ENGEN582-24X Literature Review

Integrating streamed sensor data into a distributed model of a complex system

Bert Downs

Supervised by Tim Walmsley, Mark Apperley

April 2024

“Live Data integration in Industry 4.0”. Generated by Bing AI.

Contents

1 Introduction

1.1 Background

The World Economic Forum ranked "Failure to mitigate climate change" as the number one threat to the world in the next ten years [1], due to the effects climate change has on extreme weather events, biodiversity, and climate-vulnerable economies. Decarbonisation is a crucial step in mitigating climate change. In New Zealand, process heat accounts for 8% of total greenhouse emissions [2].

Digital Twin (DT) technologies can inform efficient scheduling, optimisation, maintenance, and retrofit procedures. Thus, DT technology has the potential to decrease process heat emissions, particularly in site-wide and site-edge applications [3].

1.2 Purpose

This literature review is being conducted in conjunction with a software development project. The project is to develop a system that ingests real-time sensor data from chemical processing plants and uses a mathematical model of the plant to estimate other characteristics in real time. The project proposal may be found in appendix **??**. However, the project proposal is very high-level, explaining *what* the project is but not *why* the project is necessary or *how* the project can be implemented.

This literature review supports the software development project by answering the *why* and *how* questions about the project. It reviews the current state of research integrating real-time industrial sensor data into digital twins. It confirms and updates the overall aim and specifications of the software development project, by identifying knowledge gaps requiring further research, and methodologies or tools that would support development.

1.3 Scope

This literature review will broadly cover topics related to Big Data including Data Collection in Factories, the Internet of Things, Data Stream Processing, and Machine Learning for Data Streams. It will focus more specifically on Digital Twins (DTs) and Chemical and Process Engineering (CAPE).

1.4 Research Questions

The sources used in this literature review were selected to focus on the following research questions:

- *RQ1*: Where is real-time data collection most used in industrial processing facilities, and where is there potential for it to be used better?
- *RQ2:* What techniques and processes are required to turn real-time data into useful insights that can inform and automate decisions in an industrial facility?
- *RQ3:* How can real-time data be integrated into a Digital Twin that is based on mathematical, first-principles simulation techniques?

RQ1 focuses on the general goals and purpose of the project. By reviewing case studies and other literature that apply real-time data collection in an industrial processing facility, an overview of the most applicable areas can be achieved, enabling the project to focus on areas where the greatest improvement over existing techniques is expected.

RQ2 focuses on the pipeline to process live data in an industrial factory. This will be used to define the technical environment the software development project will pull from and contribute to. From this, the inputs and the outputs of the software development project can be clearly defined, enabling the project to be clearly situated as a part of the existing data processing pipeline.

RQ3 discusses the major technical challenge this project will need to overcome. Real-time data is by nature dynamic and imperfect. Mathematical and first principles simulations are well-defined and cannot always handle variation, missing data, and inaccuracies. By discussing methods that others have used to process real data using mathematical simulation techniques, areas for future work can be identified. This can be used to inform the development of the core algorithms of the software development project.

1.5 Methodology

Digital Twins are by nature highly integrated and context-dependent. Thus, the research and technology underpinning them can differ greatly depending on the application. Because this project is focused on industrial live data, most sourced literature is related to Digital Twins in industrial settings. This literature will identify the specific technologies and techniques used to build digital twins in industrial settings. The technologies and techniques are then reviewed in greater depth outside of the context of digital twins, to gain a general understanding of the limitations and strengths of each. This can then be used to draw conclusions about its use in Industrial Digital Twins.

Because this literature review will primarily be used to inform future development, some sections include a review of relevant products and tools that apply the techniques identified. These products have been identified based on their applicability to industrial processes and settings. As a convention, sources from product websites and documentation are referenced as footnotes, to differentiate from academic citations included in the bibliography.

The three research questions identified each focus on very different areas of industrial data processing and analysis, spanning the spectrum from big-picture needs down to challenges in developing specific algorithms. Each could be the subject of a literature review on its own. Additionally, Big Data, Artificial Intelligence, and Digital Twins have recently become very popular topics in academic literature. The diversity of sources and availability of research means there are a large number of papers that are related to this literature review. As such, it is a challenge to filter out all research on a topic to find those that are the best fit for this project - there are simply so many articles and technologies that are relevant and could be used. Care has been taken to ensure that the sources used cover the spectrum of available technologies and techniques, on each level of analysis. However, this cannot be considered a comprehensive review of all relevant research, algorithms, and implementations.

2 An Overview of Digital Twins

The concept of Digital Twins has been rising in popularity in recent years and is used in a number of industries, including Manufacturing, Energy, Transportation, Healthcare, and Networking [4]. The concept of a "Digital Twin" refers to a simulation of something in the physical world, which is kept up to date with a physical system using real-time data [5].

2.1 Purpose of Digital Twins

There are number of reasons Digital Twins may be used. These generally fall into one of the following categories:

- **Testing**. The digital twin can then be used as a test environment, as modifications can be applied to the digital twin before applying them to the physical system [6].
- **Validation**. Data can be collected to validate that a factory is working correctly. Data can also be collected to validate that a digital model accurately reflects the conditions of the factory [7].

- **Diagnostics**. Live data feeds from each stage of operations can help diagnose issues, by tracing the problem back to where it first occurred. They can also be used to measure the health of the physical system as the digital twin changes over time[7].

- **Fault Classification**. This goes a step further than diagnostics, as fault classification involves interpreting the data to find an exact cause and type of fault [8].

- **Automation**. A software system can be set up to trigger actions in response to changes in process conditions. For example, the system can initiate corrective measures autonomously[3].

- **Schedule Optimisation**. Data from operations can be used to optimise plant schedules, such as avoiding peak power usage when the electricity price is high [9].

- **Control Optimisation**. Live data can be used to identify what conditions provide the highest cost-yield ratio [10] [11].

- **Asset Optimisation**. Plant data can be useful when evaluating the return on investment of capital expenditure during expansion, retrofit, or replacement of equipment [12].

2.2 Conceptual Model of Digital Twins

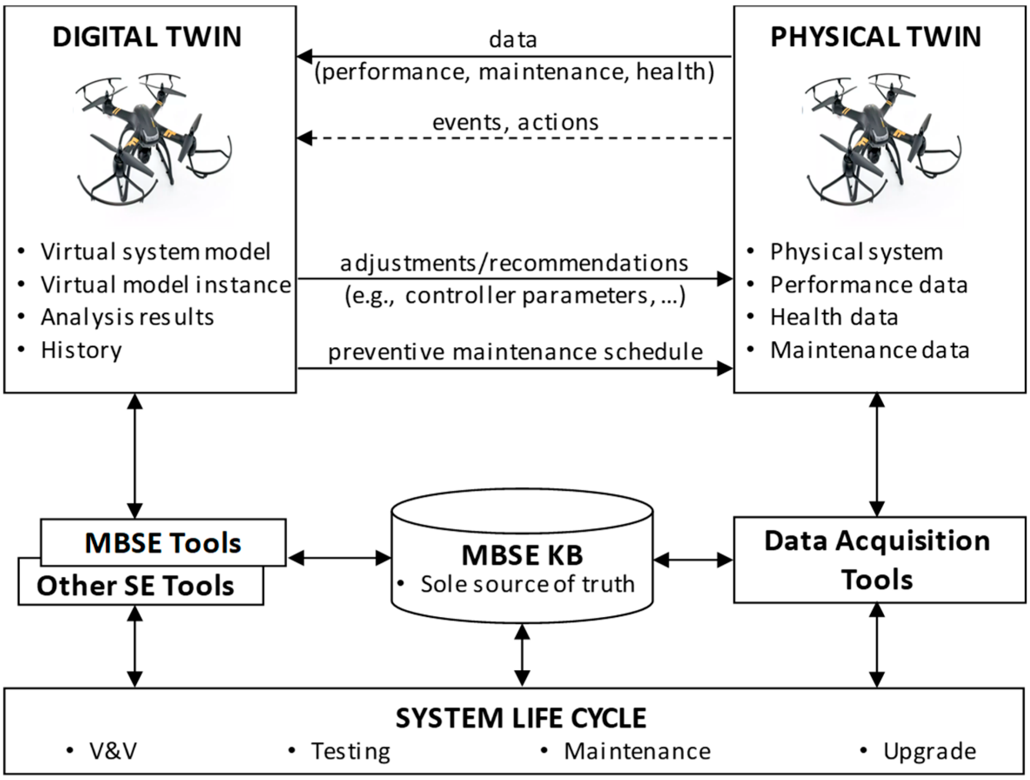

Figure 1: Digital twin concept, using the Model-Based Systems Engineering Framework (Madni et al. [7])

Madni et al. analysed the concept of digital twins from a Model-Based Systems Engineering (MBSE) perspective [7]. Figure 1 shows how the physical twin is able to send information to the digital twin, such as performance data and events. The digital twin is then able to be used to calculate recommended actions and inform decisions using simulated data beyond that which is physically possible to collect. It is able to do this by using the MBSE Knowledge Base, a database that records simulation results based on real-time data. As the digital twin is a model of a specific instance of a physical asset, rather than a generic model, it is more accurate than traditional approaches.

Yu et al. [5] classified DTs on varying levels of detail. They identified that there was a lack of research into higher-level digital twins, which contained accurate process simulations and operated at a larger scale over an entire factory or site. They also identified that much less research has been done on process and energy Digital Twins in comparison to their mechanical counterparts.

2.3 Conceptual Model of Live Data Processing

There are many steps to processing live data. One framework others have used to view the processing of live data is the Knowledge Pyramid [13]. Also referred to as the Data-Information-Knowledge-Wisdom hierarchy or Data Pyramid, this consists of four layers of abstraction:

- **Data**: Discrete, objective facts about something. Simple values or measurements.
- **Information**: Data in context - structured to provide a useful story or meaning, such as linking to a specific location, or time.
- **Knowledge**: Something that provides the why or how - explaining the relationships between information.
- **Wisdom**: Placing knowledge in a broader context, and applying it to different or novel situations.

This hierarchy is used to illustrate the progression from raw data to meaningful insights and actions [14]. It is analogous to the progression of live data in a digital twin, which starts out as raw sensor data and eventually is applied to improve the physical system. Figure 2 illustrates this flow from data to meaning as follows:

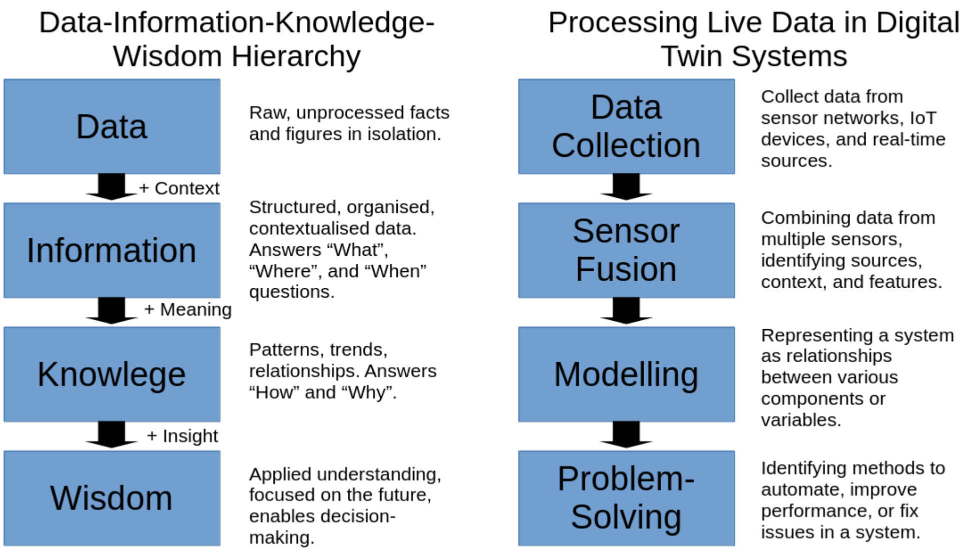

Figure 2: Comparing and applying the Data-Information-Knowledge-Wisdom Hierarchy with the flow of data through Digital Twin Systems.

- **Data Collection**: Raw Data must first be collected from the industrial process. This forms the basis of all future insights but is meaningless in isolation. This data is collected in real time.
- **Sensor Fusion**: Context is given to the data, such as type of measurement, time, and/or location. More advanced sensor fusion techniques can be used to integrate data from multiple sensors to overcome limitations such as noise, uncertainty, or individual sensor limitations. This can be viewed as converting the "Data" into "Information".

- **Modelling**: Modelling is the core of the Digital Twin. Modelling creates a "representation or abstraction of the system" [15], linking distinct pieces of information together by creating or finding relationships between them. This means modelling can be viewed as helping to convert "Information" into "Knowledge". Models enable the simulation of hypothetical scenarios.

- **Problem-Solving**: Wisdom is the least well-defined portion of the data pyramid. However, it is generally seen as future-focused, applying current knowledge to make decisions. In Digital Twins, this can be seen as the process of applying live data models in order to find insights that aid in or make decisions. This could include automating actions, optimising performance, or identifying faults in advance.

The remainder of this literature review analyses literature related to each level of data processing in this pipeline in turn. This provides an end-to-end understanding of how data is processed in a digital twin. Section 3 will cover Live Data Collection, section 4 reviews literature related to sensor fusion, section 5 covers various methods of modelling with live data, and section 6 considers how digital twins can be used to solve various real-world problems.

3 Live Data Collection

The first part of processing live data in a Digital twin is to collect the raw data from the industrial processes. This section discusses the data sources available in industrial plants, and what technologies enable collecting data from them. This information is used to discuss and update the requirements the software development project will have.

3.1 Data Sources

In Industrial Process operations, there are a number of data sources, as shown in fig. 3 [16]. These include:

- **Process Data**, which is data from on-site sensors and control systems. This can include the state of operations, the current control configuration, temperature, humidity, vibration, position, or chemical composition. This data is collected in real-time, can be automatically analysed, and is the main focus of our study.

- **ERP Databases** contain time-stamped information about usage, breakdowns, plant configuration, capacities, and external factors such as the supply chain or product specifications. This data is also collected in real-time, and can usually be automatically analysed, and thus is also central to our study.

- **Engineering Data** refers to datasheets and models, usually created during plant design. This is not a live data feed, but can be used to inform the development of sensor fusion algorithms as discussed in section 4. Factory models can also be used for real-time simulation, and this will be discussed in section 5.

- **Manual Logs** refer to manually recorded information about the plant operations. In general, these cannot be automatically analysed.

- **Quality Logs** refer to factory samples that are taken and analysed offsite. These results are not a live data source, but real-time algorithms based on plant conditions and past quality logs could be developed to target plant conditions that provide the highest quality results.

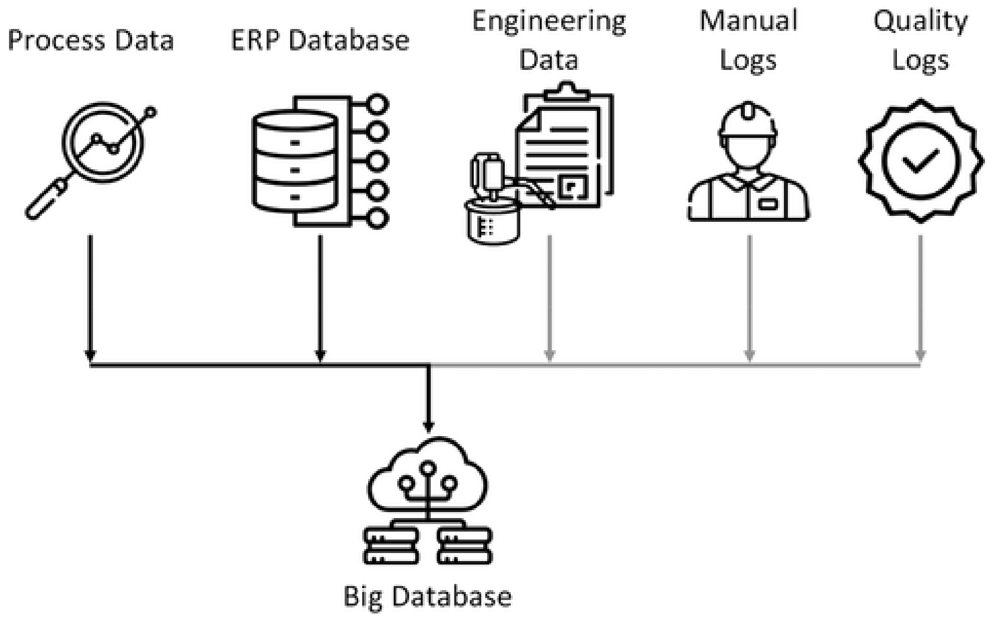

Figure 3: Typical data repositories in an industrial plant [16].

3.2 Technologies

There are a number of tools and technologies that are used to collect live data from factories. It is important to identify what industry tools exist and are already being used. Understanding these tools provides a framework future tools can build on and avoids recreating work that is already done.

IoT Communication tools provide standardised ways to interact with a variety of IoT devices. These include:

- MQTT [1], which is a publish/subscribe system that enables other services to communicate with IoT devices
- Modbus [2], an open standard to connect to Programmable Logic Controllers (PLCs) in factories.
- Telegraf [3], which can collect metrics from various protocols, transform, aggregate, and filter them, and send them to a centralised database.
- OPC [4] is a standard that abstracts PLC protocols such as Modbus into a standardised interface. This enables interoperability between devices from different vendors.

In an industrial context, Supervisory Control and Data Acquisition (SCADA) systems are used to gather and process data. It is industry standard to use SCADA Systems to help chemical plants work autonomously, optimally, and efficiently [17]. Examples of commercially available SCADA systems include (in no particular order) WinCC, Genisis64, iFix and Cimplicity. These systems perform a similar role to IoT-based solutions, and data from either can be used to inform a Digital Twin.

3.3 Discussion

From the above sources reviewed, it is clear that data is readily available from industrial sites. Therefore, in the context of the software development project, it is most appropriate to use existing sources

[1] `https://mqtt.org`

[2] `http://www.dankohn.info/projects/Fieldpoint_module/Open_ModbusTCP_Standard.pdf`. See also `https://www.modbustools.com/modbus.html`.

[3] `https://www.influxdata.com/time-series-platform/telegraf/`

[4] `https://opcfoundation.org/`

of data rather than deploying additional sensor networks. Sensor data is not the only source of live data - so the project should also process external information such as usage demands. Furthermore, there are many different platforms and protocols that data is collected from, so the ability to support multiple distinct data sources is a key requirement for the project. In particular, integration with SCADA systems will be required.

Because the project will need to aggregate so many different sources of data, it will also need to be performant enough to process the data in real-time, and reliable enough to do so without interruptions. It will need to be scalable enough to support large and small factories.

4 Converting Data to Information via Sensor Fusion

Once access to real-time data is available, the next step to processing is to make the data meaningful to a digital twin system. This is an application of the field of sensor fusion. Data needs to be tagged with relevant context information, such as source and timestamp information [16], and data streams need to be aggregated together in a way that enables extracting meaning. This is the process of converting raw data into information.

4.1 Sensor Fusion Techniques

Sensor fusion is the process of combining multiple different sensors' data in order to come up with some more meaningful parameters[18]. For example, a Virtual Reality headset can fuse accelerometer, gyroscope, and magnetometer data together to find the device's rotation - even though there is no one sensor that can perform this task [19]. In essence, this creates a "digital twin" of the headset with the same orientation parameters in the virtual environment that the physical headset has in the real world.

When building a digital twin from sensor data, it is unlikely that the measurements we are able to get will always be exactly the same as what is modelled in process dynamics. For instance, it may be hard to get the temperature of a material stream, but it may be possible to get the temperature of the pipe the stream is in - and then sensor fusion techniques can be used to estimate the temperature of the material inside based on this and other data.

Alam et al. [20] explored how IoT and data fusion technologies are related. They identified three methods of data fusion:

- Probability-based methods, such as Bayesian Analysis, Markov chains, and the Monte Carlo Method. This fuses sensor data by mathematically modelling the relationship between the various sensor readings, and weighting the probability of inaccuracies in each in order to find the most likely solution.
- AI-based methods, ranging from fuzzy logic to Artificial Neural Networks. This works better than Probability-based methods when modelling complex interactions.
- Theory of Evidence-based methods, which are a generalisation of Bayesian Probability theory that is not as complete but works well in practice, and is faster than AI methods.

Bayesian Analysis was used by Renganathan et al. [21] to fuse field measurements with simulation data. Figure 4 shows the overall methodology. Bayesian Inference is used to find the value that has the highest likelihood of being correct. The main limitation identified with Bayesian Interference is that specifying the appropriate hyperparameters, such as the prior probability distribution and the likelihood function. This requires a detailed understanding and modelling of the data source.

Fuzzy logic is an extension of boolean logic to express the concept of partial truth - allowing for representing values in between true and false. Sensor values can be expressed in terms of fuzzy logic,

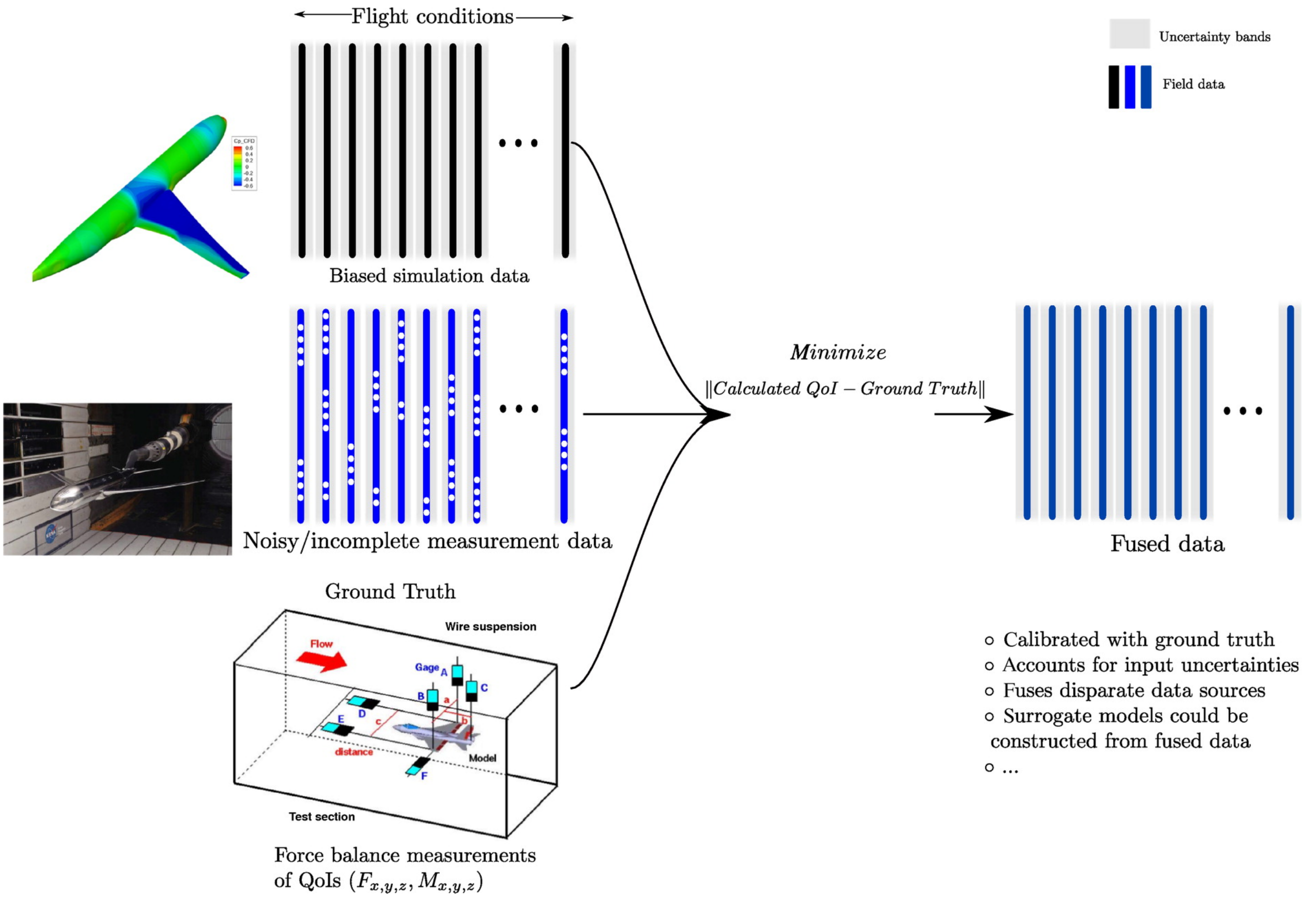

Figure 4: Sensor Fusion of data from a simulated aerodynamic model and field measurements. [21].

and then fuzzy rules can be used to fuse the signals. This multi-sensor fusion increases accuracy because decisions are not based completely on any one input [20].

Lermer and Reich [22] used fuzzy logic to augment Artificial Neural Networks, by using fuzzy rule sets to generate additional parameters. This improves training when there is little data, as the fuzzy rule sets can encode knowledge from domain experts.

Artificial Intelligence methods can be used for sensor fusion, such as to predict the ground truth from multiple competing sensors, or using Principle Component Analysis to reduce the dimensionality of the data [23]. However, when AI is used for sensor fusion, it is often also used for modelling and decision making [22]. Munir et al.[24] classified these different methods as *Sensor Fusion*, *Knowledge/Feature Fusion*, and *Decision Fusion*. Using Artificial Intelligence for Knowledge/Feature Fusion is a type of modelling and will be discussed in section 5. Using Artificial Intelligence for decision Fusion will be discussed in section 6.

Theory of Evidence models are based on Dempster-Shafer theory (DST). The main advantage Theory of Evidence models have over similar techniques such as Bayesian Analysis is that they include methods to model missing values and incomplete data[20]. Saeidi et al. [25] used DST to process data from Light Detection and Ranging (LiDAR) systems by fusing it with multispectral satellite imagery to convert it into land types, and found it was cost-effective and time-effective, because of the ability to use logical reasoning to develop the algorithm to process the data. This also means it is more interpretable than AI-based methods.

4.2 Technologies

Regardless of the methods used for data fusion, the first step is to gather data from multiple sources and tag data with additional metadata such as time or location. Data-gathering tools aggregate data from various IoT devices into one central database. Some platforms include:

- Influxdb [5], a time-series database that integrates well with a variety of platforms and protocols for storing IoT data.
- DynamoDB, [6], a NoSQL database that works well for storing flexible data from IOT sensors.
- Apache Kafka [7], an event streaming platform that enables storing and manipulating data streams.
- SCADA systems also include functionality to tag and store data, or to pass it on to other systems for further processing.

4.3 Discussion

The software development project will be able to make use of data streaming platforms such as Apache Kafka and database systems such as Influxdb to enable the collection and processing of data in real time. These technologies should be chosen to allow more advanced methods of sensor fusion, such as Bayesian Analysis, Fuzzy Logic, Artificial Intelligence, and Theory of Evidence models to be used to process raw sensor data. The exact method to use will depend on the application. Hence the project should include preprocessing functionality that is adaptable to generic or custom implementations of these algorithms.

Sensor fusion techniques also help to increase the accuracy and reliability of source data, as data from multiple sources are cross-validated. This enables higher-level modelling and simulation algorithms to be more precise, enabling better decision-making.

5 Turning Information into Knowledge via Modelling

5.1 Modelling and Simulation

Modelling is the process of defining relationships between entities. There are a number of different methods and platforms for modelling a digital twin system. This literature review will focus on those most relevant to processing live data for industrial digital twins.

The primary purpose of modelling is to enable simulation, which is the process of using a model to generate predictions of what might occur under certain conditions. De Paula Ferreira et al. [26] conducted a literature review of simulation in Industry 4.0. Simulation is used in many contexts, but some approaches were identified that are applicable to the context of processing live data. These include Analytical methods such as Agent-based modelling, Discrete Event Simulation, System Dynamics, and Data-Driven methods such as Artificial Intelligence. The literature review identified that Hybrid Modelling and real-time simulations provided opportunities for further research. They also identified that incompatibilities between simulation platforms were a major barrier to use.

The following sections first discuss Analytical modelling methods, including standards for model formulation, and Data-Driven modelling. Then research into hybrid modelling approaches and real-time simulations will be reviewed.

5.2 Analytical Modelling methods

Agent-based modelling is the process of defining rules for individual agents, or entities in a system. When simulated, emergent properties can be identified from the interactions between the entities. Abar

[5] `http://influxdata.com/`
[6] `https://aws.amazon.com/dynamodb/iot/`
[7] `https://kafka.apache.org/`

et al. [27] conducted a review of tools that provide functionality for agent-based simulation. In industry, agent-based modelling is most helpful for logistics, to optimise factory plans, and for 3d simulation. However, it is not usually appropriate for modelling chemical processes.

Discrete Event Simulation is the process of simulating a process of events as if they occurred one at a time. This method of simulation is used in industrial settings primarily related to the optimisation of scheduling and supply chains [28] such as to ensure full capacity utilisation of processing equipment [29].

System Dynamics models the change of a system over time, by modelling the relationships between components of a system. In an industrial setting, Process Systems Engineering (PSE) uses mathematical modelling and data analytics to model a chemical system. Using Mathematical Modelling in Process Systems Engineering can often be considered an application of System Dynamics. System Dynamics is useful for Process Control and enables modelling interactions such as chemical reactions or thermodynamic interactions such as thermal coupling [30].

5.2.1 Model Formulation Standards

Blaž Rodič [31] conducted several case studies on the use of simulation modelling in the industry, and found that standardised methodologies would allow better integration between platforms. Consistent ontological models allow higher-level applications to interact with different data types.

There are some efforts in the industry to come up with standardised data models. Some are open, enabling easier adoption across platforms, and some are proprietary, but provide high compatibility with other products made by the same company.

Fiware[8] is an open-source platform that enables accessing, processing, and storing data using standardised NGSI-LD models. These models have standard definitions for a variety of entities (such as buildings or IoT devices), properties (such as temperature) and relationships between entities [32]. These can be extended to handle specific use cases[9] but provide the most benefit if they are standardised.

A study on NGSI modelling for smart cities found that the added context made the data more understandable and navigable [33]. It is likely the concepts are applicable to other digital twin situations.

Thingworx [10] is a similar platform to Fiware, that uses a proprietary information model rather than NGSI. It is made by PTC and caters towards the Industrial IoT market.

Siemens Insights Hub [11] also provides a similar service. It is focused on connecting to PLCs in factories and creating a data model that can then be analysed and used in other services.

Modelica [12] is an object-oriented language for modelling cyber-physical systems. It enables defining objects and mathematical relationships between objects. The Modelica specification language is used to model many different types of physical systems, including electrical, hydraulic, heat and fluid flow, state machines, and mechanical systems. [34] There are many libraries of models available which cover these systems to varying levels of detail. A number of commercial and open-source tools provide simulation environments that can run Modelica models.

5.2.2 Modelling Tools

Anylogic [13] is a modelling solution that supports Agent-Based Modelling, Discrete Event Modeling, and System Dynamics. A whitepaper published by The Anylogic Company [35] illustrates how all

[8] `https://fiware.org`

[9] `https://ngsi-ld-tutorials.readthedocs.io/en/latest/understanding-@context.html#creating-ngsi-ld-data-models`

[10] `https://support.ptc.com/help/thingworx/platform/r9.5/en/`

[11] `https://documentation.mindsphere.io/`

[12] `https://modelica.org/`

[13] `https://anylogic.com/`

three different simulation approaches can be integrated and simulated together using Anylogic. However, Anylogic is designed for logistical and economic modelling, and while custom logic can be programmed in Java it would not be as appropriate for modelling complex chemical processes.

MapleSim [14] is a similar commercial modelling solution. It supports Model Based Systems Engineering and Modelica models, enabling it to work with a wider variety of external services and libraries, including more complex processes.

5.3 Data-Driven Simulation Modelling

Machine Learning is also applied in order to predict output conditions of a complex industrial process, given readings of input conditions. One study created an Artificial neural network that was able to predict the power output, steam and exhaust flow rates, and other parameters of a coal biomass cofired combined heat and power plant based on input air, fuel, and water conditions [36]. This was shown to have an error rate of below 5% for most operating conditions.

5.4 Hybrid approaches

More recently, research has focused on hybrid modelling that combines mathematical models and machine learning models. There are two main ways of doing this [37]: Embedding a machine learning model inside a mathematical model, or constraining a machine learning model to more closely follow a specific mathematical formulation.

Embedding machine learning models inside a mathematical model is the more common method of hybrid modelling. Ceccon et al. [38] developed an open-source package named OMLT. It is able to convert Neural Networks and Gradient Boosted Trees into models in the algebraic modelling language Pyomo. This enables embedding a machine learning model into a mathematical digital twin in order to model system interactions that are too complex or expensive to solve from first principles. However, OMLT is only able to support machine learning models that can be written as a set of algebraic constraints. It currently supports Neural Networks and Gradient Boosted Trees.

The alternative approach, constraining a machine learning model to follow a mathematical formula, was discussed by Bikmukhametov and Jäschke [39]. They presented a framework to design physics-aware machine learning models, by passing the outputs of a first principles simulation as inputs into a machine learning model. They found that generally, this type of "feature engineering" improved results, but in some cases it did not. Nevertheless, the models created were much more transparent than models that only were trained on raw data, which helps when evaluation of algorithms is required to ensure consistent performance.

Modulus [15] is a new framework developed by Nvidia to train and deploy AI models that also follow physics-based principles. These models are trained based on simulation data and equations representing the first principles simulation of the model. A case study[16] used Nvidia Modulus to accurately model a Heat Recovery Steam Generator, without having to do full fluid dynamics calculations. This enabled a Digital Twin to be created that could accurately predict when corrosion occurs.

5.5 Live Modelling

Models can be informed based on historical data, or they can be updated and used in real time. Updating and simulating models in real-time is called Live Modelling. This technology is crucial for digital twins because the models in a digital twin must be updated in real time to reflect real-world changes in the system.

[14] `https://www.maplesoft.com/products/maplesim/`

[15] `https://developer.nvidia.com/modulus`

[16] `https://www.nvidia.com/en-us/on-demand/session/gtcspring22-s41671/?playlistId=playList-c1c5c322-57be-46af-8841-e418a2f70c2c`

Live Modelling can be viewed as a combination of conventional modelling and information gained from sensor networks. Simulations are run in real time based on the information gathered from sensor data, and the results of the simulation are stored. Lugaresi and Matta [40] noted that most live modelling applications in the industry focus on scheduling, so other applications of real-time simulation could be researched further.

Various commercial modelling solutions include functionality for live modelling. Simulink [17] is a modelling software with a focus on automation and live modelling. Its focus is more on robotics, artificial intelligence, and electrical and signal processing. Anylogic supports using live data for digital twinning on a logistical and operations level and provides a good case study on how live modelling can be used. A whitepaper on digital twins [41] used models based on live data from operation stages and status, and key performance and financial indicators. Their examples showed that while simulation and modelling can identify bottlenecks and constraints, live modelling could be used to manage solutions. While the whitepaper developed a higher-level scheduling and operations simulation, these same techniques could be applied to more technical simulation platforms.

5.6 Discussion

To best fit the requirements of the software development project, the models used need to be Live Models, able to update in real time to reflect changes in the physical system. A hybrid modeling approach would best suit this. Analytical modelling forms the mathematical basis of the project because it enables greater transparency. Transparency is important in the industry where reliability needs to be assessed and the model's accuracy needs to be validated. However, elements of machine learning still need to be applied. Online machine learning can adjust the parameters of the analytical model dynamically, to keep the model up to date with the physical system. The models chosen will also need to be performant enough to run in real-time. This means choosing the model resolution is crucial - if the system is modeled at too high of a level, accuracy will be sacrificed. However, if the model is very detailed it may take too long to process the live data, delaying analysis and decision-making on the results.

Modeling forms the core of live data processing for Digital Twins. The sensor fusion techniques discussed in section 4 need to be chosen in order to get data that fits the requirements of the DT model. Likewise, the accuracy and detail of the model define how useful it is in creating strategies to solve problems, the penultimate goal of DT modeling.

6 Problem Solving through Simulation Modelling

A digital twin, consisting of a computer-based model informed by real data, provides a form of "digital knowledge" about the system. This knowledge does not provide value or improvement unless interpreted in a way that solves some problem. In the Data-Information-Knowledge-Wisdom hierarchy, interpreting knowledge is the final step to developing "wisdom", or strategies to solve problems. The problems digital twins solve were outlined in section 2.1, "Purpose of Digital Twins". This section will discuss how the "digital knowledge" can be applied to solve the problems identified, including Diagnostics and Fault Classification, and Schedule, Control, and Asset Optimisation.

6.1 Diagnostics & Fault Classification

Traditionally, Machine Learning an effective method to transform raw data from an industrial process into meaningful results, such as for fault diagnosis. Many different algorithms have been used for this. Pang et al. [42] combined data from a wind turbine SCADA system and used various convolutional

[17] https://www.mathworks.com/products/simulink.html

neural network techniques to predict faults. They used a convolutional neural network to iterate over recent readings of all sensors simultaneously. They also performed this analysis using shallow machine learning methods and found that the deep learning methods performed relatively better.

These same methods may be applied to the outputs of digital twin models. Tao et al. performed a case study on a digital twin system for a wind turbine in a power plant [8]. They designed the Digital Twin to model the gearbox, factoring in the torque, vibration frequency, and variance from the physical model, with the maximum contact stress, maximum bending stress, and gear meshing number calculated by the Digital Twin. An Extreme Learning Machine (ELM), a single hidden layer neural network, was trained to predict fault causes on the output of the digital twin. This performed up to 30% better than an equivalent ELM trained on the data from the physical model, without the extra parameters provided by the Digital Twin.

6.2 Optimisation

Because digital twins replicate not just the system inputs and outputs, but also the internal structure, they have increased flexibility to accurately simulate and optimise systems. Schulse et al. introduced the concept of "Experimental Digital Twins", which is the practice of using a digital twin model (complete with information about the digital twin's current environment) to simulate different methods of performing the same task [43]. This involved creating a virtual "testbed" for the digital twin, in order to see how it performed under different conditions. Optimisation functions can be defined and multiple experiments can be run in identical virtual conditions to see which would perform the best. In one test, a 3-D digital twin of a robot was used, and simulations of various procedures were used to find the optimal position to install a camera to maximise the visibility of the work the robot was performing. They concluded that this same technology could be used in the development of future systems, and to improve the operation of existing systems. While the techniques discussed in this paper were applied to 3-D robotic simulation, the approach of creating a virtual test bed and simulating various optimisation functions can still be applied to other types of digital twins.

In a Process Engineering context, there are three main ways to optimise a system: Asset Optimisation, Schedule Optimisation, and Control Optimisation. The following subsections will discuss each of these in turn, and analyse how Digital Twins and Live Data can be used to assist them.

6.2.1 Asset Optimisation

Asset optimisation is the process of retrofitting or upgrading assets in an industrial process to reduce operation costs. A common way to perform asset optimisation is pinch analysis [44] which analyses streams that need to be heated or cooled to find ways to transfer heat from one to the other. This decreases the amount of external work required to provide heating and cooling. Other methods include analysing waste costs, reaction quality, mass balances, or safety indicators to compare theoretical optimums and current performance [45]. In these cases, data collection is important, and digital twins can enable more accurate data collection and estimation of system properties. However, Asset Optimisation is by nature an "offline" process, so real-time analysis is not critical.

6.2.2 Schedule Optimisation

Scheduling involves finding the most efficient way to run a process with minimum overhead between tasks. Georgiadis et al. [46] reviewed the theory and current real-life industrial applications of process scheduling. They concluded that while there was substantial theory behind schedule optimisation, more integration was needed between the mathematical models and industrial Enterprise Resource Planning systems. However, digital twin technology and live data were not specifically addressed as a way to improve this, which could be an area for future investigation.

Others have investigated how schedule optimisation can become self-adaptive using live data. Qiao et al. [47] proposed a framework based on Cyber-Physical systems, where the Cyber-Physical system is enabled to analyse live data (via Digital Twin technology or another method) to identify disturbances. The system could then automatically recompute its schedules to find the optimum method of continuing operations based on the disturbance. The Cyber-Physical system used the same pattern of information processing already identified in this literature review - converting *data* into meaningful *information*, updating scheduling *knowledge*, and creating a coherent *strategy* to continue operations.

6.2.3 Control Optimisation

In this context, "Control Optimisation" refers to adjusting factory controls (such as PID controller targets) to their optimal values. This is related to the field of Model Predictive Control (MPC). Maxim et al. [48] provided an overview of Control in Industry 4.0, and identified that tools to monitor and optimise the performance of PID controllers via parameter tuning are generally available, but that there is more room to iterate in the field of Model Predictive Control and plant-wide economic optimisation. They also concluded that Digital Twins combined with data-driven control design may improve performance. Flowsheet simulation tools such as IDAES [18] are able to perform model predictive control, as demonstrated by Parker et al. [49].

Alves de Araujo Junior et al. [23] created a digital twin using a machine learning model based on automatically extracting fuzzy rules. They demonstrated it could optimise the number of fans needed for a water cooling system based on the factory's power data, the radiator inlet temperature, and ambient temperature data. This was able to have a low error response even with a relatively small amount of training data, saving an estimated 1.44MWh of energy per day.

6.3 Discussion

Problem-solving algorithms are likewise highly context-dependent. The best way to support them is to ensure that real-time data streams from DT modeling is available through standard databases or data streaming protocols. This will enable piping data into other existing control platforms, and developing custom algorithms, or custom machine learning models.

The software development project will focus mainly on control optimisation. This forms a good first use case for live data processing, because it is well established, and factory control systems can be easily compared in a timely manner, using either simulated or real data. If enough time is available, fault classification and schedule optimisation algorithms can be explored, but asset optimisation does not benefit from live data processing in the same way. By making the data from DT modelling available through standard data-streaming protocols, it will not be a technical challenge to apply the project to these different areas.

7 Summary

This literature review covers the data processing pipeline through the lens of a Data-Information-Knowledge-Wisdom hierarchy. By doing so, a broad-spectrum survey of research and technologies that enable insights from live industrial process data is presented. These insights have been discussed in the context of the software development project to develop a system that ingests real-time sensor data from chemical processing plants and use a mathematical model to simulate the plant in real time. Knowledge gaps and additional requirements have been identified, and a high-level development plan has been outlined, fulfilling the purpose of the literature review as stated in section 2.1. The research questions in section 1.4 have been answered as follows:

[18] `https://idaes-pse.readthedocs.io/en/stable/`

RQ1: Where is real-time data collection most used in industrial processing facilities, and where is there potential for it to be used better?

Real-time data is used extensively for monitoring and improving operational efficiency, particularly for schedule optimisation. It shows promise for more advanced diagnostics, fault classification, and control optimisation. Real-time data is a crucial element of the emerging field of Digital Twins because it enables the virtual model to be aware of changing conditions in the real world.

This provides key requirements for the software development project. The project will need to provide reliable software because it needs to be able to work continuously to enable crucial tasks such as diagnostics and optimisation to take place. The software must be scalable enough to work in large and small industrial settings and needs to provide interoperability with existing systems that perform some of these roles already.

RQ2: What techniques and processes are required to turn real-time data into useful insights that can inform and automate decisions in an industrial facility?

This question has been answered by outlining the data processing pipeline. Raw data is collected from SCADA systems, IoT networks, and system logs. This data is organised and labelled, tagging it with relevant context such as timestamps and source information. Sensor data from multiple sources can then be fused together to provide more meaningful, understandable interpretations. This increases the accuracy and enhances fault tolerance. These interpretations serve as a bridge between raw data and model simulation. Simulation is used to estimate other parameters of the system that may not be able to be measured directly by sensors, enabling a comprehensive view of the entire industrial process's state. Finally, manually defined or data-driven algorithms can be used to respond to the process's state, to optimise schedule, control, or maintenance tasks.

This provides context for the software development project. As the project's aim is to ingest real-time data from an industrial process and use a mathematical model to estimate other parameters in real-time, this sits the project squarely in the sensor fusion and model simulation areas of the data processing pipeline. Hence, the inputs and outputs of the project can be clearly defined. It will need to be able to interface with lower-level data collection software, as they will provide the input data of the project. It will need to be able to output simulation results in real-time in a manner that enables higher-level algorithms for process optimisation to be developed on top of them.

RQ3: How can real-time data be integrated into a Digital Twin that is based on mathematical, first-principles simulation techniques?

Some key methods to overcome these technical challenges have been identified. Advanced methods for sensor fusion, such as Bayesian Analysis, Fuzzy Logic, and Theory of Evidence models, help convert data into a more interpretable format that is appropriate for use in an analytical or mathematical model. Furthermore, Hybrid modelling has been shown to enable more flexible modelling that still retains the advantages of mathematical modelling. It is better suited to the variation and noise that is inherent in real-world sensor measurements. Hybrid Modelling also enables the modelling of complex dynamics that are difficult to capture via mathematical modelling alone. Additionally, live modelling augments models with the ability to update themselves in real time, so that the model can reflect changes in the real-world environment.

This provides a basis for developing the core features of the software development project. The project will need to incorporate existing or novel sensor fusion technologies to prepare data for modelling. A mathematical modelling framework will need to be chosen that supports hybrid modelling, and technologies will need to be developed to enable live modelling in that environment. Further research and review will be required in this area to find the optimal technologies and methods to use to develop these features. The optimal solution will depend highly on the external constraints and requirements of the stakeholders.

These insights enhance the project proposal in appendix **??** by specifying *why* and *how* the project can be built. By updating the specifications and requirements of the project based on the discussions in this literature review, the software development project is most likely to achieve its purpose to minimise industrial process energy use via efficient scheduling, optimisation, and maintenance procedures.

References

[1] World Economic Forum. Global Risks Report. January 2023. Accessible at https://www.weforum.org/publications/global-risks-report-2023.

[2] Ministry of Business, Innovation & Employment. Low Emissions Economy: Decarbonising Process Heat. August 2022. Accessible at https://www.mbie.govt.nz/building-and-energy/energy-and-natural-resources/low-emissions-economy/decarbonising-process-heat/.

[3] Timothy Gordon Walmsley, Panos Patros, Wei Yu, Brent R. Young, Stephen Burroughs, Mark Apperley, James K. Carson, Isuru A. Udugama, Hattachai Aeowjaroenlap, Martin J. Atkins, and Michael R. W. Walmsley. Adaptive digital twins for energy-intensive industries and their local communities. *Digital Chemical Engineering*, 10:100139, March 2024.

[4] M. Mazhar Rathore, Syed Attique Shah, Dhirendra Shukla, Elmahdi Bentafat, and Spiridon Bakiras. The role of ai, machine learning, and big data in digital twinning: A systematic literature review, challenges, and opportunities. *IEEE Access*, 9:32030–32052, 2021.

[5] Wei Yu, Panos Patros, Brent Young, Elsa Klinac, and Timothy Gordon Walmsley. Energy digital twin technology for industrial energy management: Classification, challenges and future. *Renewable and Sustainable Energy Reviews*, 161:112407, June 2022.

[6] Yiwen Wu, Ke Zhang, and Yan Zhang. Digital Twin Networks: A Survey. *IEEE Internet of Things Journal*, 8(18):13789–13804, September 2021. Conference Name: IEEE Internet of Things Journal.

[7] Azad M. Madni, Carla C. Madni, and Scott D. Lucero. Leveraging digital twin technology in model-based systems engineering. *Systems*, 7(1), 2019.

[8] Fei Tao, Meng Zhang, Yushan Liu, and A. Y. C. Nee. Digital twin driven prognostics and health management for complex equipment. *CIRP Annals*, 67(1):169–172, January 2018.

[9] Chenhan Zhang and Zhenlei Wang. A data-driven strategy for industrial cracking furnace system scheduling under uncertainty. *Chemical Engineering Science*, 277:118865, August 2023.

[10] Yong Zhang, Ying Yang, Steven X. Ding, and Linlin Li. Data-Driven Design and Optimization of Feedback Control Systems for Industrial Applications. *IEEE Transactions on Industrial Electronics*, 61(11):6409–6417, November 2014. Conference Name: IEEE Transactions on Industrial Electronics.

[11] Krupa Prag, Matthew Woolway, and Turgay Celik. Toward Data-Driven Optimal Control: A Systematic Review of the Landscape. *IEEE Access*, 10:32190–32212, 2022. Conference Name: IEEE Access.

[12] Mohammed Alkatheri. *Data-driven Optimization: Applications to Energy Infrastructure and Process Industry*. Doctoral Thesis, University of Waterloo, December 2021. Accepted: 2021-12-20T15:59:47Z.

[13] Murray E. Jennex. Big data, the internet of things, and the revised knowledge pyramid. *SIGMIS Database*, 48(4):69–79, nov 2017.

[14] Jennifer Rowley. The wisdom hierarchy: representations of the dikw hierarchy. *Journal of Information Science*, 33(2):163–180, 2007.

[15] Louis G. Birta and Gilbert Arbez. *Modelling and Simulation: Exploring Dynamic System Behaviour*. Simulation Foundations, Methods and Applications. Springer International Publishing, Cham, 2019.

[16] Isuru A. Udugama, Carina L. Gargalo, Yoshiyuki Yamashita, Michael A. Taube, Ahmet Palazoglu, Brent R. Young, Krist V. Gernaey, Murat Kulahci, and Christoph Bayer. The Role of Big Data in Industrial (Bio)chemical Process Operations. *Industrial & Engineering Chemistry Research*, 59(34):15283–15297, August 2020. Publisher: American Chemical Society.

[17] Omar Chamorro-Atalaya, Dora Arce-Santillan, Guillermo Morales-Romero, Nicéforo Trinidad-Loli, Adrián Quispe-Andía, and César León-Velarde. SCADA and Distributed Control System of a Chemical Products Dispatch Process. *International Journal of Advanced Computer Science and Applications*, 12(12), 2021.

[18] Sirish L Shah. Sensor-fusion strategies for process and performance monitoring. 2008.

[19] Baptiste Delporte, Laurent Perroton, Thierry Grandpierre, and Jacques Trichet. Accelerometer and Magnetometer Based Gyroscope Emulation on Smart Sensor for a Virtual Reality Application. *Sensors & Transducers.*, 14-1(Special Issue ISSN 1726-5479):p32–p47, March 2012.

[20] Furqan Alam, Rashid Mehmood, Iyad Katib, Nasser N. Albogami, and Aiiad Albeshri. Data Fusion and IoT for Smart Ubiquitous Environments: A Survey. *IEEE Access*, 5:9533–9554, 2017. Conference Name: IEEE Access.

[21] S. Ashwin Renganathan, Kohei Harada, and Dimitri N. Mavris. Aerodynamic Data Fusion Toward the Digital Twin Paradigm. *AIAA Journal*, 58(9):3902–3918, 2020. Publisher: American Institute of Aeronautics and Astronautics _eprint: https://doi.org/10.2514/1.J059203.

[22] Matthias Lermer and Christoph Reich. Creation of Digital Twins by Combining Fuzzy Rules with Artificial Neural Networks. In *IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society*, volume 1, pages 5849–5854, October 2019. ISSN: 2577-1647.

[23] Carlos Antonio Alves de Araujo Junior, Juan Moises Mauricio Villanueva, Rodrigo José Silva de Almeida, and Isaac Emmanuel Azevedo de Medeiros. Digital Twins of the Water Cooling System in a Power Plant Based on Fuzzy Logic. *Sensors*, 21(20):6737, January 2021. Number: 20 Publisher: Multidisciplinary Digital Publishing Institute.

[24] Arslan Munir, Erik Blasch, Jisu Kwon, Joonho Kong, and Alexander Aved. Artificial Intelligence and Data Fusion at the Edge. *IEEE Aerospace and Electronic Systems Magazine*, 36(7):62–78, July 2021. Conference Name: IEEE Aerospace and Electronic Systems Magazine.

[25] Vahideh Saeidi, Biswajeet Pradhan, Mohammed O. Idrees, and Zulkiflee Abd Latif. Fusion of Airborne LiDAR With Multispectral SPOT 5 Image for Enhancement of Feature Extraction Using Dempster–Shafer Theory. *IEEE Transactions on Geoscience and Remote Sensing*, 52(10):6017–6025, October 2014. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

[26] William de Paula Ferreira, Fabiano Armellini, and Luis Antonio De Santa-Eulalia. Simulation in industry 4.0: A state-of-the-art review. *Computers & Industrial Engineering*, 149:106868, November 2020.

[27] Sameera Abar, Georgios K. Theodoropoulos, Pierre Lemarinier, and Gregory M. P. O'Hare. Agent Based Modelling and Simulation tools: A review of the state-of-art software. *Computer Science Review*, 24:13–33, May 2017.

[28] Wilson Trigueiro de Sousa Junior, José Arnaldo Barra Montevechi, Rafael de Carvalho Miranda, and Afonso Teberga Campos. Discrete simulation-based optimization methods for industrial engineering problems: A systematic literature review. *Computers & Industrial Engineering*, 128:526–540, February 2019.

[29] Anupam Keshari, Anand N. Sonsale, Brij K. Sharma, and Sanjay D. Pohekar. Discrete Event Simulation Approach for Energy Efficient Resource Management in Paper & Pulp Industry. *Procedia CIRP*, 78:2–7, January 2018.

[30] E N Pistikopoulos, Ana Barbosa-Povoa, Jay H Lee, Ruth Misener, Alexander Mitsos, G V Reklaitis, V Venkatasubramanian, Fengqi You, and Rafiqul Gani. Process systems engineering – *The generation next? Computers & Chemical Engineering*, 147:107252, April 2021.

[31] Blaž Rodič. Industry 4.0 and the New Simulation Modelling Paradigm. *Organizacija*, 50(3):193–207, August 2017.

[32] Gilles Privat and Alexey Medvedev. Guidelines for modelling with ngsi-ld. *ETSI White Paper*, 42, 2021.

[33] Víctor González, Laura Martín, Juan Ramón Santana, Pablo Sotres, Jorge Lanza, and Luis Sánchez. Reshaping Smart Cities through NGSI-LD Enrichment. *Sensors*, 24(6):1858, January 2024. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[34] *Modelica ® -a unified object-oriented language for systems modeling: language specification*. 2023.

[35] The Anylogic Company. Multimethod Simulation Modelling for Business Applications: Overview with guided model building example. *Anylogic.com*, 2019.

[36] S. De, M. Kaiadi, M. Fast, and M. Assadi. Development of an artificial neural network model for the steam process of a coal biomass cofired combined heat and power (CHP) plant in Sweden. *Energy*, 32(11):2099–2109, November 2007.

[37] Joel Sansana, Mark N. Joswiak, Ivan Castillo, Zhenyu Wang, Ricardo Rendall, Leo H. Chiang, and Marco S. Reis. Recent trends on hybrid modeling for industry 4.0. *Computers Chemical Engineering*, 151:107365, August 2021.

[38] Francesco Ceccon, Jordan Jalving, Joshua Haddad, Alexander Thebelt, Calvin Tsay, Carl D. Laird, and Ruth Misener. OMLT: Optimization & Machine Learning Toolkit. *Journal of Machine Learning Research*, 23(349):1–8, 2022.

[39] Timur Bikmukhametov and Johannes Jäschke. Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models. *Computers Chemical Engineering*, 138:106834, July 2020.

[40] Giovanni Lugaresi and Andrea Matta. REAL-TIME SIMULATION IN MANUFACTURING SYSTEMS: CHALLENGES AND RESEARCH DIRECTIONS. In *2018 Winter Simulation Conference (WSC)*, pages 3319–3330, December 2018. ISSN: 1558-4305.

[41] The Anylogic Company. An Introduction to Digital Twin Development. *Anylogic.com*, 2019.

[42] Yanhua Pang, Qun He, Guoqian Jiang, and Ping Xie. Spatio-temporal fusion neural network for multi-class fault diagnosis of wind turbines based on SCADA data. *Renewable Energy*, 161:510–524, December 2020.

[43] Michael Schluse, Marc Priggemeyer, Linus Atorf, and Juergen Rossmann. Experimentable Digital Twins—Streamlining Simulation-Based Systems Engineering for Industry 4.0. *IEEE Transactions on Industrial Informatics*, 14(4):1722–1731, April 2018.

[44] Ian C. Kemp and Jeng Shiun Lim. *Pinch Analysis for Energy and Carbon Footprint Reduction: User Guide to Process Integration for the Efficient Use of Energy*. Butterworth-Heinemann, August 2020. Google-Books-ID: yW_KDwAAQBAJ.

[45] Ana Carvalho, Rafiqul Gani, and Henrique Matos. Design of sustainable chemical processes: Systematic retrofit analysis generation and evaluation of alternatives. *Process Safety and Environmental Protection*, 86(5):328–346, September 2008.

[46] Georgios P. Georgiadis, Apostolos P. Elekidis, and Michael C. Georgiadis. Optimization-Based Scheduling for the Process Industries: From Theory to Real-Life Industrial Applications. *Processes*, 7(7):438, July 2019. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

[47] Fei Qiao, Juan Liu, and Yumin Ma. Industrial big-data-driven and CPS-based adaptive production scheduling for smart manufacturing. *International Journal of Production Research*, 59(23):7139–7159, December 2021. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00207543.2020.1836417.

[48] Anca Maxim, Dana Copot, Cosmin Copot, and Clara M. Ionescu. The 5W's for Control as Part of Industry 4.0: Why, What, Where, Who, and When—A PID and MPC Control Perspective. *Inventions*, 4(1):10, March 2019. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[49] Robert B. Parker, Bethany L. Nicholson, John D. Siirola, and Lorenz T. Biegler. Model predictive control simulations with block-hierarchical differential–algebraic process models. *Journal of Process Control*, 132:103113, December 2023.