




TECHNICAL ARTICLE

Prediction Model of Yield Strength of V–N Steel Hot-rolled Plate Based on Machine Learning Algorithm

ZONGXIANG SHI,¹ LINXIU DU ^{1,4} XIN HE,¹ XIUHUA GAO,¹
HONGYAN WU,¹ YANG LIU,¹ HENG MA,² XIAOXIN HUO,²
and XUEHUI CHEN³

1.—State Key Laboratory of Rolling and Automation, Northeastern University, Shenyang 110819, Liaoning, China. 2.—Laiwu Iron and Steel Group, Yinshan Steel Co., Ltd, Jinan 271104, Shandong, China. 3.—Central Iron and Steel Research Institute Co., Ltd, Beijing 10081, China. 4.—e-mail: dulx@ral.neu.edu.cn

Mechanical properties are an essential standard for V–N steel hot-rolled plates used in steel structures such as ship hulls, paint pipelines and offshore platforms. To solve the problems of low production efficiency and low applicability of the traditional physical metallurgy (PM) model, this study proposed an adequate model, namely eXtreme Gradient Boosting based on Bayesian optimization (BO-XGBoost). First, composition-process-yield strength data of V–N steel hot-rolled plate with steel grade Q550D were collected, and K nearest neighbor (KNN), support vector machine (SVR), multi-layer perception (MLP), random forest regression (RFR), gradient boosting regression (GBR) and XGBoost machine learning (ML) models were established using preprocessed data sets. Then, the Bayesian optimization method was used to optimize the hyperparameters of the RFR and XGBoost models with better performance. Therefore, the mechanical properties prediction model was established, and the impact of feature processing and PM parameters on the model was discussed. The results show that the BO-XGBoost model can effectively predict the mechanical properties of high-dimensional industrial big data and has excellent generalization ability (testing set $Er = 93.52\%$, $MAE = 13.56$ MPa, $RMSE = 20.19$ MPa), which is suitable for large-scale and industrial production of V–N steel hot-rolled plate.

INTRODUCTION

As an essential variety of steel materials, plate plays an irreplaceable role in national economic construction, people's lives and national security.^{1–3} V–N steel hot-rolled plate has realized mass industrial production. Compared with the conventional production process, the production process saves the RH vacuum refining, quenching, tempering and other heat treatment processes, dramatically reduces energy consumption, realizes green production and produces significant economic benefits.⁴ With the rapid development of computer technology in recent years, mass data generated in V–N steel hot-rolled plate production are collected and stored.

However, due to some plates' unqualified mechanical property data in the production process, the products cannot be put into regular use. Therefore, by establishing a mechanical property prediction model to mine the considerable research value existing in industrial big data, material science theory and big data can be effectively linked, and the law between the composition, structure, performance and process of the material can be revealed macroscopically. This provides practical guidance for optimizing the chemical composition and production process of products and improves the production efficiency of products.^{5,6} However, in the face of V–N steel hot-rolled plate products with a large amount of data, complex production process and various parameters, the traditional PM model will consume much time and economic costs and be

(Received September 9, 2022; accepted February 23, 2023;
published online March 21, 2023)

prone to failure. Thus, it is necessary to establish an efficient prediction model for the mechanical properties of the V–N steel hot-rolled plate.

Artificial intelligence technology has gradually emerged as an influential power for large-scale data processing, steel performance prediction, material composition design and process preparation.^{7–10} The mechanical properties prediction model based on the ML algorithm has achieved remarkable results and advantages in the field of big data analysis for the steel industry. Hore et al. established a prediction model of mechanical properties of hot-rolled TRIP steel by combining an adaptive neural network with a fuzzy inference system and obtained the relationship between coiling temperature, tensile strength, elongation and other properties and achieved good prediction results.¹¹ Qian et al. established the DNN model based on composition and process parameters to predict the yield strength (YS), tensile strength, elongation and impact energy of hot-rolled steel plate. It is applied to the factory's manufacturing execution system to assist the regular online production monitoring.¹² Wu et al. established the tensile strength prediction model of X70 pipeline steel based on stepwise regression (SR), ridge regression (RR), Bayesian regularization neural network (BRNN), radial-basis function neural network (RBFNN), support vector machine (SVM) and random forest (RF). They found that RF had a strong nonlinear fitting ability and achieved the best prediction effect ($R = 0.95$, $MSE = 278.7$ MPa2).¹³ Diao et al. constructed five ML models based on 97 kinds of carbon steel data and successfully predicted the tensile strength, fracture stress, impact energy, hardness, fatigue strength and elongation. They found that SVR and MLP models had the best performance in predicting mechanical properties.¹⁴ In addition, to solve the problem of low interpretability of the physical mechanism of the ML model, Li et al. introduced Ac_1 , Ac_3 temperature and deformation resistance parameters to guide the ML process and successfully established a new industrial big data analysis system combining ML classification and regression model with key PM variables.¹⁵ However, most of the above studies have carried out performance predictions for multiple steel grades, which cannot profoundly mine the complex functional relationship between the composition of each steel grade, process parameters and mechanical properties and clarify the influence mechanism of different parameters on the performance. So far, there is no report on modeling V–N steel hot-rolled plate based on the ML algorithm.

Therefore, the study aims to establish a prediction model of mechanical properties of V–N steel hot-rolled plate based on the ML algorithm. This study uses the standard YS index as an example to provide modeling ideas for other mechanical property prediction models. The carefully preprocessed data are added with PM parameters for model

training and testing. The efficiency ratio (Er), mean absolute error (MAE) and root mean square error (RMSE) are used to evaluate the model's prediction performance, and the best two models are selected according to the prediction effects of KNN, SVR, MLP, GBR, RFR and XGBoost. Then, the BO method is used to optimize the selected model's hyperparameters further to improve the model's prediction accuracy and generalization ability.

EXPERIMENTAL METHOD

Data Preprocessing

Data Cleaning and Processing of Missing and Abnormal Values

The research data of the V–N hot-rolled plate in this study were obtained from Laiwu Iron and Steel Group Co., Ltd., of Shandong Province, including chemical composition, rolling process and mechanical properties. The corresponding steel grade is Q550D, and the corresponding national standards are GB/T 1591–2008 and GB/T 16270–2009. The data mainly include chemical composition, rolling process and mechanical properties. First, the data were correlated by composition-process properties and arranged according to the production process. The samples and features with many null values and zero values were eliminated, and the unqualified samples were eliminated according to the corresponding national standards GB/T 1591–2008 and GB/T 16270–2009 of Q550D steel. Then, the Pauta criterion is used to detect and eliminate abnormal values.¹⁶ When the data are normally distributed, the Pauta criterion has an excellent processing effect,¹⁷ as shown in formula (1):

$$|v_i| = |x_i - x| > 3\sigma \quad (1)$$

where x_1, x_2, \dots, x_n are YS values in plate data, x is the mean value, $v_i = x_i - x$ ($i = 1, 2, \dots, n$) is the residual error, and σ is the standard deviation. If the residual error v_i ($1 \leq i \leq n$) of the YS x_i satisfies formula (1), that is, the YS $x_i \notin [\mu - 3\sigma, \mu + 3\sigma]$, the YS corresponding to the steel is considered abnormal data and eliminated.

In addition, a small number of missing values in the data set will bring additional complexity to the model. The missing values can be processed by filling, such as artificial, mean, EM, hot card and regression filling.^{18–20} The study used methods such as mean, median, mode, KNN interpolation and RFR filling to fill in the missing values in the dataset.

Feature Scaling

Feature scaling applied to features with an extensive numerical range can standardize their scale, so all variables can make comparable contributions to the model results and accelerate the training process, which positively impacts the model

effect.²¹ Therefore, the study uses the Z-score standardization method to scale features. The Z-score standardization formula is as follows:

$$x' = \frac{x - \mu}{\sigma} \quad (2)$$

where x is the original feature, x' is the standardized feature, and μ and σ are the mean and standard deviation of all samples of the original feature, respectively.

Feature Processing and PM Element Introduction

Feature selection is selecting a relevant feature subset for model construction and discarding some irrelevant features under minimum information loss. Feature selection can effectively simplify the model, make the model easy to understand, save storage and computing overhead and reduce the risk of overfitting.²² The study used XGBoost model importance evaluation and Pearson correlation analysis to select the input features. In the first step, the XGBoost model is used to evaluate the importance of features, and the features as segmentation points are calculated according to the gain of structure score. The importance of each feature is the number of times it appears in all trees. The more times a feature is used to build decision trees in the model, the higher its importance.²³ In the second step, the Pearson correlation coefficient is used to measure the degree of linear correlation between two features x_1 and x_2 and reduce the features. Pearson correlation coefficient is calculated as follows:

$$\gamma_{xy} = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^m (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^m (y_i - \bar{y})^2}} \quad (3)$$

where \bar{x} and \bar{y} are the averages of elements in vectors 1 and 2, respectively. γ_{xy} is the correlation between different variables; m is the number of data in the sequence. γ_{xy} is between -1 and 1, where 1 is completely positive and -1 is completely negative. When $|\gamma_{xy}| > 0.9$, it indicates that the two variables have a strong correlation.²⁴ Two highly correlated features are considered to contain similar information or have similar effects on the mechanical properties of V–N steel hot-rolled plates. Therefore, we build XGBoost models with two highly correlated features as input respectively and eliminate the features with poor model effects.

According to the characteristics of the V–N steel hot-rolled plate data set, this study introduced several PM parameters to guide the ML model. Considering that the phase transformation temperature is an essential reference factor for the formulation of the heat treatment process of V–N steel, the A_{c1} and A_{c3} temperatures that can reflect the microstructure of plates are introduced in the study

as new input features.²⁵ In addition, the reduction rate will affect the microstructure evolution of plates. Controlling the reduction rate in the hot-rolling process can effectively refine the grain size, make the precipitates dispersed and fine, and improve the strength and toughness of the steel. Therefore, according to slab, intermediate slab and rolling thicknesses, the reduction rate (ε_1) in the rough rolling process and reduction rate (ε_2) in the finish rolling process are calculated as other input features.

Based on the above data processing process, the input features selected for each category include chemical composition, rolling process and PM parameters. The output feature is YS. The descriptive statistical information of the above data is shown in Table I.

Machine Learning Model

Without loss of generality, six ML models were established to predict YS, and the most suitable method was found to solve the problem in the study, including KNN, SVR, MLP, RFR, GBR and XGBoost.^{26,27} In the above model, KNN finds k training samples closest to the training set based on a specific distance metric and predicts them by K 'neighbors' information. SVR aims is to obtain a model $f(x)$ that can fit the training set samples as much as possible. By constructing a loss function between the sample label and the model's predicted value, the loss function is minimized to determine the model $f(x)$.²⁸ In the study, RBF is selected as the kernel function of SVR. MLP, also known as artificial neural network (ANN), is composed of the input layer, hidden layer and output layer, and the layers are fully connected. In the training process, the model continuously adjusts the node weight according to the difference between the measured value and the predicted value by backpropagation to obtain a smaller loss function value.²⁹ GBR, XGBoost and RFR belong to the ensemble learning algorithm. Since ensemble learning completes the learning task by constructing multiple weak learners and combining specific strategies, the main advantage is that when one or more weak learners make error predictions, other weak learners can be corrected, so the ensemble learning algorithm often has a better fitting ability and generalization performance. At present, ensemble learning is mainly divided into two categories: bagging and boosting. RFR belongs to the bagging class, and GBR and XGBoost belong to boosting class.^{30,31} Compared with GBR, XGBoost adds regularization to the objective function, which is beneficial to reducing the variance of the model, simplifying the model and preventing overfitting. It has been successfully applied to air pollution and landslide.³²

For the problem of data set partitioning, the specific solutions are the hold-out, cross validation and bootstrapping method. In the study, the data

Table I. Range of input and output parameters in data sets

No	Feature	Mean	Minimum	Maximum	Standard deviation	No	Feature	Mean	Minimum	Maximum	Standard deviation
1	C (wt.%)	0.0895	0.058	0.136	0.0124	12	^d IBT (mm)	98.63	70	132	21
2	Si (wt.%)	0.1746	0.04	0.32	0.0434	13	^e FOT (°C)	927.37	829	1011	43
3	Mn (wt.%)	1.5350	1.17	1.72	0.0860	14	^f FRT (°C)	867.64	752	955	36
4	P (wt.%)	0.0125	0.007	0.019	0.0023	15	^g COT (°C)	641.11	405	768	81
5	S (wt.%)	0.0092	0.003	0.022	0.0035	16	^h CR (°C/s)	17.17	5	56	9
6	Cr (wt.%)	0.2121	0.01	0.25	0.0291	17	Ac ₁ (°C)	727.75	722	732	1
7	V (wt.%)	0.0913	0.011	0.111	0.0126	18	Ac ₃ (°C)	881.61	869	894	4
8	N (wt.%)	0.0127	0.0049	0.0273	0.0023	19	ϵ_1	0.610	0.470	0.72	0.086
9	^a HT (°C)	1191	1163	1232	8.1660	20	ϵ_2	0.660	0.500	0.76	0.067
10	^b ROT (°C)	1172	1113	1222	14.4118	21	YS(MPa)	640	521	782	46
11	^c RFT (°C)	1121	980	1171	14.2068						

^aHT is the heating temperature ^bROT is the roughing opening temperature ^cRFT is the roughing final temperature ^dIBT is the intermediate billet thickness ^eFOT is the finishing opening temperature ^fFRT is the finishing rolling temperature ^gCOT is the cooling outlet temperature ^hCR is the cooling rate.

set was randomly divided into a training set and a testing set in the ratio of 8:2 using the hold-out method. To further analyze different models and select appropriate models for prediction, three evaluation indexes, Er, MAE and RMSE, are used to measure the model's prediction accuracy.¹⁵ Er refers to the proportion of the total number of data relative to the total amount of data whose prediction deviation of V-N steel data points is < 6% of the average value of the target attribute. The formula of the three indexes is defined as follows:

$$Er = \frac{N_e}{N_{all}} \times 100\% \tag{4}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \tag{6}$$

where y_i and \hat{y} represent the measured and predicted values of plates' mechanical properties, respectively; n is the total number of sample data, N_e represents the total number of data within the specified error range, and N_{all} represents the total amount of data. The higher the Er value of the model is, the lower the MAE value and RMSE value, and the better the fitting effect of the model.

The SVR, MLP, GBR, RFR and XGBoost models used in the study were implemented based on the scikit-learn library in Python. The experiment was carried out on a Win10 system with 64 bits, Intel (R) Core (TM) i5—6200 U CPU @ 2.30 GHz (4 CPU) and running memory of 8 GB.

Bayesian Hyperparameter Optimization

BO is a global optimization algorithm based on probability distribution. It is assumed that there is a black box function $f(x)$ from the hyperparameter to a specific objective (such as the model performance index). BO infers the information of the black box function by collecting the hyperparameter (independent variable) and the corresponding model performance index (dependent variable). The purpose is to find the hyperparameter x^* with d dimensions in the hyperparameter space that optimizes the model performance index. The study takes finding the minimum value as an example, and the expression is as follows (7):

$$x^* = \arg \min f(x) \tag{7}$$

where $x \in X$, $X \subseteq R^d$, X is a hyperparameter space with d dimensions. The BO process utilizes the Bayes theorem. The core steps are the probability surrogate model and the acquisition function. The probability surrogate model approximates the black box function $f(x)$. The acquisition function selects the next

evaluation point that optimizes the objective function value according to the evaluated domain H to reach the optimal solution quickly. The Bayes theorem and the expression of the evaluated domain H are shown in (8) and (9) as follows:

$$p(f|H) = \frac{p(H|f)p(f)}{p(H)} \quad (8)$$

$$H = \{(x_1, f(x_1)), \dots, (x_n, f(x_n))\} \quad (9)$$

where $p(f)$ and $p(H|f)$ are the prior probability distribution and likelihood distribution of $f(x)$, respectively. $p(f|H)$ is the conditional probability distribution of $f(x)$ given the evaluated domain H , that is, the posterior probability distribution.

In the study, tree-structured Parzen estimator (TPE) is selected as the probability proxy model and expected improvement (EI) is selected as the acquisition function. The $p(H|f)$ in (8) is defined as follows:

$$p(x|y) = \begin{cases} l(x), & y < y^* \\ g(x), & y \geq y^* \end{cases} \quad (10)$$

where $l(x)$ is the density formed by using the observations $x^{(i)}$ such that corresponding loss $f(x^{(i)})$ was less than y^* and $g(x)$ is the density formed by using the remaining observations.

Optimization criteria based on improving EI:

$$\begin{aligned} EI_{y^*}(x) &= \int_{-\infty}^{y^*} (y^* - y)P(y|x)d_y \\ &= \int_{-\infty}^{y^*} (y^* - y) \frac{P(x|y)P(y)}{P(x)} d_y \end{aligned} \quad (11)$$

By construction, $\gamma = p(y < y^*)$ and $p(x) = \int_{\mathcal{R}} P(x|y)P(y)dy = \gamma l(x) + (1 - \gamma)g(x)$. Therefore:

$$\begin{aligned} \int_{-\infty}^{y^*} (y^* - y)p(x|y)p(y)d_y &= l(x) \int_{-\infty}^{y^*} (y^* - y)p(y)d_y \\ &= \gamma y^* l(x) - l(x) \int_{-\infty}^{y^*} yp(y)d_y \end{aligned} \quad (12)$$

So finally: $EI_{y^*}(x) = \frac{\gamma y^* l(x) - l(x) \int_{-\infty}^{y^*} yp(y)d_y}{\gamma l(x) + (1 - \gamma)g(x)} \propto \left(\gamma + \frac{g(x)}{l(x)}(1 - \gamma)\right)^{-1}$. This last expression shows that to maximize improvement, we would like points x with high probability under $l(x)$ and low probability under $g(x)$. The tree-structured form of l and g makes it easy to draw many candidates according to l and evaluate them according to $\frac{g(x)}{l(x)}$. On each iteration, the algorithm returns the candidate x^* with the greatest EI.³³⁻³⁶

In the study, the BO algorithm is implemented based on the Optuna library in Python, and the optimization process is shown in Fig. 1. The specific

implementation steps can be summarized as follows:

- (1) According to the determined hyperparameters to be optimized, the model hyperparameter optimization range is set, and the initial parameter combination is randomly generated within the parameter range.
- (2) Based on the acquisition function EI, the next superparameter combination sampling point x_i is selected from the TPE probabilistic surrogate model.
- (3) The value of the model's hyperparameters as a new combination of hyperparameters is input into the prediction model for training, and the prediction results of the model are obtained.
- (4) If the error of the new hyperparameter combination meets the accuracy requirement, the algorithm is terminated and exited, and the optimal hyperparameter combination is output. If the accuracy requirement is not met, $(x_i, f(x_i))$ is input into the TPE model to modify the acquisition function, and steps (2) and (3) are repeated to find the hyperparameter combination that meets the accuracy requirement and stops the iteration.

RESULTS AND DISCUSSION

Data Preprocessing Results and Discussions

Results and Discussion of Data Cleaning and Processing of Missing and Abnormal Values

In the study, 5021 samples were obtained through data association, and 3856 samples were obtained after deleting samples with many null values, redundant features and abnormal values. The cleaning effect of the above abnormal values is shown in Fig. 2a. The processed data are more intensive and concentrated. Then, the missing values were filled, and the filling effect is shown in Fig. 2 (based on the XGBoost model runs ten times). As shown in Fig. 2b and c, the KNN imputation and RFR imputation methods achieved better results on the training and test sets than other imputation methods. Because the KNN interpolation method uses the average of K nearest neighbors to fill in missing values, RFR can combine multiple decision trees to make a more accurate and stable prediction of missing values than a single tree. Among them, the RFR filling achieved the best filling effect. The training sets Er, MAE and RMSE were 99.99%, 3.63 MPa and 5.18 MPa, respectively. The testing sets Er, MAE and RMSE were 93.26%, 14.25 MPa and 21.41 MPa, respectively. Therefore, the study took the data after RFR filling as the subsequent modeling data.

Feature Processing Results and Discussion

The XGBoost model importance evaluation method was used to calculate the importance of

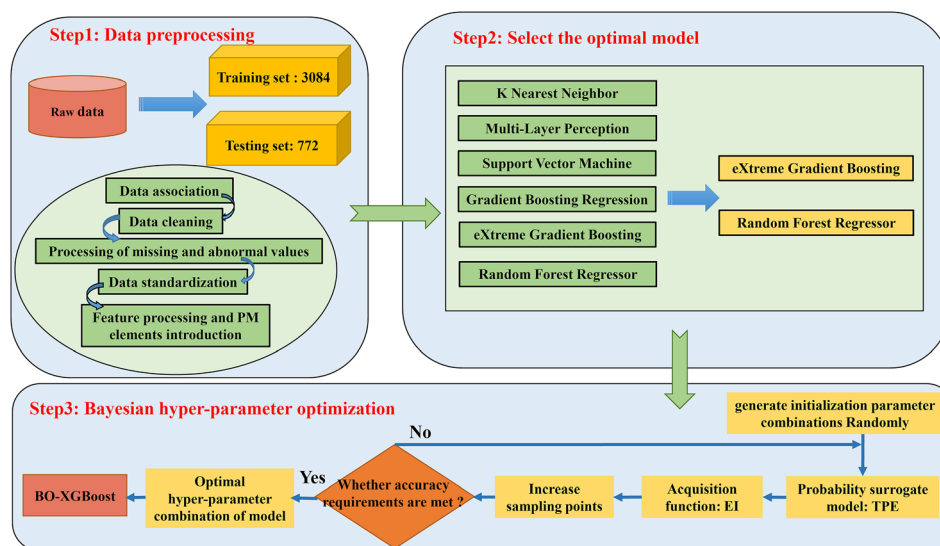


Fig. 1. Process of Bayesian hyperparameter optimization based on TPE.

input features to YS. To avoid the influence of randomness on the results, the data set was randomly divided ten times, and the obtained feature importance ranking is shown in Fig. 3. To ensure the generalization of the model, the features with an importance score < 90 , namely Cu, Nb, Mo, Al, B, Ti and Ni, were removed.

Then, the correlation coefficient value with an absolute value > 0.9 is shown in Fig. 4. When $|r_{xy}| > 0.9$, two features were used as input features respectively to establish two ML models. The RT and CIT features were removed by comparing the model's prediction accuracy. Based on the XGBoost model feature importance analysis and Pearson correlation analysis, the final input and output parameters are determined as shown in Table I in Sect. "Data preprocessing".

Figure 3 shows that the input features have different degrees of influence on the YS of the V–N steel hot-rolled plate with steel grade Q550D. The feature importance score of chemical composition from high to low is C, N, Si, Mn, Cr, P, S and V. C element can stabilize austenite and form an interstitial solid solution. Adding an appropriate amount of C element to the V–N steel of Q550D can form V (C, N), VN and other precipitates with V and N elements.³⁷ The precipitates can promote the formation of intra-grain ferrite (IF) in the austenite region and play the role of fine grain strengthening. In the ferrite region, fine dispersed particles can be formed to play the role of precipitation strengthening, thus significantly improving the YS of the steel. However, too high C content will reduce the low temperature toughness of steel and significantly destroy the weldability of materials. Therefore, by ultra-low carbon composition design, the C content is controlled between 0.058 wt.% and 0.136 wt.%. Both Si and Mn can improve the stability of austenite, thus improving the strength and

hardness of steel. However, too high Mn is prone to composition segregation, leading to too high hardness and toughness deviation. Cr is a cheap element that not only has the effect of solid solution strengthening and refining the organization but also can improve steel's hardenability, significantly improving steel's antioxidant effect. However, too much Cr has a greater tendency to temper brittleness. S and P are harmful elements in steel. S will cause serious segregation of FeS, resulting in steel cracking during hot working, which is a hot brittle phenomenon. P is mainly introduced by raw materials such as ore and pig iron. Although it has a significant strengthening effect, it will significantly reduce low temperature toughness, that is, cold brittleness. Therefore, in the actual production process, the content of S and P in steel needs to be strictly controlled in a lower range. V and N are the main microalloying elements of Q550D V–N steel. The non-metallic compound of V can form MnS + VN composite second phase with fine MnS particles in V–N steel to further promote the nucleation of IF.³⁸ On the one hand, precipitation strengthening improves the strength of the steel; on the other, acicular ferrite also greatly improves the material's toughness. V is abundant and inexpensive, and it has a strong affinity with N. N is a harmful element when it is free and can exist in the form of precipitates with V to improve the overall performance of the steel. Therefore, by designing the ratio of V and N, the steel can be guaranteed to obtain excellent comprehensive performance while greatly saving production costs.

For the rolling process, the microstructure of the steel will change continuously through the rolling process of heating, rough rolling, finishing rolling and cooling. HT, ROT, RFT and ε_1 will affect the morphology of the original austenite and austenite recrystallization behavior. As the temperature

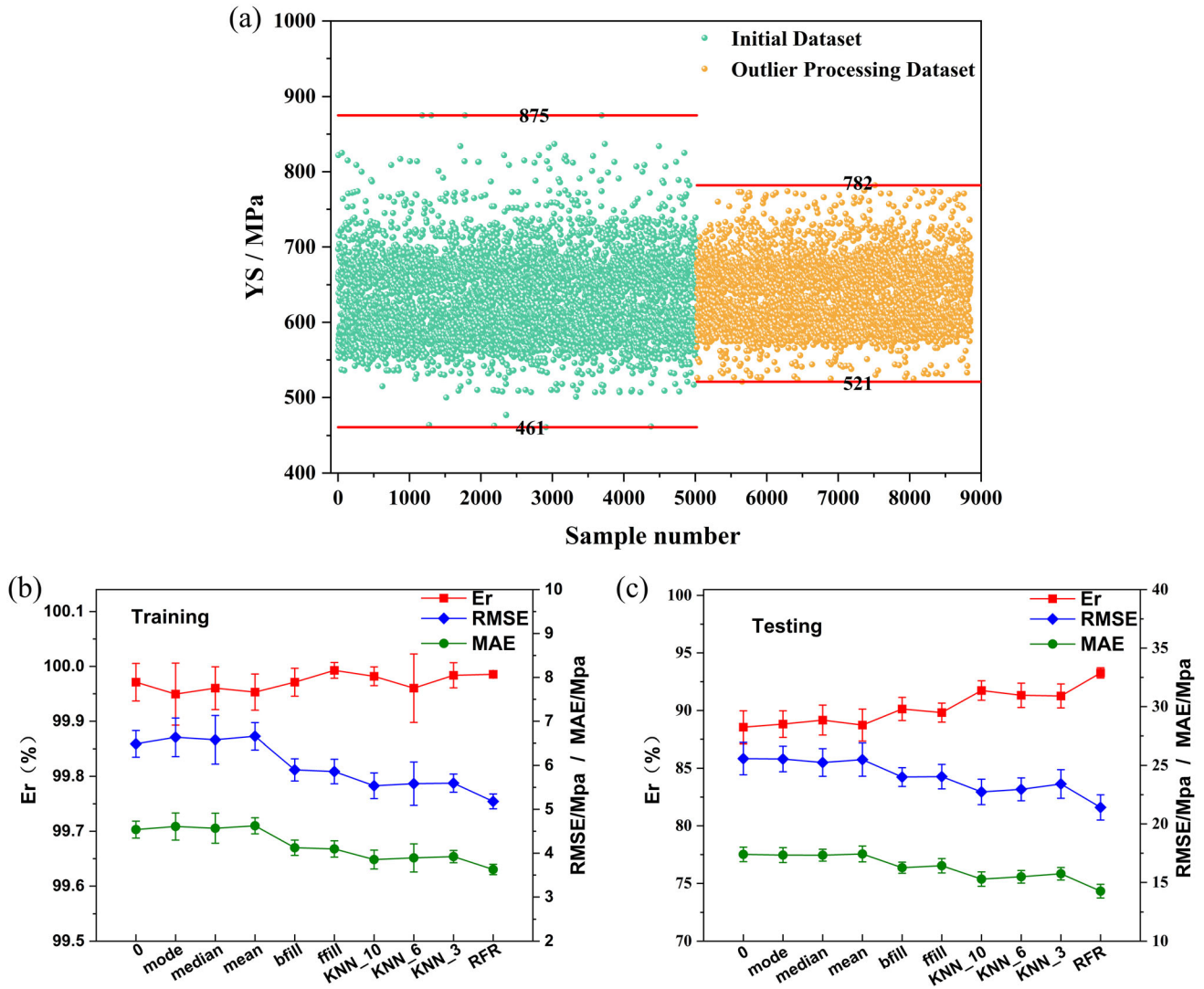


Fig. 2. Comparison of abnormal value elimination effect and effects of different missing value filling methods on modeling: (a) abnormal values elimination effect, (b) training set, (c) testing set.

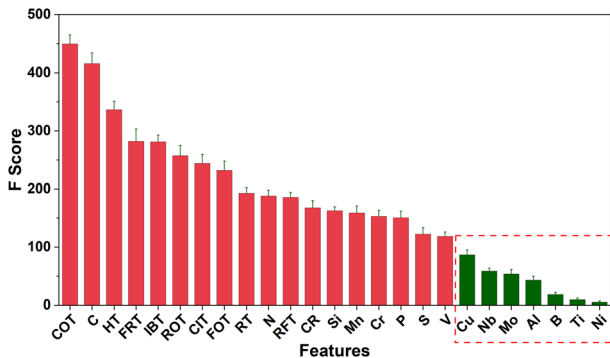


Fig. 3. Feature importance ranking based on XGBoost model.

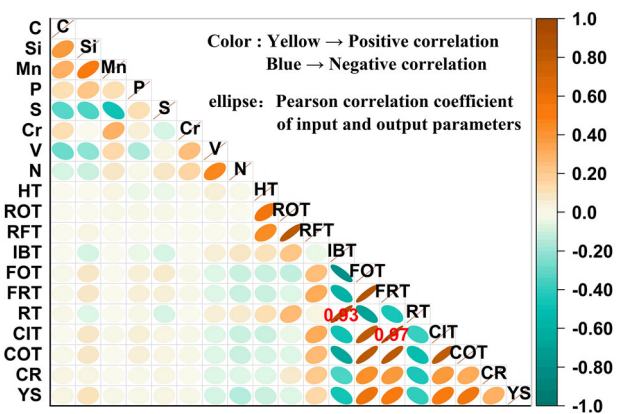


Fig. 4. Feature correlation analysis based on Pearson.

decreases, the austenite grains are gradually refined, increasing the overall austenite grain boundary area, dislocations, and substructures. FOT, FRT and ε_2 are essential factors affecting the deformation and phase transformation of V-N steel.

With the continuous decrease of temperature, plastic deformation leads to the elongation and flattening of austenite grains, the large increase of grain

boundary area and the generation of many deformation bands and strain-induced precipitates, which provide a good nucleation site for ferrite phase transformation and refine the ferrite grains. As the factor with the highest feature importance score in Fig. 3, COT will affect the microstructure refinement of austenite to ferrite transformation.³⁹ Controlling the COT can make the second phase of V–N steel fully precipitate and produce more acicular ferrite, polygonal ferrite and other microstructures, thereby improving the strength and toughness of steel.⁴⁰

A relatively higher correlation (corresponding to darker ellipses in the top right region of Fig. 4) is observed between YS and processing parameters and among processing parameters themselves. Some of it is expected since many processing parameters are inherently coupled (e.g., FRT and CIT). It shows the critical dependence of material properties on the rolling process through microstructure.⁷ Although all process parameters were highly correlated with YS, the most influential ones are FOT, FRT, CIT and COT. In particular, FOT shows a strong positive correlation with YS. Given the way the dataset was constructed, most of these reflect that performing one or more of these processing steps enhances the YS of V–N steel hot-rolled plate. IBT and RT are negatively correlated with the YS, indicating that the smaller the thickness, the worse the YS.

Analysis of Modeling Results

After data partitioning, there were 3084 samples in the training set for training the model and 772 samples in the test set for evaluating the effect and generalization of the model. Considering that the model pays more attention to improving model generalization performance in actual production, the model was evaluated by the prediction results on the testing set. Figure 5a used the histogram to compare the prediction effects of the six models in Sect. “Machine learning model”. The evaluation indexes Er, MAE and RMSE of the XGBoost and RFR models were much better than those of KNN, RBF-SVR and MLP models. The MAE and RMSE evaluation indexes of the GBR model were poor, but its Er was close to that of the XGBoost and RFR models. To show the model fitting effect, a scatter plot of the measured and predicted values of the six models were drawn. Figure 5b–g clearly shows the relationship among the measured value, predicted value and best-fitting line through the scatter plot.

In Fig. 5, the slope of the red line is 1, and the dashed black line is the boundary of the Er value. The more prediction points fall within the boundary range and are close to the line with slope 1, the better the fitting effect of the model is. Figure 5b–d indicates many prediction points outside the Er boundary line. The Er of the model was < 90%, and the MAE and RMSE were > 20 MPa and 30 MPa,

respectively, indicating the generalization ability of the KNN, RBF-SVR and MLP model was poor. The main reason for the poor generalization ability of the KNN model is that the industrial data dimension used in the study is too high. As the dimension increases, the distance between the two points in the KNN model tends to be larger. Compared with the KNN model, the RBF-SVR model can solve high-dimensional problems. However, the study’s poor generalization ability is mainly related to the parameter selection and the stability of the kernel function in high-dimensional space mapping. For the MLP model, the main reason for the poor prediction effect is that the appropriate artificial neural network parameters are not selected, such as the number of hidden_layer sizes and batch_size. The points in Fig. 5e were close to the reference line, and 90.42% of the prediction points were located within 6% of the average YS value of the testing set. However, in the range of YS 700–782 MPa, due to the small number of edge data, the frequency of occurrence in the training set was low, making the information in the data difficult to learn by the model. Therefore, with the decrease in data density, the prediction effect worsened. In Fig. 5f and g, > 90% of the data in the prediction results of the XGBoost and RFR models were located within the Er boundary, and the fitting slope was closer to 1. The MAE and RMSE were < 16 MPa and 23 MPa, respectively, indicating that the prediction effect was perfect. As described in Sect. “Machine learning model”, the ensemble learning algorithm has a strong generalization ability and better fitting effect for high-dimensional complex features. Therefore, XGBoost and RFR models were selected as the objects of BO algorithm optimization.

Hyperparameter Optimization Results Analysis

In the modeling process, the hyperparameters affecting RFR are mainly n_estimators, max_depth, min_samples_split, max_features, etc.⁴¹ The hyperparameters affecting the XGBoost model are mainly n_estimators, max_depth, min_child_weight, learning_rate, etc.⁴² The greater the number of n_estimators in the modeling process, the more information the model learns, but too many n_estimators will lead to increased running time and overfitting phenomenon, thus reducing the model prediction accuracy. If Max_depth is deeper, more feature attributes need to be divided in the modeling, and the corresponding model structure will be more complex. Therefore, the size of max_depth needs to be set with reference to how many feature attributes of the training data. The min_samples_split, which affects the random forest model, mainly restricts the conditions for further sub-tree division. If the number of samples at a node is less than the min_samples_split, no further attempts will be made to select the optimal features for division, while the max_features can control the

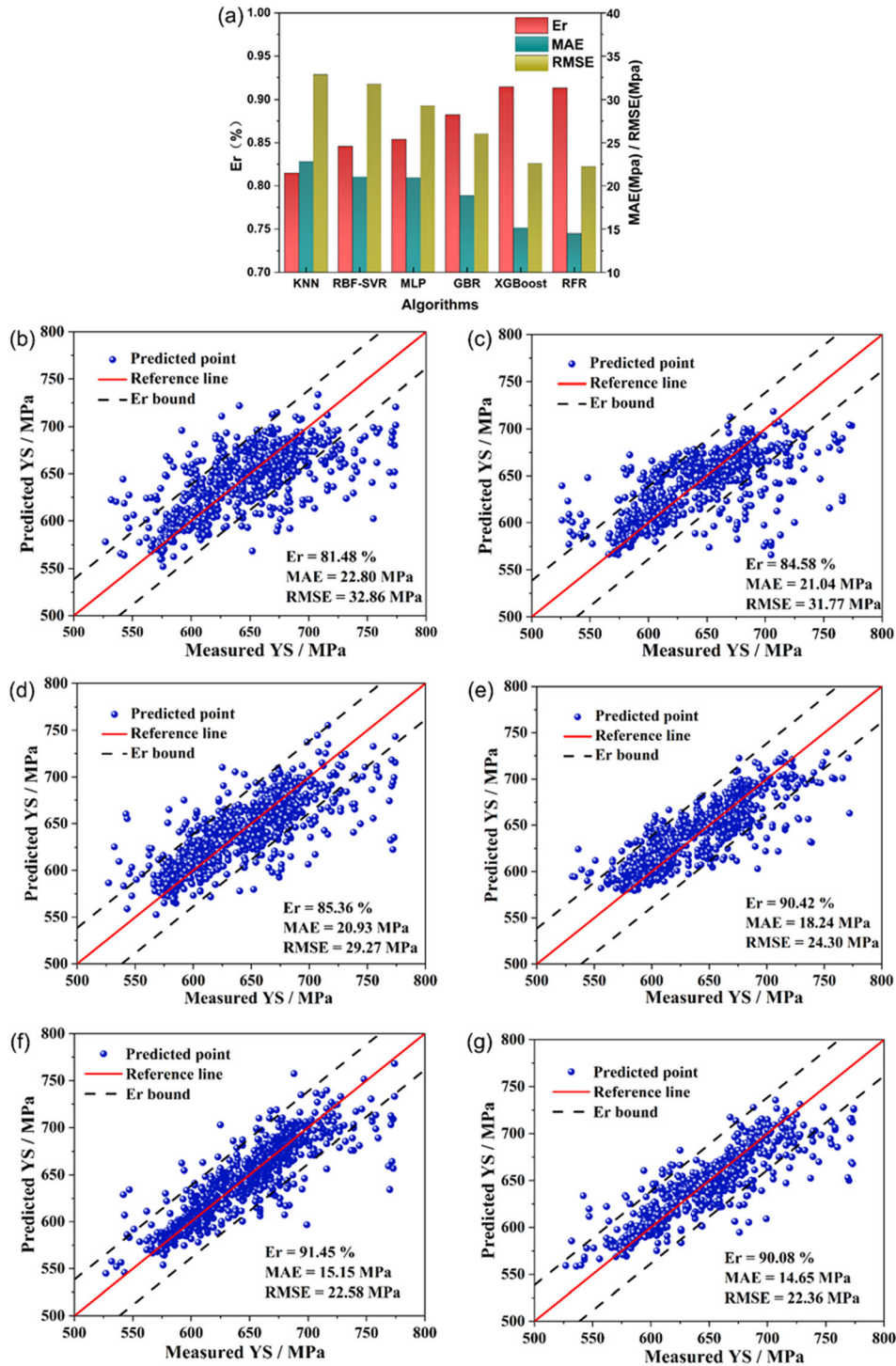


Fig. 5. Comparison of prediction results of KNN, RBF-SVR, MLP, GBR, XGBoost and RFR models and scatter plots of predicted and measured values distribution: (a) comparison results, (b)–(g) prediction results (Color figure online).

maximum number of features considered for the division to control the generation time of the decision tree. The smaller the `min_child_weight` affecting the XGBoost model, the easier the model is to overfit, while `learning_rate` controls the step size when updating the weights in each iteration, and the smaller the value, the slower the training. According to the description of the physical

mechanism of hyperparameters above, the main hyperparameter ranges and optimal hyperparameters of the RFR and XGBoost models optimized by the BO method in the study are shown in Table II. The hyperparameter range in Table II (default values for other hyperparameters) was adopted, and the hyperparameters obtained by optimization were used for modeling. The prediction effect of the

Table II. Optimized hyperparameters in three models

Algorithm	Hyperparameter	Range	Best hyperparameters
XGBoost	n_estimators	(100, 1000)	779
	max_depth	(1, 30)	7
	min_child_weight	(1, 50)	1
	learning_rate	(0.001, 0.2)	0.059
RFR	n_estimators	(100, 1000)	973
	max_depth	(1, 30)	29
	min_samples_split	(5, 300)	5
	max_features	(1, 20)	19

BO-RFR and BO-XGBoost models is shown in Fig. 6, and the points with different colors in Fig. 6b and c represent the absolute error size. Figure 6a compares the performance indexes of RFR, XGBoost, BO-RFR and BO-XGBoost models. It was found that BO made two models' Er increase by 0.29% and 2.27%, two models' MAE increase by 1.17% and 10.54%, respectively, and two models' RMSE increase by 1.27% and 10.54%, respectively. It showed that the BO algorithm had a better optimization effect on the XGBoost model and improved its accuracy, robustness and reliability because the BO algorithm can establish a probability model through the previous evaluation results of the objective function, find the value of the minimum objective function and avoid falling into the optimal local solution. In addition, as shown in Fig. 6b and c, the points of the BO-XGBoost model were more concentrated within the boundary line, while the BO-RFR model had multiple points with huge prediction absolute errors. Figure 6d, e shows that the relative error of the BO-XGBoost model was concentrated in a narrow interval. Finally, it was found that the BO-XGBoost model had higher performance indicators than the BO-RFR model. Er, MAE and RMSE were 2.70%, 6.40% and 8.35% higher, respectively. Therefore, the BO-XGBoost model had the best prediction effect and generalization performance for the YS modeling of V-N steel. The study made some data of V-N hot-rolled plate with steel number Q550D public and provided a GitHub link to test the BO-XGBoost model. The GitHub link is <https://github.com/Sstar126/data.git>. In the form of "black box function," the model directly realizes the efficient prediction of the YS of V-N steel hot-rolled plate. Compared with the traditional PM model, it has the advantages of convenience, high precision and stronger applicability. It can provide key and effective guidance information for the subsequent design of chemical composition and rolling process parameters and reduce the production of unqualified products.

Effects of Feature Processing and PM Parameters on Models

In the study, the input features were reduced from 25 to 16 dimensions by XGBoost model feature

importance analysis and Pearson correlation analysis, which greatly simplified the data structure. At the same time, four PM parameters, Ac_1 , Ac_3 , ε_1 and ε_2 , were introduced to guide the model establishment. In this section, three BO-XGBoost models were established using the original data set, the data set after feature processing and the data set after introducing PM parameters. The influence of feature processing and PM parameters on the model was illustrated by comparing the three models' evaluation indexes and modeling efficiency.

As shown in Fig. 7a, the feature-processed data set's performance was slightly worse than the original data set's performance on the BO-XGBoost model because the removed features contain helpful information for modeling. However, Fig. 7b shows that the modeling efficiency of the data set after feature processing on the BO-XGBoost model was improved by 58.07%. This case showed that feature processing could simplify the model and improve the modeling efficiency under slight information loss. When chemical composition information is removed, dimensionality is reduced, and material variability is eliminated, which provided excellent help for calculating the model and storing data in the steel production process. At the same time, the related elements removed in the dimension reduction process, such as B, Ti, Ni and other elements, have a low influence on the mechanical properties of the plate.⁴³ Through the guidance of data dimension reduction, the cost and design difficulty can be reduced in the actual material research and development. In addition, in combination with Fig. 7a and b, it could be found that the data set with the introduction of PM parameters increased the modeling time on the BO-XGBoost model by about 500 s compared with the data set after feature processing. However, its Er, MAE and RMSE were increased by 0.85%, 5.02% and 5.11%, respectively, indicating that the introduction of PM parameters improved the model's accuracy and increased the interpretability and scalability of the ML model. Ac_1 and Ac_3 are expected to provide guidance on the subsequent heat treatment of failed products, and ε_1 and ε_3 can provide assistance on the reduction of rolling passes.

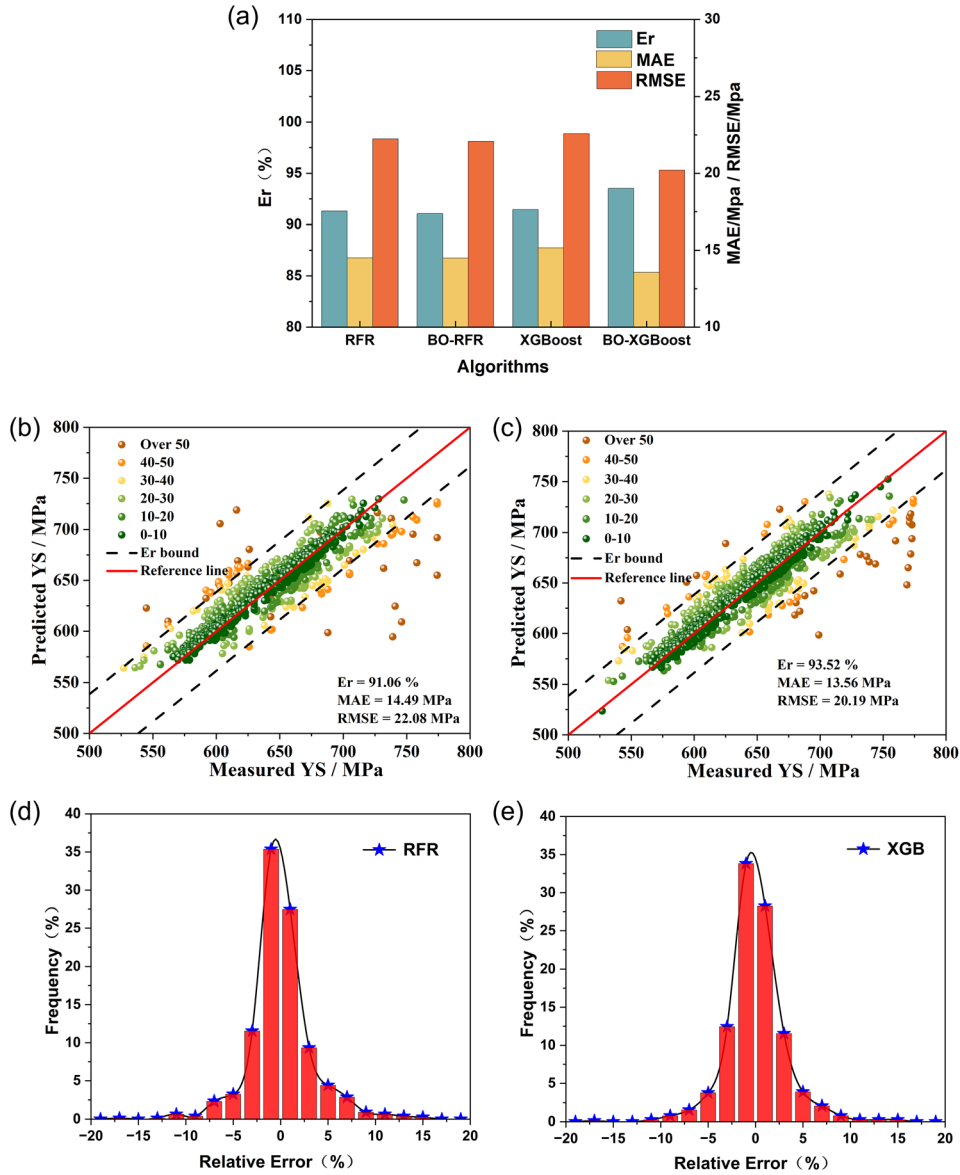


Fig. 6. Comparison of RFR, BO-RFR, XGBoost and BO-XGBoost models, scatter plot of error distribution and relative error frequency distribution of BO-RFR and BO-XGBoost: (a) comparison results, (b)–(e) prediction results (Color figure online).

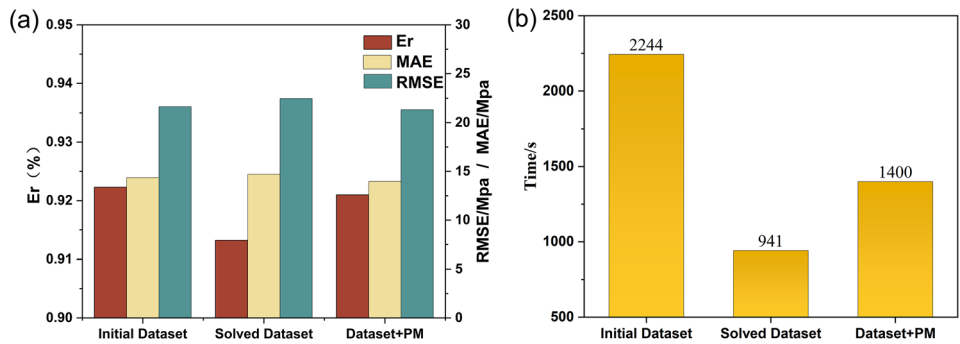


Fig. 7. Effect of feature processing and PM parameters on the model: (a) comparison results on the testing set, (b) comparison of modeling time.

CONCLUSION

A methodology was presented to train multiple ML models and perform hyperparameter optimization to predict YS using chemical composition and rolling process data of V–N steel hot-rolled plates. The framework presented includes industrial big data collection and cleaning, data dimensionality reduction and introduction of PM elements, multiple ML model building, hyperparameter optimization of ML models and evaluation of ML models. The data were obtained from the actual production process of domestic steel mills, and the high-dimensional data were reduced to low-dimensional data after cleaning and dimensionality reduction and passed to KNN, SVR, MLP, RFR, GBR and XGBoost models. The model was compared in two stages, including using six original models to ensure prediction results' comparability under the default parameters and using Bayesian optimization methods to find the hyperparameters of XGBoost and RFR models efficiently. The above two models integrated learning models that were widely used in related fields. Finally, the prediction results of BO-XGBoost and BO-RFR models were compared, and the effects of data dimensionality reduction and PM parameters on the modeling were discussed. The following key conclusions could be drawn from the above analysis.

- (1) Data preprocessing methods such as data cleaning, data normalization, Pauta criterion and RFR filling could significantly improve data quality and highlight data's regularity, which dramatically influenced data-driven modeling methods. In addition, the feature importance analysis based on the XGBoost model and Pearson correlation analysis reduced the dimensionality and introduced four physical metallurgical parameters, Ac_1 , Ac_3 , ε_1 , ε_2 , which could fully explore the guiding laws on PM.
- (2) The prediction and generalization of the XGBoost and RFR models were perfect without the hyperparameter search. The BO-XGBoost was the best model by Bayesian optimization ($Er = 93.52\%$, $MAE = 13.56$ MPa, $RMSE = 20.19$ MPa). One possible reason is that the BO-XGBoost model can effectively solve the modeling problem of high-dimensional data sets and learn the correlation between chemical composition-rolling process-yield strength.
- (3) Feature processing and PM parameter calculation were introduced into this study; the former could improve the modeling efficiency significantly, and the latter could increase the model's interpretability and effectiveness. Therefore, studying different feature processing methods and PM parameters is an essential element and direction for modeling steel data.

ML models for the actual production of V–N steel hot-rolled plates can provide a powerful way to predict the expected mechanical properties in a relatively fast way with the chemical composition parameters and rolling process parameters designed by the researchers, and more production data will augment the training data and facilitate the generalization of the model.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. D. Liu, B. Cheng, and Y. Chen, *Metall. and Mater. Trans. A*, 44, 440–455. <https://doi.org/10.1007/s11661-012-1389-9> (2013).
2. S.E. Hu, W.H. Sun, X.D. Liu, D.H. Hou, R. Zhou, F.Q. Xia, Y.Z. Liu, and G.D. Wang, *Mater. Sci. Forum* 773–774, 518–524. <https://doi.org/10.4028/www.scientific.net/MSF> (2014).
3. Z.-Q. Cao, Y.-P. Bao, Z.-H. Xia, D. Luo, A.-M. Guo, and K.-M. Wu, *Int. J. Min. Metall. Mater.* 17, 567–572. <https://doi.org/10.1007/s12613-010-0358-9> (2010).
4. H. Najafi, J. Rassizadehghani, and S. Asgari, *Mater. Sci. Eng., A* 486, 1–7. <https://doi.org/10.1016/j.msea.2007.08.057> (2008).
5. D. Shin, Y. Yamamoto, M.P. Brady, S. Lee, and J.A. Haynes, *Acta Mater.* 168, 321–330. <https://doi.org/10.1016/j.actamat.2019.02.017> (2019).
6. A. Kordijazi, T. Zhao, J. Zhang, K. Alrfou, and P. Rohatgi, *JOM* 73, 2060–2074. <https://doi.org/10.1007/s11837-021-04701-2> (2021).
7. A. Agrawal, and A. Choudhary, *Int. J. Fatigue* 113, 389–400. <https://doi.org/10.1016/j.ijfatigue.2018.04.017> (2018).
8. G. Carayannis, *JOM* 45, 43–51. <https://doi.org/10.1007/BF03222461> (1993).
9. J.-Y. Lee, M. Kim, and Y.-K. Lee, *Mater. Sci. Eng. A* 843, 143148. <https://doi.org/10.1016/j.msea.2022.143148> (2022).
10. C. Shen, C. Wang, X. Wei, Y. Li, S. van der Zwaag, and W. Xu, *Acta Mater.* 179, 201–214. <https://doi.org/10.1016/j.actamat.2019.08.033> (2019).
11. S. Hore, S.K. Das, S. Banerjee, and S. Mukherjee, *Iron-making Steelmaking* 44, 656–665. <https://doi.org/10.1080/03019233.2016.1227025> (2017).
12. Q. Xie, M. Suvarna, J. Li, X. Zhu, J. Cai, and X. Wang, *Mater. Des.* 197, 109201. <https://doi.org/10.1016/j.matdes.2020.109201> (2021).
13. S. Wu, J. Ren, X. Zhou, G. Cao, Z. Liu, and J. Yang, *Trans. Indian Inst. Met.* 72, 1277–1288. <https://doi.org/10.1007/s12666-019-01624-0> (2019).
14. Y. Diao, L. Yan, and K. Gao, *J. Mater. Sci. Technol.* 109, 86–93. <https://doi.org/10.1016/j.jmst.2021.09.004> (2022).
15. H. Li, Y. Li, J. Huang, C. Shen, C. Wang, T. Jing, Z. Liu, and W. Xu, *Steel Res. Int.* 93, 2100820. <https://doi.org/10.1002/srin.202100820> (2022).
16. H. Xiao, Y. Zhang, X. Liu, H. Yin, P. Liu, and D.C. Liu, *J. Phys. Conf. Ser.* 1769, 012009. <https://doi.org/10.1088/1742-6596/1769/1/012009> (2021).
17. J. Xia, J. Zhang, Y. Wang, L. Han, and H. Yan, *Pattern Recognit.* 121, 108177. <https://doi.org/10.1016/j.patcog.2021.108177> (2022).
18. Q. Yu, X. Guan, Y. Zhai, and Z. Meng, *IFAC-PapersOnLine* 53, 152–157. <https://doi.org/10.1016/j.ifacol.2021.04.094> (2020).
19. G. Lin, A. Lin, and D. Gu, *Inf. Sci.* 608, 517–531. <https://doi.org/10.1016/j.ins.2022.06.090> (2022).

20. R. Feng, D. Grana, and N. Balling, *Comput Geosci* 152, 104763. <https://doi.org/10.1016/j.cageo.2021.104763> (2021).
21. C.-W. Chu, J.D. Holliday, and P. Willett, *J. Chem. Inf. Model.* 49, 155–161. <https://doi.org/10.1021/ci800224h> (2009).
22. J. Xiong, S.-Q. Shi, and T.-Y. Zhang, *Mater. Des* 187, 108378. <https://doi.org/10.1016/j.matdes.2019.108378> (2020).
23. S. Ben Jabeur, N. Stef, and P. Carmona, *Comput. Econ. h* <https://doi.org/10.1007/s10614-021-10227-1> (2022).
24. H. Zhang, H. Fu, X. He, C. Wang, L. Jiang, L.-Q. Chen, and J. Xie, *Acta Mater.* 200, 803–810. <https://doi.org/10.1016/j.actamat.2020.09.068> (2020).
25. J. Trzaska, and L.A. Dobrzański, *J. Mater. Process. Technol.* 192–193, 504–510. <https://doi.org/10.1016/j.jmatprotec.2007.04.099> (2007).
26. K. Song, F. Yan, T. Ding, L. Gao, and S. Lu, *Comput. Mater. Sci.* 174, 109472. <https://doi.org/10.1016/j.commatsci.2019.109472> (2020).
27. M. Dissanayake, H. Nguyen, K. Poologanathan, G. Perampalam, I. Upasiri, H. Rajanayagam, and T. Suntharalingam, *Thin-Walled Struct.* 175, 109152. <https://doi.org/10.1016/j.tws.2022.109152> (2022).
28. S. Li, H. Fang, and X. Liu, *Expert Syst. Appl.* 91, 63–77. <https://doi.org/10.1016/j.eswa.2017.08.038> (2018).
29. F.P.V. Ferreira, R. Shamass, V. Limbachiya, K.D. Tsavdaridis, and C.H. Martins, *Thin-Walled Struct.* 170, 108592. <https://doi.org/10.1016/j.tws.2021.108592> (2022).
30. J. Yin, and N. Li, *Ore Geol. Rev.* 145, 104916. <https://doi.org/10.1016/j.oregeorev.2022.104916> (2022).
31. A. Abdulalim Alabdullah, M. Iqbal, M. Zahid, K. Khan, M. NasirAmin, and F.E. Jalal, *Construct. Build. Mater.* 345, 128296. <https://doi.org/10.1016/j.conbuildmat.2022.128296> (2022).
32. Y. Liu, K. Zhou, N. Zhang, and J. Wang, *Ore Geol. Rev.* 100, 133–147. <https://doi.org/10.1016/j.oregeorev.2017.04.029> (2018).
33. J.P. Srinivas, and R. Katarya, *Biomed. Signal Process Control* 73, 103456. <https://doi.org/10.1016/j.bspc.2021.103456> (2022).
34. R. Shi, X. Xu, J. Li, and Y. Li, *Appl. Soft Comput.* 109, 107538. <https://doi.org/10.1016/j.asoc.2021.107538> (2021).
35. J. Bergstra, R. Bardenet, B. Kégl, Y. Bengio, Algorithms for Hyper-Parameter Optimization, (2011).
36. Y. Xia, C. Liu, Y. Li, and N. Liu, *Expert Syst. Appl.* 78, 225–241. <https://doi.org/10.1016/j.eswa.2017.02.017> (2017).
37. F.P. Li, N. Li, X.L. Wang, and M.H. Liang, *Mater. Sci. Forum* 1035, 424–429. <https://doi.org/10.4028/www.scientific.net/MSF.1035.424> (2021).
38. J. Hu, L.-X. Du, J.-J. Wang, and C.-R. Gao, *Mater. Sci. Eng., A* 577, 161–168. <https://doi.org/10.1016/j.msea.2013.04.044> (2013).
39. F. Ishikawa, T. Takahashi, and T. Ochi, *Metall. Mater. Trans. A.* 25, 929–936. <https://doi.org/10.1007/BF02652268> (1994).
40. J.Q. Qing, B.Q. Wu, J.C. Wu, and Y. He, *Mater. Sci. Forum* 561–565, 45–48. <https://doi.org/10.4028/www.scientific.net/MSF.561-565.45> (2007).
41. Z. Zhao, N. Xiao, M. Shen, and J. Li, *Sci. Total Environ.* 842, 156867. <https://doi.org/10.1016/j.scitotenv.2022.156867> (2022).
42. V.-Q. Nguyen, V.-L. Tran, D.-D. Nguyen, S. Sadiq, and D. Park, *Transp. Geotech.* 37, 100878. <https://doi.org/10.1016/j.trgeo.2022.100878> (2022).
43. L. Qiao, Y. Liu, and J. Zhu, *Eng. Fracture Mech.* 235, 107105. <https://doi.org/10.1016/j.engfracmech.2020.107105> (2020).

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.