



Uncertainty Quantification of Data-driven Quality Prediction Model For Realizing the Active Sampling Inspection of Mechanical Properties in Steel Production

Yong Song¹ · Feifei Li¹ · Zheng Wang¹ · Baozhong Zhang² · Borui Zhang²

Received: 4 December 2023 / Accepted: 11 March 2024
© The Author(s) 2024

Abstract

Pre-production quality defect inspection is a crucial step in industrial manufacturing, and many traditional inspection strategies suffer from inefficiency issues. This is especially true for tasks such as mechanical performance testing of steel products, which involve time-consuming processes like offline sampling, specimen preparation, and testing. The inspection volume significantly impacts the production cycle, inventory, yield, and labor costs. Constructing a data-driven model for predicting product quality and implementing proactive sampling inspection based on the prediction results is an appealing solution. However, the prediction uncertainty of data-driven models poses a challenging problem that needs to be addressed. This paper proposes an active quality inspection approach for steel products based on the uncertainty quantification in the predictive model for mechanical performance. The objective is to reduce both the sampling frequency and the omission rate on the production site. First, an ensemble model based on improved lower and upper bound estimation is established for interval prediction of mechanical performance. The uncertainty of the specific value prediction model is quantitatively estimated using interval probability distributions. Then, a predictive model for the mechanical performance failure probability is built based on the prediction interval size and probability distribution. By determining an appropriate probability threshold, the trade-off between prediction accuracy and defect detection accuracy (recall rate) is balanced, enabling the establishment of an active sampling strategy. Finally, this functionality is integrated into the manufacturing execution system of a steel factory, realizing a mechanical performance inspection approach based on proactive sampling. The proposed approach is validated using real production datasets. When the probability threshold is set to 30%, the prediction accuracy and recall rate for failure mechanical performance samples are 75% and 100%, respectively. Meanwhile, the sampling rate is only 5.33%, while controlling the risk of omission. This represents a 50% reduction in sampling rate compared to the inspection rules commonly used in actual production. The overall efficiency of product quality inspection is improved, and inspection costs are reduced.

Keywords Quality inspection · Active sampling · Mechanical performance prediction · Uncertainty quantification

1 Introduction

In the manufacturing industry, it is generally necessary to implement effective quality inspection procedures to prevent defective products from being delivered to customers [1].

High-quality products are crucial for the long-term competitiveness of manufacturing enterprises. However, the increasing demand for product customization and complexity has resulted in a significant increase in inspection volume. This has made the inspection process a bottleneck in

✉ Yong Song
songyong@ustb.edu.cn

Feifei Li
15311449327@163.com

Zheng Wang
wz13853952310@163.com

Baozhong Zhang
zhangbaozhong@ningbosteel.com

Borui Zhang
zhangborui@ningbosteel.com

¹ National Engineering Research Center for Advanced Rolling and Intelligent Manufacturing, University of Science and Technology Beijing, Beijing 100085, China

² Manufacturing Management Department, Ningbo Iron & Steel Co., Ltd, Zhejiang 315800, China

the production-to-delivery process [2]. Specifically, in tasks such as mechanical performance inspection of steel products, which involve time-consuming processes like offline sampling, specimen preparation, and testing, the inspection volume has a significant impact on production cycles, inventory, yield, and labor costs [3, 4]. In the context of intelligent manufacturing, researching an advanced inspection strategy and solution to replace the traditional low-efficiency inspection model has become one of the keys for steel companies to improve quality and efficiency. Advanced quality inspection strategies and approaches not only reduce inspection costs but also bring overall manufacturing cost advantages [5, 6].

A comprehensive inspection can eliminate the possibility of defective products reaching customers. However, sampling inspection can reduce inspection costs. To achieve a balance between cost and risk, it is necessary to analyze the impact of actual defect rate, inspection error rate, sample size, and unit inspection cost on the choice between full inspection and sampling inspection [7]. Considering that these impacts are closely related to the complex dynamics and stochastic behavior of manufacturing systems, using dynamic sampling inspection methods can find the optimal production and maintenance control strategies, as well as the correct sampling strategies, to reduce overall manufacturing costs [8]. However, even more complex sampling inspection rules cannot completely eliminate the risk of missed inspections. Mistakenly classifying qualified products as unqualified products increases the cost risk for manufacturing companies, while releasing unqualified products poses a risk to customers and ultimately affects the reputation of manufacturing companies, with the latter's overall cost being much higher than the former [5]. Therefore, the main challenge for time-consuming and labor-intensive product quality inspection processes, such as mechanical performance testing of steel products, lies in how to achieve accurate detection and assessment of product quality through more efficient inspection strategies (with fewer offline sampling quantities), striking a balance between cost and risk.

With the development of industrial big data, researchers are trying to build data-driven models for product quality prediction and actively use sampling inspection based on the prediction results to improve the efficiency of overall product quality inspection and reduce inspection costs [9]. The active sampling inspection strategy greatly reduces the inspection volume by only inspecting products with uncertain prediction results (for steel production, this refers to products classified as unqualified and those whose qualification cannot be determined) [10]. Therefore, the core of the active sampling inspection strategy lies in the accuracy and reliability of the quality prediction model. Due to the ability of neural networks to model nonlinear dependencies, they are widely used for product quality prediction [11–13]. In

order to obtain better machine learning models, ensemble methods are used to combine models to ensure good prediction results under different conditions. Moreover, the performance of ensemble models largely depends on the selected fusion hyperparameters [14]. A more practical solution is to balance the prediction performance by considering the acceptance rate of misclassified products in different scenarios, between prediction accuracy and the accuracy of capturing defective products (recall rate) [12]. However, besides balancing the prediction performance by selecting different modeling methods based on prior knowledge, there is currently a lack of more intuitive and easy-to-implement general methods.

For steel production, both products classified as unqualified by the quality prediction model and products that cannot be highly classified as qualified need to undergo active sampling inspection. If the quality prediction model classifies too many qualified products as 'defective products', the active sampling inspection strategy may lose its meaning because it cannot effectively reduce the sampling volume. Therefore, the main optimization objective for the mechanical performance prediction model used in the active sampling inspection strategy should be the ability to accurately classify 'defective products' (mechanical performance unqualified) with high certainty. However, quantifying the prediction uncertainty of data-driven models is a challenging problem [15, 16].

This paper proposes an active sampling inspection scheme for steel product quality based on the uncertainty quantification in mechanical performance prediction models, which can simultaneously reduce the sampling frequency and the rate of missed inspections on the production site. In this scheme, an ensemble model based on improved lower and upper bound estimation (LUBE) is established for interval prediction of mechanical performance, quantifying the uncertainty of specific value prediction models through interval probability distribution estimation. Then, a prediction model for the mechanical performance failure probability is established based on the size of the prediction interval and the probability distribution. By analyzing the prediction accuracy of the model and the accuracy of capturing defective products (recall rate), a probability threshold for determining whether active sampling is needed is determined. Finally, this functionality is integrated into the manufacturing execution system of a steel factory to implement a mechanical performance inspection scheme based on active sampling strategy. The main contributions of this paper are as follows:

- Based on the improved LUBE interval prediction ensemble algorithm, a prediction model for mechanical performance failure probability in steel products is developed by incorporating interval center bias correction.

- To establish an intuitive and easy-to-implement method for balancing model prediction accuracy and the accuracy of capturing defective products, the threshold for the unqualified rate can be adjusted, thereby improving the operability of the solution.
- Designing an active sampling inspection strategy based on the prediction model and integrating the developed model into the factory's manufacturing execution system can significantly reduce the sampling volume while achieving an excellent recall rate.

The organization of this paper is as follows: Sect. 2 presents a review of the related research work on the proposed solution. Section 3 describes the model methodology framework and the associated steps. Section 4 presents a case study with real production data. Section 5 discusses the application of the algorithm in an industrial production scenario. Section 6 concludes the paper.

2 Related Work

2.1 Mechanical Property Prediction

Hot-rolled strip steel is a common metal sheet product. The steel billet, as the raw material, needs to undergo processes such as heating, rolling, and cooling to eventually be coiled into a cylindrical strip steel product. These strip steels need to be tested for dimensions, mechanical properties, surface quality and so on, to ensure that the products meet standards and customer requirements. Then, the traditional method for testing mechanical properties usually requires destructive testing, which involves applying force or strain on the specimen to measure performance indicators such as fracture point and yield point. This method requires a large number of samples, expensive experimental equipment, and has a long testing cycle and high costs. To address this issue, a mechanical property prediction method can be employed. This method utilizes existing material data and features and leverages mathematical models, statistics, or machine learning algorithms to establish a prediction model for estimating the mechanical properties of hot-rolled strip steel. Through this approach, the mechanical properties of the material can be rapidly and accurately predicted without the need for destructive testing. This enables quality control and optimization of the production process.

The relationship between the chemical composition and process parameters of hot-rolled strip steel is highly complex [17]. Despite decades of research on its mechanisms by many scholars, our understanding of the internal mechanisms of steel is still not sufficient [18]. In 1998, Shiv et al. [19] first proposed using neural network algorithms to predict the yield strength and tensile strength

of steel. However, due to the black-box nature of neural networks, which lacks explanation and guarantees for prediction results, their widespread application in practical engineering has been hindered to some extent. In 2000, Yang et al. [20] designed an integrated neural network model to predict the mechanical properties of steel materials along with confidence intervals, which partially alleviated users' doubts about the predictions. In recent years, with the emergence of various new and more adaptive artificial intelligence algorithms, as well as the progress of digital transformation and storage technology in the steel industry, the storage of production data for hot-rolled strip steel has become more standardized, complete and secure. This has significantly reduced the difficulty of obtaining data and provided researchers at home and abroad with a large amount of production data. They have conducted extensive research on the prediction of the mechanical properties of hot-rolled strip steel using artificial intelligence algorithms [21–23]. Existing experimental results have shown that modeling schemes using machine learning algorithms such as random forests [24], XGBoost [25], support vector machines [26], and logistic regression [27], as well as complex modeling schemes using deep learning algorithms such as multilayer perceptron neural networks [29], convolutional neural networks [30], and deep belief networks [31], have achieved good prediction results [28, 32]. These prediction models have the potential to improve production efficiency and optimize quality control in the steel industry, providing valuable references for engineering design and decision-making.

However, in engineering applications, providing only point predictions is not sufficient, and the lack of uncertainty estimation hampers the practical application of advanced models. By considering the uncertainty of the model, steel companies can develop more effective inspection strategies. For example, they can determine the batches of hot-rolled strip steel that need to be sampled and inspected based on the uncertainty of the prediction results. For batches with high uncertainty, stricter inspection standards or increased sampling quantities can be applied to ensure product quality. Conversely, for batches with low uncertainty in the prediction results, more flexible inspection strategies can be adopted to improve production efficiency. Therefore, in the context of intelligent manufacturing, prediction algorithms based on uncertainty estimation and advanced inspection strategies are crucial for steel companies to achieve quality improvement and efficiency enhancement. Such approaches can help companies better utilize the predictive capabilities of advanced models and optimize and make decisions based on uncertainty, thereby improving production quality, reducing costs and increasing efficiency.

2.2 Model Prediction Uncertainty

Over the past decade, machine learning methods have penetrated nearly every scientific field and become a key part of a variety of real-world applications [33–36]. As the spread of machine learning methods continues, the confidence of their predictions becomes increasingly important. However, basic machine learning models cannot provide deterministic estimates and cannot tolerate excessively high or low confidence levels. To address this issue, many researchers have focused on understanding and quantifying the uncertainty in machine learning model predictions [37]. They have proposed various methods to quantify the uncertainty of predictions, including interval prediction methods [38]. By using interval prediction, a better understanding and communication of the uncertainty in prediction results can be achieved, providing more comprehensive information support for decision-making.

Currently, there are many different interval prediction methods available, each with different characteristics and applicability in various application domains and problems. The Delta method, proposed by Hwang and Ding, is a method that constructs intervals through nonlinear regression [39]. It is based on a Taylor series expansion of the regression function and linearizes the neural network model by minimizing a cost function based on errors and the sum of square errors to optimize a set of parameters. The standard asymptotic theory is then applied to the constructed neural network model to form the prediction interval. Delta uses a fixed target variance to form the prediction interval, while the mean-variance estimation method (MVEM) uses two neural networks for the mean and variance parts, respectively [40]. Therefore, MVEM is simpler to implement and does not require computationally expensive derivatives or matrices (such as Hessian or Jacobian matrices) when constructing the prediction interval. However, this method assumes that the mean part of the neural network can accurately estimate the true average of the target, which may not be accurate in practice, leading to prediction intervals that are higher or lower than the predefined confidence, resulting in lower interval coverage rates [41]. Bayesian methods are probabilistic models where all weight parameters are represented as probability distributions rather than having a fixed value in Bayesian neural networks [42]. Compared to non-Bayesian neural networks, Bayesian neural networks are generally more challenging to implement and have slower training computation speeds. Therefore, researchers have introduced the MC-Dropout method, which uses Dropout as a regularization term to compute prediction uncertainty [43]. However, this method is highly influenced by the accuracy of the assumed distribution of parameters. The Bootstrap method is a resampling technique, also known as an ensemble method, that constructs high-quality prediction

intervals by integrating multiple neural networks [44]. The ensemble intuitively represents the model's uncertainty in predictions by evaluating the diversity of member predictions. The LUBE method, proposed by A. Khosravi [45], is a method that directly predicts the upper and lower bounds of the interval. It trains the neural network by constructing a new cost loss function that does not assume point predictions as the interval center and is applicable to various real-world problems. Compared to other methods, the LUBE method can provide more intelligent prediction intervals, particularly demonstrating advantages in quantifying uncertainty in the prediction of mechanical properties of hot-rolled strip steel. It can more accurately describe the range of prediction results and help address uncertainty issues in practical applications.

2.3 LUBE Method

The basic idea of the LUBE method is to predict the upper and lower bounds of the target variable by training the model, rather than a single point estimate [45]. In LUBE, the output of the neural network consists of an upper limit and a lower limit, and the objective of the training is to minimize the range of the output limits while ensuring that the true value falls within that range. By introducing the concept of interval prediction, LUBE is able to provide more comprehensive prediction results. This is particularly useful for dealing with uncertainty issues, especially in industrial manufacturing and other fields where accurate prediction of physical properties and assessment of risks are crucial. Figure 1 illustrates the basic principle of LUBE.

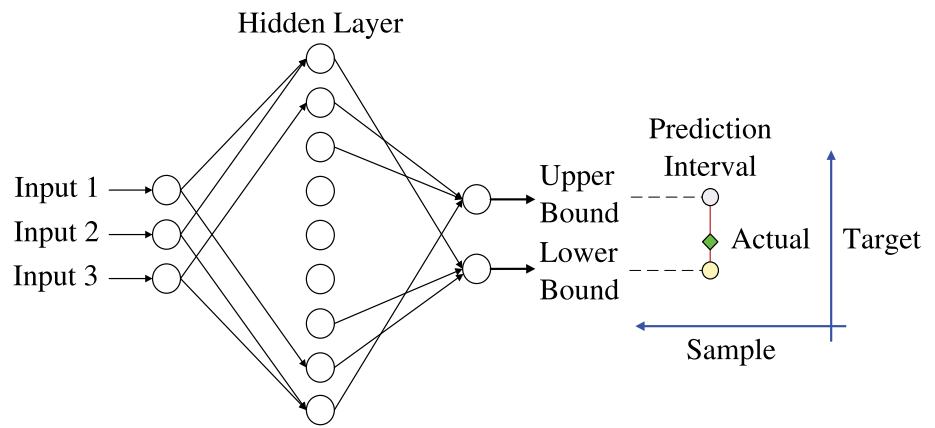
In LUBE, the quality of the prediction interval should be evaluated from two perspectives: reliability and effectiveness. The reliability is assessed by the prediction interval cover probability (PICP), which measures the probability that the true value falls within the predicted interval and is not less than the given confidence level $(1 - \alpha)\%$. The calculation of PICP is as follows:

$$PICP = \frac{1}{n} \sum_{i=1}^n k_i$$

$$k_i = \text{sigmoid}(s \cdot (y_i - \hat{y}_i^L)) \cdot \text{sigmoid}(s \cdot (\hat{y}_i^U - y_i))$$

where n represents the total number of samples, $s > 0$ is a weighting coefficient, \hat{y}_i^L and \hat{y}_i^U denote the lower and upper bounds of the predicted interval for the i -th sample, and y_i refers to the true value of the i -th sample. Indeed, overly wide prediction intervals provide no valuable information as they do not convey any insights about the variability of the target. Therefore, there is a need for another approach to quantify the width of the prediction interval: the mean prediction interval width (MPIW), defined as follows:

Fig. 1 The LUBE method predicts upper and lower bounds of the interval



$$MPIW = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i^U - \hat{y}_i^L)$$

If the target width is known, the MPIW can be standardized to facilitate objective comparison of prediction intervals developed using different methods. The normalized MPIW (NMPIW) is defined as follows:

$$NMPIW = \frac{MPIW}{r}$$

Where r represents the difference between the maximum and minimum values of the true values in the sample set. Based on the coverage width, we can derive the coverage width-based criterion (CWC):

$$CWC = NMPIW(1 + e^{-\eta(PICP-\mu)})$$

where the constants η and μ are two hyperparameters that determine the degree of penalty for prediction intervals with lower coverage probabilities. μ corresponds to the desired confidence level associated with the prediction interval and can be set as $1 - \alpha$. The role of η is to amplify any minor differences between the PICP and μ .

For samples where the prediction interval does not contain their true values, further reducing the width of the prediction interval for such samples does not affect the value of PICP. Therefore, when computing the loss function, the width of the prediction interval for samples where the true value is not included should not be further decreased. Pearce et al. [46] improved upon NMPIW and introduced the prediction interval normalized average width (PINAW), which optimizes the average interval width only for samples where the prediction interval contains the true value.

$$PINAW = \frac{1}{nr} \sum_{i=1}^n (\hat{y}_i^U - \hat{y}_i^L) \cdot k_i$$

L. Cheng et al. [47] introduced the consideration of whether the true value is at the center of the upper and lower bounds into a new penalty term, ensuring a more evenly distributed prediction interval around the true value. Therefore, in addition to considering PICP and PINAW, the quality of the center of the prediction interval is incorporated. The mean prediction interval center deviation (MPICD) is defined as follows:

$$MPICD = \frac{1}{n} \sum_{i=1}^n \left(\frac{\hat{y}_i^U + \hat{y}_i^L}{2} - y_i \right)^2$$

3 Model and Method

The proposed proactive quality inspection method for strip steel’s mechanical performance mainly includes interval prediction of mechanical performance and prediction of nonconformity probability. In this section, we first introduce how to improve the LUBE algorithm and construct an ensemble learning model for interval prediction of mechanical performance based on the improved LUBE algorithm. Then, we describe how to derive the nonconformity probability of mechanical performance using the interval prediction results, and ultimately establish an active sampling strategy.

3.1 The Improved Loss Function of LUBE

Due to the repetitive nature of manufacturing datasets, most of the samples have small center deviations in their prediction intervals. As a result, the contribution of the MPICD in the loss function is relatively small and may not effectively regulate the prediction interval distribution. Therefore, it is necessary to introduce a gradient weight in the calculation of the loss to increase the contribution of samples where

the true value is closer to the upper and lower bounds of the prediction interval. This adjustment can help improve the distribution of prediction intervals for the samples.

Based on the above analysis, this paper considers introducing the cosine function as the gradient weight in the calculation of the loss function. The original interval of the mechanical performance values in the test set is mapped to the corresponding interval of the cosine function, i.e., $[y_{\min}, y_{\max}]$ is mapped to $[0, \pi]$. The calculation method for the weight coefficient is as follows:

$$\psi_i = \cos^2 \left(\frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \cdot \pi \right)$$

where y_i represents the true mechanical performance of the i -th sample, y_{\max} and y_{\min} represent the maximum and minimum values of the mechanical performance in the sample set, respectively. Squaring is performed to avoid negative values for the weight coefficient. To simultaneously control the width and center deviation of the prediction intervals for samples where the tensile strength is within the standard range, the weight coefficient is introduced into both parts of the loss function. The calculation method is as follows:

$$PINAW_{\text{new}} = \frac{1}{n} \sum_{i=1}^n \psi_i (\hat{y}_{iU} - \hat{y}_{iL}) k_i$$

$$MPICD_{\text{new}} = \frac{1}{n} \sum_{i=1}^n \psi_i \left(\frac{\hat{y}_{iU} + \hat{y}_{iL}}{2} - y_i \right)^2$$

Then, a new form of CWC is derived:

$$CWC_{\text{new}} = (PINAW_{\text{new}} + MPICD_{\text{new}}) \cdot (1 + e^{-\eta \max(0, (1-\alpha) - PICP)^2})$$

3.2 Integrated Prediction Interval Model

Although the LUBE interval prediction method can provide prediction intervals for the mechanical performance of

samples, it only considers the uncertainty introduced by the data and does not account for the uncertainty of the model itself. The uncertainty of the model can be viewed from two aspects:

(1) Imbalanced sample label distribution: In most industrial manufacturing datasets, the distribution of label values in the samples is often imbalanced. This means that the proportion of samples with extreme label values is very small, leading to a bias in the model's training towards samples with larger proportions, resulting in biased predictions for different data.

(2) Different initialization parameters of the model: Due to the randomness of the initialization parameters, models trained with different initializations exhibit a certain level of uncertainty in predictions. This means that the same input may lead to different prediction results because the model's interpretation of the data may differ in its initial state.

To address these uncertainties of the model, this paper adopts an ensemble method, training multiple neural networks with different initialization parameters through parameter resampling. During the training process of the sub-models, subsampling can be performed from the training set to obtain relatively balanced sub-datasets and train neural networks with each sub-dataset separately. By training multiple models, we can obtain an ensemble of neural networks that captures the diversity of the training data and mitigates the bias issue to some extent. The variance of the predictions from these neural networks can be used as an estimate of the model's uncertainty. The structure of the ensemble model is illustrated in Fig. 2. Each sub-model is an independently trained neural network with different initialization parameters. During testing, the predictions from all sub-models can be aggregated, for example, by taking the average or calculating the standard deviation, to obtain the final prediction and estimate of the model's uncertainty. The implementation of the ensemble method involves the following steps:

Step 1: Interval ensemble strategy

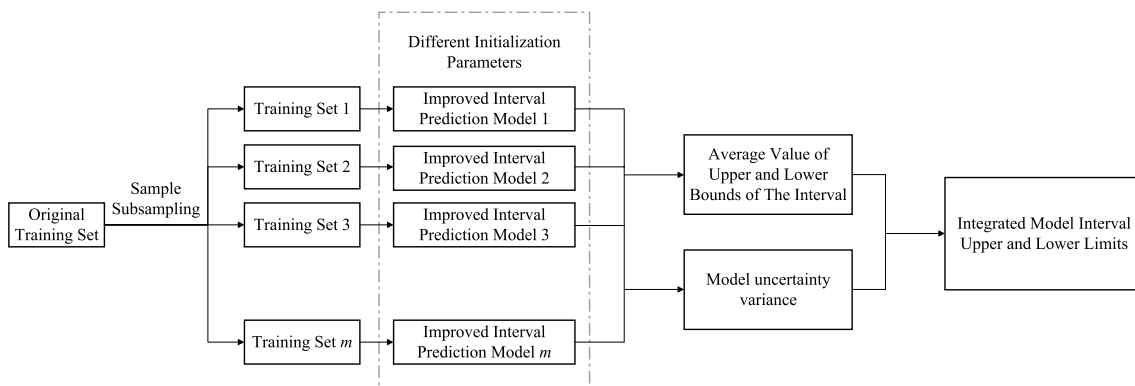


Fig. 2 The integrated model structure diagram

In the ensemble learning method using LUBE as sub-models, each sub-model's predictions consist of upper and lower interval bounds. To integrate the intervals, we adopt the strategy proposed by Pearce et al. [46] to construct an ensemble of m neural networks using the loss function as the criterion. Let \hat{y}^U and \hat{y}^L represent the upper and lower bounds of the prediction interval, respectively. The calculation of the variance for model uncertainty is as follows:

$$\bar{y}_i^U = \frac{1}{m} \sum_{j=1}^m \hat{y}_{ij}^U$$

$$\sigma_i^U = \sqrt{\frac{1}{m-1} \sum_{j=1}^m (\hat{y}_{ij}^U - \bar{y}_i^U)^2}$$

where \hat{y}_{ij}^U represents the upper prediction bound of the i -th predicted sample on the j -th neural network, and the calculation method for \bar{y}_i^L and σ_i^L is the same. According to the equation for confidence intervals in normal distribution, the upper and lower bounds of the ensemble model's interval are calculated as follows:

$$\hat{y}_i^U = \bar{y}_i^U + z \cdot (\sigma_i^U)^2$$

$$\hat{y}_i^L = \bar{y}_i^L - z \cdot (\sigma_i^L)^2$$

where z represents the quantile of the interval, which depends on the specified confidence level. When the preset confidence level is 95%, z is equal to 1.96.

Step 2: Prediction result integration strategy and uncertainty

Since the uncertainty of the model includes the uncertainty of the training data distribution and the uncertainty of the model parameters, it is necessary to randomly initialize the parameters of each model when building the prediction model. At the same time, the ensemble model is constructed by sub-sampling the training set. The sub-sampling of the training set is performed using the bootstrap method. The training steps are as follows:

- Randomly select n samples with replacement from the training sample set in a uniform manner to create a sub-dataset with a relatively balanced label distribution. Then, establish a sub-model using this sub-dataset and randomly initialize the initial parameters of the sub-model.
- Repeat the previous step for m times and train all models in parallel, resulting in m prediction models trained with different training samples.
- Use each model trained in the previous step to predict the test set samples separately. Take the average of

the m prediction results as the final result, and use the variance as the model uncertainty.

3.3 Prediction of The Probability of Mechanical Performance Failure

Assuming that the mechanical performance of the steel samples within the prediction interval follows a normal distribution, it can be denoted as $X \sim \mathcal{N}(\mu, \sigma^2)$. The probability density function is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

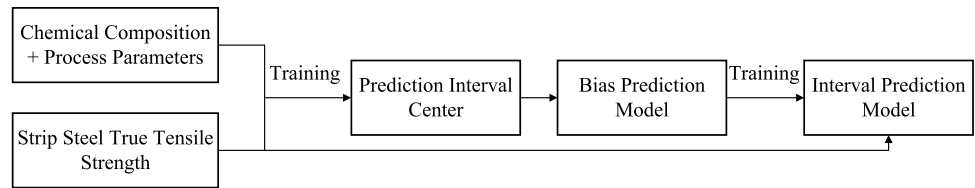
where x represents the mechanical performance, μ represents the center value of the prediction interval, and σ^2 is the unknown variance. To predict the probability of mechanical performance failure for a given sample, we need to determine the probability density function of the mechanical performance within the prediction interval. Based on the specifications and requirements provided by metallurgical standards and manufacturers, the standard range of tensile strength $[a, b]$ for the tested steel grade is determined, and the probability of the sample's mechanical performance being non-compliant is calculated. The implementation steps are as follows:

Step 1: Correcting the prediction interval center value.

Using the ensemble learning model mentioned above, we obtain the upper bound \hat{y}_U and lower bound \hat{y}_L of the mechanical performance for the sample, as well as the center value of the prediction interval. However, in the process of collecting actual steel data, there may be an imbalance in the distribution of samples in the training set. The number of samples with tensile strength below the lower limit of the standard range is relatively small, causing the model to overestimate the mechanical performance of those samples with extremely low values. Then, the center value of the prediction interval is higher than the true value, leading to a large deviation. As the tensile strength increases and the number of samples increases, the model's predictions gradually approach the true value, resulting in a smaller deviation that tends to zero. Additionally, the number of samples with tensile strength above the upper limit of the standard range is also relatively small, causing the model to underestimate the mechanical performance of those samples with extremely high values. This leads to the center value of the prediction interval being lower than the true value, with an increasing deviation in the negative direction.

Based on the above analysis, it can be concluded that there is a certain correlation between the center value of the prediction interval and the true mechanical performance of the samples. Therefore, this paper considers using a regression

Fig. 3 The regression chain model



chain model, as shown in Fig. 3, to treat the prediction interval center and the true mechanical performance of the samples as multiple targets for regression analysis. By fitting the deviation of the prediction interval center for the training set samples, we can predict the new interval center for the test set samples.

Step 2: Calculate the variance of the normal distribution.

Based on Step 1, we can obtain the mean μ in Eq. (1), but the variance is unknown. Since the prediction interval of the sample is symmetric about the center value, the original normal distribution can be transformed into a standard normal distribution.

$$X \sim \mathcal{N}(\mu, \sigma^2) \rightarrow Y = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1) \tag{2}$$

During the training of the ensemble learning model, a nominal confidence level is set, which allows us to determine the proportion of the prediction interval of the sample in the entire normal distribution interval, i.e., $\Pr(\hat{y}_L < X < \hat{y}_U)$ is known. Then, based on Eq. (2) and the probability table of the standard normal distribution, we can calculate the mechanical performance of the sample, i.e., $\frac{\hat{y}_U - \mu}{\sigma}$. Based on the mean μ and the upper bound of the interval \hat{y}_U , we can obtain the standard deviation σ corresponding to the standard normal distribution.

Step 3: Calculate the probability of mechanical performance failure for the sample.

After obtaining the mean and standard deviation of the mechanical performance normal distribution interval for the sample, we can derive the corresponding probability distribution function $f(x)$. Based on the specifications and requirements provided by the metallurgical standards and the manufacturer, the standard range of tensile strength $[a, b]$ for the tested steel grade is determined. Using the equation for calculating the probability of the normal distribution, we can calculate the probability of the mechanical performance of the sample being non-compliant as follows:

$$\Pr(a < X < b) = \int_a^b f(x)dx$$

In summary, the process for predicting the probability of mechanical performance failure for the steel sample is shown in Fig. 4.

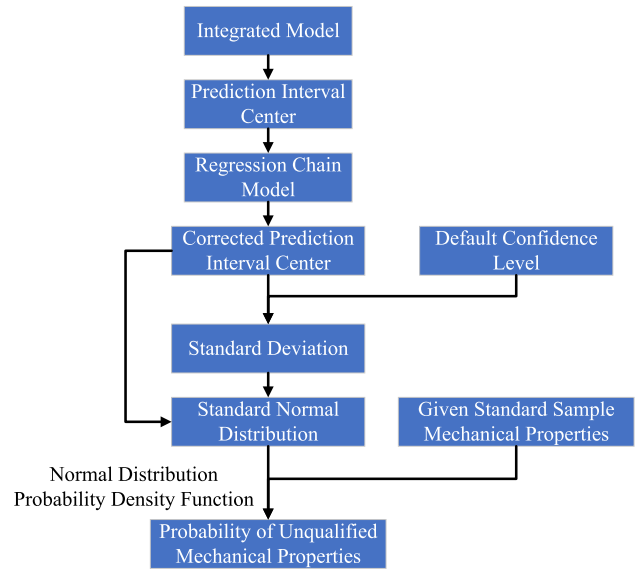


Fig. 4 The process for predicting the probability of mechanical performance failure for the steel sample

3.4 Active Sampling Inspection Strategy for Mechanical Properties

Based on Sect. 3.3, the probability of mechanical performance failure for the sample is calculated. Considering a set threshold for the probability of failure, each sample is evaluated to determine whether the calculated failure probability exceeds the threshold. If it exceeds the threshold, the sample is considered to have failed in terms of mechanical performance and requires proactive sampling inspection. If it does not exceed the threshold, the sample is considered to have passed the mechanical performance requirements and no further sampling inspection is necessary.

Table 1 The binary classification confusion matrix

	Predict positive class	Predict negative class
True positive class	TP	FN
True negative class	FP	TN

In this study, we incorporate the evaluation methods of recall and precision in a classification model to construct a corresponding confusion matrix by setting different probability thresholds. Table 1 represents the binary classification confusion matrix, where the true positive (TP) indicates the number of positive instances correctly predicted as positive, the false positive (FP) represents the number of negative instances incorrectly predicted as positive, the true negative (TN) represents the number of negative instances correctly predicted as negative, and the false negative (FN) represents the number of positive instances incorrectly predicted as negative.

To select the optimal threshold, we consider various metrics such as recall, precision, and sampling rate. Recall is defined as $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$, and precision is defined as $\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$. In the steel judging process, it is crucial to correctly identify all samples with non-compliant tensile strength. Therefore, it is important to prioritize achieving a high recall rate. However, it is also necessary to ensure that the sampling rate does not exceed the original 10% sampling rate in order to avoid significant cost increases. The algorithm implementation is shown in Algorithm 1.

Algorithm 1 Probability Threshold Optimization

Input: Predict the probability list of failure, the true label list, the threshold list
 Initialize optimal threshold
 Iterate through the threshold list

1. Compare each predicted probability of failure with the threshold, calculate the sampling rate
2. Combine with the true labels to obtain the confusion matrix
3. Calculate precision and recall

End of iteration
 Plot precision, recall, and sampling rate against threshold values
 Based on industrial requirements, determine the optimal threshold

4 Experiments and Results

4.1 Dataset Introduction

The proposed method is experimentally tested and analyzed using the QSTE420TM steel grade as an example. Table 2 is the statistics of relevant information of the dataset, the specified range for the tensile strength performance of this product is 620 MPa (upper limit) and 480 MPa (lower limit). The total number of samples in the dataset is 2124. A test set of 300 samples is obtained through random sampling, while the remaining 1824 samples are used as the training set. In addition, to make the LUBE model have good accuracy and prediction stability, while taking into account the computational efficiency of the model to a certain extent, we set the number of sub-training set samples used in each sub-model training to 300.

4.2 Model Structure and Parameters

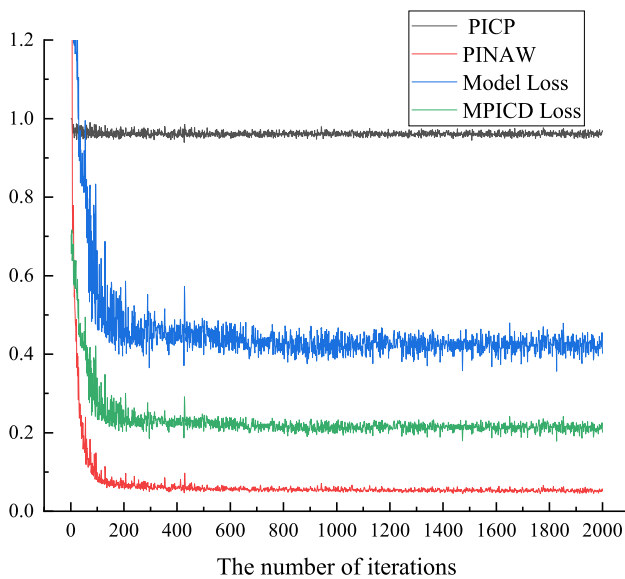
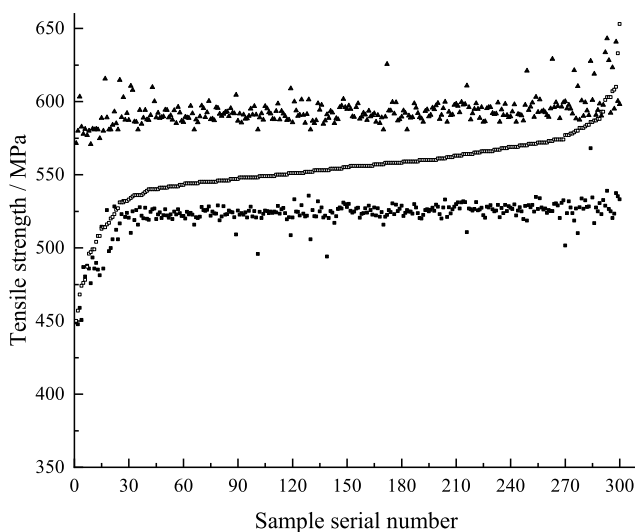
The improved LUBE, serving as the basic unit of the ensemble model, has a neural network structure with the number

Table 2 QSTE420TM Data Details

Variable name	Maximum value	Minimum value	Average value
C	0.0877	0.0579	0.0704
Si	0.0027	0.0412	0.0105
Mn	0.5859	0.4432	0.4956
P	0.0237	0.0042	0.0124
S	0.0109	0.0016	0.0050
Nb	0.0055	0.0005	0.0028
V	0.0066	0.0012	0.0029
Ti	0.063	0.0394	0.0499
Oven temperature	1256.8961	1198.5366	1230.4045
Final rolling temperature	864.19	902.892	882.5862
Coiling temperature	651.443	589.562	620.7414
Finished product thickness	10.003	2.804	6.9302
Tensile Strength	595	452	510.8798

Table 3 Neural Network Structure

Neural network layer	Number of neurons	Activation function
Input layer	15	
Hidden layer 1	64	ReLU
Hidden layer 2	32	ReLU
Hidden layer 3	8	ReLU
Output layer	2	

**Fig. 5** Changes in parameters of the LUBE model training process after the loss function is improved

of layers, number of neurons per layer, and activation functions as shown in Table 3. To prevent overfitting, reduce redundancy in the neural network and enhance the model's robustness, Dropout regularization is incorporated at each layer's connections with a probability of 0.75.

During model initialization, the maximum iteration count is set to 2000. Considering the number of samples in the training set, the batch size is set to 256 for improved training efficiency. To achieve more stable convergence in the later stages of training, the 'LearningRateScheduler' is used to dynamically adjust the learning rate of the model. The initial learning rate is set to 0.01, and the decay factor is set to 0.5. For hyperparameters in the loss function, appropriate selection is necessary to balance the accuracy and width of the prediction interval. In interval prediction, the nominal confidence level $1 - \alpha$ represents the desired interval coverage probability, indicating the probability of the true value being within the predicted interval is $1 - \alpha$. A common choice is setting α to 0.05, as it strikes a balance between interval width and prediction accuracy. Regarding hyperparameters s and λ in the model's loss function, this paper adopts a grid search approach to determine their optimal values.

4.3 Comparison of Improvement Effects of Loss Function

After improving the loss function of the LUBE model, the changes in various parameters during the model training process are shown in Fig. 5. It can be observed that as the number of iterations increases, the parameters gradually converge. This indicates that the model gradually learns the

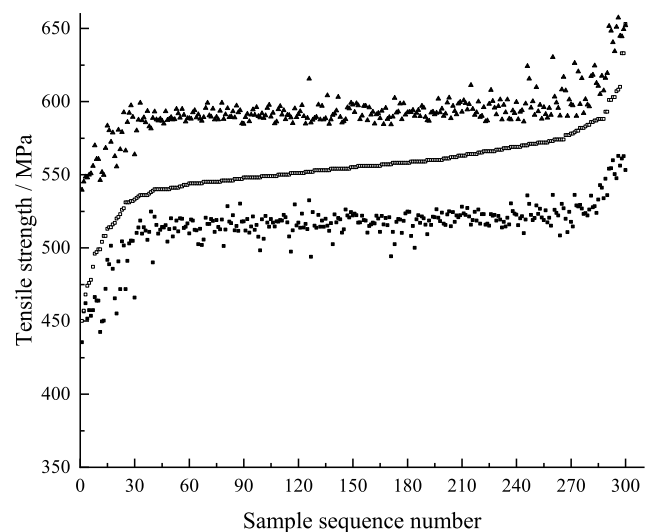
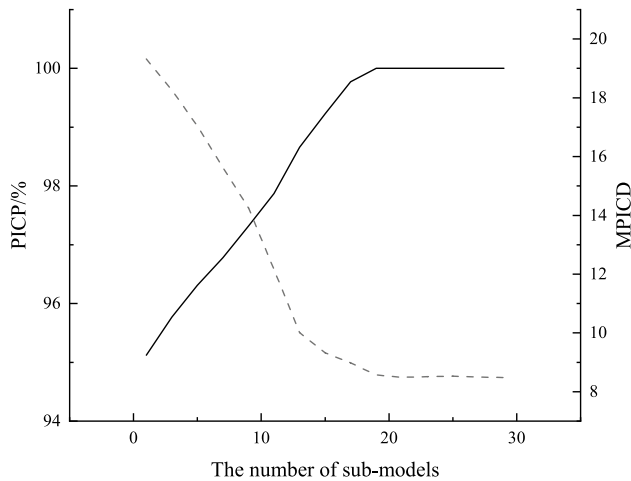
**Fig. 6** Prediction and true interval value of tensile strength of test set samples before (left) and after (right) improvement

Table 4 Neural network structure

Model	PICP	PINAW	MPICD
Original model	95.56%	0.262	11.43
Improved model	97.32%	0.251	9.12

**Fig. 7** The number of integrated model sub-models and model results**Table 5** Comparison of evaluation indicators between the integrated model and the single model

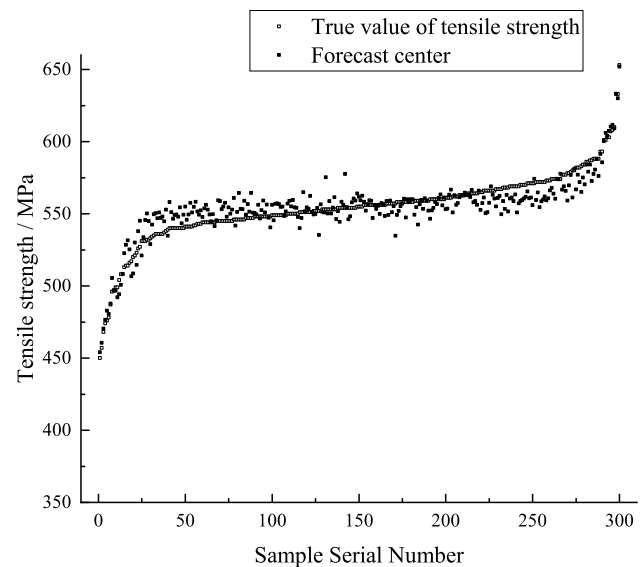
Model	PICP	PINAW	MPICD
Single model	97.32%	0.251	9.12
Integrated model	100%	0.259	8.48

optimal parameter configuration during the training process to minimize the loss function.

The comparison of the prediction performance on the test samples between the models trained with the original loss function and the new loss function is shown in Fig. 6. Compared with the original loss function, this paper introduces the NMPICD loss term and obtains the PICP of 96.33%, the PINAW of 0.249, and the NMPICD of 10.06. These improvements indicate that the predicted interval centers are closer to the true values of the mechanical performance. Meanwhile, from Table 4, it can be observed that compared to the pre-improvement model, there are significant improvements in the test set's the PICP, the PINAW and the NMPICD.

4.4 The Effect of the Integrated Approach

After optimizing the loss function for the LUBE model, multiple sub-models are trained using different initialization parameters. Figure 7 illustrates the PICP and the MPICD for

**Fig. 8** Interval center distribution of errors

different numbers of sub-models in the ensemble model. It can be seen that as the number of sub-models increases, the PICP of the ensemble model gradually increases and reaches 100%, while the MPICD decreases. When the number of sub-models exceeds 20, the model achieves a 100% PICP and the lowest MPICD. Therefore, this study selects 20 sub-models for the ensemble model.

After constructing the ensemble interval prediction model, the model predicts the test set samples, and the results are shown in Table 5. Compared to a single model, the ensemble model achieves a 100% PICP without significantly increasing the PINAW. Additionally, it reduces the PINAW, resulting in higher precision of the predicted intervals for the samples.

4.5 Failure Probability Prediction

(1) Interval center correction based on regression chain

The interval center is corrected using a regression chain model, and the results are shown in Fig. 8. It can be observed that by fitting the MPICD, the original prediction interval center is adjusted by adding the MPICD, resulting in a new interval center that closely aligns with the true value of the sample's tensile strength.

Samples with tensile strength located in the middle of the standard range, specifically with true tensile strength between 510 and 590 MPa, do not exhibit ideal results when using a cubic polynomial to fit the deviation of the prediction interval center values. This is also evident in Fig. 8, where these samples show some degree of prediction interval center deviation. However, since the true tensile strength of these samples is significantly distant from the upper and

Table 6 Model comparison

	MPICD (MPa)
Ensemble interval prediction model	16.91
Regression chain prediction model	3.46

lower limits of the standard range, the presence of some prediction interval center deviation and the resulting distribution bias in the prediction intervals would not affect the steel classification.

The prediction interval center deviations before and after statistically fitting the deviation of the prediction interval center values for samples with true tensile strength below 510 MPa and above 580 MPa are presented in Table 6. After applying the regression chain to correct the prediction interval center, significant improvements are observed for samples with extreme true tensile strength values. The interval center deviation is reduced from 16.91 to 3.46 MPa, indicating the remarkable effectiveness of the interval center correction method based on regression chain.

(2) Failure probability calculation

Figure 9 illustrates the distribution of the prediction interval after the correction for a specific sample. The original prediction interval center value is 520.11 MPa, while the corrected prediction interval center value is 508.27 MPa. The upper and lower limits of the interval are 568.21 MPa and 472.01 MPa, respectively. Based on the principle of a nominal confidence level of 95%, the normal distribution

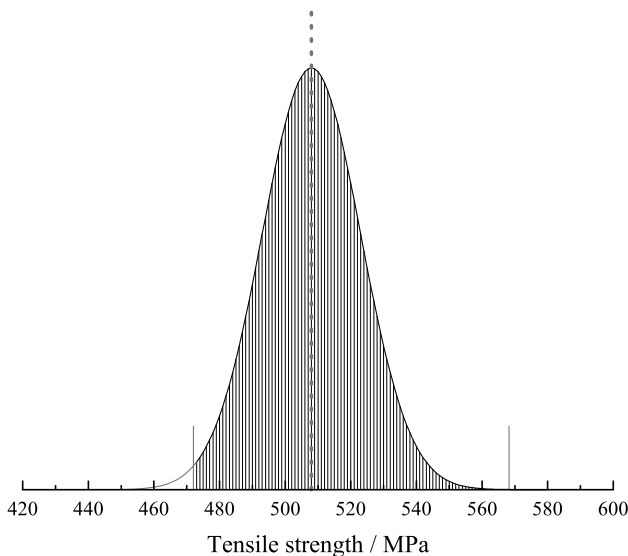


Fig. 9 Normal distribution of sample interval center and tensile strength

parameters for this sample’s prediction result are $\mu = 508.27$. Additionally, there is a 95% probability that the predicted result falls within the range of 472.01 MPa to 568.21 MPa.

Given that the normal distribution mean μ is 508.27 MPa, the probability of the interval [472.01MPa, 568.21MPa] under this normal distribution is 95%. Consequently, the mechanical performance interval of this sample can be considered to follow a normal distribution with a mean of 508.27 MPa and a standard deviation of 19.78. By calculation, the probability of the sample’s tensile strength being below 480 MPa or above 620 MPa is:

$$\Pr(X < 480 || X > 620) = 0.035$$

This implies that the probability of the sample’s mechanical performance not meeting the specifications is 3.5%.

4.6 Active Sampling Threshold Calculation

In determining the probability threshold, the evaluation method of recall and precision is used in the classification model to select the optimal threshold. In this context, positive class samples represent actual samples with non-compliant mechanical performance, while negative class samples represent actual samples with compliant mechanical performance. Predicted positive class represents samples with predicted non-compliant performance probability greater than the probability threshold, while predicted negative class represents samples with predicted non-compliant performance probability less than the probability threshold. The precision and recall at different probability thresholds are illustrated in Fig. 10.

In Fig. 10, the x-axis represents the probability threshold ranging from 30% to 90%, while the y-axis represents

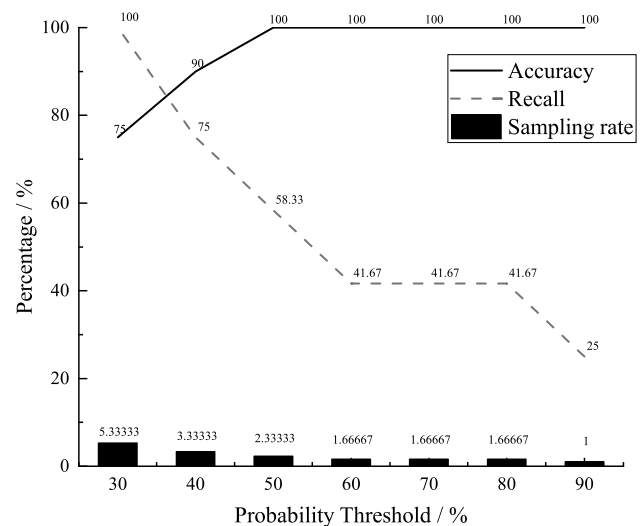


Fig. 10 Precision and recall under different probability thresholds

the percentages of precision, recall, and sampling rate. The solid line represents the variation of precision with increasing probability threshold, the dashed line represents the variation of recall with increasing probability threshold and the bar chart represents the variation of sampling rate with increasing probability threshold.

In the steel classification process, it is crucial to prioritize correctly identifying all non-compliant samples with regards to tensile strength. Therefore, a high recall rate should be ensured. Then, it is essential to maintain a sampling rate no higher than the original 10% sampling rate to avoid significant cost increases. From Fig. 10, it can be observed that to achieve a high recall rate, the probability threshold should be set as small as possible. In the test dataset, there are a total of 12 non-compliant samples with regards to tensile strength. When the probability threshold is set to 30%, the recall rate reaches 100%, indicating that all non-compliant samples are correctly identified without any false negatives. Moreover, the number of samples taken is 16 coils, resulting in a sampling rate

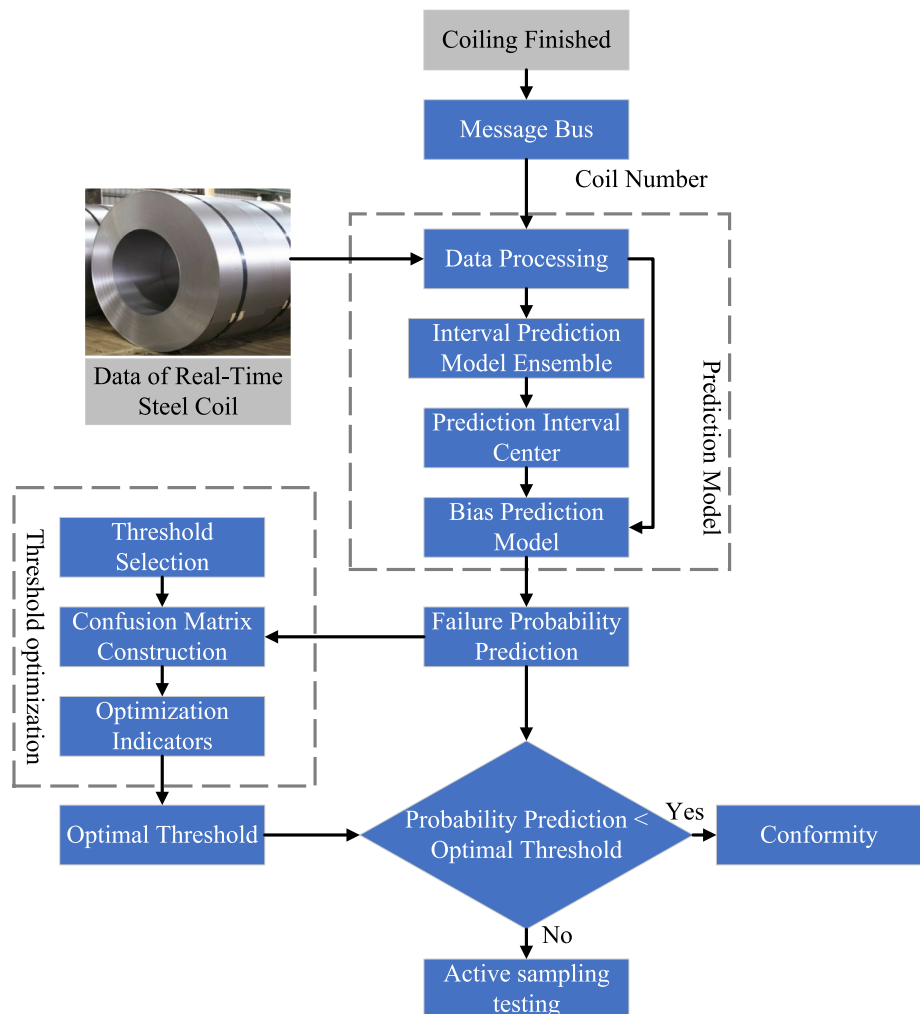
of only 5.33%, which is lower than the original sampling rate.

Based on the analysis above, setting the probability threshold to 30% is recommended. When a sample's non-compliant performance probability exceeds the threshold, it is classified as a non-compliant sample. By setting the probability threshold to 30%, all non-compliant samples with regards to tensile strength can be identified to avoid false negatives, while reducing the number of samples taken and minimizing costs.

5 Industrial Applications

The mechanical performance active sampling detection model proposed in this paper has been implemented in the manufacturing execution system of a large steel mill. Figure 11 illustrates the logical diagram of the mechanical performance active sampling detection module within the platform. This module acquires real-time data from the

Fig. 11 Logic diagram of the active sampling and testing module for mechanical properties under the platform



message bus regarding the completion of coil rolling. Subsequently, the data processing stage retrieves real-time data on chemical composition, production process, and historical coil rolling data, which are then processed into the required data structure for the model. These data are then fed into the model for local modeling and prediction, yielding the probability of non-compliant mechanical performance. When the predicted probability exceeds the set probability threshold, the detection department initiates active sampling testing. For qualified samples, the sampling process will no longer be needed, and the results of performance predictions can be directly used to judge steel, thus greatly reducing the frequency of physical sampling and improving the efficiency of product inspection.

6 Conclusions

This study proposes a data-driven proactive inspection scheme for steel product quality, addressing the inefficiency of traditional quality inspection strategies. By constructing an integrated model for predicting the mechanical performance interval and quantifying prediction uncertainty, accurate predictions of the probability of non-compliance are achieved. Compared to the traditional inspection rules used in actual production, this scheme reduces the sampling rate by 50%, resulting in a sampling rate of only 5.33% in practical applications, while ensuring controlled risks of false negatives. These results demonstrate that our proactive sampling strategy achieves satisfactory quality inspection outcomes in practical applications.

This research significantly improves the efficiency of product quality inspection, reduces production lead time, inventory, and quality disputes, and brings substantial savings in terms of labor costs. The scheme of this paper also provides guidance and inspiration for future research. For example, this scheme can be extended to the determination of multiple quality indicators to achieve a more comprehensive quality assessment. In addition, the model in this paper can also be applied to quality inspection in other fields, such as automobiles, aviation, etc., and has broad application prospects and promotion value.

Acknowledgements Conceptualization, S.Y.; methodology, L.F.; validation, W.Z.; formal analysis, Z.B. and Z.B.. All authors have read and agreed to the published version of the manuscript.

Author Contributions Conceptualization, SY; methodology, LF; validation, WZ; formal analysis, ZB and ZB. All authors have read and agreed to the published version of the manuscript.

Funding Not applicable.

Availability of data and materials Data are provided by a steel mill.

Declarations

Conflict of interest Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Schmitt, J., Bönig, J., Borggräfe, T., Beiting, G., Deuse, J.: Predictive Model-Based Quality Inspection Using Machine Learning and Edge Cloud Computing. *Adv. Eng. Inform.* **45**, 101101 (2020)
- Azamferei, V., Psarommatis, F., Lagrosen, Y.: Application of automation for in-line quality inspection, a zero-defect manufacturing approach. *J. Manuf. Syst.* **67**, 1–22 (2023)
- Belodedenko, S., Hrechanyi, O., Vasilchenko, T., Baiul, K., Hrechana, A.: Development of A Methodology for Mechanical Testing of Steel Samples for Predicting The Durability of Vehicle Wheel Rims. *Results Eng.* **18**, 101117 (2023)
- Caprili, S., Mattei, F., Mazzatura, I., Ferrari, F., Gammino, M., Mariscotti, M., Mori, M., Piscini, A.: Evaluation of mechanical characteristics of steel bars by non-destructive vickers microhardness tests. *Proc. Struct. Integrity* **44**, 886–893 (2023)
- Sarkar, B., Saren, S.: Product inspection policy for an imperfect production system with inspection errors and warranty cost. *Euro. J. Oper. Res.* **248**(1), 263–271 (2016)
- Azamferei, V., Granlund, A., Lagrosen, Y.: Multi-Layer Quality Inspection System Framework for Industry 4.0. *Int. J. Auto. Technol.*, 15(5): 641-650 (2021)
- Bose, D., Guha, A.: Economic Production Lot Sizing under Imperfect Quality, On-Line Inspection, and Inspection Errors: Full vs. Sampling Inspection. *Comput. Ind. Eng.* **160**, 107565 (2021)
- Ait-El-Cadi, A., Gharbi, A., Dhouib, K., Artiba, A.: Integrated Production, Maintenance and Quality Control Policy for Unreliable Manufacturing Systems under Dynamic Inspection. *Int. J. Prod. Econ.* **236**, 108140 (2021)
- Shim, J., Kang, S., Cho, S.: Active inspection for cost-effective fault prediction in manufacturing process. *J. Process Control* **105**, 250–258 (2021)
- Papananias, M., McLeay, T., Obajemu, O., Mahfouf, M., Kadirkamanathan, V.: Inspection by Exception: A New Machine Learning-Based Approach for Multistage Manufacturing. *Applied Soft Computing*, 97(Part A): 106787 (2020)
- Gittler, T., Relea, E., Corti, D., Corani, G., Weiss, L., Cannizaro, D., Wegener, K.: Towards predictive quality management in assembly systems with low quality low quantity data—a methodological approach. *Proc. CIRP* **79**, 125–130 (2019)
- Wang, G., Ledwoch, A., Hasani, R., Grosu, R., Brintrup, A.: A generative neural network model for the quality prediction of work in progress products. *Appl. Soft Comput.* **85**, 105683 (2019)

13. Nguyen, B., Tran, T., Nguyen, T., Nguyen, G.: An improved sea lion optimization for workload elasticity prediction with neural networks. *Int. J. Comput. Intell. Syst.* **15**(90), 1–26 (2022)
14. Struchtrup, A., Kvaktun, D., Schiffers, R.: Adaptive quality prediction in injection molding based on ensemble learning. *Proc. CIRP* **99**, 301–306 (2021)
15. Sacco, M., Ruiz, J., Pulido, M., Tando, P.: Evaluation of machine learning techniques for forecast uncertainty quantification. *Q. J. R. Meteorol. Soc.* **148**(749), 3470–3490 (2022)
16. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. 31st Conference on Neural Information Processing Systems, CA, USA, (2017)
17. Zhu, B., Zhu, J., Zhu, Z., Wang, Y., Zhang, Y.: Effect of rapid heating process in hot stamping on compact strip production hot rolled plate. *Proc. Manuf.* **15**, 1055–1061 (2018)
18. Guo, Z., Sha, W.: Modelling the correlation between processing parameters and properties of maraging steels using artificial neural network. *Comput. Mater. Sci.* **29**(1), 12–28 (2004)
19. Pettersson, F., Chakraborti, N., Singh, S.: Neural networks analysis of steel plate processing augmented by multi-objective genetic algorithms. *Steel Res. Int.* **78**(12), 890–898 (2007)
20. Yang, Y., Linkens, D., Trowsdale, A., Tenner, J.: Ensemble neural network model for steel properties prediction. *IFAC Proc Vol* **33**(22), 401–406 (2000)
21. Saravanakumar, P., Jothimani, V., Sureshbabu, L., Ayyappan, S., Noorullah, D., Venkatakrishnan, P.: Prediction of mechanical properties of low carbon steel in hot rolling process using artificial neural network model. *Proc. Eng.* **38**, 3418–3425 (2012)
22. Sui, X., Lv, Z.: Prediction of The mechanical properties of hot rolling products by using attribute reduction ELM. *Int. J. Adv. Manuf. Technol.* **85**(5–8), 1395–1403 (2016)
23. Xie, Q., Suvarna, M., Li, J., Zhu, X., Cai, J., Wang, X.: Online prediction of mechanical properties of hot rolled steel plate using machine learning. *Mater. Design* **197**, 109201 (2021)
24. Kwak, S., Kim, J., Ding, H., Xu, X., Chen, R., Guo, J., Fu, H.: Machine learning prediction of the mechanical properties of γ -TiAl alloys produced using random forest regression model. *J. Mater. Res. Technol.* **18**, 520–530 (2022)
25. Chen, J., Zhao, F., Sun, Y., Zhang, L., Yin, Y.: Prediction model based on XGBoost for mechanical properties of steel materials. *Int. J. Model. Identification Control* **33**(4), 322 (2019)
26. Wang, L., Mu, Z., Guo, H.: Application of support vector machine in the prediction of mechanical property of steel materials. *J. Univ. Sci. Technol. Beijing, Mineral, Metallurgy, Material* **13**(6), 512–515 (2006)
27. Li, F., Song, Y., Liu, C., Jia, R., Li, B.: Research on error distribution modeling of mechanical performance prediction model for hot rolled strip. *Metallurgical Ind. Auto.* **43**(6), 28–33 (2019)
28. Cheng, T., Chen, G.: Prediction of mechanical properties of hot-rolled strip steel based on PCA-GBDT method. *J. Phys. Conf. Ser.* **1774**(1), 012002 (2021)
29. Zhang, J., Gao, P., Fang, F.: An ATPSO-BP neural network modeling and its application in mechanical property prediction. *Comput. Mater. Sci.* **163**, 262–266 (2019)
30. Li, W., Xie, L., Zhao, Y., Li, Z., Wang, W.: Prediction model for mechanical properties of hot-rolled strips by deep learning. *J. Iron Steel Res. Int.* **27**(9), 1045–1053 (2020)
31. Huang, S., Tian, T.: Prediction of mechanical properties of hot rolled strip based on DBN and composite quantile regression. *Assoc. Comput. Mach.* **110**, 1–6 (2021)
32. Sui, X., Lv, Z., Li, T.: Application of High-dimensional multi-input layers GA neural network in prediction of hot-rolling product's mechanical property. *J. Inform. Comput. Sci.* **12**(3), 1159–1168 (2015)
33. Bhatti, U., Marjan, S., Wahid, A., Syam, M., Huang, M., Tang, H., Hasnain, A.: The effects of socioeconomic factors on particulate matter concentration in China's: new evidence from spatial econometric model. *J. Clean. Prod.* **417**, 137969 (2023)
34. Bhatti, U., Huang, M., Neira-Molina, H., Marjan, S., Baryalai, M., Tang, H., Wu, G., Bazai, S.: MFFCG-Multi Feature Fusion for Hyperspectral Image Classification Using Graph Attention Network. *Expert Systems with Applications*, 229(Part A): 120496 (2023)
35. Çelik, ö.: A Research on Machine Learning Methods and Its Applications. *Journal of Educational Technology and Online Learning*, 3: 25–40 (2018)
36. Zhang, Y., Chen, J., Ma, X., Wang, G., Bhatti, U., Huang, M.: Interactive Medical Image Annotation Using Improved Attention U-Net with Compound Geodesic Distance. *Expert Systems with Applications*, (2024), 237(Part A): 121282
37. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.* **110**, 457–506 (2021)
38. Chow, G.: Tests of equality between sets of coefficients in two linear regressions. *Econometrica* **28**(3), 591–605 (1960)
39. Hwang, J., Ding, A.: Prediction intervals for artificial neural networks. *J. Am. Stat. Assoc.* **92**(438), 748–757 (1997)
40. Nix, D., Weigend, A.: Estimating the mean and variance of the target probability distribution. *IEEE Int. Conf. Neural Netw. Orlando, FL, USA* **1**, 55–60 (1994)
41. Khosravi, A., Nahavandi, S., Creighton, D., Atiya, A.: Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Trans. Neural Netw.* **22**(9), 1341–1356 (2011)
42. Ungar, L., Veaux, R., Rosengarten, E.: Estimating Prediction Intervals for Artificial Neural Networks. *The 9th Yale Workshop on Adaptive and Learning Systems*, (1996)
43. Gal, Y., Ghahramani, Z.: Dropout as A Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *The 33rd International Conference on Machine Learning, Yarin Gal, Zoubin, Ghahramani*, 48: 1050–1059 (2016)
44. Carney, J., Cunningham, P., Bhagwan, U.: Confidence and prediction intervals for neural network ensembles. *Int. Joint Conf. Neural Netw. Washington, DC, USA* **2**, 1215–1218 (1999)
45. Khosravi, A., Nahavandi, S., Creighton, D., Atiya, A.: Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE Trans. Neural Netw.* **22**(3), 337–346 (2010)
46. Pearce, T., Zaki, M., Brintrup, A., Neely, A.: High-quality prediction intervals for deep learning: a distribution-free, ensemble approach. *Int. Conf. Mach. Learn. PMLR* **80**, 4075–4084 (2018)
47. Lian, C., Zeng, Z., Wang, X., Yao, W., Su, Y., Tang, H.: Landslide displacement interval prediction using lower upper bound estimation method with pre-trained random vector functional link network initialization. *Neural Netw.* **130**, 286–296 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.