APEM journal
Advances in Production Engineering & Management
Volume 20 | Number 2 | June 2025 | pp 157–172
https://doi.org/10.14743/apem2025.2.533
ISSN 1854-6250
Journal home: apem-journal.org
Original scientific paper

Enhanced product defect forecasting using partitioned attributes and ensemble machine learning

Sun, Y.Y.[a], **Yang, J.H.**[a,*], **Zhai, L.Y.**[a], **Liu, N.**[a]

[a]School of Economics and Management, University of Science and Technology Beijing, Beijing, P.R. China

ABSTRACT

This study addresses a critical challenge in industrial big data analytics for smart manufacturing: conventional machine learning methods often fail to account for data discontinuities caused by scrapped defective intermediates in multi-stage production processes, inadvertently treating non-conforming products as qualified during model training. We propose a novel process-aware data analytics framework specifically designed for process industries, featuring: (1) intelligent attribute partitioning based on information flow discontinuity points, and (2) an ensemble modelling approach combining Random Forest and C5.0 Decision Tree algorithms to generate interpretable prediction rules with quantified feature importance rankings. Validated using real-world production data from a Chinese rail steel manufacturer, our methodology demonstrates superior performance by explicitly incorporating process-specific data correlations. The proposed solution effectively mitigates information distortion caused by scrapped intermediates while maintaining operational interpretability – a crucial requirement for industrial implementation. The research results increased the accuracy rate of the test set of the random forest experiment from 88.39 % to 92.69 %, and the accuracy rate of the test set of the decision tree experiment from 71.89 % to 79.15 %. Additionally, the experimental results verify that, compared with the traditional methods, our framework has better applicability in capturing product quality in the manufacturing industry when process attributes are considered.

ARTICLE INFO

Keywords:
Intelligent manufacturing;
Process industry;
Industrial data mining;
Defect prediction;
C5.0 decision tree;
Random forest;
Process-oriented analytics;
Machine learning

**Corresponding author:*
yangjh@ustb.edu.cn
(Yang, J.H.)

Article history:
Received 27 March 2025
Revised 28 May 2025
Accepted 10 June 2025

1. Introduction

The world's leading industrial countries have adopted new manufacturing production strategies, typical of which are the Industry 4.0 strategic initiative formulated by Germany, the Advanced Manufacturing Partnership of the United States, and the Made in China 2025 strategy. The core concept of intelligent manufacturing is a deep integration of informatization and industrialization, and its key elements are integration, analysis, and use of data. In addition to the "5V" characteristics of big data, the big data associated with the process industry also have the characteristics of correlation, process, and timing [1, 2], which means that the relatively mature data processing methods used in internet-related analysis may not be applicable in the industrial field. For industrial big data, the more important problem to be solved is not one of algorithms but of data collection and preprocessing [3]. Therefore, it is crucial to consider the characteristics of industrial production data in the process of data mining [4].

In traditional industrial production, on-site workers guide production based on experience. In addition, technical personnel may use simple statistical modelling methods to analyse the

factors affecting the qualification rate of products. With the increasing amount of data, more types of data, and requirements of faster data processing, it is necessary to develop data mining technology more deeply. At present, the data mining method of industrial data involves selecting the target variable first and then taking the data of the whole production process as the independent variable to study the relationship between the independent variable and the target variable through methods such as decision tree, association rules, and neural network. The major challenge is that industrial production data involves many processes, and there are many correlations between processes. Borchert *et al*. [5] validates the significance of considering production business processes for production management analysis. To analyse the relationship between production data and the qualification rate of final products, the complexity of data must be reduced. The data is so complex that traditional machine learning algorithms cannot get meaningful analysis results, and the results deviate greatly from the actual production. The experience of workers cannot interact with the machine. The existing methods mainly improve the intelligence of the algorithm technically but fail to make improvements from the perspective of simplifying the correlation between production processes from the management level. Few studies have explored data mining methods specifically tailored to the characteristics of the process industry, and existing research shows limited applicability to the analysis of production data with complex process flows.

This study proposes an innovative data mining methodology for industrial process analysis that simultaneously incorporates data correlation structures and process characteristics. The core innovation involves a correlation-driven partitioning mechanism for high-dimensional attributes based on inter-process relationships, enabling comprehensive defect pattern extraction and probabilistic defect prediction. Specifically designed for process-intensive manufacturing sectors (e.g., steel, chemicals, pharmaceuticals), our method addresses the critical challenge of data discontinuity caused by substandard intermediate products, effectively preventing the learning of artifact patterns from interrupted data streams. The approach maintains three fundamental advantages: (1) holistic process integration that preserves information integrity, (2) systematic analysis of variable-target correlations, and (3) adaptive attribute segmentation optimized for production dynamics. Experimental validation using real rail manufacturing data demonstrates the method's superior practical utility, showing statistically significant improvements in prediction accuracy compared to conventional approaches, ultimately delivering enhanced operational guidance value for actual production environments.

2. Related works

Quality control is one of the most important aims in industrial production, reducing the probability of defects is an extremely important task. Therefore, the detection and prevention of product defects have always been a focus of researchers. With the development of intelligent manufacturing, the use of big industrial data to control production quality problems has gradually entered the field of the vision of researchers [6]. There is substantial research on manufacturing quality management, defect diagnosis, and defect prediction. Kovacic *et al*. [7] designed an improved genetic programming method to predict the proportion and length of defects on rolled bar surfaces Perzyk *et al*. [8] summarized the techniques of using data mining methods to predict defects. Hsu and Chien [9] developed a hybrid data mining method to prevent defects from occurring in a manufacturing process and to identify and extract patterns associated with manufacturing defects from models. Norrena *et al*. [10] studied rule-based decision making and machine learning algorithms for predicting crack formation and other defects in defect-prone steel grades and showed the potential for defect prevention through composition optimization. Hochbaum and Liu [11] proposed an adjacency clustering model to predict the product defect rate in integrated circuit manufacturing and improve production efficiency

Building upon defect prediction models, various production optimization methods have been developed to enhance product quality in steel manufacturing. Vukelic *et al*. [12] built predictive models and performed process optimizations to maximize dimensional accuracy and surface quality. Chongwatpol [13] used big data to diagnose defects and causes of changes in the produc-

tion process and used control charts, defect costs, and defect prediction scores based on clustering to reduce defects. Lee *et al*. [14] used fuzzy association rule mining and a recursive process mining algorithm to identify the relationship between production process parameters and product quality. Akarwal *et al*. [15] applied correlation association rules to determine the optimal processing conditions for three non-traditional machining processes: electrochemical machining, ultrasonic machining, and electrical discharge machining. Breznikar *et al*. [16] proposed to optimize and regulate the pouring temperature during the casting process. Ciarapica [17] adopted association rules to identify refinery environmental risks and take corrective measures. Li *et al*. [18] proposed a self-organizing radial basis function neural network to predict output and optimize operation. The above research used machine learning method to study quality control. Quality management has transitioned into the era of machine learning analysis from the previous era of traditional statistical analysis.

Decision tree and random forest are classification algorithms in machine learning, which have high applicability in the field of production quality management. The decision tree algorithm provides interpretable decision rules for the root cause of high or low product pass rate. Decision tree method is one of the classical methods to predict industrial production quality with process attributes. Irani *et al*. [19] first applied the decision tree model to the prediction of the quality qualification rate of industrial products. Keswani [20] applied a classifier based on decision trees for inventory management and regarded the number of defects as a fuzzy variable to predict seasonal demand. Li *et al.* [21] adopted the improved C5.0 decision tree algorithm to optimize production process parameters. The decision tree algorithm is widely used also in optimizing process parameters and extracting classification rules. The random forest algorithm is an integrated learning method for multiple decision trees; it has strong anti-noise ability and can accurately predict the importance of attributes. Grdinaru *et al*. [22] emphasized that the random forest classification algorithm classifies statistical units based on decision trees and has a relatively low error rate for identifying complex elements. Ozbalci *et al*. [23] validated the advantages of random forest algorithms in process industry data analysis. Wang *et al*. [24] analysed large-sized data using the random forest optimization algorithm, selected key production factors through information gain and predicted product quality by using sensitivity analysis, and verified the applicability of the random forest algorithm in the problem of product quality prediction in the process industry. Esteve *et al*. [25] discussed the advantages of random forest algorithm in processing high dimensional data of freely disposed hulls and proposes a new method to evaluate the importance of input variables.

The studies above used data mining methods to analyse industrial data, but the studies used data mining methods to directly model and ignored the process and correlation in the actual production. The relationship between each process is intricate, and the high-dimensional production process data with complex information often make the mining results inaccurate [26]. There is some internal correlation between the data in the process industry. Considering the correlation between processes is a top priority of industrial big data mining [27]. Wei *et al*. [28] identified key limitations in handling the inherent diversity and dynamics of complex production workflows, subsequently developing a simulation model that accurately represents real-world manufacturing processes. Yao and Ge [29] proposed a method for quality prediction of big data in the process industry. Negoita and Borangiu [30] incorporated business processes into demand forecasting models, merging process automation techniques with inventory management to enable knowledge-intensive service discovery. Novak *et al*. [31] identified the challenges of making informed decisions that positively impact business processes and proposed a new approach to integrating business processes with requirements management. Ding *et al*. [32] proposed a latent sequence correlation calculation model for anomaly detection of industrial data series. However, these studies aimed at the analysis of the time series of the whole production process of a certain process or a certain machine. Although the correlation of industrial data was considered, the overall analysis method was not provided from the perspective of the total factors of the whole production.

3. Research model

3.1 Problem description

Industrial production data is generated sequentially along the manufacturing process, often involving intermediate products that undergo multiple processing stages before becoming final outputs. Process industry data is characterized by its large volume, high dimensionality, and tight integration with business workflows, posing significant challenges for industrial data analysis. In industrial production, products are divided into multiple intermediate items manufactured across different processes. Each processing workshop is responsible for producing specific intermediates, which must pass quality inspection. Non-conforming intermediates are scrapped and do not proceed to subsequent workshops, resulting in the termination of their corresponding production data streams. All workshops consolidate their production data at the corporate level, including data from both qualified and non-conforming intermediates. However, in practice, the qualification rate of intermediates is routinely reported as 100 %, while machine learning models inadvertently train on data from defective intermediates. These non-conforming items, though excluded from the final product, participate in the machine learning analysis of final product data. This information discontinuity caused by scrapped intermediates may significantly degrade the quality of knowledge discovery.

In industrial production, product processing is divided into M stages, generating K intermediate products. Each intermediate undergoes N processing steps and must pass quality inspection. If a batch of intermediates achieves a 100 % qualification rate, it proceeds to the $N + 1$ stage. If the qualification rate is below 100 %, the non-conforming items are removed and scrapped, and the remaining batch continues to subsequent stages until the final product is completed. The process flow is illustrated in Fig. 1.

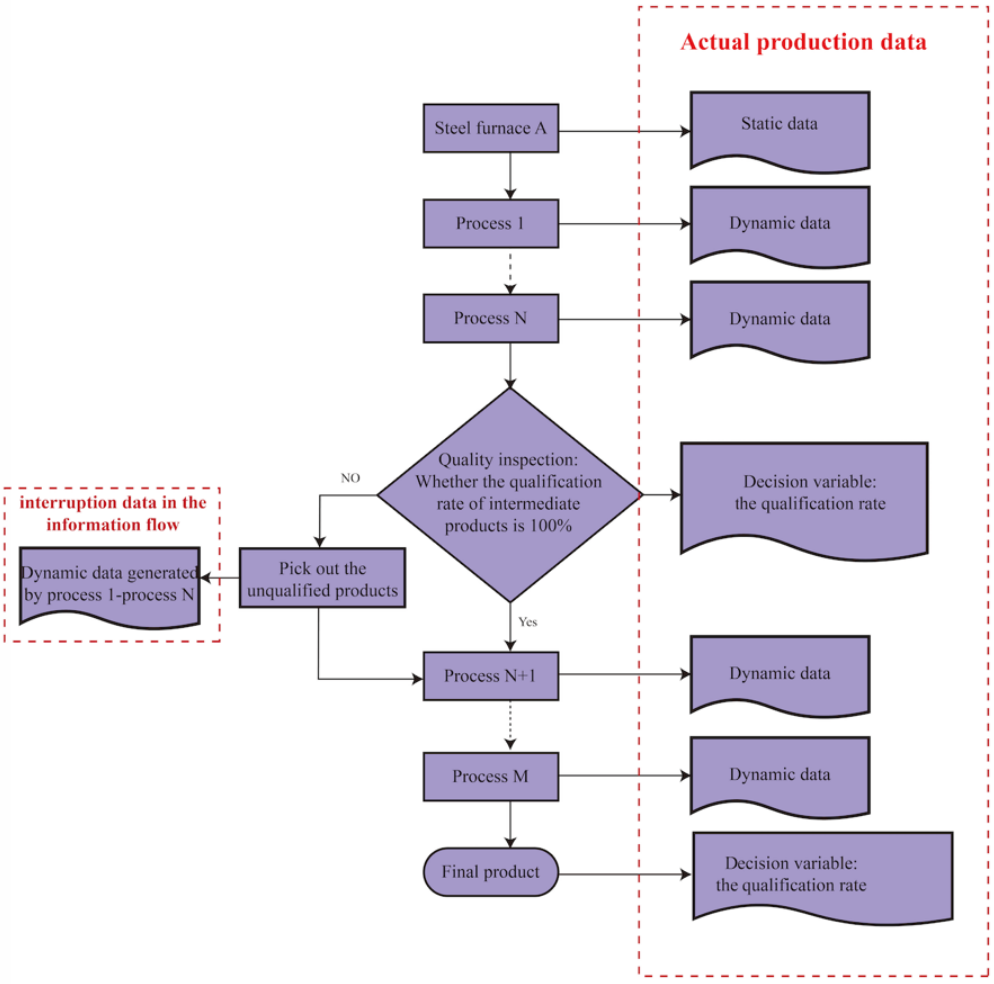

Fig. 1 Product manufacturing data flow chart

Dynamic data is generated at each production stage. However, the scrapping of non-conforming items during quality inspection leads to information discontinuities. Specifically, some dynamic data generated from the first to the Nth stage becomes irrelevant to the final product qualification rate. Moreover, all intermediates entering the $N + 1$ stage are qualified, effectively optimizing the qualification rate to 100 %. The data flow is segmented at the quality inspection node, meaning that production data from non-conforming intermediates does not proceed to subsequent stages.

Traditional data mining methods are ill-suited for analysing process-oriented industrial production data due to these discontinuities. The goal of this study is to enable more accurate industrial data analysis by preventing models from learning spurious relationships between irrelevant independent variables and target variables. To address the flow and correlation characteristics of big industrial data, we propose a process attribute-partitioning based data mining method tailored for process industries. Implementing attribute partition in the production system provides substantial operational advantages, including improvements in production efficiency and product qualification rates (for details, see the discussion of experimental results in Section 4.3). Crucially, this method requires only the minimum organizational investment. For instance, enterprises need to increase their investment in business process integration to collect key production data that affects the qualification rate of each intermediate product. Strengthen the statistical norms of production data, etc.

3.2 Defect prediction model of process industry products

Due to the possibility of defective intermediate products being scrapped during industrial production, data flow interruptions can lead to the loss of correlations between data. To address this, we propose a method of segmenting data based on the business process to reduce the complexity of industrial data mining and avoid using interrupted data for machine learning. Below, we describe our proposed industrial data mining method that considers process and correlation data characteristics.

The method consists of three main stages. The first stage is data preprocessing, which includes data cleaning, discretization of quality inspection data for intermediate products, converting industrial data into a learnable format, and sampling and partitioning the dataset. The second stage involves using attribute partitioning methods, based on the location of information flow interruptions in the quality inspection process. The third stage uses the Random Forest algorithm, which is an ensemble learning method based on decision trees. This algorithm is known for its robustness against noise and its ability to accurately predict the importance of attributes. The classification rules generated by the C5.0 decision tree algorithm are easy to understand and reflect the production process sequence more accurately. Therefore, based on the characteristics of rail production data, we perform decision tree and Random Forest experiments on the partitioned data, generating a rule set that guides actual production and obtaining a ranking of attribute importance. C5.0 decision tree algorithm transparent interpretability (white-box model) enables industrial practitioners to easily extract decision rules and efficiently handle high-dimensional data with mixed numerical and classification features. The algorithm demonstrates rapid training convergence, which proves essential for efficient model iteration and refinement in production environments. Random Forest algorithm is an integrated approach that can handle thousands of dimensions of features, has high parallelization efficiency, and is more suitable for process industry data analysis with high data dimensions. The algorithm selected in this paper strikes a balance between interpretability (for operational guidance) and scalability (for high-dimensional process data).

The attribute partitioning method divides input variables into data blocks for mining, with each block corresponding to a target variable. This method is divided into four steps. The first step is to arrange all attributes D_i in sequence according to the production process. The second step is to traverse each attribute D_i to determine if it belongs to the quality inspection process. The third step is to label the attributes, if it belongs to the quality inspection procedure, label it as D_j. The fourth step is to divide the data into N parts based on the location of D_j, and construct decision trees for N stages. Attribute partitioning is effective for production data with multiple

process steps or large datasets, where the fitting dimension N is often very high. The random forest algorithm utilizes Bagging and random feature selection to reduce variance, ensuring that the test error is very close to the training error. Decision trees rely on pruning techniques to alleviate the problem of overfitting. Adnan and Islam have verified that the set of 100 decision trees achieves a prediction accuracy of over 80 % [33]. In the manufacturing industry, the production process usually involves 5 to 8 intermediate products. The proposed attribute division method can maintain robust predictive performance even when N is large and is particularly suitable for multi-stage production systems. The target variable is the qualification rate decision data, and the independent variable is the dynamic data from each block. Each decision tree reflects the rules of different stages, and the leaf nodes represent the probabilities of each result for the target variable. Based on the result from the root node of the previous decision tree, resampling is conducted to obtain the corresponding probability samples, which are then used to construct the second decision tree. Integrate the decision trees at each stage to obtain the product defect rule set and the product qualification rate.

Table 1 presents the process of the defect product prediction algorithm for process industries based on Random Forest.

Table 1 Defect product prediction algorithm for process industries based on Random Forest

Algorithm: Product defect prediction algorithm of process industry based on random forest

Input: training set $G = \{(x_1, y_1), (x_2, y_2), \dots, (x_w, y_w)\}$;
Among, $x_i \in x \subseteq R^n$, $y_i \in y = \{1,2\}, i = 1,2, \dots, w$;
Feature space D;
Number of decision trees in a random forest t;
Random characteristic number;
Business process;
Number of training epochs (data flow interruption nodes) N.
Output: Visualize attribute importance and prediction accuracy
Course:
1: Data preprocessing
2: Sorted according to the business process, it is divided into N phases
3: The feature space D is divided into N feature subsets according to business processes N , $N \in \{1,2, \dots, K\}$
4: **for** n**=1,2,...,**N **do**
5: Bootstrap Sampling: For each tree $t \in \{1,2,...,T\}$
A sample is extracted from G_N and a subdataset is generated, generate a subdata set G_{Nt}
6: Generate a decision tree for each G_{Nt}
When each node splits, m features are randomly selected from D_N, the optimal splitting point is selected, and the splitting is repeated until the stop condition is met
7: Integrated model: Combine T trees into a random forest
8: **end for**
Output: Visualize attribute importance and prediction accuracy

As shown in Fig. 2, Experiment 1 and Experiment 2 in the figure are the traditional defect prediction algorithm and the improved defect prediction algorithm respectively. The detailed experimental process is described in Section 4.

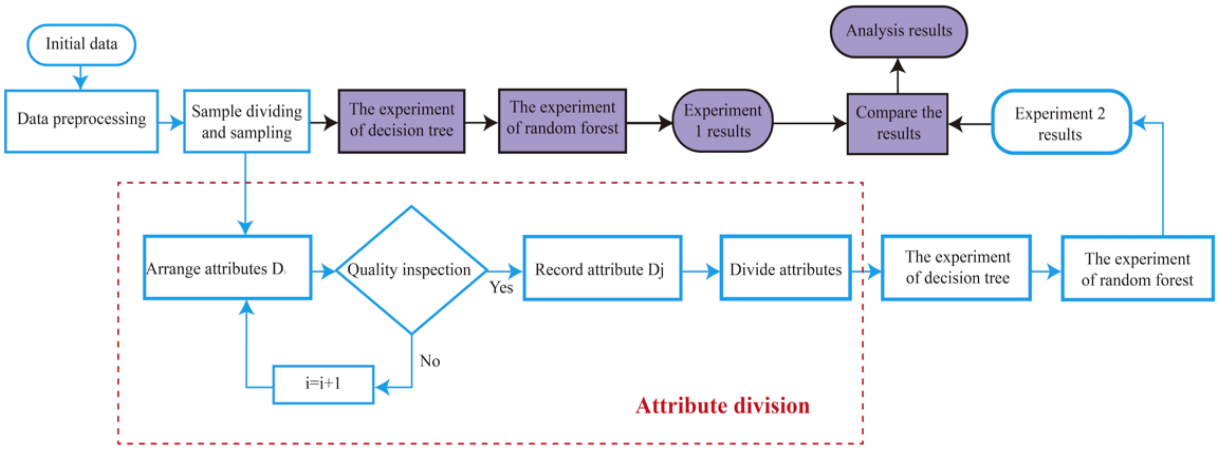

Fig. 2 Flow chart of industrial data mining method based on attribute partitioning

4. Results and discussion

4.1 Experiment design

Regarding experimental design, we first divide all attributes according to the information interruption position in the production process and then use the machine learning algorithms to analyse the industrial production data.

Experiment 1: Data mining method without considering data correlation of the process industry. The decision tree was constructed, and the random forest experiment was carried out to obtain the rule set and the ranking of attribute importance that affected the rail defects.

Experiment 2: Data mining method considering data correlation of process industry. In the first stage, the data were partitioned according to the relevance of the production process. In the second stage, the decision tree was constructed in blocks, and the random forest experiment was carried out. In the third stage, the decision tree was connected to obtain the rule set, and the ranking of attribute importance was obtained by the random forest experiment.

By comparing Experiment 1 and Experiment 2, the effectiveness of the data mining method based on attribute partitioning is verified.

In view of the correlation characteristics of industrial big data, this paper takes the actual production data of a large domestic steel mill as an example to preprocess and analyse the data. The random forest algorithm is used to sort the importance of attributes, and the C5.0 decision tree algorithm is used to construct the rule set. The model was established and evaluated by two methods, and the results of traditional data mining (Experiment 1) and those of a data mining method based on production process attribute partition (Experiment 2) were compared. The research process was as follows: (1) data preprocessing, including data cleaning and data transformation. (2) In Experiment 1, the C5.0 decision tree algorithm was used to mine the relationship between all sample data and target variables, and a decision tree rule reflecting whether there are some defective rails was obtained. The random forest algorithm was used to sort the importance of attributes. (3) In Experiment 2, attributes were divided into two parts according to the quality inspection procedure, and the C5.0 decision tree algorithm was used to obtain the two-stage decision tree rules. The random forest algorithm was used to sort the importance of the attributes of the two stages. The two decision trees obtained were connected to generate a complete rule set. (4) The results of Experiment 1 and Experiment 2 were compared and analysed, and the characteristics and advantages of industrial data mining were summarized.

4.2 Experiment process and results

Taking the industrial production of iron and steel enterprises as the research context, this study selected the real data of rail production of a large steel mill in China as the research object, predicted the probability of rail defects, and analysed the causes of defects. The data are the rail production data of the whole year of 2018, with 7,353 samples in total. For controlling variables, we selected the processing data of U75V rail in 2018 in the steel mill, 2,383 effective samples and 93 attributes were obtained. One sample corresponds to one batch, and the production operation process of each batch is the same (a batch is the minimum production unit for rail production). The sample includes 44 attributes, including refining heating time, temperature before VD, temperature after VD, R(C)L (rail carbon content distance lower boundary), qualification rate of slab, and qualification rate of rail, which covers the whole production process from steelmaking to finished product. The rail processing flowchart is shown in Fig. 3.

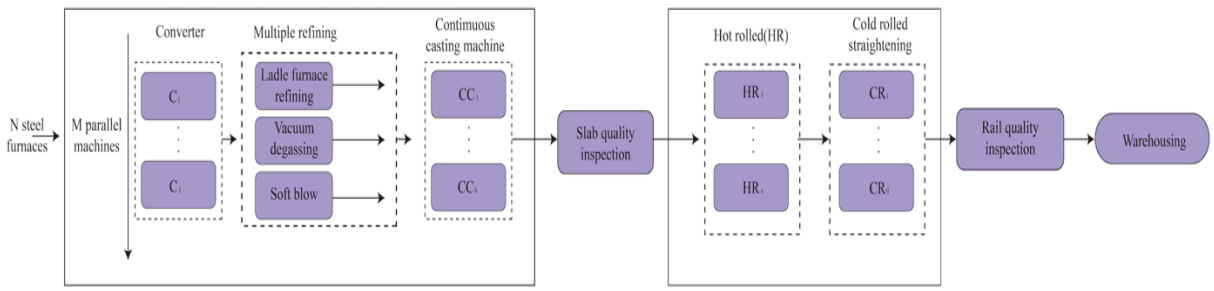

Fig. 3 Rail manufacturing process

The following is the data preprocessing process.

Data cleaning

In the original data set used in this study, there were many unrecorded entries, that is, missing values. However, for attributes such as qualification rate of slab and qualification rate of rail, missing values reflect the actual production conditions and exist as target variables in this study. If the missing values are artificially filled, the actual production situation will be changed, resulting in poor performance of the classification model. In this study, we deleted the entire row of all variables that could not be filled. Heating time, soft blowing time, temperature after VD, and other attributes correspond to data collected in the actual production process, which can be filled after analysis. Descriptive statistics were used to determine the method to fill in the missing values. The mean value was used to fill in the case of symmetric data distribution, and the median value was used to fill in the case of skewed data distribution.

Data cleaning in this study also includes processing outliers and removing noise in the data, which was performed by identifying outliers through box separation. Noise outliers were removed and smoothed by the box mean method. We deleted the entire row of all variables that could not be processed.

Data discretization

The classification model used in this paper can deal with continuous independent variables, but the target variables must be discrete. Therefore, it is necessary to discretize the target variables of continuous attributes in the experimental data, including qualification rate of rail and qualification rate of slab. Data discretization methods include box discretization, histogram discretization, entropy discretization, and clustering discretization. In this paper, the qualification rate of rail is divided into two categories: qualified rate of 100 % and qualification rate of less than 100 %, that is, whether the sample contains defective rail or not. When the qualification rate of the slab is taken as the target variable, the samples with 100 % qualification rate and the samples with less than 100 % qualification rate are taken as labels to discretize the continuous attributes.

Data conversion

Data transformation transforms data into a new form that is easier to mine. We conducted the following operations. (i) Attribute transformation: generalize the date to minutes and seconds. (ii) Attribute construction: construct new attributes such as temperature difference and time interval, etc., and add the new attribute to the attribute set. (iii) Standardization: convert attribute data into distance between upper and lower bounds according to its optimal range. The content of chemical elements in this paper has an industry standard range, so we converted all element content into distance between upper and lower bounds. Through attribute transformation, new attribute construction, and the introduction of the distance between element content and boundary, the information content of data can be extracted to the extent possible.

Sampling and partitioning of datasets

If the uneven distribution of target variables makes it difficult for the decision tree and random forest algorithms to ensure accuracy, the performance of the classifier will be biased towards most classes. To prevent a few types of oversampling from causing overfitting, this study adopts a combination of undersampling and oversampling to process unbalanced data. In order to ensure the objectivity of the results, the data were divided into training sets and test sets, accounting for 80 % and 20 %, respectively.

4.2.1 Data mining method without considering the correlation of rail production data

Experiment 1: Random forest algorithm modelling

Taking qualification rate of rail as the target variable, variables include two labels: qualification rate of 100 % and qualification rate of less than 100 %; other attributes are independent variables. The importance of attributes calculated by the model is shown in Fig. 4. The top 10 attributes of importance involve several stages of refining, vacuum degassing, continuous casting, and hot rolling. The ranking of attribute importance obtained is more accurate than the traditional statistical analysis method. The attributes ranked in the top ten in importance belong to multiple production stages and belong to different production workshops. It is impossible to accurately determine the ranking of important attributes that affect the qualification rate of rails.

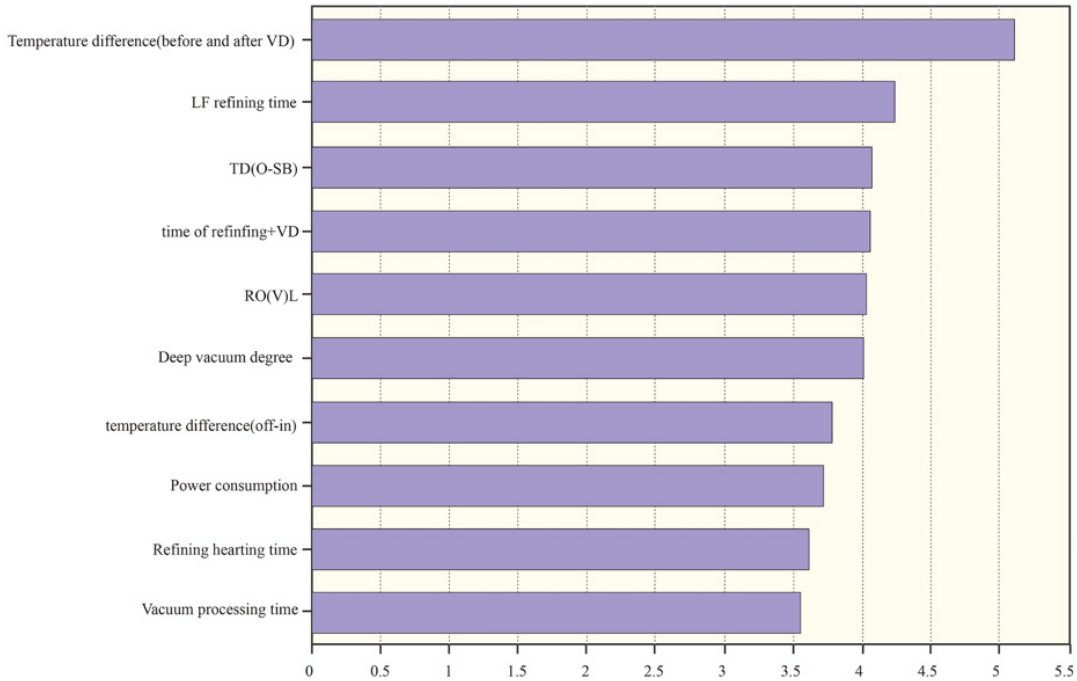

Fig. 4 The attribute importance degree that affects the qualification rate of rail

Experiment 1: Decision tree algorithm modelling

Qualification rate of rail after discretization was taken as the target variable. All the remaining attributes, which involve several stages of production from refining to rail quality inspection, were independent variables. The mapping diagram of the decision tree modelled by the C5.0 decision tree is shown in Fig. 5. With node 96 as an example, the detailed diagram is shown in Fig. 6.

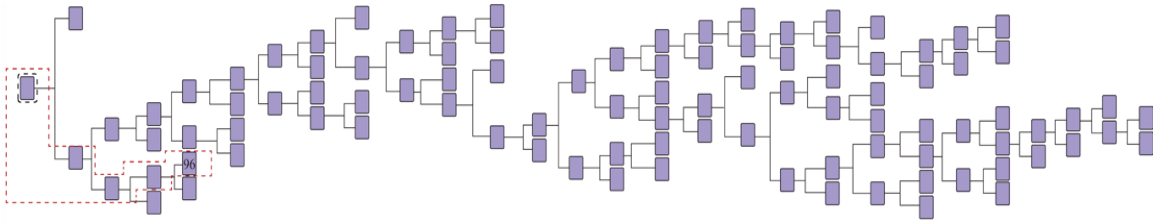

Fig. 5 Experiment 1: Decision tree mapping graph

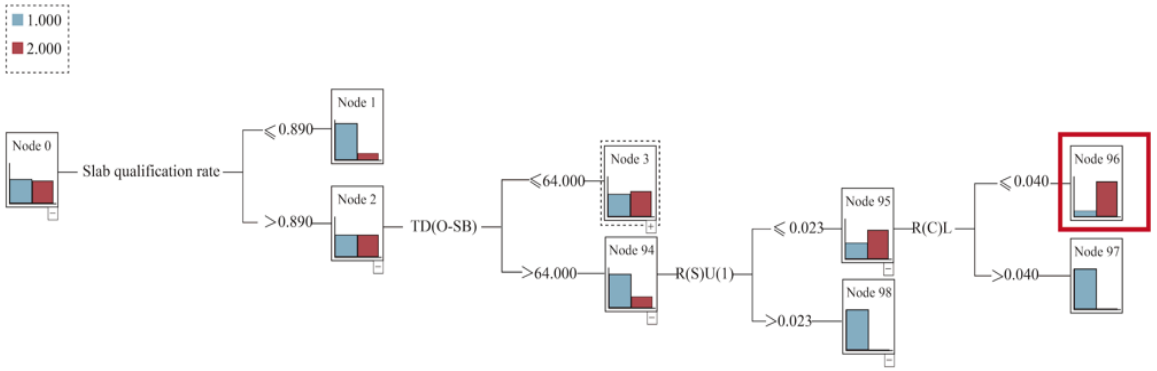

Fig. 6 Details of the decision tree for node 96

The decision tree rule obtained is illustrated in Table 2. Rule 1: The batch contains defective rails with a probability of 100 %. Rule 2: The batch contains defective rails with a probability of 86.84 %. In the obtained decision tree, each path corresponds to a rule, and leaf nodes reflect the probability of whether a batch of rails contains defective products.

The decision tree results of Experiment 1 obtained the production operation rules and corresponding probabilities of the batch containing defective rails, and the important attributes affecting the occurrence of defects were obtained from the random forest experiment results. Before the experiment, the data were divided into training sets and test sets. The accuracy of the

results of the traditional experiment (Experiment 1) is shown in Table 3 The model accuracy is greater than 70 %, which proves that the experiment has credibility. Although the experimental results were obtained based on a large amount of data and the accuracy was credible, the correlation and process of the data were not considered. The results were obtained based on all the data without attribute partition, and the influence of the interrupted data on the results was not excluded.

Table 2 Decision tree rule enumeration

Decision tree leaf node	Process rules	Result, Probability
1.node#15	If 0.89 =< qualification rate of slab <=1, TD(O-SB) <= 64, R(V)L <= 0.01, casting time <=50, LF refining time > 65, RO(V)L > 0.001, soft blowing flow <=65, and temperature before VD > 1551	The batch contains defective rails, 100
2.node#96	If 0.89 =< qualification rate of slab <=1, TD(O-SB) > 64, R(S)U <= 0.023, and R(C)L <= 0.04	The batch contains defective rails, 86.84

Table 3 The accuracy rates of the two models

Experiment	Training set (%)	Test set (%)
Experiment 1: random forest algorithm modelling	98.08	88.39
Experiment 1: decision tree algorithm modelling	82.48	71.89

4.2.2 Data mining method considering the correlation of rail production data

Based on Experiment 1, we proposed to divide attributes based on the production process, and conduct data mining for each piece of data after division and then combine the obtained results to find the rules that could not be found in Experiment 1.

We arranged all the attributes according to the rail production process, as shown in Fig. 7.

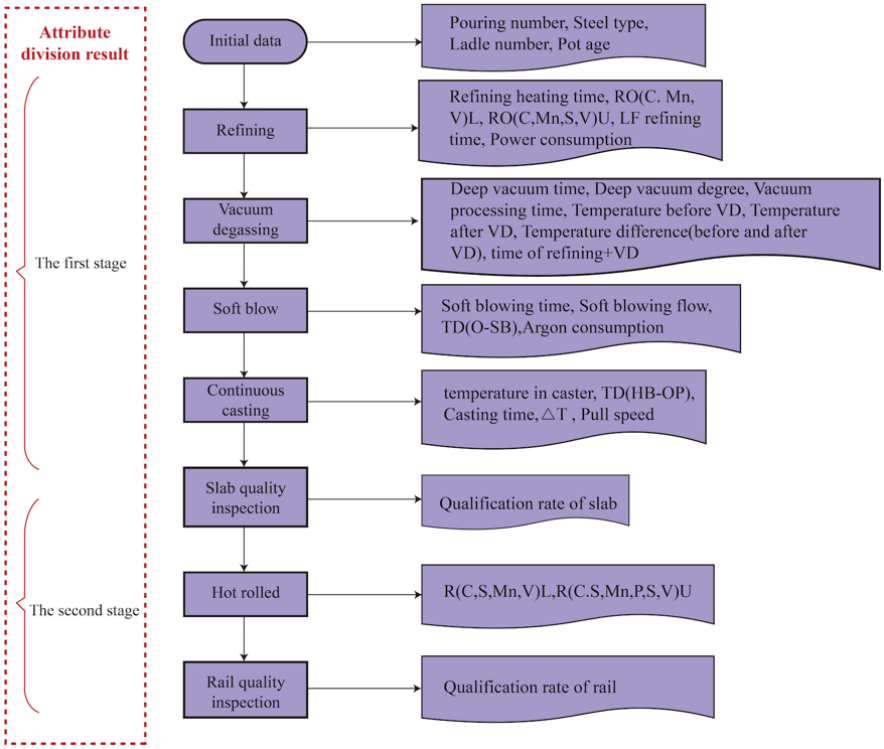

Fig. 7 Data flowchart of rail manufacturing

After examination of all the attributes, it is found that the attributes “qualification rate of slab” and “qualification rate of rail” are the data generated by quality inspection. The independent variables of Experiment 1 were divided into two parts in Experiment 2 with “qualification rate of slab” as the node. The first part of the data is from refining to slab quality inspection process in the production data; the second part of the data is between hot rolling and rail quality inspection process production data. First, the discretized “qualification rate of slab” was taken as the target variable of the first stage, and all variables of the first stage were taken as input.

Experiment 2: Random forest algorithm modelling

The random forest algorithm was used to get the ranking of attribute importance in the first stage. The results of the first stage show the ranking of importance of attributes that affect the qualification rate of slab. The top 10 attributes are shown in Fig. 8. The variables of the second stage and "qualification rate of rail" were taken as independent variables, and the discretized qualification rate of rail was taken as the target variable. The random forest algorithm was used to get the ranking of the importance of attributes of the second stage. The top 10 attributes are shown in Fig. 9. The results show the order of attribute importance of the influence of attributes from hot rolling to rail quality inspection on qualification rate of rail.

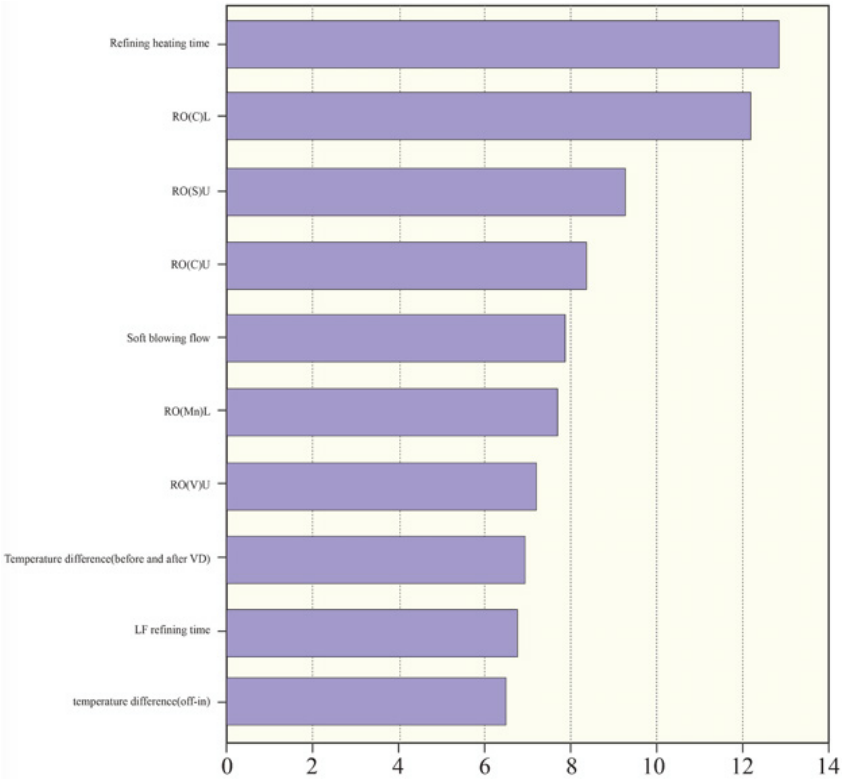

Fig. 8 The attribute importance degree that affects the qualification rate of slab

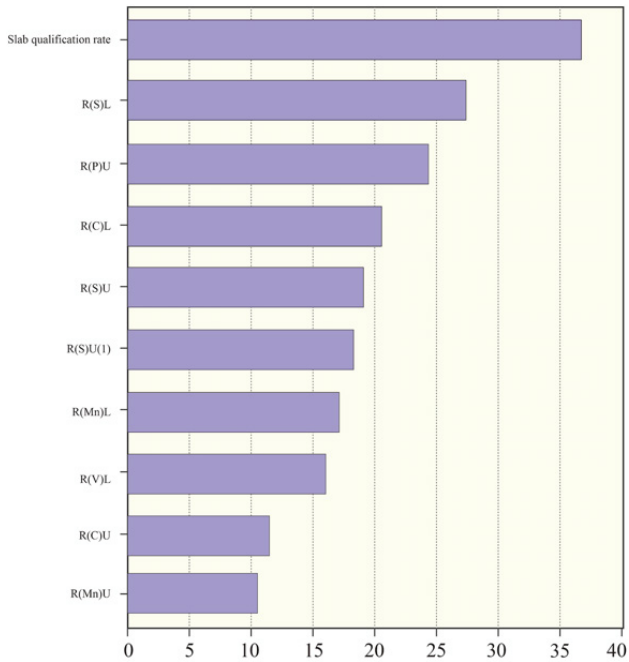

Fig. 9 The attribute importance degree that affects the qualification rate of rail

Experiment 2: Decision tree algorithm modelling

– The first stage experiment

The discretized "qualification rate of slab" was taken as the target variable of the first stage, and all the variables of the first stage were taken as independent variables. The C5.0 decision tree algorithm was used to obtain the rule mapping diagram of the first stage decision tree, as shown in Fig. 10. With node 86 as an example, the detailed diagram is shown in Fig. 11.

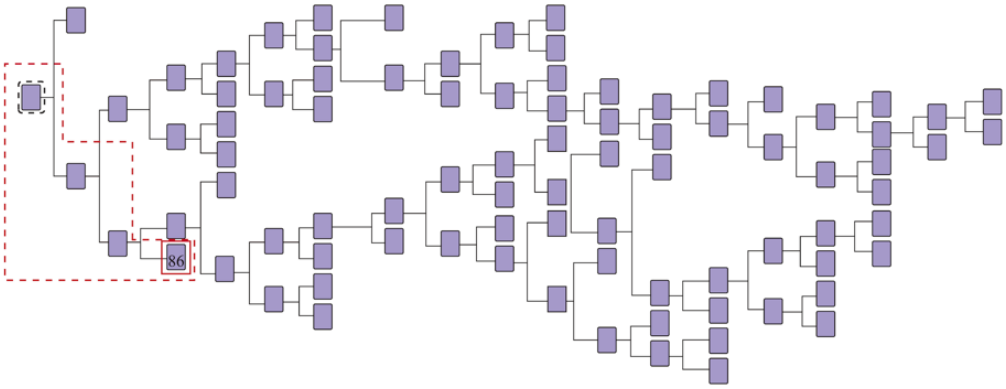

Fig. 10 Decision tree mapping graph of the first stage

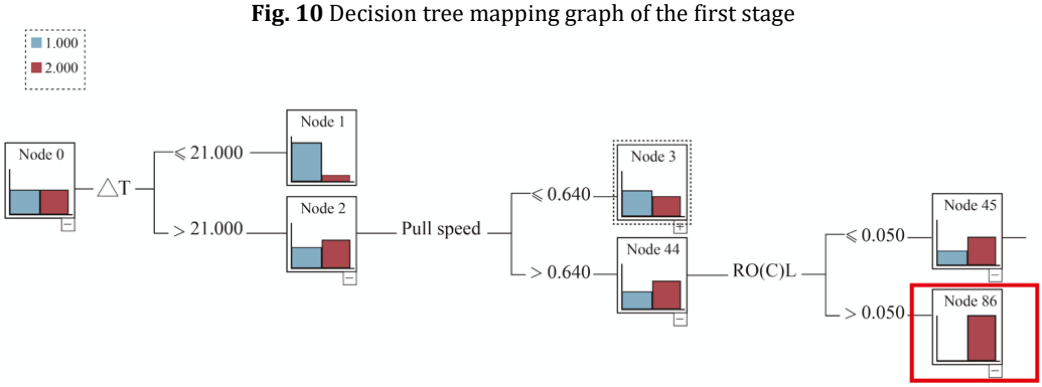

Fig. 11 Details of the decision tree for node 86

– The second stage experiment

The target variable of the second stage was the discretized "qualification rate of rail", and all the attributes of the second stage and "qualification rate of slab" were taken as independent variables. In the C5.0 decision tree algorithm, the largest information gain rate among independent variables was selected as the root node. The experimental results prove that the root node of the second stage decision tree is "qualification rate of slab". The "qualification rate of slab" was discretized into two categories: "1" (the batch containing defective rails) and "2" (the batch does not contain defective rails). All the leaf nodes of the decision tree in the first stage reflected the probability of two categories. Therefore, according to the probability of the two categories, samples of the corresponding proportion are selected from the second stage to construct the decision tree of the second stage. The root node of the two-stage decision tree is connected with the corresponding leaf node to get the rule set (not a new decision tree). Since the number of branches of the decision tree grows exponentially at this time, only the leaf nodes of the two categories "1" and "2" of the decision tree in the first stage are connected to the decision tree in the second stage with the probability of occurrence of more than 80 %.

In Fig. 12, leaf node 84 in the first stage is taken as an example. The probability of category "2" of the "qualification rate of slab" of this node is 100 %. Therefore, 1,000 samples from category "2" of qualification rate of slab are selected, and only the attributes of "qualification rate of slab" and the attributes of the second stage are retained for the decision tree experiment. The decision tree mapping diagram is shown in Fig. 13.

The results of Experiment 2 were two-stage decision tree rules after the attribute division. The results obtained after integrating the two-stage rules are listed in Table 4. The set of pro-

duction operation rules that affect the high defect rate of rail and the probability of each path correspondence are obtained. The results of the random forest experiment show that the important attributes affect the defects of rail after attribute division. Before the experiment, the data were divided into training sets and test sets. The accuracy of Experiment 2 is shown in Table 5. The accuracy of Experiment 2 is greater than 70 %, which proves the credibility of the experiment. The results exclude the influence of interruption data and prove that the attribute partition is scientific and reasonable.

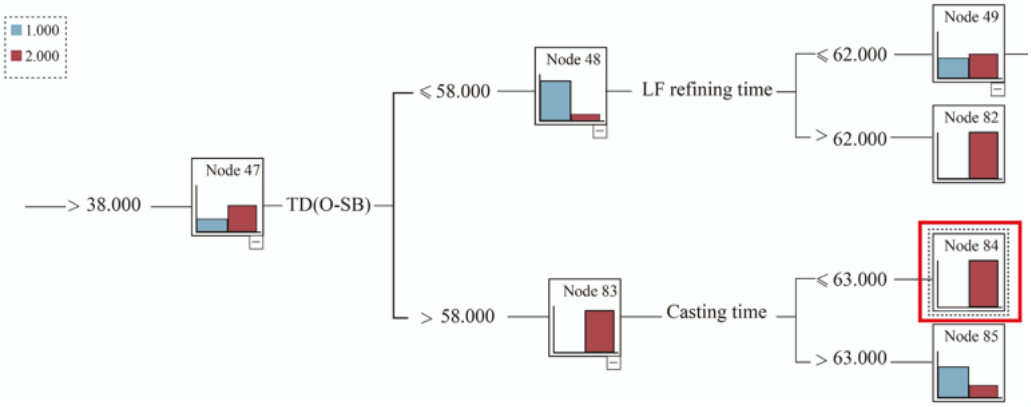

Fig. 12 Details of the decision tree for node 84

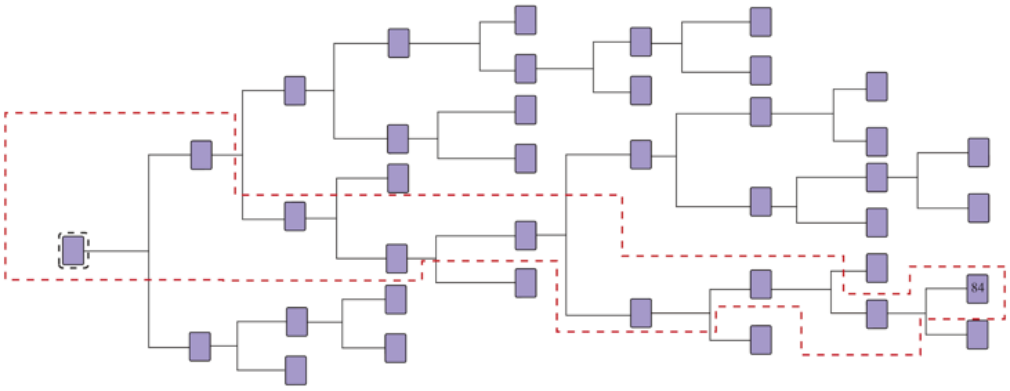

Fig. 13 Decision tree mapping graph of the second stage

Table 4 Decision tree rule enumeration

Decision tree leaf node	Process rules	Result, Probability
1.a (node#84)+b	If △T > 21, pull speed > 0.64, RO(C)L <= 0.05, casting time > 38, TD(O-SB) > 58, casting time <= 63, R(Mn)L <= 0.24, and R(V)L <= 0.006	Containing defective rails, 92
2.a (node#81)+b	If △T > 21, pull speed > 0.64, RO(C)L > 0.05, R(Mn)L > 0.24, R(V)L > 0.004, and R(P)U > 0.018	The furnace is all qualified, 100

Table 5 The accuracy rates of each model

Experiment	Training set (%)	Test set (%)
The first stage: random forest algorithm model	98.59	93.46
The second stage: random forest algorithm model	95.4	92.69
The first stage: decision tree algorithm model	89.44	72.17
The second stage: decision tree algorithm model	78.78	79.15

4.3 Experiment results discussion

Our data consists of 42 attributes, which are generated by 8 production processes and 2 production workshops. Due to the complexity of the production processes, the data dimension is high, spanning multiple production processes and even belonging to multiple production workshops. Therefore, there is a high demand for data processing technology.

The improved defect prediction model for process industrial products has obtained the attribute importance order and defect prediction rule set based on attribute division according to the business process. The processing workshop can obtain the integrated attribute importance ranking and production guidance rule set applied to each intermediate product production process. Experiment 2 shows that when the temperature difference, pull speed, carbon oxide content, manganese content, vanadium content, and phosphorus content exceed the critical threshold, all rails are qualified. However, when the carbon oxide content or manganese content falls below the critical threshold, the rail qualification rate significantly decreases. Comparing the algorithm accuracy of Experiment 1 and Experiment 2, it can be observed that the test set accuracy of the random forest experiment increased from 88.39 % to 92.69 %, and the test set accuracy of the decision tree experiment increased from 71.89 % to 79.15 %. The machine learning accuracy has been significantly improved after attribute division.

Excluding the influence of external factors on the accuracy of experiment, such as environmental factors (temperature, humidity, precipitation, industrial pollution) which can cause systematic deviations in sensor data, and the coupling effect of the interaction between environmental and industrial parameters, the accuracy of machine learning models will be further improved. We leave the deviations caused by these external factors for future research. Our improved algorithm guides enterprise practices, and at the same time, enterprises should cooperate with the optimization of data integration and process collaboration. For example, establish a unified data platform to avoid the problem of information silos. Formulate data governance strategies (such as the ISO 8000 standard) and deploy automated cleaning tools to achieve data standardization. Processes such as quality inspection and handling of non-conforming products involve multi-departmental collaboration. Traditional linear processes are difficult to adapt to the dynamic requirements of machine learning models. It is recommended to adopt a modular design and decompose the processes into configurable sub-modules.

In conclusion, in actual production, it is necessary to consider whether industrial data has process relevance and to divide attributes according to the quality inspection procedures and the sequence of production processes. Machine learning is then applied on this basis to identify issues. The experimental results prove that the defect product prediction method proposed in this study is more suitable for industrial big data mining. The management insights obtained based on actual production data have more guiding significance for actual production.

5. Conclusion

This study has developed a new data analysis framework for the process industry. This framework systematically classifies the attributes of process properties to alleviate the data interdependence effect, integrates block-based decision trees and random forest models to generate interpretable defect prediction rules, and identifies key quality attributes through feature importance analysis. Experimental validation using rail production data from a major Chinese steel mill demonstrates significant performance improvements: the proposed attribute-partitioning approach increases random forest test accuracy from 88.39 % to 92.69 % and decision tree accuracy from 71.89 % to 79.15 %. The framework reveals actionable quality rules - when the temperature difference, pull speed, carbon oxide content, manganese content, vanadium content, and phosphorus content exceed the critical threshold, all rails are qualified. While, when the carbon oxide content or manganese content falls below the critical threshold, the rail qualification rate significantly decreases.

By transforming high-dimensional correlated production data into process-aligned modular datasets, the framework demonstrates superior applicability for process-driven industries compared to conventional methods, delivering both analytical precision and interpretable operational guidance that bridges theoretical data mining with industrial practice. In the future, the defect prediction problem of multiple products can be explored, and the data analysis methods for products with overlapping processes can be considered.

References

[1] Qin, S.J. (2014). Process data analytics in the era of big data, *AIChE Journal*, Vol. 60, No. 9, 3092-3100, doi: 10.1002/aic.14523.

[2] Topal, B., Sahin, H. (2018). The influence of information sharing in the supply chain process on business performance: An empirical study, *Studies in Informatics and Control*, Vol. 27, No. 2, 203-214, doi: 10.24846/v27i2y201808.

[3] Nagy, R., Horvát, F., Fischer, S. (2024). Innovative approaches in railway management: Leveraging big data and artificial intelligence for predictive maintenance of track geometry, *Tehnički Vjesnik – Technical Gazette*, Vol. 31, No. 4, 1268-1276, doi: 10.17559/TV-20240420001479.

[4] Nguyen, T.V., Zhou, L., Chong, A.Y.L., Li, B., Pu, X. (2019). Predicting customer demand for remanufactured products: A data-mining approach, *European Journal of Operational Research*, Vol. 281, No. 3, 543-558, doi: 10.1016/j.ejor.2019.08.015.

[5] Borchert, P., Coussement, K., De Weerdt, J., De Caigny, A. (2024). Industry-sensitive language modeling for business, *European Journal of Operational Research*, Vol. 315, No. 2, 691-702, doi: 10.1016/j.ejor.2024.01.023.

[6] Wu, Z., Shi, Y. (2024). Development and digitalization of cultural industry marketing based on big data, *Environmental Engineering and Management Journal*, Vol. 23, No. 5, 1097-1108, doi: 10.30638/eemj.2024.089.

[7] Kovacic, M., Zuperl, U., Gusel, L., Brezocnik, M. (2023). Reduction of surface defects by optimization of casting speed using genetic programming: An industrial case study, *Advances in Production Engineering & Management*, Vol. 18, No. 4, 501-511, doi: 10.14743/apem2023.4.488.

[8] Perzyk, M., Kochanski, A., Kozlowski, J., Soroczynski, A., Biernacki, R. (2014). Comparison of data mining tools for significance analysis of process parameters in applications to process fault diagnosis, *Information Sciences*, Vol. 259, 380-392, doi: 10.1016/j.ins.2013.10.019.

[9] Hsu, S.C., Chien, C.F. (2007). Hybrid data mining approach for pattern extraction from wafer bin map to improve yield in semiconductor manufacturing, *International Journal of Production Economics*, Vol. 107, No. 1, 88-103, doi: 10.1016/j.ijpe.2006.05.015.

[10] Norrena, J., Louhenkilpi, S., Visuri, V.V., Alatarvas, T., Bogdanoff, A., Fabritius, T. (2023). Assessing the effects of steel composition on surface cracks in continuous casting with solidification simulations and phenomenological quality criteria for quality prediction applications, *Steel Research International*, Vol. 94, No. 5, Article No. 2200746, doi: 10.1002/srin.202200746.

[11] Hochbaum, D.S., Liu, S. (2018). Adjacency-clustering and its application for yield prediction in integrated circuit manufacturing, *Operations Research*, Vol. 66, No. 6, 1457-1759, doi: 10.1287/opre.2018.1741.

[12] Vukelic, D., Milosevic, A., Ivanov, V., Kocovic, V., Santosi, Z., Sokac, M., Simunovic, G. (2024). Modelling and optimization of dimensional accuracy and surface roughness in dry turning of Inconel 625 alloy, *Advances in Production Engineering & Management*, Vol. 19, No. 3, 371-385, doi: 10.14743/apem2024.3.513.

[13] Chongwatpol, J. (2015). Prognostic analysis of defects in manufacturing, *Industrial Management & Data Systems*, Vol. 115, No. 1, 64-87, doi: 10.1108/IMDS-05-2014-0158.

[14] Lee, C.K.H., Ho, G.T.S., Choy, K.L., Pang, G.K.H. (2013). A RFID-based recursive process mining system for quality assurance in the garment industry, *International Journal of Production Research*, Vol. 52, No. 14, 4216-4238, doi: 10.1080/00207543.2013.869632.

[15] Agarwal, S., Dandge, S.S., Chakraborty, S. (2019). Development of association rules to study the parametric influences in non-traditional machining processes, *Sadhana: Academy Proceedings in Engineering Sciences*, Vol. 44, Article No. 230, doi: 10.1007/s12046-019-1218-6.

[16] Breznikar, Z., Bojinovic, M., Brezocnik, M. (2024). Application of machine learning to reduce casting defects from bentonite sand mixture, *International Journal of Simulation Modelling*, Vol. 23, No. 4, 634-643, doi: 10.2507/IJSIMM23-4-702.

[17] Ciarapica, F., Bevilacqua, M., Antomarioni, S. (2019). An approach based on association rules and social network analysis for managing environmental risk: A case study from a process industry, *Process Safety & Environmental Protection*, Vol. 128, 50-64, doi: 10.1016/j.psep.2019.05.037.

[18] Li, Q., Li, D., Cao, L. (2015). Modeling and optimum operating conditions for FCCU using artificial neural network, *Journal of Central South University*, Vol. 22, No. 4, 1342-1349, doi: 10.1007/s11771-015-2651-2.

[19] Irani, K.B., Cheng, J., Fayyad, U.M., Qian, Z. (1993). Applying machine learning to semiconductor manufacturing, *IEEE Expert*, Vol. 8, No. 1, 41-47, doi: 10.1109/64.193054.

[20] Keswani, M. (2024). Designing a fuzzy logic-based carbon emission cost-incorporated inventory model: A comparative analysis of different machine learning algorithms for demand forecasting with memory effects, *Economic Computation and Economic Cybernetics Studies and Research*, Vol. 58, No. 4, 143-158, doi: 10.24818/18423264/58.4.24.20.

[21] Li, Z., Tie-Xin, C., Ying, M., Qi, L., Liu, M. (2016). Decision tree data mining model for welding parameters selection based on C5.0 improved algorithm and its application, *Chinese Journal of Management Science*, Vol. 2016, No. S1, 230-236.

[22] Grădinaru, G.-I., Manea, D.-I., Andreescu, F., Toma, D.-A., Paraschiv, L.-I. (2024). Identifying the main factors of elaborating "Smart City" strategy using machine learning: A comparative study among Romanian cities, *Economic Computation and Economic Cybernetics Studies and Research*, Vol. 58, No. 3, 53-71, doi: 10.24818/18423264/58.3.24.04.

[23] Özbalci, O., Çakir, M., Oral, O., Doğan, A. (2023). Machine learning approach to predict the effect of metal foam heat sinks discretely placed in a cavity on surface temperature, *Tehnički Vjesnik – Technical Gazette*, Vol. 31, No. 6, 2003-2013, doi: 10.17559/TV-20240302001366.
[24] Wang, T., Wang, X., Ma, R., Li, X., Hu, X., Cahn, F.T.S. (2020). Random Forest-Bayesian optimization for product quality prediction with large-scale dimensions in process industrial cyber–physical systems, *IEEE Internet of Things Journal*, Vol. 7, No. 9, 8641-8653, doi: 10.1109/JIOT.2020.2992811.
[25] Esteve, M., Aparicio, J., Rodriguez-Sala, J.J., Zhu, J. (2022). Random forests and the measurement of super-efficiency in the context of free disposal hull, *European Journal of Operational Research*, Vol. 304, No. 2, 729-744, doi: 10.1016/j.ejor.2022.04.024.
[26] Han, J.H., Lee, J.Y. (2023). Genetic algorithm-based approach for makespan minimization in a flow shop with queue time limits and skipping jobs, *Advances in Production Engineering & Management*, Vol. 18, No. 2, 152-162, doi: 10.14743/apem2023.2.463.
[27] Stojic, N., Delic, M., Bojanic, T., Jokanovic, B., Tasic, N. (2024). Integrated model of risk management in business processes in industrial systems, *International Journal of Simulation Modelling*, Vol. 23, No. 3, 412-423, doi: 10.2507/IJSIMM23-3-689.
[28] Wei, Z.H., Yan, L., Yan, X. (2024). Optimizing production with deep reinforcement learning, *International Journal of Simulation Modelling*, Vol. 23, No. 4, 692-703, doi: 10.2507/IJSIMM23-4-CO17.
[29] Yao, L., Ge, Z. (2018). Big data quality prediction in the process industry: A distributed parallel modeling framework, *Journal of Process Control*, Vol. 68, 1-13, doi: 10.1016/j.jprocont.2018.04.004.
[30] Negoiță, R.F., Borangiu, T. (2023). Robotic process automation of inventory demand with intelligent reservation, *Studies in Informatics and Control*, Vol. 32, No. 2, 5-14, doi: 10.24846/v32i2y202301.
[31] Novak, C., Pfahlsberger, L., Bala, S., Revoredo, K., Mendling, J. (2023). Enhancing decision-making of IT demand management with process mining, *Business Process Management Journal*, Vol. 29, No. 8, 230-259, doi: 10.1108/BPMJ-12-2022-0631.
[32] Ding, J., Liu, Y., Zhang, L., Wang, J., Liu, Y. (2016). An anomaly detection approach for multiple monitoring data series based on latent correlation probabilistic model, *Applied Intelligence*, Vol. 44, No. 2, 340-361, doi: 10.1007/s10489-015-0713-7.
[33] Adnan, M.N., Islam, M.Z. (2016). Optimizing the number of trees in a decision forest to discover a subforest with high ensemble accuracy using a genetic algorithm, *Knowledge-Based Systems*, Vol. 110, 86-97, doi: 10.1016/j.knosys.2016.07.016.