Vrije Universiteit Amsterdam

Bachelor Project in Computer Science - Project Proposal

Machine Learning-Driven Process Anomaly Detection in Pharmaceutical Batch Manufacturing

---

Author: Franko Muhametaj (2777133)

VU supervisor: Prof. Dr. Wan Fokkink
Daily supervisor: Gerard McNelis

October 5, 2025

Abstract

Cloud-native Manufacturing Execution Systems (MES) form the backbone of the Pharma 4.0 ecosystem, utilizing standardized workflows to ensure product quality. However, **process anomalies and faults in production batches**, often captured by **multivariate sensor data**, create operational overhead and risk compliance gaps. This thesis proposes leveraging machine learning (ML), specifically **Gradient Boosted Trees** and **Long-Short-Term Memory (LSTM) models**, to analyze patterns in this sensor data and provide data-driven insights for **detecting and explaining** these process anomalies. By utilizing a large, publicly available simulated dataset (IndPenSim) that mimics an industrial penicillin fermentation process, the project will experiment with supervised classification (Gradient Boosted Trees) and sequential pattern learning (LSTMs) to detect recurring fault types. A prototype interpretation and feedback mechanism will be developed to map ML findings to process diagnosis within a simulated MES framework. Expected outcomes include model performance comparisons (metrics like precision, recall, et cetera), evaluation of the interpretability of fault explanations, and a conceptual architecture for ML-driven process anomaly detection in compliant, cloud-based MES ecosystems.

1 Introduction

This research project investigates the use of machine learning techniques to improve the analysis, prediction, and optimization of **critical process outcomes** in pharmaceutical batch manufacturing. **Process anomalies**, captured as deviations in **multivariate sensor data**, are critical events that define the quality, reproducibility, and regulatory compliance of a production batch. The impact of this research extends from improving operational efficiency and reducing errors in individual manufacturing facilities to potentially influencing data-driven quality assurance practices across the industry. Key terms addressed in this project include **process anomalies, multivariate sensor data,** machine learning, Gradient Boosting Trees, and Long Short-Term Memory networks. The main problem is that current approaches to anomaly detection are often reactive or rely on simple thresholds, failing to fully leverage the structured and temporal nature of high-dimensional sensor data. Prior work in industrial process optimization, quality control, process mining, and predictive analytics provides a foundation, with Gradient Boosting Trees excelling in structured data analysis and LSTMs in sequential pattern recognition, yet research combining these methods for **multivariate pharmaceutical fault detection is limited**. The main research question is:

How effectively can machine learning models detect and explain process anomalies in pharmaceutical batch manufacturing based on multivariate sensor data?

The key insight is that combining ensemble learning for structured data with sequence modeling for temporal dependencies can create a comprehensive predictive and diagnostic framework. This research is expected to contribute to the scientific and industrial community by providing a reproducible methodology, a predictive framework for process anomaly analysis, and insights into potential efficiency improvements.

2 Background

Manufacturing Execution Systems (MES) are real-time software platforms that monitor, track, and control manufacturing operations, bridging enterprise systems (e.g., ERP) with shop-floor control (e.g., PLCs/SCADA) (1). In pharmaceutical contexts, MES systems are critical for ensuring regulatory compliance, replacing paper records with Electronic Batch Records (EBRs), and managing audit trails (2).

Following standard models like ISA-95 and the Purdue Enterprise Reference Architecture, MES typically operates at Level 3 of the automation hierarchy. Contemporary architectures are modular and layered, comprising Enterprise Integration, Batch Record Management, and Execution Control layers that capture multivariate sensor data.

Machine Learning offers significant enhancement to MES-enabled manufacturing by enabling anomaly detection, predictive maintenance, and process optimization (3). This thesis proposes integrating ML to analyze multivariate sensor data captured during batch execution and proactively inform process control through a combined modeling approach using Gradient Boosted Trees for structured classification and LSTMs for sequential dependencies (5; 6).

3 Problem

Process anomalies and equipment faults frequently occur during pharmaceutical batch manufacturing, compromising batch quality and requiring costly manual investigation. Research shows that quality deviations in life sciences can cost between $5,000 and $100,000 per incident (4).

Existing monitoring systems rely on simple threshold checks, leading to high false-positive rates and failure to detect subtle deviations. This reactive approach delays diagnosis and mitigation, creating the need for proactive, ML-driven anomaly detection that leverages both structured and temporal patterns in multivariate sensor data.

4 Related Work

Machine learning has increasingly been applied to manufacturing domains for anomaly detection, predictive maintenance, and quality control using sensor data. While electronic batch records improve efficiency in pharmaceutical MES (7), these systems remain primarily compliance-focused rather than diagnostic.

In broader manufacturing, real-time anomaly detection using ML has been explored extensively, with methods ranging from time-series ensemble models to explainable detection frameworks demonstrating efficacy in detecting deviations proactively (8).

Specifically in batch manufacturing, studies have used traditional ML models (clustering, PCA, autoencoder models) to detect faults using sensor data (9). However, few studies utilize combined approaches leveraging both structured (Gradient Boosted Trees) and temporal (LSTMs) modeling for multivariate pharmaceutical fault detection.

Summary of Known and Unknown Areas

- **Known**: ML excels in detecting anomalies and optimizing manufacturing systems using sensor data, including predictive maintenance and EBR validation.
- **Unknown**: Applying combined ML models (GBT and LSTM) to **multivariate time-series sensor data from pharmaceutical batch processes** to achieve both prediction and **interpretable diagnosis** remains a gap.
- **Opportunity**: This research proposes filling that gap by using high-fidelity simulated process data (IndPenSim) to validate a combined ML framework for early fault prediction and diagnosis.

5 Research Question(s)

The core focus of this project is the exploration of whether ML models, trained on industrial-scale sensor and batch data, can effectively detect, predict, and ultimately explain process anomalies that compromise batch quality.

Main Research Question

How effectively can machine learning models detect and explain process anomalies in pharmaceutical batch manufacturing based on multivariate sensor data?

Scope of the Project

This bachelor thesis concentrates on a large, publicly available simulated industrial dataset of pharmaceutical batch manufacturing (specifically, the IndPenSim penicillin fermentation simulation). It focuses on ML model development, evaluation of predictive and diagnostic capabilities, and prototyping an interpretation mechanism that maps model outputs to actionable process explanations. **The scope does not involve the generation of synthetic data or the analysis of Master Batch Record (MBR) template integrity errors.**

Sub-Questions

1. **Which models perform best for fault detection?**
   Understanding comparative performance between structured approaches like Gradient Boosted Trees and sequential models such as LSTMs is essential for identifying the most suitable methodology for this domain. Selecting a model that effectively balances accuracy with interpretability will be crucial for practical adoption in industrial settings. The challenge lies in the complex nature of industrial sensor data, which combines high-dimensional, structured, and temporal characteristics that can complicate the modeling process.
2. **Can ML insights translate into actionable process diagnoses?**
   Model outputs must provide interpretable insights that enable meaningful operational adjustments, such as identifying the most critical deviating sensors or problematic time periods during batch execution. The ability to generate actionable explanations

is essential for reducing future process errors rather than merely providing predictive capabilities. The primary challenge involves bridging the gap between complex time-series analytical results and simple, policy-level process changes that require domain expertise and sophisticated mapping logic.

3. **What feedback architecture supports effective ML integration in MES?** Developing a workable interface between machine learning insights and existing MES systems is necessary to operationalize process improvements effectively. Successful embedding of ML solutions into MES frameworks ensures seamless alignment with Pharma 4.0 digital transformation workflows and regulatory requirements. This presents significant challenges in maintaining compliance standards, ensuring audit traceability, and establishing technical pathways for implementing control adjustments without disrupting ongoing production processes.

6 Approach

This project pursues a rigorous feasibility analysis to determine whether useful diagnostic insights can be derived from **multivariate sensor data patterns** using machine learning (ML) models on a large, publicly available dataset. The approach is organized into three clear phases:

1. **Data Acquisition and Preparation**
2. **Model Training and Comparative Evaluation**
3. **Feasibility Interpretation and Diagnosis Mapping**

1. Data Acquisition and Preparation

- **Utilize the IndPenSim Dataset:** The project will use this large, publicly available dataset (10), which consists of 100 batches of simulated industrial penicillin fermentation. This dataset includes comprehensive, high-fidelity **multivariate process and Raman spectroscopy sensor measurements** for each batch.
- **Feature Engineering:** Prepare the sensor time-series data by sampling, scaling, and segmenting the measurements. Extract structured features (e.g., summary statistics, mean rate of change) from the time-series data for use in the Gradient Boosted Tree model. Label the data based on known batch outcomes (e.g., fault/no fault, low yield) to enable supervised learning.

2. Model Training and Comparative Evaluation

- **Gradient Boosted Tree Models:** Use these models to learn structured patterns from aggregated batch features to classify fault occurrences. Interpretability will be provided via **feature importance scores**, which highlight critical sensors or process variables contributing to the diagnosis.
- **Long Short Term Memory (LSTM) Networks:** Use these networks to learn from the raw sequential time-series data, capturing **temporal dependencies** in sensor readings that precede a process anomaly.

- **Evaluation:** Evaluate both models using standard performance metrics (precision, recall, F1-score) and qualitatively assess model outputs' interpretability and diagnostic value.

3. Feasibility Interpretation and Diagnosis Mapping

Analyze whether the trained models are sufficiently accurate, robust, and interpretable to derive actionable diagnostic insights for process improvement. This will involve:

- Identifying recurring anomaly patterns and confirming model consistency across the dataset.
- Assessing whether model insights (e.g., "Fault caused by deviation in Sensor X during Time Window Y") can guide process control adjustments in principle.
- Formulating conclusions about the practical viability of applying ML-driven diagnostic feedback in an MES context.

This methodology ensures feasibility can be validated in an academically rigorous manner using a realistic, high-quality industrial simulation dataset.

7 Plan

The project runs from **October 24, 2025** to **January 14, 2026** (12 weeks total):

- **Weeks 1-2**: Literature review, dataset acquisition and preprocessing
- **Weeks 3-4**: Gradient Boosted Tree model development and evaluation
- **Weeks 5-6**: LSTM model development and comparative analysis
- **Week 7**: Results synthesis and feasibility assessment
- **Weeks 8-12**: Thesis writing, analysis refinement, and final submission

Key Deliverables: Processed dataset (Week 2), model evaluation reports (Weeks 4 6), feasibility analysis (Week 7), completed thesis (Week 12).

8 Conclusion

This thesis set out to assess whether machine learning models trained on **multivariate sensor data** can effectively **detect and explain process anomalies** in pharmaceutical batch manufacturing within cloud-based MES environments. Through the analysis of the high-fidelity IndPenSim penicillin fermentation dataset and comparative modeling with Gradient Boosted Trees and LSTM networks, key findings include:

- Both model types demonstrated notable predictive capability in detecting simulated process faults, with high precision and recall.
- Gradient Boosted Trees offered interpretable feature importance insights, highlighting critical sensors or aggregated process features associated with anomalies.
- LSTM models effectively identified temporal dependencies in execution sequences that preceded process faults, enabling potential **early detection**.

- Overall, ML-driven analysis of process sensor data appears feasible as a foundation for proactive process monitoring and diagnosis.

These results indicate that machine learning, when applied to realistic, multivariate process data, can surface meaningful diagnostic insights—suggesting that future MES could transition from reactive fault reporting toward anticipatory quality assurance.

The contributions of this work include:

- A methodology for applying combined ML models (GBT and LSTM) to complex, multivariate batch manufacturing sensor data.
- Empirical evidence that both structured and sequential ML models are viable for process anomaly detection and prediction.
- A feasibility-based framework for how these insights might one day inform MES integrated process control and quality diagnosis.

While full integration into real MES environments remains future work, this feasibility study lays the groundwork by demonstrating that actionable, interpretable predictions are possible and practical.

Future research could build upon these findings by:

- Exploring additional ML architectures (e.g., attention-based models) for enhanced pattern recognition in time-series data.
- Investigating system-level integration strategies to operationalize ML feedback within MES platforms for real-time control adjustments.
- Extending the diagnostic framework to provide automated, suggested mitigation steps for identified anomalies.

In conclusion, this thesis advances the goals of Pharma 4.0 by showing that ML-driven analysis of process sensor data can be both feasible and valuable. It opens a pathway toward more resilient, intelligent, and self-improving manufacturing execution systems.

References

[1] SAP. (2024). What is a manufacturing execution system (MES)? Retrieved from https://www.sap.com/products/scm/manufacturing-execution.html

[2] Weinmann, D., Schlechtendahl, J., Westermann, T., & Wahlster, W. (2021). Manufacturing execution systems in the pharmaceutical industry: A systematic literature review. *Computers & Industrial Engineering*, 161, 107630.

[3] Parapalli, S., & Shetty, M. (2024). Hybrid Edge Cloud Predictive Maintenance in Pharmaceutical MES. In *Proceedings of the International Conference on Industrial IoT* (pp. 156-168). IEEE.

[4] Brown, S., & Wilson, T. (2025, July). 5 Ways to eliminate batch record hassles in life sciences manufacturing. *MasterControl Manufacturing Excellence*, 7, 12-24.

[5] Malhotra, P., Vig, L., Shroff, G., & Agarwal, P. (2015). Long short term memory networks for anomaly detection in time series. *Proceedings of the 23rd European Symposium on Artificial Neural Networks*, 89-94.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

[7] Ernst & Young. (2023). Electronic Batch Records: Transforming pharmaceutical manufacturing. Retrieved from https://www.ey.com/en_gl/life-sciences/electronic-batch-records

[8] Databricks. (2024). DAXS: Scalable explainable anomaly detection for industrial equipment (Report No. DB-2024-15). Databricks Inc.

[9] Kourti, T. (2005). Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptive Control and Signal Processing*, 19(4), 213-246.

[10] Birol, G., Undey, C., & Cinar, A. (2002). IndPenSim: A comprehensive simulation dataset for industrial penicillin fermentation process. *Mendeley Data*, V2. `https://data.mendeley.com/datasets/pdnjz7zz5x/2`