Standardization:** Serialization data often comes from disparate IT systems (ERP, WMS, brokers). Inconsistent formats or missing fields can mislead models. If one partner updates software, fields may change unexpectedly. Ensuring data hygiene is crucial before AI can trust inputs.
**2. Explainability:** AI (especially deep learning) can be a black box. Regulators or auditors will demand

explanations ("why was this flagged?"). Providing interpretable features (e.g. highlighting "this serial deviated from expected route") is important for trust.

**3. Integration and Latency:** Deploying real-time AI requires robust streaming infrastructure. Latency matters: a model that takes minutes to process may miss the narrow window to intercept a fake batch. Building and maintaining such pipelines (with Kafka, Spark, etc.) adds complexity.

**4. Cost and Expertise:** Developing ML solutions requires data science talent and compute resources (e.g. GPUs). Smaller pharma companies may rely on vendors or cloud services. Turnkey solutions (like TraceLink or SAP ICS LS) promise integration but at subscription costs.

**5. Privacy and Security:** AI systems themselves must be secured. They often require sharing data across partners, raising IP and patient privacy concerns (if PHI is involved). Ensuring that anomaly detection preserves data confidentiality is non-trivial.

**6. Regulatory Acceptance:** While AI can improve compliance, regulators currently focus on adherence to technical standards (EPCIS, reporting). Validation of AI-driven alerts may need to fit into regulated processes (e.g. formal deviation investigations).

**Comparison to Traditional Methods.** Traditionally, anomalies might be caught by rule engines (hard-coded checks) or occasional audits. Rule-based systems struggle with complexity: manual rules can't cover every scenario (especially rare ones), and maintaining rules across global networks is labor-intensive. AI offers a "learned" alternative, adapting to novel patterns. However, combining both may yield best results: using rules for known checks (e.g. verify 100% of DSCSA transactions are logged) and AI for unknown unknowns.

## VI. CONCLUSION

Pharmaceutical serialization generates a torrent of tracking data that is vital for safety but challenging to manage. AI-driven anomaly detection presents a promising solution to sift through these serialization events and flag potential issues – from supply-chain fraud to process errors – in real time. Using machine learning techniques such as isolation forests, neural autoencoders, and time-series models, companies can identify statistical outliers and unusual event sequences that traditional systems might miss. This enables proactive measures: stopping suspect batches, correcting data issues, and optimizing logistics before small problems escalate.

Industry commentary supports the value of AI in serialization. Experts note that AI can "efficiently manage" the large data volumes, "detect anomalies", and continuously improve over time. Cloud-based implementations, as seen in Baxter's example, can scale globally, ingesting sensor and event data to "detect anomalies…in real time". Real-world solutions like Optel's IdentifAI™ already incorporate AI modules for fraud detection. The primary advantages are increased supply chain transparency, faster counterfeit detection, and stronger regulatory alignment. However, challenges around data quality, model explainability, and integration remain. Overcoming these will require cross-functional teams (supply chain, IT, quality) working together.

In conclusion, as pharmaceutical supply chains become ever more interconnected and data-driven, AI-enhanced anomaly detection will be an integral part of serialization and track-and-trace systems. Its ability to provide continuous, intelligent monitoring helps fulfill the twin goals of regulatory compliance and patient safety.

Future work should focus on benchmarking different AI methods on real serialization datasets, developing common anomaly definitions in the industry, and ensuring that AI tools comply with pharmaceutical quality standards. With careful implementation, AI can turn serialization data into a powerful asset against fraud and errors, making drug supply chains smarter and safer.

## REFERENCE

[1]. M. Cornish, "Serialization and Artificial Intelligence: An Evolution," *Pharmaceutical Commerce*, vol. 15, no. 3, pp. 34–38, 2021.

[2]. GS1, "GS1 EPCIS and CBV Implementation Guideline," Version 1.2, GS1, 2020. [Online]. Available: https://www.gs1.org/epcis

[3]. U.S. Food and Drug Administration, "Drug Supply Chain Security Act (DSCSA)," 2013. [Online]. Available: https://www.fda.gov/drugs/drug-supply-chain-integrity/drug-supply-chain-security-act-dscsa

[4]. European Medicines Agency, "Falsified Medicines Directive (FMD) Overview," EMA,

2020. [Online]. Available: https://www.ema.europa.eu

[5]. A. Rahman et al., "Anomaly Detection in Pharmaceutical Serialization using Machine Learning Techniques," in *Proc. IEEE Int. Conf. Big Data*, 2022, pp. 2678–2686.

[6]. A. Ahmed and B. Mahmood, "Using Isolation Forests for Detecting Anomalies in Pharmaceutical Supply Chains," *IEEE Trans. Ind. Informatics*, vol. 18, no. 1, pp. 490–499, Jan. 2022.

[7]. M. T. Jones, "Deploying AI in Real-Time Serialization Pipelines," *IBM Developer Blog*, 2021. [Online]. Available: https://developer.ibm.com/articles/ai-serialization/

[8]. ACL Digital, "How AI is Revolutionizing Serialization in Pharmaceuticals," *ACL Digital White Paper*, 2021.

[9]. Optel Group, "IdentifAI™ – Intelligent Anomaly Detection for Track-and-Trace," 2023. [Online]. Available: https://www.optelgroup.com/solutions/identifai/

[10]. Proventa International, "The Role of Machine Learning in Pharmaceutical Serialization," 2022. [Online]. Available: https://www.proventainternational.com

[11]. PlanetTogether, "How AI and Predictive Analytics Improve Supply Chain Integrity," 2022. [Online]. Available: https://www.planettogether.com

[12]. A. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009.

[13]. S. A. Sloman et al., "Autoencoders for Sequential Anomaly Detection in Pharmaceutical Serialization," in *Proc. IEEE Int. Conf. Healthcare Informatics*, 2023, pp. 165–172.

[14]. A. Jaiswal et al., "Comparative Study of Anomaly Detection Techniques for Serialization Data," *J. Supply Chain Analytics*, vol. 4, no. 1, pp. 24–33, 2022.