International Journal of Information Management 49 (2019) 502–519

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

# Machine Learning based Digital Twin Framework for Production Optimization in Petrochemical Industry

Qingfei Min[a], Yangguang Lu[a,b], Zhiyong Liu[a,*], Chao Su[a], Bo Wang[c]

[a] School of Economics and Management, Dalian University of Technology, Dalian 116024, PR China
[b] China Wanda Group Co. Ltd., Dongying 257500, PR China
[c] Big Data & IoT Business Development Unit, Lenovo Capital & Incubator Group, Beijing 100085, PR China

ARTICLE INFO

*Keywords:*
digital twin
machine learning
internet of things
petrochemical industry
production control optimization

ABSTRACT

Digital twins, along with the internet of things (IoT), data mining, and machine learning technologies, offer great potential in the transformation of today's manufacturing paradigm toward intelligent manufacturing. Production control in petrochemical industry involves complex circumstances and a high demand for timeliness; therefore, agile and smart controls are important components of intelligent manufacturing in the petrochemical industry. This paper proposes a framework and approaches for constructing a digital twin based on the petrochemical industrial IoT, machine learning and a practice loop for information exchange between the physical factory and a virtual digital twin model to realize production control optimization. Unlike traditional production control approaches, this novel approach integrates machine learning and real-time industrial big data to train and optimize digital twin models. It can support petrochemical and other process manufacturing industries to dynamically adapt to the changing environment, respond in a timely manner to changes in the market due to production optimization, and improve economic benefits. Accounting for environmental characteristics, this paper provides concrete solutions for machine learning difficulties in the petrochemical industry, e.g., high data dimensions, time lags and alignment between time series data, and high demand for immediacy. The approaches were evaluated by applying them in the production unit of a petrochemical factory, and a model was trained via industrial IoT data and used to realize intelligent production control based on real-time data. A case study shows the effectiveness of this approach in the petrochemical industry.

## 1. Introduction

Petrochemical production is a typical process manufacturing industry. The economic efficiency of petrochemical production is largely affected by the portfolio and market prices of the final products. The fluctuation of both market demand and production prices leads to greater requirements for the controllability of the final product portfolio. Therefore, refineries must strengthen their ability to control the production of different components by engaging in continuous technological innovation. The multiple types of final products and complex control variables lead to high demands for safety, stability, and continuity in the production process, as well as strong requirements for instantaneous production control. Therefore, it is of great importance to study how to determine the portfolio of components with advanced production control methods.

Focusing on the special characteristics of petrochemical production, many studies have discussed the production control problem within this industry (Alidi, 1996; Nai-Qi & Bai, 2005; Pach, Berger, Bonte, & Trentesaux, 2014). However, most of the attention has been paid to physical product lifecycle management. Currently, with the application of new-generation information technologies in industry and manufacturing, such as internet of things (IoT)-based digital twins, the convergence between physical products and virtual space has accelerated (Esposito et al., 2018). Digital twins are composed of physical products, virtual products, and connection data that ties physical and virtual products together. The so-called smart factory in the petrochemical industry considers operational excellence its goal. Production control optimization through a high degree of automation, digitization, visualization, modeling and integration is the future direction of petrochemical enterprises (Li, Jiang, Suo, & Guo, 2015; Li, Tao, Cheng, & Zhao, 2015). However, in contrast to this urgent demand for digital twin applications and optimization models in the manufacturing industry, academic research on the application method and framework of digital twins remains in an exploratory stage (Lu, Min, Liu, & Wang, 2019).

* Corresponding author
*E-mail address:* leoliuzy@dlut.edu.cn (Z. Liu).

https://doi.org/10.1016/j.ijinfomgt.2019.05.020
Received 31 October 2018; Received in revised form 24 April 2019; Accepted 21 May 2019
Available online 31 May 2019
0268-4012/ © 2019 Elsevier Ltd. All rights reserved.

Due to the lack of convergence between physical space and virtual space, the data in product lifecycles are isolated, fragmented, and stagnant, offering low value for production control optimization in the petrochemical industry. These problems lead to low levels of efficiency, intelligence, and sustainability in the product design, manufacturing, and service phases. Therefore, we summarize the research gap in production control optimization for the petrochemical industry as follows.

First, there are few methods that can achieve fast and effective interaction between virtual models and real environments, such as addressing time series data issues in the production operation stage for the uninterrupted flow of physical and chemical reactions in closed pipes and vessels through multiple interconnected units.

Second, current data processing methods are isolated and fragmented, and the existing mathematical programming-based methods for short-term scheduling of the refinery industry are not practical in application due to the combinatorial nature and large scale of the problems (Liu, Fan, Wang, & Zhao, 2017).

Past research has indicated that enterprises are deemed successful if they can provide a broad variety of high-quality products while keeping manufacturing and distribution costs low to meet customer expectations and needs (Esposito, Castiglione, Martini, & Choo, 2016; Ferreira, Faria, Azevedo, & Marques, 2017; Santos et al., 2017). Through the digital twin-based framework, factories can control production processes more accurately and agilely in response to changes in market demand. The framework can also help to reduce the cost of inefficient production and improve the economic benefits and sustainable development capabilities of enterprises.

From the above, constructing a digital twin between the physical and cyber worlds is necessary for production simulation and control. Furthermore, offering a practical method with new techniques, such as the implementation of machine learning and IoT techniques, in the production control process is a theoretical and practical direction for the petrochemical manufacturing industry (Esposito, Castiglione, Pop, & Choo, 2017; Li, Jiang et al., 2015, Li, Tao et al., 2015). This paper proposes a digital twin framework for production control based on IoT and machine learning, including the components' architecture, basic steps of the approach, and key evaluation indices. Furthermore, this paper conducts a case study on a real petrochemical factory to examine the effectiveness and value of this framework.

This paper proposes a theoretical framework for digital twin-based production control, which is distinguished from traditional perspectives. This novel framework eliminates the dependency on expert experience and knowledge and avoids the crucial influence of single machine learning results. Industrial big data are utilized to iteratively generate dynamic models according to the changes in environments. This study provides solutions to and a reference for industrial big data analysis and model training, especially in the petrochemical industry. The framework and approach proposed in this paper can be generalized to other process manufacturing industries to introduce a new way to improve their economic benefits through production control.

The remainder of this paper is organized as follows. In section 2, there is a summary of the literature on digital twin research, production control in the petrochemical industry, and machine learning. Section 3 introduces the digital twin framework for production control, including the architecture, data processing and machine learning methods and key evaluation indices. In section 4, a case study shows the effectiveness of this framework and the approach to the petrochemical industry. Theoretical contributions and implications for practice are discussed in section 5. Conclusions, limitations, and future research directions are discussed in section 6.

## 2. Literature Review

### 2.1. Digital Twin

A digital twin as a specific applied technical framework is a realization of the cyber-physical system (CPS), which has received increasing attention from both practitioners and researchers. The concept of the digital twin was first studied in the field of aircraft and aerospace as an information mirroring model concept for spacecraft state simulation and to obtain accurate data for decision-making assistance (Glaessgen & Stargel, 2012; Tuegel, Ingraffea, Eason, & Spottswood, 2011; Tuegel, 2013). Grieves (2014) proposed the digital twin concept from the product lifecycle management (PLM) view and defined it as having three main parts: physical products, virtual products, and the connections of data and information that tie them. Schleich, Anwer, Mathieu, and Wartzack (2017) suggested thinking about digital twins through all stages of product realization and introducing a comprehensive reference model over the product life cycle to shape the digital twin for design and production engineering. Tao, Cheng et al. (2018), Tao, Qi, Liu, & Kusiak (2018), Tao, Sui et al. (2018) provided a digital twin-driven framework to utilize physical product data, virtual product data, and connected data that tie together physical and virtual products for product design, manufacturing, and service (Tao, Cheng et al., 2018; Tao, Qi, Liu, & Kusiak, 2018; Tao, Sui et al., 2018). Miller, Alvarez, and Hartman (2018) presented a sophisticated network of models with a level of interconnectivity based on the concept that a digital twin calls for virtual replicas of real world products. Liu, Zhang, Leng, and Chen (2018) presented a digital twin-driven methodology for generating rapid, individualized designs for an automated flow-shop manufacturing system.

Digital twins have been proven to be a practical method for integrating the physical world and the virtual world of manufacturing, and they support intelligent manufacturing strategies in terms of smart design, smart operations, smart controls and management. However, past research has mostly focused on the PLM view and has attempted to offer methods that achieve virtual reality between real factory/product and digital models. Some practical approaches to planning/designing domains have also been proposed, while research on providing a formal approach to constructing digital twins between the physical and cyber worlds for production control and simulation purposes has been rare. For production control optimization, while many prior studies have focused on product data management, product information tracking or supply-demand matching, the current research on product lifecycle mainly focuses on physical products rather than virtual models (Li, Jiang et al., 2015, Li, Tao et al., 2015; Nai-Qi & Bai, 2005; Tao et al., 2017; Tao, Cheng et al., 2018; Tao, Qi, Liu, & Kusiak, 2018; Tao, Sui et al., 2018).

### 2.2. Production Control in the Petrochemical Industry

Based on the particularity of petrochemical production, many studies have discussed optimization topics with this industry background. Alidi (1996) proposed a multi-objective optimization model based on the goal programming approach to assist in the proper management of hazardous waste in the petrochemical industry. Nai-Qi and Bai (2005) discussed the problems of production planning and short-term scheduling optimization in the petroleum refining industry and the difficulties in implementing production planning and short-term scheduling optimization. Saputelli, Nikolaou, and Economides (2006) provided details on the real-time identification of hybrid models and their use at the scheduling and supervisory control levels, and they proposed a decision-making approach for optimizing the profitability of hydrocarbon reservoirs. Restivo (2006) presented an agile and adaptive manufacturing control architecture that addresses the need for a fast reaction to disturbances at the shop floor level when working in volatile environments, allowing global production optimization to be combined with agile reactions to unexpected disturbances. Al-Sharrah, Elkamel, and Almanssoor (2010) used sustainability indicators as objectives for a mixed-integer optimization model to plan the development of a typical petrochemical industry, and they proved its utility in identifying a balanced petrochemical network. Pach et al. (2014) created a reactive and effective hybrid manufacturing control architecture, combining

hierarchy and heterarchy adapted to the constraints of the industrial market and environment, which they applied to a flexible manufacturing system (FMS) problem. The past research indicates that the existing mathematical programming-based methods for short-term scheduling of the refinery industry would not be practical in applications due to the combinatorial nature and large scale of the problem, and "heuristic + simulative + enumerative" techniques for the problem could therefore be a feasible path.

### 2.3. Machine Learning in Manufacturing

Many theoretical and empirical works have proven that big data-based machine learning approaches, including the use of data mining, pattern recognition, and artificial neural networks, are promising in the manufacturing industry (Carbonell, Michalski, & Mitchell, 1983; Kateris et al., 2014; Köksal, Batmaz, & Testik, 2011; Wen, Li, Lin, Hu, & Huang, 2012; Zhang, 2004). However, studies on the real application of machine learning in industry, especially machine learning applications based on industrial big data and IoT technology, remain rare. Therefore, more efforts are needed to facilitate the relevant framework, methods and applications.

Past research has indicated that machine learning and big data are being increasingly utilized in a variety of industry domains (Rehman, Chang, Batool, & Wah, 2016; Tellaeche & Arana, 2013; Yaqoob et al., 2016). Hatziargyriou (2001) summarized the machine learning applications in power systems. Monostori (2003) introduced hybrid AI and multi-strategy machine learning approaches for managing complexity, changes and uncertainties in manufacturing. Pham et al. (2004) proposed an application for data mining and machine learning techniques in the metal industry. Tellaeche and Arana (2013) analyzed machine learning algorithms for quality control in the plastic molding industry and presented a real case of an application. Rana, Staron, Hansson, Nilsson, and Meding (2014) discussed the factors that influence decision space for the adoption or acceptance of machine learning algorithms in industry. Bilal et al. (2016) discussed the current state of the adoption of big data in the construction industry and the future potential. Meanwhile, big data is seen as "a new type of strategic resource in the digital era and the key factor to drive innovation in the industry, which is changing the way of current production manufacturing" (Lim et al., 2018; Mamonov & Triantoro, 2018; Santos et al., 2017). Zhang, Ren, Liu, and Si (2017) proposed an overall architecture of big data-based analytics to improve PLM and production decisions. Wu et al. (2017) used big data to explore the decisive attributes of supply chain risks and uncertainties. Cheng, Chen, Sun, Zhang, and Tao (2018) noted that the large number of raw data collected from physical manufacturing sites or generated in various information systems caused heavy information overload problems, and most traditional data mining techniques are not yet able to process big data for smart production management. Tao, Cheng et al. (2018), Tao, Qi, Liu, & Kusiak (2018), Tao, Sui et al. (2018) discussed the role of big data in supporting smart manufacturing by providing an overview of the historical perspective on the data lifecycle in manufacturing.

Some research about petrochemical smart factories has appeared in recent years, given the increasing attention paid to advanced or intelligent manufacturing. Li (2016) suggested to thinking differently about the smart factory compared the original production systems used in the petrochemical industry and prioritizing systems thinking and systems problem solving for the smart factory. Yuan, Qin, and Zhao (2017) opined that by using real-time and high-value support systems, smart manufacturing enables a coordinated and performance-oriented manufacturing enterprise and will transform the oil refining and petrochemical sector into a connected, information-driven environment. The above research reflects the development of intelligent manufacturing in the petrochemical industry and highlights future trends in research directions; however, further practical methods of production control are not provided.

The practical value of machine learning and big data in the manufacturing industry has been outlined, demonstrated and highlighted in many existing studies. However, explorations are still needed into the implementation of machine learning techniques in the process manufacturing industry.

In the real world, the physical factory will produce according to the simulation and optimization results from the virtual factory model. Therefore, in the digital twin framework, virtual and physical factories are constantly exchanging data and promoting the continuous optimization of production control, which is a cycle practice loop between the cyber world and the real world. Based on a review of the existing literature, there are several research gaps that need further exploration and study:

1) A concrete and practical framework is still needed that can support the application of digital twin-related theories and approaches to production control and optimization. Approaches to dynamically improving the adaptation of digital twin models to the changing environment should be studied.
2) In production control optimization, several data processing issues must be considered when machine learning and industrial big data are utilized to generate digital twins, which have usually been ignored in previous research: a). time series data from the factory IoT have large dimensions; b). time series data are collected in different time cycles; c). the reaction processes in the devices will influence each other; however, they vary in the degree of correlation between time series data; d). the time lag issue must be considered, which means the time gap, $\Delta t$, that one point must pass to affect another; and e). high demand for the immediacy of process control is based on instantaneous data, as needed by the petrochemical production line.

To fill the above research gaps, a framework and an approach are proposed to generate digital twin models for production control purposes. The framework and approach provide some concrete solutions for data processing and machine learning, and they performed well in a case study from a real petrochemical factory.

## 3. Digital Twin Framework for Petrochemical Production Control

### 3.1. Digital Twin-based Architecture

In previous research on petrochemical intelligent manufacturing, systems directly related to production control consisted of subsystems from four business domains: industrial automation systems, production management systems, business and market strategy systems, and simulation and optimization systems. From the view of production control, the traditional framework between these systems is shown in Fig. 1. The data and information communications between different system domains follow the traditional plan-simulation-execution-control business transaction process logic. Simulation and optimization techniques are mainly based on expert experience and knowledge, or they continuously refer to the one-time results of previous machine learning. Data in the system are used only as the driver of transactions in the single direction of information flow and have not been fully utilized to iteratively improve and update virtual factory models to adapt to changes in production environments.

A significant difference between the previous knowledge-based method and the digital twin-based method for production control optimization is that the previous method basically depends on stable expert knowledge or is based on one-time machine learning output. In contrast, the digital twin-based method builds a continuous interactive process between the physical manufacturing factory and a virtual digital factory. In the digital twin framework, the virtual digital factory will continuously collect real-time data from the physical production line, utilize real-time and historic data for model training, model

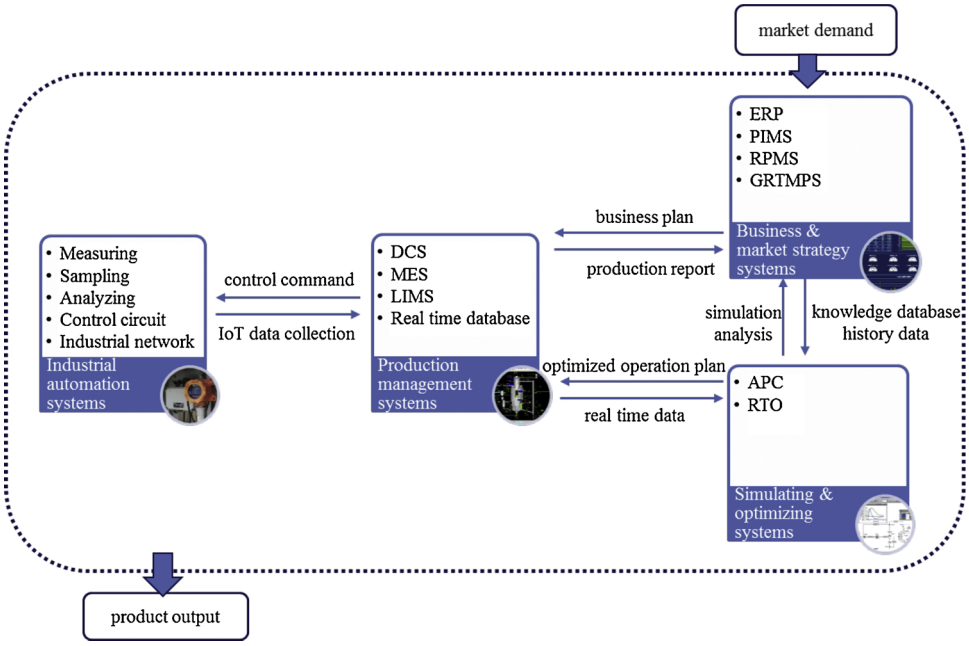


**Fig. 1.** Traditional production control architecture of a petrochemical smart factory.

verifying, and model updating, and ultimately providing feedback to the real factory for production control purposes.

Compared to the previous production control method for petrochemical intelligent manufacturing, the digital twin of the petrochemical factory consists of the following three elements:

1) The physical factory, including the production units, chips, production service systems, environment, and the interconnection between them.
2) The digital factory, including the virtual model, simulation, validation, and digital simulation systems for optimization purposes in production and operation processes.
3) The mapping between the real physical factory and the virtual digital factory.

The digital twin architecture includes the elements of the digital twin and the constant real-time data exchange between the physical factory and the digital model, the data-driven practice loop for continuous optimization and improvement in production, and the full integration of production and information elements. The petrochemical industrial IoT is an important infrastructure needed to realize all of the connections between the real and cyber worlds, as shown in Fig. 2.

The main processes applying digital twin framework into the petrochemical industry, as shown in Fig. 2, are summarized as following:

1) According to the basic framework of production process, production element and expert knowledge, the basic digital framework of process operation mechanism is constructed.
2) Based on the historical big data of existing industrial systems and production and operation systems, the digital twin model is trained by machine learning.
3) According to a series of evaluation indexes, the digital twin model is evaluated, screened and optimized.
4) The final optimized model will be deployed online, combined with the input information of market demand and the optimal solution simulated by the real-time industrial big data on the digital twin model, and feeds back to the distributed control system (DCS) to guide the production control.
5) The digital twin model will be iteratively trained and optimized based on continuously updated and accumulated data to adapt to the continuous changes in the real factory environment.
6) This creates a constant loop between virtual and reality, namely the digital twin practice loops.

The components in the digital twin-based architecture are introduced in the following sections. The data within these components and their application in generating digital twin models for production control optimization are also proposed.

### 3.1.1. *Production Service Systems*

In the petrochemical industry, production service systems are designed for production management and control purposes, which mainly include the following: a). a manufacturing execution system (MES), which realizes real-time and dynamic monitoring and control of the entire production process by collecting, storing, integrating, and utilizing the processing data; it also offers production planning, scheduling, and controlling, as well as material use, energy consumption, and equipment status data for digital twin model training; b). a laboratory information management system (LIMS), which is the laboratory data and information management system designed for the overall environment of the petrochemical laboratory; it offers product and intermediate quality data for digital twin model training; c). a real-time database, which is a system for querying and analyzing real-time information and archiving historical data; it offers centralization of the collection of all real-time and historical data for digital twin model training; and d). a distributed control system (DCS), which is a multi-level computer system comprising process control and process monitoring based on a communication network that is designed to realize the decentralized control, centralized operation and hierarchical management of petrochemical production lines. In a broad sense, a DCS in the petrochemical industry also includes the supervisory control and data acquisition (SCADA) system and the programmable logic controller system (PLC), which directly controls the reaction parameters in the production line in the digital twin practice loop.

### 3.1.2. *Industrial IoT Systems*

Industrial IoT systems consist of online data collection systems and control circuits that transfer the production process command. Online data collection systems in the petrochemical industry are formed by the sensing and measuring systems, industrial networks and various

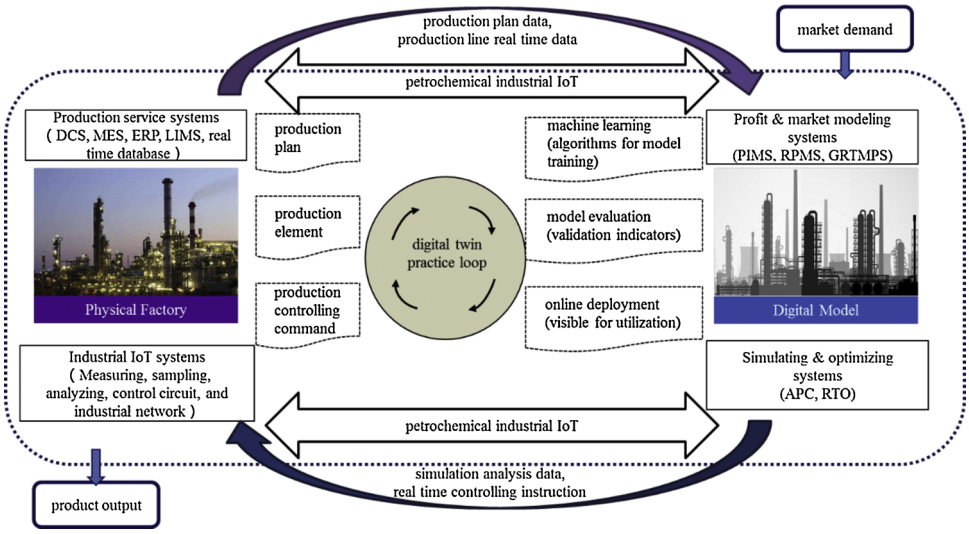


**Fig. 2.** Digital twin-based architecture of petrochemical production control.

sampling and analysis systems. The online collection and analysis system is used for measuring, calculating and detecting the purpose of the chemical composition or physical properties in the petrochemical production process, either continuously or periodically. The systems have certain edge calculation abilities and automatic analysis functions, and they usually consist of six parts:

1)

1) Online measuring instruments, which measure temperature, pressure, liquid level, flow rate and so on.
2) Sampling, preprocessing and injection systems, which take a representative sample from the production process and make it meet the requirements of the analyzer as the sample state or condition.
3) Online analyzers, which analyze the composition or physical properties of the sample and convert the results into measurable electrical signals.
4) Electric and electronic circuits, which act as a power supply to each part of the instrument, control the work of the instrument, amplify the electrical signal sent by the analyzer, and output an electronic signal to the monitor.
5) Monitors and recorders, which are used to display and record electrical signals that represent quantities or properties of components.
6) Industrial internet, which is a digital communication network connecting the production line, computers and server.

To access the production line real-time data, there are usually the following APIs (application programming interfaces) according to their classifications: OLE (object linking and embedding), OPC (OLE for process control), and UIP (user interface program), as shown in Table 1.

### 3.1.3. Profit and Market Modeling Systems

Petrochemical enterprises develop production and operation plans based on profit and market modeling systems, which include a series of software tools for economic planning and the building of crude oil evaluation databases for the process industry. The components of profit and market modeling systems include the following: a). process industry modeling systems (PIMS); b). refinery and petrochemical modeling systems (RPMS); and c). generalized refining transportation marketing planning systems (GRTMPS). The petrochemical enterprise will use these systems to optimize the selection and distribution of upstream sources to track raw material market changes, forecast and analyze market trends, and optimize raw material selection and transportation with collaborative supply chain integration of the upstream and downstream by optimal allocation and comprehensive utilization of raw materials and products. The profit and market modeling systems will offer crude oil types, quality evaluation, properties, components, cutting, reconciliation, price, transportation and usage data for digital twin model training.

### 3.1.4. Simulation and Optimization Systems

Simulation and optimization systems in the petrochemical industry include advanced process control (APC) systems and real-time optimization (RTO) systems. APC refers to a broad range of techniques and technologies implemented within industrial process control systems that address particular performance or economic improvement opportunities in the process, bringing a higher level of computational capability to the control system with valuable and advanced, but not critical, control applications. RTO is a system for realizing the comprehensive integration of planning, scheduling, optimization and control; it uses online calculations to determine the optimal values and then continually updates them in response to disturbances and process variations. It uses algorithms to provide constraint satisfaction, online diagnostics, including optimality guarantees for the plant, and rapid convergence. APC and RTO will offer a prior knowledge base that is a good reference for the digital twin data processing and model training steps, and their control logic could be used for online model deployment.

## 3.2. Digital Twin Modeling

A machine learning-based approach is proposed to form a mathematical digital twin model that simulates the control inputs and outputs. There are 5 primary steps in this method, as shown in Fig. 3.

### 3.2.1. Step 1. Preparation and data collection

Preparation is the necessary investigatory work that helps us to understand the business model in the factory, including the production processes and principles, the meaning of the reaction, what the data represent and the significance of indicator data changes. Data collection means not only collecting model training data from the petrochemical industrial IoT systems and business transaction-driven systems but also data mapping based on knowledge in business models from preparatory works.

### 3.2.2. Step 2. Data feature engineering

Data feature engineering includes data washing, data transforming,

**Table 1**
IoT data types and access API methods

| Data type | Data catalog | Access API |
|---|---|---|
| Material measurement data | Liquid level, flow rate, temperature, tank reserve, etc. | OLE |
| Energy consumption Measurement data | Energy consumption of water, electricity, gas and compressed air, etc. | OLE |
| Quality data | Composition, density, quality, moisture content, etc. | OLE |
| SIS (safety instrument system) lock data | Pump running state signal, interlock bypass signal, interlock action signal, alarm signal, equipment vibration, displacement, speed, oil pressure, etc. | OPC |
| Production processing data | Temperature, pressure, feeding speed, etc. | OPC |
| Manual data | Manual meter reading data, compensation data and manual corrected data, etc. | UIP |

and other necessary steps corresponding to the particularities of the petrochemical industry with the following goals: a). unify the time series data frequency; b). resolve the time lag issue between time series data; c). conduct correlation analysis and reduce data dimensions if necessary; d). analyze the autocorrelation and partial autocorrelation and regenerate new stable time series data if necessary; and e). create new variables from existing time series data and expand the feature indicator dimension if necessary.

#### 3.2.3. Step 3. Model training and validation

To train and validate the model, the collected and transformed data must be divided into two groups: one group for training and the other for validation. The training target is used to construct an accurate mapping relationship based on available data and current algorithms. The mapping is briefly described in formula 1.

$$\boldsymbol{Y_t} = \mathbf{F}(\boldsymbol{X_{t\pm\{\Delta\}}} + \boldsymbol{Z_{t\pm\{\Delta\}}}) \tag{1}$$

In formula 1, Y is the cluster of controlling targets (dependent variables), e.g., the yield of the specified product. X is the cluster of real-time controllable independent variables, e.g., the feeding speed. Z is the cluster of real-time uncontrollable independent variables, e.g., the crude oil quality. $\mathbf{t} \pm \{\Delta\}$ indicates the variation in the time lag between variables.

The training process includes utilizing different machine learning algorithms, e.g., random forest, AdaBoost, XGBoost, gradient boosting decision tree (GBDT), LightGBM, and neural networks. The different models and training results from different algorithms are validated by a validation data group. In most cases, the model output, especially when trained by integration model algorithms (e.g., boosting and bagging types), can be verified in terms of its validity but is not explainable by formal descriptions. The reference accuracy, error of fitting, and coefficient of determination are important indicators for evaluating the model.

If the raw material input volume is *I*, the unit price is *S*, the energy consumed value is *E*, the yield from major units is $\boldsymbol{Y_1}$, $\boldsymbol{Y_2}\ldots\ \boldsymbol{Y_n}$, and the corresponding market price is $\boldsymbol{P_1}$, $\boldsymbol{P_2}\ldots\ \boldsymbol{P_n}$, then the comprehensive economic value *V* is formula 2.

$$\mathbf{V} = \sum_{i=1}^{n} \boldsymbol{P_i} \times \boldsymbol{Y_i} \times \boldsymbol{I} - \boldsymbol{I} \times \boldsymbol{S} - \boldsymbol{E} \tag{2}$$

In the control process, which ultimately aims to maximize economic value, the machine learning target for production control purposes can be described by formula 3.

$$\mathbf{Find}[\boldsymbol{X_i}] \rightarrow \mathbf{max}\{\boldsymbol{V}\} \tag{3}$$

#### 3.2.4. Step 4. Tryout and optimization

The tryout step tests the training model in a real-time production environment with the latest data to verify its effectiveness and security. It is a necessary step before actually using the training model based on the following two considerations: first, security must always be the top priority in petrochemical production environments, so the model must pass the necessary health, security, and environmental checks before being used into production control practices; second, model training is primarily based on historical data, so it must be revalidated in the latest environment.

After the tryout, the model must be optimized according to the test results and feedback from the factory departments. That is, the final model to be deployed online is a “digital twin” model of the real production line for control optimization purposes.

#### 3.2.5. Step 5. Model online deployment

In the model online deployment process, the final digital twin model will be established with a connection to the petrochemical industrial IoT and other related systems, e.g., MES, LIMS, and the real-time database, to obtain the necessary real-time data as input. In turn, the digital twin model will output the control command to the production

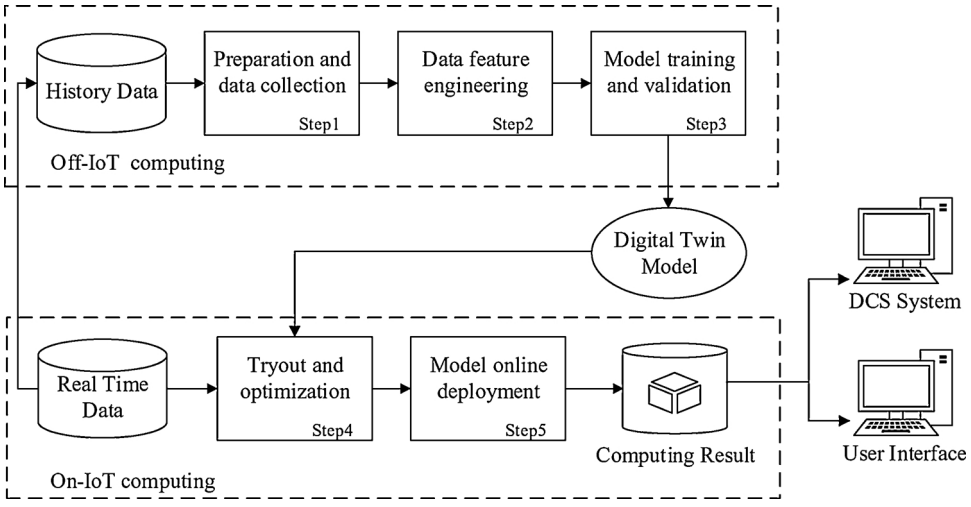


**Fig. 3.** Primary steps of machine learning-based digital twin modeling.

line directly via the DCS or indirectly via operators by integrating visible recommendation information into the MES. For security reasons, the latter approach is usually recommended, especially in experimental projects.

Since the digital twin model is deployed online, the system will continuously search the set of optimal control parameters based on the digital twin model with real-time data. However, the petrochemical production line has a high demand for immediacy in process control. It always requires adjustment to the control points, such as temperature, pressure and liquid levels, in a quick reaction to instantaneous data. If the control process defines the maximization of economic value as the ultimate goal, then the machine learning target for production control is to find the optimized real-time control set based on the final digital twin model and real-time data. There are many search algorithms to find the best control set, such as depth first search (DFS), breadth first search (BFS), grid search, or particle swarm algorithm. In general, when choosing the appropriate search algorithm for the optimized control set, it is important to consider the computing power of the servers and their balance to avoid falling into the local optimum and to ensure timeliness.

Step 1 to step 5 is not a one-time process but a repeating cycle, which is how the digital twin practice loop created for the petrochemical production line continually controls and optimizes, as shown in Fig. 4. In the first round of the practice loop, much manual work is needed for preparation and business understanding, but in the subsequent rounds, all tasks are expected to be executed automatically by the computers between the physical IoT and the cyber-network. The frequency of repetition depends both on the business requirements and on computing performance.

### 3.3. Data Processing

The original data collected from the production line IoT and the business systems are mostly time series data and will be handled through the basic steps of data washing, coding mapping, and processing abnormal and missing data. Several important data preparation actions are needed in the petrochemical industry before the data are used for model training with machine learning.

1) Unification of time series data frequency.

The data collected from the petrochemical industry IoT have continuous features. However, the original IoT data from the production line are usually collected at different frequencies according to their types, as shown in Fig. 5 and Table 2. Therefore, steps must be undertaken to achieve time series data frequency unification, either by expanding indicators that are at a lower frequency or by compressing indicators that are at a higher frequency.

Normally, we can select one of the data dimensions as a benchmark and make the other dimensions align with this benchmark. This process could be handled by servers in the machine learning method, and today, edge calculation technology also makes this process possible on the production line side.

Taking the simple linear mean value method as an example, if 2 sets of IoT data $X_1$ and $X_2$ were collected in different frequencies as $T_1$ and $T_2$, and we consider dimension $X_2$ as the benchmark, then the new data dimension $X'_1$ will be generated based on $X_1$ and aligned with the data sampling frequency of $X_2$ as in formula 4.

$$\begin{cases} m = \left\lfloor \dfrac{T_2 \times j}{T_1} \right\rfloor \\ X'_{1j} = X_{1m} + (X_{1(m+1)} - X_{1m}) \dfrac{T_2 \times j - T_1 \times m}{T_1} \end{cases} \tag{4}$$

Formula 4 shows how to generate a new time series data $X'_1$ based on the original value of time series data $X_1$ and keeps the data frequency unified with $X_2$. If the frequencies of $X_1$ and $X_2$ are different from $T_1$ and $T_2$, then the frequency ratio $\frac{T_2}{T_1}$ is an important reference for finding the new match point of $X'_{1j}$in the original $X_1$ curve. First we find the 2 reference values $X_{1m}$ and $X_{1(m+1)}$, which are before and after $X'_{1j}$ in the $X_1$ curve, and then we use the distance ratio on the time axis between $X_{1m}$, $X'_{1j}$ and $X_{1(m+1)}$ to calculate the approximate value of $X'_{1j}$ based on the weighted averages method.

The calculation procedure for Equation (4) is as following:

Input: $X_1$ and $X_2$, two sets of time series IoT data; $T_1$ and $T_2$, the data collection cycle for $X_1$ and $X_2$ respectively
Create a new time series data $X'_1$
For every $x_{2j}$ in $X_2$
  Corresponding Time point $t_{2j}$=$T_2 \times j$
  Search for $x_{1m}$ and $x_{1n}$ in $X_1$
    where $t_{1m}$= max ($t_{1i}$| $t_{1i} < t_{1j}$)
      $t_{1n}$= min ($t_{1i}$| $t_{1i} > t_{1j}$)
For every $x'_{1j}$ in $X'_1$
  $x'_{1j}$=linear mean($x_{1m}$, $x_{1n}$)
Output: $X'_1$, a new time series data instead of $X_1$ in data base

1) Time lag between time series data.

In the petrochemical industry, the production process is an uninterrupted flow of physical and chemical reactions, and the time lag in the interaction between 2 points must be accounted for, which means that there is delay in the time, Δt, that one reaction point must pass to affect another. When this issue is extended to the mutual relationship of all points in the whole production line, then there is an extremely large time alignment matrix.

To resolve the time lag problem, we propose an approach in the following steps:

Step 1. Normally in the factory technical documents, the reaction process, parameters, and time delay information have already been documented and recorded, which would provide the major information source for constructing the time alignment matrix. Currently, the detailed information factory model, which is known as the visible digital factory model, offers a better resource and data interface for acquiring and calculating the latency time.

Step 2. The prior empirical information from the production line, e.g., the experience from the onsite operators, is also important reference information to supplement the time lag.

Step 3. When the time alignment matrix is created, we set one data dimension (indicator) as the benchmark and generate the other new data dimensions by moving them front and back on the time axis as in

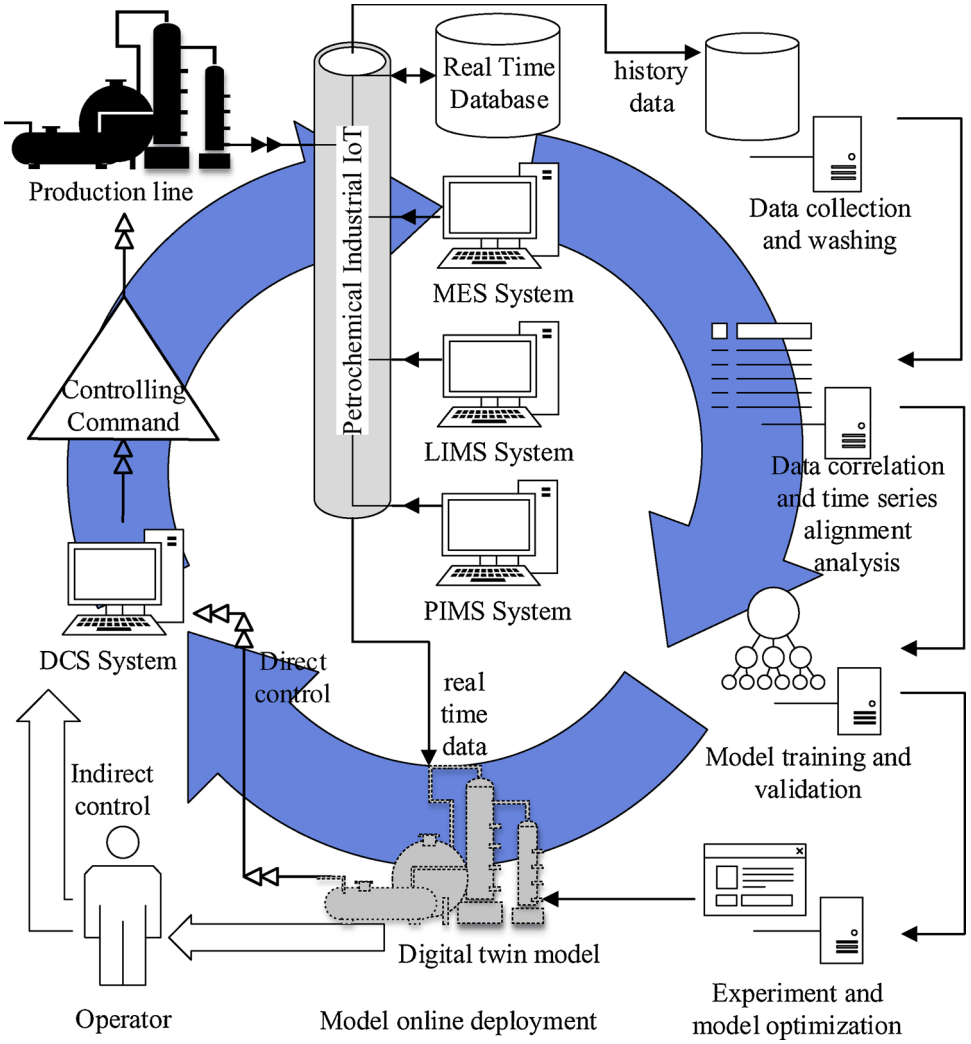


**Fig. 4.** Machine learning-based digital twin practice loop.

formula 5.

$$\begin{cases} \boldsymbol{time\_lag}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \Delta t \\ if\ benchmark = \boldsymbol{X}_1 \\ then\ \boldsymbol{X'}_{2j} = \boldsymbol{X}_{2(j+\lfloor \Delta t \rfloor)} \end{cases} \quad (5)$$

Formula 5 shows how to generate new time series data $X'_2$ based on the original value of time series data $X_2$ and keep the phase position aligned with $X_1$. If we select $X_1$ as the benchmark, and the reaction of point $X_2$ is $\Delta t$ later than $X_1$, then we must move the whole $X_2$ curve "left" by $\Delta t$ distance along the time axis and use the closest value of the original $X_2$ series data as the approximate value of $X'_2$ in each data sampling point.

The calculation procedure for Equation (5) is as following:

Step 4. The statistical errors and empirical errors could cause errors in the time alignment matrix; therefore, the MA solution in formula 6 and the data dimension expansion solution in formula 7 must be utilized as an effective solution for reducing the influence of time alignment value errors.

$$\boldsymbol{X'}_{ij} = \frac{\boldsymbol{X}_{i(j-\mathbf{n})} + \boldsymbol{X}_{i(j-\mathbf{n}+1)} + \ldots + \boldsymbol{X}_{ij} + \boldsymbol{X}_{i(j+1)} + \ldots + \boldsymbol{X}_{i(j+n)}}{2\boldsymbol{n} + 1} \quad (6)$$

$$\begin{cases} max\_error\_expect = \delta \\ dimension\_expansion\_multiple = n \\ \boldsymbol{m} \in \left[-\frac{\boldsymbol{n}}{2}, \frac{\boldsymbol{n}}{2}\right], \ \boldsymbol{and}\ \boldsymbol{X}_{it}^{\ m} = \boldsymbol{X}_{\mathbf{i}\left(t+\frac{\mathbf{m}\times\delta}{\boldsymbol{n}}\right)} \end{cases} \quad (7)$$

Input: $X_1$ and $X_2$, two sets of time series IoT data; $time\_lag$, a data table contains the time lag corresponding to $X_1$ and $X_2$;
Search from $time\_lag$, the time lag between $X_1$ and $X_2$ is $\Delta t$
Create a new time series data $X'_2$
For every $x'_{2j}$ in $X'_2$
  For every $x_{2j}$ in $X_2$
    $x'_{2j} = x_{2j} + \Delta t$
Output: $X'_2$, a new time series data instead of $X_2$ in data base

Formula 7 shows how to expand the data dimension to reduce time lag error effects. If the maximum error expectation of the time lag matrix is δ, and we plan to expand the original time series data $\boldsymbol{X_i}$ to n dimensions, then the time series data in the new dimensions will use the original value of $\boldsymbol{X_i}$ and move "left and right" along the time axis by a total of n times, and the moving length of each step is $\frac{\delta}{n}$.

The calculation procedure for Equation (7) is as following :

Input: $\boldsymbol{\delta}$, maximum error expectation of the time lag matrix; n, dimension expansion multiple; $X_i$ , original dimension data;

Create n new time series data $X_i^{-n/2}$~$X_i^{n/2}$

For every $X_i^m$ in $X_i^{-n/2}$~$X_i^{n/2}$

For every $x_i$ in $X_i$

$x_i^m = x_i + \mathrm{m} \times \frac{\delta}{n}$

Output: $X_i^{-n/2}$~$X_i^{n/2}$, n new time series data instead of $X_1$ in data base

1) Correlation analysis and dimensionality reduction.

Correlation analysis is used to analyze the correlations between different indicators, and Pearson's correlation coefficient (PCC), which is shown in formula 8, is a very common index to evaluate the degree of linear correlation between two indicators. The value of $\rho_{X,Y}$ is between -1 and 1, and the higher that its absolute value is, the greater the correlation between two indicators. This analysis is very helpful in the petrochemical industry for two reasons: a. intuitively analyzing the internal relations between the reaction devices and b. feature selection.

$$\rho_{X,Y} = \frac{\boldsymbol{COV}(\boldsymbol{X}, \boldsymbol{Y})}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(\boldsymbol{X_i} - \bar{\boldsymbol{X}})(\boldsymbol{Y_i} - \bar{\boldsymbol{Y}})}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{X_i} - \bar{\boldsymbol{X}})^2}\sqrt{\sum_{i=1}^{n}(\boldsymbol{Y_i} - \bar{\boldsymbol{Y}})^2}} \quad (8)$$

The petrochemical industry usually has a high density of data collection points, resulting in a large number of indicators and a high data dimension for machine learning. Here, we propose using PCC to generate a correlation matrix map as a more visible method for analyzing high dimension data, which is shown in the practice case.

### 3.4. Model Training Algorithm

Various machine learning algorithms could be used for digital twin model training. However, to better adapt them to the complicated and inexplicable petrochemical inner workings, the basic regression algorithms, e.g., classification and regression tree (CART), must be utilized, along with integration algorithms, e.g., random forest, AdaBoost, XGBoost, GBDT, LightGBM, and neural networks. The algorithms each have advantages and disadvantages; however, in the digital twin practice loop with the support of strong computing capacity, most of the potential algorithms can be tested and compared based on validation indicators to make the final selection of the best optimized model. The functions and features of different machine learning algorithms during the training of digital twin models are introduced in the following sections.

The CART algorithm is mostly used as a regression function that defines the gain value calculation in formula 9, which is used for control model training purposes. The tree split target of CART is to minimize the gain value in each node: $R_1$ and $R_2$ are the 2 split branches in

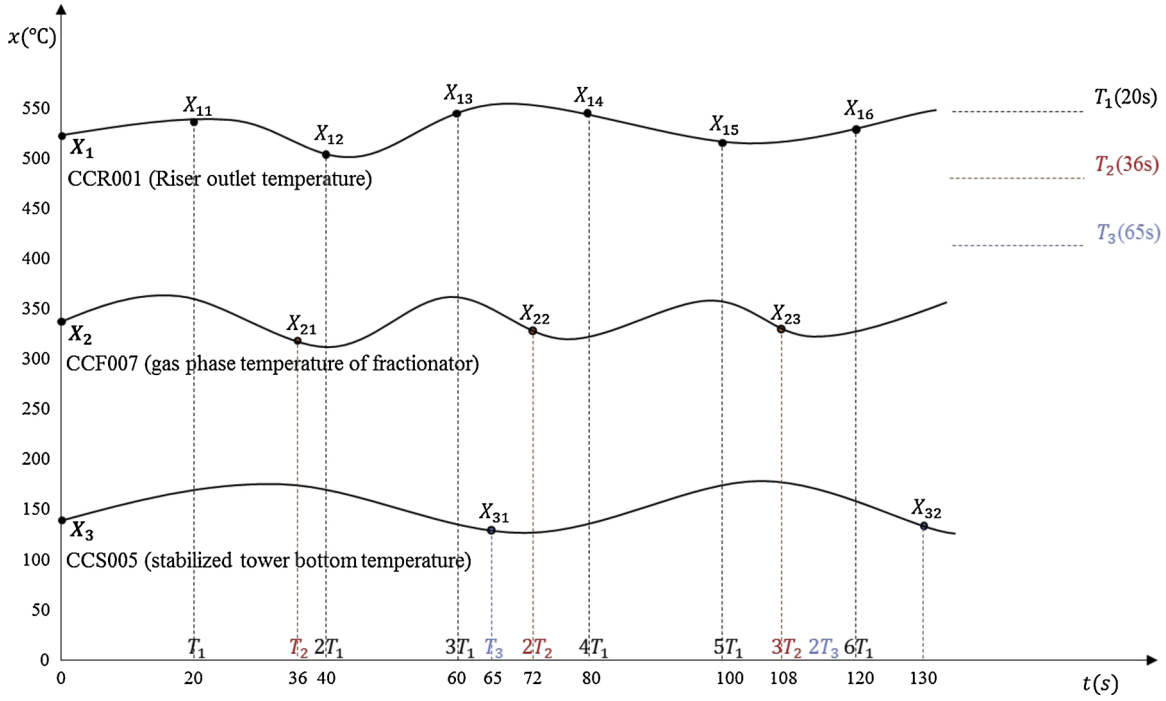


**Fig. 5.** Example of different data collection frequencies.

**Table 2**
Indicator type and data collection cycle

| Type | Specific catalog | Collection cycle |
|---|---|---|
| Critical control point | Control points that have a significant impact on the stable operation of the main unit, e.g., reactor temperature and pressure and heat exchanger temperature, and monitoring points that have an important impacts on production safety, e.g., security points and alarm points | $T_1 \in [5s,\ 30s]$ |
| Normal control point | Other control parameters other than the critical control points, e.g., auxiliary devices and tank measuring points | $T_2 \in (30s,60s]$ |
| Normal metering point | Material and utility metering points, quality inspection data, etc. | $T_3 \in (60s,120s]$ |

each node, and $c_1$ and $c_2$ are the returned regression values of the two split branches.

$$\text{Gain}=\sum_{i\in I}\sigma_i=\sum_{i\in R_1}(y_i-c_1)^2+\sum_{i\in R_2}(y_i-c_2)^2 \tag{9}$$

Random forest is a representative bagging algorithm that randomly extracts samples and features from the data to train several different decision trees and averages the prediction results of all trees to form a "forest" model, while AdaBoost is a representative boosting algorithm that can train a group of weak learners $\{h_1(x), h_2(x), ..., h_T(x)\}$and combine them by weight to generate a strong learner, as shown in formula 10. AdaBoost assigns an equal weight value to each sample at the beginning and adjusts the weight of each sample and each tree according to the error record in each calculation iteration.

$$\mathrm{H}(X)=\sum_{t=1}^{T}a_t h_t(x) \tag{10}$$

Previous research has provided the following three boosting algorisms. Comparing with AdaBoost algorisms, they have different features and performances. a). The GBDT algorithm integrates regression trees for forecasting and classification. It draws its final conclusion by summing the conclusions of all of the trees, and the core concept is that each tree learns the residuals (negative gradients) of the sum of all of the previous tree conclusions, which is the cumulative amount of the true value after adding the predicted value. b). XGBoost is an improved gradient boost with model complexity controlled by the pruning method. The formula used for tree node splitting includes shrinkage and column subsampling. It uses a presorted algorithm to reduce the variance in the model, simplify the learning model, and prevent

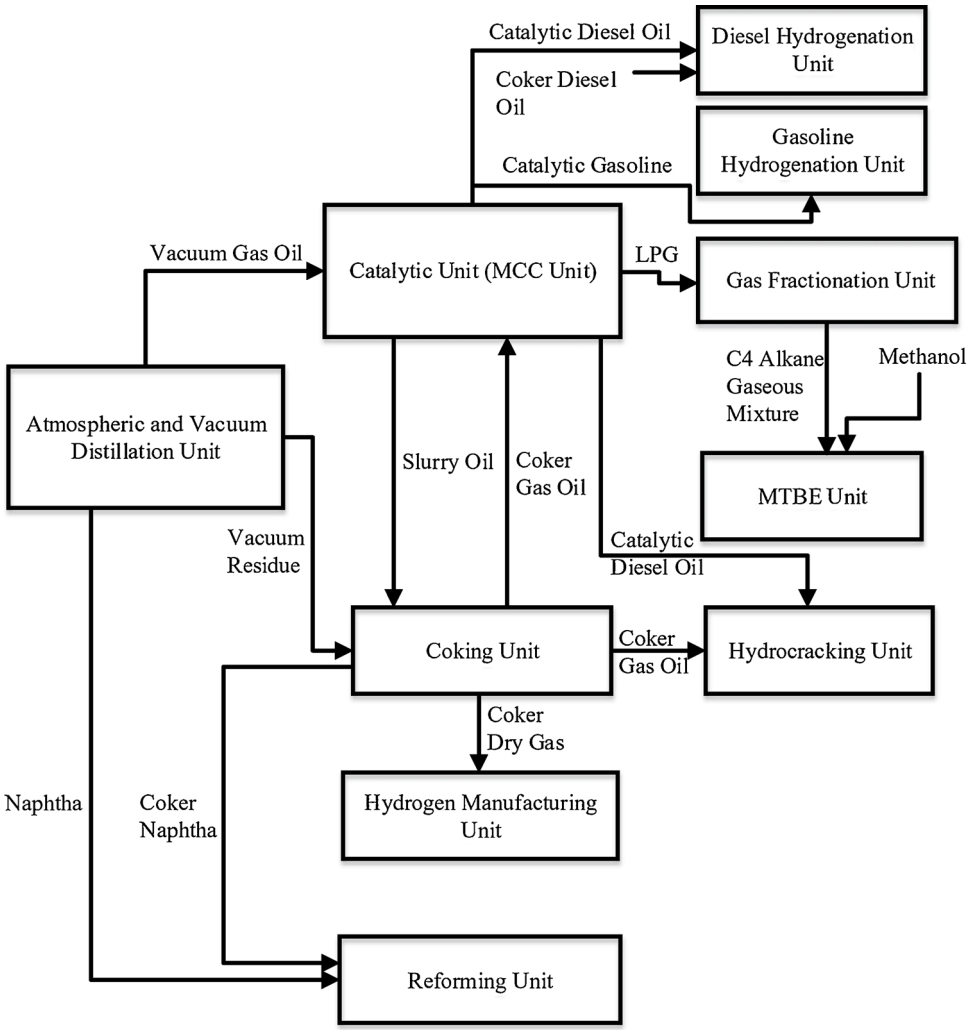


**Fig. 6.** Production units and processes in the MAYA petrochemical factory.

overfitting. c). LightGBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient, with faster training speed and greater efficiency, better accuracy and lower memory usage. It uses the histogram algorithm, thereby accelerating the construction speed, and it is capable of handling large-scale data.

### 3.5. Model Validation Indicators

Because the simulation models were trained by different machine learning algorithms with the training data group, the validation data set and appropriate evaluation indicators are used to evaluate model performance and select the final optimized model. The best digital twin model is then adopted for simulation and an optimized control plan is proposed for the production line.

To evaluate the convergence speed of different algorithms and to intuitively evaluate the prediction effect, the fitting error (FE) index is calculated in formula 11, and the true value and absolute value of FE through a curve are shown to compare the efficiency of different algorithms. $y_i$ is the true value from the training data group, and $\hat{y}_i$ is the predicted value. The smaller the absolute value of FE is, the better it is.

$$\mathrm{FE}= \frac{\log_2(1+\hat{y}_i)-\log_2(1+y_i)}{\log_2(1+y_i)} \tag{11}$$

To comprehensively evaluate the model quality, we propose using four evaluation criteria: the model accuracy ratio (MAR), the root mean square error (RMSE), the variance interpretation rate (VIR), and PCC. The calculation of the indicators is presented in formulas 12-15. Referring to formula 1 in the introductory section on the basic concept of the digital twin model, y is the controlling target (dependent variables), e.g., the yield of the specified product, and x is the prediction label (independent variables). $y_i$ is the true value from the validation data group, $\bar{y}$ is the average value, and $\hat{y}_i$ is the predicted value calculated by the trained model.

$$\mathrm{MAR}= 1-\frac{\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|}{n} \tag{12}$$

Formula 12 shows the calculation of MAR, which is the difference between 1 and the first order deviation rates. MAR can reflect the deviation of the difference between the predicted result and the actual value. The greater MAR is, the better it is, and its maximum value is 1.

$$\mathrm{RMSE}= \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \tag{13}$$

Formula 13 shows the calculation of the RMSE, which is the index used to evaluate the performance of the regression model based on the quadratic distance between the actual value and the predicted value; the smaller the RMSE is, the better it is, and its minimum value is 0.

$$\mathrm{VIR}= 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{14}$$

Formula 14 shows the calculation of VIR, which is transformed from the coefficient of determination. In the petrochemical industry, the target predicted value usually fluctuates within a range that is much smaller than its absolute value, indicating that even random prediction by average historical values will also perform at a “very good” level for accuracy. Therefore, we use this index to show how much better the model prediction is than random prediction. When the value is greater than 0, the effect of the model prediction is better than a random prediction. The greater that VIR is, the better it is, and its maximum value is 1.

$$\mathrm{PCC}= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{15}$$

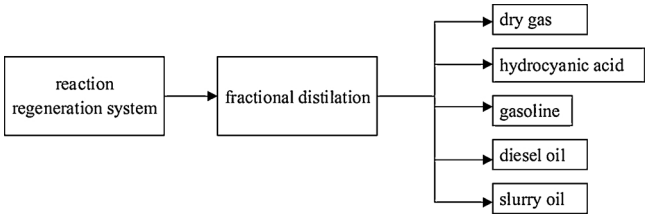


**Fig. 7.** Primary production process of the catalytic cracking unit.

Formula 15 shows the calculation of PCC, which is used to measure the linear correlation of two variables. For the model prediction performance evaluation case, PCC indicates the degree of consistency between the predicted value and the actual value. The greater PCC is, the better it is, and its maximum value is 1.

## 4. Evaluation

### 4.1. Case Background

MAYA factory is located in northern China, and it was founded in 2014, with 6 million tons/year atmospheric and vacuum units, 2 million tons/year MCC units, 1.8 million tons/year coking units, and 1.8 million tons/year diesel hydrogenation units, as shown in Fig. 6. The main products are gasoline, diesel, LPG, propane, propylene, petroleum coke, oil slurry, naphtha, and sulfate MTBE.

Production processes in the MAYA factory are controlled by the internal operations workers (operators) in the controlling hall. There are a total of 120 operating positions in different control halls, and each internal operation worker watches dozens of real-time process control indicators from the production line. Their operations are in response to the monitoring information and are accomplished according to predefined operating procedures, a specified limited range, and personal experience.

Since 2016, the MAYA factory has faced challenges caused by fluctuations in market demand and price, which have forced more competitive production control, especially in terms of the yield of components of different products. At the end of 2017, MAYA started an intelligent manufacturing project, which included the aim of improving yield control ability with machine learning. The catalytic cracking unit was selected as the target unit for a pilot project, and the factory operators set the primary target to increase the yield of light oil (gasoline and diesel oil) with a machine learning-based control optimization method.

### 4.2. Case Analysis

The catalytic cracking unit is an important component of the petrochemical production line. As a typical reaction regeneration system, its production process can be simplified, as shown in Fig. 7.

The process of the catalytic cracking unit includes 5 major facilities and 40 control points, as shown in Table 3. The control target is the yield of light oil, and 410 sets of indicators collected from the production line analytics and metering instruments and other systems could potentially influence the control target. The 40 control points are some of the 410 indicators, so there are 370 sets of uncontrollable indicators.

According to our method, we must utilize machine learning technology with the obtained data and train the model F that most accurately simulates the real factory production circumstance as $\boldsymbol{Y_t} = \mathbf{F}(\boldsymbol{X}_{t\pm\{\Delta\}} + \boldsymbol{Z}_{t\pm\{\Delta\}})$, where Y is the yield of light oil, X is the 40 controlling points, and Z is the other 370 sets of uncontrollable indicators. Then, the model is deployed online as the digital twin to simulate real-time data from the production line and to provide a control optimization proposal.

**Table 3**
The control points of the catalytic cracking unit in MAYA

| No | Facility | Control Points | Code | Unit | Threshold |
|---|---|---|---|---|---|
| 1 | Reaction | riser outlet temperature | CCR001 | ℃ | 505-525 |
| 2 | | raw oil inlet temperature | CCR002 | ℃ | ≤300 |
| 3 | | pre-upgrade steam flow | CCR003 | t/h | ≤2.3 |
| 4 | | settler pressure | CCR004 | MPa | 0.150-0.190 |
| 5 | | regenerator pressure | CCR005 | MPa | 0.180-0.220 |
| 6 | | two-device pressure difference | CCR006 | KPa | 30-50 |
| 7 | | stripping section storage | CCR007 | T | 25 ± 10 |
| 8 | Fractional distillation | oil and gas separator liquid level | CCF001 | % | 10-30 |
| 9 | | oil and gas separator boundary position | CCF002 | % | 30-50 |
| 10 | | diesel stripping tower liquid level | CCF003 | % | 40-60 |
| 11 | | bottom level of fractionating tower | CCF004 | % | 30-70 |
| 12 | | sealing tank level | CCF005 | % | 55-80 |
| 13 | | liquid temperature of fractionator | CCF006 | ℃ | ≤365 |
| 14 | | gas phase temperature of fractionator | CCF007 | ℃ | 370-400 |
| 15 | | fractionator top pressure | CCF008 | MPa | 0.12 ± 0.02 |
| 16 | | fractionator top temperature | CCF009 | ℃ | 120 ± 10 |
| 17 | | product slurry to tank temperature | CCF010 | ℃ | ≤120 |
| 18 | | product slurry to coking temperature | CCF011 | ℃ | 90—150 |
| 19 | | diesel transport temperature | CCF012 | ℃ | ≤65 |
| 20 | Stabilization | V1302 level | CCS001 | % | 30-50 |
| 21 | | V1302 bound position | CCS002 | % | 30-50 |
| 22 | | V1303 level | CCS003 | % | 20-50 |
| 23 | | absorption tower top temperature | CCS004 | ℃ | 40 ± 10 |
| 24 | | stabilized tower bottom temperature | CCS005 | ℃ | 165-180 |
| 25 | | stabilized tower top temperature | CCS006 | ℃ | 50-65 |
| 26 | | reabsorption tower top pressure | CCS007 | MPa | 0.8 ± 0.2 |
| 27 | | V1303 pressure | CCS008 | MPa | ≤1.0 |
| 28 | Thermal | deaerator liquid level | CCT001 | % | 60-80 |
| 29 | | medium pressure superheated steam temperature | CCT002 | ℃ | ≥380 |
| 30 | | medium pressure drum pressure | CCT003 | MPa | 3.8 ± 0.3 |
| 31 | | medium pressure drum liquid level | CCT004 | % | 30-60 |
| 32 | Machine set | main fan lubricating oil pressure | CCM001 | MPa | 0.26-0.38 |
| 33 | | main fan lubricating oil temperature | CCM002 | ℃ | 35 ± 5 |
| 34 | | turbocharger lubricating temperature | CCM003 | ℃ | 40 ± 5 |
| 35 | | air pressure lubricating temperature | CCM004 | ℃ | 35 ± 5 |
| 36 | | pneumatic outlet pressure | CCM005 | MPa | 0.8-1.5 |
| 37 | | pneumatic middle liquid level | CCM006 | % | ≤40 |
| 38 | | pneumatic inlet liquid level | CCM007 | % | ≤20 |
| 39 | | gas turbine outlet temperature | CCM008 | ℃ | 510-540 |
| 40 | | main fan outlet pressure | CCM009 | MPa | 0.24 ± 0.02 |

### *4.3. Result*

After obtaining the necessary historical data, the following actions are performed:

1) Set indicator CCR001 as the data sampling frequency benchmark, for which the data collection interval is $T_1 = 30$s, and establish data sampling frequency standardization for all of the other indicators according to formula 4.
2) Create the time alignment matrix, define indicator CCR001 as the time axis benchmark, and resolve the time lag issue according to formulas 5-7.
3) Analyze the internal relations between the indicators and select features with the correlation analysis method.

PCC is used to generate a correlation matrix map as a visible method to analyze the high dimension data, as shown in Fig. 8. Red indicates a positive correlation between two sets of data, blue indicates a negative correlation, and a darker color indicates a stronger correlation. With this intuitive matrix drawing, we focus on indicators that are strongly related to the final yield target and delete some redundant indicators that have a high degree of correlation with the aim of retaining effective indicators, controlling data dimension, and improving model training efficiency. Finally, we reduce the data dimensions from 410 to 100 to improve the machine learning efficiency.

1) Model training

Four algorithms, including random forest, AdaBoost, XGBoost, and LightGBM, are practiced to train the simulation model separately with historical data. Considering that the production environment and feeding materials were continuously changing and that, in a short period, the state was relatively stable, three time points are randomly selected as the reference points for training, and each algorithm is tested three times to fully compare the effect differences among the algorithms. The interval between each pair of time points is approximately 20 days. The data for 2 months ahead of each time point are the training group, and the data for 15 days after are the validation group.

To evaluate the prediction effect intuitively, FE is calculated in formula 11, and the result shown in Fig. 9 indicates that LightGBM is better than the other algorithms in terms of prediction effect.

To comprehensively evaluate the model quality, four evaluation criteria are selected: MAR, RMSE, VIR, and PCC, which were calculated using formulas 12-15, respectively. The results shown in Table 4 indicate that the model trained by the LightGBM algorithm has advantages over the other algorithms in predicting accuracy and comprehensive performance, so we choose LightGBM as the basic training algorithm of the “digital twin” model.

1) Model online deployment

For security and practical reasons, the machine learning model,

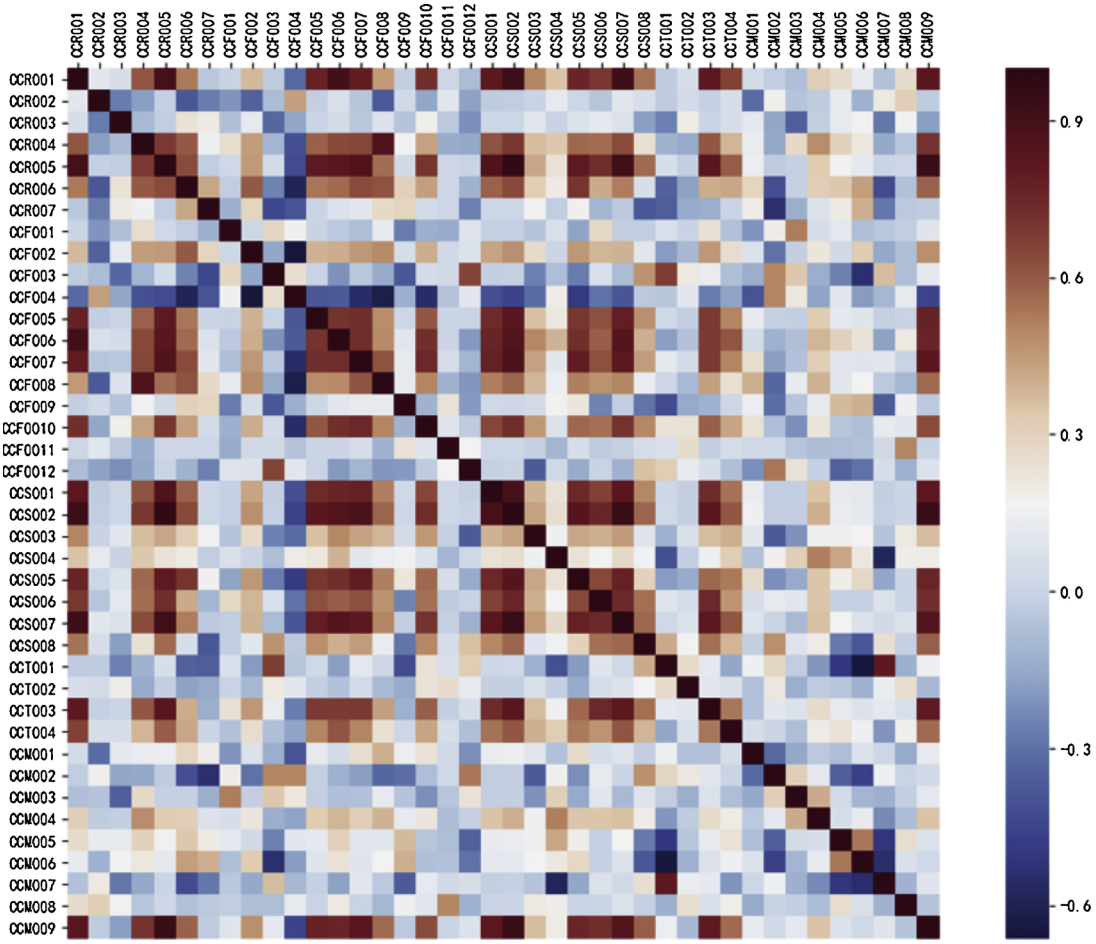


**Fig. 8.** Correlation matrix map of part catalytic cracking unit indicators.

integrated online with the MES, is deployed, and output from the digital twin, as real-time recommended control information, is viewed by operators first before being utilized in the production system.

To provide easy practice for the operators, in this pilot project, we select the 5 most important indicators for control recommendations. In the model training process, the LightGBM algorithm generates the importance index, which refers to an indicator that results in an increase in the total information gain of the model, and in the regression case, the information gain represents the reduction of a representation loss, such as the reduction of mean square error. Assuming that the $X_\alpha$ feature is used n times in the process of constructing the model and that the gain of the first order is $g_i$, then the importance $I_\alpha$ can be expressed as in formula 16.

$$I_\alpha = \sum_{i=1}^{n} g_i \tag{16}$$

Finally, the five most important controllable indicators are selected as the real-time control variable for operators, as shown in Table 5.

To validate the practical effect after the digital twin model was deployed online, three groups of experiments are conducted. Each group was in a period of relatively stable crude oil types and production environments; the first experiment produced for two weeks using traditional production control methods and then produced for two weeks with the recommended control information from the digital twin to compare the yield of light oil before and after implementing the new method. The experimental results (Table 6) show that the new method can effectively increase the yield of light oil by 0.2% and 0.5%. An independent-samples t-test was used to test the significance of these increases, and the results showed that all of these three feeding types had significant increases after implementing the new method. Fig. 10 shows the comparison of production performances before and after optimizing the technological parameters under the same production environment setting.

The results of the experiments prove that, when the yields of specific product are set as the goals for machine learning, under the same production circumstances, the digital twin-based model training approach and feedback mechanism can effectively optimize production control. This finding has theoretical and practical implications for the petrochemical and other process manufacturing industries in terms of achieving agile and effective production control.

## 5. Discussion

Existing literatures have identified the theoretical and practical values of machine learning and digital twin in manufacturing industry. By utilizing machine learning and digital twin technologies, manufacturing factories can achieve accurate and agile production control in response to the changes in market demand. These technologies also help to reduce the cost of inefficient production, improve the economic benefits and enhance sustainable development capabilities of SMEs. However, investigations on the theory, approach, process and guidelines of implementation of machine learning and digital twin in manufacturing industry are still research gaps. There are few methods that can achieve fast and effective interaction between virtual models and real environments. Current data processing methods are isolated and fragmented, and the existing mathematical programming-based methods for short-term scheduling of the refinery industry are not practical in application.

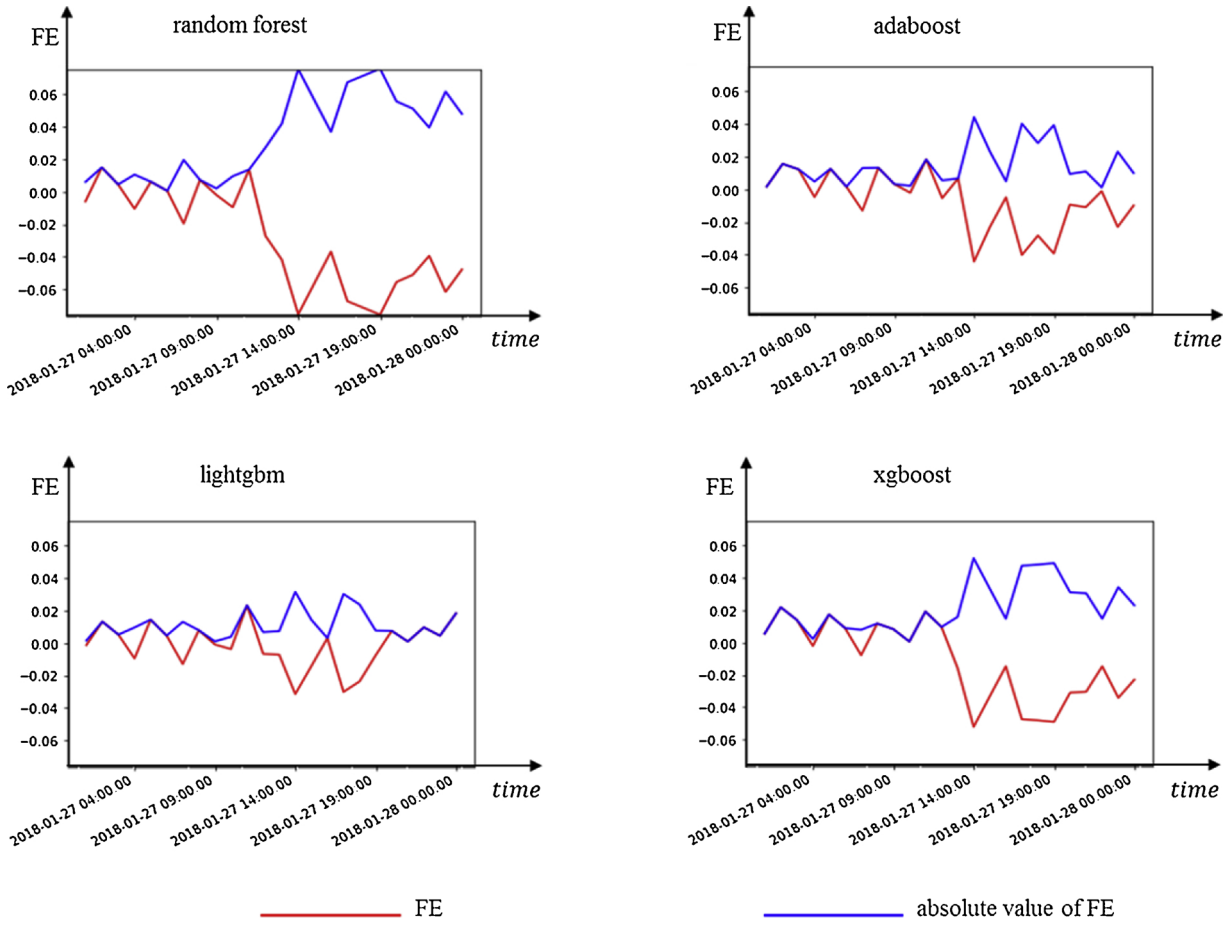


**Fig. 9.** Fitting error curves of different machine learning algorithms.

This paper fills these research gaps by proposing a digital twin framework for production control based on IoT and machine learning, including the components' architecture, basic steps of the approach, and key evaluation indices. Furthermore, a case study is also conducted in a real petrochemical factory to examine the effectiveness and value of the framework and approaches. Research contributions and practical implications are discussed in this section.

### 5.1. Research Contributions

This paper contributes to the literatures on machine learning, digital twin and production optimization, by proposing a theoretical framework for digital twin-based production control. This novel framework helps to reduce the dependency of production management decision making on expert experience and domain knowledge, and avoid the excessive influences of single machine learning results. This paper introduces guidelines on the usage of industrial big data, and provides methodological insights on building digital twin with using machine learning, training digital twin models based on industrial big data, selection of evaluation indices and digital twin model refinement.

With the rise of CPS and digital twin, increasing attention has been paid to the interaction between the physical world and the virtual world. The results of this paper provide guidelines for fast and effective interactions between virtual models and real industry environments.

**Table 4**
The model quality evaluation result (year: 2018)

| | Random forest | | | AdaBoost | | |
|---|---|---|---|---|---|---|
| Time points | 18.Jan. | 8.Feb. | 1.Mar. | 18.Jan. | 8.Feb. | 1.Mar. |
| MAR | 97.15% | 97.90% | 98.55% | 97.53% | 97.94% | 98.22% |
| RMSE | 0.0180 | 0.0122 | 0.0087 | 0.0146 | 0.0123 | 0.0102 |
| VIR | 0.4403 | 0.4316 | 0.8703 | 0.6319 | 0.4219 | 0.8221 |
| PCC | 0.6387 | 0.7466 | 0.8323 | 0.5641 | 0.7486 | 0.7573 |
| | LightGBM | | | XGBoost | | |
| Time points | 18.Jan. | 8.Feb. | 1.Mar. | 18.Jan. | 8.Feb. | 1.Mar. |
| MAR | 98.11% | 98.20% | 98.65% | 98.06% | 98.18% | 98.36% |
| RMSE | 0.0111 | 0.0104 | 0.0077 | 0.0111 | 0.0108 | 0.0091 |
| VIR | 0.7872 | 0.5854 | 0.8975 | 0.7879 | 0.5546 | 0.8564 |
| PCC | 0.7737 | 0.7700 | 0.8497 | 0.7500 | 0.7660 | 0.7942 |

**Table 5**
The control points of the catalytic cracking unit in MAYA

| No. | indicator | code | $I_\alpha$ |
|---|---|---|---|
| 1 | riser outlet temperature | CCR001 | 3373.97 |
| 2 | liquid temperature of fractionator | CCF006 | 3756.94 |
| 3 | stabilized tower bottom temperature | CCS005 | 322.59 |
| 4 | settler pressure | CCR004 | 186.36 |
| 5 | regenerator pressure | CCR005 | 104.77 |

The framework and approach for constructing a digital twin based on the petrochemical industrial IoT and machine learning complements the literatures. Approaches proposed in this paper, including model development, data processing, and model training, contribute to the methodological research on the applications of machine learning and digital twin. The effectiveness of these approaches is also proved by applying to a real case from the petrochemical industry. The modeling processes of digital twins proposed in this paper based on the industrial IoT and machine learning also offer valuable insights for other types of manufacturing enterprises.

### 5.2. Implications for Practice

From the above, it can be concluded that constructing a digital twin between the physical and cyber worlds is important for production simulation and control. Developing a formal approach on the implementation of machine learning and IoT techniques in the production control process in petrochemical manufacturing industry provides both theoretical and practical value. In the era of Big Data, there are challenges when applying artificial intelligence for decision making (Duan, Edwards, & Dwivedi, 2019). This paper addresses the challenges by introducing a digital twin-based framework which integrates machine learning and IoT techniques. This framework brings implications for achieving intelligent decision making in smart factory management with applying machine learning and digital twin.

In the context of process manufacturing industry, the features of raw materials and environment parameters are not stable during production. Using digital twin based framework, factories can quickly response to the changes, and keep an optimized output capability under the continuously changing external constraints. The implications from this research are not only on controlling production indices including temperature, pressure et al. in petrochemical industry, but also on the optimization of production control through real time data analysis in other industries. For instance, in metallurgy industry, digital twin models based on the real time data including product purity and production control parameters can help to optimize the boiler temperature and filling speed; in food processing industry, digital twin models based the real time data including quality indices and temperature can support to control quantity of additives, fermentation time et al.

In the context of discrete manufacturing industry, real-time online automatic quality detector, numerical machine bench with PLC and CNC, and new RFID technology that can integrate cutter tools, all provide sufficient industrial big data. Therefore, the same method can be used for reference to build a digital twin based system for optimizing production control. For example, in the precision machining industry, real-time data such as process precision, tool wear degree, shaft running stability and other real-time data construct digital twins, can be used to real-time optimization of cutting machine bench; in the chip production industry, the process parameters of each procedure and the screening criteria of intermediate quality control can be optimized in real time through the digital twin modeling, based on real-time data including quality grade, slicing, lithography and degumming.

Big open linked data (BOLD) can be utilized to build interpretive structural models (Dwivedi et al., 2017), not only in the production control in one firm, but also in the interactions among upstream and

**Table 6**
Experimental result for the yield of light oil (year: 2018)

| No | feeding type | average daily throughput (before)/ton | average light oil yield (before) /ton | yield ratio (before) | average daily throughput (after) /ton | average light oil yield (after) /ton | yield ratio (after) | yield increased | t-value | Sig |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Refined wax oil | 4209.65 | 2031.16 | 48.25% | 4198.73 | 2047.72 | 48.77% | 0.52% | −9.152 | $p < 0.001$ |
| 2 | Refined wax oil + tail oil | 4285.32 | 1978.96 | 46.18% | 4273.26 | 1981.08 | 46.36% | 0.18% | −4.692 | $p < 0.001$ |
| 3 | Refined wax oil + condensate oil | 4197.19 | 1964.70 | 46.81% | 4201.55 | 1980.19 | 47.13% | 0.32% | −7.13 | $p < 0.001$ |

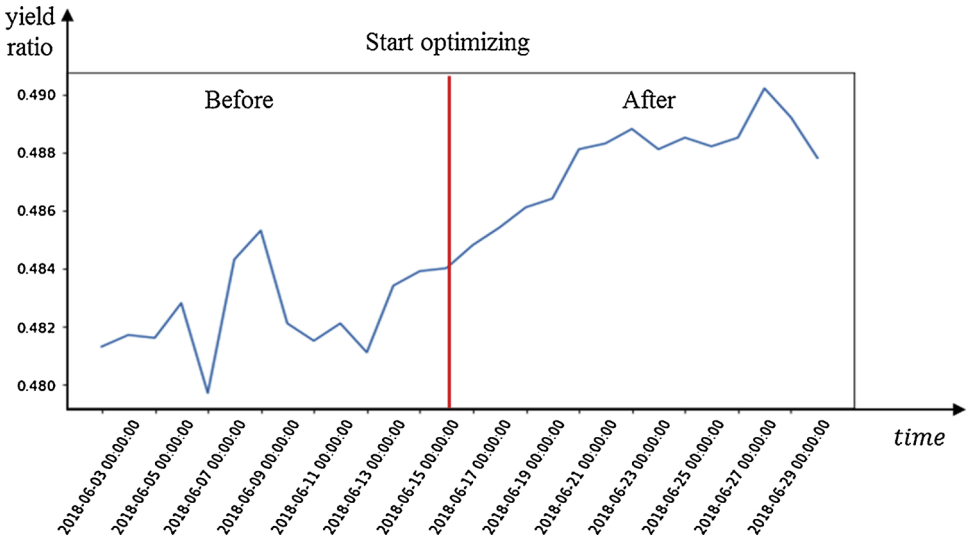


**Fig. 10.** Comparing before and after refining the wax oil feeding parameters.

downstream firms. The machine learning based digital twin modeling approach provides a practical application scheme. This novel and intelligent scheme can work in supply chain management, energy management, safety management, equipment management and many other related domains in factory. For example, forecasting raw material demand through machine learning-based digital twin, optimizing in-and-out logistics, and optimizing inventory; finding the correlation between different process points, predicting the output configuration and proportion of products, and optimizing the output distribution of products; forecasting the energy consumption and optimizing the consumption of water, electricity and other public resources; predicting human behavior and make risk forecasting for safety of people; assessing environmental performance and providing risk warning; simulating and evaluating the changing state of equipment and pipeline, predicting the degree of equipment wear and pipeline corrosion, and analyzing the root cause of failure problems.

This machine learning based digital twin framework also possesses potential value for the smart city issues. With the support of digital twin framework, smart cities can help to improve the quality of life for its citizens, the local economy, transport, traffic management, environment, and interaction with government (Ismagilova, Hughes, Dwivedi, & Raman, 2019). The digital twin framework integrating IoT big data and machine learning, will help the government and policymakers for eradicating the potential interferences in smart city development initiatives (Rana et al., 2018). The digital twin models trained by utilizing historical data, along with the real time data from IoT, can provide better solution on the issues including traffic control, public transportation and allocation of public resources.

## 6. Conclusions

In this paper, a digital twin framework for petrochemical production control optimization is proposed based on the industrial IoT and machine learning. The architecture, practice loop, machine learning methods, and key evaluation indices are proposed. The approach can be considered a concrete solution for the specifics of the petrochemical industry environment. This paper also provides approaches to time series data processing issues in digital twin modeling, such as frequency unification, time lag issues, and the demand for immediacy. Finally, the digital twin framework and approach were implemented and evaluated in the catalytic cracking unit of a petrochemical factory. The results prove the effectiveness of this framework and approach for production control optimization.

### 6.1. Limitations

This research is still with some limitations, which can be summarized as following:

First, the optimum yield of the unit does not represent the best overall economic benefit but must also consider the effects of energy consumption, market changes, and other lower yields. Therefore, a more complex decision model could be established to make instantaneous process control decisions based on consideration of overall economic benefit.

Second, as the digital twin practice loop was described, the digital twin modeling is not a one-time process but a repeating circle practice loop which is formed to continually control and optimize the petrochemical production line. The formation of a continually improving and self-adaptive digital twin mechanism requires more theoretical research and practical testing.

Third, this research focuses on the petrochemical industry; however, the digital twin framework for production control purposes could be widely used in the manufacturing industry. Finding the commonalities in different industries and creating a universal method for constructing digital twins for production control is a meaningful direction for extension.

### 6.2. Future Research Directions

In the future, the machine learning based digital twin modeling approach proposed by this paper can be extended in the following directions:

First, although this approach is proposed to address the production control problems in petrochemical manufacturing industry, it has the potential to be applied in other industries. This approach can be easily extended to support production management in other manufacturing industries, including process manufacturing and discrete manufacturing industries. Public service domains such as building “smart city” will also benefit from the proposed digital twin modeling approach.

Second, the selection of the evaluation indices and the evaluation method for digital twin models can be further improved. In the context of high dimensional data, the method determining the importance of evaluation indices of digital twin models still needs further study. The evaluation standard and method of digital twin models with combining the priori expert knowledge and machine learning results will be another important issue for future research.

Third, the approach can be further improved by developing better

method on the parameter setting. In the case with infinite data set, the optimized weight of data dimensions during digital twin modeling should be determined following a formal method. When the historical data cannot meet the requirements in the range of thresholds, the algorithm that simulates the digital twin model and predict result will be another important research issue.

## Acknowlegements

This work was partly supported by the National Natural Science Foundation of China [71772002, 71431002, 71421001, 71872033]; Natural Science Foundation of Liaoning Province [20180550433]; Philosophy and Social Science Planning Fund of Liaoning Province [L18CGL015].

## Reference

Al-Sharrah, G., Elkamel, A., & Almanssoor, A. (2010). Sustainability indicators for decision-making and optimisation in the process industry: The case of the petrochemical industry. *Chemical Engineering Science, 65*, 1452–1461.

Alidi, A. S. (1996). A multiobjective optimization model for the waste management of the petrochemical industry. *Applied Mathematical Modelling, 20*, 925–933.

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., et al. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. *Advanced engineering informatics, 30*, 500–521.

Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). *An overview of machine learning. Machine Learning, Volume I*, Elsevier3–23.

Cheng, Y., Chen, K., Sun, H., Zhang, Y., & Tao, F. (2018). Data and knowledge mining with big data towards smart production. *Journal of Industrial Information Integration, 9*, 1–13.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. *International Journal of Information Management, 48*, 63–71.

Dwivedi, Y. K., Janssen, M., Slade, E. L., Rana, N. P., Weerakkody, V., Millard, J., et al. (2017). Driving innovation through big open linked data (BOLD): Exploring antecedents using interpretive structural modelling. *Information Systems Frontiers, 19*(2), 197–212.

Esposito, C., Castiglione, A., Martini, B., & Choo, K.-K. R. (2016). Cloud manufacturing: security, privacy, and forensic concerns. *IEEE Cloud Computing, 3*, 16–22.

Esposito, C., Castiglione, A., Palmieri, F., Ficco, M., Dobre, C., Iordache, G. V., et al. (2018). Event-based sensor data exchange and fusion in the Internet of Things environments. *Journal of Parallel and Distributed Computing, 118*, 328–343.

Esposito, C., Castiglione, A., Pop, F., & Choo, K.-K. R. (2017). Challenges of connecting edge and cloud computing: a security and forensic perspective. *IEEE Cloud Computing*, 13–17.

Ferreira, F., Faria, J., Azevedo, A., & Marques, A. L. (2017). Product lifecycle management in knowledge intensive collaborative environments: An application to automotive industry. *International Journal of Information Management, 37*, 1474–1487.

Glaessgen, E., & Stargel, D. (2012). *The digital twin paradigm for future NASA and US Air Force vehicles. 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference 20th AIAA/ASME/AHS Adaptive Structures Conference 14th AIAA (pp. 1818)*.

Grieves, M. (2014). *Digital Twin: Manufacturing Excellence Through Virtual Factory Replication.* Florida Institute of Technology.

Hatziargyriou, N. (2001). *Machine Learning Applications to Power Systems.* Berlin Heidelberg: Springer.

Ismagilova, E., Hughes, L., Dwivedi, Y. K., & Raman, K. R. (2019). Smart cities: Advances in research—An information systems perspective. *International Journal of Information Management, 47*, 88–100.

Kateris, D., Moshou, D., Pantazi, X.-E., Gravalos, I., Sawalhi, N., & Loutridis, S. (2014). A machine learning approach for the condition monitoring of rotating machinery. *Journal of Mechanical Science and Technology, 28*, 61–71.

Köksal, G., Batmaz, İ., & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems with Applications, 38*, 13448–13467.

Li, D. (2016). Perspective for smart factory in petrochemical industry. *Computers & Chemical Engineering, 91*, 136–148.

Li, D., Jiang, B., Suo, H., & Guo, Y. (2015). Overview of smart factory studies in petrochemical industry. *Computer Aided Chemical Engineering, 37*, 71–76.

Li, J., Tao, F., Cheng, Y., & Zhao, L. (2015). Big data in product lifecycle management. *The International Journal of Advanced Manufacturing Technology, 81*, 667–684.

Lim, C., Kim, K.-H., Kim, M.-J., Heo, J.-Y., Kim, K.-J., & Maglio, P. P. (2018). From data to value: A nine-factor framework for data-based value creation in information-intensive services. *International Journal of Information Management, 39*, 121–135.

Liu, Z., Fan, S., Wang, H. J., & Zhao, J. L. (2017). Enabling effective workflow model reuse: A data-centric approach. *Decision Support Systems, 93*, 11–25. https://doi.org/10.1016/j.dss.2016.09.002.

Liu, Q., Zhang, H., Leng, J., & Chen, X. (2018). Digital twin-driven rapid individualised designing of automated flow-shop manufacturing system. *International Journal of Production Research,* 1–17.

Lu, Y., Min, Q., Liu, Z., & Wang, Y. (2019). An IoT-enabled simulation approach for process planning and analysis: a case from engine re-manufacturing industry. *International Journal of Computer Integrated Manufacturing,* 1–17. https://doi.org/10.1080/0951192X.2019.1571237.

Mamonov, S., & Triantoro, T. M. (2018). The strategic value of data resources in emergent industries. *International Journal of Information Management, 39*, 146–155.

Miller, A. M., Alvarez, R., & Hartman, N. (2018). Towards an extended model-based definition for the digital twin. *Computer-Aided Design and Applications,* 1–12.

Monostori, L. (2003). AI and machine learning techniques for managing complexity, changes and uncertainties in manufacturing. *Engineering applications of artificial intelligence, 16*, 277–291.

Nai-Qi, W. U., & Bai, L. P. (2005). Scheduling optimization in petroleum refining industry: a survey. *Computer Integrated Manufacturing Systems, 11*, 90–96.

Pach, C., Berger, T., Bonte, T., & Trentesaux, D. (2014). ORCA-FMS: a dynamic architecture for the optimized and reactive control of flexible manufacturing scheduling. *Computers in Industry, 65*, 706–720.

Pham, D. T., Packianather, M. S., Dimov, S. S., Soroka, A., Gerard, T., Bigot, S., et al. (2004). *An application of datamining and machine learning techniques in the metal industry*.

Rana, N. P., Luthra, S., Mangla, S. K., Islam, R., Roderick, S., & Dwivedi, Y. K. (2018). Barriers to the development of smart cities in Indian context. *Information Systems Frontiers,* 1–23. https://doi.org/10.1007/s10796-018-9873-4.

Rana, R., Staron, M., Hansson, J., Nilsson, M., & Meding, W. (2014). A framework for adoption of machine learning in industry for software defect prediction. *In Software Engineering and Applications (ICSOFT-EA), 2014 9th International Conference on (pp. 383-392)*.

Rehman, M. H. U., Chang, V., Batool, A., & Wah, T. Y. (2016). Big data reduction framework for value creation in sustainable enterprises. *International Journal of Information Management, 36*, 917–928.

Restivo, F. (2006). *ADACOR: A holonic architecture for agile and adaptive manufacturing control.* Elsevier Science Publishers B. V.

Santos, M. Y., e Sá, J. O., Andrade, C., Lima, F. V., Costa, E., Costa, C., et al. (2017). A Big Data system supporting Bosch Braga Industry 4.0 strategy. *International Journal of Information Management, 37*, 750–760.

Saputelli, L., Nikolaou, M., & Economides, M. J. (2006). Real-time reservoir management: A multiscale adaptive optimization and control approach. *Computational Geosciences, 10*, 61–96.

Schleich, B., Anwer, N., Mathieu, L., & Wartzack, S. (2017). Shaping the digital twin for design and production engineering. *CIRP Annals, 66*, 141–144.

Tao, F., Cheng, J., Cheng, Y., Gu, S., Zheng, T., & Yang, H. (2017). SDMSim: a manufacturing service supply–demand matching simulator under cloud environment. *Robotics and computer-integrated manufacturing, 45*, 34–46.

Tao, F., Cheng, J., Qi, Q., Zhang, M., Zhang, H., & Sui, F. (2018). Digital twin-driven product design, manufacturing and service with big data. *The International Journal of Advanced Manufacturing Technology, 94*, 3563–3576.

Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems*.

Tao, F., Sui, F., Liu, A., Qi, Q., Zhang, M., Song, B., et al. (2018). Digital twin-driven product design framework. *International Journal of Production Research,* 1–19.

Tellaeche, A., & Arana, R. (2013). Machine learning algorithms for quality control in plastic molding industry. *Emerging Technologies & Factory Automation (ETFA), 2013 IEEE 18th Conference on (pp. 1-4)*.

Tuegel, E. (2013). *The Airframe Digital Twin: Some Challenges to Realization. Aiaa/asme/asce/ahs/asc Structures, Structural Dynamics and Materials Conference Aiaa/asme/ahs Adaptive Structures Conference Aiaa*.

Tuegel, E. J., Ingraffea, A. R., Eason, T. G., & Spottswood, S. M. (2011). Reengineering aircraft structural life prediction using a digital twin. *International Journal of Aerospace Engineering, 2011*.

Wen, J., Li, S., Lin, Z., Hu, Y., & Huang, C. (2012). Systematic literature review of machine learning based software development effort estimation models. *Information and Software Technology, 54*, 41–59.

Wu, K.-J., Liao, C.-J., Tseng, M.-L., Lim, M. K., Hu, J., & Tan, K. (2017). Toward sustainability: using big data to explore the decisive attributes of supply chain risks and uncertainties. *Journal of Cleaner Production, 142*, 663–676.

Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., et al. (2016). Big data: From beginning to future. *International Journal of Information Management, 36*, 1231–1247.

Yuan, Z., Qin, W., & Zhao, J. (2017). Smart Manufacturing for the Oil Refining and Petrochemical Industry. *Engineering, 3*, 179–182.

Zhang, H. (2004). The Application of Data Mining in Petrochemical Enterprise. *Computer Engineering & Applications, 40*, 208–210.

Zhang, Y., Ren, S., Liu, Y., & Si, S. (2017). A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. *Journal of Cleaner Production, 142*, 626–641.

**Qingfei Min** is Professor of information systems at School of Economics and Management, Dalian University of Technology, China. His research interests include IT/IS implementation and adoption, e-commerce / m-commerce behavior and strategies, global virtual team and social medias. He received his Ph.D. in Information Systems from Dalian University of Technology. He has published several studies in Information & Management, International Journal of Mobile Commerce, Computers in Human Behavior as well as in some Chinese academic journals and international conferences.

**Yangguang Lu** is a Ph.D. candidate of Management Science and Engineering at the School of Economics and Management, Dalian University of Technology, China. His

research focuses on digital factory, simulation, and big data utilization in manufacturing industry. He has over 12 years of digital twin / intelligent manufacturing / simulation and IT planning project experience in the automobile industry and petrochemical industry.

**Zhiyong Liu** is Associate Professor of information systems at School of Economics and Management, Dalian University of Technology, China. His research interests include business process modeling, workflow technology, cross-border e-commerce, blockchain technology. He received his Ph.D. in Information Systems from a joint program of City University of Hong Kong and University of Science and Technology of China. He has published several studies in Decision Support Systems, Electronic Commerce Research, Journal of Information Science as well as some leading international conferences.

**Chao Su** is a Ph.D. candidate of Management Science and Engineering at School of Economics and Management, Dalian University of Technology, China. His research focuses on information system adoption and social commerce.

**Bo Wang** is data mining senior specialist of Big Data & IoT Business Development Unit, Lenovo Capital & Incubator Group, China. His research focuses on industrial big data utilization and IoT technology.