Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **7**, Number **3**, August **2022**

Research on Quality Control Application of Whole Process Intelligent Manufacturing in Steel Industry 4.0 Based on Big Data Analysis

Feng Zhao

School of Management Science and Engineering,
Anhui University of Technology
Xiushan District,Maanshan City, Anhui Province, China
zhaofeng@tech.ahut.edu.cn

Chen Yin*

School of Management Science and Engineering,
Anhui University of Technology
Xiushan District,Maanshan City, Anhui Province, China
297914542@qq.com

Xiaoxiao Huo

School of Aerospace Engineering,
Xiamen University
Xiang'an District, Xiamen City, Fujian Province, China
1056058740@qq.com

Yingying Xu

College of Science,
Hohai University
Nanjing China
837004231@qq.com

*Corresponding author: Chen Yin

Received March 1, 2022, revised April 22, 2022, accepted June 10, 2022.

Abstract. *In this paper, in the project of upgrading the intelligent manufacturing Industry 4.0 of B steel plant, a large amount of measured historical sample data and real-time data collected by sensors are processed to solve the need of the enterprise to change from univariate to multivariate monitoring of the full process operation of the whole process of the strip in the period of industrial development of big data cloud platform. In this paper, the monitoring and diagnostic identification of the quality of the whole process are achieved by applying a set of algorithm structures such as improved ZC4.5 decision tree algorithm - MSPM algorithm (comparing performance with MICA algorithm) - improved statistical result comparison analysis Granger analysis. This paper proposes to redevelop the general idea and methodology of data processing based on the theoretical idea of the MES system of steel plants to upgrade the big data technology is conducive to the transformation of a single process to full process quality control mode, improve fault diagnosis accuracy, reduce manual work, lay the foundation for product quality improvement of enterprises, and can be extended to most steel enterprises.*
Keywords: Big data analytics; intelligent Manufacturing; Whole process quality management; MSPM algorithm

1. **Introduction.** In recent years, because of the cyclical changes and adjustments in the steel industry, a mega-combined enterprise has been gradually integrated and established, and the production line has increased from the initial steel treatment to the subsequent production processes. However, the corresponding companion is still very far apart. With the advent of Industry 4.0 (including technologies related to Industrial Internet of Things (IIoT), Artificial Intelligence (AI), process simulation and optimization, cognitive computing, and cloud computing [1]) paradigm, the original MES (manufacturing execution system) system can no longer meet the requirements of the existing information system development[2]. In addition, this research experiment required is very high, because the equipment for iron and steel enterprise investment is huge, so to realize the whole process of monitoring all the way, must go to all involved in the process of plant collection in each production line PLC (programmable logic controller) sensor data, the production line PLC said there are hundreds of little quantity. The data of thousands of sensors need to be synchronized with the clock (PLC data corresponds to the time of the same product through different processes), and the transmission data is linked with the characteristic equation processing, sorting out the dimensionless. Jaskó et al.[3] mentioned in their literature review that next-generation MES solutions need to have such machine learning (ML) data mining capabilities. The purpose of this paper is to show the development and results of relevant data mining functions, as can be seen from the review of relevant industrial informatics papers by Chen[4].

Due to the increase in production processes and the development of sensor technology, the amount of industrial-level data has gradually entered into explosive growth, which can no longer be solved by pure statistical models, and Kano and Nakagawa [5] pointed out that statistical models are not suitable for processing large amounts of data when looking at the future of the steel industry, so the research method of control charts based on statistical theory is not a practical application development direction for future steel enterprises.

The new quality control system must adapt to the existing MES system and ERP(Enterprise Resource Planning) system, and realize the monitoring, control, and optimization of THE manufacturing process by the MES system. The information provided by MES helps decision-makers understand the interconnections among the various subsystems involved in the production, and this knowledge can promote continuous improvement of the manufacturing industry. Therefore, it is necessary to refer to the previous development system of quality management theory when developing a new multivariate monitoring model. This paper mentions the use of Six Sigma theory about the whole product life cycle theory, similar to what happens in the production chain mentioned by SzilardJasko, AdriennSkrop. In product Lifecycle Management (PLM) systems, there is a common basic idea about the industry 4.0 paradigm, everything must support computerization. Use computer-based controls throughout the production chain.

Data-driven methods are complementary techniques that can be used in conjunction with classical and advanced control algorithms. Their main focus in complex processes is to ensure compliance of product quality [6], presents interactions between its different variables, time-varying parameters of mass and heat transfer nonlinearities dynamics, and large time delay due to the slow chemical reactions [7]. These intrinsic features make the use of model-based techniques for fault detection and control purposes very difficult it is not unrealistic [8]. Furthermore many cement plants worldwide still use centralized manual control methods to ensure operation. Therefore, it would make sense to integrate appropriate fault detection and diagnosis systems that can easily and accurately monitor this type of industrial process.

Regarding the methods applied in this paper for quality monitoring, Qin [9] mentioned in a literature review that due to the data-based nature of the Statistical Process Monitoring SPM method, it is relatively easy to apply to real processes of considerable scale compared to other methods based on system theory or rigorous process models. Traditional SPM methods include principal component analysis (PCA) and partial least squares (PLS) [10], and it is generally assumed that the system variables are static and follow a Gaussian distribution. Recent developments in SPM also include dynamic modelling, non-Gaussian distributions, and non-linear methods [11, 12], they can be applied to a wider range of industrial processes. Due to the data-based nature of the SPM methods, it is relatively easy to apply to real processes of rather large scale comparing to other methods based on systems theory or rigorous process models. The mature use of control charts tends to focus on univariate monitoring, and the commonly used univariate control charts are mainly Shewhart control charts, cumulative sum control charts, and exponentially weighted moving average control charts. The use of Multivariate Quality Control techniques is usually avoided by practitioners because of the complexity involved in the design, implementation, and maintenance of the control system. Sepú lveda and Nachlas [13] proposed a simulation approach to multivariate quality control. From now on, it does not apply to the application of large-scale industrial-level data, and the main mechanism of control charts is based on statistical theory, so it must be discarded, also in line with the academic trend that machine learning theory will surpass statistical theory in future research. Bakdi et al. [14] used PCA analysis in chemical applications applied to cement chemical production, some scholars believe that machine learning algorithms such as PCA and some evolutionary principal meta-analysis methods cannot indicate which process variables are responsible for the anomalies, the contribution of this paper lies in such a large data magnitude of a day's update in just over 100,000 industrial-grade data, through decision tree algorithm-MSPM algorithm-Granger analysis, to achieve the monitoring and diagnosis of the quality of the whole process Identifying, the experiments locate the final process variables through a simplified Granger analysis to achieve the identification of abnormal sources. A set of identification variables for the process window is provided for validation by researchers with similar product lines The analysis results are shown in Table 1.

Table 1. A set of experimental process window monitoring variables

Number	Case 1: Pressure oxygen quality monitoring point
1	Si, As content monitoring
2	Air-fuel ratio and heating time of the first heating section, second heating section, third heating section and soaking section
3	Total heating time, air excess coefficient
4	Pressure, flow rate and effect of one-time phosphorus removal
5	Rough rolling phosphorus removal pass, pressure
6	Exit temperature of roughing mill and maximum temperature of finishing mill entrance
7	F1-F4 slot spray open state
8	Rolling rhythm and predicted temperature of wheel face
9	Phosphorus removal by finishing rolling
10	F1-F4 working rod cooling water, etc.

Why combine decision tree algorithm with MSPM algorithm? Because multivariate statistical methods use only few possible independent sources mostly causing the variation in the process and rendering these methods to be very efficient in monitoring large scale processes [15]. The kernel entropy component analysis method can also be considered here in combination with the MSPM algorithm. However, for large scale systems including many complex units, data-driven methods make a superior alternative solution for fault detection and diagnosis systems [16, 17]. Generally, large size multivariate data with highly correlated variables present low statistical rank. Since the number of process variables on the production line is very large, even if the scope is reduced through metallurgical experience, it is necessary to further determine the main variables. In this paper, the decision tree algorithm is firstly used to screen out three main process variables: the shape size of 2# rolling mill, steel passing rate, and the height of 1# ∼5# looper sleeve. Because of the steel passing rate (PLC drawing speed), this process variable read all of his PLC related data acquisition volume, including: Column1: Setting value of drawing speed; Column2:Actual drawing speed; Column3: Set liquid level of molten steel; Column4: Actual liquid level of molten steel; Column5: Opening and closing value of plugrod. Then the MSPM algorithm is analyzed for the data on the 1000 time points of these five data collection points (note that this step of the collection should be clock synchronized, because there is a time difference when the steel is run through different PLCs in the production line). Finally, the source of the anomaly is introduced in reverse by Granger analysis. This paper focuses on the structure of this algorithm, in which the formula code of the decision tree algorithm - MSPM algorithm - Granger analysis is modified in order to adapt to the line data of different process variables in different production lines, whether the data is selected for dimensionless processing depends on the process variables in the metallurgical process.

As general-purpose steel, the plate has a wide range of uses in various fields. The plate can be processed into various components and products, so it has a large number of customers. However, there are still many problems with sheet and strip products. There are many individual customer needs, and the process of producing products through the "identify-verify-cure" process is extremely complex. The sheet market is still facing high demand and low supply, despite the availability of raw materials, capital, and labor. Technological progress is the main breakthrough direction, the whole process of integrated design and personalization of the digital twin model needs to be established. The quality control technology of big data analysis of the whole process studied in this paper can provide an important means of accurate service and ultimately achieve the ultimate goal of improving the efficiency of enterprises.

FIGURE 1. Actual production line

1.1. **Existing device information system.** L1: Basic automation system refers to the PLC control unit used for automatic equipment control in the strip production line. This system is mainly used for recording process curve data, critical event status data, etc. L2: Process automation systems cover many mechanized operations on the production line. Including the mathematical model and material tracking system, storage material and process parameter set value, measured value (feedback value), statistics, and other logical correspondence. At the same time, it records the processing time information of products in each piece of equipment. The steelmaking area includes (converter - desulfurization - pouring - KR - refining - flame cleaning machine - 1, 2# continuous casting machine, 3# continuous casting machine) - hot rolling area - cold rolling area (acid rolling - 1# galvanizing - 2# galvanizing - cover receding - leveling - shearing). Another external property line is (acid rolling - 1# continuous rewind - 2# continuous rewind - recoil - 1# annealing furnace - 2# annealing furnace). In the process of information transformation, it is also a complex and arduous project to build the digital twin model of the site. This paper focuses on data mining, in which the model is similar to human skeleton and data mining is human blood. The actual production line is shown in Figure 1. The production diagram of crude steel and its one - to - one ratio digital twin model is shown in Figure 2.

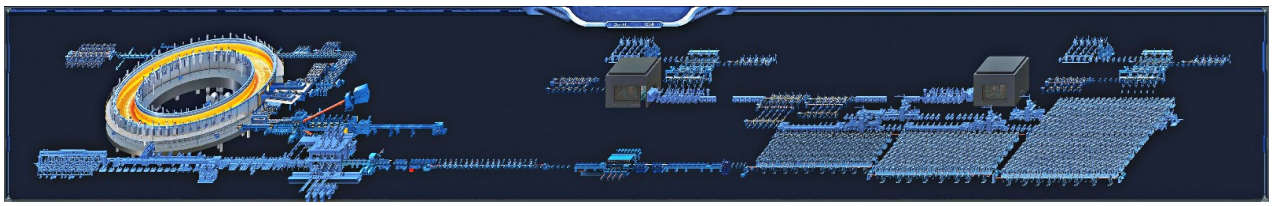


FIGURE 2. Production diagram of crude steel and its one - to - one ratio digital twin model

1.2. **Problems existing in the strip process quality system.** (1) The existing control systems lack attention to the quality determination process data, and each data is relatively isolated in the whole process flow operation process of the strip. Since there are many factors affecting product quality, there is a high-dimensional and multivariate coupling problem in the production process. It is difficult to discover the root causes affecting quality through simple threshold measurements. In particular, the relationship between the problems occurring in hot rolling and steelmaking, and the relationship between the problems occurring in cold rolling and hot rolling require further cluster analysis and cross-process correlation analysis of the process data. There is a lack of effective correlation analysis tools for exploring quality problems in the analysis of multidimensional factors.

(2) The production depletion of sheet and strip is high. It is difficult to match the individual needs of customers with the whole process of product quality production. The only remaining transactions are still in the form of paper agreements. The entire strip production process cannot be quality controlled for different needs. Because of the existing problems, there is an urgent need to improve the way of exchanging opinions to increase efficiency in the sheet trade.

(3) External disturbances in process input and quality variation in the process window are still difficult to control. At the production site, the formulation of process windows such as speed, temperature, heating temperature, arrival time, arrival composition, etc. is still determined empirically. The above-mentioned effects on the surface quality of the strip have not been included in the analysis. The strip production quality lacks long-term

stability, data traceability is inefficient, and the boundaries between processes and departments are unclear. Due to departmental divisions, quality problems cannot be effectively tracked and analyzed between different processes

2. **General idea.** (1) The researchers improved the PCA algorithm and designed the algorithm multivariate statistical process monitoring (MSPM), which is suitable for determining the surface quality of plates.
(2) To solve the problems arising in the process, the researchers used decision trees to filter and analyze the relational data, screened the three most representative operational process variables to provide the basis for the new algorithm, and selected the training data set.
(3) To meet real-time monitoring requirements, the screened data are entered into a database and modelled to visualize the data (process curve control charts). The plate production data is monitored for abnormalities during the production process.
(4) The MSPM algorithm and supporting software were developed by solving the actual problem of plate production in company B, thus verifying the purpose of the MSPM algorithm. As a result, it both meets the actual production needs and accelerates the pace of moving towards Industry 4.0.

2.1. **Concept Statement.** (1) Decision tree model: the introduction of penalty parameter Z improves the ID3 algorithm to form a new C4.5 algorithm more suitable for process variable screening.
The decision tree algorithm used in this paper for screening key factors is a very classical machine learning algorithm, which is suitable for integrated learning such as random forest algorithm and can be used as a regression algorithm, and also can be used as a classification. The decision tree model is a typical tree structure, and its learning process consists of feature selection, decision tree generation and pruning. Since this paper applies the decision tree algorithm to screen the process variables that play a major role in the quality problem, the pruning process is not considered in the paper.
The decision tree algorithm uses a tree model, where the linear model is a linear model where all features are given weights to sum up to get a new value, while the tree model is a partition for each feature. Decision trees can find non-linear divisions. The tree model is closer to the human way of thinking and can produce visual classification rules that produce models with interpretability. The function fitted by the tree model is a step function of the partition. The decision tree starts at the root node, which iterates from the top down to produce multiple internal nodes and leaf nodes. Each node of the tree represents a test of a feature, and the branches of the tree represent the results of each test of that feature. Each leaf node of the tree represents a category, and the final child node that cannot be split is called a leaf node and can represent the final category.
As can be seen above, the core of the decision tree algorithm lies in how to make the optimal attribute selection, and there are three main criteria for the optimal selection of decision trees, which are maximum information gain, maximum information gain ratio and Gini coefficient. The algorithms corresponding to these three criteria are ID3 algorithm, C4.5 algorithm and CART algorithm.
(2) Information gain: the difference between the empirical entropy of set D and the empirical conditional entropy $H(D|A)$ of set D under the given conditions of feature $A.H(D)$ denotes the empirical entropy of data set $D, H(D|A)$ denotes the empirical conditional entropy of set D under the given conditions of feature $A, g(D, A)$ denotes the information gain, and the information gain is calculated as follows:
Let A be - a discrete random variable with finite values and its probability distribution

is :

$$P(A = x_i\ ) = p_i, i = 1, 2, \cdots, n \tag{1}$$

Then the first degree of the random variable A is defined as:

$$H(A) = -\sum_{i=1}^{n} p_i \log p_i \tag{2}$$

The conditional entropy function $H(D|A)$ is defined as the uncertainty of a sample set D of random variables conditional on a random variable.

$H(D|A)$ is the mathematical expectation of the entropy of the conditional probability distribution of D of A under the given conditions on A.

$$H(D \mid A) = \sum_{a \in A} p(a) H(D \mid A) \tag{3}$$

Where $p(a)$ denotes the probability of occurrence of $A = a$.

$$g(D, A) = H(D) - H(D \mid A) \tag{4}$$

$$H(D) = -\sum_{k=1}^{k} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \tag{5}$$

$$H(D \mid A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H\left(D_i\right) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{k} \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|} \tag{6}$$

For the sample set D, the random variable A is the category of the sample, i.e., the sample is assumed to have k categories and the probability of each category is $|Ck|/|D|$, where $|Ck|$ denotes the number of samples of category k and $|D|$ denotes the total number of samples.

Information gain - $G(D, A)$ indicates the degree of uncertainty reduction in the classification of the dataset due to the selection of feature A. The more the reduction, the lower the uncertainty in the classification of the dataset. The more the reduction, the lower the uncertainty of the dataset classification.

$H(D)$ indicates the entropy of the label category of the dataset, i.e., the uncertainty when each label takes the value of a different category.

$H(D|A)$ denotes the entropy of the category label of the dataset under the condition that features A is selected. In this case, it can also represent the mutual information of category labels and features. (3) ID3 algorithm calculates the information gain of all node's technical features and selects the feature with the largest information gain for splitting.ID3 algorithm tends to select features with more values, and sometimes this tendency brings some error in the construction of decision trees. Wang et al. [18] proposed an RLBOR algorithm, which considers The number of nodes in the decision tree model to optimize the decision tree optimization ratio, Decision Tree Optimization Ratio but still can not avoid the error, such as extreme conditions, based on a certain attribute after splitting, a subset corresponds to a data, when the information gain is maximum and the information entropy is 0, but this division has no value. Because metallurgical process variables have a relatively small range of data variation due to the variables themselves, there is continuous data, and ID3 will favor that feature A. To correct this error in the ID3 algorithm, the C4.5 algorithm that uses the information gain ratio as the optimal attribute selection index is proposed. Mu et al. [19] mentioned the application of the C4.5 algorithm and pointed out that in supervised classification, large training data are very common and decision trees are widely used. However, many supervised classifiers (including the classical C4.5 tree) cannot directly handle large data due to some bottlenecks such

as memory limitations, time complexity, or data complexity. In supervised classification, large training data are very common, and decision, However, as some bottlenecks such as memory restrictions, time complexity, or data complexity, many supervised classifiers One solution for this problem is to design a highly parallelized learning algorithm. One solution for this problem is to design a highly parallelized learning algorithm. And in this paper, we need a C4.5 algorithm for feature selection of metallurgical process variables, and then choose other machine learning algorithms for big data processing by combining the characteristics of metallurgical data. The proposed ZC4.5 algorithm is that when a feature corresponds to too many values, and improved penalty parameter Z is taken to multiply the information gain so that the information gain ratio is small. The information gain ratio is defined as the ratio of the information gain brought by feature A to set D to the entropy of feature A itself.

Information gain ratio C4.5: The ratio of the information gain brought by feature A to set D to the entropy of feature A itself. The information gain ratio is calculated as follows:

$$g_r(D, A) = \frac{g_r(D, A)}{H_A(D)} \tag{7}$$

$$H_A(D) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \log_2 \frac{|D_i|}{|D|} \tag{8}$$

Information gain ratio = penalty parameter * information gain.

Penalty parameter info=$1/H_A(D) * Z$

$$Z = \begin{cases} 0 & R < 1\% \\ 1 & 1\% < R < 30\% \\ \text{null} & 30\% < R \end{cases}$$

When the value of Z changes in the range of less than 1%, the code set $Z = 0$, to facilitate the writing of the model code calculation automatically screens invalid process variables to reduce the pressure of the calculation.

when the range of value change of Z is greater than 30% (this PLC has an error), the code sets $Z = Null$ at this time to indicate that the data model this PLC has an error.

When the value of Z changes in the range of 1 to 30%, indicating that the data can be operated normally, $Z = 1$.

In the formula, n indicates the number of subsets after splitting based on feature A, and Di indicates the number of samples in each subset.

(4) PCA(Principal Component Analysis), PCA, is a technique for data analysis.PCA can extract the main factors of multivariate things, remove the noise and redundancy in the whole data, and reveal its essential characteristics.

3. **Experimental analysis.** In this paper, a new multivariate statistical process monitoring (MSPM) algorithm is formed by adding Hotelling- T2 and squared prediction error (SPE) as indicators for fault determination based on PCA dimensionality reduction, which transforms the univariate monitoring algorithm of metallurgical production lines into a multivariate monitoring algorithm.

3.1. **Initial screening model of indicators based on decision tree algorithm.** In this paper, the historical data of 13 production process parameters were obtained from steel plant B. The decision tree algorithm was used to feature select these parameter variables, and the three most representative and independent operation variables were selected from these variables , to retain the main features for the next step of model training and ignore some minor factors. The actual data of each operational variable are limited

to a certain range according to the needs of the actual production environment. For the acquired 60,000 sample data, the quality parameters of qualified samples are marked as 1, and the quality parameters of unqualified products are marked as 0. The truncated part of the data is shown in Table 2 below. We used the C4.5 classification tree for feature

Table 2. Historical data of process parameters of strip production line

Sample number	1	2	3	4	5	6	7	8
Steel speed	128.53	159.92	156.64	141.45	125.03	117.87	168.81	151.6
Descaling pressure	16.18	15.3	16.94	14.88	14.74	17.78	15.09	15.72
1#∼5#Looper height	110.66	126.38	109.25	116.37	129.91	133.62	128.17	113.56
Tension control	1.98	1.42	0.76	0.88	0.63	0.72	0.86	1.37
Heating section temperature	1140.52	1153.61	1173.83	1152.34	1121.4	1137.25	1125.36	1120
Average heat segment temperature	1165.74	1195.66	1166.72	1194.48	1181.12	1185	1197.23	1162.36
Advance quantity of pinch roll	9.45%	5.07%	5.93%	2.31%	3.98%	9.11%	4.16%	7.18%
Finishing mill speed	81.66	19.36	37.73	16.88	83.21	91.84	82.88	39.5
2#Rolling mill material size	17.16	17	16.9	16.84	17.31	17.23	17.05	16.75
Roller speed	78.98	71.83	61.35	79.2	82.87	60.56	83.54	60.55
Fan delivery	11.97	23.02	0.37	39.53	64.08	28.61	85.09	94.92
Guideway alignment accuracy	0.2	-0.32	0.33	0.16	-0.08	-0.38	-0.06	-0.48
Water tank	2.06	1.45	2.01	0.3	2.14	0.84	1.39	2.47
Quality	1	0	0	1	1	1	0	1

selection on the above sample data. Since the samples were classified as qualified (noted as 1) and unqualified (noted as 0), a binomial classification tree was used as the training model. We calculated the information gain ratios corresponding to the 13 feature variables and sorted them in order from largest to smallest, and selected three main factors based on the information gain ratio of each feature variable.

The experimental environment was set up as Intel(R) Core(TM) i5-5200U_CPU_@3.60GHz, and the operating system was Windows 10, which was implemented in the Jupyter notebook platform using the Python language. The analysis results are shown in Table 3. According to the quality inspection department's records of historical samples, the main factors affecting the surface quality of the strip are the size of the mill shape (2#), the oversteel rate, and the height of the live sleeve (1#∼5#) produced by the equipment. The variation of the above three variables is mainly monitored at the time of going online. We define thresholds in our experiments, above which an alarm is generated. Through

TABLE 3. Information gain analysis results

Characteristic variables of process parameters	Information gain ratio
2# rolling mill material shape size	0.1645
Steel speed	0.1389
1#∼5# looper height	0.1229

process control, surface quality problems such as periodic wax hanging, iron oxide, and cracks can be reduced for this process parameter. To meet the needs of the enterprise, the univariate monitoring problem is transformed into a multivariate monitoring problem after extracting the important influencing variables, and a multivariate monitoring model is developed.

3.2. **Dimensionality reduction model for quality inspection data based on principal component analysis.** Principal component analysis (PCA) is one of the basic projection models in multivariate statistical analysis. PCA transforms the original data into a set of linearly independent representations of each dimension through a linear transformation, which can be used to extract the main characteristic components of the data and is often used for dimensionality reduction of high-dimensional data. If there are n variables in the original data table, PCA will consider the information in this data table to readjust the combination and extract P main feature variables from it ($p < n$).
Let $x \in R^m$ be the measurement samples of m sensors, and each sensor has n independent samples. Construct matrix $X = [x_1,\ x_2, \ldots\ldots, x_n]^T \in R^{(n\times m)}$ where each column represents a measurement variable and each row represents a sample. Treat the columns of X as variables with unit variance and zero mean. Define S as the normalized sample X covariance matrix. Decompose the eigenvalues and arrange them in descending order. The improved PCA model is used to decompose X as follows:

$$X = \hat{X} + E = TP^T + E \tag{9}$$

$$T = XP \tag{10}$$

Where $P \in R^{m\times A}$ is the load matrix composed of the first A feature vectors of S. $T \in R^{n\times A}$ is the scoring matrix. The columns of T are called principal variables, and A represents the number of principal elements.
PCA model divides the variable space into two orthogonal and complementary subspaces. The subspace composed of all columns of P is called the principal component space (PCS), and the orthogonal complement of PCS is called the residual subspace (RS). Any sample vector can be decomposed into a projection on a host subspace and a residual subspace:

$$x = \hat{x} + \widetilde{x} \tag{11}$$

$$x^{\wedge} = \mathbf{P}\mathbf{P}^T x \in R_p \equiv \mathbf{P} \tag{12}$$

$$\widetilde{x} = \left(\mathbf{I} - \mathbf{P}\mathbf{P}^T\right) x \in R \equiv \mathbf{P}^{\perp} \tag{13}$$

$x^{\wedge}$ denotes the projection of the sample in the principal element space that is modeled, and $\widetilde{x}$ denotes the projection of the sample in the residual space that is not modeled.
The input: N dimensional sample set $D = (x^{(1)}, x^{(2)}, \ldots, x^{(m)})$, the dimension reduced to p;
The output: Sample set D' after dimensionality reduction;
1) Centralize all samples: $x^{(i)} = x^{(i)} - \frac{1}{m}\sum_{j=1}^{m} X^{(j)}$;
2) Calculate the covariance matrix of the sample xx^T;
3) Do the eigenvalue decomposition of xx^T;

4) Take the eigenvectors $\{\omega_1, \omega_2, \ldots\ldots, \omega_{n'}\}$, after all the feature vectors are normalized, the feature vector matrix W is formed.
5) To a new sample $Z^{(i)} = W^T x^{(i)}$;
6) Get $D' = (z^{(1)}, z^{(2)}, \ldots, z^{(m)})$.

3.3. **Calculation of statistical indicators and principles of multivariate process monitoring.**

(1) T^2 statistics

Hotelling-T^2 statistics measure the variation of variables in the principal metric space.

$$T^2 = x^T P / pTx \leq T_\alpha^2 \tag{14}$$

$\wedge = diag\{\lambda_1, \ldots\ldots, \lambda_A\}, T_\alpha^2$ stands for T^2 limit of control with a confidence of α. Assuming that the samples under normal operation of the process obey multivariate normal distribution, the control limits can be calculated by following the equation.

$$T_a^2 = \frac{A(n^2-1)}{n(n-A)} F_{A,n-A;a} \tag{15}$$

$F_{A,n-A;a}$ is the critical value of F distribution with A and $N-A$ degrees of freedom and the confidence interval of α.

(2) Square prediction error index

The SPE metric measures the change in the projection of the sample vector in the residual space.

$$SPE = \| \left(I - PP^T\right) x \|^2 \leq \delta_a^2 \tag{16}$$

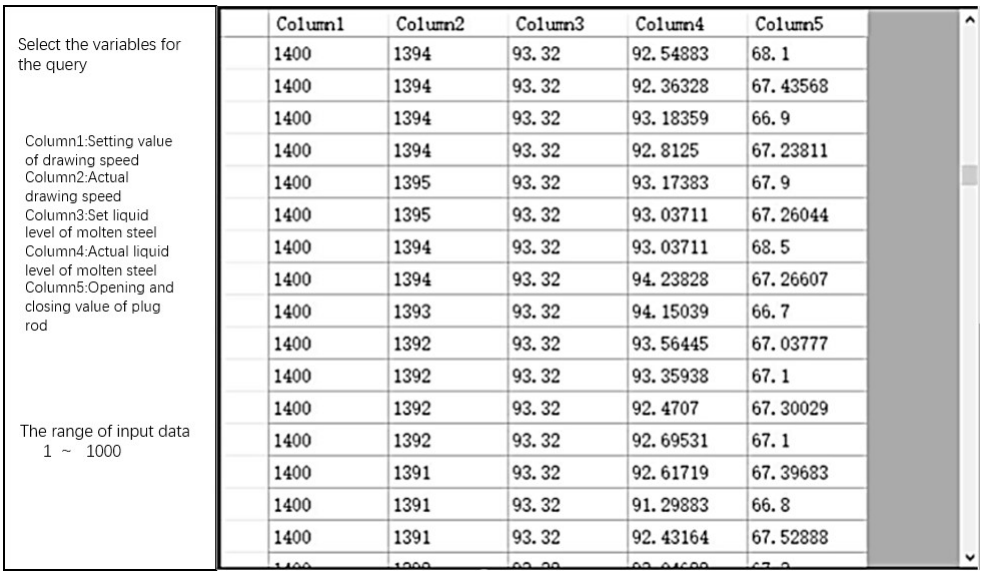
δ_a^2 represents the control limit with the confidence interval of α. The process is considered normal when the SPE is within the limits of control. The calculation formula of the control limit is

$$\delta_a^2 = \left(\frac{c_a \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0^2 (h_0 - 1)}{\theta_1^2} \right)^{\frac{1}{h_0}} \tag{17}$$

Where $\theta_i = \sum_{j=A+1}^{m} \lambda_j^i (i = 1, 2, 3), h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_1^2}, \lambda_j$ is the eigenvalue of the covariance matrix of X, C_a is the threshold of the standard normal distribution at the confidence level, A is the number of principal elements of the PCA model, and m is the dimension of sample X. If $SPE \leq \delta_a^2$ occurs, then this part of the metallurgical process is normal in terms of the pull speed value process. When a fault occurs, the projections of the faulty problem part and the normal part are combined into a process process problem sample vector X. This makes $SPE > \delta_a^2, \delta_a^2$ denote the control limit of SPE, and this expression is similar to the third order moment distribution of SPE.

(3) Model Building

We divide the multivariate process monitoring modeling process into four steps: model building, data selection, data processing, and model updating. The variable names in the data table are first defined and displayed on the development interface, and then the production operator selects the desired variables and the corresponding data ranges to build the model. For this experiment, we select the five process variables in the table shown in Figure 3 for analysis. After determining the data range of the model, the operator saves the information of the created model in the form of an XML file to a specified address for the next call to the model information. Using c# programming software to import the above multivariate process monitoring model based on the improved PCA algorithm, and develop the normality curve module, and then use Minitab and other software to test the normality of the data of the above five variables again (Stat-Basic Statislics-Normality Test), resulting in data P = 0.46, which is greater than 0.05. This means that the data are normally distributed. If the operator chooses not to introduce new variables in the



Column1	Column2	Column3	Column4	Column5
1400	1394	93.32	92.54883	68.1
1400	1394	93.32	92.36328	67.43568
1400	1394	93.32	93.18359	66.9
1400	1394	93.32	92.8125	67.23811
1400	1395	93.32	93.17383	67.9
1400	1395	93.32	93.03711	67.26044
1400	1394	93.32	93.03711	68.5
1400	1394	93.32	94.23828	67.26607
1400	1393	93.32	94.15039	66.7
1400	1392	93.32	93.56445	67.03777
1400	1392	93.32	93.35938	67.1
1400	1392	93.32	92.4707	67.30029
1400	1392	93.32	92.69531	67.1
1400	1391	93.32	92.61719	67.39683
1400	1391	93.32	91.29883	66.8
1400	1391	93.32	92.43164	67.52888

FIGURE 3. Range of data for analysis of five process variables for the selected example

data processing interface, the data processing interface will directly call the model created by himself, and the following results will be obtained for the training sample projection data processing in Figure 4, as well as the results of the indicator analysis in Figure 5 and Figure 6. We monitored T^2 to determine whether the data were abnormal or not,

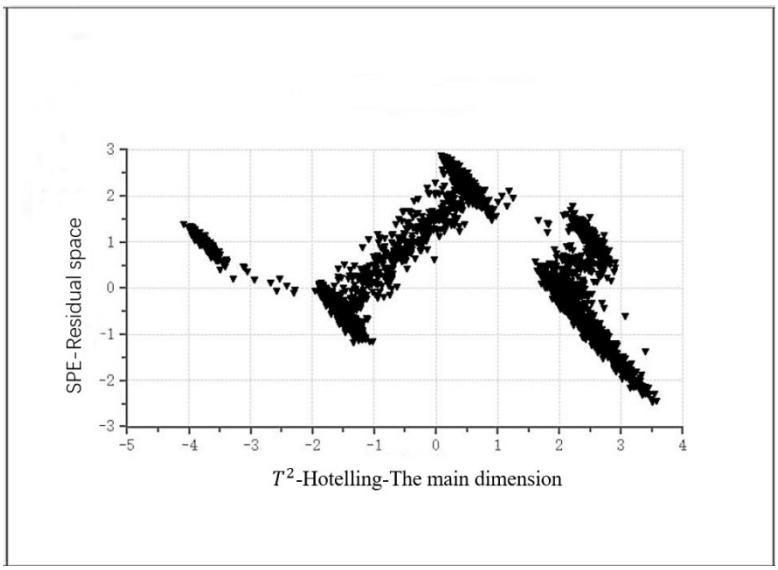


FIGURE 4. Two-dimensional scatter of PCA projection data for multivariate monitoring data processing interface

and judged whether a fault occurred according to the SPE index, and combined with the PCA principal component analysis method to construct the MSPM model. After testing, as shown in Figure 5, the five variables of pulling speed test value, actual pulling speed value, level setting value, actual level value, and plug bar opening and closing value were stable in the first 1000 sample points of the principal element projection T^2, and rarely exceeded the upper limit value. As shown in Figure 6, the SPE has a larger variation and easily exceeds the upper limit. After theoretical analysis, this is because most of

the data of these five variables tend to be normally distributed in the steel metallurgical process, and due to a large amount of sample size data, the T^2 control limit T_α^2 not only tends to be F-distributed but also will tend to be more chi-squared. This leads to the possibility that the control limits of the SPE indicators for theoretical process monitoring are not realistic, but we have included the examination of the local sample range in writing the entire module code, which can be used to observe whether the specific fault variable samples are consistent with the SPE and T^2 alarms through statistical test plots.

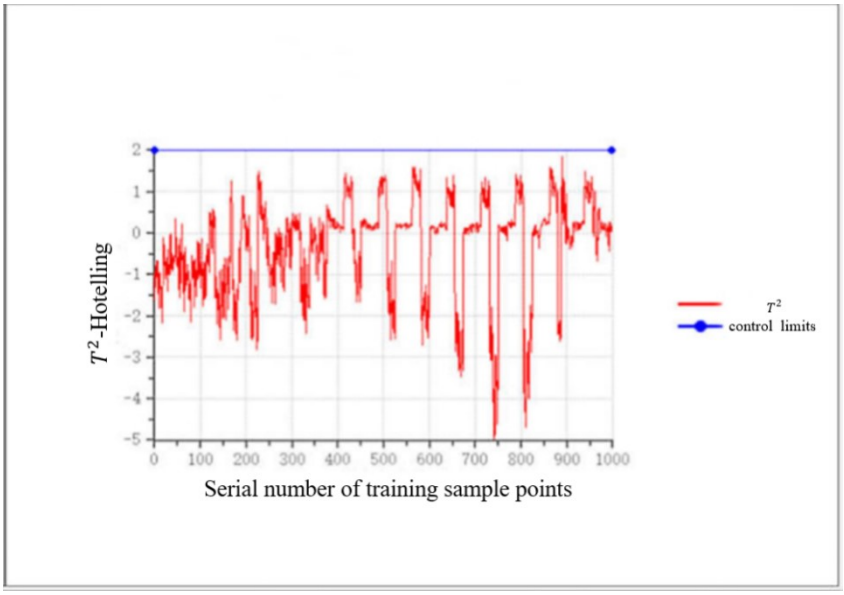


FIGURE 5. T^2 control chart of test sample data for multivariate monitoring data processing interface

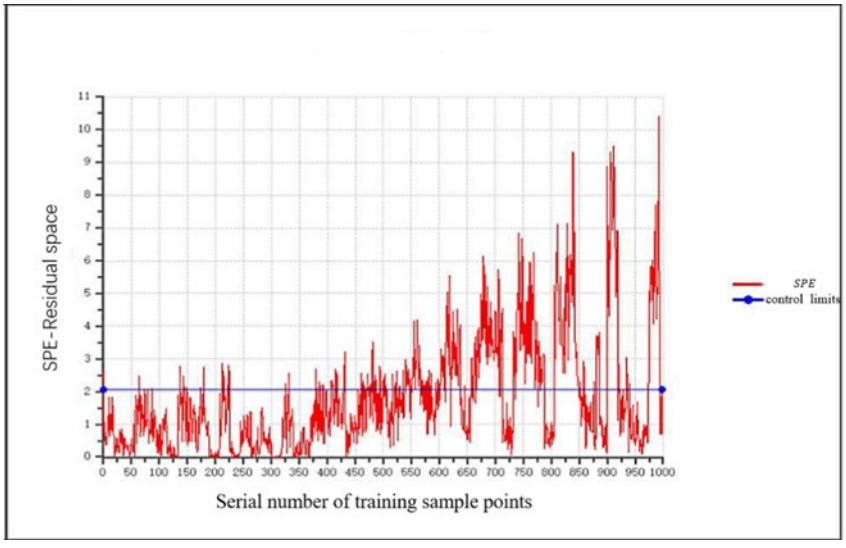


FIGURE 6. SPE control chart of test sample data for multivariate monitoring data processing interface

As shown in Figure 7, the sample data with problematic SPE statistics are 11, 20, and 31, respectively, while the variables with T^2 overruns on the left side after local T_α^2 control limit adjustment are also: 11, 20, and 31, in line with the practical test accuracy. Figure

8 shows the respective contributions of the two indicators in this projection, while some papers propose the use of a composite indicator to combine the two indicators, which is not in line with the actual operation of the production staff and is not conducive to their judgment, so it is not used.

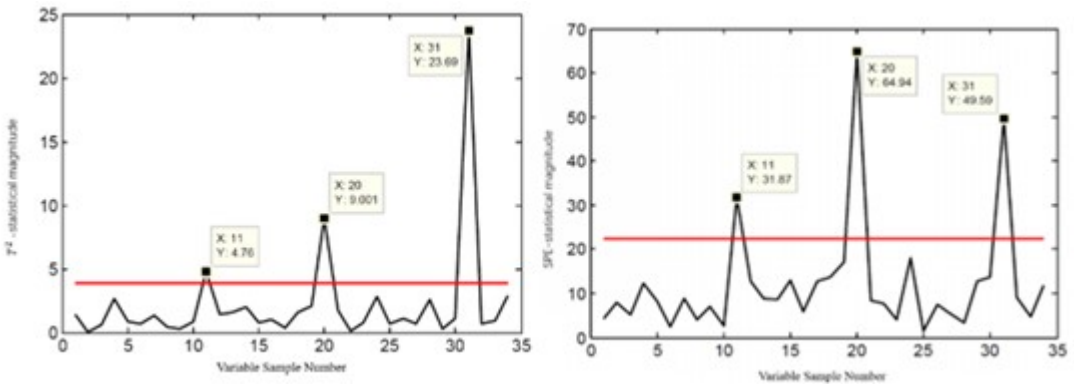


FIGURE 7. Comparison of SPE and T^2 statistic consistency for training samples

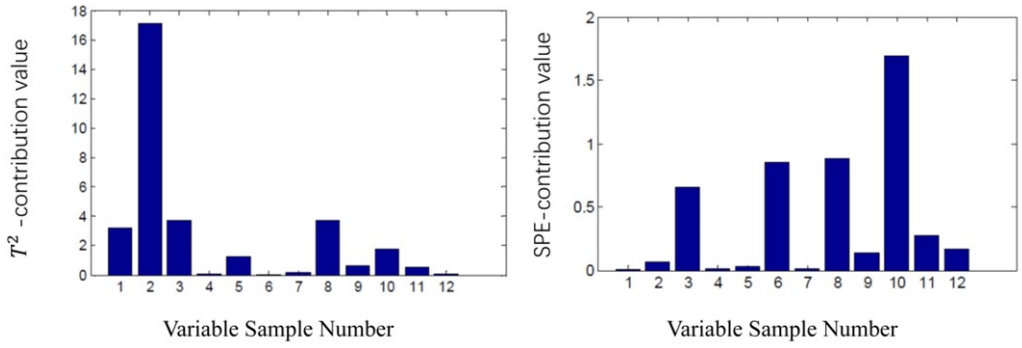


FIGURE 8. Contribution values of T^2 and SPE statistics of training samples

From the data obtained from the production operation and management department, the problem of quality monitoring variation of each batch of steel strip may be related to the influence of special differential working conditions such as duty personnel and process procedures, while the previous univariate monitoring obviously had many other indicators that chose to be ignored when a quality problem occurred, and therefore could not detect the problem of its correlation with important process variables. The improved PCA algorithm model provides a multivariate monitoring approach, which well improves the detection accuracy of the model and effectively solves the problem.

(4) Model Validation

To study and verify the effect of the algorithm, we collected records of batches with warpage in steel strip production in a quarter of B steel enterprise and analyzed the data. In the process of analysis and training, 200 faulty batches, as well as 100 normal batches, were collected as the test data set, and 14 process values such as mill speed difference, live sleeve set amount, and temperature difference for each batch in this data set were collected as training data set, as shown in Table 4. The previous univariate fault detection rate FDR as well as the fault false alarm rate FAR are compared, and the kernel parameters of the conventional PCA algorithm for unilateral quantities as well

as the improved MSPM algorithm are unit normalized, and the kernel density is used to estimate the control limits, i.e., the empirical values are used to estimate the specific values. The corresponding values of the Fault Fault Reporting Rate are given in Table 5.

TABLE 4. The 14 process value indicators collected

	Process value	Variable name	Variable number
procedure parameter	1zone∼3zone	Temperature value of heating zone	3
	1#∼4#	Looper height	4
	1#∼8#frame speed	frame speed difference	7

TABLE 5. Univariate detection statistics of warped skin and detection rate of MSPM software

Algorithm	Variable indexes	Texting indexes	The rate of checkout(%)
MPSM Algorithm	SPE	FAR	0.0
		FDR	18.5
	T^2	FAR	0.0
		FDR	12.0
MICA Algorithm	I^2	FAR	6.0
		FDR	11.5
	T^2	FAR	7.0
		FDR	3.0

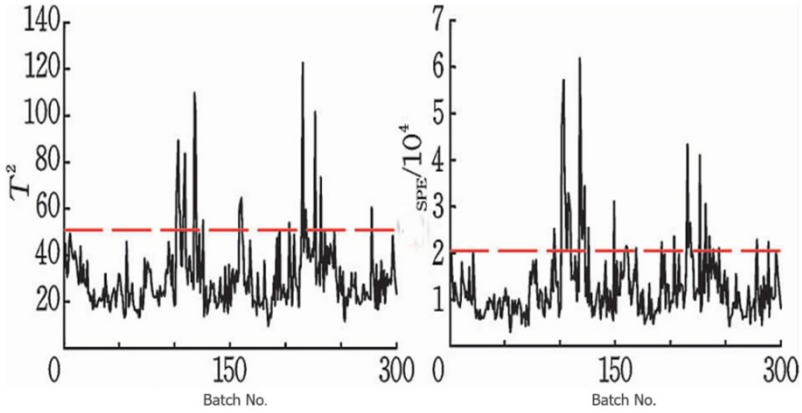


FIGURE 9. T^2 versus SPE data in MPSM algorithm (red line is alarm threshold)

(5) Comparison of the improved MPSM algorithm with the traditional multiplexed independent component analysis MICA algorithm

Figures 9 and 10 above demonstrate that the fault detection rate FDR of MSPM reaches 18.5%, while the detection result of traditional MICA is only a modest 3%. MPSM is obviously better than the traditional monitoring means, and the false alarm rate has also improved significantly, with 0% false alarms. Because there is a multivariate projection in the monitoring process of the analysis of the operational process, so after getting the fault alarm data, can not immediately determine which variable is the problem, the 14

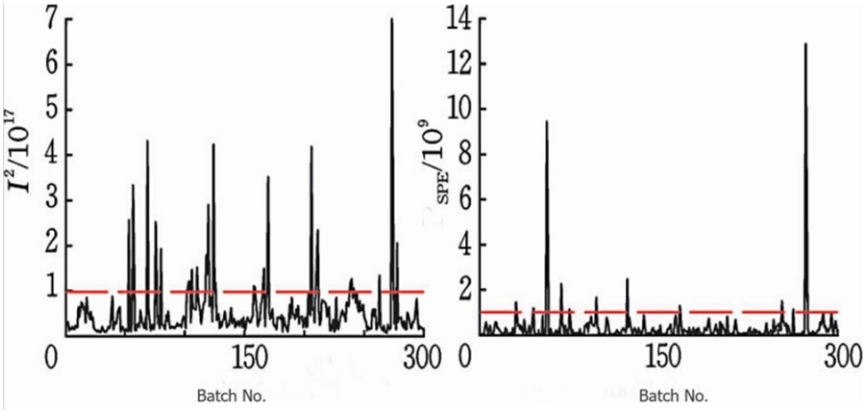


FIGURE 10. T^2 versus SPE data in MICA algorithm (red line is alarm threshold)

variables there are multivariate coupling role through Granger causality analysis, the variable at the beginning of the arrow triggered the change of the variable at the end of the termination, the number of times each variable triggered other variables in descending order ranking, locating the root cause as shown in Figure 11 of.

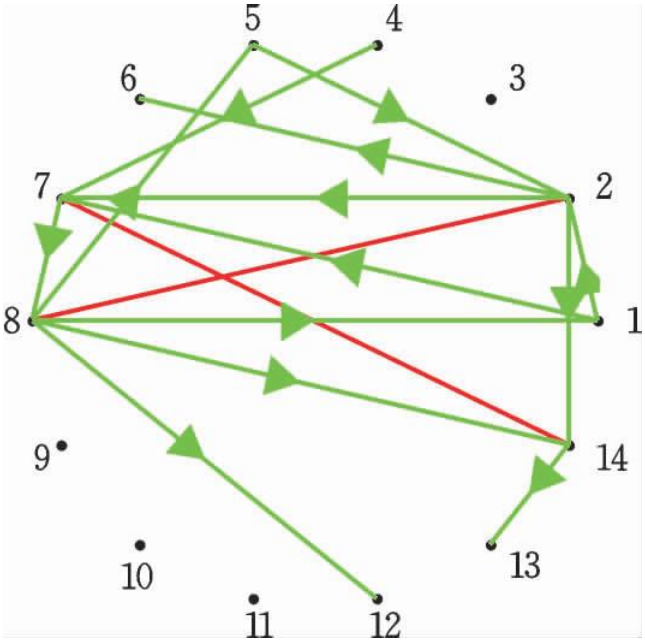


FIGURE 11. Interaction of 14 variables

For this situation, we can analyze by Granger analysis, Granger mentioned that his theory has a very limited scope of application and cannot be used as a real method to obtain the final results, Tiwari, and Aviral. [20] used a modified asymmetric Granger analysis to solve the relationship of the actual energy consumption, so this paper considers the complexity of the actual production process, the product anomalies are often influenced by multiple variables or coupled between multiple variables. Obviously, if only one failure variable is identified as the root cause of the failure, it cannot meet the actual production needs. Therefore, Granger analysis should be improved here to adapt metallurgical process variables for analysis, and the results of statistical analysis should be compared with Granger causality analysis to analyze the root cause of failure in detecting faulty

batches for further analysis. Figure 12 shows the statistical appearance of the 14 process variables that produce the impact of the abnormal warpage situation against Figure 11.

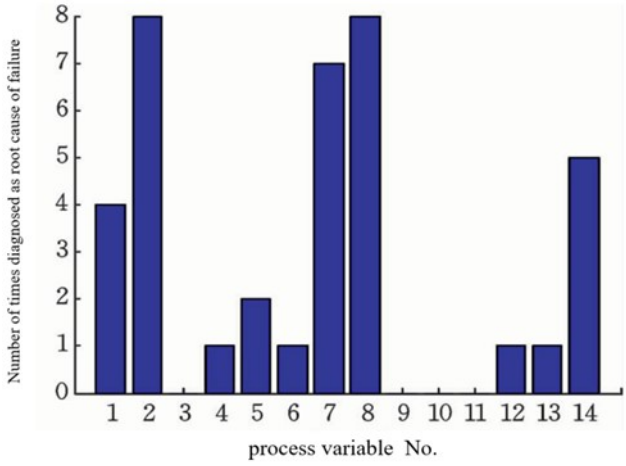


FIGURE 12. Statistics of 14 process variables producing effects on abnormal warpage

The statistics show that the most influential variables are 2, 7, and 8. The sample size of this experiment was then expanded to 100 data, as shown in Figure 13.

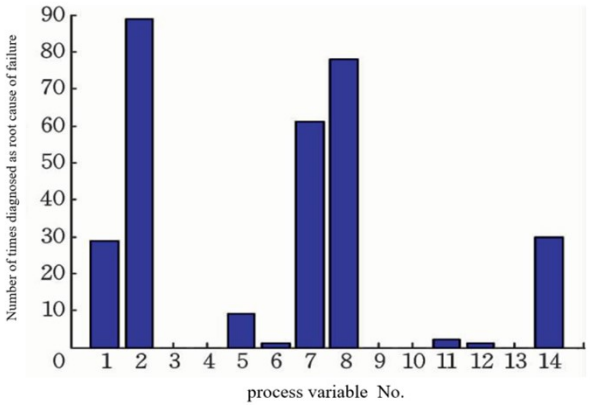


FIGURE 13. 100 statistics of 14 process variables producing effects on abnormal warping cases

The results show that 2, 7, and 8 are still the most influential variables, and 2, 7, and 8 correspond to the temperature difference of the average heat section, 2# live sleeve height, and 7# rack speed difference, respectively. This shows that the MSPM algorithm can monitor the data abnormality more accurately, so as to alarm the fault, but also through Granger analysis and statistical analysis combined to further determine the specific variables that cause the abnormality, so as to reduce the false alarm rate.

4. **Conclusion.** The target set by the China Industry and Information Technology Bureau by 2020 is to compress crude steel production capacity by 100-150 million tons, while according to the research in this paper, it is expected that China's crude steel consumption will also decline by about 3% per year after 2020, and the current oversupply situation in the steel industry will not change shortly. In this context, the concept of whole-process quality management is proposed and practiced, forcing enterprises to continuously upgrade their quality management level and improve the core competitiveness of products is

the most urgent need at present. At present, the Internet technology continues to break through, injecting new momentum for industrial development, Chinese steel enterprises to achieve the curve to overtake, need to be in the process of information technology will be customized data mining big data analysis technology and production process for deep integration, thus helping China in the process of production line automation information technology to develop and master more core independent technology.

The main problems we want to solve now are the accurate implementation of customer requirements, multi-module collaboration to achieve in-event control of product quality and consistent quality, one-click traceability of defects, rapid adjustment and optimization of process parameters with the help of AI technology and visualization technology such as digital twin, and the provision of customized business to achieve collaborative manufacturing of quality, process, and equipment operation. Nowadays, the generalized Chinese steel mill informatization architecture is mainly a five-layer hierarchical quality information system architecture. The current idea is to make it a flat quality information system architecture, which can meet both the production requirements in management and the quality needs in production, in line with the currently recognized more mainstream trend of steel mill informatization. The contribution of this paper is to narrow down the selection range of influencing process variables through the improved Z4.5 algorithm (the limitation of the algorithm does not allow for large data processing), to apply the MSPM algorithm in chemical production to metallurgical process data analysis and compare the algorithm performance with the MICA algorithm, to monitor multiple variables of production data simultaneously, and finally to accurately diagnose anomalies through the improved Granger analysis of comparative statistical results. The anomalies are diagnosed accurately and a set of monitoring points is provided for the engineers of the relevant product lines to verify. This paper also combines experimental kernel entropy component analysis (KECA) with DISSIM, a statistical process monitoring method for process data variability analysis (DISSIM) proposed by KANO et al. [21] to replace the combination of the Z4.5-MSPM algorithm to test the actual effect.

Because the experimental conditions in this paper are very demanding, so the research workers conducting related aspects of the published literature is very small, and many of them declare it as a patent, or a means of industry profitability, for smart manufacturing Industry 4.0 many scholars who are not practicing on the front line consider the further upgrade of the MES system [22]. This paper tends to integrate many of the original upstream and downstream factories in the steel manufacturing industry, and inject the original design MES and ERP concepts into the full computer control system by integrating the concept of the whole product life cycle, and in the future, by establishing a series of algorithmic models, data analysis and data The future will be through the establishment of this series of algorithmic models, and data analysis and data visualization, to achieve the digital twin, which can be a large amount of today's data for rapid cloud computing processing, so as to guide production. Consider network security issues for enterprise cloud data, enterprises can refer to key Agreement Protocol in Cloud-based Smart Healthcare Environments proposed by Wu et al. [23].

The general idea and methodology of this paper for plant data processing can be of guidance and reference value in continuous production process continuous optimization, yield improvement, quality improvement, equipment warning, energy optimization management, etc.

REFERENCES

[1] M. Hermann, T. Pentek, B. Otto, "Design principles for industrie 4.0 scenarios," *in 49th Hawaii International Conference on System Sciences (HICSS),* 2016,pp. 3928-3937

[2] K. Wang; P. Liu; A. Zhao; Q. Zhang; L. Wang; Y. Xue; X. Gao; D. Gao, "Development and Application of MES Based on Cloud Platform for Steel Structure Enterprises," *in IEEE International Conference on Industrial Engineering and Engineering Management .IEEM* , 2019,pp. 521-525.
[3] S.Jaskó, A.Skrop, T. Holczinger, T. Chován, J. Abonyi, "Development of manufacturing execution systems in accordance with Industry 4.0 requirements: A review of standard- and ontology-based methodologies and tools," *Computers in Industry,* vol. 123, 103300, 2020.
[4] Y. Chen, "Industrial information integration—A literature review 2006–2015," *Journal of Industrial Information Integration,* vol. 2, pp.30-64, 2016.
[5] M.Kano, Y. Nakagawa, "Data-based process monitoring, process control, and quality improvement: Recent developments and applications in steel industry," *Computers and Chemical Engineering,* vol. 32, pp.12–24, 2008.
[6] A.-K.Pani, H.-K. Mohanta, "Online monitoring of cement clinker quality using multivariate statistics and Takagi-Sugeno fuzzy-inference technique," *Control Engineering Practice,* vol. 57, pp. 1-17, 2016.
[7] K.-S.Stadler, J.Poland, E.Gallestey, "Model predictive control of a rotary cement kiln," *Control Engineering Practice,* vol. 19,pp. 1-9, 2011
[8] J.Zhang, G.Rizzoni, A.-C.Arenas, A. Amodio, B.-A.Guvenc, "Model-based diagnosis and fault tolerant control for ensuring torque functional safety of pedal-by-wire systems," *Control Engineering Practice,* vol. 61 , pp. 255-269, 2017.
[9] S. -J. Qin, "Survey on data-driven industrial process monitoring and diagnosis," *. Annual Reviews in Control .* vol. 36, pp. 220–234, 2012.
[10] S. -J. Qin. "Statistical process monitoring: Basics and beyond," *Journal of Chemometrics,* vol. 17 , no. 8-9, pp. 480-502, 2003.
[11] Z. Li, U. Kruger, L. Xie, A. Almansoori, H. Su. "Adaptive KPCA modeling of nonlinear systems," *IEEE Transactions on Signal Processing,* vol. 63, no. 9, pp. 2364-2376, 2015.
[12] L. Xie, J. Zeng, U. Kruger, X. Wang, J. Geluk. "Fault detection in dynamic systems using the Kullback–Leibler divergence," *Control Engineering Practice,* vol. 43, pp. 39-48, 2015.
[13] A .Sepúlveda, J .-A. Nachlas, "A simulation approach to multivariate quality control," *Computers & industrial engineering,* vol. 33 no.1, pp. 113-116, 1997.
[14] A.Bakdi,A.Kouadri ,A.Bensmail, "Fault detection and diagnosis in a cement rotary kiln using PCA with EWMA-based adaptive threshold monitoring scheme," *Control Engineering Practice,* vol. 66, pp. 64 -75, 2017.
[15] L.-H.Chiang, E.-L.Russell, R.-D.Braatz ," Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis," *Control Engineering Practice,* vol. 50, no.2, pp.243-252, 2000.
[16] S. Joe Qin. "Statistical process monitoring: Basics and beyond," *Journal of Chemometrics,* vol.17,no. 8–9, pp. 480-502, 2003.
[17] V. Venkatasubramanian, R. Rengaswamy, S.N. Kavuri, K. Yin. "A review of process fault detection and diagnosis: Part III: Process history based methods," *Computers & Chemical Engineering,* vol.27 vol. 3, pp. 327-346, 2003.
[18] H. Wang, T. Wang, Y. Zhou, L. Zhou, H. Li. "Information classification algorithm based on decision tree optimization," *Computers & Chemical Engineering* vol 22, pp .7559–7568, 2019.
[19] Y.-S.Mu, X.-D. Liu, Z.-H. Yang, X.-L. Liu, "A parallel C4.5 decision tree algorithm based on MapReduce," *Concurrency and Computation-practice & Experience.* 2017.
[20] Tiwari, Aviral ," *The asymmetric Granger-causality analysis between energy consumption and income in the United States." Renewable and Sustainable Energy Reviews .* vol. 36.pp.362-369.
[21] M. Kano,H. Ohno,S. Hasebe,I. Hashimoto , " Process-Monitoring Based on Dissimilarity of Time ScricsData." *Kagaku Kogaku Ronbunshu.* vol. 25 no.6,pp.1004-1009,1999.
[22] A. Moeuf , R. Pellerin, "The industrial management of SMEs in the era of Industry 4.0. International, " *Journal of Production Research* vol .56 no.3, pp. 79-87,2018.
[23] T.-Y. Wu, L. Yang, J.-N. Luo, J. M.-T. Wu. "A provably secure authentication and key agreement protocol in cloud-based smart healthcare environments," *Security and Communication Networks,* vol. 2021, 2299632, 2021.