AIMS Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 9(11): 33000–33022.
DOI: 10.3934/math.20241578
Received: 23 August 2024
Revised: 11 November 2024
Accepted: 12 November 2024
Published: 21 November 2024

Research article

NGBoost algorithm-based prediction of mechanical properties of a hot-rolled strip and its interpretability research with ANOVA values

Hongyi Wu, Jinwen Jin and Zhiwei Li*

Institute of New Iron and Steel Materials, Ningbo Iron & Steel Co., Ltd, Zhejiang, 315800, China

* **Correspondence:** Email: lizhiwei@ningbosteel.com.

Abstract: Hot-rolled strip steel is an essential material extensively used in various industrial fields, with its mechanical properties being critical to product quality and engineering design. This article presents a method for predicting the mechanical properties of hot-rolled strip steel using the NGBoost (natural gradient boosting) algorithm. The study focused on predicting tensile strength, yield strength, and elongation of hot-rolled strip steel and compared the predictive results with those obtained from the gradient boosting algorithm, Lasso regression, and decision tree algorithms. The results indicated that the NGBoost algorithm performs well on average coverage error (ACE) and prediction interval absolute width (PIAW) values at different confidence levels, demonstrating strong predictive performance. Furthermore, the analysis of variance (ANOVA) method was employed to identify factors that significantly impact mechanical performance, providing theoretical support for optimizing design schemes and enhancing structural safety and reliability.

Keywords: hot-rolled strip steel; mechanical properties; NGBoost algorithm; ANOVA value
Mathematics Subject Classification: 62F07, 62J05

1. Introduction

As a key component of the national economy, the steel industry plays a crucial role in aerospace technology, production and manufacturing, and national defense construction. Its vitality has driven the steady growth of China's economy and society. The extensive production and application of steel products cover many national construction fields, including machinery manufacturing, automobile production, construction equipment, and electrical appliance manufacturing. However, in response to the new era and its development challenges, the steel industry is undergoing a historic transformation. With the continuous emergence of new situations, formats, and models, the steel industry has entered a new stage of development [1].

As an important metal material, hot-rolled strip steel plays a significant role in the engineering field.

Predicting its mechanical properties can not only shorten the product development cycle and improve production quality, stability, and efficiency but also help reduce emissions and save resources [2–7]. Although system identification can estimate model parameters well in complex environments, it cannot provide uncertainty quantification [8, 9]. Accurately predicting the mechanical properties of hot-rolled strip steel is of practical significance for material design, processing, and final product performance. With the continuous advancement of industrial technology and evolving needs, research on predicting the mechanical properties of hot-rolled strip steel has become increasingly urgent. Although traditional experimental methods can provide some performance data, they are limited by factors such as time, cost, and feasibility. Consequently, developing prediction technologies based on theoretical models and computational methods has become a prominent research focus. Reference [10] uses a deep belief network combined with a quantile regression loss function and an ε-insensitive loss function to build two models. These models employ unsupervised pre-training of the underlying restricted Boltzmann machine and supervised fine-tuning with the backpropagation (BP) algorithm to enhance prediction capabilities and avoid local optima. Reference [11] proposed an improved stacked autoencoder model, combining denoising autoencoders and sparse autoencoders to effectively process noise in the data and address the overfitting problem. Reference [12] introduced a hot-rolled strip mechanical prediction model based on extreme gradient boosting (XGBoost), utilizing a genetic algorithm to globally optimize the parameters of the BP neural network, yielding good prediction results. Reference [13] described a method based on a support vector machine quantile regression model. By exploring the relationship between the chemical composition and process parameters of the strip and the tensile strength, the least squares fitting algorithm was used to solve the parameters, achieving a smaller prediction error. Reference [14] introduced a method that combines the random forest algorithm with mechanistic modeling to establish a performance prediction model by ranking the importance of each factor and gradually adding independent variables, thereby improving the accuracy of the prediction model.

Traditional methods for predicting mechanical properties mainly rely on physical models and empirical formulas. Although these methods can reflect the characteristics of materials to some extent, they often suffer from issues such as complex models, high computational demands, and limited prediction accuracy. With the development of big data technology and machine learning algorithms, data-driven prediction methods have gradually become a research hotspot. By analyzing and modeling large datasets from the production process, these methods can predict the mechanical properties of materials more accurately, thereby guiding production practices more effectively.

NGBoost (natural gradient boosting) is an emerging probabilistic prediction model based on the gradient boosting framework that can perform both regression and classification tasks simultaneously. Unlike traditional gradient boosting methods, NGBoost not only provides prediction values but also measures the uncertainty of these predictions, offering high flexibility and robustness [15]. Reference [16] studied the application of NGBoost in the field of civil engineering and verified the feasibility of the algorithm in strength prediction and damage classification. Reference [17] developed a bond strength prediction model for reinforced concrete structures based on NGBoost and provided an assessment of safety risk uncertainty. Reference [18] applied NGBoost for uncertainty estimation of long-term creep fracture life prediction in specific steel. Reference [19] employed NGBoost for seismic vulnerability analysis of building structures, thereby reducing model uncertainty and enhancing model prediction accuracy. In the field of materials science, using the NGBoost model to predict the mechanical properties of hot-rolled strip steel can effectively combine multi-dimensional feature information, such as chemical composition and process parameters, to deliver more accurate and reliable prediction results.

The main purpose of this study is to use the NGBoost algorithm to predict the mechanical properties of hot-rolled strip steel and verify its effectiveness and stability through experiments. This paper first introduces the data collection and preprocessing methods, followed by a detailed description of the feature engineering process and model construction method. The model is then trained and evaluated, with the prediction results assessed using the average coverage error (ACE) and prediction interval absolute width (PIAW). The research results demonstrate that the prediction model based on NGBoost exhibits excellent performance in terms of accuracy and stability, providing an effective solution for predicting the mechanical properties of hot-rolled strip steel.

2. Prediction of steel mechanical properties based on the NGBoost algorithm

2.1. Prediction of mechanical properties of hot-rolled strip steel

As shown in Figure 1, the production of hot-rolled strip steel involves several key processes, including raw material preparation, heating, and rough rolling. Starting from raw material preparation, the steel undergoes rough rolling after being heated to a suitable temperature, forming a thick strip of steel plate. This is followed by finishing, which thins the plate and improves surface quality and dimensional accuracy. The cooling process needs to control the speed to regulate the structure and performance of the steel. After cooling, the steel is cut, possibly subjected to rust removal, and finally coiled for easy transportation and storage.

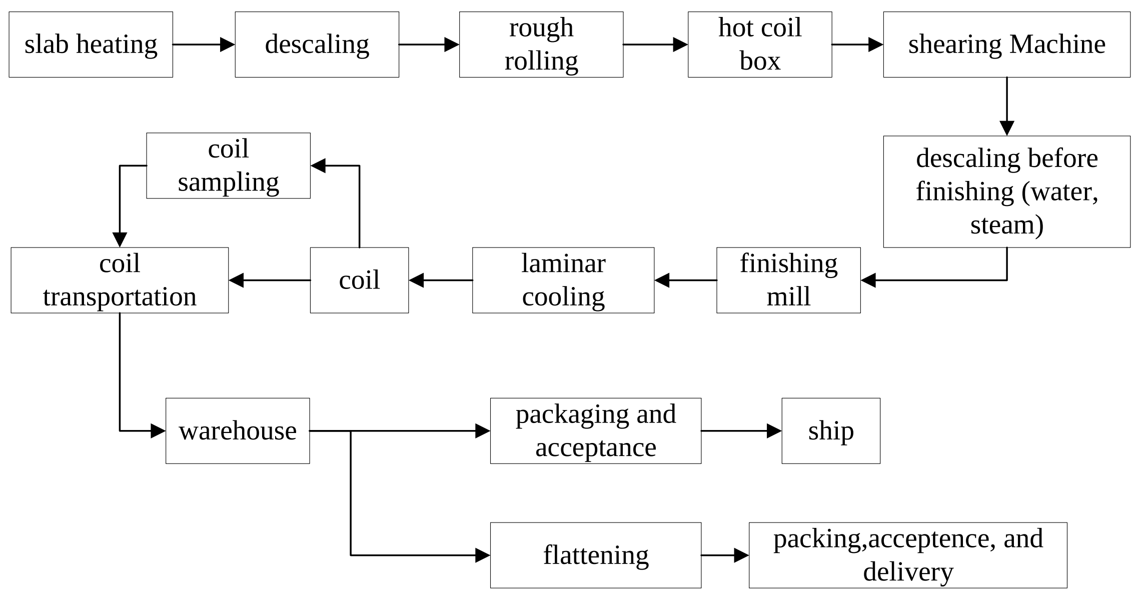

Figure 1. Main process diagram of hot-rolled strip production.

The mechanical properties of steel materials are usually measured by indicators such as yield strength, tensile strength, and elongation. Yield strength refers to the stress level at which steel materials begin to undergo plastic deformation, marking the transition from the linear elastic stage to the plastic deformation stage during a tensile test. Tensile strength refers to the maximum tensile stress that the material can withstand under tension. Elongation reflects the ductility of the material; the greater the elongation, the better the ductility of the steel. These three performance indicators are among the most critical parameters in steel materials and must be strictly controlled in accordance with relevant standards during the production process [12].

Prediction of the mechanical properties of hot-rolled strip steel involves forecasting and evaluating the material's performance during its use. By assessing these properties under various working conditions, engineers and designers can better understand the stress conditions that materials will experience, allowing them to optimize design schemes and enhance structural safety and reliability. Predicting the mechanical properties typically considers factors such as the material's characteristics, working conditions, and stress distribution. This can be achieved through numerical simulations, experimental testing, and empirical formulas.

2.2. Construction of the steel mechanical properties prediction model based on the NGBoost algorithm

The NGBoost algorithm is a new ensemble algorithm proposed by Andrew Ng's team in October 2019, also known as the natural gradient-based boosting method. This ensemble learning method has a unique feature, namely, it can directly generate a complete probability distribution in the output space, so that probability can be used to predict uncertainty [20].

In NGBoost, base learners usually use decision trees or other simple regression models. Each base learner learns the residuals (i.e., the difference between the current model prediction and the actual value) during training and fits these residuals to gradually improve the prediction accuracy of the overall model. Assuming that the mechanical properties of hot-rolled strip steel follow a normal distribution, choosing a suitable probability distribution is crucial to accurately reflect the characteristics of the data and the performance of the model. The model fits the data by optimizing the parameters of these probability distributions (such as mean and standard deviation).

Assume that the prediction $y|x$ for a new input x is in the form of a conditional probability distribution P_θ, where the parameter θ is obtained by combining the outputs of M base learners with the initial value, and θ completely determines the probability prediction $y|x$,

$$y|x \sim P_\theta(x) \quad \theta = \theta^{(0)} - \eta \sum_{m=1}^{M} \rho^{(m)} \cdot f^{(m)}(x). \tag{2.1}$$

A proper scoring rule $\mathcal{S}$ will take as input the predicted probability distribution P and the actual observed outcomes y and assign a score to each prediction such that the true distribution of outcomes has the best score expectation if and only if it satisfies:

$$E_{y \cdot Q}[S(Q, y)] \leqslant E_{y \cdot Q}[S(P, y)], \quad \forall P, Q, \tag{2.2}$$

where Q represents the result of the true distribution and P represents any other distribution. During the training process, an appropriate scoring rule is used as the loss function to encourage the model to output calibrated probability predictions. In addition, it is necessary to restrict the parameter family of probability distributions and use its parameters to determine a specific distribution. In this case, the most commonly used scoring rule is the logarithmic score $\mathcal{L}$, also known as maximum likelihood estimation (MLE):

$$\mathcal{L}(\theta, y) = -\log \mathrm{P}_\theta(y)\,. \tag{2.3}$$

First, a dataset containing mechanical property data of hot-rolled strip samples is collected. This dataset includes tensile strength, yield strength, and elongation as target variables, along with various chemical compositions and process parameters as feature variables related to these mechanical properties.

Next, the NGBoost algorithm is employed to build a prediction model. NGBoost enhances prediction performance by iteratively fitting weak prediction models. In each iteration, it adjusts the model parameters to minimize losses based on the gradient direction of the negative log-likelihood loss function. The NGBoost model is used to predict the target variable by learning the relationship between the input features and the target variables.

Finally, the models prediction results are analyzed, and the prediction process is explained as needed. The feature importance information provided by NGBoost is used to identify which features most significantly contribute to the prediction of mechanical properties. This helps in gaining a deeper understanding of the relationship between the features and the mechanical properties of hot-rolled strip steel.

The input of the NGBoost algorithm for hot-rolled strip mechanical property prediction based on MLE and the natural gradient is the actual test data $D = \{x_i\ , y_i\}_{i-1}^{n}$ of a steel plant. Among them, n represents the total data sample size, the subscript i represents the i-th sample (the same below), the number of iterations is M, the learning rate is η, and the loss function is $\mathcal{L}$. Assume that the mechanical properties of hot-rolled strip steel obey the normal distribution with parameter $\theta = (\mu, \sigma)$, where μ represents the mathematical expectation and σ represents the standard deviation.

2.3. Calculation of prediction interval of the steel mechanical properties

According to the prediction probability distribution of the mechanical properties of hot-rolled strip steel, the prediction interval of the mechanical properties with a confidence level of $I_\alpha = [L_\alpha, U_\alpha]$ can be obtained, where [21]:

$$\begin{cases} L_a = \mu - Z_{a/2}\sigma, \\ U_a = \mu + Z_{a/2}\sigma. \end{cases} \tag{2.4}$$

2.4. Implementation process

The entire implementation process is shown in Algorithm 1, including data collection and preparation, feature engineering, etc.

Algorithm 1 Implementation process of using the NGBoost algorithm to predict mechanical properties of hot-rolled strip steel

Require: Dataset, including chemical composition, rolling process parameters, and mechanical properties of hot-rolled strip steel.

Ensure: Trained NGBoost model and use ACE and PIAW for model evaluation

1: **Data collection and preparation:**
- Collect data from a steel plant, including chemical composition, rolling process parameters, and mechanical properties of hot-rolled strip steel.
- Data preprocessing, handling missing values, identifying and handling outliers, and scaling features.

2: **Feature engineering:**
- Perform feature selection on raw data based on domain knowledge and actual conditions.
- Select main features for analysis to improve the prediction ability of the model.

3: **Data partitioning:**
- Divide the data set into a training set and test set.
- A cross-validation or holdout method is usually used.

4: **NGBoost model establishment:**
- Use the NGBoost library in Python to build a prediction model.
- Select appropriate parameters to achieve good prediction results.

5: **Model training:**
- Train the NGBoost model using the training set.
- Learn the probability distribution of the data by iteratively optimizing the loss function.

6: **Model evaluation:**
- Evaluate the performance of the model using the test set.
- Use indicators such as ACE and PIAW to evaluate the accuracy and stability of the prediction results.

3. Interpretation of the mechanical property prediction results based on the ANOVA value

3.1. Definition of the ANOVA value

Analysis of Variance (ANOVA), also known as the F test, is a statistical method presented by Sir Ronald Fisher. It is used to compare whether there are significant differences between the means of two or more samples [22]. As a commonly used statistical method, ANOVA analyzes the impact of different factors on data variation and determines which factors have a significant effect [23].

In predicting the mechanical properties of hot-rolled strips, multiple influencing factors, such as chemical composition and rolling process parameters, can affect the mechanical properties of the strips. Therefore, ANOVA can be used to determine which factors have a significant impact on the mechanical properties and to explore interactions between factors. For example, factors like chemical composition and rolling temperature can be used as influencing factors, while mechanical properties (such as yield strength) can be used as response variables for ANOVA.

The ANOVA value is a result of multi-factor analysis of variance and measures the influence of each factor on the variable. A larger ANOVA value indicates a more significant impact of the factor on the variable; conversely, a smaller value suggests a less significant impact. In predicting the mechanical properties of hot-rolled strip steel, the ANOVA value can identify which factors have a crucial influence on mechanical properties and their relative importance.

Studying the ANOVA value provides an in-depth understanding of the relationship between the mechanical properties of hot-rolled strip steel and various factors. This analysis helps optimize process parameters and improve the prediction accuracy and stability of the mechanical properties.

3.2. Calculation of the ANOVA value

The main factors affecting the mechanical properties of hot-rolled strip steel are expressed by ANOVA [24, 25], which represents the main factors affecting the mechanical properties of hot-rolled strip steel. The specific analysis is as follows:

(1) Calculate the mean of each sample

Calculate the average value of each sample, the sample average value corresponding to the influencing factor and the total mean of all observations. The calculation formula is as follows:

$$\bar{x}_{ik} = \frac{\sum\limits_{j=1}^{n_i} x_{ijk}}{n_i}, \quad i = 1, 2, ..., n, \quad k = 1, 2, 3,$$
$$\bar{x}_k = \frac{\sum_{i=1}^{n} n_i x_{ik}}{n}, \quad k = 1, 2, 3, \tag{3.1}$$

where n_i is the sample size of the i-th influencing factor, x_{ijk} is the k-th observed value of the j-th type of influencing factor under the i-th performance indicator, $\bar{x}_{ik}$ is the sample mean of the k-th type of influencing factor under the i-th performance indicator, and $\bar{x}_k$ is the total mean of all observations under the k-th performance indicator.

(2) Calculate the sum of squares of each error

E_{SSAk} is the sum of squares of the errors between the mean of each group $\bar{x}_{ik}$ and the total mean under the k-th performance index, E_{SSEk} is the sum of squares of the errors between the sample data of each group and its group mean under the k-th performance index, and the calculation formula is as follows:

$$E_{SSAk} = \sum_{i=1}^{n} n_i(\bar{x}_{ik} - \bar{x}_k)^2, \quad k = 1, 2, 3,$$
$$E_{SSEk} = \sum_{i=1}^{n} \sum_{j=1}^{n_i} (x_{ijk} - \bar{x}_{ik})^2, \quad k = 1, 2, 3, \tag{3.2}$$

(3) Calculate statistics

In order to eliminate the effect of the number of observations on the size of the error sum of squares, E_{SSAk} and E_{SSEk} need to be averaged. The E_{SSAk} degrees of freedom is $i-1$, and the degrees of freedom is $n - i$. The mean square and statistic calculation formula are as follows:

$$E_{MSAk} = \frac{E_{SSAk}}{i - 1},$$
$$E_{MSEk} = \frac{E_{SSEk}}{n - k},$$

$$F = \frac{E_{MSAk}}{E_{MSEk}}, \tag{3.3}$$

where F is the test statistic. The larger of the value of F, the more significant the influence of this influencing factor on the mechanical properties of the strip steel.

4. Case analysis

This paper takes the hot-rolling production line data of a steel plant as an example to build a prediction model to predict the mechanical properties of hot-rolled strip steel.

4.1. Experimental data and prediction algorithm settings

The hot-rolling data used in this article includes 138 features and 6300 samples, and a description of all features is visible in Table 1. Due to the harsh environment of the data source, data preprocessing is performed first. Data preprocessing plays a vital role in the entire data analysis process because it is directly related to the accuracy of subsequent model building and performance prediction results. In order to ensure the robustness of the model, missing values processing, outlier processing, and feature selection become three important links.

Table 1. Category of features.

Features	Technique
chemical composition	heating
temperature	heating, cooling, and curl
velocity	rolling and cooling
time	heating, cooling, and curl
sizes	rolling

First, missing values in the data are a common problem, especially in industrial production environments where some sensors may fail to record data due to malfunction or other reasons. Missing values that are not properly handled may lead to bias in model training or even cause the algorithm to fail to operate properly. We adopt the method of missing value deletion for data rows with few records and a large number of missing values to ensure data quality. After processing, 5834 samples were obtained with the same dimensions.

Second, outliers are data points that are significantly different from the majority of the data and are usually caused by measurement errors or extreme operating conditions. In order to identify and deal with outliers, we use the 3-sigma principle, which assumes that the data follow a normal distribution, and data points that are more than 3 standard deviations away from the mean are considered outliers and are removed. This prevents outliers from adversely affecting the model. In this way, we obtained 5331 samples.

Finally, production data often contain a large number of features, some of which have a weak correlation with the prediction target and may even introduce noise. Therefore, feature selection is a very critical step in data preprocessing. We utilize recursive feature elimination to screen features. By recursively training the XGBoost model, evaluating the importance of features, and gradually eliminating unimportant features, the most representative features are finally selected for model training. The influencing factors selected for constructing the prediction model are shown in Table 2.

Table 2. Influencing factors of selecting the performance model for the training prediction of the mechanical properties of hot-rolled strip steel.

Influencing factors	Illustrate
C, Si, Mn, P, S, Nb, V, Ti, Cr, Ni, Mo, Cu	Chemical composition
temp_dischg	Out-of-furnace temperature
temp_fdt_avg	Average finishing rolling outlet temperature
temp_ct_avg	Average coiling temperature
size_product_thk_avg	Average finished product thickness

After preprocessing, a sample set consisting of 5331 data with a dimension of 16 is obtained, which is divided into a training set and a testing set in a 7:3 ratio. For the training set, a random search strategy based on the Scikit-learn library is used as the optimization method for NGBoost hyperparameters using decision trees (DT) as regression models. Through 10-fold cross-validation, optimal parameter selection is obtained as shown in Table 3.

Table 3. Main parameter setting of the NGBoost algorithm.

Parameter name	Parameter description	Settings
Number of DT	Number of base estimators	1000
Learning rate	The step size or scaling factor used during gradient boosting for each base model	0.01
Max depth	Maximum depth of DT	3
Tolerance value Tol	Control the stopping of the training process	0.0001

4.2. Steel performance prediction and evaluation index

ACE and PIAW are used to evaluate the prediction performance of the algorithm. ACE reflects the reliability of the prediction, while PIAW reflects the sensitivity of the prediction [26].

ACE is defined as the difference between the empirical coverage probability (ECP) and the nominal coverage probability (NCP) where the actual value falls within the predicted range. The calculation formula is shown in (4.1).

$$\begin{cases} D_{ACE} = P_{ECP} - P_{NCP}, \\ P_{ECP} = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i, \\ \varepsilon_t = \begin{cases} 1, y \in [L_a, U_a], \\ 0, y \in [L_a, U_a], \end{cases} \\ P_{NCP} = 1 - \alpha, \end{cases} \tag{4.1}$$

where D_{ACE}, D_{ECP}, and D_{NCP} represent the values of ACE, ECP, and NCP, respectively. U_α and N_α correspond to the upper and lower limits of the prediction interval. N represents the number of data points. y represents the actual value of the mechanical properties of each data point, and ε_α represents the indicator function, which is 1 only when the actual value of the mechanical properties is within the prediction interval, and otherwise it is 0. The smaller the absolute value of ACE, the higher the credibility of the prediction interval.

PIAW is defined as the average width of the prediction interval, and its calculation is shown in (4.2).

$$P_{PIAW} = \frac{1}{N}\sum_{i=1}^{N}(U_{i,\alpha} - L_{i,\alpha}). \tag{4.2}$$

The smaller the value of PIAW is, the better the prediction result is at aggregating uncertain information.

4.3. Analysis of the prediction results of the steel mechanical properties

As can be seen from Figures 2–7, the NGBoost algorithm has the largest error in predicting elongation and the smallest error in predicting tensile strength, and converges very quickly. After about 175 iterations, the loss function changes slightly. ACE and PIAW are used as probability prediction indicators, and the root mean squared error (RMSE) and R-squared (R^2) are used as single-point prediction indicators (the mean is treated as a predict value). Table 4 shows the performance of the algorithm on the training set and test set.

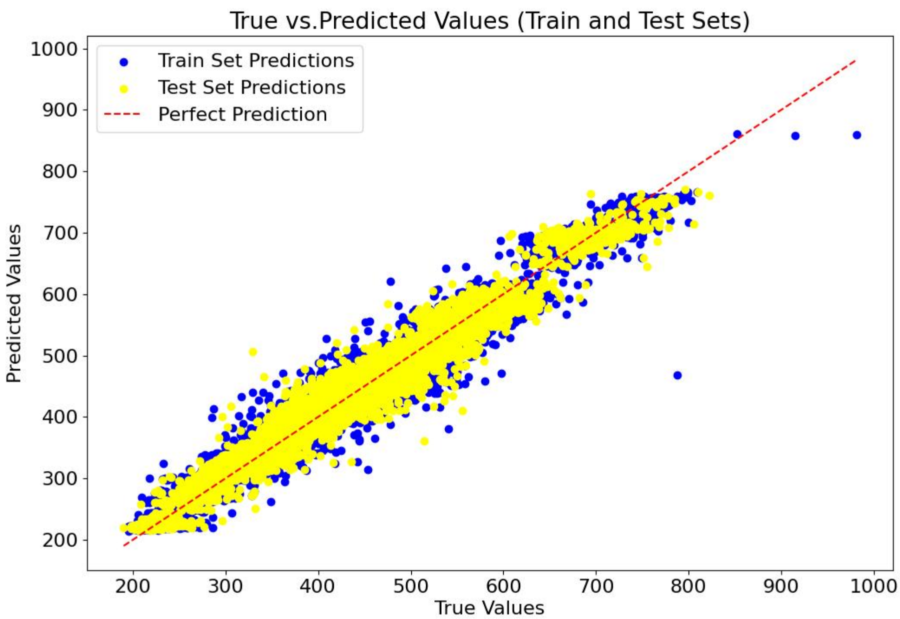

Figure 2. Yield strength prediction comparison.

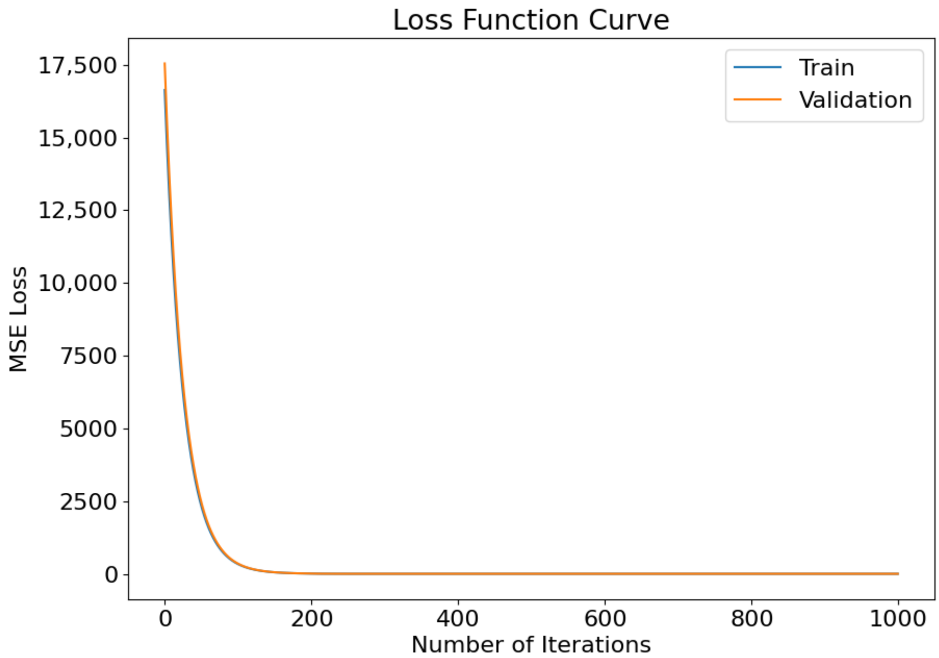

Figure 3. Yield strength prediction iterative process.

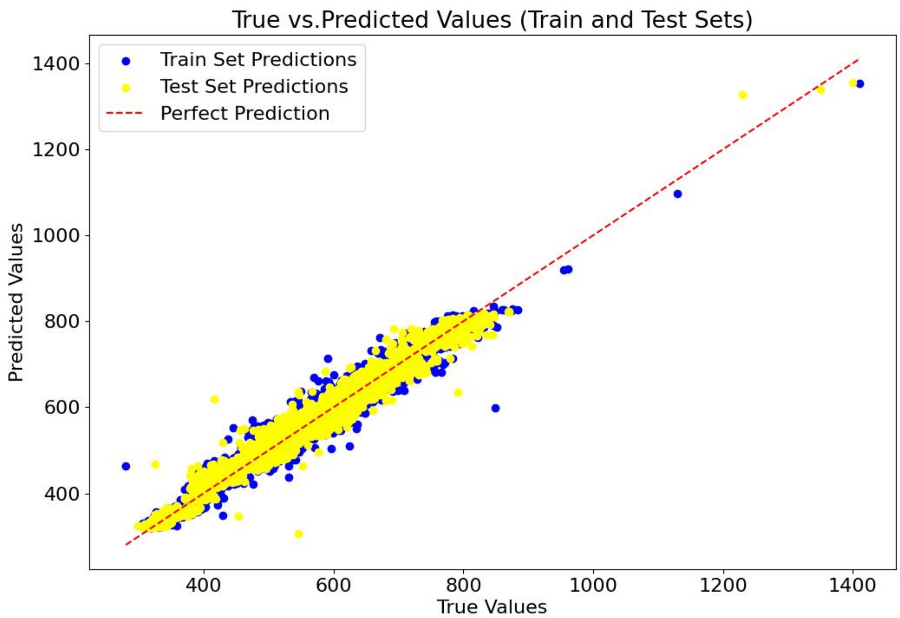

Figure 4. Tensile strength prediction comparison.

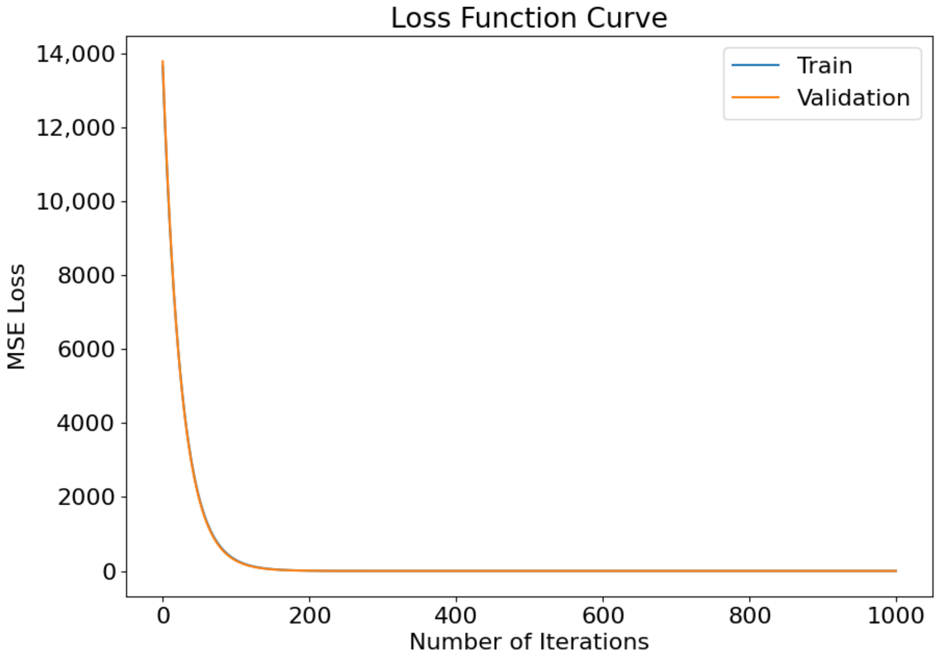

Figure 5. Tensile strength prediction iterative process.

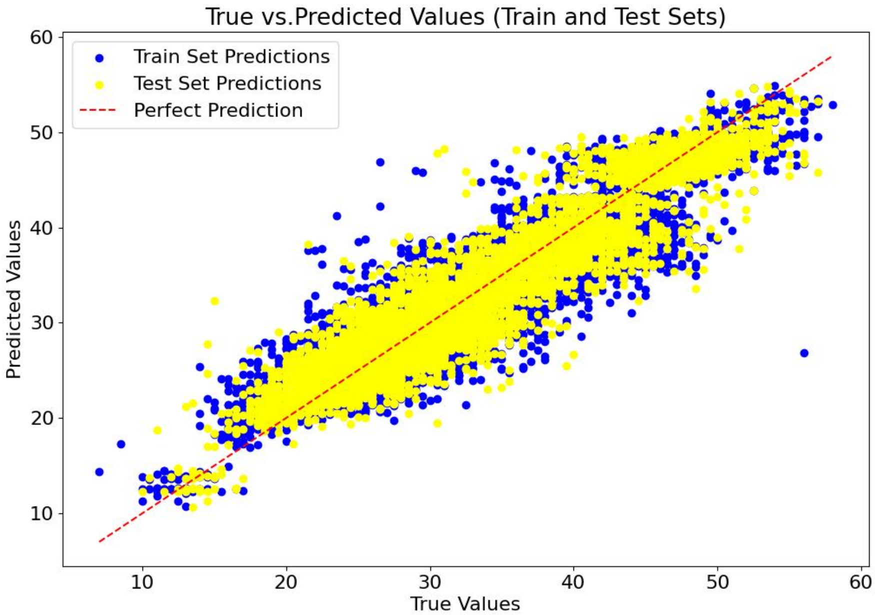

Figure 6. Elongation prediction comparison.

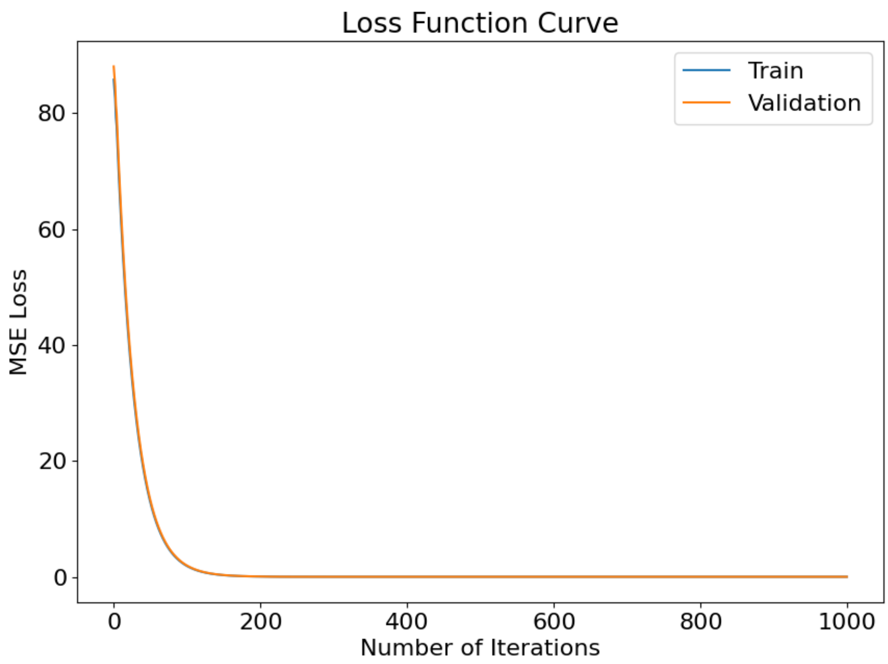

Figure 7. Elongation prediction iterative process.

Table 4. Performance of NGBoost on training and test sets.

Set	Yield Strength				Tensile Strength				Elongation			
	ACE	PIAW	RMSE	R^2	ACE	PIAW	RMSE	R^2	ACE	PIAW	RMSE	R^2
Train	22	49	19	95	32	45	18	95	-4.5	5	1.4	96
Test	27	52	21	93	35	52	19	95	-4.6	6	2.1	95

4.4. Prediction and comparative analysis of the mechanical properties of steel

Since NGBoost not only provides probabilistic predictions but also gives point estimates, therefore, we perform a comparison between the probabilistic prediction metrics ACE and PIAW, and likewise for the RMSE and R^2 metrics for mean prediction.

Remark 1. *Since most regression models do not provide probabilistic predictions, here we use residual analysis to construct prediction confidence intervals. Specifically, the regression model is trained with a training set, the residuals of the predicted and true values are computed, the distribution of the residuals is fitted to obtain the mean and variance, and the confidence intervals are constructed for the predicted values of the new test data.*

The regression performance and probability prediction performance of the NGBoost algorithm are compared with that of the traditional gradient boosting (GB) algorithm, the Lasso regression algorithm, and the DT regression algorithm. The results of the RMSE and R^2 metrics for these algorithms are shown in Table 5. According to the probability distribution prediction results of yield strength, tensile strength, and elongation using different algorithms, the ACE and PIAW of the prediction interval are calculated at the confidence levels of 80%, 90%, and 95%, respectively, and the results are shown in Tables 6–8.

From the results in Tables 6 and 8, it can be seen that when the confidence level is set to 80%, 90%, and 95%, the NGBoost algorithm obtains the smallest absolute ACE and PIAW values compared with other algorithms. Therefore, the reliability of the NGBoost algorithm is higher than that of other algorithms.

According to the results in Table 7, when the confidence level is 80%, the reliability of the DT algorithm is relatively higher than that of other algorithms (the absolute value of ACE is the smallest), but compared with the NGBoost algorithm, its prediction interval is wider (the PIAW value is larger), indicating that its ability to aggregate uncertain information is relatively weak. Specifically, the NGBoost algorithm not only performs well in terms of reliability, but also has a narrow prediction interval (a smaller PIAW value), indicating that it has a strong ability to aggregate uncertain information.

Table 5. RMSE and R^2 of yield strength, tensile strength, and elongation predicted by each algorithm at different confidence levels.

Algorithm	Yield strength		Tensile strength		Elongation	
	RMSE/Mpa	R^2/%	RMSE/Mpa	R^2/%	RMSE/%	R^2/%
NGBoost	20.093	93	19.129	95	2.176	96
GB	21.71	93	22.7	93	2.47	94
Lasso	34.015	85	27.92	89	3.211	87
DT	25.38	91	23.63	90	2.67	92

Table 6. ACE and PIAW of prediction intervals of yield strength predicted by each algorithm at different confidence levels.

Algorithm	Confidence level: 80%		Confidence level: 90%		Confidence level: 95%	
	ACE/%	PIAW/Mpa	ACE/%	PIAW/Mpa	ACE/%	PIAW/Mpa
NGBoost	27.0	51.383	27.0	65.950	27.0	78.584
GB	34.7	54.932	34.7	70.503	34.7	84.009
Lasso	69.2	99.045	88.8	153.292	105.8	182.658
DT	41.7	59.622	53.5	75.523	63.7	91.183

Table 7. ACE and PIAW of prediction intervals of tensile strength predicted by each algorithm at different confidence levels.

Algorithm	Confidence level: 80%		Confidence level: 90%		Confidence level: 95%	
	ACE/%	PIAW/Mpa	ACE/%	PIAW/Mpa	ACE/%	PIAW/Mpa
NGBoost	36.9	40.931	37.7	52.535	37.3	62.596
GB	39.8	43.687	38.9	55.878	39.1	66.599
Lasso	49.4	73.270	63.3	94.041	75.5	112.056
DT	30.7	45.511	39.3	58.412	46.9	69.093

Table 8. ACE and PIAW of prediction intervals of elongation predicted by each algorithm at different confidence levels.

Algorithm	Confidence level: 80%		Confidence level: 90%		Confidence level: 95%	
	ACE/%	PIAW/%	ACE/%	PIAW/%	ACE/%	PIAW/%
NGBoost	-4.6	7.007	-4.6	5.950	-4.6	10.584
GB	-4.7	7.011	-4.7	8.998	-4.7	10.722
Lasso	9.9	14.731	12.7	18.907	15.2	22.528
DT	5.4	8.006	7.0	10.082	8.3	12.336

In order to verify the performance of the NGBoost algorithm in the probability prediction of mechanical properties, this paper uses this algorithm and other methods to predict the probability distribution of mechanical properties, and calculates the prediction intervals at different confidence levels (50%, 80%, 90%, 95%, 98%, 99%). In order to see the comparison results of different algorithms more clearly and intuitively, the tabular data is presented in the form of a histogram, and the results are shown in Figures 8–13. It can be seen from Figures 8–10 that at different confidence levels, the absolute value of ACE of the NGBoost algorithm is small, showing higher reliability, and according to Figures 11–13, it can be seen that the PIAW of the NGBoost algorithm is small, showing better sensitivity. In summary, the NGBoost algorithm not only provides higher reliability, but also shows better sensitivity.

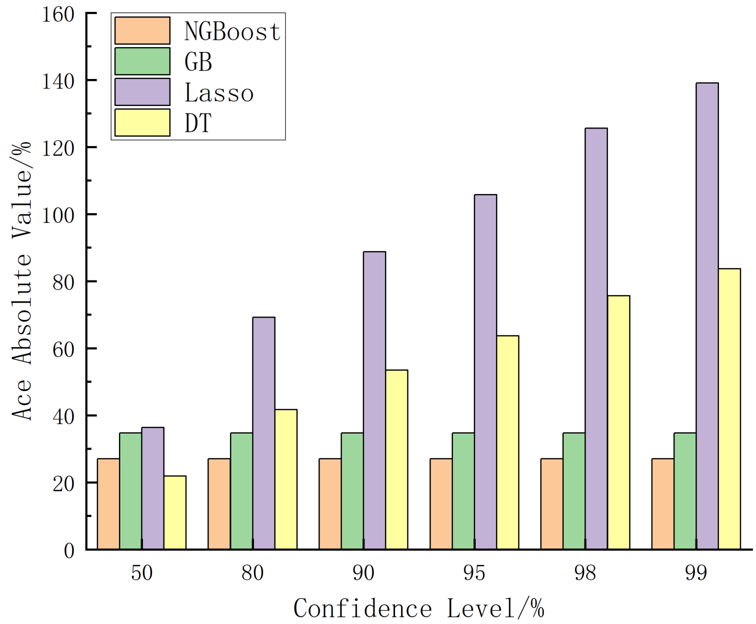

Figure 8. ACE value of yield strength.

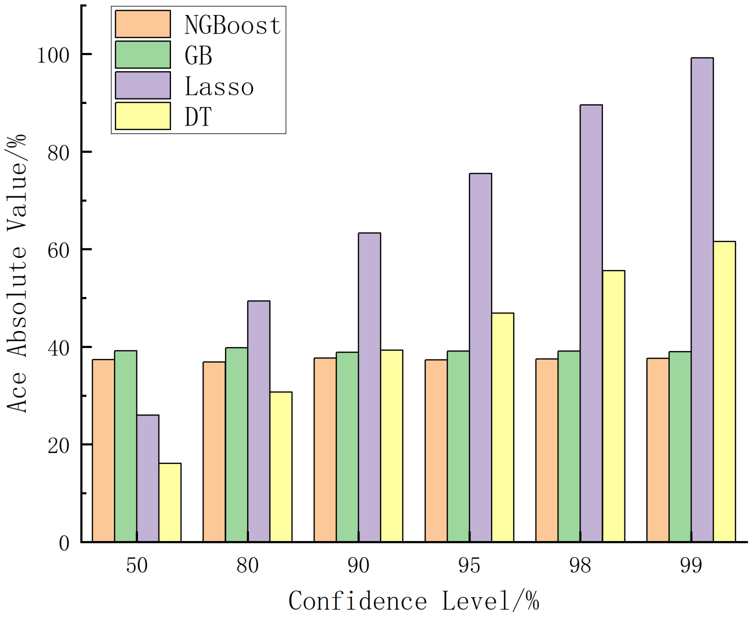

Figure 9. ACE value of tensile strength.

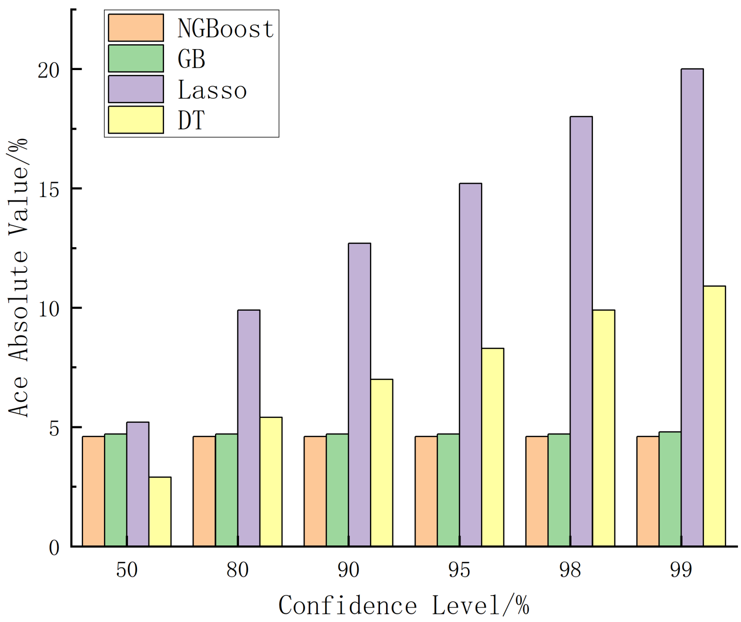

Figure 10. ACE value of elongation.

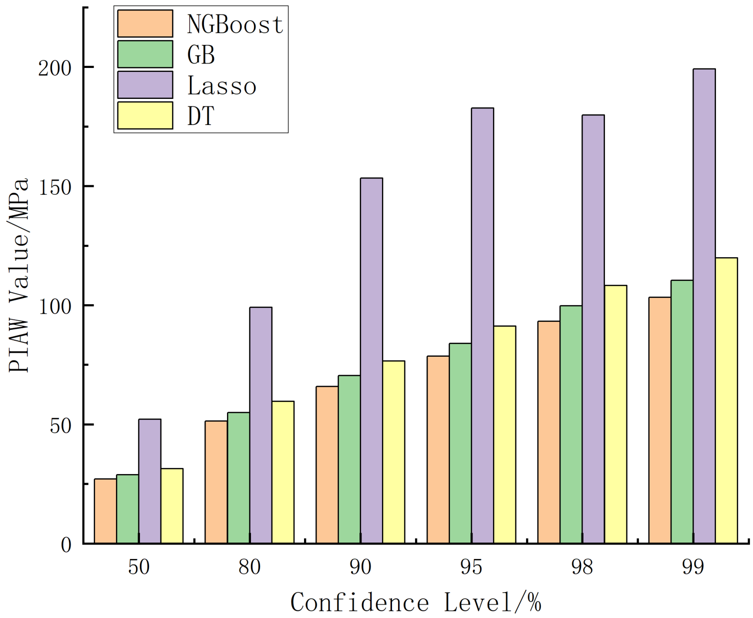

Figure 11. PIAW value of yield strength.

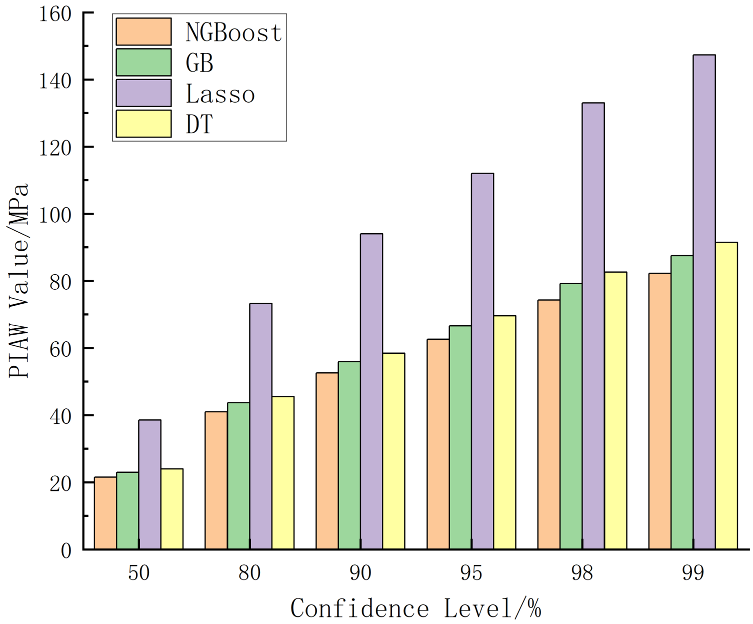

Figure 12. PIAW value of tensile strength.

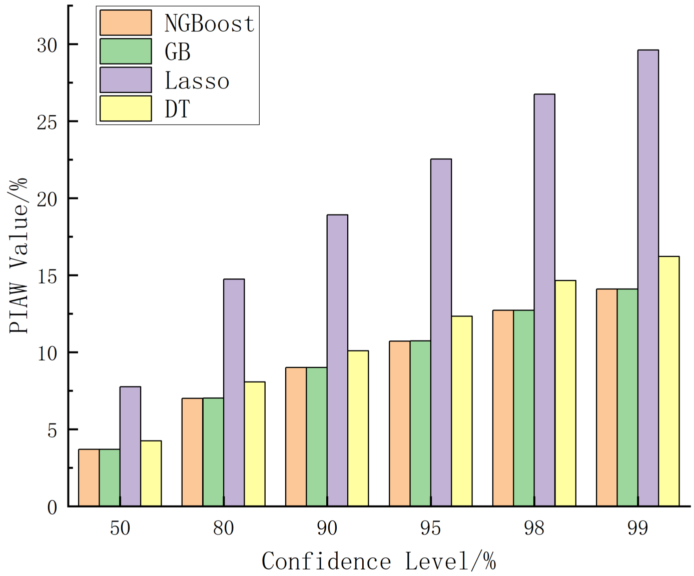

Figure 13. PIAW value of elongation.

Remark 2. *To prevent model overfitting, we adopt various measures, including cross-validation and an early stopping mechanism, that stops model training when the performance of the validation set is no longer improving. These methods effectively improve the generalization ability of the model.*

4.5. Importance analysis of the characteristics affecting the mechanical properties of steel

After performing the ANOVA, the results can be interpreted in terms of F-values and p-values, where the F-value represents the degree of difference in the variance between different groups. The larger the F-value, the more significant the difference in mean values between different groups, and the p-value indicates the probability of whether the observed results are caused by random factors. Generally, if the p-value is less than the significance level (such as 0.05), the null hypothesis is rejected and the difference is considered significant.

The F-values and p-values of different mechanical properties are shown in the Figures 14–16 below:

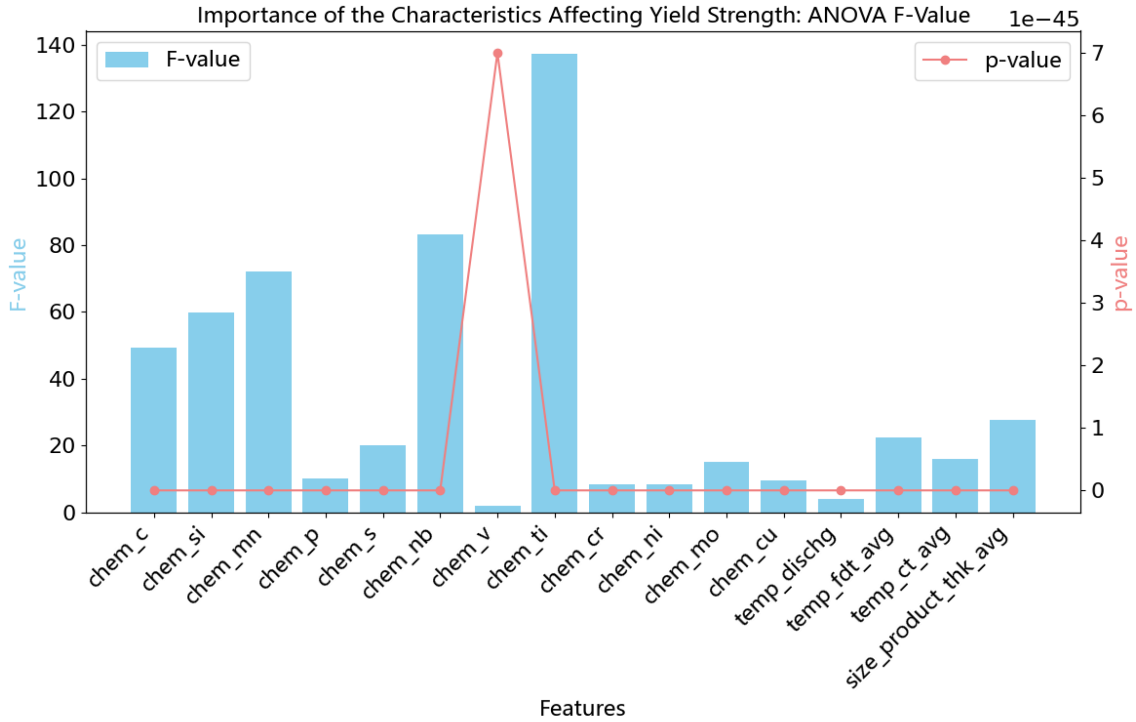

Figure 14. Importance of the characteristics affecting yield strength.

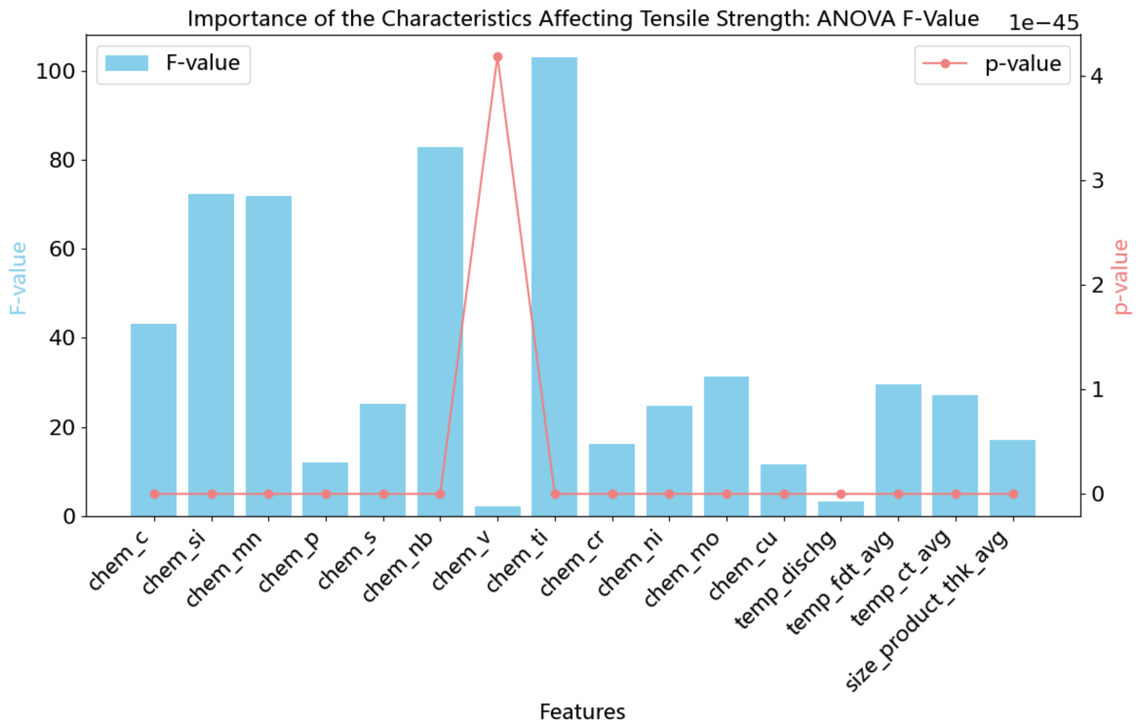

Figure 15. Importance of the characteristics affecting tensile strength.

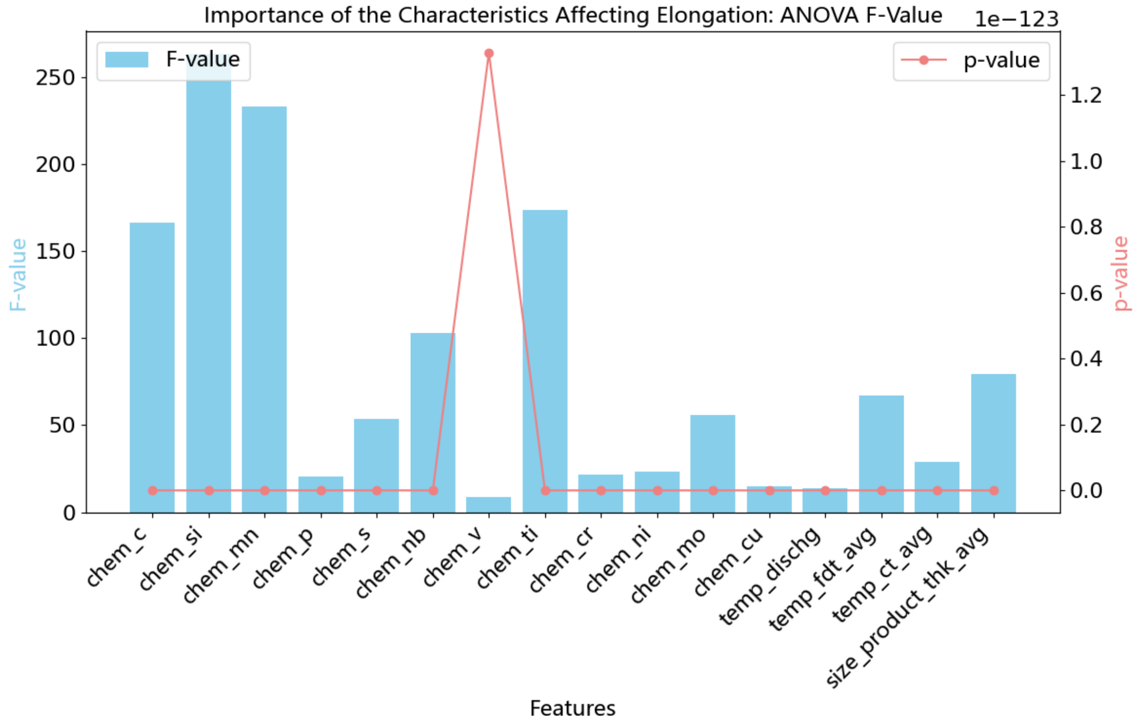

Figure 16. Importance of the characteristics affecting elongation.

It can be seen from Figure 14 that the top six characteristics with the greatest influence on yield strength are Ti, Nb, Mn, Si, C content, and the average finishing outlet temperature. It can be seen from Figure 15 that the top six characteristics with the greatest influence on tensile strength are Ti, Nb, Si, Mn, C, Mo content. It can be seen from Figure 16 that the top six characteristics with the greatest influence on yield strength are Si, Mn, Ti, C, Nb content, and the average finished product thickness.

The results of ANOVA show that chemical composition is a key factor affecting the mechanical properties of steel, especially for improving the tensile strength and toughness of steel. These results have the following references for engineering decisions in the steel industry.

1) Chemical composition optimization: Engineers can adjust the ratio of alloying elements based on the analysis results to balance the mechanical property requirements under different working conditions. For example, increasing the carbon content can increase tensile strength, but may reduce ductility. Therefore, by optimizing the chemical composition, engineers can optimize the material properties according to the usage scenarios of the final product.

2) Production cost control: Some rare alloying elements (e.g., chromium, nickel) are costly, and ANOVA can help producers decide whether to use these elements based on their actual needs, thus achieving a balance between performance and cost.

3) Quality control and consistency: By analyzing chemical compositions, manufacturers can better control the consistency of materials during the production process and ensure that the final product meets requirements across batches and working conditions.

5. Conclusions

To predict the mechanical properties of hot-rolled strip steel, analyze the composition (chemical elements) of the strip, and assess the influence of process parameters during rolling, this paper proposes a prediction method based on NGBoost and ANOVA values. The actual production data from a steel plant was used for verification. The results show that:

1) The proposed method can effectively predict the mechanical properties of hot-rolled strip steel, demonstrating fast convergence speed and good prediction performance.

2) The influencing factors of the mechanical properties were analyzed using the ANOVA value, revealing that the main factor affecting the mechanical properties is the chemical composition.

Author contributions

Hongyi Wu: Conceptualization; Jinwen Jin: Methodology; Zhiwei Li: Validation. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. X. L. Wang, J. D. Chi, Intelligent manufacturing promotes the transformation and upgrading of the iron and steel industry, *Metall. Ind. Autom.*, **42** (2018), 1–5.

2. H. Yada, *Prediction of microstructural changes and mechanical properties in hot strip rolling*, In: Proceedings of the Metallurgical Society of the Canadian Institute of Mining & Metallurgy, 1988, 105–119. https://doi.org/10.1016/B978-0-08-035770-6.50012-X

3. M. Suehiro, K. Sato, Y. Tsukano, H. Yada, T. Senuma, Y. Matsumura, Computer modeling of microstructural change and strength of low carbon steel in hot strip rolling, *Transact. ISIJ*, **29** (1987), 439–445. https://doi.org/10.2355/isijinternational1966.27.439

4. T. Senuma, M. Suehiro, H. Yada, Mathematical models for predicting microstructural evolution and mechanical properties of hot strips, *ISIJ Int.*, **32** (1992), 423–432. https://doi.org/10.2355/isijinternational.32.423

5. P. D. Hodgson, R. K. Gibb, A mathematical model to predict the mechanical properties of hot rolled C-Mn and microalloyed steels, *ISIJ Int.*, **32** (1992), 1329–1338. https://doi.org/10.2355/isijinternational.32.1329

6. Y. B. Xu, X. H. Liu, G. D. Wang, Prediction-control model for microstructure and property of hot-rolled steel strip and their application, *J. Iron Steel Res.*, **14** (2002), 65–68. https://doi.org/10.1080/09503150208411543

7. Y. Gan, The R & D of process modeling in thin slab hot strip rolling, *Iron Steel*, **38** (2003), 10–15.

8. J. Guo, R. Jia, R. Su, Y. Zhao, Identification of FIR systems with binary-valued observations against data tampering attacks, *IEEE T. Syst. Man Cy.-S.*, **53** (2023), 5861–5873. https://doi.org/10.1109/TSMC.2023.3276352
9. J. Guo, X. Wang, W. Xue, Y. Zhao, System identification with binary-valued observations under data tampering attacks, *IEEE T. Automat. Contr.*, **66** (2021), 3825–3832. https://doi.org/10.1109/TAC.2020.3029325
10. S. Y. Huang, Prediction of tensile strength of hot rolled strips based on deep belief network, *Wuhan Univ. Sci. Technol.*, 2022.
11. Y. Song, B. Li, C. Liu, F. F. Li, Prediction model of mechanical properties of hot rolled strip based on improved stacked self-encoder, *Metall. Ind. Autom.*, **44** (2020).
12. F. Zhao, Research on prediction of mechanical properties for hot rolled strip based on big data and XGBoost, *Metall. Autom. Design Res. Inst.*, 2020.
13. X. Zhang, Prediction of mechanical properties of hot rolled strips based on support vector quantile regression, *Wuhan Univ. Sci. Technol.*, 2020.
14. W. Yang, W. Li, Y. Zhao, B. K. Yan, W. B. Wang, Mechanical property prediction of steel and influence factors selection based on random forests, *Iron Steel*, **53** (2018), 44–49.
15. E. Díaz, G. Spagnoli, Natural gradient boosting for probabilistic prediction of soaked CBR values using an explainable artificial intelligence approach, *Buildings*, **14** (2024), 352. https://doi.org/10.3390/buildings14020352
16. S. Z. Chen, D. C. Feng, W. J. Wang, E. Taciroglu, Probabilistic machine-learning methods for performance prediction of structure and infrastructures through natural gradient boosting, *J. Struct. Eng.*, **148** (2022), 04022096.
17. X. R. Ma, X. L. Wang, S. Z. Chen, Trustworthy machine learning-enhanced 3D concrete printing: Predicting bond strength and designing reinforcement embedment length, *Automat. Constr.*, **168** (2024), 105754. https://doi.org/10.1016/j.autcon.2024.105754
18. O. Mamun, M. F. N. Taufique, M. Wenzlick, J. Hawk, R. Devanathan, Uncertainty quantification for Bayesian active learning in rupture life prediction of ferritic steels, *Sci. Rep.*, **2083** (2022), 12. https://doi.org/10.1038/s41598-022-06051-8
19. J. Y. Ding, D. C. Feng, E. Brunesi, F. Parisi, G. Wu, Efficient seismic fragility analysis method utilizing ground motion clustering and probabilistic machine learning, *Eng. Struct.*, **294** (2023), 116739. https://doi.org/10.1016/j.engstruct.2023.116739
20. T. Duan, A. Anand, D. Y. Ding, K. K. Thai, S. Basu, A. Ng, et al., *NGBoost: Natural gradient boosting for probabilistic prediction*, In: Proceedings of 37th International Conference on Machine Learning(ICML), 2020, 2690–2700.
21. B. S. Li, C. J. Pang, D. C. Cheng, Interpretable wind power probabilistic prediction based on NGBoost, *Zhejiang Elec. Power*, **42** (2023), 28–36.
22. *Statistics Group, Institute of Mathematics, Chinese Academy of Sciences, Variance Analysis*, Science Press, 1977.
23. H. C. Shen, Y. Zhao, M. Lang, Analysis on the application of variance analysis in the detection of moisture in iron ore, *China Sci. Technol. Period. Database Ind. A*, 2023, 20–22.

24. J. Kang, S. X. Zhang, Q. P. Zhang, B. Gao, Z. H. Yan, Fault diagnosis method of transformer based on ANOVA and BO-SVM, *High Volt. Eng.*, **49** (2023), 1882–1891.

25. F. Li, Y. Song, C. Liu, R. Jia, B. Li, Research on error distribution modeling of mechanical performance prediction model for hot rolled strip, *Metall. Ind. Autom.*, **43** (2019), 28–33.

26. C. Wan, Y. H. Song, Theories, methodologies and applications of probabilistic forecasting for power systems with renewable energy sources, *Autom. Electr. Power Syst.*, **45** (2021), 2–16.