SCIEPublish

Article

Knowledge-data Collaborated Digital Twin Model of Papermaking Process

Zejun Liu, Mengna Hong and Jigeng Li *

State Key Laboratory of Pulp and Paper Engineering, South China University of Technology, Guangzhou 510640, China; msliuzejun@mail.scut.edu.cn (Z.L.); femnhong@scut.edu.cn (M.H.)

* Corresponding author. E-mail: jigengli@scut.edu.cn (J.L.)

Received: 13 December 2023; Accepted: 23 January 2024; Available online: 27 February 2024

ABSTRACT: The structure of the drying section in papermaking process is complex and too compacted to install sensors. In order to monitor the parameters in dynamic and manage the process practically with virtual simulations instead of physical experiments, a digital twin-based process parameter visualization model is constructed in this study. Regarding to the possible missing data in the modeling framework, it is proposed to combine industrial data, and knowledge of mechanism with intelligent algorithms to fill in the missing parameters. Upon which, a digital twin-based data visualization model is established using CADSIM Plus simulation software. Both of the knowledge -based mechanism solution model and the random forest-based parametric prediction model perform well, and the predicted parameters can support the digital twin visualization model in CADSIM Plus. Visual modeling of surface condenser in the paper drying section was realized for example, and results show that the model is capable of monitoring the dynamic changes of parameters in real time, so as to support the optimization and decision making of papermaking process such as formation, drying, et al.

Keywords: Digital twin; Model; Papermaking; Parameter prediction; Simulation

1. Introduction

The papermaking process involves a large number of complex physical and chemical reactions [1], accompanied with characteristics of multi-variable, strong coupling and non-linear etc., hindering the intelligent development process of the papermaking industry [2]. At present, the process models and control systems, established by paper-making enterprises based on a new generation of information technology, have not yet solved the problem of "data silos" and cannot integrate material flow and energy flow information to achieve real-time monitoring and control of paper production, which affects the sustainability of the process dramatically.

With the development of Industry 4.0 era [3,4], digital twin is gradually being studied and applied in the process industry [5]. In the steel industry, with the applications of technologies such as intelligent data sensing, multi-source heterogeneous data integration, efficient data transmission, digital twin creation, enhanced interaction, and conversion applications, a production line that combines reality with virtuality can be established to realize the optimization of the production process [6]. In the machine building industry, simulation and optimization based on the digital twin's dynamic perception of the physical machine tool, it can be effectively optimized machining conditions such as cutting parameters and reduce carbon emissions [7]. It is worth noting that, the establishment of a digital twin model of the process industry, can facilitate the simulation, analysis, monitoring and optimization of manufacturing processes in real time, turns out the dynamic management without physical efforts [8–11].

The papermaking industry is a typical process-oriented industry [12,13], and the information technologies such as big data and machine learning have been widely used in its energy-saving renovation [14,15], modeling [16], scheduling [17,18], fault prediction [19], decision-making support [20], and process optimization [21–23] sectors. However, most of the previous studies only focus on a single process or a single equipment of it, without considering the modeling and control of the whole process systematically, which tends to achieve local optimizations. Therefore,

https://doi.org/10.35534/amsm.2024.10003

there is an urgent need to develop the digital twin model for papermaking process that can perform online sensing, analysis, simulation, optimization, and decision-making of the process. The emerging of digital twin in the industrial domain provides feasible solution ideas [24–26]. As aforementioned, although a large number of scholars have conducted modeling studies on digital twins and achieved some results in related fields [27–29], it remains empty of a robotic general digital twin frameworks for the chemical industry, especially for the papermaking industry. Therefore, this paper proposes a digital twin-based data visualization model based on CADSIM Plus simulation software for the papermaking industry to promote the process performance.

2. Methods

2.1. A Digital Twin Framework for the Papermaking Process

The papermaking process has the following characteristics: diverse raw materials with different nature and complex interrelationships in the process, and there are a large number of parameters that are difficult to be detected along the process; various complicated physical and chemical reactions occur in its process [30].

The core of designing a digital twin of a papermaking process is to perform mechanistic or data-driven modeling of the process to rationally describe the production process [31–33]. Industrial data are first collected, pre-processed, and stored in a database. Then data-driven methods are used to infer parameters that cannot be directly measured, find relationships in the production process, and simulate the process to enable assessment of the current state, diagnosis of past problems, prediction of future trends, provide more reasonable and reliable process parameters and optimize the process g process [34,35]. Figure 1 illustrates the digital twin framework for the papermaking process, and the individual modules are described correspondingly.

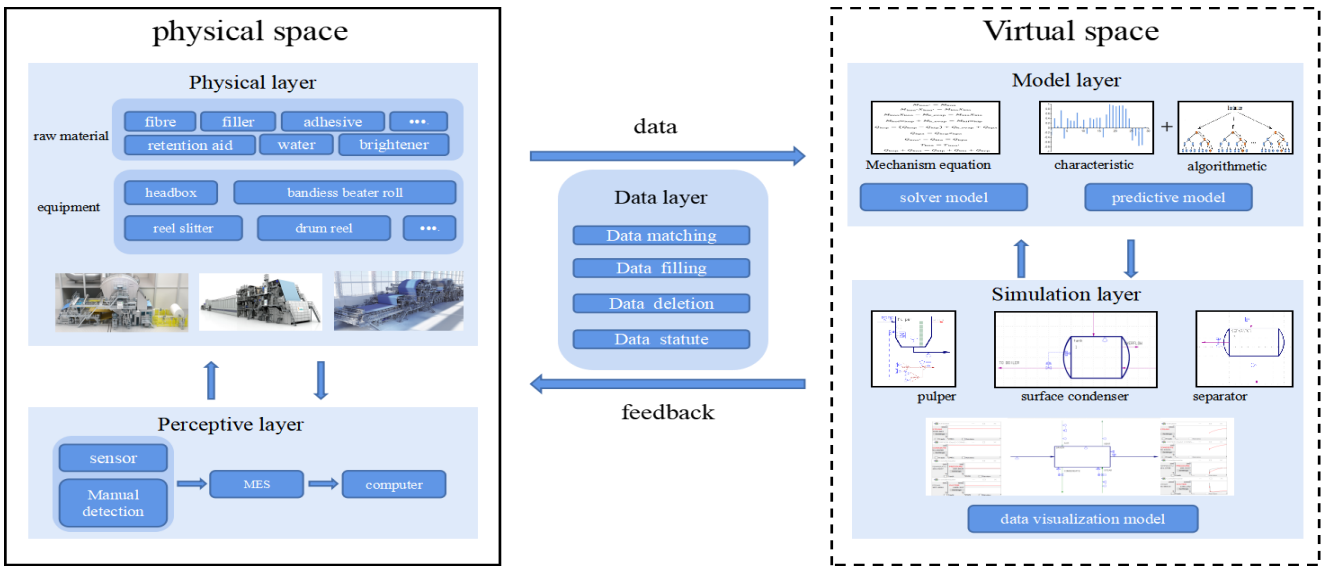

Figure 1. Digital twin framework of papermaking process.

The physical layer mainly includes materials, equipment, production lines, environment and personnel, etc. It mainly provides data on all production factors and production elements such as materials, equipment and processes in the papermaking process. Materials such as fibers, fillers, retention aids, dry strength agents, etc., equipment such as pulper, headbox, white part, core winder and cutter etc., production lines are also divided into different production lines such as newsprint, household paper, boxboard, etc., as well as the temperature and humidity of the environment, personnel operation, etc.

The sensing layer is mainly responsible for data collection and transmission. The original data in the process is collected by the sensors on site to the distributed control system, and then uploaded to the enterprise's cloud service platform through the net gate, and the data can be downloaded directly through the cloud service platform when used. The data collection also includes the data of offline inspection such as the quality index of raw paper, etc.

The data layer is responsible for receiving the original data. Due to the different frequency of data collection, spatial location of sensors, etc., as well as the fact that some of the data are obtained by offline laboratory tests, and there are many uncertainties in the process, such as process abnormalities, equipment failures, or irregularities in operation, there are time differences and delays in the collected data. Therefore, after operations such as data

matching, data filling, and outlier deletion, various information needs to be stored in the database, which can later be used as input for the construction of the digital twin model in virtual space.

The model layer is the digital modeling of physical entities. The model layer includes a solver model for the process parameters, a prediction model, and a simulation model of the physical entity [36]. Based on the principle of mass and heat transfer, a parameter prediction model is built by analyzing the data generated in the process and simulating and calculating the production process to visualize the process parameters and process monitoring.

The application layer includes functions such as process optimization, parameter optimization, operation guidance, testing and diagnosis, and inventory calculation [37,38]. Finally, it is fed back to the physical entity to realize the control management and closed-loop optimization of the whole process. Based on the massive data, it can be provided more reasonable and reliable process parameters and optimize the process through prediction and optimization models to achieve the evaluation of the current state, or the prediction of future trends.

2.2. Research on Digital Twin-based Modeling Techniques

The above research provides a method for constructing a digital twin framework for the papermaking process, but due to the complexity and variety of variables in the papermaking process, cost, and technical constraints that prevent access to certain key information on process variables, there is still a need to find suitable methods to address the problem that certain variables in the process are difficult to measure or cannot be measured directly by sensors.

Therefore, it is proposed a method for parameter solution and parameter prediction in which, when the process mechanism can be constructed, constraints are set on the unknown parameters, and the missing parameters are solved by combining the mechanism equation with a multi-objective optimization algorithm; in the case where the process mechanism is difficult to construct, the parameters are predicted by using a data-driven model by screening the characteristic variables through correlation analysis. Figure 2 shows the technical route of the method. It starts from the preprocessing treatment of the raw data with data matching, outlier deletion, and data filling. The derived data could be used for parameter prediction and solution obtaining. When there are not mechanistic equations about the process, LR, GBR, and RF models would be used to present the unclear interrelationships of variables, and turning the output with data. Whereas, when there are equations, the outputs would be calculated in different scenarios based on the availabilities of parameters. Specifically, NLP and NSGA-II were used to find the optimal parameter setting when it is not fully available.

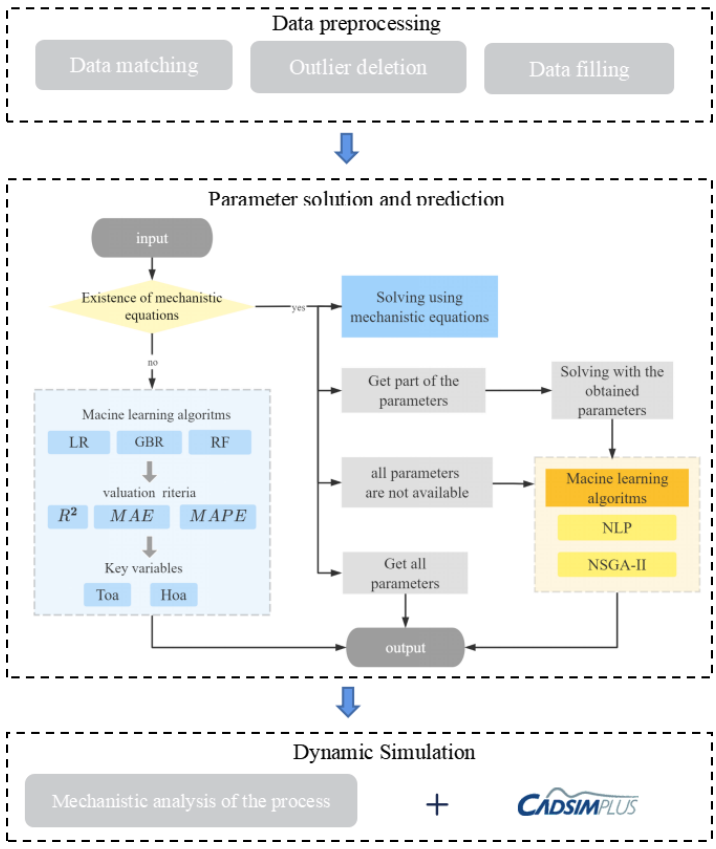

Figure 2. Technical route of the constructed digital twin papermaking model.

2.2.1. Data Processing

The raw data required for this paper is collected from the mill's distributed control system, which collects process data such as Vehicle speed, steam pressure, temperature, and humidity for four quantifications, as shown in Table 1.

The data collected had null values, missing values or data redundancy, and some of the data were collected at different frequencies, making modeling more complex [39]. To address the above characteristics, the originally collected data were matched in time series, with most of the data being sampled in 60 s, 10 min and 20 min sets, and the collected data were matched in 5 min sets to take the average value within the sampling interval. Missing values are filled by averaging and interpolation. Outliers are removed using box plots.

Paper production can be divided into working conditions based on basic weight, as shown in Table 2. In this paper, four basic weights are considered as four working conditions are modeled separately to compare the predictive effect of the model under different working conditions.

Table 1. Modeling basic data table.

Variables		
R_s	Reel speed	$m \cdot min^{-1}$
P_s	Pressure of steam	kPa
P_h	Pressure of headbox	kPa
BW_s	Basis weight before drying sizing	$g \cdot m^{-2}$
M_{dz}	Moisture before drying sizing	%
M_p	Moisture content of paper	%
C_p	Pulp consistency of top wire	%
T_{ia}	Temperature of entering air	℃
H_{ia}	Humidity of entering air	%
T_{oa}	Temperature of exhaust air	℃
H_{oa}	Humidity of exhaust air	%
W_{oa}	Weight of exhaust air	$kgd \cdot a \cdot s^{-1}$
F_s	Flow of steam into cylinder	kg/s
P_{sc}	Pressure of steam into cylinder	kPa
T_s	Temperature of steam into cylinder	℃
BW	Basics weight	$g \cdot m^{-2}$
V_f	Former Vacuum	kg
V_{cr}	Couch Roll Vacuum	kpa
V_{pr}	Press Roll Vacuum	kpa

Table 2. Working conditions and basic weight comparison.

Condition	1	2	3	4
Basic weight ($g \cdot m^{-2}$)	115~130	105~115	80~95	65~80

2.2.2. Parametric Solver Model Based on Digital Twin

For processes where the mechanistic equations have been constructed, the solve function of the math module in Python is used to calculate the unknown solution. If only a partial solution can be found, the found parameters can be used together with the known parameters as known parameters, and the other parameters can be assigned values until the error in each mechanistic equation is minimized, and when the solution is stable, the value at this point can be considered to be the value of the sought parameter. Since there are multiple equations in which the computational error needs to be minimized simultaneously, this problem can be transformed into a multi-objective, multi-constraint optimization problem. Classical nonlinear programming (Sequential Least Squares Programming, SLSQP) and NSGA-II are used in this paper [16].

SLSQP [23] is one of the nonlinear planning algorithms for constrained problems, which can solve iterative methods for nonlinear optimization problems with constraints, such as boundary constraints, equation constraints, and inequality constraints [40]. Its method has the following characteristics: it can handle any degree of nonlinearity, including nonlinearity in constraints, and it must choose appropriate parameters such as initial solution, shift acceptance criterion, and step control strategy to ensure convergence and efficiency.

NSGA-II is an improvement on the first generation of non-dominated sorting genetic algorithm: it proposes a fast non-dominated sorting algorithm [41–43], which reduces the computational complexity on the one hand, and merges the parent population with the offspring population on the other hand, retaining all the best individuals; it introduces an elite

strategy to ensure that some good population individuals will not be discarded in the evolutionary process, and improves the accuracy of the optimization results; it adopts the crowding degree and crowding degree comparison operators as the criteria for comparison among individuals in the quasi-Pareto domain to ensure that the diversity of individuals in the quasi-Pareto domain can be evenly extended to the whole Pareto domain. Adopting crowding degree and crowding degree comparison operator as the comparison criteria among individuals in the population, so that the individuals in the quasi-Pareto domain can be evenly extended to the whole Pareto domain and ensure the diversity of the population.

2.2.3. Parametric Prediction Model Based on Digital Twins

When it is difficult to construct a mechanistic model to find the relationship between these parameters, a predictive model based on machine learning is established. Taking the ventilation system of the drying department as an example, the exhaust air temperature and humidity is an important index that reflects the operational status of the drying department and the rationality of the process, and the mechanism of the process is difficult to construct, so the prediction models for its exhaust air temperature and humidity are established. In this paper, it is chosen the more classical linear regression (LR), gradient boost regression (GBR) and random forest (RF) methods [44]. The proposal of these algorithms involves the considerations of that GBR and RF are typical ensemble learning algorithms of bagging and boosting, respectively. And they good at generalization with learning from small data sets. On the other hand, LR is relatively easy to conducted with good robustness, which could be seen as a baseline in this study.

LR is a type of regression analysis that models the relationship between one or more independent and dependent variables using a least square function called a linear regression equation [44]. Linear regression models are very easy to understand and the results are very interpretable, which facilitates decision analysis, but for nonlinear data or polynomial regression with correlation between data features is difficult to model and difficult to represent highly complex data well. GBR use a continuous approach to constructing trees, with each tree trying to fix the errors of the previous tree [45]. By default, gradient boosted regression trees are not randomized, but use strong prepruning, and gradient boosting usually uses trees with very small depths. Such models have a small memory footprint and faster prediction speed [46]. The main idea of gradient boosting is to merge multiple simple models, where each tree makes good predictions for only part of the data, and the more trees added, the more iterations can be made to improve performance.

RF is an algorithm that integrates multiple trees through the idea of integrated learning, based on which the samples are trained and predicted [47,48]. When used as a regression algorithm, the output value of the sample prediction is weighted by the regression value of the decision trees that make up the random forest as the output. It can handle both discrete and continuous data, and operates with high efficiency and accuracy.

2.3. Visualization Model Based on Digital Twin

Most of the traditional simulation models of the papermaking process have only focused on a single section, without considering the whole process, or have only calculated a relatively small number of parameters; The visualization model of the papermaking process is only based on a mechanistic model, but does not consider processes that cannot be modelled mechanistically, and does not incorporate advanced information technology. The built simulation process can only be calculated in a specific order. In view of the above problems, this paper uses the chemical simulation software CADSIM Plus to build the simulation process.

CADSIM Plus is a chemical process simulation software that has built-in physical property system database and operation unit module library of papermaking. It consists of a lot of physical property methods and physical property parameters. It can be flexibly used all the arrangement and combination of these methods to develop process simulation. At the same time, based on the previous work, it can be applied the mechanism model combined with process data to build modules of various parts in the software and store them in the model library. In the future, when building different processes, it can be continued to reuse this module, which improves the simplicity of modeling and reduces the time required for modeling.

3. Results and Discussion

3.1. Solver Model

This paper takes the surface condenser of the drying process as the object of study. Firstly, the mechanistic process is analyzed: the secondary steam entering the surface condenser is discharged as condensate through heat exchange, and the incoming cold water is heated and discharged as hot water. Figure 3 is a schematic diagram of the surface condenser module, and Table 3 shows the input and output parameters of the surface condenser module.

In the ideal state, the mass balance and energy balance are Equations (1)–(3):

$$M_{is} = M_{oc} \tag{1}$$

$$M_{ow} = M_{iw} \tag{2}$$

$$Q_{is} + Q_{iw} = Q_{ow} + Q_{oc} \tag{3}$$

Among them, secondary steam, cold water, hot water and condensate heat calculation Equations (4)–(7), C is the specific heat capacity of water:

$$Q_{is} = M_{is} \times (28.054 \times \ln(P_{is}) + 2583.6) \tag{4}$$

$$Q_{ow} = M_{ow} \times T_{ow} \times C \tag{5}$$

$$Q_{iw} = M_{iw} \times T_{iw} \times C \tag{6}$$

$$Q_{oc} = M_{oc} \times T_{oc} \times C \tag{7}$$

After constructing the mechanistic equations, the first assumption is made for a specific case and the two methods are compared with different numbers of known parameters for the solution. The number of known parameters is assumed to be 5, 4, 3, 2, and 1. Each parameter has different combinations and the number of known parameters is decremented in the order of Table 3 until the number of known parameters is one. The idea of the solution is to output the results directly if the known parameters can be calculated exactly by the mechanism equation. If the missing parameters do not give the full result, the solution continues in other ways. When substituting the unknown parameters into the mechanism equation, if the error and error sum of each equation is relatively small, the parameters can be considered as acceptable calculated values at this time.

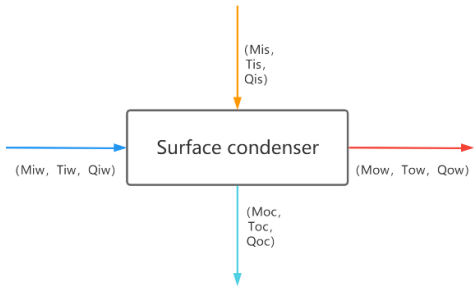

Figure 3. Surface condenser module.

Table 3. Input and output variables.

Input Variables		Output Variables	
M_{iw} (kg)	Mass of cold water,	M_{ow} (kg)	Mass of warm water
T_{iw} (°C)	Temperature of cold water	T_{ow} (°C)	Temperature of warm water
Q_{iw} (kJ)	Heat of cold water	Q_{ow} (kJ)	Heat of warm water
M_{is} (kg)	Mass of secondary steam	M_{oc} (kg)	Mass of condensated water
P_{is} (kpa)	Temperature of secondary steam	T_{oc} (°C)	Temperature of condensated water
Q_{is} (kJ)	Heat secondary steam	Q_{oc} (kJ)	Heat condensated water

As the relative error has the advantage of reflecting the magnitude and direction of the error and can more accurately reflect the level of confidence in the measurement, it is set in terms of relative error when setting the objective function over the error. The objective function is as follows: Equation (8).

$$\begin{cases} f1 = |(Mis - Moc)/Mis + (Mis - Moc)/Moc| \\ f2 = |(Mow - Miw)/Miw + (Mow - Miw)/Mow| \\ f3 = \left|\begin{matrix}(Qis - Mis \times (28.054 \times \mathrm{np.log}(Pis) + 2583.6))/Qis + \\ (Qis - Mis \times (28.054 \times \mathrm{np.log}(Pis) + 2583.6))/ \\ (Mis \times (28.054 \times \mathrm{np.log}(Pis) + 2583.6))\end{matrix}\right| \\ f4 = \left|\begin{matrix}(Qow - Mow \times Tow \times C)/Qow + \\ (Qow - Mow \times Tow \times C)/(Mow \times Tow \times C)\end{matrix}\right| \\ f5 = \left|\begin{matrix}(Qoc - Moc \times Toc \times C)/Qoc + \\ (Qoc - Moc \times Toc \times C)/(Moc \times Toc \times C)\end{matrix}\right| \\ f6 = \left|\begin{matrix}(Qiw - Miw \times Tiw \times C)/Qiw + \\ (Qiw - Miw \times Tiw \times C)/(Miw \times Tiw \times C)\end{matrix}\right| \\ f7 = \left|\begin{matrix}(Qis + Qiw - Qow - Qoc)/(Qis + Qiw) + \\ (Qis + Qiw - Qow - Qoc)/(Qow + Qoc)\end{matrix}\right| \end{cases} \tag{8}$$

The constraints are the upper and lower limits of the range of each parameter in the actual case. To obtain better results, the limits of the error function are added to the constraints, as in Equation (9):

$$\begin{cases} f1 < 0.05 \\ f2 < 0.05 \\ f3 < 0.05 \\ f4 < 0.05 \\ f5 < 0.05 \\ f6 < 0.05 \\ f7 < 0.05 \end{cases} \quad (9)$$

Table 4 shows a comparison table of the errors in solving the parameters under different methods with the number of known parameters of 5, 4, 3, 2 and 1. The errors of the values in the table are in the form of percentages. From the table it can be seen that the solution result of nonlinear programming is better than that of NSGA-II. Nonlinear programming needs to be assigned reasonable initial values before running the model, so it is already closer to reasonable values in the first run. But it is difficult to provide feasible initial solutions to determine upper and lower limits of parameters or when there are more parameters. So nonlinear programming is assumed already closer to the reasonable values in the first run.

When the upper and lower limits of parameters are not easy to determine or when there are more parameters to assign, it will not be possible to determine a suitable initial value. NSGA-II, on the other hand, uses random assignment of values in the first run, which can be applied to all cases. Normally, the more parameters are known, the more accurate the solution will be, whereas if there are only two or one known parameter values, the solution will no longer be credible. In this paper only five of these cases are listed, the rest are similar.

It can be concluded that the method can, to some extent, solve the problem of missing parameter values in the process of model calculation. Moreover, it is a relatively general method that can be used to solve the missing parameters in any process where the mechanistic equations can be established. When the range of parameters is known, the nonlinear programming solution is used; when the range of parameters is unknown and initial values are difficult to assign, NSAG-II is used to solve.

Table 4. Comparison of solution error results.

	5		**4**		**3**		**2**		**1**	
	NLP (%)	**NSGA-II (%)**	**NLP (%)**	**NSGA-II (%)**	**NLP (%)**	**NSGA-II (%)**	**NLP (%)**	**NSGA-II (%)**	**NLP (%)**	**NSGA-II (%)**
M_{is}	-	-	-	-	-	-	-	-	-	-
P_{is}	-	-	-	-	-	-	-	-	369	891
M_{iw}	-	-	-	-	-	-	2	47	2	61
T_{iw}	-	-	-	-	10	22	8	15	7	27
M_{oc}	-	-	0	0	0	0	0	0	0	0
T_{oc}	9	11	9	12	18	24	18	1	18	3
T_{ow}	1	1	1	1	7	13	5	22	3	31
M_{ow}	0	0	0	0	0	0	2	47	2	61

3.2. Prediction Model

Considering that the production process is a relatively complex, non-linear, highly coupled and uncertain process, a correlation analysis between the parameters of the collected characteristics is necessary to eliminate irrelevant variables. The Pearson correlation coefficients of each collected variable with the exhaust air temperature and humidity are calculated, and it is generally considered that a correlation coefficient with an absolute value greater than 0.4 represents a moderate correlation. In this study, the characteristic variables with correlation coefficients greater than 0.4 were selected as inputs to the model. Figure 4 shows the results of the correlation analysis.

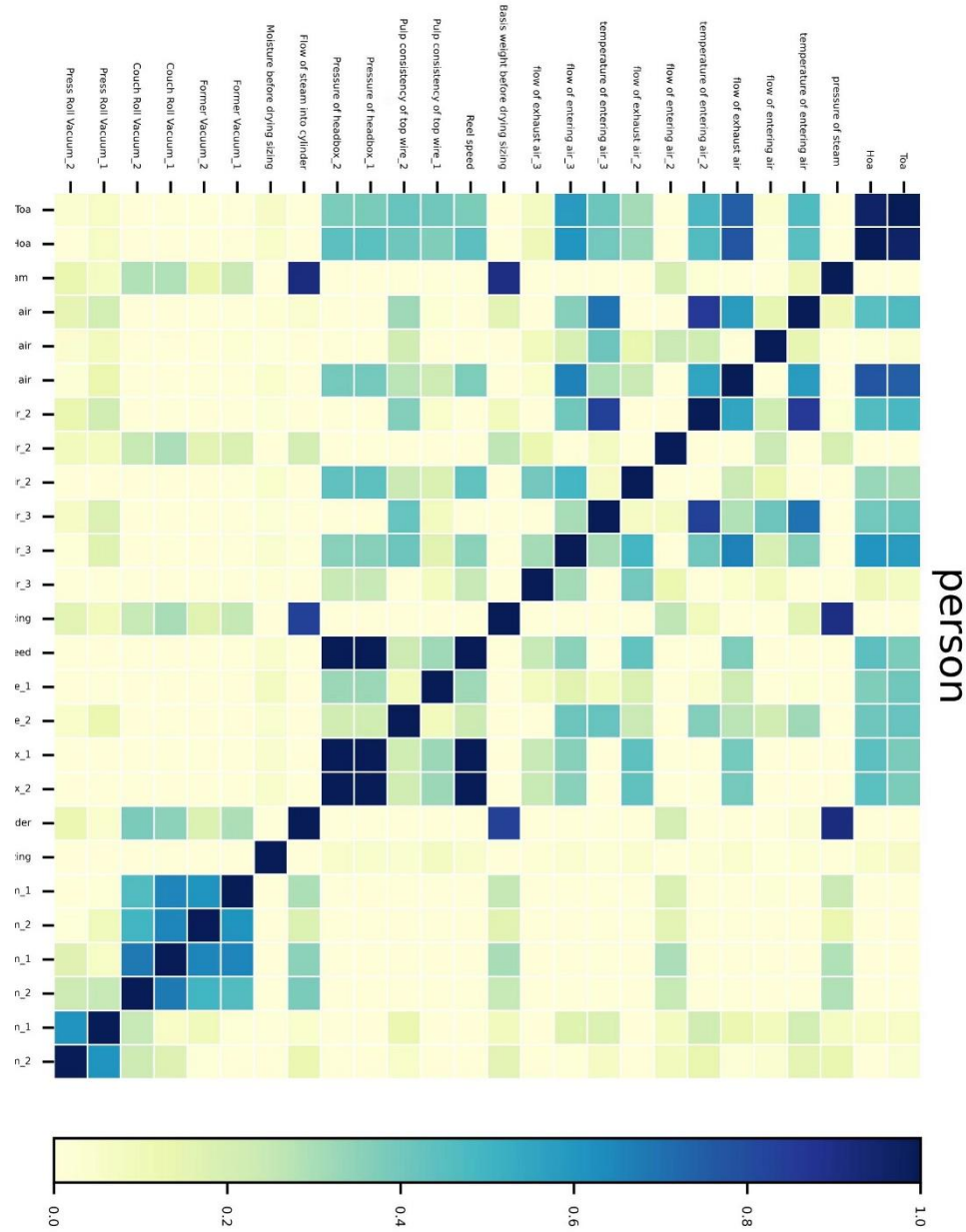

Figure 4. Correlation analysis results of T_{oa} and H_{oa}.

The prediction results of the prediction model for the first twenty data sets under the four operating conditions are shown in Figure 5. It is noted that the predicted values of random forest and gradient boosting regression are closer to the actual values, and the prediction of linear regression is less effective. In principle, linear regression is suitable for situations where there are few variables and the relationships between the variables are not particularly complex. However, papermaking is a complex process involving mass and heat transfer, and the relationship between variables is complex. In most cases, the relationship between parameters and parameters is non-linear, so linear regression is less effective in making regression predictions for non-linear data.

Both Random Forest and Gradient Boosting Decision Tree are integrated learning algorithms, they both consist of multiple decision trees, and the final result needs to be decided by all decision trees together. However, Random Forest and Gradient Boosting Decision Tree differ in their ideas. Random Forest adopts the idea of Bagging in machine learning, in which, Bagging draws samples from the training set to train weak classifiers by uniform sampling with put-back, and the training sets of each classifier are independent of each other. The training sets of decision trees are independent of each other, and the trees of Random Forest can be generated in parallel with each other. The gradient boosting decision tree uses the Boosting idea, where the training sets of each classifier are not independent of each other. The composed trees need to be generated serially in order, and the training sets of each weak classifier are sampled from the results of the previous weak classifier.

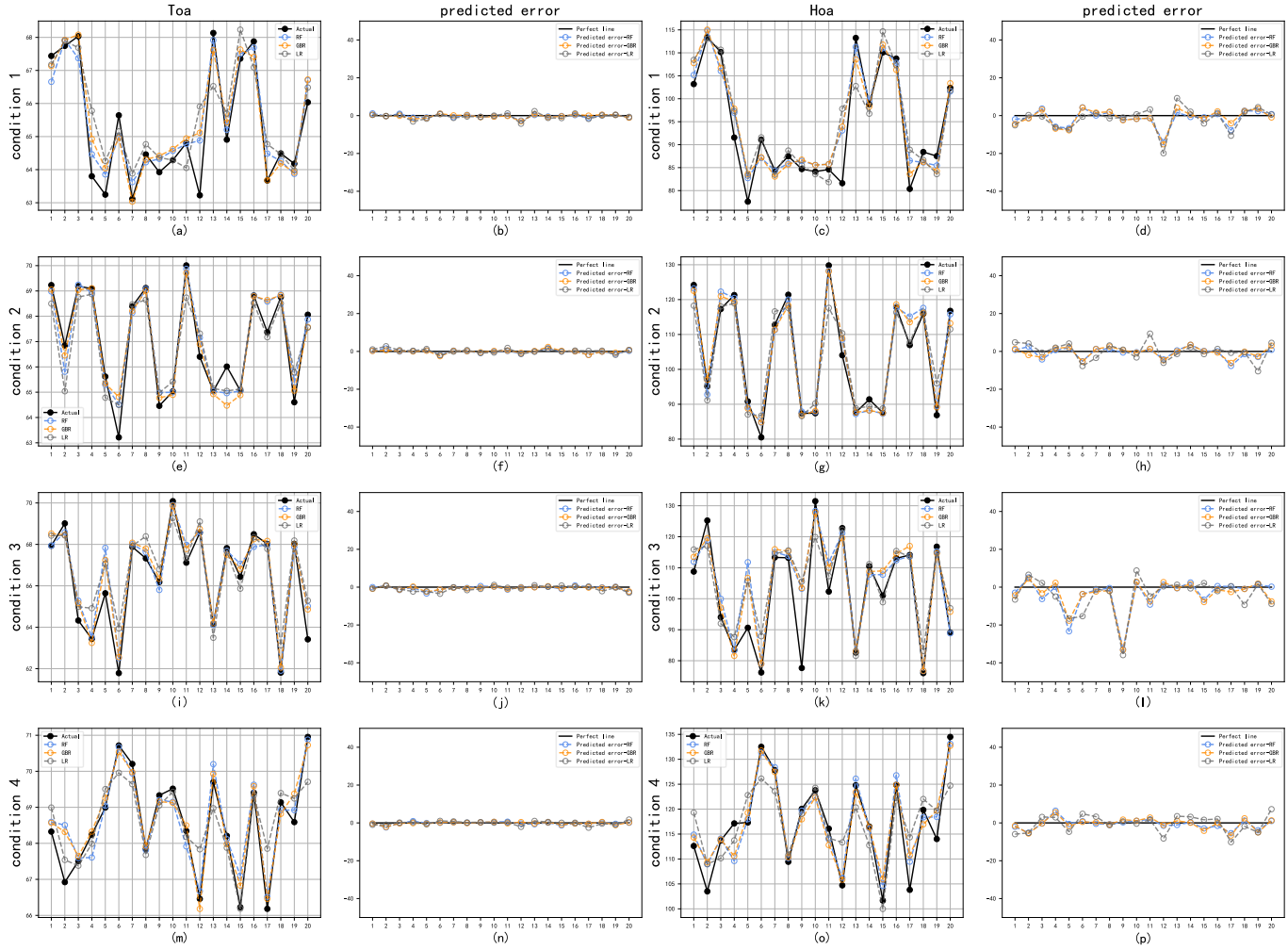

Figure 5. Prediction comparison of different models under four working conditions.

Figure 6 shows the evaluation of the effectiveness of the exhaust fan temperature and humidity prediction model, which is a comparison of the area of a triangle consisting of three angular evaluation indicators (R^2, MAE and MAPE, the calculation details refer to Eqs. (10)–(12)), with MAE and MAPE as the inverse and their actual values as an increasing trend from the center to the apex of the radar plot. As shown in Table 5, it means that the larger the area of the triangle in the figure, the better the performance of the model in all dimensions. It can be seen that in most cases the triangles of Random Forest and Gradient Boosting Regression have a high overlap and the area is larger than that of Linear Regression.

$$R(e,p) = \frac{\sum_{i=1}^{n}(e_i - \bar{e})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{n}(e_i - \bar{e})^2 \cdot \sum_{i=1}^{n}(p_i - \bar{p})^2}} \tag{10}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_i - p_i| \tag{11}$$

$$\text{MAPE} = \frac{\sum_{i=1}^{n}|(e_i - p_i)\,/\,e_i|}{n} \cdot 100\% \tag{12}$$

where e_i is the real targets, whereas p_i is the predicted output of the model.

According to the analysis above, it can be concluded that the prediction errors of both Random Forest and Gradient Boosting Regression are relatively small. From the values of the evaluation indices, the predicted values of random forest are better than those of gradient boosting regression, with better agreement with the actual value curve and a closer trend. It shows that the random forest model has better robustness and generalisation ability, and also shows more stable effect under different working conditions. So, it can be regarded that the random forest can be applied to complex drying conditions to provide reference for index prediction. It is also shown that the model is real-time and the prediction results can be used for online process monitoring.

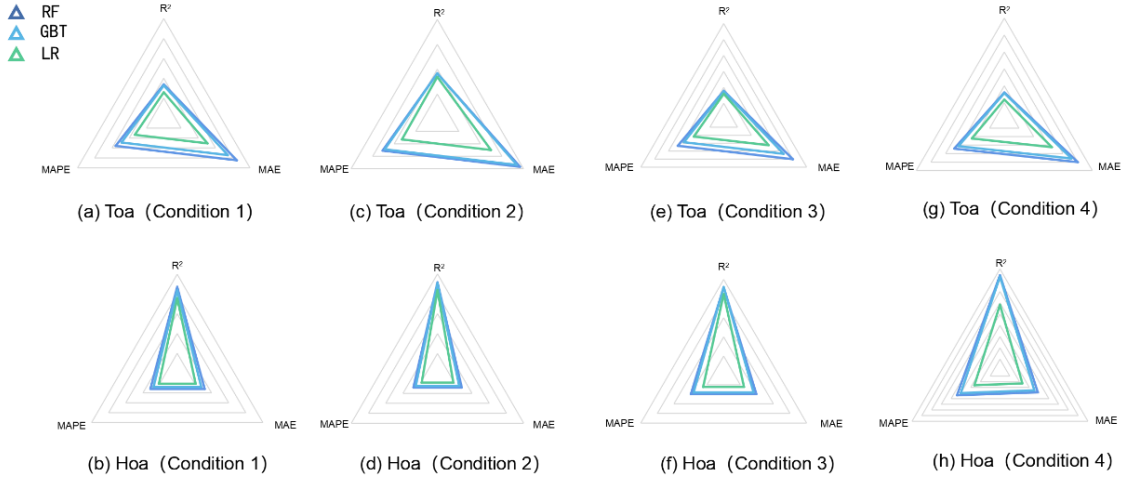

Figure 6. Triangle diagram of model performance in terms of R^2, MAPE, and RMSE.

Table 5. Comparison of the effects of exhaust temperature and humidity prediction models.

Variable	Model	Valuation Criteria	1	2	3	4
T_{oa} (℃)	RF	R^2	0.85	0.89	0.92	0.81
		MAE	0.47	0.40	0.52	0.40
		MAPE	0.72	0.60	0.79	0.59
	GBR	R^2	0.82	0.88	0.92	0.78
		MAE	0.54	0.46	0.54	0.44
		MAPE	0.82	0.69	0.82	0.64
	LR	R^2	0.65	0.80	0.85	0.59
		MAE	0.79	0.61	0.80	0.62
		MAPE	1.19	0.92	1.22	0.90
H_{oa} (%)	RF	R^2	0.87	0.92	0.92	0.85
		MAE	3.12	2.54	3.56	2.58
		MAPE	3.20	2.52	3.58	2.27
	GBR	R^2	0.84	0.92	0.91	0.83
		MAE	3.56	2.77	3.96	2.85
		MAPE	3.62	2.75	3.96	2.49
	LR	R^2	0.76	0.85	0.84	0.58
		MAE	4.65	4.08	5.44	4.37
		MAPE	4.68	4.03	5.45	3.82

3.3. Visualisation Model

The following simulation model is based on the example of a surface condenser, where the input and output processes of the surface condenser are analyzed and the input and output streams are identified based on the mechanistic equations. A new module is created in CADSIM Plus as a "Polygon", which is a closed shape used to represent a unit of equipment and, when connected to the flow lines, forms the basis of the simulation model. Then to select the streams to be used and draw the flow lines, select "Create Stream Definition" to access the "Stream Definition Wizard" and select the streams to be used in the process Select the stream components to be used in the process, such as liquids, gases and solids, or some property values such as temperature and pressure. Select "Paper" for the "Stream Type", "STEAM" for the secondary stream and "Temperature" for the other three streams, depending on the mechanism equation. The other three streams are selected as "WATER" and temperature, and are constructed as shown in Figure 7.

CADSIM Plus communicates with the COM server through the COMCLIENT module. External numerical data can be used in simulation calculations as part of the simulation model. CADSIM Plus can provide numerical data to an external application (e.g. Excel), which can perform its own calculations and then return the results to CADSIM Plus for further processing. Based on the pre-solved data, the accuracy of the simulation is verified by bidirectional data transfer via the software's COM function, and the operation is observed for different input parameters. It can be

found that the simulation results are the same as the mechanistic calculations when the heat loss is set to zero, and the module can accurately simulate the surface condenser. The performance is affected when operating parameters such as the temperature and mass of the incoming secondary steam or the temperature and mass of the cold water are changed. Any change in these parameters requires a completely new simulation.

In the real production processes, there is a certain amount of heat loss in the heat transfer process. In CADSIM Plus, it is possible to set the heat loss ratio for the equipment to obtain simulation results that are closer to the real situation. Table 6 shows that after adjusting the heat loss ratio for the equipment to 3%, the output term changes in a similar way to the real situation.

Figure 8 shows the interface of the surface condenser simulation run, the curves shown in the figure are the variation curves of condensate and temperature. In practical use, real-time data can be imported via EXCEL and the dynamic changes of output parameters can be observed after simulation calculations, and parameter variation curves can be constructed for each parameter in the model interface to provide reference for actual process decisions.

Table 6. Interface display of surface condenser model.

Heat Less	M_{iw}	T_{iw}	M_{is}	T_{is}	M_{ow}	M_{oc}	T_{ow}
0%	0.51	11.42	100	30	0.51	11.42	55.87
	0.58	11	95	32	0.58	11	62.38
	0.66	11.55	99	35	0.66	11.55	68.00
	0.54	11.61	101	33	0.54	11.61	59.94
	0.67	11.32	102	32	0.67	11.32	66.29
3%	0.51	11.42	100	30	0.51	11.42	55.84
	0.58	11	95	32	0.58	11	62.35
	0.66	11.55	99	35	0.66	11.55	67.97
	0.54	11.61	101	33	0.54	11.61	59.91
	0.67	11.32	102	32	0.67	11.32	66.25

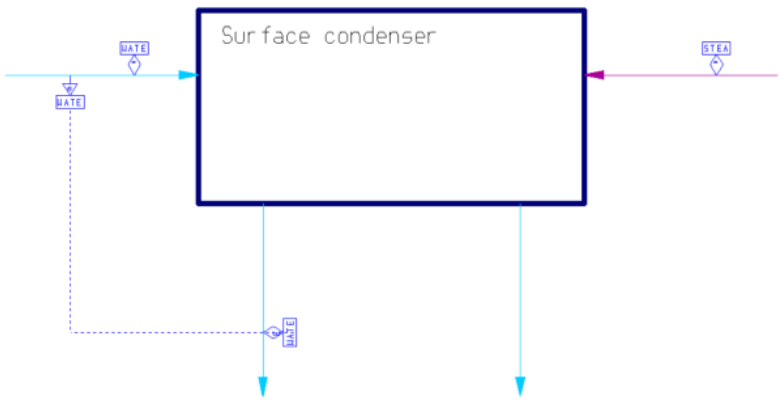

Figure 7. Interface display of surface condenser model.

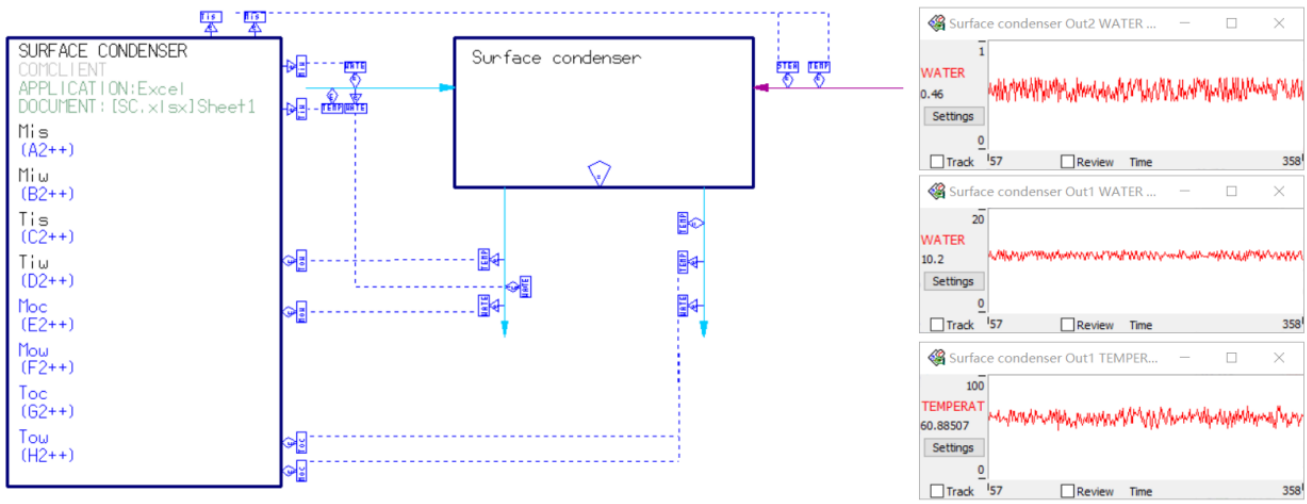

Figure 8. Real-time parameter monitoring interface of surface condenser model.

4. Conclusions and Future Works

At present, the papermaking industry lacks efficient means of whole life-cycle control. Therefore, this paper proposed a digital twin modelling framework for the paper manufacturing process, and developed two data completion methods for addressing the possible missing data in the modeling framework, including parameter solving based on mass and heat transfer mechanisms, and parameter prediction based on random forest.

The following conclusions are obtained: the mechanism-based parameter solution can be used as a general method to solve a variety of parameter missing problems and usually obtains better results; the random forest-based parameter prediction model is robust and has high accuracy, with the average value of R^2 above 0.9. Base on which, this paper implemented a visual modelling of the surface condenser in the dry section of papermaking process based on CADSIM Plus and the digital twin framework. The model runs well and is able to monitor the dynamic change of parameters in real time.

The digital twin-based visualization model proposed in this paper can reduce the coupling complexity of the physical entity modules, and has good scalability and generality, and can be subsequently extended to the whole process of paper production.

However, there are certain limitations of the model should be further considered in the future study. Current studies mostly focused on the drying section, pay scarce attention to other sections, which is hard to integrate the processes as a whole. Meanwhile, certain simplified assumption is too ideal to be applied in the industry, which should be studied deeper in the future. In addition, artificial intelligence and big data analysis technology can be furthermore to be explored by online analysis on the basis of interaction with the production site thorough data. And applying the established models to implement management of production process for optimization and decision making, and so on. The integration of these techniques could significantly improve the production efficiency and reduce production costs, and ultimately achieve sustainable development of the process.

Author Contributions

Made substantial contributions to conception and design of the study and performed data analysis and interpretation: Z.L.; Performed data acquisition, as well as provided administrative, technical, and material support: J.L. and M.H.

Ethics Statement

Not applicable.

Informed Consent Statement

Not applicable.

Funding

This research received no external funding.

Declaration of Competing Interest

All authors declared that there are no conflicts of interest.

References

1. Man Y, Han Y, Li J, Hong M. Review of energy consumption research for papermaking industry based on life cycle analysis. *Chin. J. Chem. Eng.* **2019**, *27*, 1543–1553.
2. Qian F, Bogle D, Wang M, Pistikopoulos S, Yan J. Artificial intelligence for smart energy systems in process industries. *Appl. Energy* **2022**, *324*, 119684.
3. Antonino PO, Capilla R, Pelliccione P, Schnicke F, Espen D, Kuhn T, et al. A Quality 4.0 Model for architecting industry 4.0 systems. *Adv. Eng. Inform.* **2022**, *54*, 101801.
4. Fan Y, Dai C, Huang S, Hu P, Wang X, Yan M. A life-cycle digital-twin collaboration framework based on the industrial internet identification and resolution. *Int. J. Adv. Manuf. Technol.* **2022**, *123*, 2883–2911.
5. Hu Y, Li J, Hong M, Ren J, Man Y. Industrial artificial intelligence based energy management system: Integrated framework for electricity load forecasting and fault prediction. *Energy* **2022**, 244, 123195.

6. Li Y, Yang C, Zhang H, Li J. Discussion on key technologies of digital twin in process industry. *Acta Automat. Sin.* **2021**, *47*, 501–514. doi:10.16383/j.aas.c200147 (In Chinese).
7. Zhao L, Fang Y, Lou P, Yan J, Xiao A. Cutting parameter optimization for reducing carbon emissions using digital twin. *Int. J. Precis. Eng. Manuf.* **2021**, *22*, 933–949.
8. Huynh TA, Zondervan E. Process intensification and digital twin-the potential for the energy transition in process industries. *Phys. Sci. Rev.* **2022**, *8*, 4859–4877.
9. He Z, Xu J, Tran KP, Thomassey S, Zeng X, Yi C. Modeling of textile manufacturing processes using intelligent techniques: a review. *Int. J. Adv. Manuf. Technol.* **2021**, *116*, 39–67.
10. Li M, He Z, Xu J. A comparative study of ozonation on aqueous reactive dyes and reactive-dyed cotton. *Color. Technol.* **2021**, *137*, 376–388.
11. He Z, Li M, Zuo D, Xu J, Yi C. Effects of color fading ozonation on the color yield of reactive-dyed cotton. *Dye Pigments* **2019**, *164*, 417–427.
12. Hu Y, Li J, Hong M, Ren J, Lin R, Liu Y, et al. Short term electric load forecasting model and its verification for process industrial enterprises based on hybrid GA-PSO-BPNN algorithm—A case study of papermaking process. *Energy* **2019**, *170*, 1215–1227.
13. Man Y, Yan Y, Wang X, Ren J, Xiong Q, He Z. Overestimated carbon emission of the pulp and paper industry in China. Energy **2023**, *273*, 127279.
14. Lai C, Wang Y, Fan K, Cai Q, Ye Q, Pang H, et al. An improved forecasting model of short-term electric load of papermaking enterprises for production line optimization. *Energy* **2022**, *245*, 123225.
15. He Z, Liu C, Wang Y, Wang X, Man Y. Optimal operation of wind-solar-thermal collaborative power system considering carbon trading and energy storage. *Appl. Energy* **2023**, *352*, 121993.
16. He Z, Qian J, Li J, Hong M, Man Y. Data-driven soft sensors of papermaking process and its application to cleaner production with multi-objective optimization. *J. Clean. Prod.* **2022**, *372*, 133803.
17. Zhang H, Li J, Hong M, Man Y, He Z. Cost Optimal Production-Scheduling Model Based on VNS-NSGA-II Hybrid Algorithm—Study on Tissue Paper Mill. *Processes* **2022**, *10*, 12072.
18. Zhang Z, He X, Man Y, He Z. Multi-objective scheduling in dynamic of household paper workshop considering energy consumption in production process. *J. Smart Environ. Green Comput.* **2023**, *3*, 87–105.
19. He Z, Chen G, Hong M, Xiong Q, Zeng X, Man Y. Process Monitoring and Fault Prediction of Papermaking by Learning from Imperfect Data. *IEEE Trans. Autom. Sci. Eng.* **2023**. doi:10.1109/TASE.2023.3290552.
20. He Z, Tran KP, Thomassey S, Zeng X, Xu J, Yi C. A deep reinforcement learning based multi-criteria decision support system for optimizing textile chemical process. *Comput. Ind.* **2021**, *125*, 103373.
21. Zhang Y, Hong M, Li J, Ren J, Man Y. Energy system optimization model for tissue papermaking process. *Comput. Chem. Eng.* **2021**, *146*, 107220.
22. He Z, Hong M, Zheng H, Wang J, Xiong Q, Man Y. Towards low-carbon papermaking wastewater treatment process based on Kriging surrogate predictive model. *J. Clean. Prod.* **2023**, *425*, 139039.
23. Soares RM, Câmara MM, Feital T, Pinto JC. Digital twin for monitoring of industrial multi-effect evaporation. *Processes* **2019**, *7*, 537.
24. Liu M, Fang S, Dong H, Xu C. Review of digital twin about concepts, technologies, and industrial applications. *J. Manuf. Syst.* **2021**, *58*, 346–361.
25. Schroeder GN, Steinmetz C, Rodrigues RN, Henriques RV, Rettberg A, Pereira CE. A methodology for digital twin modeling and deployment for industry 4.0. *Proc. IEEE* **2020**, *109*, 556–567.
26. Shabbir I, Mirzaeian M, Sher F. Energy efficiency improvement potentials through energy benchmarking in pulp and papermaking industry. *Clean. Chem. Eng.* **2022**, *3*, 100058.
27. He Z, Tran KP, Thomassey S, Zeng X, Xu J, Yi C. Multi-objective optimization of the textile manufacturing process using Deep-Q-Network based multi-agent reinforcement learning. *J. Manuf. Syst.* **2021**, *62*, 939–949.
28. Xu J, He Z, Li S, Ke W. Production cost optimization of enzyme washing for indigo dyed cotton denim by combining Kriging surrogate with differential evolution algorithm. *Text. Res. J.* **2020**, *90*, 1860–1871.
29. Xu J, Liu F, He Z, Zhang Z, Li S. Cost optimization of sodium hypochlorite bleaching washing for denim by combining ensemble of surrogates with particle swarm optimization. *J. Eng. Fiber. Fabr.* **2021**, *16*. doi:10.1177/15589250211022331.
30. Li J, Tian X, Liu J. Dynamic Data Scheduling of a Flexible Industrial Job Shop Based on Digital Twin Technology. *Discrete Dyn. Nat. Soc.* **2022**, *2022*, 1009507.
31. Bamunuarachchi D, Georgakopoulos D, Banerjee A, Jayaraman PP. Digital twins supporting efficient digital industrial transformation. *Sensors* **2021**, *21*, 6829.
32. Zhuang C, Miao T, Liu J, Xiong H. The connotation of digital twin, and the construction and application method of shop-floor digital twin. *Robot. Comput. Integr. Manuf.* **2021**, *68*, 102075.
33. Negri E, Berardi S, Fumagalli L, Macchi M. MES-integrated digital twin frameworks. *J. Manuf. Syst.* **2020**, *56*, 58–71.

34. Tao F, Sui F, Liu A, Qi Q, Zhang M, Song B, et al. Digital twin-driven product design framework. *Int. J. Prod. Res.* **2019**, *57*, 3935–3953.
35. Zhang Y, Wang W, Zhang H, Li H, Liu C, Du X. Vibration monitoring and analysis of strip rolling mill based on the digital twin model. *Int. J. Adv. Manuf. Technol.* **2022**, *122*, 3667–3681.
36. Ding G, Guo S, Wu X. Dynamic Scheduling Optimization of Production Workshops Based on Digital Twin. *Appl. Sci.* **2022**, *12*, 10451.
37. Goodwin T, Xu J, Celik N, Chen CH. Real-time digital twin-based optimization with predictive simulation learning. *J. Simul.* **2022**, doi:10.1080/17477778.2022.2046520.
38. Yin Y, Liu J, Wang Y, Zhuo Y, Meng Y. Modeling of Ventilation's Influence on Energy Consumption in Multi-cylinder Dryer Section Part1: Theoretical Model. *Int. J. Comput. Intell. Syst.* **2022**, *15*, 1–13.
39. Marques JP, Cunha DC, Harada LM, Silva LN, Silva ID. A cost-effective trilateration-based radio localization algorithm using machine learning and sequential least-square programming optimization. *Comput. Commun.* **2021**, *177*, 1–9.
40. Liu Y, Shen W, Man Y, Liu Z, Seferlis P. Optimal scheduling ratio of recycling waste paper with NSGAII based on deinked-pulp properties prediction. *Comput. Ind. Eng.* **2019**, *132*, 74–83.
41. Jadidi A, Menezes R, de Souza N, de Castro Lima AC. Short-term electric power demand forecasting using NSGA II-ANFIS model. *Energies* **2019**, *12*, 1891.
42. Verma S, Pant M, Snasel V. A comprehensive review on NSGA-II for multi-objective combinatorial optimization problems. *IEEE Access* **2021**, *9*, 57757–57791.
43. Ao Y, Li H, Zhu L, Ali S, Yang Z. The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. *J. Pet. Sci. Eng.* **2019**, *174*, 776–789.
44. Ciulla G, D'Amico A. Building energy performance forecasting: A multiple linear regression approach. *Appl. Energy* **2019**, *253*, 113500.
45. Rathore SS. An exploratory analysis of regression methods for predicting faults in software systems. *Soft Comput.* **2021**, *25*, 14841–14872.
46. Otchere DA, Ganat TOA, Ojero JO, Tackie-Otoo BN, Taki MY. Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *J. Pet. Sci. Eng.* **2022**, *208*, 109244.
47. Cai J, Xu K, Zhu Y, Hu F, Li L. Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. *Appl. Energy* **2020**, *262*, 114566.
48. Sun L, Ji Y, Zhu X, Peng T. Process knowledge-based random forest regression for model predictive control on a nonlinear production process with multiple working conditions. *Adv. Eng. Inform.* **2022**, *52*, 101561.