Scilight

Journal of Artificial Intelligence for Automation

https://www.sciltp.com/journals/jaia



Check for updates

Article

Human-in-the-Loop Hybrid Intelligent Decision for Automated Strip Steel Flatness Detection

Qi Kang [1,2,*], Kefan Wang [2] and Junyu Shang [1]

[1] The Department of Control Science and Engineering, Tongji University, Shanghai 201804, China
[2] Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai 200092, China
* Correspondence: qkang@tongji.edu.cn

How To Cite: Kang, Q.; Wang, K.; Shang, J. Human-in-the-Loop Hybrid Intelligent Decision for Automated Strip Steel Flatness Detection. *Journal of Artificial Intelligence for Automation* **2026**, *1*(3), 15. https://doi.org/10.53941/jaia.2026.100015

Received: 13 February 2026
Revised: 16 August 2026
Accepted: 29 September 2026
Published: 14 October 2026

Abstract: Intelligent decision-making in high-risk industrial scenarios is critical for maximizing production efficiency and ensuring operational safety. Strip flatness detection is a representative task because missed flattening decisions may lead to downstream quality defects. To address this problem, this paper proposes a Human-in-the-Loop Intelligent Decision Framework (HIL-IDF) and develops a Multimodal Hybrid Intelligent Strip Detection method (MHI-SD) as its task-specific implementation. The proposed framework consists of an expert-rule layer, an intelligent model layer, and a risk-aware decision layer. In MHI-SD, the expert-rule layer rapidly handles samples with deterministic characteristics, while the intelligent model layer combines a time-series branch and an image-based auxiliary branch. The time-series branch extracts segment-wise statistical features and performs weighted ensemble learning using k-nearest neighbors (KNN), random forest (RF), and Light Gradient Boosting Machine (LightGBM). The image branch converts paired bending-leveling curves into two-dimensional morphology images and uses a convolutional neural network (CNN) with a support vector machine (SVM) for auxiliary verification. The risk-aware decision layer is designed from a cost-sensitive perspective and implemented as a threshold-based sequential routing mechanism. Low-confidence samples, high-risk predictions, and samples with model disagreement are routed to auxiliary checking or manual verification. Based on historical production-line data, MHI-SD achieves 96.32% accuracy and 95.19% recall. Together with the ablation and threshold-sensitivity analyses, these results support the effectiveness of the proposed risk-aware decision design and suggest its practical potential for strip flatness decision support under the studied production conditions.

Keywords: strip shape detection; machine learning; expert rules; ensemble learning; multimodal fusion

1. Introduction

Intelligent judgment and decision-making are pivotal to the automation and safe operation of modern industrial production systems [1,2]. In contrast to offline analyses or low-risk applications, industrial decision-making carries inherently high safety stakes, where erroneous actions can cause severe equipment damage, compromise product quality, or even trigger catastrophic production incidents [3,4]. This high-risk context imposes rigorous demands on intelligent systems: they must not only deliver accurate predictions but also possess the capacity to perceive, quantify, and actively manage potential risks [5–7].

Although data-driven intelligent methods have achieved significant advances in industrial tasks [8,9], fully relying on end-to-end automated models remains inadequate for ensuring decision safety and controllability in

Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

high-risk industrial environments. On the one hand, the limited availability of high-quality annotated data forces models to operate beyond their training distribution [10], undermining reliable generalization under extreme or unforeseen operating conditions [11]. On the other hand, directly using model outputs as final decisions, without explicitly accounting for predictive uncertainty and operational risk [12], provides limited fault tolerance and inadequate buffering under strict safety-cost constraints [13].

To address these challenges, we propose a Human-in-the-Loop Intelligent Decision Framework (HIL-IDF), which integrates expert knowledge, intelligent model predictions, and risk-aware human intervention to improve decision reliability in high-risk industrial scenarios. As shown in Figure 1, HIL-IDF consists of three layers: an expert-rule layer, an intelligent model layer, and a risk-aware decision layer. Human experts play a critical and irreplaceable role [14–16]. In the expert-rule layer, expert experience and engineering constraints are incorporated as prior knowledge to constrain samples with deterministic decision characteristics to a low-risk decision subspace [17], enabling rapid and interpretable judgments. For samples that the rule layer cannot cover, the intelligent model layer uses data-driven methods to generate predictions and confidence or uncertainty indicators, which provide a basis for subsequent risk assessment [18,19]. In the risk-aware decision layer, expert decision-making serves as the final arbitration mechanism for high-risk samples [20], intervening when model confidence is insufficient or when multiple models produce inconsistent predictions. In this way, HIL-IDF combines the efficiency of automated models with the reliability of expert intervention, thereby improving robustness and controllability in complex production environments.

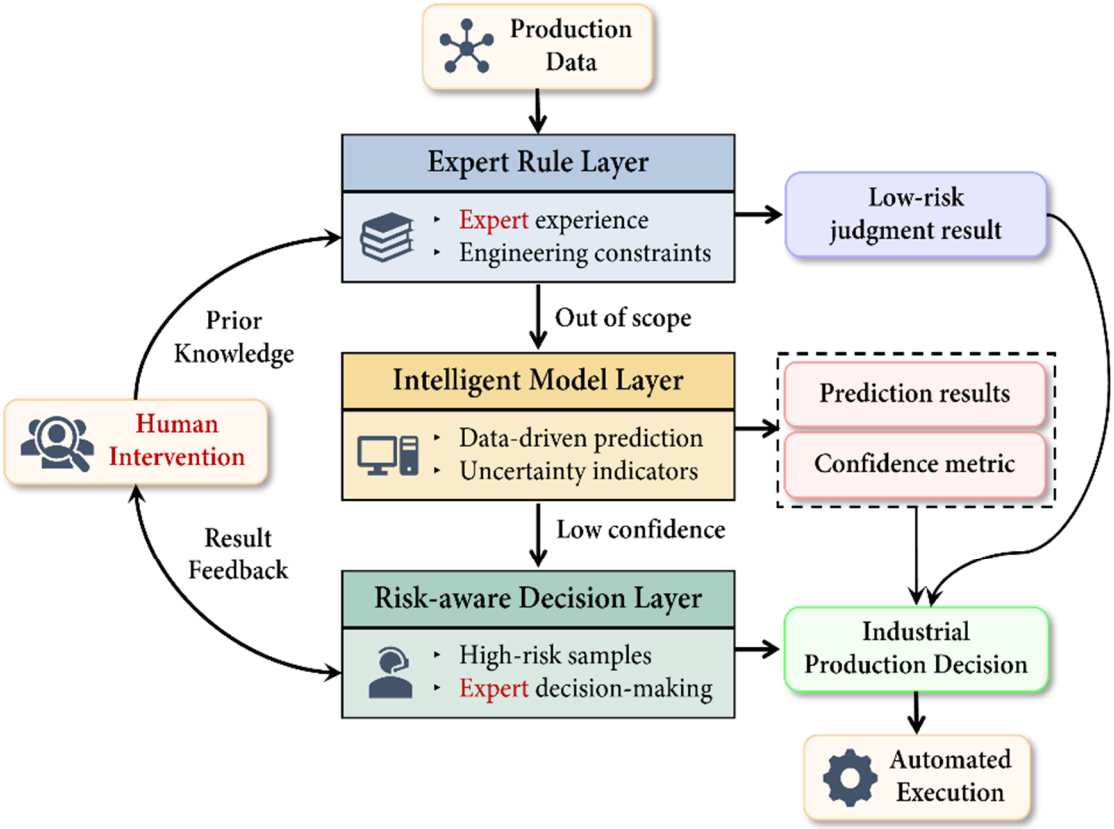


Figure 1. Human-in-the-Loop Intelligent Decision Framework (HIL-IDF).

In steel production, strip quality directly affects downstream product quality [21,22], making the determination of whether a strip requires flattening a critical production decision [23,24]. Currently, real-time data collected by strip-shape measurement sensors are used to generate bending and leveling curves that characterize deformation at different strip locations. However, determining whether subsequent flattening is required still relies on manual judgment, leading to low efficiency and high labor costs [25,26]. To this end, this paper focuses on the strip flatness judgment task and proposes a Multimodal Hybrid Intelligent Shape Detection method (MHI-SD) based on the proposed HIL-IDF framework. Here, HIL-IDF denotes the general three-layer human-in-the-loop decision framework, whereas MHI-SD is its task-specific implementation for strip flatness detection.

As shown in Figure 2a, conventional strip shape detection methods based on decision trees, neural networks, or clustering analysis [27–30] primarily rely on single models and single-modality features, resulting in reduced

reliability under complex operating conditions or noise interference. Within the HIL-IDF framework, MHI-SD addresses this limitation by incorporating multi-representation feature extraction and multi-model ensemble mechanisms. These representations are derived from the same production signals, including the original curve-based time-series representation and the transformed morphology-image representation.

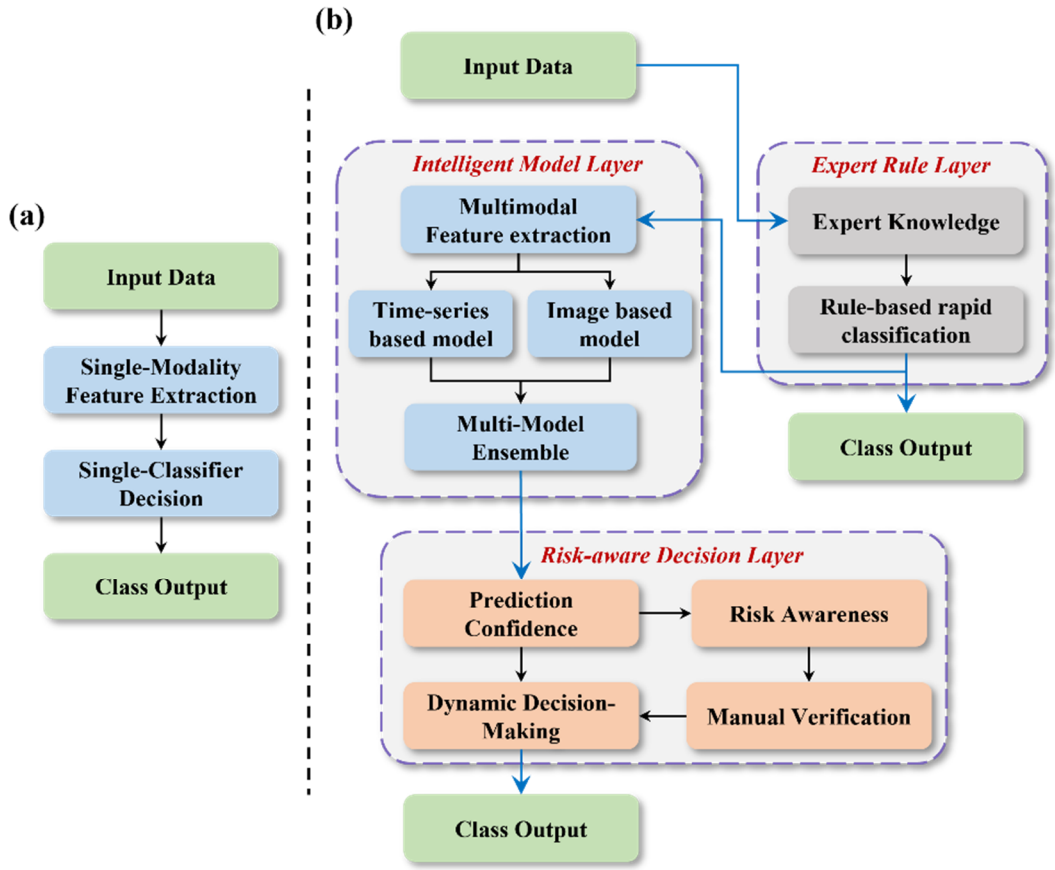


Figure 2. (**a**) Traditional fixed decision-making process using a single modality and a single model; (**b**) the multimodal hybrid intelligent shape detection method (MHI-SD) based on the HIL-IDF.

Specifically, as illustrated in Figure 2b, MHI-SD first applies expert-defined rules to rapidly classify samples with salient discriminative characteristics. Samples not covered by the expert rules are subsequently handled by the intelligent model layer, which includes both a time-series-based model and an image-based model to capture complementary representations. The time-series-based model employs an ensemble approach: it first extracts statistical features from segmented curves and then performs weighted fusion of classification probabilities from three base models, KNN, Random Forest, and LightGBM [31–33], to obtain more reliable classification results. For the image-based model, the curves are first converted into two-dimensional images. The image-based branch then extracts features using a CNN [34]. To further improve classification performance, an SVM [35] is employed as the final classifier, providing the result of the image-based classification. In the risk-aware decision mechanism, the time-series-based model serves as the primary probabilistic pathway, while the image-based model is used as an auxiliary risk-checking channel. Because a false no-flattening decision may cause downstream quality defects, the routing logic is designed from a cost-sensitive perspective and implemented as a threshold-based sequential routing mechanism. Instead of forcing all samples into automatic decisions, samples with low confidence in the time-series branch, potentially high-risk no-flattening predictions, or disagreement between the time-series and image-based branches are routed to auxiliary image checking or manual verification.

The main contributions of this work are summarized as follows:

(1) A human-in-the-loop intelligent decision framework is proposed, in which human experts are incorporated at two critical stages, rule-based prior and risk-aware decision-making. This design provides an uncertainty-aware risk-mitigation mechanism for ambiguous or high-risk samples while preserving automation efficiency within the validation scope of this study.

(2) For strip flatness judgment, a multimodal hybrid intelligent shape detection method is designed within the HIL-IDF framework, integrating time-series modeling and image-based analysis with a risk-aware dynamic decision mechanism to improve robustness and safety under complex and noisy conditions.
(3) The proposed method is evaluated using real production-line datasets collected from steel plants. The results show that MHI-SD improves classification accuracy and recall over the evaluated baselines, indicating its practical potential for strip flatness decision support.

2. Related Work

Based on different recognition principles, strip shape detection methods can be categorized as contact-based, non-contact-based, mechanism-based, and artificial intelligence-based methods. Table 1 summarizes the advantages and limitations of each method type.

Table 1. Advantages and limitations of different methods.

Methods	Advantages	Limitations
Contact-based [36]	Directly reflects strip state	Easily scratches strip steel
Non-contact-based [37–40]	High detection speed	Limited accuracy and highly environment-dependent
Mechanism-based [41]	Strong theoretical guidance	High computational cost and poor real-time performance
Artificial intelligence-based [22,26,30]	Data-driven and high efficiency	Relying on a single modality with limited generalization capabilities

Contact-based strip shape detection primarily relies on tension measurement [36], which measures the tension applied to the strip surface to determine whether flattening is required. However, direct physical contact may scratch or damage the strip surface. Non-contact strip shape detection methods can be further classified into vibration-based measurement [37], optical methods [38], and machine-vision-based inspection methods [39,40]. The vibration-based method requires the strip to pass through a sensing device equipped with air nozzles and vibration amplitude sensors, and it evaluates strip shape using tension-induced vibration responses. However, its measurement accuracy is relatively low. Optical methods use laser scanning to measure the strip surface profile; by analyzing the laser-line projection captured from the surface, strip flatness can be evaluated. However, these methods involve complex operating procedures and require high-precision instruments. Machine-vision-based methods use industrial cameras or other imaging devices to acquire visual data, which are then processed and analyzed using image-processing algorithms. Although these methods offer high detection speed and low cost, their performance is sensitive to ambient lighting. Mechanism-based strip shape detection mainly relies on the finite element method (FEM) [41] to simulate key physical processes during strip production, such as stress states and temperature distributions. However, this category of methods depends heavily on manually defined constraints, parameter settings, and underlying mathematical models.

With the widespread adoption of machine learning technologies, artificial intelligence models can be trained using historical data collected from strip steel production lines, thereby reducing redundant processing steps and significantly improving production efficiency [42–44]. Wang et al. [22] further addressed the multilayer nonlinear characteristics of strip shape prediction using a convolutional neural network (CNN), and designed an Inception-ResNet-based architecture. With fewer parameters and lower computational complexity, the proposed model exhibited outstanding predictive performance. Wang et al. [26] proposed a strip shape prediction model based on CNN. By incorporating an isolation forest-based outlier removal method and a data folding strategy, the model achieved improved prediction accuracy and reduced forecasting errors, demonstrating strong applicability in cold rolling production scenarios. Building on real industrial data from the cold rolling process, Yang et al. [30] employed ensemble learning approaches, including Bagging, Boosting, and Stacking, to experimentally evaluate the predictive performance of strip shape forecasting. They further developed a Stacking-based ensemble model, EALB-LR, for primary and secondary strip shape matching at the mill exit. The proposed model achieved higher prediction accuracy in strip shape forecasting tasks.

Recent industrial inspection studies have increasingly explored multimodal perception and uncertainty-aware prediction. Multimodal methods combine process signals, visual information, or domain knowledge to improve robustness under noisy production conditions [11,19,25]. Uncertainty-aware methods further estimate predictive confidence or identify unfamiliar samples to support more reliable industrial decisions [5,8,10]. However, in many existing studies, multimodal evidence and uncertainty estimates are mainly used as additional model inputs or diagnostic indicators, while their connection to downstream operational decisions remains insufficiently specified.

Several related industrial AI studies have addressed complementary manufacturing problems. Bi et al. [45] investigated dynamic job-shop scheduling, and Zhang et al. [46] studied industrial time-series prediction under variable operating states. In steel-related applications, Luo et al. [47, 48] focused on visual surface-defect inspection and classification, while Dong et al. [49] studied data-driven prediction and optimization of hot-rolled strip shape. These studies have advanced industrial AI from the perspectives of scheduling optimization, time-series prediction, visual inspection, and strip-shape modeling. Nevertheless, most of them mainly aim to improve prediction, classification, or optimization performance, and the subsequent decision stage is often simplified as directly accepting the model output.

This simplification is insufficient for safety-critical strip flatness judgment, where different error types have asymmetric consequences. In particular, an incorrect no-flattening decision may lead to downstream quality defects. Therefore, the key gap addressed in this work is not only to improve average classification accuracy, but also to develop a risk-aware routing mechanism that links model confidence, asymmetric error consequences, and human arbitration for uncertain or high-risk samples.

3. Proposed Method

This section introduces the proposed hybrid intelligent architecture. The method integrates expert rules, curve-based time-series features, and morphology-image-based auxiliary outputs derived from the same production signals to support reliable strip flatness classification and risk-aware decision routing.

3.1. Task Definition

On strip steel production lines, shape detection relies on sensor-derived bending and leveling signals. These signals are recorded as two one-dimensional time-series curves:

$$X_B = \{x_b(s) | s = 1,2,\dots,S\}, \quad X_L = \{x_l(s) | s = 1,2,\dots,S\} \tag{1}$$

where s denotes the position along the strip length, and the amplitude $x_i(s)$ represents the deviation relative to a reference horizontal baseline. Here, X_B corresponds to the bending curve, while X_L represents the leveling curve. The strip shape inspection task can be formulated as a binary classification problem. Given the pair of curves (X_B, X_L), the goal is to determine whether a strip requires flattening or can directly proceed without additional processing (no flattening required). Formally, let the classifier be defined as:

$$f_\theta : (X_B, X_L) \to y, \tag{2}$$

where

$$y = \begin{cases} 1, & \text{requires flattening (positive case)} \\ 0, & \text{no flattening required (negative case)} \end{cases} \tag{3}$$

Here, θ represents the model parameters to be learned from labeled production samples. From the perspective of an intelligent model, the task corresponds to learning a parameterized mapping function $f_\theta(\cdot)$ such that the empirical risk is minimized:

$$\theta^* = arg\min_\theta \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y^{(i)}, f_\theta(X_B^{(i)}, X_L^{(i)})) \tag{4}$$

where N denotes the number of training samples, $\mathcal{L}(\cdot)$ denotes the loss function. It is important to note that the requires-flattening class is the safety-critical class. A false negative, i.e., predicting no flattening when the strip actually requires flattening, may cause downstream product defects and production disruptions. Therefore, the proposed system does not simply aim to maximize average classification accuracy, but also seeks to reduce high-risk false negatives.

To account for the greater risk of missed flattening, MHI-SD adopts a conservative cascaded routing strategy rather than a single automatic classifier. The expert-rule layer first checks whether a sample satisfies deterministic production rules. If a rule-based decision can be made, the sample is directly assigned to either the requires-flattening class or the no-flattening class. Otherwise, the sample is deferred to the model layer. For rule-deferred samples, the time-series branch estimates the posterior probabilities $p_t(y = 1|x)$ and $p_t(y = 0|x)$ where $y = 1$ denotes requires flattening and $y = 0$ denotes no flattening. It then computes the odds score:

$$s(x) = \frac{p_t(y = 1|x)}{p_t(y = 0|x) + \epsilon} \tag{5}$$

and the confidence score is $c_t(x) = max\{p_t(y=1|x), p_t(y=0|x)\}$. Here, $s(x)$ is used for the preliminary class decision: the sample is predicted as requiring flattening if $s(x) \geq \theta_o$, and as no flattening otherwise. The confidence score $c_t(x)$ measures the certainty of this preliminary decision. Samples with low confidence or a preliminary no-flattening decision are further checked by the auxiliary image branch because missed flattening is the more critical error. If the time-series and image-based branches disagree, or if $c_t(x)$ is below the manual-verification threshold, the sample is routed to human experts for final arbitration. Otherwise, the automatic decision is accepted. In this way, expert-rule filtering, probabilistic model inference, auxiliary checking, and human arbitration are integrated into a practical risk-aware routing mechanism.

3.2. Framework

The hybrid intelligent method proposed in this study is illustrated in Figure 3. The input data X_B, X_L are first evaluated through a rule-based check to determine whether their characteristics fall within the scope of the expert rules. If the rule conditions are satisfied, the output result y is directly produced according to these rules. Otherwise, the data are further analyzed by the intelligent models, which include a time-series-based model and an image-based model. The expert rules mainly include the following categories: whether the bending value exceeds 300; whether the minimum bending value is less than 50 while the leveling value is greater than the bending value; and whether the leveling signal contains any values greater than 50. It should be noted that, due to industrial confidentiality, the complete rule set cannot be fully disclosed.

For samples not covered by the expert rules, the time-series-based model serves as the primary probabilistic branch. As shown in Figure 3, it outputs a pair (y_i^T, c_i^T), where y_i^T denotes the predicted class label and c_i^T represents the corresponding confidence score. Equivalently, for the i-th sample x_i, the model estimates the posterior probabilities $p_t(y=1|x_i)$ and $p_t(y=0|x_i)$, where $y=1$ denotes the requires-flattening class and $y=0$ denotes the no-flattening class. The confidence score is defined as $c_i^T = max\{p_t(y=1|x_i), p_t(y=0|x_i)\}$. The risk-aware decision layer uses (y_i^T, c_i^T) to determine whether the preliminary decision can be accepted automatically or should be further checked. Specifically, low-confidence samples and preliminary no-flattening decisions are conditionally passed to the image-based auxiliary branch, which outputs an auxiliary label $g(x_i)$. If the time-series and image-based branches disagree, or if c_i^T is below the manual-verification threshold, the sample is routed to human experts for final arbitration. Otherwise, the automatic decision is accepted. In this way, the image-based branch functions as an auxiliary checking module, while human experts provide final arbitration for uncertain or inconsistent cases.

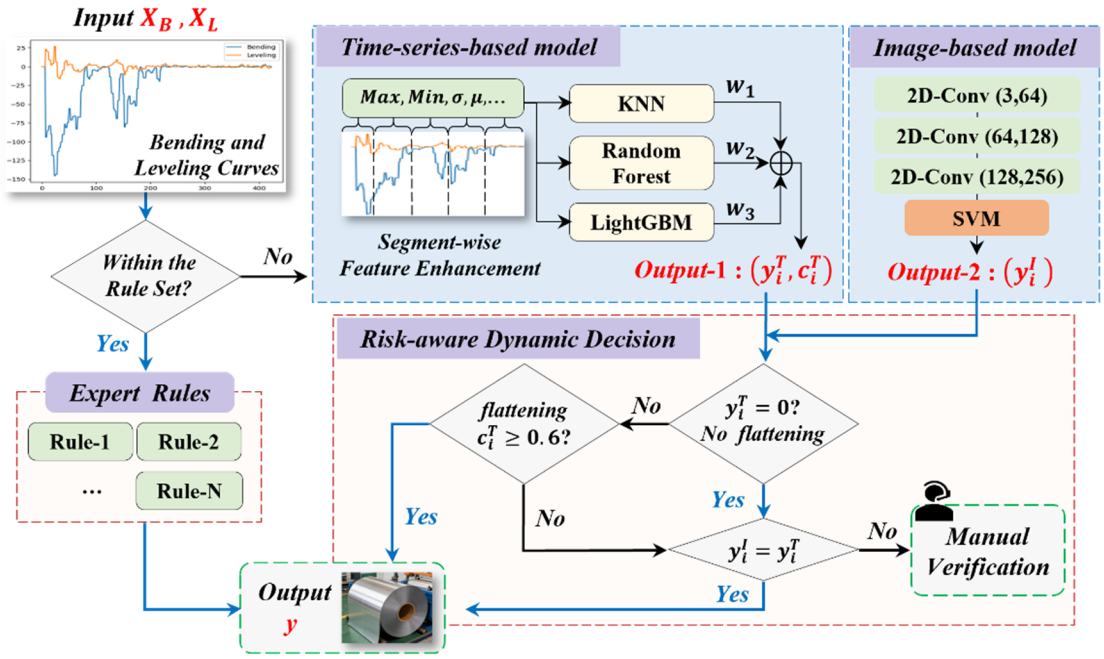


Figure 3. The hybrid intelligent method, consisting of expert rules, a time-series-based model, an image-based model, and a risk-aware dynamic decision mechanism.

3.3. Time-Series-Based Model

To enrich the data representation, we segmented the Bending and Leveling curves and extracted statistical features from each segment. Specifically, the curves were divided into segments of length 40, and for each segment, we computed the maximum, minimum, variance, and mean of features X_B and X_L, as well as the position of the minimum Bending value and the maximum absolute difference between Bending and Leveling. These additional features X_{add} were then combined with the original features X_B and X_L to form an enhanced feature set $X_D = \mathrm{Concat}(X_B, X_L, X_{add})$. The subsequent time-series models were trained using this new feature set X_D.

To enhance the stability and generalization capability of time-series classification, this study adopts a weighted ensemble strategy that integrates the prediction outputs of three base models: KNN, Random Forest (RF), and LightGBM (LGB). Specifically, each base learner produces a probability vector for the two classes (positive and negative), denoted as:

$$p_i^{(KNN)} = \left[p_{i,0}^{(KNN)}, p_{i,1}^{(KNN)}\right],\ p_i^{(RF)} = \left[p_{i,0}^{(RF)}, p_{i,1}^{(RF)}\right],\ p_i^{(LGB)} = \left[p_{i,0}^{(LGB)}, p_{i,1}^{(LGB)}\right] \tag{6}$$

where $p_{i,1}$ and $p_{i,0}$ denote the predicted probabilities that sample i belongs to the positive and negative classes, respectively. The outputs of the three models are then combined through a weighted average using predefined weights to obtain the fused probability $\hat{p}_i$. In this study, the weights for the three models are set to 0.40, 0.45, and 0.15 for RF, LGB, and KNN, respectively.

$$\hat{p}_i = w_{KNN} \cdot p_i^{(KNN)} + w_{RF} \cdot p_i^{(RF)} + w_{LGB} \cdot p_i^{(LGB)} \tag{7}$$

$$y_i^T = \begin{cases} 1, & s_i \geq \theta_o \\ 0, & s_i < \theta_o \end{cases}, \quad s_i = \frac{\hat{p}_{i,1}}{\hat{p}_{i,0} + \epsilon} \tag{8}$$

The time-series decision is made by comparing the odds score with an odds threshold θ_o. Specifically, the sample is predicted as requiring flattening if $s_i \geq \theta_o$, and as no flattening otherwise. The odds threshold is not a posterior-probability threshold; its equivalent posterior threshold is $p^* = \theta_o/(1+\theta_o)$. Thus, $\theta_o = 0.5$ corresponds to $\hat{p}_{i,1} \geq 0.33$. Since missed flattening is the more critical error, an odds threshold below the neutral value of 1 is used to favor the requires-flattening class.

3.4. Image-based Model

In the image-based classification channel, the original pair of bending and leveling curves is transformed into a two-dimensional morphology image X_g. This transformation is not intended to introduce new physical measurements; instead, it provides a different representation space for describing curve morphology. The two-dimensional representation enables convolutional filters to capture local curve structures, including fluctuations, crossings, abnormal peaks, and relative shape differences between the bending and leveling curves. After feature extraction by a convolutional neural network (CNN), the image representation is obtained as:

$$z = \mathrm{CNN}(X_g) \tag{9}$$

The CNN consists of three two-dimensional convolutional layers, with the number of channels increasing from 3 to 256. Each convolutional layer is followed by batch normalization, a ReLU activation function, and a max-pooling layer. The extracted feature vector z is then fed into a support vector machine (SVM) for classification:

$$y_i^I = \mathrm{SVM}(z) \tag{10}$$

The image-based branch is used as an auxiliary verification module rather than a replacement for the time-series branch. It provides a morphology-oriented check for low-confidence samples or preliminary no-flattening cases, and its output is used for consistency checking with the time-series branch. When the two branches disagree, the sample is routed to manual verification.

During model training, the CNN is optimized using the Adam optimizer with a learning rate of 0.001, while the SVM employs a radial basis function (RBF) kernel to handle nonlinear class boundaries. The dataset is partitioned into training and test sets at a 7:3 ratio to ensure reliable performance evaluation. Training and inference are conducted on a workstation equipped with an NVIDIA RTX 4090 GPU with 24 GB of memory.

4. Experiments and Results

In this study, a total of 4441 steel strip samples with Bending and Leveling curve data were used, including 3201 samples for training and 1240 samples for testing. Long Short-Term Memory (LSTM) [50], signal decomposition (SD) model [51], LightGBM, Random Forest (RF), and a one-dimensional convolutional neural network (1D-CNN) were selected as baseline models. Accuracy, recall, F1-score [52], false-negative rate (FNR), and false-positive rate (FPR) were used to evaluate both overall classification performance and safety-related error types. Here, recall denotes the proportion of samples that actually require flattening and are correctly identified. To improve reproducibility, fixed settings were used for the conventional tabular models, while the weighted fusion model combined RF, LightGBM, and KNN with weights of 0.4, 0.45, and 0.15, respectively, using an odds threshold of 0.5. The CNN-SVM image branch was constructed with a 2D CNN feature extractor and an RBF-SVM classifier, and was used as an auxiliary verifier for risk checking.

4.1. Comparative Experiments

The comparative experimental results are presented in Table 2. "Ensemble" denotes the weighted fusion of RF, LightGBM, and KNN based on segment-wise statistical features. "Rules + Ensemble" represents the proposed rule-filtered weighted time-series ensemble branch, which integrates expert-rule filtering, segment-wise statistical feature extraction, RF/LightGBM/KNN weighted fusion, and the odds-threshold decision. "Rules + 1D-CNN" introduces the same expert-rule layer before the 1D-CNN baseline, so that samples with deterministic production patterns are first handled by expert rules. "Ours-Full" denotes the automatic MHI-SD pipeline before expert correction, which further incorporates the image-based auxiliary branch for consistency checking. The performance of Ours-Full in Table 2 is reported before expert correction, so that all methods are compared under the same automatic evaluation setting.

Table 2. The comparative experimental results.

Model	Accuracy (%)	Recall (%)	F1-Score (%)
LSTM	52.91	47.48	53.45
SD	75.66	82.87	79.89
LightGBM	78.31	83.80	81.89
RF	74.60	74.49	77.07
Ensemble	87.58	80.00	83.15
1D-CNN	88.09	80.70	83.84
Rules + Ensemble	94.35	89.26	92.37
Rules + 1D-CNN	94.06	90.25	92.09
Ours-Full	96.32	95.19	94.58

Compared with the conventional baselines, Ours-Full achieves the best overall performance, with 96.32% accuracy, 95.19% recall, and 94.58% F1-score. The Ensemble model outperforms the single tabular models in terms of F1-score, indicating that segment-wise statistical features and ensemble learning can capture complementary flatness patterns. The 1D-CNN baseline also achieves competitive results. After applying the same expert-rule layer, Rules + Ensemble obtains 94.35% accuracy, while Rules + 1D-CNN obtains 94.06% accuracy. This comparison shows that both feature-based ensemble learning and deep sequence modeling benefit substantially from expert-rule filtering. Under the same rule-filtered setting, Rules + Ensemble achieves slightly higher accuracy and F1-score, whereas Rules + 1D-CNN obtains slightly higher recall. Therefore, the proposed feature-based ensemble branch provides a comparable but more interpretable solution for the current limited-sample industrial scenario. Compared with Rules + Ensemble, Ours-Full further improves the accuracy, recall, and F1-score to 96.32%, 95.19%, and 94.58%, respectively. The improvement is modest but consistent, which is reasonable because the image-based branch is only activated for uncertain or preliminary no-flattening cases.

The comparisons among Ensemble, 1D-CNN, Rules + Ensemble, Rules + 1D-CNN, and Ours-Full provide a progressive framework-level analysis. The results indicate that the final improvement is achieved through the cascaded integration of expert-rule filtering, segment-wise feature representation, weighted ensemble learning, and risk-aware routing, rather than by simply combining multiple classifiers. Specifically, expert rules reduce the ambiguity of deterministic samples, segment-wise statistical features capture local extrema and bending-leveling differences, the weighted ensemble integrates the complementary strengths of LightGBM, RF, and KNN, and the routing mechanism avoids forcing automatic decisions on uncertain samples.

4.2. Validation of the Risk-Aware Routing Mechanism

To examine the behavior of the risk-aware routing mechanism, two threshold-sensitivity studies were conducted on the test set. The first study varies the odds threshold used by the time-series branch, while the second study varies the confidence threshold for manual verification. Both studies are based on the time-series model before image-assisted routing, and are used to evaluate the trade-off between classification performance, conservative risk control, and manual-verification workload rather than to claim an analytically optimal threshold.

Table 3 reports the sensitivity of the odds threshold. The odds threshold controls the tendency of the time-series branch to predict the safety-critical requires-flattening class. A lower odds threshold corresponds to a lower posterior-probability threshold and therefore makes the model more conservative toward the requires-flattening class. The results in Table 3 show that the odds threshold directly affects the recall-oriented safety-performance trade-off. When the threshold is reduced from 0.50 to 0.25, recall increases from 89.26% to 90.11%, indicating that the model identifies more samples requiring flattening. In contrast, increasing the threshold leads to lower recall, although the overall accuracy remains relatively stable. These results indicate that the odds threshold is a controllable parameter for adjusting the balance between conservative detection and classification performance.

Table 3. Odds-threshold sensitivity of the risk-aware hybrid decision mechanism.

Odds Threshold	Equivalent Probability Threshold	Accuracy (%)	Recall (%)	F1-Score (%)
0.25	0.20	93.76	90.11	91.89
0.50	0.3333	94.35	89.26	92.37
1.00	0.50	93.95	87.58	91.73
1.50	0.60	94.03	86.95	91.78
2.00	0.6667	93.55	85.68	91.05

In Table 4, “Manual” denotes the proportion of samples routed to human experts, and “Final Acc.” and “Final Rec.” are calculated under the assumption that manually verified samples are corrected by experts. As shown in Table 4, increasing the confidence threshold routes more samples to manual verification and improves the final accuracy under the expert-correction assumption, but it also increases manual workload. The final recall remains unchanged from the 55% to 65% thresholds, indicating that the additionally routed samples in this range mainly affect overall classification accuracy rather than correcting additional false-negative cases. When the threshold is increased to 70%, the final recall slightly improves from 89.26% to 89.89%, suggesting that the more conservative setting begins to cover additional missed-flattening cases. Therefore, the 60% threshold is used as the default high-automation setting, maintaining an automation ratio of 99.11% with only 0.89% of samples requiring manual verification. The 70% threshold is reported as a more conservative alternative for scenarios with stricter safety requirements. Overall, the sensitivity results support threshold-based routing as a practical risk-control mechanism, while the final threshold can be adjusted according to production requirements.

Table 4. Manual-verification sensitivity of the time-series branch under different confidence thresholds.

Confidence Threshold	Manual (%)	Auto (%)	Final Acc. (%)	Final Rec. (%)
<55%	0.56	99.44	94.60	89.26
<60%	0.89	99.11	94.68	89.26
<65%	1.61	98.39	94.92	89.26
<70%	2.74	97.26	95.16	89.89

5. Conclusions

This work presents a Human-in-the-Loop Intelligent Decision Framework (HIL-IDF) for high-risk industrial decision-making and develops a Multimodal Hybrid Intelligent Strip Detection method (MHI-SD) for strip flatness classification in steel production. By integrating expert rules, time-series modeling, morphology-image auxiliary verification, and a risk-aware routing and human-arbitration mechanism, the proposed framework provides a structured way to combine automatic inference with expert intervention. Experiments on the available real production-line dataset show that MHI-SD improves classification performance over the evaluated single-model and single-representation baselines. The ablation study, threshold-sensitivity analysis, and deployment-oriented discussion further suggest that expert-rule filtering, morphology-image auxiliary checking, and confidence-based manual verification can help reduce high-risk automatic decisions for uncertain or ambiguous samples.

Future work will focus on improving the methodological rigor and generalizability of the proposed framework, including uncertainty calibration, adaptive threshold selection, and cost-sensitive routing to better balance decision safety and automation efficiency. In addition, systematic production-line validation will be conducted under more diverse operating conditions, with complete records of auxiliary-checking cases, manual-verification cases, expert agreement, and abnormal events.

Author Contributions

Q.K.: conceptualization, methodology, writing—reviewing and editing, supervision, funding acquisition; K.W.: data curation, writing—original draft; J.S.: visualization, investigation. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported in part by the Natural Science Foundation of Shanghai (23ZR1466000).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The datasets analyzed in this study were obtained from industrial steel plate shape detection data provided by the corresponding data provider. Due to proprietary, confidentiality, copyright, and licensing restrictions, the raw data are not publicly available. The processed data and code used to support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. Zhong, R.Y.; Xu, X.; Klotz, E.; et al. Intelligent Manufacturing in the Context of Industry 4.0: A Review. *Engineering* **2017**, *3*, 616–630.
2. Perez-Cerrolaza, J.; Abella, J.; Borg, M.; et al. Artificial Intelligence for Safety-Critical Systems in Industrial and Transportation Domains: A Survey. *ACM Comput. Surv.* **2024**, *56*, 1–40.
3. Basora, L.; Viens, A.; Chao, M.A.; et al. A Benchmark on Uncertainty Quantification for Deep Learning Prognostics. *Reliab. Eng. Syst. Saf.* **2025**, *253*, 110513.
4. Windmann, A.; Wittenberg, P.; Schieseck, M.; et al. Artificial Intelligence in Industry 4.0: A Review of Integration Challenges for Industrial Systems. In Proceedings of the 2024 IEEE 22nd International Conference on Industrial Informatics (INDIN), Porto, Portugal, 18–22 August 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–8.
5. Abdar, M.; Pourpanah, F.; Hussain, S.; et al. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Inf. Fusion* **2021**, *76*, 243–297.
6. Zhou, H.; Chen, W.; Cheng, L.; et al. Trustworthy Fault Diagnosis with Uncertainty Estimation through Evidential Convolutional Neural Networks. *IEEE Trans. Ind. Inform.* **2023**, *19*, 10842–10852.
7. Lin, Y.H.; Li, G.H. A Bayesian Deep Learning Framework for RUL Prediction Incorporating Uncertainty Quantification and Calibration. *IEEE Trans. Ind. Inform.* **2022**, *18*, 7274–7284.
8. Yang, J.; Zhou, K.; Li, Y.; et al. Generalized Out-of-Distribution Detection: A survey. *Int. J. Comput. Vis.* **2024**, *132*, 5635–5662.
9. Wang, X.; Zhang, H.; Qiao, X.; et al. Generalized Out-of-Distribution Fault Diagnosis (GOOFD) via Internal Contrastive Learning. *IEEE Trans. Ind. Inform.* **2024**, *20*, 9987–9996.

10. Wei, D.; Zuo, M.; Tian, Z. Open-Set Fault Diagnosis for Industrial Rotating Machines Based on Trustworthy Deep Learning. *IEEE Trans. Ind. Cyber-Phys. Syst.* **2025**, *3*, 181–189.
11. Wang, J.; Ma, Y.; Zhang, L.; et al. Deep Learning for Smart Manufacturing: Methods and Applications. *J. Manuf. Syst.* **2018**, *48*, 144–156.
12. Yang, Y.; Zhang, Y.; Zou, W.; et al. Synthesizing Control Barrier Functions with Feasible Region Iteration for Safe Reinforcement Learning. *IEEE Trans. Autom. Control* **2023**, *69*, 2713–2720.
13. Khan, Z.; Nasir, A.; Mekid, S. Fault-Tolerant Control Strategies for Industrial Robots: State of the Art and Future Perspective on AI-Based Fault Management. *Artif. Intell. Rev.* **2025**, *58*, 362.
14. Chen, H.; Li, S.; Fan, J.; et al. Human-in-the-Loop Robot Learning for Smart Manufacturing: A Human-Centric Perspective. *IEEE Trans. Autom. Sci. Eng.* **2025**, *22*, 11062–11086.
15. Hu, D.; Huang, C.; Wu, J.; et al. Toward Multi-Task Generalization in Autonomous Navigation: A Human-in-the-Loop Adversarial Reinforcement Learning with Diffusion Policy. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 19493–19507.
16. Hu, D.; Zhou, G.; Wu, J.; et al. Trust-Calibrated Human-in-the-Loop Reinforcement Learning for Safe and Efficient Autonomous Navigation. *IEEE Internet Things J.* **2025**, *12*, 39029–39041.
17. Wang, J.; Zhang, Z.; Huang, Y.; et al. Proactive Dispatch of Rail Maintenance Equipment under Weather-Induced Disruptions: An Optimization Framework Integrating Machine Learning-Driven Risk Assessment. *Adv. Eng. Inform.* **2026**, *69*, 103990.
18. Zhang, X.; Yan, S.; Wang, C.; et al. Generalizable Fault Diagnosis under Distribution Shifts Induced by Unseen Working Conditions via Synthetic and Adversarial Sample Learning. *IEEE Trans. Reliab.* **2025**, *74*, 5517–5530.
19. Alzarooni, A.; Iqbal, E.; Khan, S.U.; et al. Anomaly Detection for Industrial Applications: Challenges, Solutions, and Future Directions. *arXiv* **2025**, arXiv:2501.11310.
20. Wiggerthale, J.; Reich, C. Human-in-the-Loop Visual Inspection for Safety-Critical Systems. *Preprints* **2025**, 2025112107. https://doi.org/10.20944/preprints202511.2107.v1.
21. Geng, Y.F.; Zhao, C.L.; Zang, X.M.; et al. Feeding Steel Strip Technology in Continuous Casting Process: A Review. *Steel Res. Int.* **2021**, *92*, 2100051.
22. Wang, Q.; Sun, J.; Hu, Y.; et al. Deep Learning-*Based* Flatness Prediction via Multivariate Industrial Data for Steel Strip During Tandem Cold Rolling. *Expert Syst. Appl.* **2024**, *237*, 121777.
23. Hu, X.; Song, H.; Yang, K.; et al. Research on the Flatness Control Performance and Parameter Influence Law of Four-High Skin Pass Rolling Mill. *J. Mater. Eng. Perform.* **2025**, *34*, 8141–8151.
24. Wu, C.; Zhang, L.; Qu, P.; et al. Effect of the Elastic Deformation of Rolls on the Surface Texture Transfer in Skin-Pass Rolling. *Int. J. Mech. Sci.* **2021**, *198*, 106358.
25. Xu, Y.; Wang, D.; Liu, H.; et al. Intelligent Representation Method of Image Flatness for Cold Rolled Strip. *J. Iron Steel Res. Int.* **2024**, *31*, 1177–1195.
26. Wang, Y.; Li, C.; Peng, L.; et al. Application of Convolutional Neural Networks for Prediction of Strip Flatness in Tandem Cold Rolling Process. *J. Manuf. Process.* **2021**, *68*, 512–522.
27. Cui, D. Application of SOCP-SVM Algorithm on Strip Steel Surface Defect Recognition. *Int. J. Comput. Intell. Syst. Appl.* **2016**, *4*, 1–7.
28. Hu, H.; Li, Y.; Liu, M.; et al. Classification of Defects in Steel Strip Surface Based on Multiclass Support Vector Machine. Multimed. *Tools Appl.* **2014**, *69*, 199–216.
29. Zhang, X.W.; Ding, Y.Q.; et al. A Vision Inspection System for the Surface Defects of Strongly Reflected Metal Based on Multi-Class SVM. *Expert Syst. Appl.* **2011**, *38*, 5930–5939.
30. Yang, W.; Zhao, Z.; Zhu, L.; et al. Strip Flatness Prediction of Cold Rolling Based on Ensemble Methods. *J. Iron Steel Res. Int.* **2024**, *31*, 237–251.
31. Guo, G.; Wang, H.; Bell, D.; et al. KNN Model-Based Approach in Classification. In Proceedings of the OTM Confederated International Conferences: On the Move to Meaningful Internet Systems, Catania, Sicily, Italy, 3–7 November 2003; Springer: Berlin/Heidelberg, Germany, 2003.
32. Salman, H.A.; Kalakech, A.; Steiti, A. Random Forest Algorithm Overview. *Babylonian J. Mach. Learn.* **2024**, *2024*, 69–79.
33. Ke, G.; Meng, Q.; Finley, T.; et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–9.
34. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; et al. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, and Future Directions. *J. Big Data* **2021**, *8*, 53.
35. Chandra, M.A.; Bedi, S.S. Survey on SVM and Their Application in Image Classification. *Int. J. Inf. Technol.* **2021**, *13*, 1–11.
36. Wang, X.C.; Yang, Q.; Jiang, Z.Y.; et al. Research on the Improvement Effect of High Tension on Flatness Deviation in Cold Strip Rolling. *Steel Res. Int.* **2014**, *85*, 1560–1570.

37. Usamentiaga, R.; Garcia, D.F.; Molleda, J.; et al. Vibrations in Steel Strips: Effects on Flatness Measurement and Filtering. In Proceedings of the 2013 IEEE Industry Applications Society Annual Meeting, Lake Buena Vista, FL, USA, 6–11 October 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 1–10.
38. Krauhausen, M.; Priem, R.; Claßen, R.; et al. Sub-Micron Inline Thickness Measurement of Cold-Rolled Metal Strips by Multi-Wavelength Interferometry and Laser Triangulation. *Opt. Express* **2023**, *31*, 43804–43820.
39. Tang, B.; Chen, L.; Sun, W.; et al. Review of Surface Defect Detection of Steel Products Based on Machine Vision. *IET Image Process.* **2023**, *17*, 303–322.
40. Yan, S.; Wang, X.; Yang, Q.; et al. Online Deviation Measurement System of the Strip in the Finishing Process Based on Machine Vision. *Measurement* **2022**, *202*, 111735.
41. Chai, X.; Li, H.; Zhang, J.; et al. Flatness Analysis and Control of Strips with Different Thickness in 2250 mm Hot Tandem Rolling. *Steel Res. Int.* **2018**, *89*, 1800404.
42. Shi, X.; Kang, Q.; An, J.; Zhou, M. Novel L1-Regularized Extreme Learning Machine for Soft-Sensing of an Industrial Process. *IEEE Trans. Ind. Inform.* **2022**, *18*, 1009–1017.
43. Shi, X.; Kang, Q.; Zhou, M.; et al. Dual Attention-Aided Cooperative Deep Spatiotemporal Feature Extraction Network for Semi-Supervised Soft Sensing. *IEEE Robot. Autom. Lett.* **2025**, *10*, 2184–2190.
44. Shi, X.; Kang, Q.; Bao, H.; et al. Principal Component-Based Semi-Supervised Extreme Learning Machine for Soft Sensing. *IEEE Trans. Autom. Sci. Eng.* **2024**, *21*, 3966–3976.
45. Bi, J.; Wang, C.; Wang, Z.; et al. Dual-GNN-Driven Cooperative Optimization for Makespan-Minimized and Large-Scale 3C Dynamic Job-Shop Scheduling. *IEEE Trans. Autom. Sci. Eng.* **2026**, *23*, 4350–4364. https://doi.org/10.1109/TASE.2026.3661178.
46. Zhang, M.; Bi, J.; Yuan, H.; et al. An Enhanced Hybrid Deep Neural Network Method for Adjusted Industrial Time Series Prediction with Variable Operating States. *Expert Syst. Appl.* **2026**, *299*, 130029. https://doi.org/10.1016/j.eswa.2025.130029.
47. Luo, Q.; Fang, X.; Liu, L.; et al. Automated Visual Defect Detection for Flat Steel Surface: A Survey. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 626–644. https://doi.org/10.1109/TIM.2019.2963555.
48. Luo, Q.; Fang, X.; Sun, Y.; et al. Surface Defect Classification for Hot-Rolled Steel Strips by Selectively Dominant Local Binary Patterns. *IEEE Access* **2019**, *7*, 23488–23499. https://doi.org/10.1109/ACCESS.2019.2898215.
49. Dong, J.; Yang, X.-Y.; Xu, Q.-Q.; et al. A Data-Driven Integrated Framework for Collaborative Prediction and Optimisation of Hot-Rolled Steel Strip Shape. *Int. J. Prod. Res.* **2026**, 1–26. https://doi.org/10.1080/00207543.2026.2623538.
50. Chen, Y.; Peng, L.; Wang, Y.; et al. Prediction of Tandem Cold-Rolled Strip Flatness Based on Attention-LSTM Model. *J. Manuf. Process.* **2023**, *91*, 110–121.
51. Jiang, Z.; Ding, Z.; Liu, Y.; et al. A Data-Driven Based Decomposition–Integration Method for Remanufacturing Cost Prediction of End-of-Life Products. *Robot. Comput. Integr. Manuf.* **2020**, *61*, 101838.
52. Kang, Q.; Chen, X.S.; Li, S.S.; et al. A Noise-Filtered Under-Sampling Scheme for Imbalanced Classification. *IEEE Trans. Cybern.* **2016**, *47*, 4263–4274.