Proceedings of the 34th European Modeling & Simulation Symposium (EMSS), 006
19th International Multidisciplinary Modeling & Simulation Multiconference

ISSN 2724-0029 © 2022 The Authors.
doi: 10.46354/i3m.2022.emss.006

Data-driven recipe optimisation based on unified digital twins and shared prediction models

René Wöstmann[1] *, Thorbjörn Borggräfe[1], Sascha Janßen[1], Josef Kimberger[2], Solomon Ould[3], Nick Bennett[3], Victor Hernandez Moreno[3] and Jochen Deuse[1,3]

[1]RIF Institute of Research and Transfer e. V., Joseph-von-Fraunhofer-Str. 20, Dortmund, 44227, Germany
[2]Bitburger Braugruppe GmbH, Römermauer 3, Bitburg, 54634, Germany
[3]Centre for Advanced Manufacturing, University of Technology Sydney, 15 Broadway, 2007 Ultimo, Australia

*Corresponding author. Email address: rene.woestmann@rif-ev.de

Abstract

The importance of cross-process multivariate data analysis for improving products and processes is continuously increasing. Artificial intelligence and machine learning offer new possibilities to represent complex cause-effect relationships in models and to use them for optimisation. For consistent and scalable usage, unified data structures and representations of products, processes and resources are required in order to be able to use larger data populations as well as deploy these models in different application contexts. The paper presents an approach of shared prediction models for recipe optimisation based on unified digital twins in the beverage industry. For this purpose, a central generic data model was created, which is the basis for unified digital twins and thus the integration of physical and digital entities, as well as the foundation for cross-process data analysis.

Keywords: Machine Learning, Recipe Optimisation, Digital Twins, Beverage Industry

1. Introduction

In a globalised world, food and beverage manufacturers are confronted with rising costs for energy, resources and wages as well as diverging customer and ecological requirements. Existing optimisation paradigms are increasingly reaching their limits, as simple cause-and-effect relationships are usually analysed from production perspective and customer demands are assessed on the marketing side without closing the loop to the shopfloor.

On the other hand, *digital twins (DT)* as the basis for concepts like *Industry 4.0* or the *Industrial Internet of Things (IIoT)*, offer new possibilities for intelligent data usage in the food and beverage industry (Schmitz & Hagemann, 2020). Approaches of *artificial intelligence (AI)* based on *machine learning (ML)* methods offer new possibilities for building multivariate prediction models in order to optimise recipes regarding productivity indicators and customer requirements, and provide the ability to test recipe changes before producing them (Robinson & Dehbi, 2021). The main challenges are a lack of data availability and quality, creating deployment-oriented ML architectures and the scalable usage of models within applications and services (West et al., 2021; Wöstmann et al., 2020). In the present case of the beverage industry, the problem relates to the fact that individual data silos for the same or similar processes exist in particular, but are not merged and used across locations. Integrating data silos across locations and value chains can significantly increase the quality of models and in some cases make application scenarios possible in the first place.

The authors present an approach to use shared prediction models across locations based on unified

DTs in order to optimise recipes. The approach uses data on the product ingredients, performed processes and customer feedback on the final product. Section 2 introduces existing work on cyber-physical systems, DTs and ML as well as recipe optimisation. Section 3 contains the concept for building unified DTs for recipe optimisation across locations, the implementation of which is presented in section 4.

2. State of the art and related work

2.1. Cyber-physical systems and digital twins

Increasingly cost-effective sensors and data storage technologies, embedded systems, new possibilities for data analysis and simulation, communication standards and cloud services are the main enablers for building *cyber-physical systems (CPS)* (Derler et al., 2012). A central element of CPS is the DT that on the one hand is specified from product, process and resources viewpoints and on the other hand, contributes to standardisation activities like the *asset administration shell (AAS)* (IEC 63278-1 ED1), *digital factory framework (DFF)* (IEC 62832-1:2020) or *equipment behaviour catalogues (EBC)* (ISO 16400-1:2020). A further differentiation is made in the integration of data and information flows. As shown in Figure 1, precursors of a DT are the *digital model* with a manual data flow between physical and digital objects, and the *digital shadow* with an automated unidirectional data flow from the physical to the digital object (Kritzinger et al., 2018). In the context of this work, the DT is defined as a digital, dynamic representation of real products, processes and resources with an automated data flow between physical and digital objects (Fuller et al., 2019). In the beverage industry, implementation has so far only been based on individual machines or processes (Perno et al., 2022) or supervising value chains (Werner et al., 2021). There is currently a lack of comprehensive data linkage of different processing steps along the process chain as well as feedback of analysis and modelling results to enable holistic data-driven product and process optimisation.

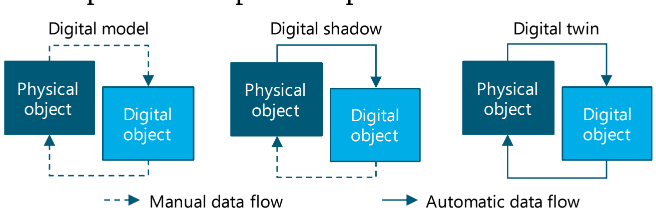

Figure 1. From digital model to digital twin (Kritzinger et al., 2018)

2.2. Artificial intelligence and machine learning

Machine learning as a subset of AI offers vast potential in the data-driven analysis and explication of multivariate cause-effect relationships in large cross-process data sources (West et al., 2021). A distinction can be made between *supervised ML methods* in which the result of an observation (so-called label) is known (e.g. classification and regression), *unsupervised ML methods* in which the result is unknown (e.g. clustering, association analysis, dimension reduction), *semi-supervised ML methods* if a label is available for parts of the observations, and *reinforcement learning*, in which an agent learns a strategy independently from historical and new data sets based on rewards for actions (Alpaydin, 2020). (Ge et al., 2017) provide an overview of the most commonly used ML methods in the process industry. In particular, *supervised methods* are suitable for making inferences about causes and thus for initiating recipe adaptations. There are numerous applications for specific process steps, but there is no systematic application across the process chain. Also, the concept of *transfer learning* is of interest, as it consists of applying models to different framework conditions (e.g. analogue *products* or *processes*). *Unsupervised* and *semi-supervised learning* is primarily used for anomaly detection. *Reinforcement learning* applications in the industry are still in the early stages, but emerging (Xia et al., 2021).

At the implementation level, ML requires the use of appropriate frameworks and platforms that must harmonise with existing IT systems. The market for concrete solutions is diverse and complex. On the one hand, there is a tendency towards open programming languages like *Python* and *R*. On the other hand, graphical programming interfaces and data science platforms like *RapidMiner* are becoming increasingly important, as they offer easy access for those without data science expertise. In addition, hyperscalers like *Amazon (AWS/Sage Maker)*, *Microsoft (Azure Machine Learning)* or *Google (Vertex AI or TensorFlow)* offer comprehensive frameworks (Grum et al., 2020). A reference architecture for ML in the process industry was developed in the *DaPro* research project, which assists users in the selection and design of ML frameworks and infrastructures while building on existing IT landscapes in companies (Wöstmann et al., 2020).

2.3. Data-driven recipe optimisation

In general, in the food and beverage industry, various *reactants* are converted into *products* through *processes*. Whereas in discrete manufacturing technical drawings and bills of materials describe *products*, in food and beverage industry this information is stored within recipes. The recipe model of *ISA-88* resp. (IEC 61512-1:1997) is the leading international standard and defines a recipe as "*the necessary set of information that uniquely defines the production requirements for a specific product*", specified in the four types of *general*, *site*, *master* and *control recipes*. Recipes thus describe both *products* and their *processing* route.

Existing approaches to recipe optimisation primarily address quality and productivity indicators, e.g. higher outputs with minimised use of resources or increased process stability. However, data-driven approaches have so far only considered simple statistics to be used as soft sensors, without focusing on a predictive multivariate model application (Ge et al., 2017). First approaches of using ML are emerging, but they rarely get into deployment. The most common problems are inconsistent data structures and poor data quality. In addition, often only isolated data silos are evaluated, resulting in individual use cases whose initial effort does not scale (West et al., 2021).

In a second model-driven research area, simulation models are used to predict reactions and results of processes, e.g. (Ban et al., 2021; Garrido-Merchán & Albarca-Molina, 2018). Here, simulation methods are based on known interrelationships and behavioural modelling approaches. Weaknesses are long calculation times and a lack of adaptability to varying conditions within the deployment in practice. In addition, they neither address the discovery of complex unknown patterns and relationships nor are they applied across process chains with the integration of the customer.

In a third field of related work, marketing departments carry out customer analyses in order to recognise market trends and to convert them into *product* features (Tarallo et al., 2019). Numerous start-ups such as *IntelligentX* or *BRAUERAI* are currently being formed for this purpose (Brewer World, 2020). They represent the largest overlap to the presented approach, but do not provide direct control loops to machine and plant parameters nor adaptable solution patterns for the industry.

The present approach contributes to mapping the entire process chain from the customer via the production processes to the ingredients downstream using digital product, process and resource twins, to use them for collaboratively recipe optimisation.

3. Concept of digital twin for recipe optimisation

As described in section 2 a DT can have different manifestations relating to its context. Commonly it consists of the three *components physical (PE)* and *digital entity (DE)* and their *interconnection*, that are enlarged in recent research to five components, adding *DT data* and *applications* (often related as services) (Tao, Nee, & Zhang, 2019). The requirements for the DT define the scale and extent to which the PE is considered. A first decision addresses the hierarchical level (Tao, Zhang, et al., 2019). Recipe optimisation can be an overarching task at *enterprise*, *site* or *process cell* level, but also lead to interference in *unit*, *equipment* or *control modules* of *ISA 88*. Therefore, a DT can consist of various sub-models, for which the interaction and dependencies have to be considered. EBC and AAS provide generic possibilities of describing equipment and asset compositions and capabilities. Figure 2 shows the general concept of the presented approach. In this section, the individual components are discussed in more detail for recipe optimisation and the possible manifestations of these entities.

3.1. Physical entity

The PE represents all relevant elements of the physical world. For the recipe optimisation, the focus is on products (e.g. ingredients and the final product), processes (e.g. procedures, reactions and process steps) and resources (e.g. asset and equipment). The product and process to be optimised takes place in an asset and is performed with equipment, which are summarised as resources. An asset can carry out several processes by varying equipment or process parameters. Since the processes have a dependence among themselves, all processes, which are accomplished by this asset belong also to the consideration on unit level. In addition, the resources that are made available to the asset for the execution of the process and the product that is created by the *process* belong to the unit on this observation level.

The characteristics of the product range from biological ingredients as in the food and beverage industry, to chemicals in the chemical industry, to mineral raw materials in the steel industry, but combining a batch-oriented processing. Intermediate products from other processes can also be a part of a final product. The process is mapped in the DE by bringing together the accumulated knowledge about the process and a correct process execution. This includes all the individual process steps that represent the process as a whole, the ingredients required for it as well as characteristics of a final product. The recipe optimisation addresses processes as well as associated products, conducted by resources. Therefore, an adequate mapping has to be considered by creating a DE.

3.2. Digital entity

The DE can be based on simulation software and shopfloor IT systems, that represent product, process and resources in digital models. In the food and beverage industry, processes have many complex influencing variables from causes to effects, leading to a large history of simulation and automation systems. A general distinction can be made between simulation approaches for product and process engineering as well as shopfloor IT systems for running production systems (prostep, 2019). Also external supplier or customer data can be part of the DE of a product. All of these models can, but do not necessarily have to, represent part of a DT.

Geometric models consider geometrical dimensions of the DE, e.g. two- and three-dimensional technical drawings. In process industries, these models can be used to plan operations before a process is carried out, so that equipment is correctly dimensioned, including piping and instrumentation. Other simulation methods address the physical or behavioural simulation of production processes. Examples are computational fluid dynamics (CFD) or thermodynamically simulations (Bröckner et al., 2021). For a more in-depth review of simulation tools and their application to DTs, see (Longo et al., 2021).

While numerous simulation and modelling approaches are applied in experimental product development and the early stages of production system planning, the shopfloor IT provides the second pillar of the DE. Based on the automation pyramid of the ISA 88, the systems are usually hierarchically structured and go from the programmable logical controllers (PLC) for the direct control of actuators, including high granularity information and data points over to process control systems (PCS) and manufacturing execution systems (MES), in which key figures for reporting are aggregated and control tasks are taken over (Wöstmann et al., 2020). There are also parallel systems for manufacturing data acquisition, laboratory information management system (LIMS), batch management, logistics and enterprise resource planning (ERP), which have different perspectives and subcomponents on the product, process and resources in real production and therefore use different data models. The multitude of modelling methods of product, process and resources as well as their underlying systems and data management make it impossible to postulate a universal definition and implementation of the DT. The following section therefore presents a data model for the DT for recipe optimisation, which serves to harmonise the required data across locations for using shared prediction models.

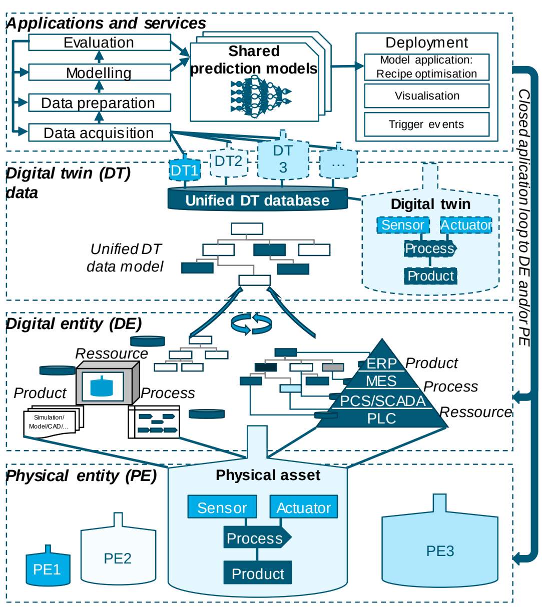

Figure 2. Technical concept of unified DT for shared prediction models

3.3. Digital twin data model

In the context of data-driven recipe optimisation, it is required to specify the data for the desired DT viewpoint. Important recipe data can be classified into planned and measured data about *product*, *process* and *resource*.

The planned recipe data can be derived from the recipe model of *ISA 88*, which introduces four hierarchical recipe types with *general*, *site*, *master* and *control recipes* from *enterprise*, *site*, *process area* to *process cell*. The (IEC 61512-2:2001) contains data models for the standardisation of interfaces in batch-oriented processing. Unique recipe and batch IDs are important, however, there are degrees of freedom in the definition of parameters. Information on the plant and its equipment is stored in the asset data, which can include dimensions, location, connections and equipment behavioural capabilities. Data about installed sensors and units are also part of a hierarchical DT representation (Barthelmey et al., 2019) .

Within the presented approach, measured data include time series of processes as well as product samples, both based on sensors. Sensor data on product and processes as well as equipment conditions represent the DT at any time. All execution steps that are carried out by humans must also be documented and stored. Data about the product can be product specific, but also contain batch information. Product data include laboratory values, shelf-life data, key figures from tests or other parameters that characterise the product and its ingredients. These data can be measured or determined by the manufacturer or provided by suppliers, e.g. via outbound analyses.

A generical data model is presented in appendix A (cf. Figure 7) for describing required data for recipe optimisation in a unified structure. In each case, it has to be mapped with existing IT structures as well as a DT database for unified data acquisition. SQL databases are suitable for all types of structured data. For semi-structured or unstructured data (e.g. images or files), other forms of storage must be chosen. In order to map existing IT systems and databases to the DT data model, the next section addresses the connections between the DT elements.

3.4. Connection

A technical implementation has to take care of the interconnection of assets and the connectors that translate the perception and manipulation of the PE to the DE (cf. Figure 2) as well as applications and services (section 3.5). On a local level the task is usually covered by the domain of automation technology. Now, however, considering the increasingly distributed nature of IT infrastructure, which can also host the DT, responsibility is shifting towards the domain of IIoT (Heidrich & Luo, 2016). Leveraging IIoT technologies allows unique addressing of individual models and services of and by the DT in near real time, negating eventual physical distances between assets and services hosted in cloud-environments in various industrial contexts (Schmitt et al., 2020).

For this interconnection, all layers of the *OSI model* (cf. Haupt, 2020) need to be addressed. While physical and transport layers are standardised in the industrial context (e.g. by IEEE 802.3 & .11, IP and TCP/UDP), technologies addressing the *machine-to-machine (M2M)* communication in the application layers are transitioning away from proprietary protocols to standardised, partially open-source alternatives such as *OPC UA* or *MQTT*. Also, more unspecific IT-related protocols and *application programming interfaces (APIs)* found in classical "web-stacks" such as *WebSocket* or *REST* can be utilised where resource and bandwidth allows (Al-Fuqaha et al., 2015). Combined with efforts to the likelihood of AAS, DFF and EBCs, this allows a manufacturer-independent, demand-based link of assets and services, optionally also including additional enterprise IT such as ERP or MES, where PE and DEs are equivalent and interchangeable.

There is an abundant supply of standardised protocols for M2M-communication in the IIoT, covering both long existing (e.g. *Siemens TCP* on top of *RFC 1006* or *Modbus*) as well as new ones like *6LoWPAN*, addressing the specific needs of IIoT networks on all OSI layers (cf. Arndt, 2018). This abundance currently complicates construction, commissioning and reconfiguration of CPS, suggesting a necessary concentration on the utilisation of unified, well-established openly available protocols and interfaces to ensure easy interoperability of entities (cf. Pethig et al., 2017). Perceived as particularly well established are e.g. the *Message Queuing Telemetry Transport-Protocol (MQTT)* and *OPC UA*, standardised in (IEC TR 62541-1:2020), which unifies M2M and *human machine interface (HMI)*-communication, abstracting the complete information model. Furthermore, there are standards also addressing the physical levels such as *IO-Link* (IEC 61131-9:2013) and *Bluetooth Low-Energy (BLE)*. *REST*-based implementations like the *Constrained Application Protocol (CoAP)* close the arc to "classical" web-protocols. These can and should be used for application-related communication purposes like database interactions and UI/UX. Furthermore, there are also XML-based protocols for more complex data like *XMPP* or *AutomationML*. (Ahrend et al., 2019; Al-Fuqaha et al., 2015; Barthelmey, 2021)

Combined, these approaches and technologies allow an interoperable IT architecture to ensure secure and targeted delivery of relevant data as well as the easy development of new features, functions and models in the context of DTs.

3.5. Applications and services

To create value, a DT has to provide applications and services for decision-making, e.g. intelligent production planning, anomaly detection, quality prediction or process control. As shown in Figure 1, the goal of a DT is a closed loop between PE and DE. Applications addressing recipe optimisation of the PE are mainly concerned with controlling or optimising the processes and product composition, based on data sources building the DE.

Also, simulation methods can be the base for applications and services, when integrated in closed loops. A disadvantage is the focus on specific simulation tasks and high computing times, which make rapid deployment across process chains and multi-dimensional problems difficult. For this reason, the presented approach proposes the use of ML as basis for the application layer. It offers the following advantages: First, complex multivariate patterns and correlations can be obtained cross-process based on historical data from the PE with large amounts of influencing factors. Second, the underlying methods and procedures are able to represent knowledge in the form of statistic models, that can provide readable outputs for both humans and machines. Third, ML offers great advantages in deployment applicability and can be provided (e.g. encapsulated as services) in different ecosystems. In this context, first, the technical design of an ML environment is important, in which different DTs can be managed in unified structures. (Wöstmann et al., 2020) provide assistance for the architectural design. Second, a major challenge is the deployment up to the triggering of events. Models and their outputs must either be made available to the employees in such a way that they can generate added value in the operational work, or interact directly with the technical control (e.g. mixing ratios in the MES or adaptive process control at PCS or at PLC level). Another challenge is the continuous monitoring of performance and quality metrics of the models as well as the closing of ML control loops, e.g. re-training of models when new, larger data sets are available. Here, new research disciplines are emerging around *MLOps*, for which technical solutions are also to be provided in the ML environment (Renggli et al., 2021). In addition, a suitable business model must be established. Within larger companies, this could be, for example, cross-location production networks in which assets of the same or similar type learn from each other. On the other hand, there are efforts by machine and plant manufacturers (e.g., Syskron in the beverage industry) to manage similar assets of their customers in digital platforms and to provide data-driven services. While Figure 2 summarises the technical concept, the following Section 4 presents an exemplary case study in the brewing industry.

4. Data-driven recipe optimisation in cyber-physical brewing labs

The case study consists of two identical cyber-physical pilot breweries located in Dortmund, Germany and Sydney, Australia. The aim is to predict and optimise the quality of the product beer, which is assessed by sensory analysis in human tasting. Within the process chain of brewing, numerous multivariate influencing factors arise on the final product. In the following, the paper demonstrates how unified data structures and models in the context of the DT constitute the basis for shared prediction models and their application for recipe optimisation. While Figure 3 gives a representation of the PE in Dortmund, Figure 4 provides an overview of the case study.

4.1. Physical entity of the case study

The PE consists of all products, processes and resources that can have an impact on the final product quality. The cyber-physical brewing assets consist of respectively one *Ziptech NANO* brewhouse of 0,5 hl and three fermentation tanks with a capacity of 1,2 hl each. Each asset carries out brewing (mashing, boiling, lautering, whirlpool) as well as fermenting and maturing processes. The brewhouse consists of two boilers, an HMI to control the asset, a pump and a piping system. In addition, temperature, flow and volume sensors are installed for condition monitoring and automation. The fermentation tanks also provide temperature sensors and a cooling unit for temperature control.

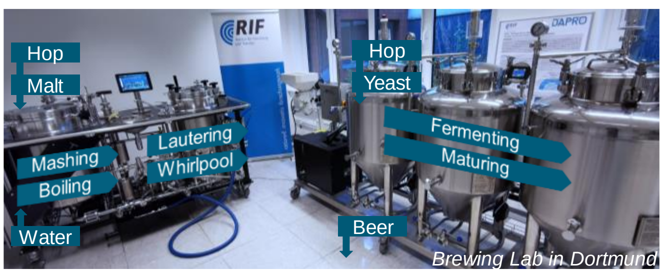

Figure 3. PE of the cyber-physical brewing lab in Dortmund. An identical system exists in Sydney.

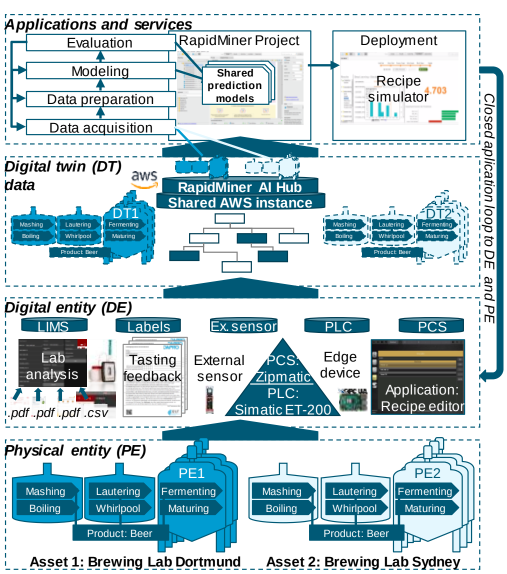

Figure 4. Overview of the case study of shared prediction models for recipe optimisation in cyber-physical brewing labs

The *product* is beer, which is characterised by its ingredients and processes, that are formalised in a recipe. There are numerous works on quality assessment of separate indicators, which address the influence of selected attributes on specific quality figures. For this purpose, the test criteria for beer of the *German Agriculture Society (DLG)* were used, which in the form of a so-called sensory analysis (tastings) include foam stability, clarity and smell as well as the taste parameters of full-mouth feel, freshness and bitterness impression (Narziß et al., 2017). Since the authors brew according to the German purity law, the ingredients consist of hops and malt of different types and mixing ratios as well as yeast and water, which can be varied in quantity in the case study. Malt is the most expensive ingredient and is available in many different varieties. Hop is very important for the quality, foam and taste of the beer. While hop acids influence bitterness and freshness, the ethereal oils from aroma hops have a major influence on the flavour profile. Yeast is of high importance, since it converts fermentable sugars into alcohol and produces numerous fermentation side-products, which contribute to the flavour profile as well. The following section describes how the physically existent products, processes and resources are digitally represented.

4.2. Digital entity of the case study

In order to map the PE to a DE, various IT systems are required. Within the brewing assets, the PCS *Zipmatic* is the central system for defining control recipes, controlling the automated brewing processes, as well as performing condition monitoring and process data acquisition. Aggregated process and asset data are stored locally in a *PostgreSQL* database of the PCS in each location. Higher resolution and more direct access to process data is provided by a *Simatic ET-200EP* PLC, which is controlled by and returns information to the PCS. Since the PLC only communicates bidirectional with the PLC in its original state, but does not have its own database, an edge device was implemented to collect raw data of the asset. Furthermore, in the delivery state of the plant, the fermentation process cannot be measured. Therefore, another external sensor (*Tilt Pro Hydrometer*) was installed, which continuously measures the respective density from the tilt angle. The data is stored in *Google Spreadsheets* using the edge device.

On the *product* side, the data acquisition is heterogeneous. Each manufacturer of yeast, hops and malt provides different product sheets and laboratory analyses in .pdf format. In addition, various measurement processes for intermediate products take place in the brewing process. They include, among others, an iodine sample, but also density measurements (e.g. original gravity, kettle full wort, malt wort, ...) using an *Anton Paar EasyDens*, which enables to export samples as .csv files from the corresponding application. For the integration of the heterogeneous product data, a LIMS with a user interface (GUI) was created that enables the manual product data creation as well as the linkage with the brewing process. The product data of the DE are completed by the final quality assessment through sensory analyses. The data for the tastings is systematically stored on the *Microsoft Forms* platform.

4.3. Digital twin data model of the case study

In order to unify DE's heterogeneous data landscape in terms of required data and provide the basis for shared prediction models, a DT data model was specified based on the data model presented in Section 3. It is presented in the appendix A in Figure 8. It forms the basis for two implementation approaches of data management and distribution. On the one hand, a relational *MariaDB* is used to capture all product and batch-related data as the backbone for the DT product data. On the other hand, an *influxDB* is implemented for the time series acquisition of the DT process data.

The database schema derived from the data model includes individual tables and connections between the tables. The product, in this case the beer, is at the centre of the data collection. Each beer has a name and a unique ID. Each brewing process is only used for one beer, so the ID of the beer can be stored directly in the brewing process and the relationship between beer and brewing process can be established. In addition, a fermenting process can be assigned to each beer and each fermenting process can be uniquely assigned to a beer. The ID of the beer can therefore also be stored directly in the fermenting process. The quality characteristics of the beer are also recorded and can be uniquely assigned to each beer via the stored beer ID. All other tables were linked to the product according to the same principle.

AWS has been selected as cross-site and unified platform. Here, both unifying databases as well as the ML environment for implementing recipe optimisation services are hosted. For creating and using shared prediction models, the collaborative data science platform *RapidMiner AI Hub* was selected. In the following section, the connections are presented in more detail.

4.4. Connection of the case study

The architecture of the brewing lab demands the ability to connect various entities to collect both planned and measured product, process and resource data (ref. 3.3), spread widely in geographical terms. The choice of methods of connection was, next to the given opportunities by the physical assets, driven by compatibility and open availability. The brewhouse asset with the fermentation tanks already utilise *OPC UA* for communication purposes between PLC and HMI, however warranty and security concerns prohibited direct manipulation or interference with internal communication.

The physical data acquisition thus had to be delegated to a dedicated edge device, which can also handle additional connections to external sensors such as *Tilt Hydrometers*. It is possible to employ a vast selection of low-cost development boards such as *Nvidia Jetson* or, in this case, *Raspberry Pis* for this. Its main task is to physically connect to the PLC and additional sensors and act as a gateway to global network and the *AWS* cloud-instance. The devices are securely managed remotely by utilising fleet management solutions, in this case *openbalena*. This abstracts specific hardware requirements for the edge devices, enabling one-click deployment and rapid provisioning.

Remotely deployed, dedicated containerised micro services handle individual tasks. A low-code development environment hosts firmware that pulls necessary process data from the PCS. It can also host the standardised input GUI for recipe and product data that users can access with any browser if necessary. It is also possible to integrate *Tilt Hydrometer* administration and data polling via *BLE-Beacon* firmware. The services also include a robust and lightweight middleware to scrape high-frequency PLC process data via *OPC UA* and sensor data from *Tilt Hydrometer* via *MQTT*. It thereby also serves as the abstraction layer to route individual asset data into the unified data model. The service sends all data securely to the cloud-based NoSQL *InfluxDB* via *WebSocket*. A NoSQL-database utilisation at this point merits vastly improved performance and reduced data footprint for high frequency time series data. It is also possible to pre-process this data stream to optimise footprint even further with an additional dedicated service. A local backup database service handles individual asset data for redundancy. Deployment approaches on the edge and cloud are similar to reduce complexity and allow scaling (cf. Ahrend et al., 2019). In this case, together these services form the proven *TICK* stack (cf. Nasar & Abu Kausar, 2019). Overall, this approach allows rapid adaption and deployment for the different PEs.

The cloud-hosted *MariaDB* holds static process and recipe data pulled from the HMI and provided by manual input (see Appendix A). Applications and services such as GUIs for dashboarding and business intelligence as well as the data science platform *RapidMiner AI Hub* for data-driven recipe optimisation (detailed in section 4.5) are able to access data directly. The cloud hosting approach allows high scalability and uncoupling of data acquisition and processing (cf. Ahrend et al., 2019).

The focus of work so far is concentrated on the connection of the PE and DE to a DT model. For closing the loop back to the PE and DE, the application layer is detailed in section 4.5. The connection from services and applications is implemented as an assistant system for brewing experts in the form of a recipe simulator. It is directly accessible by the user and supports the optimisation of chosen ingredients before the next batch as well as parameter optimisation of process control via the HMI.

4.5. Application of the case study (recipe optimisation)

For the task of optimising recipes, the use of ML is chosen. The application can have different scopes, addressing on the one hand improving the final product taste (e.g. for craft breweries and experimental batches) and on the other hand improving the utilisation of resources while getting a desired result (e.g. for a large brewery that wants to provide stable and consistent taste). The application is characterised as a regression problem in which a large number of factors influence one or more target variables. The goal is to use a common data set for analysis, from which shared prediction models are trained. Within the case study, 12 batches of beer have been brewed on both sites, resulting in $n = 208$ tasting examples from sensory analysis. First, the number of >120 possible influencing factors was narrowed down to 58 recipe parameters, that can be influenced in advance, including amounts of ingredients, process temperatures and processing times. Second, different prediction models have been trained, including neural networks, generalised linear models and decision trees. The choice of models was based on an analysis of the most commonly used approaches to regression problems, based on the level of scales of the final data set, addressing mostly ratio and nominal scales. For the evaluation of the model performance, different performance metrics can be chosen. In this case, the *root mean squared error (RMSE)* was chosen for the regression problem. Furthermore, the training and scoring times can be compared (see Table 1). The training and testing of the models is based on a 7-fold cross-validation, whose mean RMSE value was calculated. In this case, no clear best model is obvious. The small fluctuations may also be due to the fact that the target value of the mean assessment of the beer parameters compensates for variation. Other target variables, e.g. specifically the foam stability, but also the availability of new data sets, can lead to other models performing better over time. However, a GLM was chosen as a base for the application, shown in Figure 5. Based on the GLM regression algorithm of H2O 3.30.0.1, after 30 Iterations and a heuristic based lambda search on the training data it consists of 17 active predictors. As a result of the study, the amount of "Hop 3" in the dry hopping phase (*Ekuanot*) has the largest influence on the beer rating, followed by the alpha acid content of "Hop 3" and the duration of the second rest time in mashing, followed by the quantity of the second hop (*Mosaic*). The model can be used by both sites to plan new improved recipes and is continuously enriched by new data sets from both breweries.

Table 1. Comparison of different trained models within the case study

Model	RMSE	Standard deviation	Training time [ms]	Scoring time [ms]
Generalised Linear Model	0.58	± 0.11	560	50
Deep Learning	0.58	± 0.09	1810	0,0
Decision Tree	0.55	± 0.06	210	150
Random Forest	0.57	± 0.21	230	100
Gradient Boosted Trees	0.56	± 0.02	1920	75
Support Vector Machine	0.55	± 0.19	130	225

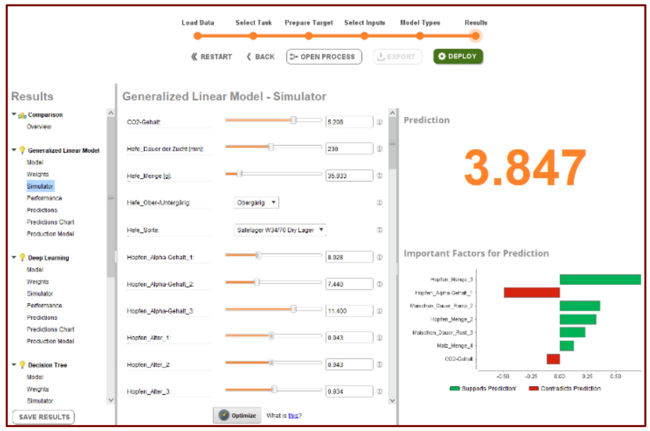

Figure 5. Recipe simulation within the model simulator environment of RapidMiner

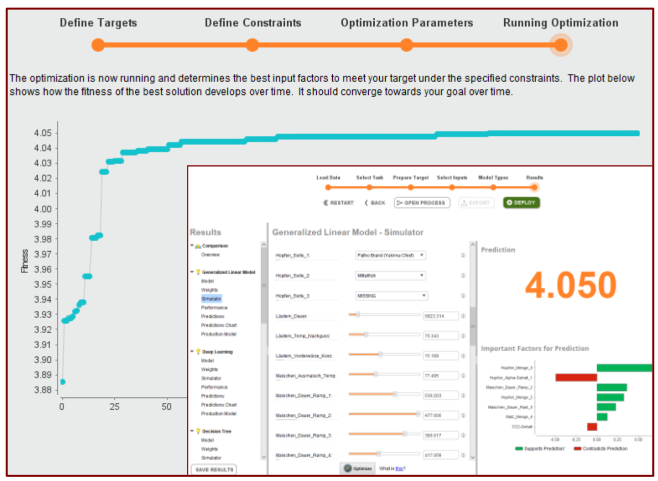

Figure 6. Data-driven generation of recipe optimisation recommendations

The resulting models can be used for taste predictions, to detect characteristic values or to generate optimised recipe parameters. The effect of recipe changes on target variables can be simulated via sliders (Figure 5). However, single parameter variations should also be treated with caution, since multivariate influencing factors are the reason for applying ML. The model application therefore also contributes to the possibility of optimisation under definable limits of parameter variation. Figure 6 shows how a data-based recipe recommendation is generated that increases the average expected prediction from 3.847 to 4.050 out of a maximum of 5.0. The changes are discussed by interdisciplinary teams, including domain experts, and then transferred to a new recipe in PCS *Zipmatic*. In this way, a cross-site control loop for continuous recipe improvement is created on the basis of unified DTs.

5. Conclusion and outlook

This paper presents an approach to create shared prediction models based on unified DTs of products, processes and resources, and shows how to use them for data-based recipe optimisation in the food and beverage industry. A central generic data model was created, which is the basis for unified DTs and thus the integration of physical and digital entities, as well as the basis for cross-domain data analysis. Further research is needed, especially on the transferability of models and services in different application contexts and IT architectures, which supports economic and ecological applications of ML. In the ML context, the collection of larger amounts of data is necessary, e.g. by cross-company data use. Open research questions include the quantification of the benefits of such an approach for each individual partner, e.g. better forecasts and services. In addition, there is a need for a detailed examination of closed control loops, on the one hand of a technical nature (e.g., automated interventions of the application in physical and digital entities), but also the role of humans in the loop.

Funding

Appendix A. Digital twin data models

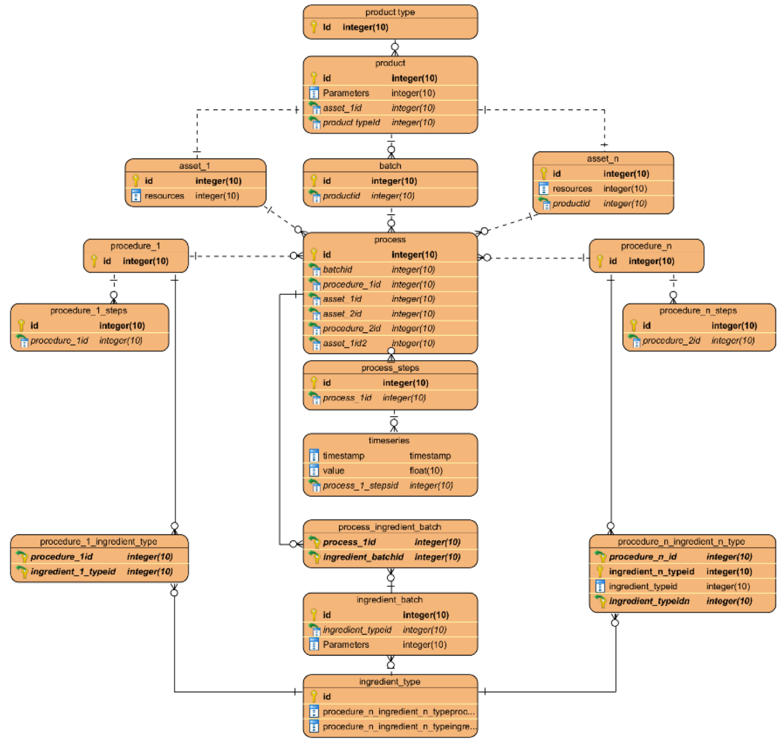

Figure 7. Generical digital twin data model for data-driven recipe optimisation

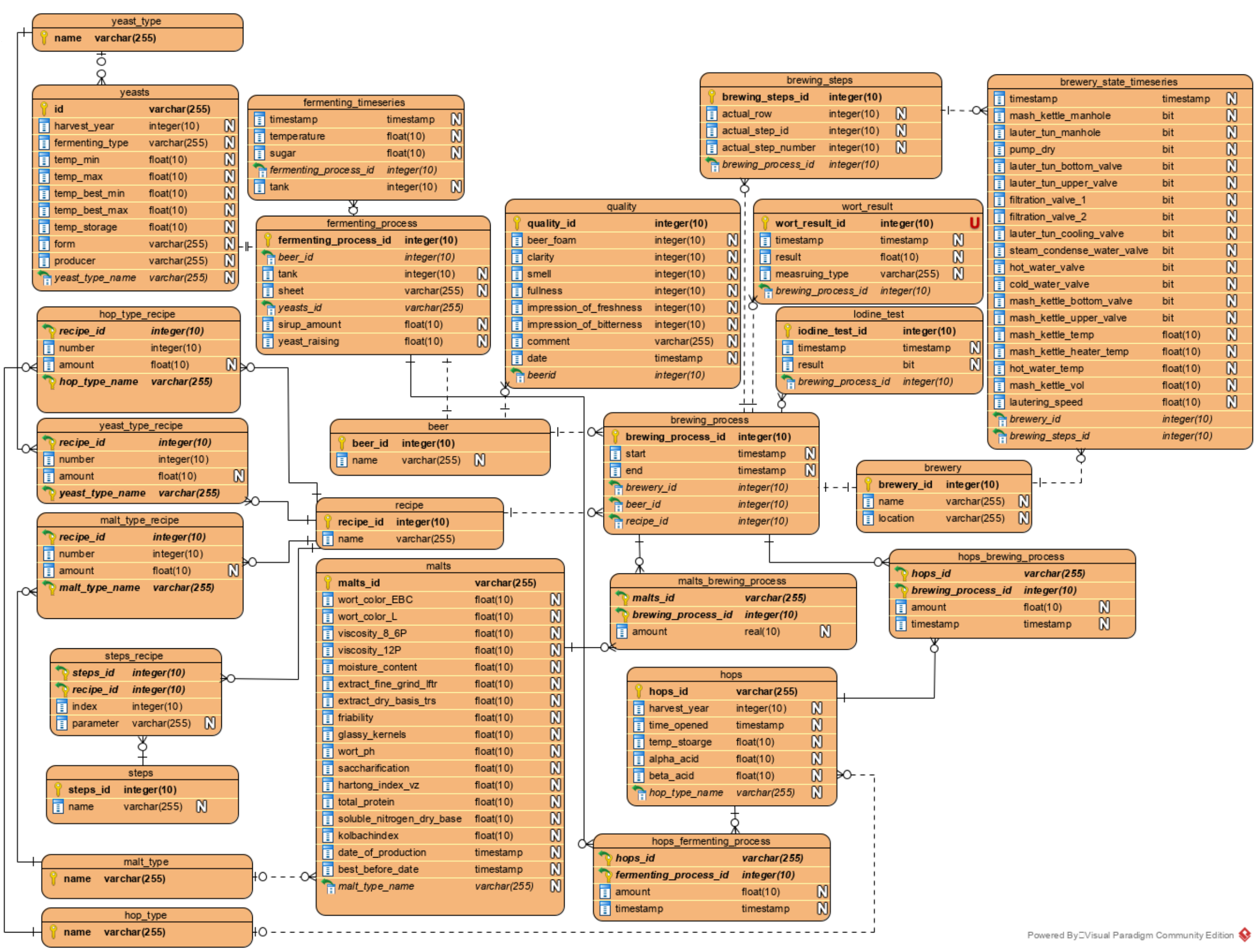

Figure 8. Digital twin data model within the case study as schema for the implemented databases

References

Ahrend, U., Aleksy, M., Berning, M., Gebhardt, J., Mendoza, F., & Schulz, D. (2019). Challenges of the digital transformation: The role of sensors, sensor networks, IoT-devices, and 5G : Invited Paper. In *2019 First International Conference on Societal Automation (SA)* (pp. 1–12). IEEE. https://doi.org/10.1109/SA47457.2019.8938077

Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications. *IEEE Communications Surveys & Tutorials*, *17*(4), 2347–2376. https://doi.org/10.1109/COMST.2015.2444095

Alpaydin, E. (2020). *Introduction to Machine Learning* (Fourth edition). *Adaptive computation and machine learning*. The MIT Press.

Arndt, S. (2018). Protokolle und Standards. In N. Gronau (Ed.), *Industrial Internet of Things* (pp. 39–80). GITO.

Ban, Y., Kearney, K., Sundahl, B., Medina, L., Bonnecaze, R. T., & Chopra, M. J. (2021, February 22–27). Fast etch recipe creation with automated model-based process optimization. In J. Bannister & N. Mohanty (Eds.), *Advanced Etch Technology and Process Integration for Nanopatterning X* (p. 22). SPIE. https://doi.org/10.1117/12.2583868

Barthelmey, A. (2021). *Digitaler Zwilling für durchgängige Datenbereitstellung und -nutzung im Lebenszyklus wandlungsfähiger Produktionssysteme*. *Schriftenreihe Industrial Engineering: Vol. 35*. Shaker.

Barthelmey, A., Lee, E., Hana, R., & Deuse, J. (2019). Dynamic digital twin for predictive maintenance in flexible production systems. In *2019 IEEE 45th Annual Conference of the IEEE Industrial Electronics Society (IES).* Symposium conducted at the meeting of Institute of Electrical and Electronics Engineers (IEEE), Lissabon.

Brewer World. (2020). *Brauer AI Serves Machine Learning And AI On Tap, Shows Us How To Brew The Perfect IPA*. https://www.brewer-world.com/brauer-ai-serves-machine-learning-and-ai-on-tap-shows-us-how-to-brew-the-perfect-ipa/

Bröckner, S., Benfer, R., Borzt, M., Engell, S., Knösche, C., & Kröner, A. (2021). *Process Simulation - Fit for the future? Position paper of the ProcessNet working Committe Process SImulation, Process Synhesis and Knwoledge Processing, 2021.* https://dechema.de/en/Positionspapier_Processsimulation-path-123211,124930/__/PP_Process_Simulation_2021_engl.pdf

Derler, P., Lee, E. A., & Sangiovanni Vincentelli, A. (2012). Modeling Cyber-Physical Systems. *Proceedings of the IEEE*, *100*(1), 13–28. https://doi.org/10.1109/JPROC.2011.2160929

Fuller, A., Fan, Z., Day, C., & Barlow, C. (2019, October 29). *Digital Twin: Enabling Technologies, Challenges and Open Research*. http://arxiv.org/pdf/1911.01276v5 https://doi.org/10.1109/ACCESS.2020.2998358

Garrido-Merchán, E. C., & Albarca-Molina, A. (2018). Suggesting Cooking Recipes Through Simulation and Bayesian Optimization, *11314*(1), 277–284. https://doi.org/10.1007/978-3-030-03493-1_30

Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access*, *5*, 20590–20616. https://doi.org/10.1109/ACCESS.2017.2756872

Grum, M., Sultanow, E., Friedmann, D., Ullrich, A., & Gronau, N. (2020). *Tools des Maschinellen Lernens*. GITO Verlag. https://doi.org/10.30844/grum_2020

Haupt, R. L. (Ed.). (2020). *Wireless communications systems: An introduction* (First edition). Wiley-IEEE Press. https://doi.org/10.1002/9781119419204

Heidrich, M., & Luo, J. J. (2016). *Industrial Internet of Things: Referenzarchitektur für die Kommunikation*. https://www.iks.fraunhofer.de/content/dam/esk/dokumente/Whitepaper_IoT_dt_April16.pdf

IEC (63278-1 ED1). *Asset administration shell for industrial applications:* Part 1: Administration shell structure.

IEC (61512-1:1997). *Batch control:* Part 1: Models and terminology (IEC 61512-1:1997). Berlin. VDE Verlag.

IEC (61512-2:2001). *Batch Control:* Part 2: Data structures and guidelines for languages (IEC 61512-2:2001). Berlin. VDE Verlag.

IEC (61131-9:2013). *Programmable controllers:* Part 9: Single-drop digital communication interface for small sensors and actuators (SDCI). Berlin. VDE Verlag.

IEC (62832-1:2020). *Industrial-process measurement, control and automation - Digital factory framework:* Part 1: General principles. Berlin. VDE Verlag.

IEC (TR 62541-1:2020). *OPC Unified Architecture:* Part 1: Overview and concepts. Berlin. VDE Verlag.

ISO (16400-1:2020). *Automation systems and integration - Equipment behaviour catalogues*

for virtual production system: Part 1: Overview.
Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital Twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine*, *51*(11), 1016–1022. https://doi.org/10.1016/j.ifacol.2018.08.474
Longo, F., Padovano, A., Nicoletti, L., Mohaiad, Elbasheer, & Diaz, R. (2021). Digital twins for manufacturing and logistics systems: is simulation practice ready? In *Proceedings of the 33rd European Modeling & Simulation Symposium* (435-442.). CAL-TEK srl. https://doi.org/10.46354/i3m.2021.emss.062
Narziß, L., Back, W., Gastl, M., & Zarnkow, M. (2017). *Abriss der Bierbrauerei*. Wiley. https://doi.org/10.1002/9783527812820
Nasar, M., & Abu Kausar, M. (2019). Suitability Of Influxdb Database For IoT Applications. *International Journal of Innovative Technology and Exploring Engineering*, *8*(10), Article J92250881019, 1850–1857. https://doi.org/10.35940/ijitee.J9225.0881019
Perno, M., Hvam, L., & Haug, A. (2022). Implementation of digital twins in the process industry: A systematic literature review of enablers and barriers. *Computers in Industry*, *134*, 103558. https://doi.org/10.1016/j.compind.2021.103558
Pethig, F., Schriegel, S., Maier, A., Otto, J., & Jasperneite, J. (2017). *Industrie-4.0-Kommunikation mit OPC UA*. VDMA Verlag. https://www.researchgate.net/publication/316432275_Industrie-40-Kommunikation_mit_OPC_UA
prostep. (2019). *Production Lifecycle Information Management: Information management from engineering to the shop floor*. White Paper. https://www.prostep.org/fileadmin/downloads/WhitePaper_PLIM_EN.pdf
Renggli, C., Rimanic, L., Gürel, N. M., Karlaš, B., Wu, W., & Zhang, C. (2021). *A Data Quality-Driven View of MLOps*. http://arxiv.org/pdf/2102.07750v1
Robinson, S., & Dehbi, A. (2021). *How sweet it is: Using Cloud AI to whip up new treats with Mars Maltesers*. https://cloud.google.com/blog/products/ai-machine-learning/baking-with-maltesers-marmite-and-cloud-ai
Schmitt, J., Bönig, J., Borggräfe, T., Beitinger, G., & Deuse, J. (2020). Predictive model-based quality inspection using Machine Learning and Edge Cloud Computing. *Advanced Engineering Informatics*, *45*, 1–10.
Schmitz, S., & Hagemann, M. (2020). *Insights from the food and beverage industry: Industrie 4.0 at scale*. Aachen. https://i40mc.de/wp-content/uploads/sites/22/2020/03/2020-03-09_I40MC_Whitepaper_Food-Industry.pdf
Tao, F., Zhang, M. & Nee, A. Y. C.(2019). *Digital Twin Driven Smart Manufacturing*. Academic Press. https://www.sciencedirect.com/science/book/9780128176306
Tao, F., Zhang, H., Liu, A., & Nee, A. Y. C. (2019). Digital Twin in Industry: State-of-the-Art. *IEEE Transactions on Industrial Informatics*, *15*(4), 2405–2415. https://doi.org/10.1109/TII.2018.2873186
Tarallo, E., Akabane, G. K., Shimabukuro, C. I., Mello, J., & Amancio, D. (2019). Machine Learning in Predicting Demand for Fast-Moving Consumer Goods: An Exploratory Research. *IFAC-PapersOnLine*, *52*(13), 737–742. https://doi.org/10.1016/j.ifacol.2019.11.203
Werner, R., Takacs, R., Geier, D., Becker, T., Weißenberg, N., Haße, H., Sollacher, R., Thalhofer, M., Schumm, B., & Steinke, I. (2021). The Challenge of Implementing Digital Twins in Operating Value Chains. *Advances in Biochemical Engineering/biotechnology*, *177*, 127–166. https://doi.org/10.1007/10_2020_153
West, N., Gries, J., Brockmeier, C., Göbel, J. C., & Deuse, J. (2021). Towards integrated Data Analysis Quality: Criteria for the application of Industrial Data Science. In *IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI).* Symposium conducted at the meeting of Institute of Electrical and Electronics Engineers (IEEE), Virtually.
Wöstmann, R., Schlunder, P., Temme, F., Klinkenberg, R., Kimberger, J., Spichtinger, A., Goldhacker, M., & Deuse, J. (2020). Conception of a Reference Architecture for Machine Learning in the Process Industry. In IEEE Computer Society (Chair), *2020 IEEE International Conference on Big Data (Big Data).* Symposium conducted at the meeting of Institute of Electrical and Electronics Engineers (IEEE), Atlanta, GA. https://doi.ieeecomputersociety.org/10.1109/BigData50022.2020.9378290
Xia, K., Sacco, C., Kirkpatrick, M., Saidy, C., Nguyen, L., Kircaliali, A., & Harik, R. (2021). A digital twin to train deep reinforcement learning agent for smart manufacturing plants: Environment, interfaces and intelligence. *Journal of Manufacturing Systems*, *58*, 210–230. https://doi.org/10.1016/j.jmsy.2020.06.012