A framework for data-driven digital twins for smart manufacturing

Friederich, Jonas; Francis, Deena P.; Lazarova-Molnar, Sanja; Mohamed, Nader

Published in:
Computers in Industry

Link to article, DOI:
10.1016/j.compind.2021.103586

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
https://doi.org/10.1016/j.compind.2021.103586

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
A framework for data-driven digital twins of smart manufacturing systems

Jonas Friederich\textsuperscript{a}, Deena P. Francis\textsuperscript{b,}\textsuperscript{*}, Sanja Lazaro\textsuperscript{a}va-Molnar\textsuperscript{a}, Nader Mohamed\textsuperscript{c}

\textsuperscript{a}Mærsk Mc-Kinney Møller Institute, University of Southern Denmark, Campusvej 55, Odense 5230, Denmark
\textsuperscript{b}DTU Compute, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
\textsuperscript{c}Department of Computer Science, Information Systems and Engineering, California University of Pennsylvania, California, PA 15419, USA

ABSTRACT

Adoption of digital twins in smart factories, that model real statuses of manufacturing systems through simulation with real time actualization, are manifested in the form of increased productivity, as well as reduction in costs and energy consumption. The sharp increase in changing customer demands has resulted in factories transitioning rapidly and yielding shorter product life cycles. Traditional modeling and simulation approaches are not suited to handle such scenarios. As a possible solution, we propose a generic data-driven framework for automated generation of simulation models as basis for digital twins for smart factories. The novelty of our proposed framework is in the data-driven approach that exploits advancements in machine learning and process mining techniques, as well as continuous model improvement and validation. The goal of the framework is to minimize and fully define, or even eliminate, the need for expert knowledge in the extraction of the corresponding simulation models. We illustrate our framework through a case study.

1. Introduction

Over the years, several definitions of the term “digital twin” have been proposed (Grieves, 2015). This term was first introduced in a talk in 2003 at the University of Michigan. All of these definitions have a few commonalities, such as there being a “real world entity” that is being “digitally modeled” in order to achieve certain goals, and the “real-time” nature of the digital model, which is a simulation model that can be executed to evaluate various what-if scenarios. The real-time aspect refers to the update of the simulation model as a consequence of changes occurring in the real entities. This enables the digitally extracted simulation model to be a high-fidelity replica of the real entity that it is modeling in real-time (Kritzinger et al., 2018), (Tao et al., 2019). Major differences in the definitions are mainly due to the requirement of a digital twin to “interact” in some way with the physical entity. Some works describe this interaction as a “two-way data flow” (Söderberg et al., 2017), resulting in changes in both the digital and real entities. Tools that enable such connectivity such as the Internet of Things (IoT) can also be considered an important component of a digital twin (Tao et al., 2019). In the interest of improving clarity and reaching a consensus regarding what we mean by a digital twin in this work, we define the following constituent parts for a digital twin:

- Real-world entity, such as a machine, a group of machines, or an entire factory;
- Data-driven simulation model, consisting of: (a) algorithms that describe the simulation model, machine learning and data mining for model extraction from data, and their software implementation, and (b) connectivity components, i.e., IoT systems; and
- Data that has been or is being generated by the real-world entity. This third component is a crucial part of a digital twin, which impacts how well the goals of digital modeling are achieved, as we discuss in Section 2.2.

The concept of digital twins can be seen as a natural extension of traditional simulation modeling coupled with increased data availability, connectivity, and evolving needs of end users. The ability to understand, monitor and experiment with complex physical systems is common to simulation modeling (Law and Kelton, 1982) and digital twins. An additional capability of digital twins is that they utilize the knowledge extracted from simulations to provide feedback to the physical systems. This feedback can then be used to

\*Corresponding author.
E-mail address: dfra@dtu.dk (D.P. Francis).

https://doi.org/10.1016/j.compind.2021.103586
0166-3615/© 2021 The Author(s). Published by Elsevier B.V.
CC_BY_4.0
control or manipulate some aspects of it to optimize some end-user given parameters. There are several uses for digital twins in areas ranging from the automotive industry to healthcare.

Some of the applications of digital twins in smart manufacturing are shown in Table 1. A comprehensive reference model for design and production was proposed by Schleich et al. in (Schleich et al., 2017).

There has also been notable work by Tao et al. on digital twin-driven product design (Tao et al., 2018). Qamsane et al. proposed a platform that uses software-defined control system for real-time monitoring of manufacturing systems (Qamsane et al., 2019). There has also been work done on process evaluation within a product development scenario (Liu et al., 2019), adaptive modeling of parameter sensitivity studies related to fault diagnosis (Wang et al., 2019) and framework for an autonomous production system (Ding et al., 2019). In this paper, we consider the case of data-driven digital twin modeling for smart manufacturing. The area of data-driven digital twin is related to the concept of automated simulation modeling. Some of the notable existing works that deal with automated simulation modeling are provided in Table 2.

The previous works on automatic or semi-automatic simulation modeling range from knowledge/data driven to event modeling approaches. In addition, these works utilize different kinds of data and manufacturing scenarios. The focus of our work is on Product Life Cycle (PLC) in the context of reconfigurable manufacturing systems. This leads us to the following relevant challenges:

1) **Entire PLC management**: all aspects of a product lifecycle, commencing from its design/idea conception to its sale can be handled by specialized digital twins (Kritzinger et al., 2018), (Lattner et al., 2010). PLC is of interest because it encompasses all aspects of a manufacturing system, which is especially important in the generation of simulation models for an entire smart factory.

2) **Reconfigurable manufacturing**: the need for this paradigm of manufacturing arises from the changing demands for the product (Koren et al., 2018), which requires that the configuration and structure of the factory change rapidly. Digital twins are especially suited to this scenario (Radziwon et al., 2014; Gola and Święc, 2016; Zhang et al., 2019). While reconfigurable manufacturing can provide many benefits, such as high degree of agility to satisfy various market demands, it includes many technical and non-technical challenges. These challenges are related to production efficiency, reliability, and cost-effectiveness. Using robust digital twins that reflect the real manufacturing environment can significantly improve production efficiency, reliability, and cost-effectiveness.

3) **Remanufacturing**: this aspect of manufacturing is associated with the conversion of old and/or used products to new ones. Digital twins have been deployed in recovery scenarios for waste electrical and electronic equipment (WEEE) (Wang and Wang, 2019), (Yang et al., 2018).

In this work, we develop a framework for data-driven digital twins for smart manufacturing. This framework provides many advantages to deal with the current challenges of smart manufacturing and production.

### Table 1: Digital twins in smart manufacturing.

<table>
<thead>
<tr>
<th>Work</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Schleich et al., 2017)</td>
<td>Modeling design and manufacturing products</td>
</tr>
<tr>
<td>(Tao et al., 2018)</td>
<td>Product design</td>
</tr>
<tr>
<td>(Qamsane et al., 2019)</td>
<td>Real-time monitoring and evaluation of smart manufacturing systems</td>
</tr>
<tr>
<td>(Wang et al., 2019)</td>
<td>Fault diagnosis of rotating machinery</td>
</tr>
<tr>
<td>(Ding et al., 2019)</td>
<td>Autonomous production systems</td>
</tr>
<tr>
<td>(Liu et al., 2019)</td>
<td>Process planning</td>
</tr>
</tbody>
</table>

### Table 2: Comparison of previous works in automated simulation modeling.

<table>
<thead>
<tr>
<th>Work</th>
<th>Purpose</th>
<th>Fully/semi-automated simulation modeling?</th>
<th>Knowledge based automated structural changes to model</th>
<th>Develops variants of a given model</th>
<th>Good model variants can be identified</th>
<th>Pros</th>
<th>Cons</th>
<th>Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Lattner et al., 2010)</td>
<td>Simulation of a given product</td>
<td>Yes, but not for reconfigurability</td>
<td>Knowledge based automated structural changes to model</td>
<td>No</td>
<td>A model template that includes simulation models and procedures, and a database</td>
<td>No, but does not develop automatic simulation models</td>
<td>No</td>
<td>No, not for reconfigurability</td>
</tr>
<tr>
<td>(Charpentier and Véjar, 2014)</td>
<td>Progressive construction of model based on contextual and object-based knowledge (where an object is and when)</td>
<td>Yes</td>
<td>Knowledge based automated structural changes to model</td>
<td>No</td>
<td>A model template that includes simulation models and procedures, and a database</td>
<td>No, but does not develop automatic simulation models</td>
<td>No</td>
<td>No, not for reconfigurability</td>
</tr>
<tr>
<td>(Uhlemann et al., 2017)</td>
<td>Open architecture for reconfiguration of manufacturing systems</td>
<td>Yes</td>
<td>Knowledge based automated structural changes to model</td>
<td>No</td>
<td>A model template that includes simulation models and procedures, and a database</td>
<td>No, but does not develop automatic simulation models</td>
<td>No</td>
<td>No, not for reconfigurability</td>
</tr>
</tbody>
</table>
manufacturing. This work is an extended version of our previous work (Francis et al., 2021) on data-driven digital twins. In this paper we discuss the main components of this framework, its benefits, challenges, and potential opportunities. We, further illustrate our approach through a case study.

The paper is organized as follows. In Section 2, we start with an overview of the problems of traditional simulation modeling, an introduction to digital twins for smart factories and the data-driven simulation modeling paradigm for smart factories. In Section 3, we describe the proposed framework for data-driven digital twins for smart factories, detailing the key constituent elements. In Section 4, we provide a case study aimed at digital twins for reliability analysis. The challenges and opportunities of the proposed framework are discussed in Section 5. The summary and general outlook of future developments are described in Section 6.

2. Background

Digital twins have been named one of the pillars of Industry 4.0 (Masood and Sonntag, 2020). Essentially, digital twins are evolving simulation models of the historical and current behaviors of the physical objects or processes that assist in optimizing on relevant performance measures. Operations of digital twins are enabled by the new technologies, such as Internet of Things (IoT), Big Data Analytics and Smart Sensors.

The concept of digital twins assists with the typical challenges associated with today's manufacturing, listed as follows:

• **Frequently changing demands**: needs of the customers and markets are constantly changing, leading to frequently changing demands in requested quantities and types of products. To reflect the changing needs, production lines change too. Consequently, the traditional manufacturing paradigm has moved from rigid to customizable, flexible and reconfigurable systems, which has led to tailor-made simulation models becoming obsolete soon after they are developed (Yelles-Chaouche et al., 2020).

• **Need for real-time monitoring**: manufacturing systems, as noted above, experience frequent changes to meet the demands, which increases the necessity for their continuous monitoring in real-time. The real-time monitoring allows for better and more informed decisions that could impact the fulfillment of the associated performance goals (Shahbazi and Byun, 2021).

• **Complexity of manufacturing systems**: due to the paradigm shift from regular to reconfigurable systems, the layouts and operations of the complete manufacturing systems turn out to be exceptionally complex. Automated and data-assisted way to control and optimize these systems to meet demands of customers, that digital twins can enable, help face this challenge (Malik and Brem, 2021), (Tao et al., 2019).

• **Need for more cost-effective productions**: Due to high competitions in the manufacturing sector, there are needs for more innovative ideas to reduce cost and make the production processes more cost-effective. Improving cost-effectiveness is impossible without accurate and timely simulation models of the different aspects of production processes and other related processes. Having accurate and timely models enable conducting more effective and useful descriptive, predictive, and prescriptive analytics to achieve cost-effectiveness in a competitive market (He and Bai, 2021).

Digital twins assist with coping with these challenges by providing a mean to continuously assess, control and optimize the performance of a given manufacturing system. A number of aspects have been, however, overlooked. We describe these overlooked aspects in detail in Section 3, after presenting the relevant background.

2.1. Challenges with traditional simulation models

Simulation modeling is a popular technique to evaluate the design, operation, and performance of complex systems. It provides methods and mechanisms to model behaviors of systems to the lowest level of detail and utilizing it to predict future behavior with the aim of enabling decision support. Discrete-event simulation models are typically employed for modeling complex systems, and they are manually designed by experts who select appropriate abstractions. This results in fully tailored simulation models.

This method of manual simulation modelling is not able to accurately reflect changes that occur in reconfigurable manufacturing systems, which undergo continuous modifications to their physical and software structure, in an attempt to match changing demands in the market (e.g., which is typical for mass customization (Qi et al., 2020)). Furthermore, systems often exhibit new and unexpected behaviors as they evolve and interact with their environments. Hence, manually created simulation models are typically rendered obsolete soon after their development. This creates need to develop new simulation models constantly and repeatedly, or manually update existing ones, which is a very difficult and expensive process. In contrast to this, data-driven simulation models, i.e., digital twins, would be able to capture these modifications on-the-fly (Lazaro-Molnar and Li, 2019).

2.2. Data-driven simulation modeling

Simulation models that are extracted and parametrized from data are referred to as data-driven simulation models (Lattner et al., 2010). Data-driven simulation modeling carries along a number of advantages, and some more relevant ones are the following:

• **Closeness to the real system**: data-driven models tend to more accurately reflect the actual systems that they model, when compared to traditional methods. Namely, real systems can exhibit behaviors that are not easy to be foreseen and modeled in advance, and such behaviors can only be captured through data (Wang et al., 2021).

• **Possibility to use Machine Learning (ML) and Artificial Intelligence (AI) for model enhancement**: data enables possibilities to use ML and AI methods to model phenomena that are problematic to reflect with standard simulation modelling. Furthermore, ML and AI methods can be combined with simulation and integrated to obtain deeper understanding of some systems’ behaviors (Cavalcante et al., 2019), (Alexopoulos et al., 2020).

A shift in focus from traditional simulation modeling to a more data-driven approach has been facilitated by the above-mentioned points, along with several successful adoptions of data-driven approaches. This shift has been made possible by the advancements in data generation and storage, algorithms and ever-changing needs of the manufacturing market. Previous research (as shown in Table 2) in data-driven simulation modeling suffers from the following drawbacks:

• **Lack of dominantly data-driven model generation**: Majority of the available model generation approaches cannot be classified as automatic simulation modelling, excluding those described in (Radziwon et al., 2014), (Goła and Świć, 2016). In addition, these noted works do not deal with the whole PLC of a typical reconfigurable smart factory (Lattner et al., 2010; Charpentier and Véjar, 2014; Rodić and Kanduc, 2015; Uhlemann et al., 2017; Goodall et al., 2019). The PLC is of interest because it gives importance to all aspects of a manufacturing system with regard to a product’s life.

• **Industry 4.0**: The real-time synchronization between models and real-world entities has become possible via the enabling
technologies of Industry 4.0. Previous works in data-driven modeling are not specifically designed to accommodate the Industry 4.0 settings in which modern factories are designed (Aheleroff et al., 2021).

- **Data heterogeneity**: Large number of data sources are associated with smart factories. Moreover, data that is delivered via the Internet of Things (IoT) devices comes in many different formats, e.g., logs, images, videos, etc. Utilizing this wide variety of data requires deployment of integration methods (Zheng et al., 2019).

Most of the existing works focus on homogenous data sources.

- **Data validation**: There are a number of challenges associated with the data collected from smart factories, as it often contains noise and missing values (very typical for sensor data). Both the volume and frequency of this data are often corrupted by noise, introduced by the factory environment and human errors. For this reason, having mechanisms to perform automated validation checks on the streaming data is of foremost importance because the quality of the extracted models highly depends on the quality and validity of the data (Bokrantz et al., 2018). Most of the previous works do not focus on data validation, which is crucial for ensuring correctness of models.

- **Reconfigurable systems**: To deal with the demands of the market to support mass customization, the adoption of reconfigurable manufacturing systems has taken place (Bortolini et al., 2018). Ideas for simulation modeling described in previous methods are not suitable for frequently changing and challenging scenarios such as reconfigurable manufacturing systems.

### 2.3. Digital twins for smart factories

Some of the key problems in production facilities with dynamic and rapidly changing boundary conditions can be solved by flexible and adaptive processes as part of a smart factory. In complex manufacturing systems, the digital twins can be adopted within a smart factory environment to efficiently model and control various aspects of it. Some of the benefits of adopting digital twins are the following: ability to monitor real-time status of machines, improved reliability, possibility for predictive maintenance, customized services to stakeholders, etc. (Rasheed et al., 2020). These benefits are realized within a smart factory by means of the data that are continuously collected and transferred by means of IoT devices. This data is subsequently used for key operations such as scheduling, maintenance, logistics, and decision making. The attainment of intelligent and data-driven decision-making within the smart manufacturing environment to improve profits and sustainability is the ultimate goal of Industry 4.0.

### 3. Data-driven digital twins for smart factories

The idea of a data-driven digital twin is motivated by the ever-increasing amounts of data that is being generated in smart factories. The emphasis is on data because it is the key element that can be used to determine and monitor states of complex systems within the factory environment. Effectively exploiting this data will aid in the development of high-fidelity models for manufacturing systems. ML and process mining tools are key methods that can be used to discover knowledge from such data (Van Der Aalst, 2012), (van der Aalst, 2018). Their role in simulation, and hence, the digital twin framework, will be that of enhancing the model's capabilities using the knowledge that was extracted. This knowledge will be online updated to reflect any changes occurred in the real manufacturing environment. This will enable many features to continuously improve different aspects of manufacturing processes even with occurrence of some changes. The novelty of the proposed framework lies in the fact that we utilize the advancements in machine learning and process mining to overcome the drawbacks of previous approaches in automatic/semi-automatic simulation model development, and to build data-driven models that are continuously validated and updated.

#### 3.1. Key elements of data-driven digital twins

Key elements that enable a data-driven digital twin for a smart factory are as follows.

1) **Data collection**.

The smart factory environment is associated with large volumes of data obtained through many sources. The first task to be performed before initiating data collection is entity identification. This step involves locating and collecting data about the entities relevant for the goal of the digital twin within the real-world system that is being modeled. On completion of this task, the data that is continuously being generated by the identified entities within the factory’s manufacturing systems are stored within a database. The key sources of data are as follows:

- **Sensors**: manufacturing components have various kinds of sensors attached to them, and these record data that is valuable for performing health monitoring and fault detection functionalities.
- **End-user inputs**: the stakeholders of the manufacturing systems such as customers often provide data that constitute their preferences, constraints, and needs.
- **Internal network**: the interconnections between the devices and parts of the factory enable the collection of comprehensive data about the factory as a whole. This has been possible by the advancements in cloud and fog computing and IoT.

The various types of data that are being generated include the following:

- **Diagnostic data**: such data are useful for monitoring the health of the manufacturing systems. It is often collected via sensors attached to the machines.
- **Logs and traces**: the history of the working of components or machines as a whole are recorded in the form of logs, which can prove to be invaluable in understanding the processes existing within the factory environment.
- **Market/end-user data**: the feedback, constraints and needs of the end-users are often used to adjust the working of the machines so as to meet predefined goals.

Another important part of the data collection process is data integration. It refers to the process of converting multi-modal data in the form of images, logs, audio etc. into usable and unified format suitable to the task that we are interested in performing.

2) **Data validation**.

Since data is the main entity that determines the correctness of the model, it is crucial that the data first undergoes a thorough inspection. This could be in the form of validity checks and cleaning. Some of the problems associated with data include the following:

- **Noise**: the data could be corrupted by errors introduced by the measuring device, the environment and human errors.
- **Missing values**: due to various reasons like device faults, networking problems and other factors, data could end up with missing values.
- **Multiple modalities**: different kinds of data such as image, text and audio are captured by the sensors present within the smart factory framework.
- **Large size**: often huge amounts of data are generated within the factory environment. The potentially high dimensionality as well as the size could render model generation difficult and non-accurate. Many ML algorithms do not deal well with data of huge dimensionality.
Data validity is of paramount importance, and as such data preprocessing with this purpose is an essential task. Accuracy and quality of extracted models can be seriously affected if the underlying data contains errors that have not been caught (Breuck et al., 2021). Some of the preprocessing tasks include imputation and cleaning. Imputation refers to the task of filling in the missing values using statistical or ML approaches. Data cleaning refers to all other tasks that convert data to a format that can be used by the knowledge extraction step.

In order to tackle the issue of data validation, we propose using an anomaly detection scheme in which there is a notion of “normal” and flagging anything that deviates from this normal as anomalous. Several anomaly detection methods have been proposed in the past, with varying approaches for distinguishing between normal and anomalous data. Several classification-based approaches have been proposed that uses a classifier such as Support Vector Machines (SVM) to discover a decision boundary between the normal data and the anomalous ones. The issue with using a supervised classifier-based anomaly detection is that the imbalance of the number of anomalous data. Several classification-based approaches have been validated on standard benchmark datasets, their performance in real-world data tends to be rather application-dependent. It has been discussed in the literature that there is no clear winner among these different approaches. Thus, in our framework, we will use a combination of these approaches to ensure data validity using anomaly detection. Concretely, depending in the type of data that we obtain from the sensors within the smart factory, the most appropriate anomaly detection method would be chosen.

3) Knowledge extraction.

The task of extracting knowledge from validated data obtained from the previous step. This task entails the following processes.

- Event detection and labeling: There are certain relevant events that are inputs to the simulation model, and therefore identifying these simulation goal-relevant events that take place in a smart factory is an important part of data-driven simulation modeling. Events are key elements that drive the development of discrete-event simulation models, and the ability to automatically detect events that occur within the smart factory determines the effectiveness of simulation modeling. ML offers various methods such as classification (Bauer et al., 2019; Becker et al., 2020; Moraru et al., 2020) and clustering (Maniak et al., 2020), (Subramaniyan et al., 2020) based approaches for detecting events from the smart factory data. The detection of special events, such as faults, is considered to be an important part of reliability analysis of manufacturing systems. The works of (Lazarova-Molnar and Mohamed, 2019), (Farsi and Zio, 2019) have stated that measuring and ensuring reliability within a smart manufacturing environment to be a complex and non-trivial task. The responsiveness and accuracy of the smart factory model is impacted by the ability to discover the occurrence of faults. From a machine learning perspective, this task of detecting faults can be categorized as the task of anomaly detection, for which a number of algorithms exists in the literature (Chandola et al., 2009), (Francis and Raimond, 2018).

- Process discovery: Besides events, the processes that occur within the smart factory are another source of knowledge for understanding the state of the factory by aiding in understanding parameters such as speed, duration, reliability etc. Discovering such processes have been subjected to study in the past few years. One key data source for process discovery is logs, or more specifically event logs. To discover the processes from the event logs, a formal description of the process flow in the system is derived using some modeling language or a process specification language, such as Petri nets (Van Der Aalst, 2012). The formal specifications of the processes can be derived from the raw event logs using these modeling languages. In addition to process discovery, its associated conformance checking is performed using these formal specifications.

4) Model development.

Development of a unified data-driven model is the next key component of the proposed scheme. The events and processes that were discovered serve as inputs to this step. The simulation model is developed using this input, initially with some level of human intervention. This initial model will then be updated automatically to reflect the changes in the smart factory. To facilitate this update, links between the model and the smart factory data streams need to be clearly identified. Furthermore, algorithms for model extraction and model updates will need to be developed to support the automatic/semi-automatic simulation modeling processes.

5) Model validation.

A rigorous model validation is an integral part of the data-driven simulation modeling approach, where the goal is to construct high-fidelity models that assist in correct and timely decision-making. Since the data is being continuously collected in real-time, an opportunity to continuously validate the simulation models is clear. This availability of data presents an opportunity to develop new model validation approaches that approach the validation process in an incremental manner, allowing for parts of the simulation models to be validated as soon as there is “sufficient” data on them. The full reliance on data for model validation demands that data itself is validated rigorously, such as to guarantee that the models are derived from high-quality data. This challenge will need to be resolved to support the proposed framework for data-driven simulation modeling.

The purpose of model validation is to ensure that the model’s performance on unseen data is good enough, in addition to achieving good performance on already seen data. Several validation techniques for the model have been proposed in the literature. It has been proposed that out-of-sample bootstrap techniques tend to give estimates that are the least biased (Tantithamthavorn et al., 2017). In our proposed framework we will use an out-of-sample bootstrap technique that works as follows. A sample of data instances is drawn randomly with replacement form the original dataset. Model training is done on this sample, and testing is done on the original dataset. This process is repeated k times and the average value of the performance metrics obtained in each training-testing process is reported. The model validation procedure will be done at predefined intervals to ensure that model is up to date with the data that has been generated so far. In our framework, the issue of the models becoming obsolete when new data arrives as a consequence of system changes is handled by the loop involving data validation, knowledge extraction, model development and validation. More precisely, using the data obtained at time t, knowledge is extracted, with which a model is developed. This model undergoes the validation step, in which the correctness of the model in terms of the data it has seen and the target of the model is checked. If the model conforms to the target and the data, then the parameters of the model are stored. At another time, t + k, after which the system changes and generates new data, the same procedure as above is repeated to ensure the current model conforms to the current data.

3.2. Framework for data-driven digital twins

We propose a framework for data-driven digital twins for smart factories, as shown in Fig. 1. The smart factory, which is the real-world entity that is being modeled continuously produces data
through its IoT devices and sensors. This data is the starting point of the data-driven modeling approach. The data collection processes include relevant entity identification and data storage in databases. Entity identification includes explicitly listing the relevant entities such as production systems, control technologies as well as other components within the smart factory. One relevant approach for collection and interpretation of manufacturing data is the Haystack standard (Home Project Haystack, 2021). This standard is an open-source initiative that offers standard semantic data models with the target of simplifying the extract value from the data generated by different IoT devices including devices in manufacturing facilities. The next logical step is data validation by means of general data cleaning or preprocessing and integration. The data implicitly captures relevant events occurring in the factory. In order to make these events more explicit and to aid the model development stage, event labeling is performed semi-automatically, i.e., by human intervention.

Event detection and labeling processes can be supported by a supplementary process of unsupervised learning (e.g., clustering) of events. In this supplementary process, all detected events are characterized and clustered based on their similarities. Conclusively, an expert would observe the generated clusters and their characterizations and interfere, if necessary, as well as provide labels for the clusters. As a result of these processes of event detection and labeling, a number of relevant events are identified and labeled. Machine learning models can be used to continuously and automatically detect events using previously labeled data. As a result, event logs are produced that are further used for process discovery using process mining algorithms (Yang et al., 2018). Both processes and events that have been identified are then used to build a simulation model for a given smart factory. A fundamental stage of the model development stage is the continuous validation of the model, which takes place in an incremental manner, validating the model as it is being extracted.

Once an extracted model is validated, the relevant model parameters are stored for future use. This model can then be utilized for performing various what-if scenarios as part of a simulation study, which are then evaluated with respect to predefined performance indicators. The results of these experiments can support stakeholders in making informed decisions regarding the optimal working of a given smart factory.

4. Case study

In this Section, we present a case study of assembling a quadcopter drone part in order to motivate and point out the possibilities of data-driven simulation modeling for digital twins in manufacturing. The goal of the simulation case study is studying the reliability of an assembly line. We, further, provide background information on the assembly process, and describe a Petri net simulation model that would be extracted from this manufacturing facility in order to illustrate the process and discuss possible extensions and implications. Finally, we demonstrate the capabilities of our proposed framework by applying it to data generated using the described assembly line.

4.1. Smart manufacturing facility for the case study

The University of Southern Denmark recently deployed a laboratory that serves as an educational, research and innovation environment for students, scientists and businesses, termed as Industry 4.0 Lab (Jepsen et al., 2020). The facility provides resources and infrastructure that contribute to a range of technologies related to the Industry 4.0 initiative. The goal is to create a vibrant and creative work environment where representatives from robot and production companies work closely together with scientists and students to develop novel and effective production technologies (Jepsen et al., 2020).

4.2. Production line description

Currently, the laboratory facilitates a production line that assembles part of a quadcopter drone. Fig. 2 illustrates the sub-assembly and its three components. The assembled drone part consists of a rotor, a motor and a chassis. Four of these assembled parts are required to build a quadcopter drone.

The production line consists of 5 assets, i.e.: 1) A warehouse with automated order picking and admission, 2) An autonomous mobile
robot with a robot arm that can load, transport and unload items, 3) A high-speed assembly track with magnetically attached transport devices, 4) Two assembly cells equipped with collaborative robot arms capable of carrying out specific tasks, and 5) A human machine interface (HMI) for controlling and monitoring the production process (Jepsen et al., 2020). We provide an overview of the production line in Fig. 3.

In the following, we briefly describe the individual production steps: 1) A user triggers the production process by submitting an order from the HMI. 2) The warehouse prepares the parts for the order. 3) The mobile robot transports the parts to the assembly track. 4) The track transports the parts to the first assembly cell. 5) Cell one executes the first assembly step, i.e. connecting motors with chassis. 6) The track transports the parts to the second assembly cell. 7) Cell two executes a second assembly step, i.e., attaching the rotors to the motors. 8) The track transports the assembled product to the end of the track. 9) The mobile robot transports the product to the warehouse. 10) The warehouse puts the product into storage. 11) The user is informed about the end of the process.

4.3. Reliability simulation model

To demonstrate the model that we expect to extract from the production line using our proposed framework, we designed a Petri net (Fig. 4). The Petri net depicts the strictly sequential nature of the process. Since there is no redundancy or buffer in the current setup of the production line, the failure of one of the assets brings the entire production to a standstill.

We modeled possible failures during asset runtime as inhibitor arcs preventing specific transitions from firing. Once an asset fails, the assets’ Fail transition fires and creates a token in the specific “Failed” place, which activates the inhibitor arcs. Thus, no transition depending on the failed asset can fire. As soon as the asset is repaired, the token in the assets’ fail place gets destroyed and the inhibitor arcs are deactivated.

We investigated what are the common faults that can occur in this quadcopter production line that may cause system failure, which we list in the following:

- The mobile robot does not recognize the box with the raw material provided by the warehouse.
- The mobile robot drops the raw material next to the transport devices from the assembly track.
- Because of the high speed of the assembly track, parts occasionally fall out of the transport devices.
- A part is tilted or rotated on the base plates - the collaborative robot arms may not be able to pick up the part.
- In cell 2, a small spring is getting attached to the motor before the rotor is attached – in almost half of the cases the spring cannot be picked up; the rotor is still attached to the motor.
- In cell 3, the rotors are not attached correctly, the rotor is attached – in almost half of the cases the spring cannot be picked up; the rotor is still attached to the motor.
- A similar Petri net for a production line is proposed by Horváth and Molnár (2016) to demonstrate and validate the performance of a timed Petri net simulation software. The authors also model a maintenance loop, similar to the asset failure loops, in order to reduce production line downtime. During the maintenance of specific assets, the corresponding timed transitions are blocked using an inhibitor arc.

In the future, further production steps and assembly cells will be added to the production line. For example, these could be cells that enable robots and humans to work together, and cells that are based solely on human interaction. In addition, buffers and redundancy could be added to make the production more error prone. Regular asset maintenance, like in the work of (Horváth and Molnár, 2016), would also contribute to the overall stability of the system.

The framework that we propose in this paper, along with the proposed methodology, allows for extraction of models, such as our reliability Petri net model from Fig. 4 from the data that is gathered through sensors, with limited expert knowledge support.

4.4. Data-driven reliability and simulation modeling

The Petri net displayed in Fig. 4 is developed manually and serves as a benchmark for what our data-driven digital twin framework, along with the proposed methodology, should enable to extract from data. In the following, we elaborate on how such a reliability-focused Petri net could be extracted automatically using data and process mining techniques.

In Friederich et al., 2021, we studied the data requirements for several hardware reliability models to make them data-driven instead of expert-driven. We found that data from smart factories that support data-driven reliability modeling can be categorized as either state data, event data, or condition monitoring data. All three data types have a time series format and thus each record consists of a timestamp (ts) and type-specific data. Fig. 5 displays the identified data types using a schematic diagram and an example data set.

- **State data:** record of the operational states of the individual assets of a system. The table in the figure shows exemplary states for each asset.
- **Event data:** record of discrete events generated by the assets of a system. The table in the figure shows case identifier, asset and event. The case identifier groups all activities which need to be
performed to manufacture a product (Lugaresi and Matta, 2021). Events mark the beginning of each activity.

- **Condition monitoring data:** record of relevant health data of a system. The table in the figure shows exemplary values recorded by sensors.

Based on event data, the process model describing the production of the drone subassembly can be extracted using process discovery methods such as the Alpha Miner (van der Aalst et al., 2004), Heuristics Miner (Weijters et al., 2006) or Inductive Miner (Leemans et al., 2013). The state data can be used to extract and model the operational state changes of the assets. Both modeling aspects are indicated in the Petri net (Fig. 4). After extracting the reliability-focused Petri net, its quality should be evaluated using appropriate criteria. This can either be done based on the data logs the Petri net was extracted from (conformance checking) or based on the ground truth Petri net. For the former, the measures fitness and appropriateness have been introduced by Rozinat and van der Aalst (2008). If the ground truth process model is known (like in our case), we can validate and assess the quality of the mined models using the common measures precision $P$, recall $R$ and their harmonic mean $F_1$ (Pourmirza et al., 2017). Precision represents the ratio of the number of correctly assigned edges to the total number of assigned edges. Recall is the ratio of the number of correctly assigned edges to the number of all edges present in the original model. We can mathematically formulate these measures as follows:

$$P = \frac{TP}{TP + FP}$$

Fig. 4. Reliability-focused petri net of the production plant in its current state.
where TP is the set of edges that exist both in the extracted model and in the original model; FP is the set of edges that exist in the extracted model, but do not exist in the original model; and FN is the set of edges that do not exist in the extracted model, but do exist in the original model.

To utilize the extracted Petri net for reliability analysis, it is necessary to know the probability distributions of the failure and repair times for each production asset ("Fail" and "Repair" transitions in Fig. 4). The state data provides this information by calculating the time to each failure and the time for each repair based on the corresponding operational state changes. To estimate the theoretical distribution that the empirically collected data follows, methods such as Maximum Likelihood Estimation (MLE) (Myung, 2003) can be used. MLE estimates the parameters of a probability distribution by maximizing a likelihood function, such that the assumed theoretical distribution best describes the observed data. Once the right distribution and the corresponding distribution parameters are estimated, the reliability function can be plotted. The reliability function (often referred to as survival function) is the complementary cumulative distribution function of the estimated distribution and displays the probability that a production asset will survive past a certain time (Kapur and Lamberson, 1977).

Besides Petri nets, fault trees are commonly used to assess reliability of a system (Friederich and Lazaro-Molnar, 2021). In Lazaro-Molnar et al. (2020), a method for data-driven fault tree analysis based on times series data is described. In a manufacturing scenario, events causing a system to fail would first need to be identified to generate fault trees. If the event data is not granular enough to capture fault events, this could be done using fault detection and classification methods as described by Heo and Lee (2018) based on condition monitoring data of a production plant.

### 4.5. Practical Demonstration

For demonstration of the application of our proposed framework for data-driven digital twins, we use data from the production line described in Section 4.2., to which we applied process mining techniques to extract the corresponding Petri net model and the relevant reliability insights, referring to failure and repair times of the production assets. In the following, we describe the performed steps in more detail.

To generate data, we produced 1000 units on our drone part production line and recorded the relevant events and operational state changes of the production assets. Tables 3 and 4 display an excerpt of the event log and state log that were generated during this process.

To extract the Petri net depicted in Fig. 4 and to estimate the distributions for the failure and repair transitions we utilized the two Python libraries “pm4py” (Berti et al., 2021) and “reliability” (Reid, 2020). “pm4py” is an open-source process mining platform with the goal to make state-of-the-art process mining algorithms available to the public and to support collaboration between academia and industry (Berti et al., 2021). Besides implementations of several process discovery and conformance checking algorithms, pm4py also offers an extensive code base for modeling and simulating Petri nets. “reliability” is a library for reliability engineering and survival analysis (Reid, 2020). The complete extracted Petri net, using the aforementioned Python libraries, is displayed in Fig. 8.

Based on the generated event log, we first applied the Alpha Miner (van der Aalst et al., 2004) algorithm to extract the material flow model of the production line (left part of the Petri net in Fig. 4). Subsequently, based on the state log, we extracted the failure and repair loops for each asset (right part of the Petri net in Fig. 4) and linked them to the corresponding asset working transitions (e.g., “Transport to Track”, “Cell 1 Operation”). This was done by first creating two places (“failed” and “OK”) and transitions (“fail” and “repair”) for each asset and connecting them using arcs. Second, the created “failed” place was connected to all transitions that are utilizing the respective asset by an inhibitor arc. The extracted model is identical to the ground truth model. This is due to the relatively simple nature of the sample production line: the process is sequential, no reworks or parallel operations are included, and there is only one product type. Thus, an evaluation based on the presented metrics Precision, Recall and F1 would result in a perfect score of 1. The metrics are more useful for models capturing a more complex production line.

### Table 3

<table>
<thead>
<tr>
<th>Excerpt of the recorded event log.</th>
</tr>
</thead>
<tbody>
<tr>
<td>timestamp</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>02/11/2021 09:36</td>
</tr>
<tr>
<td>02/11/2021 09:36</td>
</tr>
<tr>
<td>02/11/2021 09:38</td>
</tr>
<tr>
<td>02/11/2021 10:11</td>
</tr>
<tr>
<td>02/11/2021 10:13</td>
</tr>
<tr>
<td>02/11/2021 10:23</td>
</tr>
<tr>
<td>02/11/2021 10:42</td>
</tr>
<tr>
<td>02/11/2021 11:19</td>
</tr>
</tbody>
</table>

### Table 4

<table>
<thead>
<tr>
<th>Excerpt of the recorded state log.</th>
</tr>
</thead>
<tbody>
<tr>
<td>timestamp</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>02/11/2021 09:36</td>
</tr>
<tr>
<td>02/11/2021 09:38</td>
</tr>
<tr>
<td>02/11/2021 09:38</td>
</tr>
<tr>
<td>02/11/2021 09:46</td>
</tr>
<tr>
<td>02/11/2021 10:13</td>
</tr>
<tr>
<td>02/11/2021 10:13</td>
</tr>
<tr>
<td>02/11/2021 10:23</td>
</tr>
<tr>
<td>02/11/2021 14:30</td>
</tr>
</tbody>
</table>
Subsequently, we estimated the failure and repair distributions for each asset based on the state log. To do so, the time to each failure and repair needed to be calculated. For the failure times, this was done by calculating the time difference between each repair (“repaired” state) and the subsequent failure (“failure” state) or, in case of the first failure, the time difference to the start time of the production line. For the repair times, the time difference between each failure and the subsequent repair was calculated. To estimate the failure and repair distributions, we used the MLE method. For each asset, we fitted several probability distribution functions which are commonly used in reliability engineering and assessed their statistical likelihood. For example, Fig. 6 displays the probability density function and cumulative distribution function for each fitted distribution for the assembly track failures. After fitting the distribution functions, we added the most likely ones and their parameters to the extracted Petri net. Lastly, we calculated the reliability functions for each asset (Fig. 7).
5. Challenges/discussions

The presented framework for data-driven digital twins in smart factories has a number of associated challenges, which are listed as follows:

- Identifying and fully defining the appropriate extent of human intervention to facilitate the data-driven approach: this challenge refers to the use of a predominantly data-driven approach that needs and depends on human intervention to enable the automation. Well-defined points of human input will ensure a proper setup of the data-driven simulation modelling, as well as its interpretability and efficiency. Interpretability refers to the aspect of a data-driven/intelligent approach that offers some principled way of understanding the decisions made by the digital twin from the point of view of a stakeholder. Efficiency refers to the aspect of the data-driven digital twin that includes accuracy and timeliness.
- Real-time decision-making: digital twins can assist stakeholders in making advantageous decisions in terms of different performance indicators. Such decision-support is made possible because data-driven digital twins extract relevant knowledge from all data that has been collected, which could prove to be invaluable to the stakeholder. In regard to timeliness and criticality of decisions, needs with respect to timeliness should be matched to the criticality and weight of the associated decisions. This should be reflected in the granularity of the data that is being collected. Decisions related to safety, especially in safety-critical systems would require higher timeliness than decisions related to production.
  - Evolving goals: smart manufacturing systems aim to warrant stable and cost-effective production frequencies. This aim is closely correlated with other characteristics, such as reliability of manufacturing systems, decreasing of material waste and energy consumption, etc. Apart from these, environmental goals, such as decreasing greenhouse gas emissions is another objective that is and will be an important to be considered. Building simulation models to reflect, update and serve such varying and evolving goals is highly complex and a huge challenge to be addressed.

The presented framework, however, also presents new opportunities, listed as follows:

- Development of high-fidelity digital twins: due to the data-driven aspect of the proposed framework, models tend to be accurate representations of the corresponding production systems, reflecting up-to-date behaviors of processes within smart factories. In addition, the ongoing validation of both data and models ensures that this accuracy is maintained consistently.
- Improved understanding of processes and decision-making: process discovery is an important part of our proposed approach. This is done by means of mining the event logs that have been collected, in addition to other data-driven components. This helps to enhance the understanding about process flows in the system and will be useful in improved decision making. By utilizing the process flows in the system, the stakeholders can make relevant decisions about the future of the system as well as the digital twin.
- Increased reliability: reliability of the proposed framework is attributed to the continuous validation of data and model. The emphasis on continuous rather than validation at fixed intervals ensures that the system is evolving appropriately with the speed of newly arriving data. The validation system ensures that the data which is a crucial part of our framework is correct and also that the model faithfully represents the data it is modeling.
- Agile manufacturing: the data-driven digital twin can offer a sound platform for adopting the concept of agile manufacturing which enables manufacturing enterprises to respond rapidly to market changes while still maintaining good quality and controlling costs.

In summary, the adoption of digital twins offers several opportunities within the smart factory, while also presenting various challenges. These opportunities can be looked upon as motivation for the adoption of data-driven digital twins in manufacturing systems.

6. Conclusion

Data-driven digital twins are necessity for successful Industry 4.0 adoption. Manual simulation modelling is not an option with the modern manufacturing systems that undergo numerous and rapid reconfigurations during their lifetimes. For this reason, we investigated the requirements that would enable a data-driven development of simulation models for smart manufacturing systems, as a basis for their digital twins. Real-time data-driven digital twins, automatically developed from data gathered from smart factories’ IoT devices can fulfill significant needs, e.g., always-up-to-date
simulation models, accurately reflecting ongoing changes in factory layouts. In addition, data-driven digital twins will enable integrated ongoing model validation, that can be started as soon as parts of the model are being extracted. The availability of data from the real system will make this possible. The opportunities bring along a number of challenges and needs that need to be met too. Some processes will depend on human intervention even after everything has been automated. However, their clear designation and distinction will be helpful in their mainstreaming. For instance, setting simulation goal and determination of relevant events will need to be performed by experts. The ideal ratio of human expert knowledge and data would depend on the system in place and the goals of the data-driven digital twin. In attempt to address the needs of the data-driven digital twin development, we developed a high-level conceptual framework too. We anticipate that this data-driven simulation modeling will gain more attention as more mechanisms for ensuring data quality become available. We also provide a high-level case study to illustrate our framework, and discuss the associated challenges and opportunities.

**Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**References**


Becer, F., Roth, C., van der Aalst, W., Dillmann, R., Acoustic Anomaly Detection in Additive Manufacturing with Long Short-Term Memory Networks in 2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA), Apr. 2020, pp. 921–926. doi: 10.1109/ICIEA49774.2020.9100202


