MACHINE LEARNING PIPELINE FOR APPLICATION IN MANUFACTURING

Antje Fitzner[1], Tom Hülsmann[1], Thomas Ackermann[1], Kevin Pouls[1], Jonathan Krauß[1]
[1]Fraunhofer Research Institution for Battery Cell Production FFB, Bergiusstraße 8,
48165 Münster, Germany
`antje.fitzner@ffb.fraunhofer.de`

Hendrik Mende[2], Lars Leyendecker[2], Robert H. Schmitt[2,3]
[2]Fraunhofer Institute for Production Technology IPT, Steinbachstraße 17,
52074 Aachen, Germany
[3]Laboratory of Machine Tools and Production Engineering WZL, RWTH Aachen University,
Campus-Boulevard 30, 52074, Aachen, Germany

ABSTRACT

The integration of machine learning (ML) into manufacturing processes is crucial for optimizing efficiency, reducing costs, and enhancing overall productivity. This paper proposes a comprehensive ML pipeline tailored for manufacturing applications, leveraging the widely recognized Cross-Industry Standard Process for Data Mining (CRISP-DM) as its foundational framework.

The proposed pipeline consists of key phases, namely business understanding, use case selection and specification, data integration, data preparation, modelling, deployment, and certification. These are designed to meet the unique requirements and challenges associated with ML implementation in manufacturing settings. Within each phase, sub-topics are defined to provide a granular understanding of the workflow. Responsibilities are clearly outlined to ensure a structured and efficient execution, promoting collaboration among stakeholders. Further, the input and output of each phase are defined.

The methodology outlined in this research not only enhances the applicability of CRISP-DM in the manufacturing domain but also serves as a guide for practitioners seeking to implement ML solutions in a systematic and well-defined manner. The proposed pipeline aims to streamline the integration of ML technologies into manufacturing processes, facilitating informed decision-making and fostering the development of intelligent and adaptive manufacturing systems.

Keywords Machine learning · ML pipeline · Manufacturing · Data mining

1 Introduction

In the domain of modern manufacturing and Industry 4.0, the integration of machine learning (ML) has emerged as a driver of efficiency and innovation in manufacturing and across industries [1].ML has proven effective across a spectrum of applications in manufacturing, including process optimization, monitoring, control, and predictive maintenance [2]. The integration of ML techniques into manufacturing processes comprises an direct impact on physical systems. This transformative shift extends beyond automation, since ML systems have the capability to adapt a dynamic manufacturing environment. For example, in adaptive production workers receive real-time insights and are enabled to optimize the manufacturing process. This not only optimizes processes, but also enhances product quality, and minimizes downtime. However, the integration of ML into manufacturing also comprehends unique challenges such as certification of systems, safety protocols, guarantees of performance and reliability. The integration of ML into existing, but also new, processes has its challenges [1] that can be eased by ML pipelines. They serve as a systematic approach to streamline the end-to-end process of building, training, and deploying ML models. By delineating a clear path for the development of ML models, pipelines not only streamline processes but also enhance scalability and reliability within manufacturing environments.

1.1 Machine Learning Pipeline: CRISP-DM and Other Approaches

The Cross-Industry Standard Process for Data Mining (CRISP-DM; [3]) has long been recognized as a robust framework for guiding the development of data mining and ML projects. The development of CRISP-DM involved collaboration funded by the EU in the 90s, including prominent companies. This ensures that the standard is shaped by diverse perspectives, best practices, and real-world experiences, making it a robust framework for various industries that is still widely used. CRISP-DM consists of the following phases: *Business understanding*, *data understanding*, *data preparation*, *modelling*, *evaluation*, and *deployment*.
While CRISP-DM is a widely used and recognized framework, there are several alternative frameworks and methodologies, each with its own strengths and focus areas. Some notable alternatives are:
ASUM-DM (Analytics Solutions Unified Method for Data Mining/Predictive Analytics, [4]) is a standard process model developed by IBM for the application of data mining and predictive analytics. It serves as a revised and expanded version of CRISP-DM. ASUM-DM comprises five phases (analyze, design, configure & build, deploy, and operate & optimize) along with a project management stream. **CRISP-ML(Q)** (Cross-Industry Standard Process for the development of ML applications with Quality assurance methodology; [5]) is an extension of the CRISP-DM process, that includes two more phases: quality assurance for ML applications and monitoring & maintenance. In focus here are the identification of risks and dealing the the quality of ML applications. **KDD** (Knowledge Discovery in Databases; [6]) is an foundational methodology from the 90s that encompasses the entire process of knowledge discovery from data, including data selection, preprocessing, data trans-formation, data mining and interpretation leading to knowledge. It serves as a broader umbrella under which CRISP-DM and other methodologies operate. **SEMMA** (Sample, Explore, Modify, Model, Assess; [7, 8]) developed by SAS (Statistical Analysis System) that shares similarities with CRISP-DM. It outlines the following sequence of steps: data sampling, data exploration, modification of features, model building, and assessing their value according to the selected metrics. **DMME** (Data Mining Methodology for Engineering Applications; [9]) stands as a comprehensive extension of the CRISP-DM model, offering enhancements on the technical

part of the pipeline. Future work here suggests a more detailed definition of sub-tasks, interconnections and responsibilities.

This is a selective overview, considering the manifold of pipelines and processes designed for specific applications. Notably, some remain unpublished, like the very comprehensive **DMIE** [10], often documented only in theses. [11] and [12] compare multiple ML pipelines, providing insights into their strengths, weaknesses, and overall performance. [11] critiques CRISP-DM for neglecting tasks in project management, organization, and quality in data mining engineering, proposing an own process model. A more recent initiative describes the Data Science Process Model (**DASC-PM**, [13]) with the main objective to organize existing knowledge. The newest version concentrates on the project initiation phase, emphasizing the early establishment of crucial decisions and framework conditions for data science activities.

In conclusion, the state of the art in ML pipelines reflects a holistic approach. As the field evolves, data scientists continue to adapt and enhance each phase of the pipeline to meet the challenges and opportunities presented by the dynamic area of ML and their specific application domain.

1.2 Challenges of Machine Learning Pipelines in Manufacturing

One of the most compelling aspects of CRISP-DM and other similar frameworks is the fact that is was designed to be industry agnostic. Its design allows it to be used in a large variety of different applications [14], which explains its popularity and wide adoption across industries. Conversely, peculiarities of specific industries are not considered and can therefore not be adequately served by existing process frameworks.

The manufacturing domain has significantly profited from advances in data science and ML technology in recent time [15]. The increasing volume of data collected during production processes is used in order to decrease machine downtime and optimize processing times while at the same time improving product quality [16]. However, data driven applications in manufacturing are characterized by complex interactions between the physical and virtual world [17]. This leads to high reliability and safety requirements, as well as a high potential risk. At the same time, data sources are extensive, multimodal and often require deep process understanding, involving numerous stakeholders, to make them accessible for ML applications [18].

KRAUSS ET AL. identified unclear application areas, the availability of comprehensive, open datasets, and the lack of manufacturing specific development frameworks as the key challenges slowing the spread of ML use cases in manufacturing today [18]. They introduce a CRISP-DM based framework, in the form of the *ML Pipeline in Production*, which aims to address some key requirements of manufacturing environments. For example, by introducing two additional pipeline phases, *use case selection* and *certification*, as well as introducing superficial roles, describing the stakeholders for different phases of the pipeline [18].

While guidance on the practical implementation is given, some crucial aspects, like responsibilities within the phases and their output, are not considered by the authors. Drawing from lessons learned in conducted ML projects in the manufacturing domain, the proposed pipeline builds on the existing framework and extends it by addressing the aforementioned restrictions, as well as further refining the definitions of the pipeline phases. The extended pipeline therefore aims to help bridge the gap between theory and practical application even further and therefore facilitate the use of ML in the domain.

2 Machine Learning Pipeline for Manufacturing

This section introduces the proposed ML Pipeline, emphasizing its role as a comprehensive framework for guiding data mining and ML projects within the manufacturing domain.

2.1 General Overview & Contributions

The proposed ML pipeline provides a structured and collaborative approach, integrating the expertise of various roles, to facilitate the successful development, deployment, and ongoing maintenance of ML models. The pipeline as depicted in Figure 1 spans from *use case selection & specification* to *data integration*, *data preparation*, *modeling*, *deployment* and *certification*. It provides clarity on responsibilities to promote cross-functional collaboration, and to deliver robust and effective ML solutions tailored to use cases from the production domain.

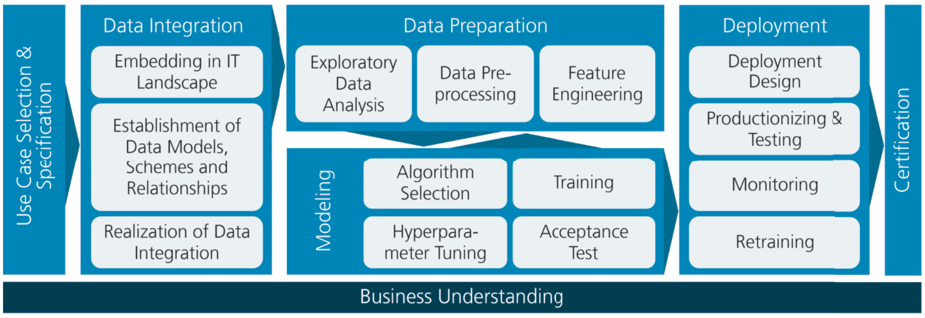

Figure 1: Overview over the proposed ML pipeline.

The pipeline therefore extends existing approaches used in the manufacturing sector in two key areas: Responsibilities and results. This is done by defining a responsible person and persons that are participating or have a consulting function for each of the pipeline sub-steps. This ensures that responsibilities are optimally allocated and provides accountability for all stakeholders and pipeline phases. These roles include persons directly involved in the pipeline, such as data scientists or engineers, as well as supporting roles such as machine operators and process experts.
Additionally, for each phase, the expected results, e.g., a document, concept or certain file, are defined. This improves the accessibility and transparency of the pipeline. As the goal of each sub-step is well defined, the transfer of the pipeline from concept into concrete ML projects is simplified. Furthermore, the expected results can be used as a reporting tool to track the progress of the active pipeline phase.

In both cases, special attention was given to the peculiarities of the manufacturing domain, such as the interplay of experts from different fields with different educational backgrounds and the strong feedback loop between the digital and physical world present in cyber physical systems (e.g. compared to recommender-systems in online shops).

2.2 Responsibilities and Results at each Sub-Step

The roles specified by the pipeline, as well as all sub-steps and their results are described in detail in the following.
The *data scientist* develops and applies machine learning models to extract insights from

data. The *data engineer* designs and maintains the infrastructure for collecting and storing data. The *IT professional* ensures the integration and security of the ML pipeline within the overall IT architecture. The *operator* oversees the operation and maintenance of the machines. The *process and production expert* oversees the production and provides insights into process and product data.

Use Case Selection & Specification: Use Case Selection

Responsible	**Participating**
Depends on the use case. Typically data scientist, data engineer and experts.	All

Results

Prioritized list of use cases with an selection of promising use case to be implemented.

Use Case Selection & Specification: Use Case Specification

Responsible	**Participating**
Depends on the use case. Typically data scientist, data engineer and experts.	All

Results

Definition of the goal of the service to be developed and of requirements and of the (qualitative) metrics. Prerequisites for implementation (data basis) are set. The scope is planned (capacity and hardware resource planning) and definition of termination criteria is finalized.

Data Integration: Embedding in existing IT landscape

Responsible	**Participating**
Data Engineer	Data Scientist, IT professional

Results

Concept how to embed the ML system in the existing IT landscape considering interactions with e.g. MES & IIoT, other existing or planned services, and enable access to the relevant data bases.

Data Integration: Establishment of Data Models, Schemes and Relationships

Responsible	**Participating**
Data Engineer	Data Scientist, Operator, Production Expert, Process Expert

Results

Definition of existing or new data models.

Data Integration: Realization of Data Integration

Responsible	**Participating**
Data Engineer	None

Results

Data defined in data model are available, accessible and prepared (e.g., basic synchronisation (of time),standardized nomenclature and units according to parameter list).

Data Preparation: Exploratory Data Analysis

Responsible
Data Scientist, Data Engineer

Participating
Operator, Production Expert and Process Expert

Results
Aggregated report about the data such as statistical analysis and visualizations. Assessment of the data quality and quantity. Definition of termination criteria based on to achieved data (e.g., quality or quantity).

Data Preparation: Data Pre-processing

Responsible
Data Scientist and Data Engineer

Participating
Operator, Production Expert, Process Expert

Results
Pre-processed data and ready for model development. The structural pre-processing is done by the data engineer, while the case-related pre-processing based on prior knowledge of the features is done by the data scientist primarily.

Data Preparation: Feature Engineering

Responsible
Data Scientist

Participating
Data Engineer, Operator, Production Expert, Process Expert

Results
New features developed for model development for this use case based on the domain knowledge of the use case. In addition, the feature extraction is performed.

Modeling: Algorithm Selection

Responsible
Data Scientist

Participating
None

Results
Determined set of suitable algorithms for this use case.

Modeling: Hyperparameter Tuning

Responsible
Data Scientist

Participating
None

Results
Strategies for efficient and automated hyperparameter-tuning, suitable for the selected algorithm.

Modeling: Training

Responsible
Data Scientist

Participating
Data Engineer, Operator, Production Expert, Process Expert

Results
Determine the conditions for the training. This covers the decision on the data split or the usage of cross validation and the definition of training parameters such as learning rates. The technical metrics to evaluate the model performance in accordance to the qualitative requirements for the use case are set.

Modeling: ACCEPTANCE TEST

Responsible	**Participating**
Data Scientist	Data Engineer, Operator, Production Expert, Process Expert

Results
Selection of the final algorithm, hyper-parameter and training parameters and evaluation of the model performance in alignment with use case requirements. All is validated with respect to the existing process.

Deployment: DEPLOYMENT DESIGN

Responsible	**Participating**
Data Engineer	Data Scientist, IT professional

Results
Selection of the deployment pipeline, e.g. the usage of Docker or other process managers, and the final embedding into the overall IT landscape as planned in the earlier phases.

Deployment: PRODUCTIONIZING & TESTING

Responsible	**Participating**
Data Engineer, Data Scientist	Operator, Production Expert, Process Expert

Results
Performed integration (including egde cases), penetration and performance test.

Deployment: MONITORING

Responsible	**Participating**
Data Scientist	Data Engineer, Operator, Production Expert, Process Expert

Results
Definition of the monitoring strategy and implementation (including retraining).

Deployment: RETRAINING

Responsible	**Participating**
Data Scientist	Data Engineer

Results
Definition of retraining strategy, e.g. implementation of automated retraining, based on the defined requirements.

Certification: CERTIFICATION

Responsible	**Participating**
Production Expert (May be third party)	All

Results
Certification of the system. System is ready to be used.

3 Discussion and Validation

In ML projects, the participation of different roles often leads to misaligned expectations and communication gaps [19], as well as decision-making complexities. To overcome this, it is essential to encourage collaboration, implement clear communication, and ensure a common project vision. CRISP-DM lacks clear definition of responsibilities among team members and falls short regarding project management compared to e.g., ASUM-DM. CRISP-DM also overlooks the crucial aspect of data acquisition and does not incorporate the concept of a *proof of concept* as an initial project phase.
The proposed ML pipeline ensures that every project phase aligns with the specific

needs and challenges of the manufacturing environment. One of the key advantages is the explicit definition of responsible roles throughout the project, by clearly outlining the roles of five key stakeholders. In this way, the pipeline promotes a collaborative and interdisciplinary approach. Furthermore, the pipeline provides clarity regarding the expected output at each phase of the project. This allows for better planning in advance, improved monitoring capabilities, and helps overcome communication challenges.

The number of iterations needed between the data preparation and modeling phases is not specified, as it can vary significantly based on the specific use case. Use cases involving extensive data exploration, cleaning, and feature engineering may require additional iterations compared to simpler tasks, again emphasizing the importance of adapting the pipeline to the requirements of each specific use case. The same is true for the weight given to different phases. In some situations, entire phases e.g., certification may be skipped, leading to a tailored approach based on given requirements.

Fraunhofer IPT and FFB used the ML pipeline as the structure for about 20 ML-projects, such as various work on glass molding and laser structuring or production in general [20, 21, 22]. The results from these projects validated the functionality of the presented ML pipeline. In MENDE ET AL. [20] feedback from the project members emphasized the effectiveness of the ML Pipeline through its proficient facilitation of model development and deployment within the manufacturing domain. A notable result was the creation of a clear communication that encouraged effective teamwork among various groups, such as machine operators, process planners, and data scientists. The pipeline showed its applicability to user-centred development of ML solutions by incorporating feedback from machine operators, tailored to their specific needs in daily operations.

Explicitly not addressed by the proposed pipeline are concrete timelines for the different phases. The timeline for any given ML project significantly depends on the complexity of the task at hand. However, flexibility is maintained by the pipeline to accommodate variations in project timelines based on varying task complexities.

In summary, the integration of ML into manufacturing processes has a significant impact on physical systems, going beyond automation to enable adaptive production. This results in real-time insights. Hence, the implementation of a tailored ML pipeline for manufacturing provides a structured framework facilitating the integration of ML technologies into manufacturing operations.

4 Conclusion

In ML projects, addressing the challenges arising from diverse roles in the manufacturing domain requires establishing a clear project management approach. The proposed pipeline does so by complementing existing pipelines through the addition of stakeholder responsibilities and anticipating the concrete output of the different phases. Some aspects, like timelines or number of iterations, which are mostly use case dependent, have purposefully been omitted. Validation across different use cases is yet to be completed, with a basic use case already having been addressed. Ongoing efforts are directed towards extending the validation to further scenarios, ensuring the robustness and applicability of the proposed method to a wide array of manufacturing use cases.

Acknowledgments

This work was supported under project “FoFeBat - Forschungsfertigung Batteriezelle Deutschland (03XP0256, 03XP0416, 03XP0501A) and “AI-NET-ANIARA” both funded by the German Federal Ministry of Education and Research (16KIS1275).

References

[1] Wuest et al.: Machine learning in manufacturing: advantages, challenges, and applications, Production & Manufacturing Research, 4:1, 23-45 (2016).

[2] Plathottam et al.: A review of artificial intelligence applicationsin manufacturing operations, J. Adv. Manuf. Process., 5(3), e10159 (2023).

[3] Shearer: The CRISP-DM model: the new blueprint for data mining. In Journal of data warehousing 5 (4), pp. 13–22 (2000).

[4] Angée et al.: Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary Multi-organization Big Data & Analytics Projects. In: Knowledge Management in Organizations. Cham: Springer International Publishing, pp. 613–624. (2018)

[5] Studer et al.: Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology (2020).

[6] Fayyad et al.: From Data Mining to Knowledge Discovery in Databases. In American Association for Artificial Intelligence (1996).

[7] SAS Institute: SEMMA data mining methodology (2008) http://www.sas.com

[8] Azevedo and Santos: KDD, SEMMA and CRISP-DM: A parallel overview (IADIS), (2008).

[9] Huber et al.: DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model, Procedia CIRP, Volume 79, Pages 403-408, ISSN 2212-8271 (2019).

[10] Solarte: A proposed data mining methodology and its application to industrial engineering. Master’s thesis, University of Tennessee (2002)

[11] Marbán et al.: Toward data mining engineering: A software engineering approach. In Information Systems 34 (1), pp. 87–107 (2009).

[12] Mariscal et al.: A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review.; 25(2):137-166 (2010)

[13] Schulz et al.: "DASC-PM v1.1 - A Process Model for Data Science Projects", NORDAKADEMIE gAG Hochschule der Wirtschaft, Hamburg 2022, ISBN: 978-3-9824465-1-6

[14] Mariscal et al.: A survey of data mining and knowledge discovery process models and methodologies, The Knowledge Engineering Review. Cambridge University Press, 25(2), pp. 137–166 (2010)

[15] Jourdan et al.: Machine Learning For Intelligent Maintenance And Quality Control: A Review Of Existing Datasets And Corresponding Use Cases. CPSL 2021. Hannover: publish-Ing., (2021), S. 499-513.

[16] Kim et al. Smart Machining Process Using Machine Learning: A Review and Perspective on Machining Industry. Int. J. of Precis. Eng. and Manuf.-Green Tech. 5, 555–568 (2018).

[17] Monostori et al., Cyber-physical systems in manufacturing, CIRP Annals, Volume 65, Issue 2, 2016, Pages 621-641, ISSN 0007-8506 (2016).

[18] Krauß et al.:Application Areas, Use Cases, and Data Sets for Machine Learning and Artificial Intelligence in Production (2023).

[19] Nahar et al.: Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process (2021).

[20] Mende et al.: Multi-target regression and cross-validation for non-isothermal glass molding experiments with small sample sizes, Proc. SPIE 12778, Optifab 2023, 127780S (2023).

[21] Motz et al.: Benchmarking of hyperparameter optimization techniques for machine learning applications in production, Advances in Industrial and Manufacturing Engineering, ISSN 2666-9129 (2022).

[22] Leyendecker et al.: Predictive Quality Modeling for Ultra-Short-Pulse Laser Structuring utilizing Machine Learning, Procedia CIRP, Volume 117, Pages 275-280, ISSN 2212-8271 (2023).