J. Iron Steel Res. Int. (2023) 30:915–925
https://doi.org/10.1007/s42243-023-00934-4

ORIGINAL PAPER

Hot metal quality monitoring system based on big data and machine learning

Ran Liu[1] · Zhi-feng Zhang[1] · Xin Li[1] · Xiao-jie Liu[1] · Hong-yang Li[1] · Xiang-ping Bu[2] · Jun Zhao[3] · Qing Lyu[1]

Received: 29 September 2022 / Revised: 15 November 2022 / Accepted: 20 November 2022 / Published online: 3 May 2023

Abstract

The system of hot metal quality monitoring was established based on big data and machine learning using the real-time production data of a steel enterprise in China. A working method that combines big data technology with process theory was proposed for the characteristics of blast furnace production data. After the data have been comprehensively processed, the independent variables that affect the target parameters are selected by using the method of multivariate feature selection. The use of this method not only ensures the interpretability of the input variables, but also improves the accuracy of the machine learning process and is more easily accepted by enterprises. For timely guidance on production, specific evaluation rules are established for the key quality that affects the quality of hot metal on the basis of completed predictions work and uses computer technology to build a quality monitoring system for hot metal. The online results show that the hot metal quality monitoring system established by relying on big data and machine learning operates stably on site, and has good guiding significance for production.

Keywords Hot metal · Big data · Machine learning · Quality monitoring · Feature engineering

1 Introduction

The integration of Germany's "Industry 4.0" and "Made in China 2025" brings a historical opportunity for China's traditional industry to transform to intelligentization and informatization [1]. At the same time, severe climate change forces countries around the world to launch low-carbon economic plans to cope with the industrial development trend of the new era. Worldwide, the steel industry is beginning to transform. For example, the intelligentization levels of BAOWU, BRS, Posco, etc., are in the leading position in the steel industry [2–5]. Relatively, the intelligent development of Chinese iron and steel enterprises is relatively slow, and it has gradually become the industry with the largest carbon emission among the 31 manufacturing categories [6]. In this context, the transformation of China's blast furnace (BF) smelting process to intelligentization has become an irreversible trend.

The quality of hot metal in BF is an important basis for achieving the goal of "high yield, high quality and low consumption". However, BF ironmaking is a very complex process. Although sensors can collect a lot of production data, the data have the characteristics of nonlinearity, time delay and strong coupling because of the wide range of data sources and the complex relationship between data [7]. These reasons indirectly lead to the slow development of intelligent ironmaking technology in BF. With the advent of the era of Industry 4.0, many researchers have begun to explore how to use computer technology and big data technology to predict key indicators, which affect the quality of hot metal. They want to improve the quality of hot metal by accurately predicting the change trend of relevant indicators.

From the current research status, the research content of most scholars is mainly to predict a key index. Martin et al.

✉ Xiao-jie Liu
xiaojie19851003@163.com

[1] College of Metallurgy & Energy, North China University of Science and Technology, Tangshan 063009, Hebei, China

[2] Hangzhou Pailie Technology Co., Ltd., Hangzhou 310000, Zhejiang, China

[3] Hebei Iron and Steel Group, Tangshan Iron and Steel Co., Ltd., Tangshan 063009, Hebei, China

[8] used the thermal simulation prediction model based on fuzzy tool to predict the hot metal temperature (HMT) of BF. In order to improve the prediction accuracy of HMT, Zhao et al. [9] used the least square vector machine based on chaotic particle swarm optimization as the HMT prediction model, and got a good prediction effect. Diaz et al. [10] improved the multivariate adaptive regression spline model to predict the HMT of BF. It is proved that the average absolute error is kept within 11.2 °C by testing and verifying the production data of the steel plant for one year. Using the dynamic relationship method between input process variables and output variables based on attentional mechanism module, Jiang et al. [11] proposed a dynamic attentional deep migration network and realized the online prediction of silicon content in hot metal. Wang et al. [12] used the dynamic neural network based on principal component analysis to predict the silicon content of hot metal. The experimental results show that the prediction accuracy of the model reaches 89.12%. Diniz et al. [13] used a nonlinear autoregressive network to implement an 8-h prediction model for hot metal and silicon content with an acceptable error range. Of course, the comprehensive evaluation of hot metal quality is not without research. And other reference uses different modeling methods to achieve multi-angle prediction of hot metal quality [14–16].

Although many scholars have done a lot of research work in predicting the quality of hot metal in BF, there are still some areas that can be improved. (1) Accurate data connection. There is serious time lag in the process of BF production [17]. It may be wrong to use the current hot metal quality index to correspond to the current operating parameters, because a certain reaction process is required for the iron-containing raw material to be transformed into hot metal. Therefore, the data processing should be delayed according to the actual production level of the enterprise. (2) Feature selection diversity. In the modeling stage, the method of selecting characteristic variables by traditional correlation analysis is not necessarily explicable from the perspective of manufacturing process and may omit some important parameters that remain unchanged for a long time [18]. In order to build an industrial machine learning model suitable for enterprise production, big data technology and process theory methods should be fully integrated in the feature engineering stage. This method can not only avoid missing the operation index which is of special concern to the production site, but also improve the interpretability of input variables and model accuracy. (3) Comprehensive prediction indicators. The quality of hot metal is a comprehensive index affected by many factors [19]. The traditional unitary prediction can only explain part of the situation and cannot reflect the complex situation of hot metal quality and BF production. Therefore, it is more persuasive and credible to be able to simultaneously predict the indicators that enterprises are concerned about. (4) Industrialization of research results. The lack of a complete system and field application is the biggest flaw at present. Researchers should apply the model to the actual production after the completion of the research [20]. As can be concluded from the above, on the one hand, opinions (1) and (2) provide a new method of data processing and analysis for iron quality prediction studies, which is the basis for improving the accuracy of prediction. On the other hand, opinions (3) and (4) meet the needs of the staff and are of great significance in guiding production in a timely manner.

Aiming at the deficiencies in the current research process of hot metal quality, this paper establishes a hot metal quality monitoring system based on big data and machine learning. The main distribution of the article is as follows.

2 Data collection and processing

In order to ensure the accuracy of the model, the data used in this study are the historical production data of a steel enterprise in China (This iron and steel enterprise mainly smelts vanadium and titanium ore. Therefore, the evaluation indicators of hot metal quality are HMT, the silicon and titanium content ([Si + Ti]) and the sulfur content ([S])). For a long time in the past, the BF ironmaking process has accumulated massive production data. Because these data have the characteristics of different storage locations and complex data structure, it is necessary to sort out these data before the research begins.

2.1 Data sources

The data used in the study were all from the factory database files. As shown in Fig. 1, all data of BF ironmaking process are stored in Oracle 10.2, SQL Server 2008 and Wonderware Historian 2014 databases. Data records and storage frequencies of each database are different. Wonderware Historian database stores real-time monitoring data of BF production equipment; SQL Server database stores the operation state parameters generated in the BF ironmaking process; Oracle database stores the performance of raw materials and test results of BF products. Historical production data from February to August in 2021 were selected from the three databases mentioned above for the study. There are 544 types data and nearly three million data items.

2.2 Data pre-processing

As we all know, the quality of data directly affects the prediction ability and generalization performance of the

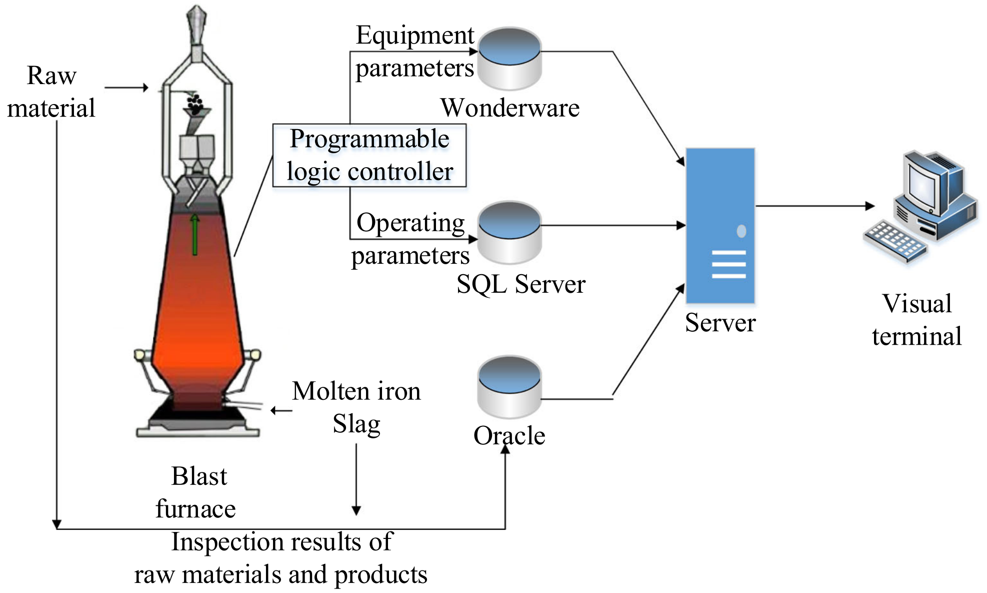


Fig. 1 Distribution of data sources

model. Therefore, data pre-processing is particularly important as the first step in the whole data processing process. It mainly aims at data integrity, credibility, timing and standardization. After full analysis and processing, standard, clean and continuous data are obtained to lay a solid foundation for subsequent work. The production parameters need to be delayed by one production cycle to correspond to the predicted target before data processing, because the process of turning the raw material into hot metal is one production cycle (8 h).

2.2.1 Data integrity processing

Integrity processing is mainly to fill the missing values of data, but the filling method depends on the situation. First of all, the data missing rate of each parameter needs to be counted. We will delete the fields whose data missing rate is greater than 50%, and then analyze and process the missing types of the remaining parameters one by one. The analysis shows that there are two reasons for missing data in the remaining parameters: BF condition maintenance and equipment failure. Furnace condition overhaul will cause large area data loss. Even filling in these data would lose authenticity, so that we are going to delete them. Data loss caused by device failure mainly includes timing loss and correlation loss. Temporal absence can be divided into short-term absence and long-term absence. If the missing value type is short-time missing, it can be filled with the previous value or linear interpolation method. However, the data missing for a long time needs to be analyzed in detail according to big data technology and process theory, and then filled with simple models or theoretical formulas. Some data that cannot be populated out of the dataset should be temporarily moved rather than be deleted. If we need these data to support the model, we can analyze them. Data integrity processing is shown in Fig. 2.

2.2.2 Data credibility processing

There are many factors that affect the data to maintain high or low fluctuation in a certain period of time, e.g., production environment fluctuation, emergency operation regulation, sampling anomaly, etc. For example, the number of batches of material, blast volume, hot air pressure and other related parameters will be successively reduced

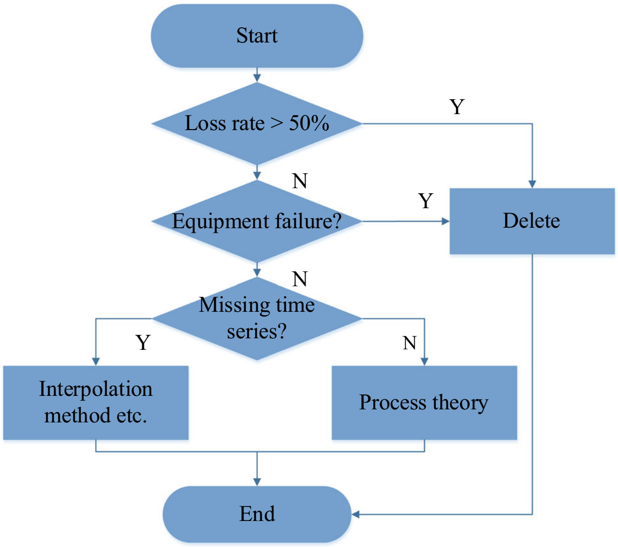


Fig. 2 Data integrity process flow

before the rest maintenance. An emergency reduction in blast volume can cause data to suddenly drop and then rise again. In these cases, traditional boxplot processing will exclude these data as outliers, resulting in the loss of valuable data. Therefore, we combine the production level of the enterprise and the process theory to deal with the outliers. The method of global check first and then local check of process theory is used to analyze the abnormal points screened out by the boxplot. In addition, when encountering some uncertain situations, it is necessary to deduce the production status according to the parameter changes and judge whether the data belong to the abnormal value.

2.2.3 Data timeline processing

It can be seen from Sect. 2.1 that the data of this factory are mainly from the three databases mentioned above, but their sampling frequencies are different. The data frequency in Wonderware database is second. The data frequency in the SQL Server database is uncertain. It contains not only some data from Wonderware and Oracle databases, but also some manually entered data. The data in the Oracle database are all manually entered, and these data are mainly laboratory results. Since the purpose of this study is to predict the hourly frequency variation trend of hot metal quality, we convert the data from Wonderware and SQL Server databases into hourly data. Then, the batch number is used as the standard to match the data in the other two databases with the data in Oracle.

2.2.4 Data standardization processing

Because of the different properties of each parameter in the dataset, it usually has different dimensional levels. If the original data are directly used for analysis, high-value fields may weaken the contribution of low-value fields. Therefore, it is necessary to standardize the data with certain rules and eliminate the limitation of data units. The Z-Score normalization rule was selected for processing in the study, and the following changes were made to the data columns $x_1, x_2, x_3, \ldots, x_n$ so that the new series $y_1, y_2, y_3, \ldots, y_n$ was scaled to [0, 1] interval. In Eq. (1), $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the overall mean of the data column samples and $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ is the overall standard deviation of the data column samples (n is the sample size and x_i and y_i is the sample individual value).

$$y_i = \frac{x_i - \overline{x}}{s} \tag{1}$$

2.3 Feature engineering

There are still 526 fields of BF ironmaking parameters after data pretreatment. If all these parameters are modeled as input variables, not only dimension disaster but also difficult learning and high time complexity problems will occur. Therefore, data must go through feature extraction before further work. In this study, feature extraction is divided into two steps: data reduction and feature selection. In order to reduce the workload of feature selection stage, data dimension reduction stage adopts filtering, principal component analysis (PCA) and pearson correlation coefficient (PCC) to achieve data dimension reduction after preprocessing. In the feature selection stage, we will use important ranking and Granger causality test to analyze the relationship between parameters after dimensionality reduction and target parameters. Finally, the combined process theory is used to supplement.

2.3.1 Data dimension reduction

Process parameters can be divided into three types: raw fuel parameters, BF status parameters and laboratory results parameters. According to the field production experience, the following parameters are showed: one part is a small part of the key data of raw fuel parameters and inspection results, and the other part is all the data of BF status parameters. They are added to the data dimension reduction table. The data dimension reduction work in this paper applies the following three methods (Table 1).

2.3.1.1 Filter Filter is simple and convenient. It can remove parameters with variance variation of less than 0.1 by calibration threshold. Although filter conditions are too extreme, it can remove some parameters that remain unchanged for a long time. For example, according to the process theory, it can be known that the basicity and other parameters of sinter in the furnace will affect the temperature of hot iron. However, in the actual production of a factory, these parameters do not change much, just fluctuating within a small range. Therefore, it is feasible to use

Table 1 Hierarchical data dimension reduction methods

Step	Name	Purpose
Step 1	Filter	Remove parameter fields with variances less than 0.1
Step 2	PCA	Integrate parameters of same type and realize data derivation
Step 3	PCC	Remove parameters with low correlation with target parameters

the filter as a preliminary dimension reduction method for data.

2.3.1.2 PCA PCA is mainly to recombine a large number of the same type data with a certain correlation. The raw data are replaced with a derived set of unrelated composite indicators. The main steps are as follows:

(1) The original data of m rows and n columns that need dimensionality reduction are transformed into n rows and m columns matrix $\boldsymbol{X}$;
(2) Zero mean is performed on each row of matrix $\boldsymbol{X}$ and each row represents an attribute;
(3) Calculate covariance matrix, eigenvalue and eigenvector;
(4) The eigenvectors are arranged into a matrix according to the corresponding eigenvalues in rows from top to bottom, and then, the first K rows are taken to form a matrix $\boldsymbol{P}$;
(5) The dimension reduction matrix is $\boldsymbol{Y} = \boldsymbol{PX}$.

In this paper, PCA is used to reduce the dimension of parameters of the same type on the premise that the information value is more than 90%. Specific dimension reduction targets and derived parameters are shown in Table 2.

2.3.1.3 PCC This method measures the degree of correlation between two parameter variables X and Y. Therefore, Pearson correlation coefficient was used in this paper to remove the data with low correlation. The correlation coefficient R ranges from –1 to 1. The closer the R approaches 1, the higher the positive correlation. The closer to –1, the higher the negative correlation. At this stage, $|R|> 0.5$ parameters were selected.

2.3.2 Feature selection

Feature selection stage is a key task of data processing. Although there are many feature selection methods based on the level of relevance as a judgment criterion, such methods are not necessarily interpretable in industry. Therefore, the method of multivariate feature selection is proposed in this paper. Firstly, recursive feature elimination with support vector machines (SVM-RFE) model is used to establish the importance ranking of the target parameters. Then, Granger causality test is used to verify the causal relationship between parametric covariable and dependent variable. Finally, the feature selection results are analyzed with process theory.

Table 2 Specific objectives and derived parameters of dimensionality reduction using PCA

Derivative parameter	Dimension reduction target	Explanation
LSJY	GL04_LSJY26275	Static pressure at 26 and 32 m of BF body
	GL04_LSJY32595	
LDWD_D1	GL04_LDWD5160	Temperature of the bottom at different heights
LDWD_D2	GL04_LDWD5700	
	GL04_LDWD6200	
LGWD_1	GL04_LGWD7210	Temperature of hearth at different heights
LGWD_2	GL04_LGWD7700	
	GL04_LGWD8200	
	GL04_LGWD8700	
	GL04_LGWD9200	
	GL04_LGWD10200	
	GL04_LGWD11700	
LQBWD_1	GL04_LQBWD6200	Temperature of cooling wall at different heights
LQBWD_2	GL04_LQBWD8110	
LQBWD_3	GL04_LQBWD10135	
	GL04_LQBWD17187	
	GL04_LQBWD19507	
	GL04_LQBWD21704	
	GL04_LQBWD24145	
	GL04_LQBWD26275	
	GL04_LQBWD29835	
	GL04_LQBWD32595	
	GL04_LQBWD34800	

2.3.2.1 SVM-RFE SVM-RFE was proposed by Guyon [21]. It generates feature rankings by using coefficient vectors. The algorithm process is as follows.

SVM-RFE algorithm

1. Initialize the original feature set $S = \{1, 2, ..., D\}$ and feature sorting set $R = []$;

2. The new training sample X_j is obtained by pairing $l\,(l - 1) / 2$ training sample with different categories (l is the number of samples).

3. The following process is repeated until $S = []$:

(a) l training subsamples X_j ($j = 1, 2, ..., l(l - 1) / 2$) were obtained;
(b) X_j was used to train SVM and coefficient vector w_j (j= 1, 2, ..., l) was obtained;
(c) Calculate the sorting criteria score $c_k = \sum_j w_{jk}^2$ ($k = 1, 2, ..., |S|$);
(d) Find the minimum feature of sorting score P = arg min kc_k;
(e) Update feature set $R = [P, R]$;
(f) Remove the feature $S = S/P$.

The algorithm removes the features with the lowest score in each cycle, and then retrains the remaining features to obtain a new feature ranking. After continuous recursion, a feature ordering table can be obtained. The feature sorting table was used to define several nested subsets to train SVM, and the optimal feature subset was obtained by evaluating the advantages and disadvantages of the subset according to the prediction accuracy. It is worth noting that the top features are the result of a combination of features. Therefore, SVM-RFE algorithm can be used to pick out complementary feature combinations.

2.3.2.2 Granger causality test Granger causality test is defined as follows: in the case of time series, the past information of variable *X* and variable *Y* is used to predict the future change of variable *Y* (variable *X* can help explain the future change of variable *Y*), and independent variable *X* is considered to be the Granger cause of dependent variable *Y* [22]. The characteristics of this test method determine that it can only be applied to the time series data model test. BF ironmaking is a continuous and high-strength process. It has the following characteristics: (1) BF parameters are nonlinear; (2) each parameter is coupled and interacts with the result; (3) the production state of BF is greatly affected by the previous state. The second and third characteristics just meet the conditions of Granger causality test. Considering the limitation of the first feature, we use the multivariate nonlinear Granger causality test for feature selection [23]. It is worth affirming that although the conclusion of the Granger causality test is only "causality" in the statistical sense, it does not hinder the reference value it brings. It is more interpretable if a parameter is selected by both methods.

This paper explains the whole characteristic selection process by taking the HMT as an example. After data dimensionality reduction and feature selection, only 32 parameters remain in the data sample. Although the number of parameters has been reduced, the number of parameters is still too large for machine learning, which easily affects the model learning time and accuracy. Therefore, we perform importance ranking and Granger causality test on these parameters. The specific results are shown in Table 3.

2.3.2.3 Theory of BF process From Table 3, after using the Granger causality test to test the 32 parameters that have completed the importance ranking, there are 14 parameters of which the Granger causality test value is $p > 0.05$. However, these parameters that do not meet the level of significance testing are meaningful in field production. Therefore, these parameters need to be discriminated using BF ironmaking process theory.

Based on the theory of the BF ironmaking process, we have the following conclusions [24]. The change of LFLL and RFYL can quickly change the heat system of the hearth and indirectly affect the change of HMT. LGWD_1 and LGWD_2 are derived parameters generated by the PCA dimension reduction of the temperature values in different directions of the BF hearth. In actual production, the temperature of hot metal is affected by the heat of the hearth. MQLYL can have a direct effect on gas distribution, heat exchange and reduction reaction, which also causes fluctuations in HMT. The amount of PCML can change the gas flow distribution in the furnace and the thermal state distribution in the hearth area. The variation trend of [Si + Ti] and the HMT are positively correlated in a large proportion in the long-term range, and the HMT is often judged by its content in the field production process.

Through cross-validation, it is known that the model works best when there are 20 parameters. First, we select 13 parameters among the top 20 parameters in feature importance, because they not only meet the Granger causality test criteria but also have a satisfactory importance ranking. In addition, seven parameters were selected from the remaining parameters in combination with the process theory. The specific results are shown in Table 4.

3 Establishment of hot metal quality prediction model

In the machine learning phase, we utilize the ensemble learning model to predict the relevant indicators that affect the quality of hot metal. As we all know, ensemble learning models are very popular in well-known data analysis competitions like Kaggle and Driven Data. It mainly combines the prediction results of multiple weak machine learning algorithms, using a certain strategy to combine them and obtaining a satisfactory effect for the researcher [25].

3.1 Base model selection and stacking

Due to the complicated BF production process, although we have performed feature screening, the data volume of the 20 parameters selected in Sect. 2.3.2.3 is also very large (These 20 parameters contain data values of 10,040), so that we use multiple base models to make preliminary predictions on the selected parameters, and use the stacking learning method [26]. This method can effectively combat overfitting by adding regular terms and does not require much adjustment to make it superior. By using the prediction results of the previous stage to perform secondary prediction, the optimal prediction results are obtained. In this paper, five base model individual learners [27–31], random forest (RF), extra trees (ET), AdaBoost, gradient boosting (GDBT) and SVM, will be selected as the first-layer learners, and then, the XGBoost algorithm will be used as the second-layer learners [32]. The XGBoost regular term can not only prevent overfitting, but also adopt a parallel optimization algorithm, which greatly improves the efficiency of the algorithm and provides a basic guarantee for the realization of HMT prediction.

Stacking, as a stacking model that relies on the results of multiple base models, usually outperforms a single strong model. The main idea is as follows (Fig. 3).

3.2 Analysis of model performance

In order to verify whether the characteristics of the ensemble model are better than other prediction models, this paper selects three other types of prediction models for comparison. Table 5 shows the prediction performance results of four different models [33–35]: stacking, gray models (GM), back-propagation network (BP) and long short-term memory (LSTM). For the above models, the article evaluates them from three angles: accuracy rate, root mean square error (RMSE) and modeling time. It is easy to find by comparison that the predicted hit rate of the stacking model is about 10% higher than that of other models. Although the modeling time of the stacking model is slightly longer than that of other models in terms of time complexity, considering that it is a stacking model, its modeling time is acceptable. In summary, from the perspective of model prediction and the development of the hot metal quality prediction and evaluation system, the stacking model has better performance than other models, and is worthy of being used as a hot metal quality prediction model.

After verifying the superiority of the model, we take the test. The results of using the stacking ensemble model to

Table 3 Importance ranking of SVM-RFE features and Granger causality test results between independent variable parameters and HMT

Parameter	Important degree	p	A/R	Definition
QLYC	1	0.3015	A	BF pressure difference
GFDN	2	0.0231	R	Blast momentum
TQXZS	3	0.0256	R	Index burden permeability of BF
LDYL_U	4	0.0482	R	Top pressure
MQGLL	5	0.0452	R	Gas flow rate
LFYL	6	0.0541	A	Pressure of cold air
RFYL	7	0.5581	A	Pressure of hot air
LGWD_1	8	0.5549	A	1-temperature of hearth
LQBWD_1	9	0.0259	R	1-temperature of stave cooler
RFH	10	0.0123	R	Thermal load
FZWD	11	0.0356	R	Seat temperature
RFWD	12	0.0304	R	Temperature of hot air
FYL	13	0.0354	R	Oxy-enriched rate
LQSGSLL	14	0.0639	A	Cooling book inlet water flow
LDWD_D1	15	0.4551	A	1-temperature of BF bottom
PJDW	16	0.0146	R	Average top temperature
LGWD_2	17	0.0615	A	2-temperature of hearth
LQBWD_2	18	0.0048	R	2-temperature of stave cooler
LLRSWD	19	0.0453	R	Theoretical combustion temperature
SJFS	20	0.0284	R	Actual wind speed
MQLYL	21	0.0017	R	Gas utilization
LQSPSLL	22	0.8792	A	Cooling book outlet water flow
LQBWD_3	23	0.0523	A	3-temperature of stave cooler
CO_LYL	24	0.2989	A	CO utilization
LQSGSWD	25	0.0002	R	Cooling water inlet temperature
LQSPSWD	26	0.0001	R	Cooling water outlet temperature
PCML	27	0	R	Injecting coal quantity
LSJY	28	0.6113	A	Static pressure of stack
LDWD_D2	29	0.7361	A	2-temperature of BF bottom
RSLL	30	0.4872	A	Soft water flow
PJYC_Z	31	0.2218	A	Average pressure difference in middle
Si + Ti	32	0.0499	R	Content of [Si + Ti] in hot metal

During Granger causality test, null hypothesis H_0 is set: dependent variable parameter X has no effect on target outcome variable Y. If significance test level $p > 0.05$, null hypothesis is accepted; otherwise, null hypothesis is rejected. "A/R" in Table 3 stands for Accept (A) or Refuse (R)

Table 4 Main characteristic parameters affecting temperature of hot metal

Operating parameter	Temperature parameter	Mixing parameter	Inspection parameter
GFDN	LGWD_1	TQXZS	[Si + Ti]
LDYL_U	LGWD_2	FYL	
MQGLL	LQBWD_1	MQLYL	
LFYL	LQBWD_2		
RFYL	RFH		
SJFS	FZWD		
PCML	RFWD		
	PJDW		
	LLRSWD		

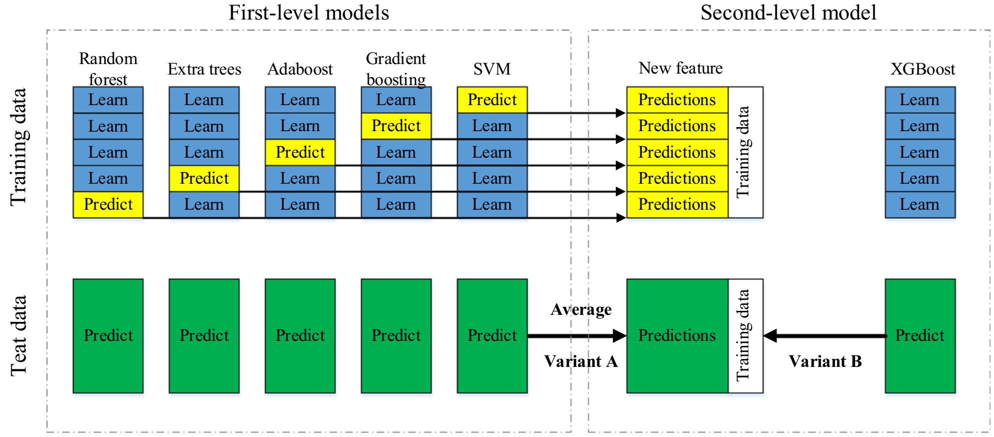


Fig. 3 Structure diagram of stacking model

Table 5 Performance comparison results of stacking model and other models

Evaluation index		Stacking	GM	BP	LSTM
HMT	Accuracy/%	89.6	62.6	76.3	81.5
	RMSE	1.8841	3.3662	2.5846	2.0597
	Modeling time/s	4.2281	2.9831	3.2545	3.5634
[Si + Ti]	Accuracy/%	88.7	65.2	75.8	84.1
	RMSE	0.0212	0.0342	0.0315	0.0272
	Modeling time/s	4.8261	2.9049	3.3842	3.5513
[S]	Accuracy/%	86.3	58.7	79.1	83.9
	RMSE	0.0083	0.0132	0.0093	0.0087
	Modeling time/s	4.5002	3.6673	3.1889	3.4819

predict the HMT are shown in Fig. 4a. It can be clearly seen that although the test set is randomly selected, the error between the test result and the real value is basically kept within ± 5 °C and the accuracy rate reaches 89.6% within the allowable error range. Similarly, we have also predicted [Si + Ti] and [S], and the prediction results have also achieved very good results. As shown in Fig. 4b, the hit rate of [Si + Ti] prediction results is 88.4%. Although there are some obvious data deviation points, its fluctuation is between 0 and 0.03. Figure 4c shows the [S] prediction results. Since the exact value of [S] itself is in the thousandths, predicted hit rate for [S] of the model is significantly reduced. It is worth affirming that the deviation between the predicted value and the actual value is within ± 0.005, and the accuracy rate is 86.3%. From the perspective of technology, the result error and accuracy can be accepted.

4 Hot metal quality monitoring system

In order to monitor the quality of hot metal, this section establishes the quality monitoring system of hot metal of BF based on the prediction of relevant indicators of hot metal.

The system uses the indicators (HMT, [Si + Ti], [S]) concerned in BF production to comprehensively evaluate the quality of hot metal. When the evaluation result exceeds the acceptable range of the factory, the abnormal parameters affecting the evaluation result are found by analyzing the missing items, and the production status is adjusted in time. Based on BF process theory, the system sets the full scores of HMT, [Si + Ti] and [S] as 35, 35 and 30, respectively, and formulates scoring rules suitable for the factory according to the data distribution map of target parameters. Steps are explained using [S] as an example. The inspection and testing department will detect the sulfur content of the furnace once every two hours. The historical distribution and frequency distribution of sulfur content are

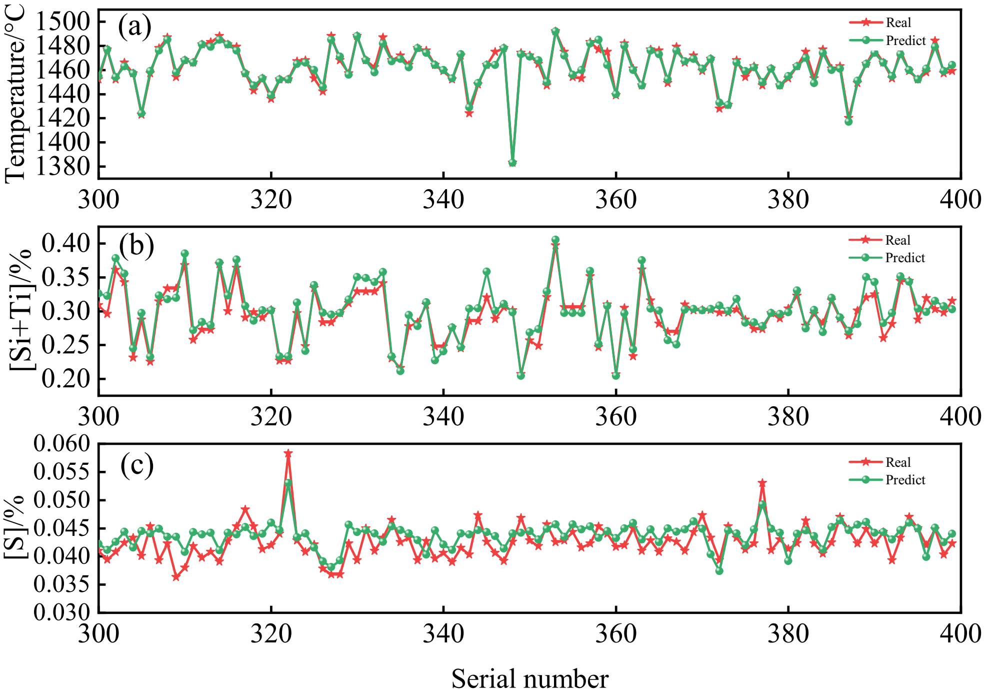


Fig. 4 Results of HMT (**a**), [Si + Ti] (**b**) and [S] (**c**)

shown in Fig. 5a, b. It can be seen from Fig. 5 that the sulfur content of the enterprise was in a stable state as a whole from February 2021 to August 2021, mainly concentrated in the interval of 0.04–0.06, and a small part was greater than 0.06 or less than 0.04. Combining with the process theory, it can be seen that in the BF ironmaking process, hot metal is mainly desulfurized. The substandard sulfur content in the hot metal will affect the fluidity of the hot metal, prevent the decomposition of iron carbide, and then affect the structure and performance of castings and steel. Therefore, take the middle interval 0.04–0.06 as the normal production level of the factory, take 0.005 as the step and take the number of steps multiplied by 1.5 as the step score. Correspondingly, respective evaluation rules are formulated for HMT and [Si + Ti]. Finally, the cumulative result of the scores of these three evaluation indicators is the final score of the hot metal quality of the heat. For the convenience of classification, four grades of S, A, B and C can be set for the final score of hot metal quality. The classification can not only be adjusted at any time

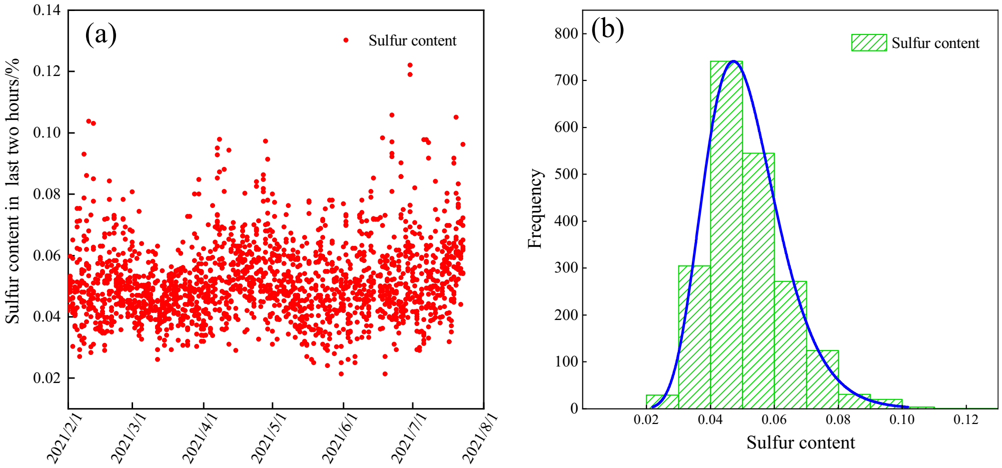
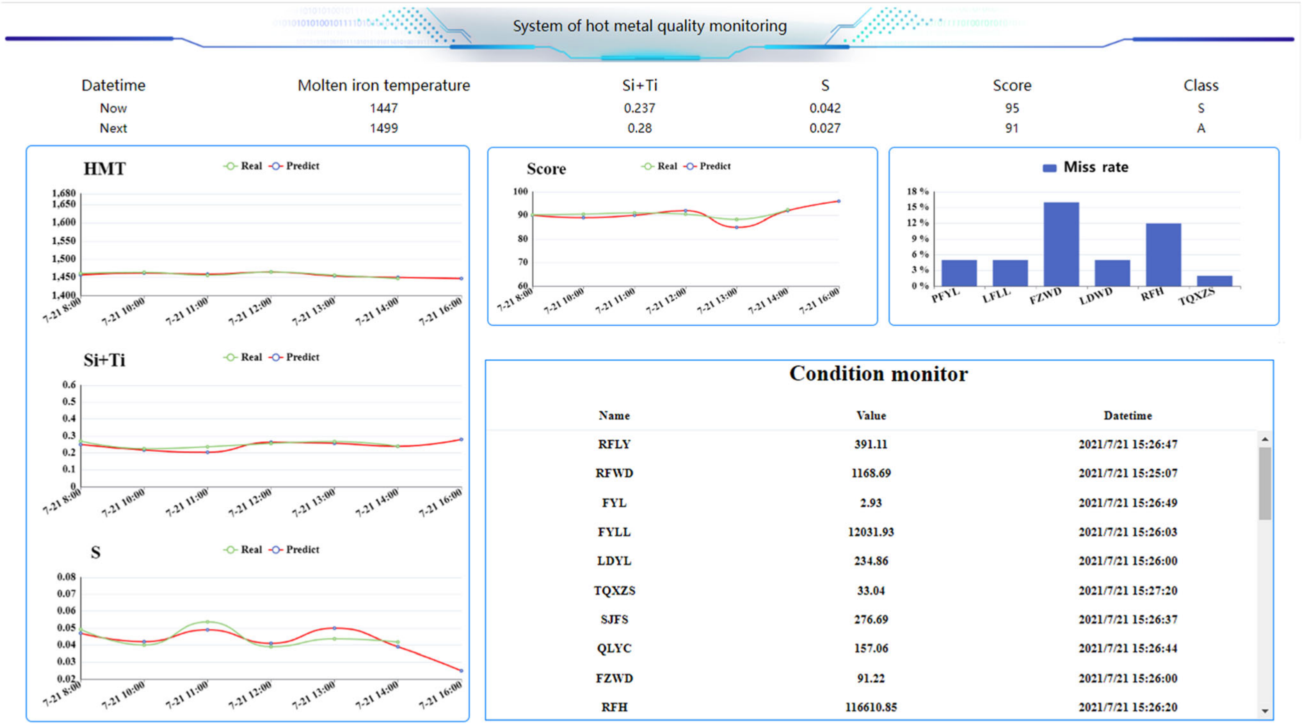


Fig. 5 Historical distribution map of sulfur content (**a**) and frequency distribution of sulfur content (**b**)



System of hot metal quality monitoring

Datetime	Molten iron temperature	Si+Ti	S	Score	Class
Now	1447	0.237	0.042	95	S
Next	1499	0.28	0.027	91	A

Condition monitor

Name	Value	Datetime
RFLY	391.11	2021/7/21 15:26:47
RFWD	1168.69	2021/7/21 15:25:07
FYL	2.93	2021/7/21 15:26:49
FYLL	12031.93	2021/7/21 15:26:03
LDYL	234.86	2021/7/21 15:26:00
TQXZS	33.04	2021/7/21 15:27:20
SJFS	276.69	2021/7/21 15:26:37
QLYC	157.06	2021/7/21 15:26:44
FZWD	91.22	2021/7/21 15:26:00
RFH	116610.85	2021/7/21 15:26:20

Fig. 6 Schematic diagram of hot metal quality monitoring system

according to the actual production status of the factory, but also can ensure the classification of data, which is convenient for data mining in the future.

On the basis of the establishment of the evaluation system, the BF hot metal quality monitoring system was designed. The system mainly relies on Linux, Webstorm2021, Xshell7 and other software as the development environment. Figure 6 shows the hot metal quality monitoring system. The interface of the system mainly includes functions such as comprehensive evaluation of hot metal quality, prediction of key indicators, monitoring of data missing rate and monitoring of core parameters. As we can see, the top of the interface of the system displays the evaluation results and prediction information of the hot metal quality at the current and the next moment, mainly including the hot metal quality evaluation indicators and comprehensive scores mentioned in this paper. The left side of the interface of the system shows the variation trends and prediction trends of the three target parameters of HMT, [Si + Ti] and [S] in detail. The evaluation module uses the prediction module information to comprehensively score the quality of hot metal in each furnace at the same frequency, which can effectively identify the quality of hot metal and take timely adjustment measures for the BF. In addition, other problems like data transmission or equipment failure may cause discontinuities in parameters, resulting in erroneous evaluation results. Therefore, a missing statistics module is set up to count the missing rate of some state parameters that are valued by operators, and play an indirect role in monitoring equipment stability. The state parameter monitoring module is used to display most of the parameters in the BF iron-making process, which are also of concern to operators, e.g., hot blast pressure, furnace top pressure and so on. The establishment of this module greatly improves work efficiency: on the one hand, it is convenient for field operators to view the status of each parameter in real time; on the other hand, it can help operators adjust strategies in time according to the trend of predicted results.

5 Conclusions

1. Disordered data frequency, long time lag and outliers are common characteristics of industrial production data. In the process of realizing industrial informatization and intelligence, when processing data, it cannot be directly deleted by partial generalization or cannot be simply processed and used directly. When excavating the potential value law of industrial data, it is necessary to fully consider the influence of various factors and conditions and use the method of combining big data processing technology and process theory to carefully process the original data.
2. The data used in the modeling process of this study are the result of mutual fusion and joint screening based on

more than 540 original fields using methods such as de-redundancy, data derivation, importance ranking, causal analysis and process experience theory. Diversified feature selection methods just meet the needs of BF process modeling. This feature selection method is not only applicable to the steel industry, but also to other manufacturing industries.

3. In this paper, a hot metal quality monitoring system is established according to the three hot metal quality evaluation indexes concerned by the enterprise. The test results of the system using the existing platform of the factory show that not only the system running state is stable during the test, but also the evaluation score error is basically stable within ± 5 points, which is in line with the expected effect. It meets the needs of intelligentization development of BF in iron and steel enterprises, and at the same time, it has the effect of guiding production in a timely manner and ensuring the smooth running of the BF.

Acknowledgements Thanks are given to the financial supports from the Basic Research Program of National Natural Science Foundation of China (52004096), and the Natural Science Foundation of Hebei Province (E2020209208).

Declarations

Conflict of interest The authors declare no conflict of interest.

References

[1] X.C. Li, C.T. Shi, F. Zhao, Iron and Steel 50 (2015) No. 11, 1–7+13.
[2] X.D. Liu, Q.L. Zhang, World Metal, A06 (Accessed: 2022–01–25). http://app.worldmetals.com.cn:10008/epaper/show.do?paper=sjjsdb&date=20220125&pageid=14923.
[3] X.L. Qu, N. Xing, W. Huang, R.F. Ding, Metall. Econ. Mange. 205 (2020) 54–56.
[4] S. Lee, C. Lee, in: 2013 International Conference on Quality, Reliability, Risk, Mantenace, and Safety Engineering (QR2MSE), 2013, pp. 595–598.
[5] Z.D. Lu, H.Z. Gu, L.K. Chen, D.L. Liu, Y.D. Yang, Ironmak. Steelmak. 46 (2019) 618–624.
[6] J.W. Zhang, L.Q. Niu, New Economy Leader 281 (2021) 41–46.
[7] X. Liu, W.J. Zhang, Q. Shi, L. Zhou, J. Northeast. Univ. 41 (2020) 1153–1160.
[8] R.D. Martin, F. Obeso, J. Mochón, R. Barea, J. Jiménez, Ironmak. Steelmak. 34 (2007) 241–247.
[9] H. Zhao, D.T. Zhao, Y.J. Yue, H.J. Wang, in: Proceedings of 2017 IEEE International Conference on Mechatronics and Automation, Takamatsu, Japan, 2017, pp. 316–321.
[10] J. Diaz, F. Fernandez, M. Prieto, Metals 10 (2020) 41.
[11] K. Jiang, Z.H. Jiang, Y.F. Xie, D. Pan, W.H. Gui, Acta Automatica Sinica (2021) https://doi.org/10.16383/j.aas.c210524.
[12] Y.T. Wang, Q.Y. Yan, G. Yang, W.R. Xu, Chinese Journal of Scientific Instrument 11 (2006) 1448–1451.
[13] A.P.M. Diniz, K.F. Coco, F.S.V. Gomes, J.L.F. Salles, Metals 11 (2021) 1001.
[14] J.P. Li, C.C. Hua, J.L. Qian, X.P. Guan, Fuzzy Sets Syst. 421 (2021) 178–192.
[15] P. Zhou, H.D. Song, H. Wang, T.Y. Chai, IEEE Trans. Control Syst. Technol. 25 (2017) 1761–1774.
[16] J.J. Liu, P. Zhou, L. Wen, Control. Theory Appl. 37 (2020) 987–998.
[17] Z.N. Li, M.S. Chu, Z.G. Liu, G.J. Ruan, B.F. Li, High Temp. Mater. Processes 38 (2020) 884–891.
[18] H.Y. Li, X. Li, X.J. Liu, X.P. Bu, H.W. Li, Q. Lyu, Ironmak. Steelmak. 48 (2021) 283–296.
[19] M. Yuan, P. Zhou, M.L. Li, R.F. Li, H. Wang, T.Y. Chai, J. Iron Steel Res. Int. 22 (2015) 487–495.
[20] X.G. Liu, F. Liu, Blast furnace ironmaking process optimization and intelligent control system, Metallurgical Industry Press, Beijing, China, 2003.
[21] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Mach. Learn. 46 (2002) 389–422.
[22] C.W.J. Granger, Econometrica 37 (1969) 424–438. https://doi.org/10.2307/1912791.
[23] C. Hiemstra, J.D. Jones, J. Finance 49 (1994) 1639–1665.
[24] M. Geerdes, H. Toxopeus, C. Vliet, Modern blast furnace ironmaking, IOS Press, Amsterdam, The Netherlands, 2015.
[25] X.B. Dong, Z.W. Yu, W.M. Cao, Y.F. Shi, Q.L. Ma, Front. Comput. Sci. 14 (2020) 241–258.
[26] Z.H. Zhou, Ensemble methods, foundation and algorithms, Publishing House of Electronics Industry, Beijing, China, 2020.
[27] L. Breiman, Mach. Learn. 24 (1996) 123–140.
[28] M.W. Ahmad, J. Reynolds, Y. Rezgui, J. Clean. Prod. 203 (2018) 810–821.
[29] Y. Freund, R.E. Schapire, in: Proceedings of the 13th Conference on Machine Learning, The MIT Press, Cambridge, USA, 1996, pp. 148–156.
[30] J.S. Yang, C.Y. Zhao, H.T. Yu, H.Y. Chen, Procedia Comput. Sci. 174 (2020) 161–171.
[31] Z.H. Zhou, Machine learning, Tsinghua University Press, Beijing, China, 2016.
[32] T.Q. Chen, C. Guestrin, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Assoc. Comp. Machinery, San Francisco, USA, 2016, pp. 785–794.
[33] S.F. Liu, J. Forrest, Y.J. Yang, J. Grey Syst. 25 (2013) 1–18.
[34] G.S. Du, Z.X. Liu, H.F. Lu, J. Comput. Appl. Math. 386 (2021) 113260.
[35] Y. Yu, X.S. Si, C.H. Hu, J.X. Zhang, Neural Comput. 31 (2019) 1235–1270.