International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 08 | Aug - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Real-Time Anomaly Detection in Biopharmaceutical Manufacturing: A Machine Learning Approach

Name of Author: Ravi Kiran Koppichetti
Email ID of Author: koppichettiravikiran@gmail.com

Abstract:

Integrating the Internet of Things (IoT) and Machine Learning (ML) within smart manufacturing facilities has significantly transformed anomaly detection processes, thereby ensuring predictive maintenance, optimizing processes, and enhancing operational efficiency. This paper reviews existing research regarding real-time anomaly detection in IoT-enabled manufacturing environments, specifically emphasizing biopharmaceutical production. A variety of Machine Learning (ML) techniques, including convolutional neural networks (CNNs), hidden Markov models (HMMs), and statistical methodologies, are examined to identify deviations from standard operating conditions. The research identifies distinct challenges associated with anomaly detection, including issues related to data collection errors, class imbalance, and constraints in the selection of ML models, which can potentially impact the accuracy of predictions. Furthermore, the paper delineates a systematic methodology for integrating anomaly detection in biopharmaceutical Active Pharmaceutical Ingredient (API) manufacturing, highlighting the importance of infrastructure assessment, data acquisition, model training, and ongoing improvement. The findings accentuate the necessity of bridging the gap between Information Technology (IT) and Operational Technology (OT), securing IoT networks, and enhancing ML models to mitigate false positives and false negatives. Ultimately, this study advances data-driven decision-making in smart manufacturing, promoting a more robust and resilient industrial ecosystem.

Keywords:
Industrial Internet of Things, machine learning, Industry 4.0, anomaly detection, Biopharmaceutical manufacturing, process monitoring

I. Introduction

Recent developments in Industry 4.0 have revolutionized traditional manufacturing into smart factories through the integration of the Internet of Things (IoT), Artificial Intelligence (AI), and Machine Learning (ML). A pivotal component of smart manufacturing is anomaly detection, which identifies discrepancies in machinery operations that may signal potential failures, inefficiencies, or security risks. Early anomaly detection is essential for predictive maintenance, minimizing unplanned downtime, and boosting overall operational efficiency.

Manufacturing plants equipped with IoT produce large volumes of sensor data from machinery, monitoring key metrics such as temperature, pressure, vibration, speed, and energy consumption. Conventional rule-based methods for anomaly detection often struggle with the complexity and scale of this data. Consequently, machine learning techniques, including convolutional neural networks (CNNs), hidden Markov models (HMMs), and various statistical approaches, have become more prominent due to their ability to identify real-time subtle, complex patterns. These AI-powered methods significantly improve fault detection, process optimization, and product quality control within industrial environments.

Nonetheless, challenges and research gaps remain in applying anomaly detection across diverse sectors, particularly in biopharmaceutical manufacturing. Biopharmaceutical plants feature highly sensitive production processes, stringent regulatory requirements, and rigorous quality control protocols, which complicate and amplify the importance of effective anomaly detection. Previous research has predominantly centered on predictive maintenance, cybersecurity in IoT networks, and general anomaly detection in the industry, with little focus on the unique demands of biopharmaceutical API manufacturing. This underscores the need for a specialized strategy tailored to this heavily regulated and data-centric field.

This paper investigates anomaly detection within smart industrial machinery, emphasizing biopharmaceutical API production. It thoroughly reviews current methodologies, highlights major challenges, and proposes a systematic framework for integrating ML-based anomaly detection in IoT-enabled settings. The study aims to enhance operational resilience, mitigate production risks, and improve compliance with regulations in smart manufacturing environments [1,2,3].

II. Literature Review of Similar Work

Anomaly detection in similar environments has been a critical research topic over the last decade. Numerous works and projects have been developed to address this vital area's challenges and solutions. Previous research has highlighted the critical aspect of detecting anomalies in real-time in IoT environments. The study [4] focused on detecting failures in industrial machinery by analyzing time series data from sensors. It demonstrated how machine learning algorithms, particularly recurrent neural networks, can identify unusual patterns in machinery operations, helping to prevent unplanned downtime. Similarly, the study [5] has focused on providing a comprehensive review of the recent advancements in ML techniques widely applied to predictive maintenance in smart manufacturing. The authors have explained data acquisition, classification of data, key contributors, etc., thus offering community guidelines.

Furthermore, the study [6] has extensively focused on utilizing machine learning models for biopharmaceutical batch process monitoring utilizing minimal historical data. This article [6] addressed the problem of real-time statistical batch process monitoring with limited production history, also referred to as the Low-N problem. The authors of [7] focused on surveying the existing machine learning solutions built on cloud/ fog/ edge architectures.

III. Identifying the gap

Despite technological advancements, there remains a significant gap in research related to biopharmaceutical manufacturing. This gap arises from the absence of approaches specifically tailored to the unique characteristics of this advanced manufacturing sector. Most previous studies have focused on securing IoT applications, enhancing cybersecurity, implementing advanced IT/OT solutions, or have concentrated on specific domains such as healthcare or general manufacturing. However, there has been insufficient research on biopharmaceutical manufacturing, particularly regarding using advanced equipment to produce Active Pharmaceutical Ingredients (APIs). This oversight has resulted in a lack of understanding of the specific challenges and optimal solutions within this context [8,9]

IV. Concepts used

Anomaly detection involves identifying data points that notably diverge from anticipated patterns in the manufacturing process. Such deviations often indicate possible quality or production consistency concerns.

A. **Anomaly**: An anomaly is a data point or pattern that significantly deviates from what is expected or considered normal within a dataset. Essentially, it is an outlier that stands out from the established patterns. In IoT systems, these anomalies can appear as unusual fluctuations in sensor measurements, atypical device behavior, or unexpected variations in data traffic [10].

B. **Statistical Methods**: Statistical methods are mathematical techniques used to collect, organize, analyze, interpret, and present data. They are essential for transforming raw data into actionable insights. They are categorized into Descriptive Statistics, which summarize data from a sample using measures such as mean or standard deviation, and Inferential Statistics, which conclude from data subject to random variation, such as observational errors or sampling variation.

C. **Machine Learning Algorithms:** A series of rules or processes computer systems employ to examine data, recognize patterns, and predict outcomes. This approach enables systems to learn from information independently, without specific programming instructions. It involves explicitly using algorithms like support vector machines, random forests, or neural networks to identify unusual patterns in data gathered from IoT sensors. These algorithms draw insights from historical data, automatically recognizing anomalies without requiring predetermined guidelines [11].

D. **Time series analysis**: Time series analysis is a statistical method used to analyze and interpret data points collected consistently over time. This technique involves examining patterns, trends, and relationships in data that may change over time or be influenced by temporal factors. It is beneficial for analyzing non-stationary data, which fluctuates or varies throughout the observation period. The significance of time series analysis lies in its capacity to extract valuable insights from temporal data, allowing organizations to make informed decisions and adapt to changing circumstances. It uncovers patterns that might not be obvious in raw data, leading to more accurate predictions and better strategic planning [11].

E. **Multivariate analysis**: Multivariate analysis is a statistical technique that simultaneously analyzes and interprets complex relationships between multiple variables. It examines more than two variables to uncover patterns, correlations, and interdependencies within datasets. There are several types of multivariate analysis techniques, including regression analysis, factor analysis, principal component analysis, discriminant analysis, and multivariate analysis of variance. Multivariate analysis is instrumental when working with complex datasets where multiple factors may influence outcomes. This approach enables researchers and analysts to uncover insights that might not be evident through more straightforward analytical methods [11].

F. **False positives and false negatives**: False positives and false negatives are types of errors that can occur in binary classification or testing scenarios. A false positive, also known as a Type I error, happens when a test incorrectly indicates that a condition is present when it is absent. Conversely, a false negative, or Type II error, occurs when a test mistakenly indicates that a condition is absent when it is actually present. These

errors are significant in various fields, including manufacturing, information security, and scientific research. The implications of false positives and false negatives can be considerable. Therefore, understanding and minimizing these errors is crucial for improving the accuracy and reliability of tests and classification systems across different domains.

V. Assumptions

This paper assumes an innovative manufacturing facility producing Active Pharmaceutical Ingredients (API). This facility is highly automated and operates around the clock, running production for multiple manufacturing processes. This highly specialized environment presents unique challenges related to security and efficiency.

Every machine and device in the plant is equipped with sensors that collect real-time data on essential parameters such as temperature, pressure, speed, and other factors critical to the performance and quality of the final product. The IoT infrastructure within the plant plays a vital role in its operation. This infrastructure consists of various sensors and devices that continuously gather data. Sensors embedded in machinery and equipment capture detailed information about their operations, while IoT devices on the network facilitate data transmission and ensure secure storage.

Data communication occurs over a secure IoT network utilizing encrypted communication protocols. Sensor data is transmitted in real-time to a central platform that stores, processes, and analyzes this information. This platform is the system's core, where machine learning algorithms for anomaly detection are implemented.

The selection of machine learning algorithms is driven by the need to detect anomalies in real time from the data generated by the plant's sensors. A combination of algorithms, including convolutional neural networks (CNN) and hidden Markov models (HMM), has been chosen for their ability to identify complex patterns and subtle changes in sensor data.

CNNs are utilized to analyze data from sensors that capture thermal and visual images of industrial machinery, enabling the detection of optical and thermal anomalies in machine operation. On the other hand, HMMs are applied to time series data that track the performance of devices over time, allowing for the identification of anomalies in machine operations. This combination of algorithms facilitates accurate, multidimensional anomaly detection. Additionally, this paper assumes that the IoT infrastructure analysis of all the equipment of interest and the IT Cyber Security assessment is done.

VI. Method Design

The method design focuses on improving efficiency in an Active Pharmaceutical Ingredients (API) manufacturing facility. This paper aims to enhance the efficiency of smart manufacturing plants by implementing anomaly detection solutions. Given the exponential growth of the data generated by the IoT devices in smart manufacturing plants, Machine-learning models have become a primary tool in detecting anomalies and enabling preventive maintenance instead of corrective maintenance. This systematic mapping was guided by the methodology proposed by [7], which we utilized to identify the process steps related to using ML techniques for detecting anomalies in biopharmaceutical manufacturing with the aid of IoT devices [14]. These steps are depicted in Fig. 1.

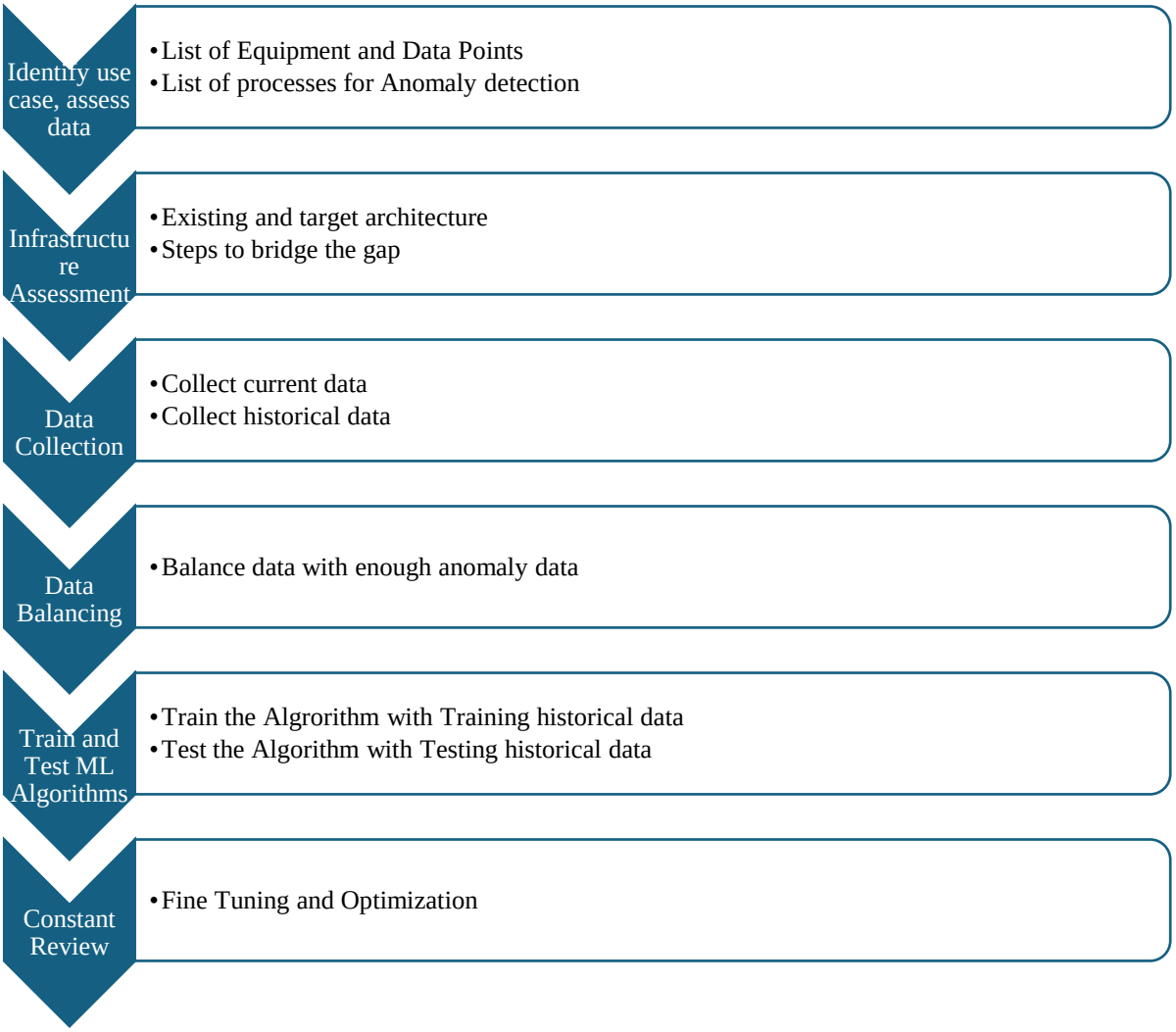

Fig. 1. Flowchart of methodological process [14]

A. **Identify use case and assess data**: The primary objective of this step is to identify the most commonly used equipment, IoT devices, and sensors across the manufacturing facility and explore the data they generate. This data must be available to export to an external database for processing and analysis. In this step, we need to investigate the application of machine learning models for anomaly detection with subject matter experts in process and manufacturing.

After the investigation, a list of equipment, IoT devices, sensors, and their data details is captured. This list will be the reference document for determining the equipment or process of interest.

B. **Infrastructure Assessment**: This step evaluates the current infrastructure to check the real-time data transfer and collection capabilities. Determining the security layers involved in data generation, transfer, and storage steps is also critical for the assessment.

After the assessment, an architecture document showcasing the existing ports, data flows, servers, devices, etc., is created. Along with this, a target architecture is built, and steps needed to bridge the gap between the current and target architecture should be of priority.

C. **Data Collection:** After identifying the use cases, assessing the generated data, and assessing the infrastructure, this step focuses on the available data for each use case. Historical data collection provides the basis for training and validating anomaly detection models in biopharmaceutical manufacturing. The historical data is generated using the manufacturing facility's sensors, computerized equipment, and IoT devices. The data is collected from these telemetry sensors, which monitor parameters such as temperature, flow rates, and pressure. The data is collected almost in real-time with a sample interval usually set at 1 second. The data can be collected with a 0.01 or 0.001 deadband for any high-frequency changes, depending on the requirements. The deadband lets the facility capture quality high-frequency data, enabling users to build more reliable machine-learning models. The users can also set maximum value, minimum value, precision, time zone, and unit of measure for each parameter. Since the data collection process is set in biopharmaceutical manufacturing, due to regulations, users should set the data confidentiality category for every parameter collected from every telemetry sensor or IoT device.

The data collection process aims to collect all required parameter data from necessary telemetry sensors or IoT devices. So, every parameter identified as needed should have a unique name containing its physical location in the manufacturing facility. The distinctive character helps the facility identify bad sensors and equipment during anomaly detection. In addition, we can investigate to find any historical data available from IoT devices or telemetry sensors, which we can utilize to capture historical data and build models.

After gathering all necessary details and setting up data collection infrastructure, users will start collecting and storing data in external databases. We will also find historical and valuable data for the model building and load it into the database in the correct format.

D. **Data Balancing**: Data Balancing is a critical step in addressing the issues faced by the uneven distribution of classes in the available dataset. Data balancing involves altering the proportions of different courses to ensure all categories are represented equally or in appropriate ratios. This process is vital for classification tasks, where an imbalanced dataset can lead to biased models favoring the majority class. During the historical and current data collection process for building machine learning models, we must ensure that regular events and anomalies are captured and labeled appropriately.

After assessing the data and use case in the first step, this information helps balance the data and ensures that anomalies and sound data are present in the database.

E. **Train and Test ML Algorithms**: Machine learning algorithms are divided into three categories: Supervised, Unsupervised, and Heuristics. There are two types of supervised models: Classification and Regression. Classification models are again divided into categories, such as neural networks, SVMs, and time-series models. The unsupervised models are divided into three categories: Outlier detection, clustering, and density estimation. Based on the review of many similar studies, we found that Neural network methods, Outlier detection, and Heuristic approaches are the most commonly used machine learning models for anomaly detection in biopharmaceutical manufacturing.

The machine learning algorithm training depends on historical data collected in the manufacturing plant. The data is carefully divided into training and test sets to ensure an objective algorithm evaluation. A typical training data set consisting of approximately 80% of the total volume of historical data available. The data

set will be divided into training and test data sets to ensure that algorithms are trained on the training data set and evaluated on the test data set. This process ensures that the algorithms can detect anomalies in the future [11].

After setting up the required infrastructure and training the models on the training data set, the models will be evaluated using the test data. The team will analyze the evaluation results to choose and deploy a suitable model for the anomaly detection process.

F. **Constant Review**: This step ensures recurring improvement in the anomaly detection process. To constantly review, a ceaseless data flow from telemetry sensors to data storage is established. Along with IoT data, the team should collect incidents, generate alerts, respond to incidents, and take corrective action.

Based on the reviews, the team must continuously improve the models to make them more efficient. These fine-tuning and optimization steps are critical to the anomaly detection process.

VII. Limitations of Anomaly detection

A. **Identifying use cases**: As mentioned earlier, identifying the use cases is one of the most critical steps in this process. Companies can mistakenly exclude the relevant details and data or ignore crucial components. This issue can be avoided by carefully fine-tuning questions for process, equipment, and business use case subject matter experts.
B. **Errors in Data Collection:** Data collected from numerous IoT devices or telemetry sensors is stored in an external database, either on-prem or in the cloud. The sensor data needs extensive metadata to be helpful in the external database for training and testing machine-learning models [12].
C. **Error in Data Balancing:** The data collected in the external database is further divided into training and testing data sets. These data sets should contain enough anomalies and sound data to train and test machine-learning models effectively.
D. **Issues with ML models:** After training and testing the models, users should assess which are more effective in finding anomalies. Although we used our best judgment to find an effective model, we may still make a mistake [12].

VIII. Conclusions

Anomaly detection in smart industrial machinery plants through the Internet of Things (IoT) and Machine Learning (ML) presents an exciting opportunity to enhance predictive maintenance, improve operational reliability, and boost overall productivity. This paper examines current research and highlights significant gaps, particularly in the biopharmaceutical manufacturing sector, where complex processes and stringent regulations demand specially designed anomaly detection frameworks. By leveraging advanced ML techniques, real-time data monitoring, and structured methodologies, industries can notably reduce downtime, enhance product quality, and mitigate operational risks. However, challenges such as data inconsistencies, model selection, cybersecurity concerns, and infrastructure limitations require continual attention to improve accuracy and efficiency. Future research should prioritize integrating artificial intelligence (AI) models, implementing real-time adaptive learning techniques, and fortifying cybersecurity frameworks to better anomaly detection in smart manufacturing. Through collaboration

and the integration of domain expertise, advancements in IoT technology, and AI-driven insights, organizations can cultivate smarter, more resilient, and adaptive industrial ecosystems [12,13]

References

[1] C. Ündey, S. Ertunç, T. Mistretta, and B. Looze, "Applied advanced process analytics in biopharmaceutical manufacturing: Challenges and prospects in real-time monitoring and control," *Journal of Process Control*, vol. 20, no. 9, pp. 1009-1018, 2010.
[2] K. Kammerer, B. Hoppenstedt, R. Pryss, S. Stökler, J. Allgaier, and M. Reichert, "Anomaly detections for manufacturing systems based on sensor data—insights into two challenging real-world production settings," *Sensors*, vol. 19, no. 24, p. 5370, 2019.
[3] A. Gupta, A. Giridhar, V. Venkatasubramanian, and G. V. Reklaitis, "Intelligent alarm management applied to continuous pharmaceutical tablet manufacturing: an integrated approach," *Industrial & Engineering Chemistry Research*, vol. 52, no. 35, pp. 12357-12368, 2013.
[4] S. G. Kim, D. Park, and J. Y. Jung, "Evaluation of one-class classifiers for fault detection: Mahalanobis classifiers and the Mahalanobis–Taguchi system," *Processes*, vol. 9, no. 8, p. 1450, 2021.
[5] Z. M. Çınar, A. Abdussalam Nuhu, Q. Zeeshan, O. Korhan, M. Asmael, and B. Safaei, "Machine learning in predictive maintenance towards sustainable smart manufacturing in Industry 4.0," *Sustainability*, vol. 12, no. 19, p. 8211, 2020.
[6] A. Tulsyan, C. Garvin, and C. Undey, "Machine-learning for biopharmaceutical batch process monitoring with limited data," *IFAC-PapersOnLine*, vol. 51, no. 18, pp. 126-131, 2018.
[7] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *12th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, BCS Learning & Development, June 2008.
[8] K. Zope, K. Singh, S. H. Nistala, A. Basak, P. Rathore, and V. Runkana, "Anomaly detection and diagnosis in manufacturing systems: A comparative study of statistical, machine learning and deep learning techniques," in *Annual Conference of the PHM Society*, vol. 11, Sept. 2019.
[9] F. Leal, A. E. Chis, S. Caton, H. González–Vélez, J. M. García–Gómez, M. Durá, et al., "Smart pharmaceutical manufacturing: Ensuring end-to-end traceability and data integrity in medicine production," *Big Data Research*, vol. 24, p. 100172, 2021.
[10] R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, "Real-time big data processing for anomaly detection: A survey," *International Journal of Information Management*, vol. 45, pp. 289-307, 2019.
[11] A. Angelopoulos, E. T. Michailidis, N. Nomikos, P. Trakadas, A. Hatziefremidis, S. Voliotis, and T. Zahariadis, "Tackling faults in the Industry 4.0 era—A survey of machine-learning solutions and key aspects," *Sensors*, vol. 20, no. 1, p. 109, 2020.
[12] A. Kharitonov, A. Nahhas, M. Pohl, and K. Turowski, "Comparative analysis of machine learning models for anomaly detection in manufacturing," *Procedia Computer Science*, vol. 200, pp. 1288-1297, 2022.
[13] J. Hochenbaum, O. S. Vallis, and A. Kejariwal, "Automatic anomaly detection in the cloud via statistical learning," *arXiv preprint arXiv:1704.07706*, 2017.
[14] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *12th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, BCS Learning & Development, June 2008.