IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 1, 2020, accepted September 8, 2020, date of publication September 18, 2020,
date of current version September 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024582

Fault Location of Strip Steel Surface Quality Defects on Hot-Rolling Production Line Based on Information Fusion of Historical Cases and Process Data

ZHAOPING WANG, JIAN WANG, AND SEN CHEN
Computer Integrated Manufacturing System Research Center, College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China
Corresponding author: Zhaoping Wang (1193671148@qq.com)

This work was supported by the National Science and Technology Innovation 2030 of China Next-Generation Artificial Intelligence Major Project, Data-Driven Tripartite Collaborative Decision-Making and Optimization, under Grant 2018AAA0101801.

ABSTRACT Surface quality is the most important index to improve the overall quality of strip steel. In order to implement the fault location on the hot-rolling line with surface defects of strip steel, a fault tracing model based on information fusion of historical production cases and process data is proposed. For historical cases, the model determines the defect cause labels through text similarity calculation, and fuzzy semantic inference is used to obtain the probability distribution of defect causes on this basis; for the process data, the model uses L1 regularization method for feature selection, and XGBoost integration method is used to train the correlation model between process data and defects to determine the contribution of each feature in the data source. Finally, based on the D-S evidence theory, different rules are set to merge the two judgments to determine the probability of each source of failure on the hot-rolling production line. The model is applied to the real production environment of iron and steel enterprises, and it is verified that the proposed method can effectively assist experts in decision-making, which greatly improves the efficiency of tracing the source of faults on the hot-rolling production line.

INDEX TERMS Fault location, fuzzy semantic inference, process data analysis, feature selection, feature importance, information fusion.

I. INTRODUCTION

The production technology of strip steel is an important symbol of the development level in the steel manufacturing industry. Strip steel has been widely used in the automotive industry, defense industry, aerospace, chemical equipment and light industry manufacturing due to its good surface quality and mechanical properties, and the downstream users have increasingly higher requirements on the surface quality of strip steel [1], [2]. According to statistics, the majority of quality objections and complaints in recent years are related to surface quality defects, so how to effectively control strip surface defects and improve the quality has become the main task of product quality improvement in the steel manufacturing industry [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaochun Cheng.

Due to the influence of many factors during the production such as raw materials, rolling process, equipment problems, and system control, defects like roll marks, scratches, and stains often occur on the surface. These defects have different degrees of influence on the wear resistance, fatigue resistance, corrosion resistance and electromagnetic properties of strip steel. In production, surface defects are not only likely to cause serious production accidents such as tape breakage, accumulation, and parking, but also seriously wear out the rollers possibly, causing inestimable economic and social impacts on production enterprises [4].

According to the mechanism of defect generation, the surface quality defects of strip steel can be divided into material defects, process defects and corrosion defects. Material defects refer to problems with the blank itself; process defects mainly mean defects caused by improper operation and equipment failure during the rolling process, and

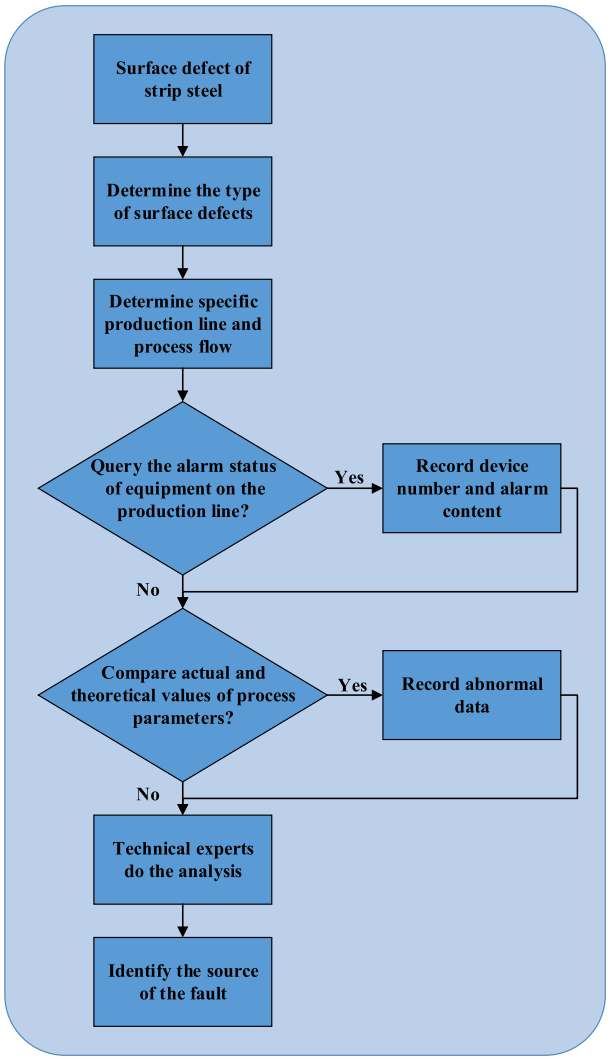

FIGURE 1. Traditional strip quality traceability process.

corrosion defects refer to the strip steel affected by the external environment. In order to control surface defects in time, it is necessary to locate the source of the fault according to the symptoms of the defect. The fault location technology is the basis for repairing production faults and the premise for ensuring efficient and stable operation of the whole system [5].

The quality defects of strip steel can be detected by the on-site staff through naked eye observation or related instruments, but in order to solve the problem thoroughly, it is necessary to trace the source of failure as soon as possible [6]. The traditional strip quality traceability process is shown in Figure 1. The fault location is usually determined by the analysis of experts according to the equipment alarm and process parameter comparison results.

The problems in the traditional fault location process are mainly reflected in the following two points:

(1) High production automation results in huge data volume, and it is difficult to distinguish fault-related features. Manual analysis is inefficient.

(2) In complex practical business scenarios, the differences in the experience of technicians and the asymmetry of information between departments have led to different fault analysis standards and processes, which are highly subjective.

At present, the existing fault location methods are mainly based on a single data source, especially the data from sensors on the production line. Therefore, a fault location model for strip surface quality defects is proposed in this paper based on historical cases and process data. The historical cases and process data were analyzed respectively, and then D-S evidence theory was used to fuse the two judgments to determine the probability of each fault source on the production line finally. The fusion judgment method based on multi-source data makes the judgment more objective and comprehensive, and improves the time lag of the traditional method through automatic analysis.

The remaining parts of this paper are organized as follows. Section II reviews the literature on the data mining technology and big data applications on fault location. In Section III, fault location method based on information fusion is proposed. Section IV describes experiments of fault location to evaluate performance of the proposed method. Finally, the conclusion and prospect are given in Section V.

II. RELATED WORK

With the development of machine learning and data mining, data-driven fault location methods have gradually become the research hotspot and development direction in this field. The idea behind methods is to provide reliable evidence for solving practical problems in industry through the analysis of process data and the excavation of essential information, which greatly reduces the reliance on accurate mathematical models and expert experience [7]. This section mainly introduces the related work of data-driven fault location research.

Many experts and scholars have made relevant explorations in this field and obtained many valuable research results. Some research work mainly focuses on extracting effective fault features and then constructing classifiers to identify equipment faults [21]. A method of fault feature extraction based on intrinsic mode function (IMF) envelope spectrum is proposed by Yang *et al.* [8] and the support vector machine (SVM) classifiers was used to provide the possibility of machinery faults. And Georgoulas *et al.* [9] analyzed the motor fault characteristics and identified them with time-frequency characteristics, in this research Markov distance classifier was used to identify the motor health. A new vibration spectral imaging (VSI) feature enhancement method was proposed by Amar *et al.* [10] under the condition of low signal-to-noise ratio, and artificial neural network (ANN) was used as the fault classifier according to these enhanced fault features. Prieto *et al.* [11] proposed a method for classifying the health status of bearing hierarchically using neural networks and statistical characteristics were taken into account. Based on spectral kurtosis (SK) and cross correlation, a method was proposed by Tian *et al.* [12] to extract fault features and form a health index using principal

component analysis (PCA) and a semi-supervised k-nearest neighbor (KNN) distance measure. However, it is difficult to avoid the limitation of single data source and necessary consideration is not always given to the complexity of production environment [21].

Industrial production lines often have related data from different systems in order to manage the whole production process [22]. Many researches tend to use the fusion of multi-source information to make fault location decisions. Sun *et al.* [13] proposed a production line fault diagnosis and maintenance decision system based on human-machine multi-information fusion, and used multi-source data to conduct the final fault location. Lyu *et al.* [14] proposed a six-step data-driven solution with decision tree and associate rules to determine the causes of product defects and the product defect rate decreased from 20% to 5%. And Pei *et al.* [15] contributed to the fault discovery by proposing an integrated approach combining the Taguchi quality loss function (QLF), the signal-noise ratio (SNR), and the relief method. In general, the fusion of multi-source information can make decisions on fault detection and location in production line more convincing and robust.

The steel hot rolling production line has the characteristic of typical industrial production line under the environment of intelligent manufacturing, so it is worthy of further study in many aspects, such as fault detection and location. Ding *et al.* [16] proposed a data-driven scheme of key performance indicator (KPI) prediction and diagnosis, which was applied to an industrial hot strip mill to improve the prediction performance. A new kernel independent and principal components analysis (ICA–PCA) based process monitoring approach [17] and a new statistical monitoring technique based on efficient projection to latent structures (EPLS) for quality-relevant fault detection [18] were proposed by Peng and his team in succession [22]. And in 2020 a framework was also proposed by Zhang *et al.* [19] for quality-based fault detection and diagnosis for nonlinear batch processes with multimode operating environment. But their all research work focused on quality-relevant fault monitoring in the hot strip mill process. Zhang *et al.* [20] address the problem of fault diagnosis in the aspect of nonlinear activation in hot rolling automation system by using a KPCA-based method. The detection is achieved by comparing the subspaces between the reference and a current state of the system.

The motivation to choose the information fusion method is mainly to solve the practical problems in hot rolling production from multiple dimensions and perspectives. It is a prominent feature of hot rolling production process to generate a large amount of multi-source data. Through information fusion, valuable information in multi-source data can be integrated to improve the feasibility of decision making.

III. RESEARCH METHODOLOGY

This section proposes a fault location model for strip surface quality defects based on historical cases and process data. The historical cases and process data were analyzed respectively, and then D-S evidence theory was used to fuse the two judgments to determine the probability of each fault source on the production line finally. The overall architecture of this model is shown in Figure 2.

A. DEFECT CAUSE CLASSIFICATION AND FUZZY SEMANTIC INFERENCE

The traditional traceability process of surface quality defects on strip steel mostly ignores the role of historical solutions. In the historical production process, experienced technical experts have identified, analyzed, processed, and solved a series of such problems, and at the same time formed a large number of solution-related documents stored in the enterprise's information system. This kind of data is accumulated from the experience of human activities in the real production environment and it often contains a lot of rules and knowledge [23]. Historical solutions can be excavated to provide auxiliary decision-making for experts, but it is difficult to extract effective information due to the unstructured and diverse characteristics of the textual data.

In the history cases of strip surface faults, the description of defect symptoms and causes is often subjective and vague in semantic expression. For example, description one is ''The heating temperature is too high and the heating time is too long. The oxidizing atmosphere in the furnace is too thick, which forms serious oxide sheet'' and description two is ''The rolling temperature is too high and regenerated oxide sheet is easy to be pressed into plate'', both of these descriptions represent the defect symptom called rolled-in scale. Therefore, it is difficult to directly determine the corresponding relationship.

Targeting the current situation where expert experience is difficult to reuse automatically, this paper collects and integrates historical solutions related to surface quality defects to form a structured historical case document, which contains multiple cases of fault location of the same type. The historical case document is the summary of a maintenance event of strip surface fault location, which mainly includes recording surface defect symptoms, initial inspection and detection, in-depth fault analysis, proposal and implementation of solutions, and final summary. And it can be defined as a unified format:

CaseID (*CaseTime, DefectCategory, DefectDescription, DefectCause*)

This unified format describes the core content of the historical case and provides a basic framework for fuzzy semantic inference. CaseID represents the unique identifier of the case about surface quality defects, *CaseTime* represents the time of occurrence of the case, and *DefectCategory*, *DefectDescription*, *DefectCause* respectively indicate the corresponding defect category, defect detailed description, and cause of the defect.

In order to characterize the association between defect categories and causes, *DefectCause* in the historical case documents should be classified firstly, so that the source of each case can be identified with a clear label. The cause of defects

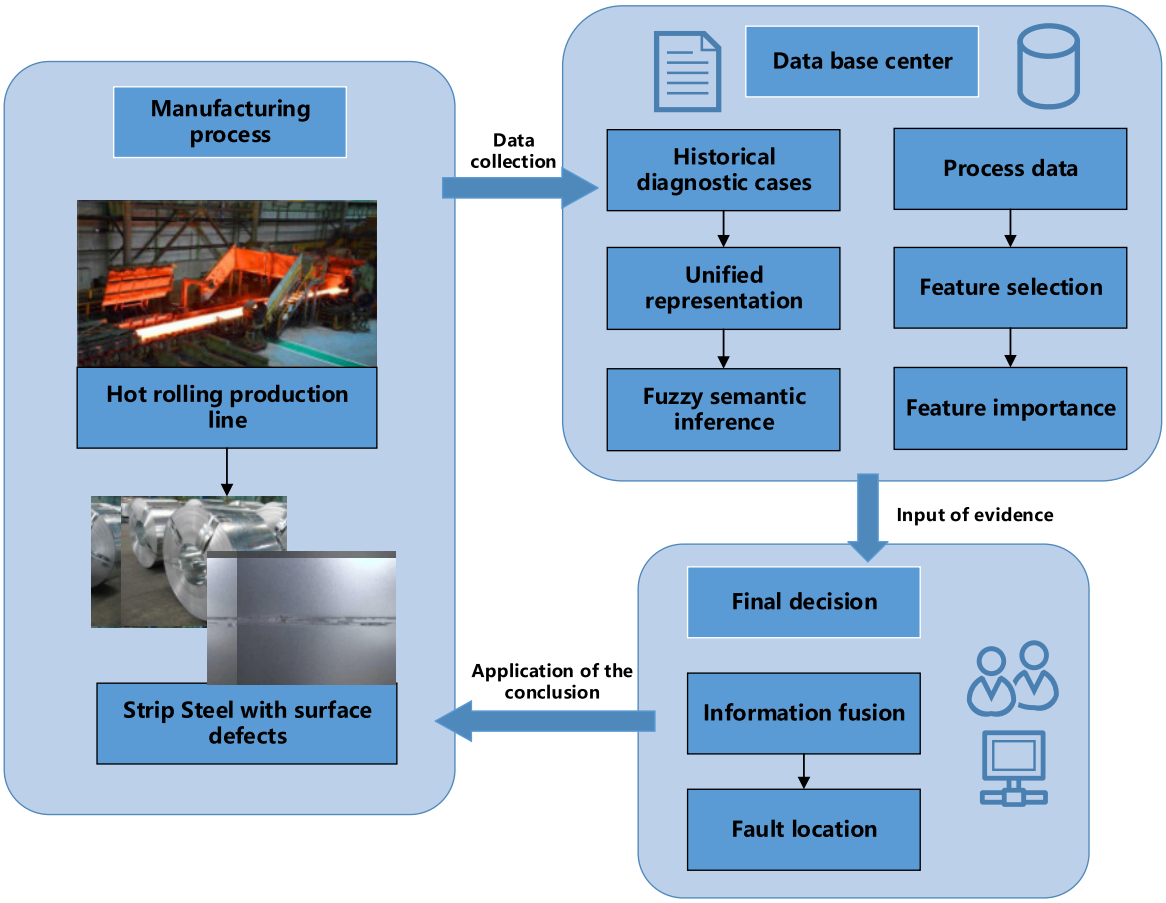

FIGURE 2. The overall architecture of the model proposed in this paper.

often involves the description of operating equipment and process. Entities in the hot-rolling field occur frequently, and each equipment has a standard process description document that can be referred to, therefore the degree of association between the two can be described by calculating text similarity. TF-IDF (Term Frequency-Inverse Document Frequency) algorithm is a statistical analysis method for keywords, which is used to evaluate the importance of a word to a corpus. In this paper, TF-IDF algorithm is adopted to select *DefectCause* and key words in device reference documents, paving the way for text similarity calculation.

Assume that the document to be processed is d_j, and the frequency of each word can be calculated after word segmentation. The word frequency is defined as follows:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

where $n_{i,j}$ is the occurrence times of the word in document d_j, and the denominator of Eq. (1) is the sum of the occurrence times of all words in document d_j. It can be seen that word frequency is the normalization of word number to measure the importance of the word. And the inverse document frequency IDF value is a measure of the general importance of one word in a corpus, which can be defined as:

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}| + 1} \tag{2}$$

In the Eq. (2), D represents the number of documents in the corpus composed of defect causes and standard reference documents, and j is the number of documents containing the word t_i, which can then be obtained as follows:

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{3}$$

High word frequency in a particular document, and low document frequency in the entire corpus, can always produce a high-weight TF-IDF value. Therefore, TF-IDF algorithm tends to filter out common words and retain important words, which are often professional terms in hot-rolled field. After the calculation of TF-IDF value, the words in two documents are arranged in descending order according to the results, and keywords in the front are selected referring to the hot-rolling field glossary. Then the word vector $\boldsymbol{s}$ and $\boldsymbol{t}$ are obtained respectively. The cosine similarity of the two vectors is shown in Eq. (3):

$$\cos(\theta) = \frac{\sum_{i=1}^{m} (s_i \times t_i)}{\sqrt{\sum_{i=1}^{m} s_i^2} \times \sqrt{\sum_{i=1}^{m} t_i^2}} \tag{4}$$

TABLE 1. Fuzzy score reference.

Defect level	Very serious	Quite serious	Serious	Normal	Light
Fuzzy score	0.8—1	0.6—0.8	0.4—0.6	0.2—0.4	0—0.2

Cosine similarity measures the difference between two individuals using cosine of angle between two vectors, and it pays more attention to the difference between two vectors in direction than Euclidean distance. The larger the cosine similarity of the two vectors, the higher the similarity between the defect cause and the equipment standard reference document, so it can be considered that hot-rolling production line equipment corresponding to maximum value is the fault source.

Strip surface defects often have similar symptoms and causes when they occur at a similar time, the accuracy of marking can be improved by the judge of *CaseTime* and *DefectDescription*. If the interval of *CaseTime* is less than one day and the text similarity of *DefectDescription* is higher than the preset threshold, the category of defect causes can be considered identical.

Fuzzy set theory and fuzzy logic are suitable forms to deal with imprecise semantic knowledge [24], [25]. In the real environment, the surface defects of strip steel are often complex, and there is not a one-to-one correlation between defect types and causes, or even a coexistence of multiple defect types, which makes it more difficult to trace source of fault [26].

Fuzzy semantic inference is to mine the fuzzy logic between input and output through semantics, and the inference rules are obtained by semantics to make decision support, so as to reduce the dependence on expert experience. In the previous chapter A, the historical cases of steel strip surface defect are divided into *CaseTime, DefectCategory, DefectDescription* and *DefectCause* four parts according to the semantics, and the corresponding standard form has been built by similarity calculation. And then the fuzzy relation matrix can be established through the statistical analysis of semantic relationships, at last the mapping from fuzzy defect symptom vector to result vector can be completed. The specific steps are described below. The entire process of fuzzy semantic inference is shown in Figure 3.

Assume $\boldsymbol{S}$ represents a set of defect symptoms, $\boldsymbol{S} = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_n\}$, and $\boldsymbol{s_i}$ represents the state variable of $\boldsymbol{S_i}$. This paper gives a fuzzy score reference for the severity of defect symptoms determined by expert experience [28], as shown in Table 1, so the fuzzy defect symptom vector $\boldsymbol{f}$ can be expressed as:

$$f = [f_1, f_2, \cdots, f_n] = [\mu_{S_1}(s_1), \mu_{S_2}(s_2), \cdots, \mu_{S_n}(s_n)] \quad (5)$$

where $\mu_{\mathbf{Si}}(\boldsymbol{s_i})$ is the membership function, and its value represents the membership of the current state. According to the fuzzy scoring criteria of Table 1, quality inspection personnel on hot- rolling production line can evaluate steel strip surface defects, determining membership degree of each defect symptoms. Fuzzy vector $\boldsymbol{f}$ can be used as input in the whole process of semantic inference, so the human experience in the production process can be quantified as a specific value to deal with the problem of uncertainty.

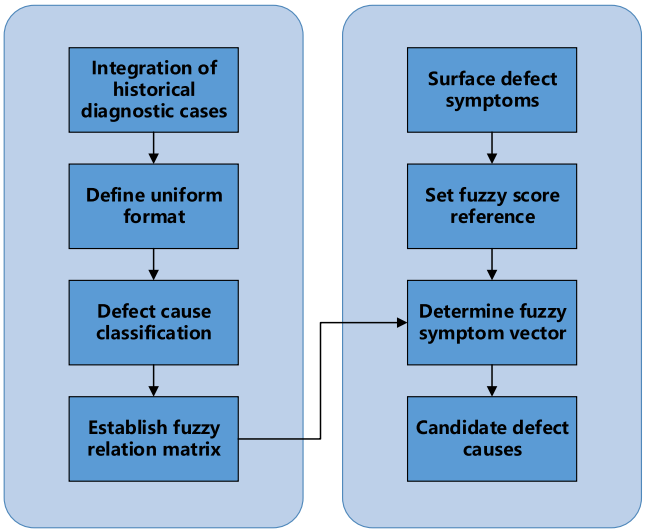

FIGURE 3. Flowchart of fuzzy semantic inference.

The relationship between defect symptoms and defect causes is represented by the fuzzy relation matrix R, as shown in Eq. (6), which depends on the statistical results of historical cases [27].

$$\mathrm{R} = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{pmatrix}, \quad 0 \le r_{ij} \le 1, \quad 1 \le i \le m, \quad 1 \le j \le n \quad (6)$$

The rows of R represent defect causes, and the columns represent defect symptoms. The element $\boldsymbol{r_{ij}}$ means degree of membership of symptom $\boldsymbol{j}$ to cause $\boldsymbol{i}$ [28]. A typical statistic method to determine $\boldsymbol{r_{ij}}$ as follows:

$$r_{ij} = \frac{n_{ij}}{n_j} \quad (7)$$

where $\boldsymbol{n_{ij}}$ represents the number of historical records of symptom $\boldsymbol{j}$ belonging to cause $\boldsymbol{i}$, and $\boldsymbol{n_j}$ represents the number of historical records that belong to cause $\boldsymbol{i}$. Then fuzzy semantic inference is carried out according to fuzzy defect symptom

vector $\boldsymbol{f}$ and fuzzy relation matrix R, as shown in Eq. (8):

$$v = f \circ R = (f_1, f_2, \cdots, f_m) \circ \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{pmatrix} = (v_1, v_2, \cdots, v_m) \tag{8}$$

$$v_j = \min(1, \sum_{i=1}^{m} f_i \cdot r_{ij}), \quad 1 \le j \le n \tag{9}$$

The symbol "$\circ$" in Eq. (8) is the fuzzy synthetic operator, whose operation logic is shown in Eq. (9). The element $\boldsymbol{v_i}$ in the output vector $\boldsymbol{v}$ represents the membership degree of the current symptom corresponding to defect cause $\boldsymbol{i}$. Finally, the output vector $\boldsymbol{v}$ is normalized, as shown in Eq. (10):

$$p_i = \frac{v_i}{\sum_{i=1}^{m} v_i}, \quad p = (p_1, p_2, \cdots, p_m) \tag{10}$$

The element $\boldsymbol{p_i}$ in vector p represents the probability of defect cause $\boldsymbol{i}$ corresponding to the current defect symptom. The defect cause with maximum value of $\boldsymbol{p_i}$ can be selected as the source of fault, or candidate defect causes can be determined according to the designed threshold. In this paper, the result of fuzzy semantic inference is taken as one evidence in final information fusion.

B. EVALUATION OF FEATURE CONTRIBUTION BASED ON XGBOOST

Hot-rolling production line has produced a large amount of process data during the actual production, these data include the basic attributes of slab (such as weight, length and thickness of slab, etc.), all kinds of processing technology parameters of the equipment (such as melt temperature, roughing thickness and finishing width, etc.), environmental information from sensors (such as temperature, humidity and pressure, etc.), as well as the quality of strip steel (such as average crown, thickness deviation, surface quality, etc.). The surface quality in each production record can be regarded as a label for training. Problems with slab, equipment failure, and improper setting of process parameters are all associated with process data, and eventually lead to surface quality problems of strip. By establishing the correlation model between process data and defects of strip steel, the characteristics of the process data can be deduced from the defects, so as to provide important data support for determining the fault source.

It is assumed that n rolls of strip are produced in the hot-rolling production line in a period of time, and the process data features generated during the processing of each steel strip have m dimensions, then the sample matrix $X \in \Re^{n \times m}$ can be obtained through collection and integration, where the row i represents the strip sample i and the column j represents features j of process data. Min-max Normalization method is used to normalize raw data and surface quality is considered as a label for data training.

Part of the process data features have a highly consistent change trend and data redundancy. In order to conduct efficient training, it is necessary to make feature selection for data. In this paper, the embedded type feature selection method is adopted, and the linear regression model with L1 regularization penalty term is used as the base model for feature selection, and the square error is used as the loss function. The optimization objective is shown in Eq. (11):

$$w' = \arg\min_{w} \sum_{i=1}^{m} (y_i - w^T x_i) + \lambda \|w\|_1 \tag{11}$$

where w is the weight matrix and λ is the regularization parameter. L1 regularization is easy to obtain sparse solutions, and the result of solving norm regularization is to get a model that uses only a part of the initial features, that is, the feature selection process is integrated with the training process. When w gets a sparse solution, it means that only the features corresponding to non-zero components in the initial features will appear in the final model, thus realizing feature selection.

Compared to other machine learning methods, XGBoost method is more suitable for industrial big data scenarios with the advantage of distributed computing [29], [30]. Therefore, the XGBoost training model is used in this paper to fit the correlation between strip steel process data and surface defects. CART decision tree is used as the base learner in boosting method, and grid search method is applied to determine the learning_rate, n_estimatores, tree_depth, subsample, max_depth and other key parameters. The classification model is obtained after training.

In general, the importance score of a feature measures its value in decision tree construction, and the more a feature is used to build a decision tree in the model, the more important it is. The importance of features is obtained by calculating each feature in process data separately and then sorting it. The importance of a feature is calculated by the degree to which each feature split point improves the performance measurement, and the *Gini* coefficient is selected to measure the performance:

$$Gini(D) = \sum_{k=1}^{N} p_k (1 - p_k) \tag{12}$$

where D represents the subset of samples corresponding to a branch in the decision tree, N is the number of defect categories in D, and p_k represents the proportion of defect type k in the entire subset of samples. According to the feature F, D can be divided into two subsets D_1 and D_2:

$$Gini(D, F) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{13}$$

$$Gain(A) = Gini(D, F) - Gini(D) \tag{14}$$

Finally, the weighted sum of the calculation result of one feature in all decision trees is taken to get the importance score. Therefore, the importance degree reflects the correlation between the feature and the surface defect of strip steel, that is, the greater the importance degree is, the more likely the equipment corresponding to the feature will have

problems. Each feature of the process data sample is derived from one equipment in the hot-rolling production line or from the slab itself, so the importance of multiple features from the same source can be added up as the final feature contribution degree, and the probability can be obtained by normalization. The value reflects the potential causal relationship between defect cause and defect category.

C. INFORMATION FUSION

Through fuzzy semantic inference based on historical cases and analysis of process data, two evidences of possible problems in some equipment of hot-rolling production line are obtained. The method of information fusion is needed to synthesize the judgment of two pieces of evidence and deal with the contradictions. Finally, the probability of problems in a certain equipment can be obtained. Therefore, this paper sets fusion rules based on D-S evidence theory to conduct information fusion [31].

According to the scenario in this paper, the identification framework θ of D-S evidence theory is a set of all devices involved in the strip production process, and the Basic Probability Assignment function is defined as:

$$\sum_{A\subseteq\theta} m(A) = 1, \quad m(\emptyset) = 0 \tag{15}$$

The fusion rule based on two pieces of evidence can be defined as Eq. (16) and Eq. (17):

$$m_1(A) \oplus m_2(A) = \frac{1}{R} \sum_{B\cap C=A} m_1(B)m_2(C) \tag{16}$$

$$R = \sum_{B\cap C\neq\emptyset} m_1(B)m_2(C) \tag{17}$$

where R represents the normalized coefficient, A is an element of θ, and B or C is a subset of θ. When the contradiction between two pieces of evidence is slight and equally important, information fusion is carried out according to the rule in Eq. (16) to calculate the final probability of defect cause.

However, the above two pieces of evidence may have different weights in the real production environment, and their influence on the final result may be in dynamic change. Therefore, another fusion rule is set as shown in Eq. (18):

$$m_1(A) \oplus m_2(A) = \alpha \cdot m_1(A) + (1-\alpha)m_2(A) \tag{18}$$

where α is the weight coefficient, which represents the trust degree of technical experts to the evidence. Experienced technicians can assign the weight by setting the value of α, which is more suitable for rapid positioning requirements in the field environment.

IV. EXPERIMENT AND ANALYSIS

This section introduces the case study background and verifies the validation of the proposed method in this paper. The results are analyzed at the end of this section.

A. PROBLEM DESCRIPTION

A steel company's 1580mm hot-rolling production line in China mainly produces high-quality strip steel. However, the surface defects of strip steel have occurred from time to time due to various factors. Whether it can quickly locate cause of the defect and reduce the further generation of defective products has a huge impact on the company's production activities.

The main equipment configuration of the production line is shown in Figure 4, including: three walking beam heating furnaces, roughing entrance descaling box, fixed width large side press, E1/R1 roughing mill, E2/R2 roughing mill, EH edge heating Furnace, flying shears, descaling box at the entrance of finish rolling, small vertical rollers for finish rolling, seven-stand finishing mill, laminar cooling and two underground coilers. R1 is a two-roller reversing mill, and R2 is a four-roller reversing mill. The finishing mills F2-F4 use a PC cross-rolling mill and F5-F7 mill flat rolls can move.

Surface quality is one of the most difficult quality indicators to control to improve the overall quality level of the company. The categories and causes of strip surface defects involved in the hot- rolling line are shown in Table 2 and Table 3:

B. CASE STUDY

The historical cases and process data used in this paper were all from the hot-rolling production line of this steel company in 2016. There were more than 400 cases of strip surface defects after collection and integration, mainly recording the whole process from discovering the surface defect by quality inspection to determining the cause. Firstly, the defect categories defined by experts and the description of the cause were extracted, then according to the proposed methods in section III. A, the textual similarity between defect case and standard document could be calculated to determine the label of defect cause, and then the fuzzy relationship matrix shown in Table 4 was statistically derived according to historical cases.

In order to verify the effectiveness of fuzzy inference, technical experts set fuzzy defect symptom vectors for five groups of strips with common surface defects according to Table 1, and the specific values were shown in Table 5. The result vectors were calculated to determine the fault sources according to Eq. (8), and the results compared with experts' opinions were shown in Table 6. It can be found that the probabilistic Top 3 selections all hit the conclusions given by experts.

5032 strip steel production data were selected as experimental samples, including slab properties, process parameters, quality inspection results and other information. Each row of data corresponds to a roll of strip steel after data preprocessing, which contained a total of 183 features. The sample variance was calculated firstly, and all zero variance features were eliminated. A variance of 0 indicated that this

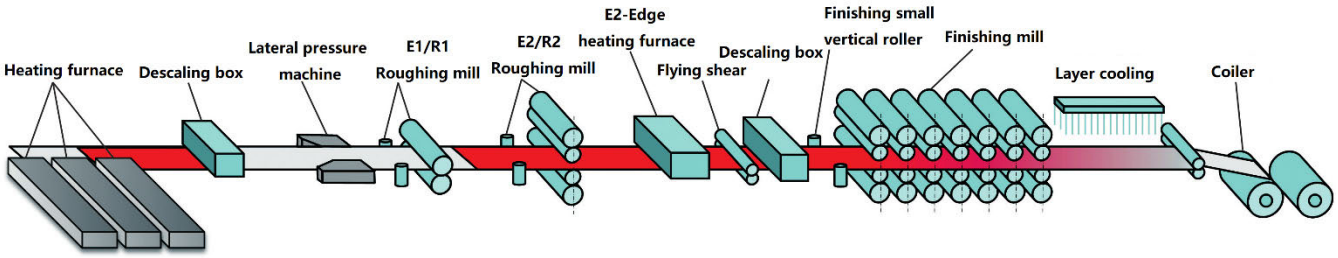

FIGURE 4. Equipment configuration of the production line.

TABLE 2. Codes of defect category of strip steel.

Code	Defect category	Code	Defect category
D_1	Wavy configuration	D_6	Edge crack
D_2	Roll marks	D_7	Pitted surface
D_3	Scratch	D_8	Camber
D_4	Rolled-in scale	D_9	Inclusion
D_5	Oil stain		

TABLE 3. Codes of defect cause of strip steel.

Code	Defect cause	Code	Defect cause
C_1	Slab	C_6	Finishing small vertical roller
C_2	Heating furnace	C_7	Finishing mill
C_3	Descaling box	C_8	Layer cooling
C_4	Lateral pressure machine	C_9	Other causes (Environment etc.)
C_5	Roughing mill(E1/R1&E2/R2)		

TABLE 4. Degrees of membership between symptoms and causes.

Defect cause	Defect category								
	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9
C_1	0.111	0	0	0.259	0.731	0.583	0.227	0.111	0.945
C_2	0.028	0	0	0.222	0	0.194	0	0.037	0
C_3	0	0	0	0.481	0.115	0	0	0	0.055
C_4	0	0.016	0.038	0	0	0	0	0.667	0
C_5	0.389	0.672	0.203	0	0	0	0.182	0	0
C_6	0.111	0.066	0	0	0	0.084	0	0	0
C_7	0.278	0.213	0.645	0	0	0.139	0.591	0	0
C_8	0.056	0	0.063	0	0	0	0	0.148	0
C_9	0.028	0.033	0.051	0.038	0.154	0	0	0.037	0

TABLE 5. Fuzzy defect symptom vectors given by experts.

No.	Defect category								
	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9
1	0.85	0	0	0	0.25	0	0	0	0.12
2	0	0	0.55	0	0	0	0	0.75	0
3	0	0.98	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0.08	0.68	0
5	0.45	0	0	0.18	0	0.78	0	0	0

feature did not contribute to the model. Then the remaining 138 features used linear regression model with L1 regularization penalty term as the base model for feature selection according to the method in section III. B. After excluding 81 features, the remaining 57 features were grouped, corresponding to defect causes C_1-C_9 according to their sources.

TABLE 6. Comparison of results before information fusion.

No.	Fault source selected by experts	Fault source selected by fuzzy inference TOP K result		
		K = 1	K = 2	K = 3
1	C_5	C_1	**C_5**	C_7
2	C_4	**C_4**	C_7	C_8
3	C_5	**C_5**	C_7	C_6
4	C_1	C_4	C_8	**C_1**
5	C_7	C_1	**C_7**	C_2

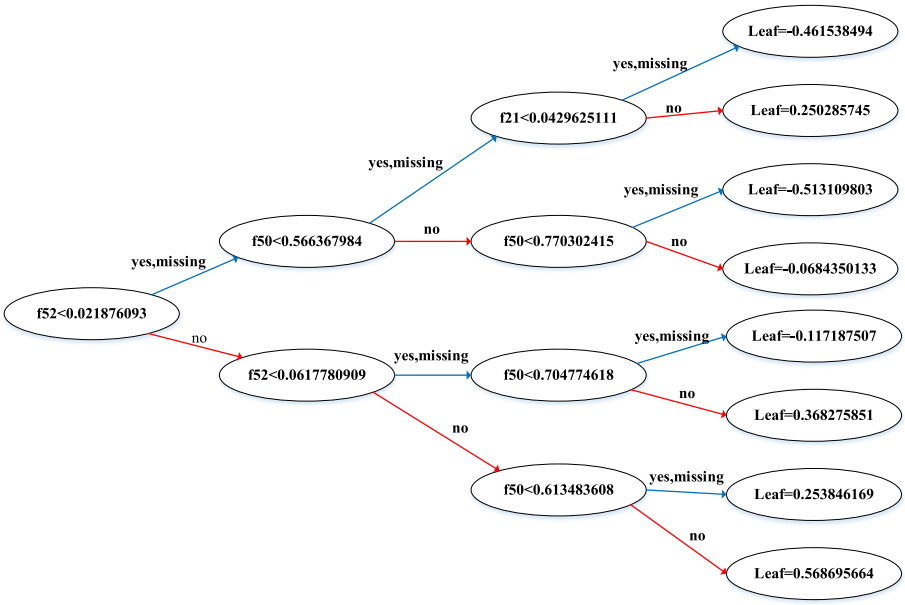

FIGURE 5. One of the decision trees generated by XGBoost.

After feature selection, XGBoost method was used to train the model between process data and surface defects. Objective function was set to binary: logitraw, and evaluation index was set to logloss. Grid search was used to determine other key parameters which were shown in Table 7 in sequence. N_estimatores represents the number of decision trees, max_depth is the maximum depth of the decision trees, and subsample means the percentage of the total training set used when training each tree. Generally speaking, the greater the value of these parameters, the easier the overfitting is. Reasonable adjustment of these parameters can effectively prevent model overfitting, so as to obtain more accurate evaluation of feature contribution. Alpha and lambda are regularization parameters, which can reduce the complexity of the whole model and thus improve the performance of it. One of the decision trees generated by XGBoost was shown in Figure 5. The contribution degree of each feature to the model classification was obtained according to the method described in section III. B, and Figure 6 provided part of the results. After summing up and normalizing the contribution degree of each feature according to the source, the occurrence probability of each cause was obtained, as shown in Table 8.

TABLE 7. Key parameters for training XGBoost model.

Training parameters	Value
learning_rate	0.2
n_estimatores	50
max_depth	3
subsample	0.8
alpha	2
lambda	2

It can be concluded that Roughing Mill (E1/R1&E2/R2) is the equipment most likely to cause surface defects of strip steel, and Slab, Layer Cooling and Finishing mill are also noteworthy sources of defects according to Table 8.

9 common causes of surface defects on hot-rolling production line were taken as the identification framework of D-S evidence theory. According to the method proposed by section III. C, the probability of defect causes obtained from fuzzy semantic inference and process data analysis was fused. After communication with technical experts, it was found that the selection in practical application was more inclined to historical cases. Therefore, Eq. (18) was selected as the

TABLE 8. Defect cause probabilistic ranking according to process data.

Defect cause	Importance degree	Defect cause	Importance degree
C_5	0.437	C_4	0.044
C_1	0.211	C_3	0.021
C_8	0.125	C_2	0.006
C_7	0.087	C_9	0.002
C_6	0.067		

TABLE 9. Comparison of results after information fusion.

No.	Fault source selected by experts	Fault source selected by information fusion Top K result		
		K = 1	K = 2	K = 3
1	C_5	**C_5**	C_1	C_7
2	C_4	**C_4**	C_7	C_5
3	C_5	**C_5**	C_7	C_6
4	C_1	C_4	**C_1**	C_5
5	C_7	C_1	C_5	**C_7**

TABLE 10. Comparison of F1 score between proposed method in this paper and other methods.

Defect cause	F1 score of methods based on single data source			F1 score of proposed method	
	LightGBM	AdaBoost	XGBoost	Rule for fusion (Eq. 16)	Rule for fusion (Eq. 18)
C_1	68.17	65.41	66.66	**70.59**	69.36
C_2	65.16	62.11	68.36	**75.63**	75.42
C_3	70.68	62.51	71.67	72.49	**72.71**
C_4	64.89	67.09	70.57	**75.82**	75.20
C_5	70.85	63.02	68.23	70.42	**75.78**
C_6	67.40	70.23	67.22	69.28	**72.85**
C_7	65.59	66.99	**72.62**	69.95	71.82
C_8	65.72	69.68	71.39	**73.86**	71.37
C_9	66.77	64.35	69.05	71.65	**75.75**
Total	67.25	65.71	69.53	72.19	**73.36**

fusion rule, and the weight coefficient was 0.7. The final fusion calculation was carried out according to the cases given by experts in Table 5, which shown in Table 9. It can be concluded that the hit ratio is improved compared with the judgment result of a single information source, and the comprehensive judgment is more reliable. In addition, information fusion can easily introduce other models and has strong extension ability.

At last, a comparison experiment between some typical methods based on the single data source and the method proposed in this paper. LightGBM, AdaBoost and XGBoost were used for the single data source, and two fusion rules () are applied respectively in the method proposed in this paper. The experimental results are shown in Table 10, and F1 score for each category of fault cause are listed in the table for necessary comparison. It can be concluded that the performance of the method proposed in this paper is generally better than methods based on single data source. In particular, the fusion rule based on Eq.18 achieves the best traceability effect.

C. ANALYSIS OF EXPERIMENTAL RESULTS

The proposed fault location method is applied to the real hot-rolling line in this section, which can analyze the process data and integrate the experience accumulated by people to obtain the comprehensive probability of defect case.

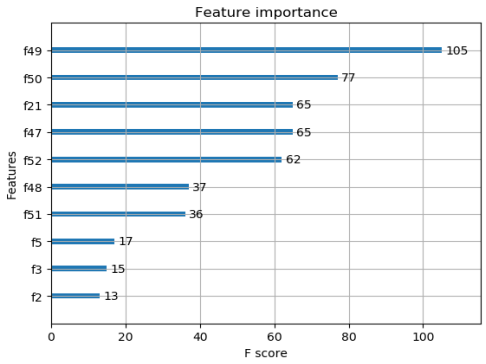

FIGURE 6. **Feature importance (Top 10).**

Traditional fault location process always spends hours on finding out the source of surface defect. But when the proposed method is applied to the actual production line, the time of preliminary determination of fault location can be reduced to a few minutes, and the automatic calculation process greatly reduces reliance on technical experts. The data-based process standards are unified, which is in accordance with the results derived by experts from the rules of experience. The entire process can be completed automatically in the system, thereby greatly improving the efficiency of fault location, shortening the unplanned downtime of the production line, and increasing corporate profits.

V. CONCLUSION AND PROSPECT

This paper analyzes the bottlenecks encountered by traditional fault location methods on hot- rolling production lines, and proposes a fault location model for strip steel surface quality defects based on fuzzy semantic inference and process data analysis. The deficiencies of traditional method in analyzing ability, judging standard and locating speed are improved. It can quickly reduce the scope of fault source and improve the effectiveness and timeliness of fault location.

Especially in the new era of manufacturing industry's transformation from automation and informatization to digitalization and networking, faced with the massive data generated by intelligent manufacturing, manual analysis will encounter many bottlenecks, and the fault location based on data may become a link in intelligent operation and maintenance in the future, which has a huge application prospect. At the same time, due to the universality of historical cases and process data on the production line, the model in this paper can be extended to other production lines and has a wide range of practical application values.

REFERENCES

[1] D. Preuveneers and E. Ilie-Zudor, "The intelligent industry of the future: A survey on emerging trends, research challenges and opportunities in industry 4.0," *J. Ambient Intell. Smart Environ.*, vol. 9, no. 3, pp. 287–298, Apr. 2017.

[2] M. Pinzone, F. Albè, D. Orlandelli, I. Barletta, C. Berlin, B. Johansson, and M. Taisch, "A framework for operative and social sustainability functionalities in human-centric cyber-physical production systems," *Comput. Ind. Eng.*, vol. 139, Jan. 2020, Art. no. 105132.

[3] Y. Liu, L. Wang, X. V. Wang, X. Xu, and L. Zhang, "Scheduling in cloud manufacturing: State-of-the-art and research challenges," *Int. J. Prod. Res.*, vol. 57, nos. 15–16, pp. 4854–4879, Aug. 2019.

[4] L. Zhou, L. Zhang, L. Ren, and J. Wang, "Real-time scheduling of cloud manufacturing services based on dynamic data-driven simulation," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5042–5051, Sep. 2019.

[5] J. Wan and M. Xia, "Cloud-assisted cyber-physical systems for the implementation of industry 4.0," *Mobile Netw. Appl.*, vol. 22, no. 6, pp. 1157–1158, Dec. 2017.

[6] N. Qin, W. Jin, J. Hung, and Z. Li, "Ensemble empirical mode decomposition and fuzzy entropy in fault feature analysis for high-speed train bogie," *Control Theory Appl.*, vol. 31, no. 9, pp. 1245–1251, Sep. 2014.

[7] S. Kang, E. Kim, J. Shim, S. Cho, W. Chang, and J. Kim, "Mining the relationship between production and customer service data for failure analysis of industrial products," *Comput. Ind. Eng.*, vol. 106, pp. 137–146, Apr. 2017.

[8] Y. Yang, D. Yu, and J. Cheng, "A fault diagnosis approach for roller bearing based on IMF envelope spectrum and SVM," *Measurement*, vol. 40, nos. 9–10, pp. 943–950, Nov./Dec. 2007.

[9] G. Georgoulas, V. Climente-Alarcon, J. A. Antonino-Daviu, I. P. Tsoumas, C. D. Stylios, A. Arkkio, and G. Nikolakopoulos, "The use of a multilabel classification framework for the detection of broken bars and mixed eccentricity faults based on the start-up transient," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 625–634, Apr. 2017.

[10] M. Amar, I. Gondal, and C. Wilson, "Vibration spectrum imaging: A novel bearing fault classification approach," *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 494–502, Jan. 2015.

[11] M. D. Prieto, G. Cirrincione, A. G. Espinosa, J. A. Ortega, and H. Henao, "Bearing fault detection by a novel condition-monitoring scheme based on statistical-time features and neural networks," *IEEE Trans. Ind. Electron.*, vol. 60, no. 8, pp. 3398–3407, Aug. 2013.

[12] J. Tian, C. Morillo, M. H. Azarian, and M. Pecht, "Motor bearing fault detection using spectral kurtosis-based feature extraction coupled with K-nearest neighbor distance analysis," *IEEE Trans. Ind. Electron.*, vol. 63, no. 3, pp. 1793–1803, Mar. 2016.

[13] Z.-H. Sun, R. Liu, and X. Ming, "A fault diagnosis and maintenance decision system for production line based on human-machine multi-information fusion," in *Proc. IEEE Int. Conf. Cloud Comput. Technol. Sci.*, Dec. 2018, pp. 151–156.

[14] J. Lyu, C. W. Liang, and P.-S. Chen, "A data-driven approach for identifying possible manufacturing processes and production parameters that cause product defects: A thin-film filter company case study," *IEEE Access*, vol. 8, pp. 49395–49411, Feb. 2020.

[15] F.-Q. Pei, Y.-F. Tong, and D.-B. Li, "Multi-level welding quality fault discovery of an intelligent production line by using Taguchi quality loss function and signal-noise ratio," *IEEE Access*, vol. 6, pp. 40792–40803, Jul. 2018.

[16] S. X. Ding, S. Yin, K. Peng, H. Hao, and B. Shen, "A novel scheme for key performance indicator prediction and diagnosis with application to an industrial hot strip mill," *IEEE Trans. Ind. Informat.*, vol. 9, no. 4, pp. 2239–2247, Nov. 2013.

[17] K. Peng, K. Zhang, G. Li, X. He, and X. Yang, "New kernel independent and principal components analysis-based process monitoring approach with application to hot strip mill process," *IET Control Theory Appl.*, vol. 8, no. 16, pp. 1723–1731, Nov. 2014.

[18] K. Peng, K. Zhang, B. You, and J. Dong, "Quality-relevant fault monitoring based on efficient projection to latent structures with application to hot strip mill process," *IET Control Theory Appl.*, vol. 9, no. 7, pp. 1135–1145, Apr. 2015.

[19] K. Zhang, K. Peng, S. X. Ding, Z. Chen, and X. Yang, "A correlation-based distributed fault detection method and its application to a hot tandem rolling mill process," *IEEE Trans. Ind. Electron.*, vol. 67, no. 3, pp. 2380–2390, Mar. 2020.

[20] F. Zhang, S. Zong, and Z. Ling, "Fault diagnosis using kernel principal component analysis for hot strip mill," *J. Eng.*, vol. 2017, no. 9, pp. 527–535, Sep. 2017.

[21] X. Wang, C. Peng, and Z. Zhang, "Application of EEMD-based resonance demodulation technology in train bearing fault diagnosis," *Mod. Electron. Technol.*, vol. 38, no. 21, pp. 24–27, Nov. 2015.

[22] K. Peng, K. Zhang, B. You, J. Dong, and Z. Wang, "A quality-based nonlinear fault diagnosis framework focusing on industrial multimode batch processes," *IEEE Trans. Ind. Electron.*, vol. 63, no. 4, pp. 2615–2624, Apr. 2016.

[23] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, pp. 37–54, 1996.
[24] J. A. Goguen, "L-fuzzy sets," *J. Math. Anal. Appl.*, vol. 18, no. 1, pp. 145–174, 1967.
[25] L. A. Zadeh, "Similarity relations and fuzzy orderings," *Inf. Sci.*, vol. 3, no. 2, pp. 177–200, Apr. 1971.
[26] A. Afify, "A novel algorithm for fuzzy rule induction in data mining," *Proc. Inst. Mech. Eng. C, J. Mech. Eng. Sci.*, vol. 228, no. 5, pp. 877–895, Apr. 2014.
[27] A. A. Afify, "A fuzzy rule induction algorithm for discovering classification rules," *J. Intell. Fuzzy Syst.*, vol. 30, no. 6, pp. 3067–3085, Apr. 2016.
[28] G. Niu and H. Li, "IETM centered intelligent maintenance system integrating fuzzy semantic inference and data fusion," *Microelectron. Rel.*, vol. 75, no. 75, pp. 197–204, Aug. 2017.
[29] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
[30] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. Data. Min. Knowl. Discov.*, 2016, pp. 785–794.
[31] R. R. Yager, "On the dempster-shafer framework and new combination rules," *Inf. Sci.*, vol. 41, no. 2, pp. 93–137, Mar. 1987.

ZHAOPING WANG received the B.S. degree in automation from Tongji University, Shanghai, China, in 2018, where he is currently pursuing the M.S. degree in control engineering. His research interests include the industrial knowledge graph and intelligent decision-making.

JIAN WANG is currently the Director of the CIMS Research Center, Tongji University. He has been long involved in research and development in the field of automatic control. His current discipline is system engineering. In recent years, he has been mainly involved in enterprise informatization, CIMS, business process management, workflow technology, energy and transportation systems, networked manufacturing, and system integration.

SEN CHEN received the M.S. degree in engineering management from Donghua University, Shanghai, in 2017. He is currently pursuing the Ph.D. degree in control science and engineering (the direction is systems engineering) with Tongji University, Shanghai, China.

He has 15 years of work experience in American and French enterprises, one of which belonged to the Fortune 500 companies. Since 2019, he has been working with the Team of CIMS Research Center, Tongji University, and doing the research of ternary data fusion and knowledge graph. He holds a dozen invention patents.