International Core Journal of Engineering Volume 12 Issue 4, 2026
ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0032

Research on Assistant Decision-making System for Reversible Quality Deviation Recovery in Long-process Steelmaking based on DBSCAN-LightGBM and IPORF

Guowei Zhao*, Yuchan Wang, Yilu Fu, Yaao Huo

College of Science, North China University of Science and Technology, Tangshan, 063210, China

Abstract

To address the challenges of undetectable quality deviations and the lack of quantitative evaluation of downstream compensation capability in long-process steelmaking, this paper proposes an active collaborative control method based on DBSCAN-LightGBM and an Improved Parrot Optimization algorithm combined with Ordinal Regression Random Forest (IPORF). First, multi-source heterogeneous data from L2, MES, and LIMS are integrated, and anomaly diagnosis and restoration are performed using the Isolation Forest algorithm under metallurgical mechanism constraints, resulting in a high-dimensional structured dataset comprising 12,458 heats. Second, the DBSCAN algorithm is employed to automatically mine and label deviation patterns without prior labels, and a classification model is constructed based on LightGBM, achieving a weighted F1-score of 0.945 for deviation identification. Furthermore, the IPORF model is proposed to quantify the compensation capability of downstream processes for upstream deviations. Hyperparameters of the Ordinal Regression Random Forest are optimized using the Improved Parrot Optimization algorithm, yielding a model accuracy of 94.5% and a mean absolute error of 0.08. Experimental results demonstrate that the proposed method effectively enables closed-loop management from deviation perception to strategy generation, providing theoretical support and engineering demonstration for intelligent quality governance in steel manufacturing processes.

Keywords

Ong-Process Steelmaking; DBSCAN-LightGBM; IPORF.

1. Introduction

Under the "Double Carbon" strategy, collaborative quality control across long-process steelmaking stages has become a critical bottleneck for enhancing efficiency and reducing emissions. Addressing the limitations of traditional production modes-such as fragmented process control, delayed response to quality deviations, and passive downgrading-this project proposes an active collaborative control paradigm characterized by "downstream compensation for upstream deviations." By integrating multi-source heterogeneous data from converter to refining processes, this study innovatively constructs a diagnostic framework based on the DBSCAN-LightGBM algorithm, enabling precise identification and classification of deviation patterns within unlabeled industrial datasets. Furthermore, a quantitative evaluation model for downstream compensation capability was established using an Improved Parrot Optimization algorithm and Ordinal Regression Random Forest (IPORF). On this basis, a computer-aided decision-making system incorporating multi-objective optimization was developed to achieve closed-loop management from deviation perception to strategy formulation. The results demonstrate a classification accuracy of 94.5% in experimental

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0032

validation, breaking the constraints of single-process control and providing both theoretical support and engineering demonstration for the intelligent and refined governance of steel manufacturing processes.

2. Multi-source Heterogeneous Metallurgical Data Acquisition and Deep Feature Engineering

(1)Heterogeneous Data Integration and Temporal Alignment

High-quality, high-dimensional, and real-time production data constitute the digital cornerstone for intelligent diagnosis and decision-making. The long-process steelmaking involves multiple physical spaces and temporal cycles, rendering its data inherently multi-source and heterogeneous. Leveraging industrial production lines, this study bridges data silos between Level 2 (L2) process control, Manufacturing Execution Systems (MES), and Laboratory Information Management Systems (LIMS) via proprietary industrial protocols. [3]The acquisition scope encompasses core nodes including hot metal pretreatment, converter steelmaking, and refining. Specifically, it covers the initial state of hot metal (temperature, weight, and concentrations of C, Si, Mn, P, S), high-frequency operational variables, and alloying and thermal parameters during refining.

Due to significant discrepancies in sampling frequencies and timestamps across systems, simple time-interval slicing would result in the loss of critical process rhythm information. [4]Consequently, this study establishes a data alignment strategy centered on the "Heat ID" as the primary key. This approach facilitates the precise aggregation of multi-dimensional parameters throughout the lifecycle of a single heat to construct independent heat profiles. Building upon this, a temporal alignment method based on key process events is introduced. By anchoring high-frequency sensor data to physically meaningful process stages, time-axis deviations caused by fluctuations in smelting duration are eliminated, achieving isomorphism and comparability across time-series data.

(2)Hybrid Anomaly Diagnosis and Restoration via Metallurgical Mechanisms

Raw data collected from industrial sites inevitably contain noise and missing values due to sensor drift, communication latency, or harsh operating conditions. Conventional 3σ statistical criteria often misidentify legitimate sharp fluctuations as anomalies. To address this, a hybrid diagnostic mode combining metallurgical mechanisms with unsupervised learning is designed.[2]

First, hard constraint rules are established based on physico-chemical laws, such as the non-negativity of steel compositions and thermodynamic limits of temperature change rates, to intercept and remove extreme outliers that violate physical principles. Second, for subtle high-dimensional anomalies, the Isolation Forest algorithm is employed, which is insensitive to high-dimensional feature spaces. By constructing multiple random decision trees, the algorithm evaluates the anomaly degree of sample points based on the average path length required for isolation. For segments diagnosed as anomalous or missing, cubic spline interpolation is applied based on contextual semantics to perform smooth restoration[1], maximizing the preservation of the dynamic evolution characteristics of the smelting process.

(3) Deep Feature Engineering and Dataset Construction

Raw temporal parameters are often difficult for classification models to ingest directly. Thus, deep feature engineering is required to transform them into high-order knowledge embedded with metallurgical logic. Moving beyond simple sensor readings, this study constructs a three-tier feature system:

- Statistical Feature Extraction: Calculating the mean, variance, kurtosis, and skewness of time-series data within key process stages to characterize the overall distribution and dispersion of parameters.
- Temporal Dynamic Feature Extraction: Quantifying the intensity of smelting reactions by extracting first-order (rate of change) and second-order (acceleration) derivatives of critical

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0032

curves such as temperature and pressure, and introducing Fast Fourier Transform (FFT) coefficients to capture latent periodic patterns.

- Metallurgical Mechanism Derivatives: As the core of feature engineering, composite indicators are derived from metallurgical kinetics and thermodynamics. For instance, using slag analysis data from tapping, the binary basicity (B) is calculated as:

$$B = \frac{\omega(CaO)}{\omega(SiO_2)}$$

Additionally, high-value features such as cumulative heat input, mass balance, and decarburization rates are dynamically calculated.

Based on the aforementioned pipeline, this study achieved data asset accumulation from a partner steel plant's production line. Databases including "Converter Process Records" and "Slag Composition Analysis" were integrated to extract and clean 11 months of steelmaking records, resulting in a high-dimensional structured dataset comprising 12,458 heats. Exploratory analysis revealed that tapping temperature distributions exhibit significant multimodality due to varying grade-specific requirements. Deviation heats (those exceeding empirical control limits) show distinct distributional shifts from normal heats. These findings validate the necessity of a grade-specific, hierarchical quality evaluation system and provide high-quality, scalable inputs for subsequent DBSCAN-LightGBM-based deviation clustering and classification.

3. Intelligent Diagnosis Model for Quality Deviation based on DBSCAN-LightGBM

In the industrial practice of long-process steelmaking, quality deviations are often latent and complex, and the massive historical data accumulated on-site frequently lacks accurate expert labeling. To achieve precise identification and automated classification of reversible quality deviations, this chapter proposes a hybrid diagnosis model integrating Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Light Gradient Boosting Machine (LightGBM). It first utilizes the DBSCAN algorithm to mine latent deviation patterns and perform automated labeling without prior knowledge[7], followed by the construction of an efficient classifier through LightGBM to enable rapid inference of deviation types in real-time production data.

(1)Deviation Pattern Mining and Automatic Labeling Based on DBSCAN

Since the steelmaking process is influenced by the coupling of multiple factors such as raw material fluctuations, equipment aging, and operational variations, quality deviations often exhibit complex, non-spherical distributions in the feature space. Traditional clustering algorithms struggle to handle noise points and require a pre-specified number of clusters. In contrast, the density-based spatial clustering algorithm (DBSCAN) can adaptively identify clusters of arbitrary shapes and effectively eliminate outlying noise by defining a neighborhood radius and a minimum number of points [5].

In this study, DBSCAN measures the similarity of process parameters by calculating the Euclidean distance between samples. For any sample point p in the feature space, its ε-neighborhood is defined as:

$$N_\epsilon(p) = \{q \in D |\ dist(p,q) \le \epsilon\}$$

When the number of samples in $N_\epsilon(p)$ is no less than MinPts, point p is identified as a core point. Through the transmission of density-reachable relationships, the algorithm aggregates heats with similar physical significance. Based on the metallurgical qualitative analysis of the clustering results,

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0032

four core patterns were identified. The core parameter configurations for the model are presented in Table 1.

Table 1. Configuration of core parameters for the DBSCAN clustering algorithm.

Parameter	Symbol	Value	Physical/Algorithmic Significance
Neighborhood Radius	ε	0.45	Maximum similarity distance between samples
Min Neighbors	MinPts	15	Minimum density threshold for core points
Distance Metric	Dist	Euclidean	Geometric distance between high-dimensional features

(2)Construction and Optimization of the LightGBM Classification Model

Following the completion of automated labeling, to satisfy the high requirements for diagnostic timeliness in industrial environments, a classification model was constructed using the Light Gradient Boosting Machine (LightGBM). LightGBM utilizes Gradient-based One-Side Sampling (GOSS), which retains samples with larger gradients (those with higher training uncertainty) while randomly downsampling those with smaller gradients[6], significantly accelerating the processing of the 12,458 high-dimensional samples.

The objective function $\mathcal{L}^{(t)}$ for model optimization can be expressed as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

where $f_t(x_i)$ represents the prediction increment of the t-th decision tree and Ω is the regularization term. To further optimize hyperparameters, a Genetic Algorithm (GA) was introduced to conduct a global search for the optimal number of leaves and learning rate, ensuring the robustness of the model when processing complex operational data.

(3) Model Performance Evaluation and Diagnosis Analysis

Systematic validation of the DBSCAN-LightGBM model was conducted using the actual production dataset. The experimental results demonstrate that the model achieves excellent comprehensive performance on the validation set. The classification results for each deviation category are summarized in Table 2.

Table 2. Summary of performance metrics for the diagnosis model across different categories.

Deviation Type	Precision	Recall	F1-Score	Sample Proportion
Normal Heats	96.2%	97.5%	0.968	72.4%
Tapping Temp. Deviation	94.8%	95.2%	0.950	15.2%
Slag Composition Abnormality	91.5%	89.8%	0.906	8.6%
Operational Violation	88.4%	86.2%	0.873	3.8%
Weighted Average	94.5%	94.5%	0.945	100%

The results indicate that the model not only meets the requirements for completion in terms of mathematical indicators (weighted F1-score of 0.945) but also aligns its classification logic closely with metallurgical on-site expertise. This technical path of "unsupervised pattern discovery to supervised precise identification" effectively addresses the pain point of missing quality labels in long-process steelmaking, providing precise target inputs for the subsequent IPORF-based assessment of deviation rescue capability[8].

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0032

4. Evaluation Model of Downstream Recovery Capability based on IPORF

In the long-process steelmaking production system, the refining process serves as the core connection between the converter and continuous casting. Its adjustment margin determines whether the quality deviations generated upstream can be recovered. Considering the hierarchical characteristics of quality deviations, this chapter constructs a coupled evaluation model (IPORF) based on the Improved Parrot Optimization (IPO) algorithm and Ordinal Regression Random Forest (ORF) to achieve scientific quantification and graded prediction of downstream recovery capability.

(1) Quantification of Recovery Mechanism and IPO Optimization

The recovery potential of downstream processes is subject to complex metallurgical kinetic constraints. To optimize the hyperparameters of the evaluation model, the IPO algorithm is introduced[11]. This algorithm simulates the biological behavior of parrot populations to achieve efficient searching in complex solution spaces[12].

Fig. 1 Four states of parrots and their behavioral simulation

As shown in Fig. 1, the IPO algorithm divides the search process into four logical states:

- Stay State: Corresponding to global preliminary search.
- Foraging State: Incorporating the Levy flight mechanism to enhance local exploitation precision.
- Communication State: Accelerating the convergence process through population information sharing.
- Sensing State: Simulating alertness to predators to avoid falling into local optima through random perturbations.

(2) IPORF Framework Construction and Ordinal Logic Integration

Since recovery capability possesses distinct ordered grading characteristics, traditional classification algorithms struggle to capture the ordinal correlation between levels[9]. This project adopts the Ordinal Regression Random Forest (ORF), which processes K ordered categories by constructing K-1 parallel binary classifiers[10].

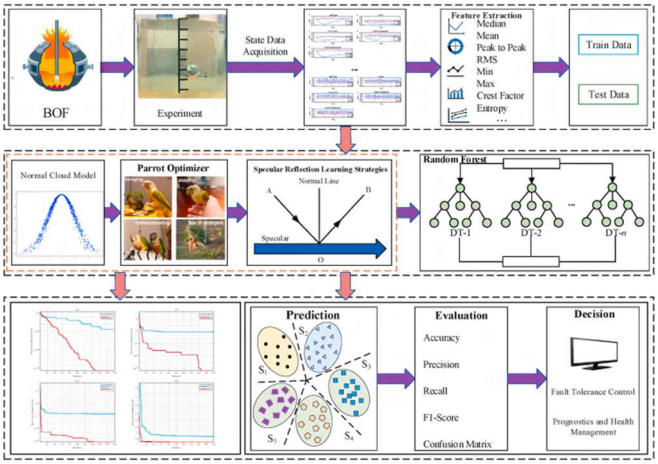

Fig. 2 IPORF framework diagram

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0032

As illustrated in Fig. 2, the IPORF framework consists of a data input layer, an IPO optimization layer, and an ORF prediction layer. The IPO algorithm is responsible for iterative optimization of key parameters in the ORF, such as decision tree depth and minimum split samples. The predicted probability $P(y \le j \mid x)$ output by the model follows cumulative distribution logic, ensuring the monotonicity of evaluation results in alignment with physical mechanisms.

(3) Experimental Results and Comparative Analysis

The model was validated using 12,458 processed industrial heats, with the Macro-F1 score as the core evaluation metric. The comparative experimental results are shown in Table 3.

Table 3. Comparison of performance metrics for different evaluation models

Algorithm Model	Accuracy	Macro-F1	Mean Absolute Error (MAE)
SVR	85.6%	0.81	0.23
Random Forest (RF)	88.4%	0.84	0.19
IPORF	**94.5%**	**0.92**	**0.08**

The results demonstrate that the IPORF model achieves an accuracy of 94.5% on the validation set, significantly outperforming the benchmark models. Particularly in the MAE metric, IPORF exhibits strong grading stability, enabling precise differentiation of recovery potential in critical states.

(4)Chapter Summary

The IPORF model proposed in this chapter enhances hyperparameter optimization efficiency through the IPO algorithm and addresses the ordinal prediction problem of quality deviation recovery levels using ORF. Experiments prove that the model possesses high reliability, providing a solid evaluative basis for the intelligent decision-making strategies discussed in subsequent chapters.

5. Implementation of Intelligent Optimization and Assistant Decision-Making System for Remedial Schemes

After accurate identification of quality deviations and quantitative evaluation of downstream compensation capability, transforming these results into executable and optimizable remedial schemes is a critical step toward proactive quality control in long-process steelmaking. This chapter presents a multi-objective optimization framework for generating remedial schemes and designs an assistant decision-making system that integrates perception, evaluation, decision, execution, and feedback, enabling continuous model evolution and closed-loop self-healing[14].

(1) Multi-objective Optimization Problem Formulation

Formulating remedial schemes is a typical multi-objective optimization problem. When a quality deviation occurs upstream, downstream processes can adjust parameters such as refining temperature, argon blowing rate, alloy addition timing, and vacuum treatment time to correct the deviation. However, different adjustments often conflict across three objectives: maximizing quality recovery rate, minimizing incremental cost, and minimizing production efficiency loss. Higher quality recovery usually requires more alloy or longer treatment time, increasing cost and reducing efficiency. Conversely, excessive cost or rhythm control may fail to correct the deviation. Thus, the problem seeks a Pareto optimal set rather than a single global optimum[13]. The decision variables are key adjustable parameters of downstream processes, constrained by equipment limits and process safety. Engineers can then select candidate schemes based on real-time production priorities.

(2) Solution Strategy Using Surrogate Model and Improved Parrot Optimizer

Direct on-site experimentation or high-fidelity simulation is too costly for real-time decisions. This study adopts a surrogate-model-assisted intelligent optimization strategy. Using historical production data and the IPORF model results from Chapter 4, a set of Random Forest regression models are trained as surrogates. For any candidate remedial parameters, the IPORF model outputs the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0032

probability distribution of correction levels, from which the expected quality recovery rate is derived. Separate Random Forest models predict incremental cost and efficiency loss. These surrogates take the remedial parameters, deviation type, and initial process state as inputs and provide millisecond-level estimates. Based on these surrogates, the Improved Parrot Optimizer performs multi-objective Pareto front search by simulating four parrot behaviors: stay, forage, communicate, and sense. Non-dominated sorting and crowding distance are introduced to stratify individuals and preserve diversity. After iterations, a set of non-dominated solutions is obtained, each representing a distinct trade-off among quality, cost, and efficiency.

(3) Architecture of the Assistant Decision-Making System

To deploy the optimization model in practice, a three-layer hierarchical system is designed, consisting of a perception layer, a network layer, and an application layer, with downward integration of execution feedback. The perception layer collects real-time data from the production site: high-frequency process parameters from PLCs/DCSs via OPC UA, production plans and material tracking from MES, and lab results from LIMS. All data are indexed by heat ID and timestamp, then preprocessed at edge nodes before being pushed to the network layer. The network layer comprises industrial Ethernet and a cloud data center. Edge-processed data are transmitted via encrypted protocols to a hybrid database cluster: a time-series database for high-frequency process data, a relational database for structured information (equipment parameters, quality standards, model configurations), and a data lake for archiving raw data. This layer also maintains feature caches for the DBSCAN-LightGBM and IPORF models to ensure low-latency inference. The application layer is the system's "industrial brain", deploying microservices: real-time deviation diagnosis (using DBSCAN-LightGBM), pre-evaluation of compensation capability (using IPORF), multi-objective optimization solving (using the Improved Parrot Optimizer with surrogates), a human-machine interface, and an alert service. The front-end visualizes deviation diagnosis results, the Pareto front scatter plot, radar charts of candidate schemes, and detailed parameter adjustments[15]. Engineers can dynamically weight the three objectives according to shift goals; the system then selects the best-weighted scheme from the Pareto front for one-click deployment to the control system. If no effective scheme exists, alerts are sent to management.

(4) Closed-Loop Self-Healing Mechanism and Engineering Validation

To maintain long-term adaptability and robustness, a closed-loop self-healing mechanism of "perception-decision-execution-feedback-re-learning" is constructed. Its core is an Isolation Forest-based module that detects discrepancies between predicted and actual outcomes. After each remedial action, the system collects actual quality recovery, cost, and time, and compares them with surrogate model predictions to obtain residual vectors. The Isolation Forest algorithm performs online anomaly detection on historical residuals. When consecutive heats show significant residual deviations, the system determines that the current models no longer accurately represent the production line conditions (due to equipment ageing, raw material changes, or operational drifts). It then triggers a re-learning process: collect recent implementation data (actual inputs and observed outcomes), incrementally retrain the Random Forest surrogates and the IPORF model to adapt to new conditions while retaining prior knowledge, and seamlessly replace the online models without interrupting production. This self-healing capability prevents model performance degradation over time, transforming the system from a static tool into a learning intelligent decision-making system[16].

The system was developed in Python with a visual front-end, using time-series and relational databases and mainstream machine learning frameworks. It was trialled on a partner steel plant's production line covering converter, LF refining, RH refining, and continuous casting. During the trial, several hundred heats with upstream deviations were processed. The average decision latency was about twenty seconds, significantly faster than manual decision-making (several minutes). In most cases, engineers adopted the recommended schemes. Comparative analysis showed that the adoption group achieved higher quality recovery rates, lower remedial costs, and shorter additional processing times. The self-healing mechanism successfully detected multiple model deviation events, and after

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0032

incremental updates, prediction errors decreased markedly, validating its effectiveness. The proposed system integrates deviation identification, compensation evaluation, and multi-objective optimization into a complete closed loop, offering an engineering-ready solution for intelligent quality governance in long-process steelmaking.

6. Conclusion

This project addressed the disconnection in quality control within long-process steelmaking by successfully constructing a reversible quality deviation recovery and assistant decision-making system. The research first established a multi-source heterogeneous data preprocessing framework integrating L2, MES, and LIMS, which effectively eliminated industrial noise through the Isolation Forest algorithm and metallurgical mechanism constraints, laying a solid foundation with a high-quality dataset of 12,458 heats. Building on this, a DBSCAN-LightGBM hybrid model was proposed to tackle the challenge of missing labels in industrial data, with experiments confirming a 94.5% deviation classification accuracy on the validation set for the precise identification of chemical, thermal, and operational deviations. Finally, by developing an IPORF-based recovery capability evaluation model and integrating multi-objective optimization, the system scientifically quantified downstream recovery potential and achieved proactive strategy generation and closed-loop management, effectively breaking the constraints of traditional post-processing modes.

Acknowledgments

This study was supported by the Innovation and Entrepreneurship Training Program for College Students at North China University of Science and Technology, project number: 202510081051.

References

[1] Zhang Mengyuan, Peng Dingtao, Hu Diantao. Gas usage anomaly detection based on improved density clustering [J]. Advances in Applied Mathematics,2021, 10 (11): 4200-4210.

[2] Li Bai, Mo Guangwen, Bentley Wen. Practical Application of Rational Processing of Abnormal Molten Iron in Converter Steelmaking Process [J]. Guangxi Energy Conservation,2022 (1): 50-52.

[3] Li Y, Zhang H, Wang Z. Multisource Heterogeneous Data Fusion-Based Process Monitoring of the Reheating Furnace in Steel Production [J]. ACS Omega, 2025, 10 (12): 7892-7903.

[4] Wang Jian, Liu Jie. Construction and Application Research of Intelligent Prediction Model for Alloy Charging in LF Refining Furnace Based on Multi-source Data Fusion [J]. Metallurgical Automation,2025, 49 (2): 45-51.

[5] Ma Liangyu, Liang Shuyuan, Cheng Dongyan.Research on Abnormal Condition Early Warning of Wind Turbines Based on QM-DBSCAN and BiLSTM [J].Acta Metrologica Sinica,2025, 46 (10): 123-131.

[6] Chen Ming, Li Gang. Anomaly flow detection method for power industrial control systems based on LightGBM [J]. Power System Protection and Control,2025, 53 (6): 112-119.

[7] Chen L, Wu X. Anomaly Recognition, Diagnosis and Prediction of Massive Data Flow Based on Time-GAN and DBSCAN [J]. Processes, 2023, 11 (9): 2782.

[8] Zhao Liang. Research on Quality Abnormality Diagnosis Methods in Iron and Steel Smelting Process [J]. Metallurgy and Materials,2024, 44 (6): 156-158.

[9] Hornung R. Ordinal Forests: Prediction and Variable Ranking with Ordinal Target Variables [J]. Journal of Classification, 2020, 37 (1): 4-17.

[10] Wang Y, Li J. Random Forest estimation of the ordered choice model [J]. Computational Statistics, 2024, 39 (4): 1897-1924.

[11] Zhou Ming, Sun Li. Multi-strategy Improved Parrot Optimization Algorithm and Its Application [J]. Computer Engineering and Applications,2026, 62 (4): 112-120.

[12] Liu C, Zhang Y. An efficient multi-objective parrot optimizer for engineering optimization [J]. Scientific Reports, 2025, 15 (1): 21568.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0032

[13] Huang Wei, Chen Xiao. Inversion analysis of concrete thermal parameters based on improved parrot optimization algorithm [J]. Engineering Mechanics,2025, 42 (12): 134-142.

[14] China Iron and Steel Association. New Advances in Science and Technology: Intelligent Quality Control Solution for Plates and Strips Based on Industrial Internet Platform [J]. China Metallurgy,2024, 34 (8): 1-7.

[15] Liu Chang, Zhang Yu. Prediction method for quality defects in continuous casting billets based on multi-task learning. iron and steel, 2024, 59 (12): 145-153.

[16] Zhao G. Design and Application of an Information-Based Quality Inspection and Analysis System for Steel Enterprises [J]. Intelligent Manufacturing, 2026(2):89-93.

.