Computational Materials Science 160 (2019) 95–104

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

A predicting model for properties of steel using the industrial big data based on machine learning

Shun Guo[a], Jinxin Yu[b,a,*], Xingjun Liu[c,d], Cuiping Wang[b], Qingshan Jiang[a]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518000, PR China
[b] College of Materials and Fujian Provincial Key Laboratory of Materials Genome, Xiamen University, Xiamen, Fujian 361000, PR China
[c] State Key Laboratory of Advanced Welding and Joining, Harbin Institute of the Technology, Harbin, Heilongjiang 150001, PR China
[d] Institute of Materials Genome and Big Data, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong 518000, PR China

ARTICLE INFO

Keywords:
Big data
Machine learning
Regression
Steel properties
Nonlinear programming

ABSTRACT

Extracting the valuable information about the connections between the overall properties and the related factors from the industrial big data of materials is of significant interest to the materials engineering. At present, most data-driven approaches focus on building a relation model for a single property of the materials, where it may ignore the restrictive boundaries of other properties. In this paper, we propose a machine-learning-based method using nonlinear programming for multiple properties of the materials, and solve the problem by using the Interior Point Algorithm. The key idea is to take the mapping functions corresponding to the properties of the materials as the constraints of the nonlinear programming problem, thus it is capable of processing the restrictions of these properties. Moreover, with our method, the possible boundaries of these properties under certain conditions can be calculated. Experiments results on steel production data demonstrate the rationality and reliability of the proposed method.

1. Introduction

Machine learning methods have been widely used in materials design and engineering [1,2]. The design of the materials requires considering many influence factors, such as the chemical compositions, process parameters, microstructures [3]. Most of the relationships between these factors are nonlinear [4,5], where the traditional regression methods used in this field are not capable of dealing with such problems.

To accelerate the design of the materials, machine learning methods were firstly introduced in 2006 by Ceder et al. [6]. They used the machine-learning-based methods to predict the crystal structure, which is the arrangement of the atoms in materials. Crystal structure could be predicted with quantum mechanics, but it is very costly. It is challenge to identify the appreciate structure efficiently from a huge amount of possible structures for the traditional research methods [7]. However, with the help of the machine learning methods, this process could be accelerated significantly. After the Ceder's work, machine learning methods were gradually accepted and applied in materials researches, such as the studies about solar materials [8], the water photosplitting materials [9], the carbon capture and gas storage materials [10], the nuclear detection and scintillators materials [11], the topological insulators materials [12], the thermoelectric materials [13]. However, all the materials mentioned before are the functional material, which means that they have some special properties and their properties are mainly dominated by the chemical compositions. Moreover, the production flows of these materials are highly complicated and could not be synthesized through the industrial production.

Structural material [14], including steel, aluminum, and so on, is another type of materials. This kind of materials is widely used in many engineering areas and is produced in large quantities by factories. Unlike the functional material, the process parameters may have a major impact on the properties of structural material [15]. However, the influence mechanisms of the process parameters on the properties have not been studied clearly. It may be due to the reason that the number of the process parameters is too large to be studied via the experimental methods or the traditional regression methods.

Steel is one of the most important materials in human civilization, it has been used since the 19th century till nowadays [16]. The most important properties of steel are the strength and the plasticity, where the former decides the upper limit to the forces that can be applied on the materials while the largest deformations of the materials are determined by the latter. However, the improvements of the properties of steel are quite difficult. The properties of steel are mainly dependent on

* Corresponding author at: Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518000, PR China.
E-mail address: 20720170154799@stu.xmu.edu.cn (J. Yu).

https://doi.org/10.1016/j.commatsci.2018.12.056
Received 28 November 2018; Received in revised form 25 December 2018; Accepted 28 December 2018

the chemical compositions and the process parameters. The main chemical composition of steel is iron (Fe), carbon (C), and more than 10 kinds of alloying elements [17]. Moreover, the process parameters of steel are of vital complicated. For example, the phase transition phenomenon, which changes the microstructures and properties of steel, occurs during the heating process [18]. The most used heating processes of steel are annealing, tempering, quenching, and normalizing [19]. To obtain the optimal microstructure, the heating temperature, the heating time, and many other parameters need to be studied. Therefore, there would be over millions of candidates with different elements combinations and process parameters, and it is almost impossible to study all of them currently. Additionally, these influence factors should be chosen carefully to reduce the production costs of steel [20].

To improve the overall properties of steel, lots of works have been done to analyze the relationships between steel properties and related influence factors using informatics methods. The early studies applied simple linear-regression methods with experimental data. For instance, Liang et al. [21] studied the influence of the annealing temperature and the carbon content on the mechanical properties of the plate steel. The predicted results of their regression model agreed with the engineering experience well but they only studied two features (influence factors).

Singh et al. [22] studied the influence of the rolling process parameters on the properties of steel with the neural network analysis method, where their model gave some useful information. To improve their model, the genetic algorithm was applied to optimize the structure parameters of the neural network [23–25]. In the similar way, the chemical compositions [26,27] and process parameters [28] of steel were studied. In [29], this method was utilized to deal with the noisy data.

With the improvement of the data collection technology, the production data of the steel can be collected more efficiently. The data could include the contents of more than 20 elements and the complicated process parameters [30,31]. Agrawal et al. [32] predicted the fatigue strength of steel with this kind of data.

As the machine-learning-based method is capable of dealing with multiscale problems, the relationships between macro-features and micro-properties could be predicted by the method. Jha et al. [33] studied the linkages between processing parameters and microstructural characteristics of steel by coupled CALPHAD and the machine-learning-based method.

All these machine-learning-based methods can be viewed as a regression problem as follows:

$$\gamma_R = \arg\min_{\gamma_R} L(R(\gamma_R, \widetilde{X}), \widetilde{Y}) \tag{1}$$

where $R(\cdot)$ is the regression model (such as Neural Network [22,23]), γ_R is the parameters of $R(\cdot)$, and $L(\cdot)$ is the loss function; $\widetilde{X} \in R^{m\times d}$ represents m samples of steel data, where each sample has d features (such as element content or process parameter); $\widetilde{Y} \in R^{m\times 1}$ represents m values of the property corresponding to $\widetilde{X}$.

However, only considering one property of steel is not enough. Most of the ways that can improve the strength of steel would decrease the plasticity. Thus, it is important to balance two properties by choosing suitable chemical compositions (contents of different elements) and process parameters. In conclusion, the disadvantages of these studies are as follows:

1. The influence factors (features) that they studied were not enough;
2. Only considered single property of steel.

In this study, we propose a machine-learning-based method for predicting properties of steel with 27 related features. With our method, the possible boundaries of three properties of steel as well as the potential optimum chemical compositions and the process parameters with different alloy contents were calculated, where it would provide the guidance for designing the suitable chemical compositions and the process parameters. And in this way, the expected properties of steel would be obtained.

2. Method

In this section, to solve the problems described above, we present a novel model for predicting properties of the materials based on nonlinear programming. Different from the previous studies, which typically constructed the relation model between the chemical compositions (or the process parameters) and a single property of the materials, here, we focus on the multiple properties of the materials and transform the problem into a constrained nonlinear programming problem.

2.1. Dataset

Our method was applied on the steel production data, which was collected by the Shanghai Meishan Iron and Steel Corporation Ltd. of Bao Steel Group. The original data contained 65,288 samples, while there were 2151 samples missing necessary information or recorded incorrectly. Finally, 63,137 samples were chosen for studying and all of them had the follows characteristics: (1) the datasets had 27 influence factors (features), which were the process parameters and chemical compositions, as shown in Table 1; (2) Each sample had three properties, which were the yield strength (YS), the tensile strength (TS), and the elongation (EL) (plasticity). In addition, we normalized the data by transforming the values of all samples to zero mean and unit standard deviation.

The strength and plasticity of steel are strongly influenced by the content of the carbon, while the contents of other alloying elements would affect the properties of steel as well. To characterize the properties more conveniently, carbon equivalent, which is obtained through the statistics of large number of experimental tests, is widely used. Carbon equivalent is the combination of the contents of carbon and other alloying elements. Carbon equivalent has many types, such as Ceq, Ce, Pcm. In this study, our data provider used Ceq and Pcm. Thus, these features are used in our study. The formula of the Ceq and the Pcm are as follows [34]:

$$Ceq = C + Mn/6 + (Cr + Mo + V)/5 + (Ni + Cu)/15. \tag{2}$$

$$Pcm = C + Si/30 + Mn/20 + Cu/20 + Cr/20 + Mo/15 + V/10 + 5B. \tag{3}$$

Table 1
Features and their numbers of steel production data.

Number	Feature	Number	Feature
1	Furnace temperature	15	Titanium content (Ti)
2	Exist temperature	16	Boron content (B)
3	Annealing temperature	17	Tin content (Sn)
4	Thickness	18	Arsenic content (As)
5	Width	19	Zirconium content (Zr)
6	Sulfur content (S)	20	Calcium content (Ca)
7	Copper content (Cu)	21	Lead content (Pb)
8	Nickel content (Ni)	22	Ceq (Carbon Equivalent #1)
9	Chromium content (Cr)	23	Pcm (Carbon Equivalent #2)
10	Molybdenum content (Mo)	24	Antimony content (Sb)
11	Vanadium content (V)	25	Nitrogen content (N)
12	Niobium content (Nb)	26	Oxygen content (O)
13	Total Aluminum content (Al)	27	Tungsten content (W)
14	Acid soluble Aluminum content		

* Ceq and Pcm are two types of carbon equivalent. Carbon equivalent is the combination of the contents of carbon and other alloying elements, which is used to characterize the properties of steel. The definitions of the Ceq and the Pcm are shown in Eq. (2) and Eq. (3) [34].

2.2. Model definition

Let $X = [X_1, X_2, \cdots, X_d] \in R^d$ represents a set of values of the chemical compositions and the process parameters, $Y = [y_1, y_2, \cdots, y_n] \in R^n$ represents n properties of steel corresponding to X. And f_i represents the map from X to y_i. Given Y and f_i, the problem becomes that whether a X could be found to satisfy $f_i(X) = y_i$ for all $i = 1, 2, \cdots, n$. To this end, we defined the model as:

$$\begin{cases} \varphi(X) = \min_X(\rho(f_1(X) - y_1)) \\ s.\ t.\ \ f_2(X) = y_2, \\ \vdots \\ f_n(X) = y_n, \\ l^j(X) \leqslant 0, \quad j = 1, \cdots, m \\ h^1(X) = 0. \end{cases} \tag{4}$$

where $\rho(\cdot) \geq 0$ is a loss function, $l^j(\cdot)$ and $h^1(\cdot)$ are constraint functions about X (for example, $l^j(\cdot) = X_i - a_i$, $h^1(X) = \sum_k \alpha^k X_k$). In our experiments, we define $\rho(x) = x^2$, where other loss functions may also be used. Thus, the problem changes to solve the solution of $\varphi(X) = 0$. Moreover, if f_i is differentiable, the equation can be solved using the Interior Point Algorithm [35–38]. However, another problem is that how to determine the appropriate f_i. Given a group of samples of X and corresponding y_i, the problem can be regarded as a regression problem. In this study, we compared different regression models for the industrial steel production data (see Section 3.2) and found that relative high prediction performance can be obtained for all of them. Although tree-based models (such as Random Forest [39]) showed higher performance than other linear models, they are non-differentiable. Furthermore, due to the unavoidable error of the data (such as the measurement error and the instrument error), the model with higher accuracy may be overfitting.

To this end, we choose Ordinary Least Square (OLS) model [40] as the function f_i in our model and define as follows:

$$f_i:\ \beta_0^i + \sum_{j=1}^{d} \beta_j^i X_j = y_i \tag{5}$$

where β_0^i and β_j^i are the model parameters that can be obtained by minimizing the residual squared error of the training data. Substituting (5) in (4), we have:

$$\begin{cases} \varphi(X) = \min_X(\rho(\beta_0^1 + \sum_{i=1}^d \beta_i^1 - y_1)) \\ s.\ t.\ \ \beta_0^2 + \sum_{i=1}^d \beta_i^2 = y_2, \\ \vdots \\ \beta_0^n + \sum_{i=1}^d \beta_i^n = y_n, \\ l^j(X) \leqslant 0, \quad j = 1, \cdots, m \\ h^1(X) = 0. \end{cases} \tag{6}$$

2.3. Solution of the proposed model

To solve Eq. (6), we introduced the Interior Point Algorithm, and reformed Eq. (6) as the approximate problem:

$$\begin{cases} \delta(X, s) = \varphi(X) - u\sum_{j=1}^m In(s^j) \\ s.\ t.\ \ h^2(X) = 0, \\ \vdots \\ h^n(X) = 0, \\ l^j(X) + s^j = 0, \quad j = 1, \cdots, m \\ h^1(X) = 0. \end{cases} \tag{7}$$

where $u > 0$ is the barrier parameter and the slack variable $s = [s^1, \cdots, s^m]$ is assumed to be positive; where $h^k(X) = \beta_0^k + \sum_{i=1}^d \beta_i^k X_i - y_k, \quad k = 2, \cdots, n$.Note that if u converge to zero, the solution of Eq. (6) should normally converge to a stationary point.

To characterize the solution Eq. (7), we introduced its lagrangian as follows:

$$\zeta(X, s, \lambda, \lambda_g) = \delta(X, s) + \sum_{i=1}^{n} \lambda^i h^i(X) + \sum_{j=1}^{m} \lambda_g^j G^j(X) \tag{8}$$

where λ^i and λ_g^j are the Lagrange multipliers, $G^j(X) = l^j(X) + s^j$. Rather than solving Eq. (8) accurately, an approximate solution $(\widehat{X}, \hat{s})$ satisfying $E(\widehat{X}, \hat{s}; u) \leq \varepsilon$ would be obtained using iterations, where E is defined by:

$$E(X, s; u) = \max\left(\left\| \nabla\varphi(X) + \sum_i \lambda^i \nabla h^i(X) + \sum_j \lambda_g^j \nabla l^j(X) \right\|_\infty, \|S\lambda_g - u\Lambda\|_\infty, \|G(X)\|_\infty \right) \tag{9}$$

where $\Lambda = [1, \cdots, 1]^T$, $S = diag(s^1, \cdots, s^m)$. And in this definition, the vectors λ, λ_g are least squares multiplier estimates [38]. Note that the terms in Eq. (9) correspond to each of the equations of the Karush-Kuhn-Tucker (KKT) conditions. To guarantee the tolerance ε and u converge to zero, here, a simple strategy was used for reducing both ε and u in each iteration by introducing a constant factor $\vartheta \in (0, 1)$.

To solve Eq. (8) using iteration, one common practice is applying a sequential quadratic programming (SQP) method with trust regions [36], and the subproblem is formulated as:

$$\begin{cases} \min_{\Delta X, \Delta s}\ \nabla\varphi(X)^T \Delta X + \frac{1}{2}\Delta X^T \nabla_{XX}^2 \zeta(X, s, \lambda, \lambda_g)\Delta X - u\Lambda^T S^{-1}\Delta s + \frac{1}{2}\Delta s^T H(s)\Delta s \\ s.\ t.\ \ h(X) + J_h \Delta X = \varepsilon_h, \\ l(X) + s + J_l \Delta s + \Delta s = \varepsilon_l, \\ (\Delta X, \Delta s) \in R. \end{cases} \tag{10}$$

where $H(s)$ represents the Hessian of Eq. (7) about s or an approximation to it. J_h and J_l represent the Jacobian of the $h(X)$ and $l(X)$, respectively. Ideally, we would like the constrains in Eq. (10) to satisfy the linearized constraints, such as the $(\epsilon_h, \epsilon_l) = 0$. R defines the region around X where the linearized constraints would be good approximations to the problem. In this study, if $\nabla_{XX}^2 \zeta(X, s, \lambda, \lambda_g)$ is positive definite, we use the step in [38] for computation, where the solution $(\Delta X, \Delta s)$ satisfies the linear system as:

$$\begin{bmatrix} \nabla_{XX}^2 \zeta & 0 & J_h^T & J_l^T \\ 0 & H(s) & 0 & I \\ J_h & 0 & 0 & 0 \\ J_l & I & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta X \\ \Delta s \\ \lambda \\ \lambda_g \end{bmatrix} = \begin{bmatrix} -\nabla\varphi(X) \\ uS^{-1}\Lambda \\ -h(X) \\ -G(X) \end{bmatrix} \tag{11}$$

The Eq. (11) can be considered as a Newton iteration on the KKT conditions of Eq. (7). If $\nabla_{XX}^2 \zeta(X, s, \lambda, \lambda_g)$ is not positive definite, we use a preconditioned conjugate gradients algorithm [41] to solve the problem in Eq. (10), where we will not go into the details and refer the reader to [41,42].

3. Results and discussions

In this section, we conducted several experiments to evaluate our method and analyze the results generated by our method. All experiments are implemented with MATLAB (R2016a edition) and run in PC (Microsoft Windows 7, Intel Core (TM) i7-6700U, CPU 3.40 GHz, 16 GB of RAM). We used the MATLAB implementation ***fmincon*** for solving the nonlinear programming problem in our model. For other methods, we used the build-in functions of the MATLAB, including ***regress***, ***fitrsvm***, ***fitrtree.*** For Random Forest, we used the package of ‘RF_Reg_C’ of the MATLAB.

Table 2
Evolutions of different YS-predicted models.

Method	R^2	R	MAE	$RMSE$
Ordinary least square	0.8867	0.9416	30.1986	41.2999
Support vector machine	0.8737	0.9347	29.5711	43.6058
Regression tree	0.9086	0.9532	25.519	37.4626
Radom forest	0.9452	0.9722	19.4481	28.7189

3.1. Function selection and analysis

To determine the appropriate functions that mapping the features to the three properties of steel respectively, four widely used regression methods were chosen for the comparison purpose as follows:

1. Ordinary Least Square (OLS) [40]
2. Support Vector Machine (SVM) [43]
3. Regression Tree (RT) [44]
4. Random Forest (RF) [39]

We evaluated the prediction performance of these methods based on the three properties, where four metrics (the explained variance (R^2), the coefficient of correlation (R), the mean absolute error (MAE), and the root mean squared error ($RMSE$)) were employed for the evaluation and the results were shown in Table 2, Table 3 and Table 4, respectively.

In our experiments, 5-fold cross validation was used to obtain the averaged prediction performance. That is, we separated the original dataset into 5 equal sized subsets; one single subset was set as the testing dataset, while the remaining 4 subsets were set as the training dataset; the process would repeat 5 times to make sure each of the subsets used exactly once as the testing dataset. Since the results in the experiments were independent of the training dataset, the problem of overfitting would be alleviated.

From the results, we can observe that the tree-based models (the RF and the RT) perform better than the linear models (the OLS and the SVM), while the OLS perform better than the SVM.

As previously mentioned, the OLS models corresponding to the three properties were chosen as the f_i of Eq. (4) in this study due to the differentiability and relative high prediction performance of them. And the parameters of these models (such as the β_0^i and the β_j^i) were substituted in Eq. (6) for forming our model. Fig. 1 presents the scatter plots of the three OLS models with predicting and actual properties. Although other mapping functions may also be chosen, here, we focus on studying the relationship between features (the chemical compositions and the process parameters) and **multiple** properties with our method.

To further analyze the effect of a single feature on the corresponding property, based on the parameters of the three OLS models, we plotted Fig. 2. In Fig. 2, the relative importance of the features on the property were calculated and ranked according to the absolute values of the parameters of the OLS model. As observed, the relative importance of the feature with the number 22 (the Ceq content) ranked first for the YS and the TS, and ranked second for the EL. Since the Ceq is mainly dominated by the carbon content, and in the context of the metallurgical principles and the engineering practice, the carbon content is commonly regarded as the most important factor in the chemical compositions that influences the overall properties of steel, the result predicted by our model agrees well with this.

Table 3
Evolutions of different TS-predicted models.

Method	R^2	R	MAE	$RMSE$
Ordinary least square	0.941	0.9705	21.1555	29.804
Support vector machine	0.9308	0.9685	20.9151	30.5569
Regression tree	0.9484	0.9739	18.2244	28.0055
Radom forest	0.9692	0.9845	13.9336	21.5441

Table 4
Evolutions of different EL-predicted models.

Method	R^2	R	MAE	$RMSE$
Ordinary least square	0.7298	0.8542	3.5494	4.7262
Support vector machine	0.7302	0.8545	3.5282	4.7298
Regression tree	0.7236	0.8506	3.5767	4.9238
Radom forest	0.8259	0.9088	2.7523	3.7924

Table 5 summarized the information about the positive and negative correlations between all features and the three steel properties. There are several ways to improve the overall properties of steel, such as the fine grain size strengthening, the solid solution strengthening, and the precipitation strengthening. The phase transition point and the range of different phases, which are mainly influenced by the chemical compositions, are other two important factors affecting the overall properties of steel. However, one element could induce multiple strengthening mechanisms at the same time, and it is difficult to identify which one plays a decisive role. The positive and negative (the different colors in Fig. 2) of the relative importance of the features based on the OLS models could provide a direction to resolve the problem.

For example, most chemical compositions could lead to the fine grain size strengthening, the precipitation strengthening, and the solid solution strengthening at the same time in the process of steel making. In Fig. 2, N (the feature with the number 25) appears positive correlations with all properties (the YS, the TS, and the EL), which is consistent with the characteristics of the fine grain size strengthening. Thus, we tend to believe that this strengthening mechanism plays the dominant role for N. On the other hand, for most other elements (such as Cr, Nb and Ti), the correlations with strength (the YS and the TS) and plasticity (the EL) are one the contrary, which means that strength improvement caused by these elements would lead to the decrease of the plasticity, and this is in accord with the characteristics of the precipitation strengthening and the solid solution strengthening. Therefore, the precipitation strengthening or the solid solution strengthening could be viewed as a decisive role for these elements. Most of the influence directions of the features (see Table 5) predicted by our model are in accordance with the previous studies [45–48].

An interesting finding was that the influence directions of two Al related features were different. The total aluminum content (the feature with number 13) includes not only acid soluble aluminum (the feature with number 14) but also some compounds that cannot be dissolved by acid. Although the strengthening mechanism of Al is still unclear now, in the light of Table 5, the main strengthening mechanism of acid soluble aluminum tend to be fine grain size strengthening. To this end, it is necessary to increase the content ratio of acid soluble aluminum in the total aluminum content in the process of steel making for improving both strength and plasticity.

3.2. Model validation and analysis

Alloy content, which represents the sum of the chemical contents of all the elements except the Fe, C, O, N, and S in steel, is considered to be one of the major influence factors on the strength and the plasticity of steel. Therefore, we took it into account as a constraint condition in our model for studies. In addition, the contents of each element are limited according to their true maximum contents in iron. Similarly, the process parameters have boundaries due to the technical factors. To this end, in our model, the constraint functions were setting as:

$$l^j(X) = X * A_j - b_{up}^j, \quad or$$
$$l^j(X) = b_{low}^j - X * A_j, j = 1, 2, \cdots, m \tag{12}$$

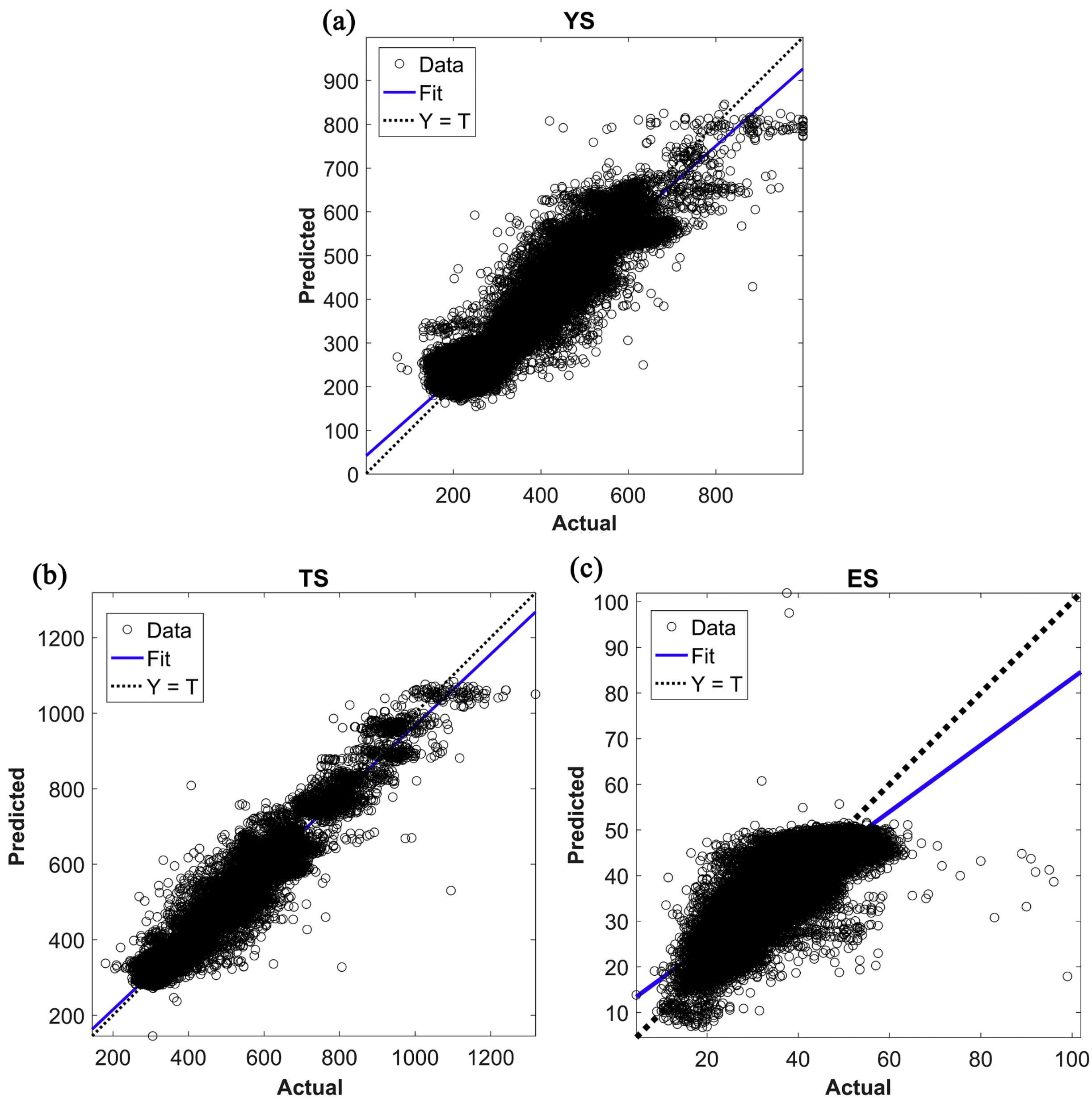

Fig. 1. 5-fold cross validation plots for the three OLS models with the predicting and actual properties. X-axis is the actual property and Y-axis is the predicting property, when the actual property equals to the predicting property, the data would distribute on the dot line.

where b_{up}^{j} and b_{low}^{j} were the upper boundary and lower boundary defined according to the boundaries of the original datasets for the constraint function; $A_j = [a_{j1}, a_{j2}, \cdots, a_{jd}]^T$, where $a_{ji} \in \{0, 1\}$ and it was used for selecting the features that should be limited.

In our experiments, without loss of generality, we predefined the YS/TS to be 0.8, as it is generally considered that the steel with this ratio would present good mechanical properties based on the engineering experience [49]. With the different boundary settings of the alloy content, we calculated the possible solutions of X in our model that given predefined properties of steel. Moreover, in the light of the value of $\varphi(X)$ in our model, we plotted the possible boundaries of the three properties responding to the alloy content, and the results were shown in Fig. 3. Ideally, if $\varphi(X) = 0$, it means that the solution X satisfies all constraint conditions in our model, and the errors for each property would be the same as that in the three OLS models (see Tables 2–4). In practice, the solution X of our model was chosen as the $\sqrt{\varphi(X)} \le 1e^{-3}$.

As it is shown in Fig. 3, the error limit that based on the $\sqrt{\varphi(X)}$ is marked in red, within which the area represents the boundaries of the three properties predicted by our method corresponding to the alloy content. For example, in Fig. 3(a), when EL is 40%, based on the limit error, the predicted YS that can be arrived are ranged from 200 to 450 MPa under the condition that the alloy content = 0.1%. It can also be observed that with the increase of alloy content, the area within the error limit becomes larger.

Fig. 4 shows the relationship between the proportion of the area and the alloy content. It can be found that the gradient of the curve in Fig. 4 drops and tends to be gentle with the increase of the alloy content, which indicates that the effect of the alloy content on the properties becomes weaker. As the increase of the alloy content would rise the cost of steel, the alloy content should be carefully chosen. Thus, this result may provide some insights about how to choose the appropriate alloy content under certain conditions.

The mechanical properties of steel are typically evaluated by the strength (the YS and the TS) and the plasticity (the EL). However, most

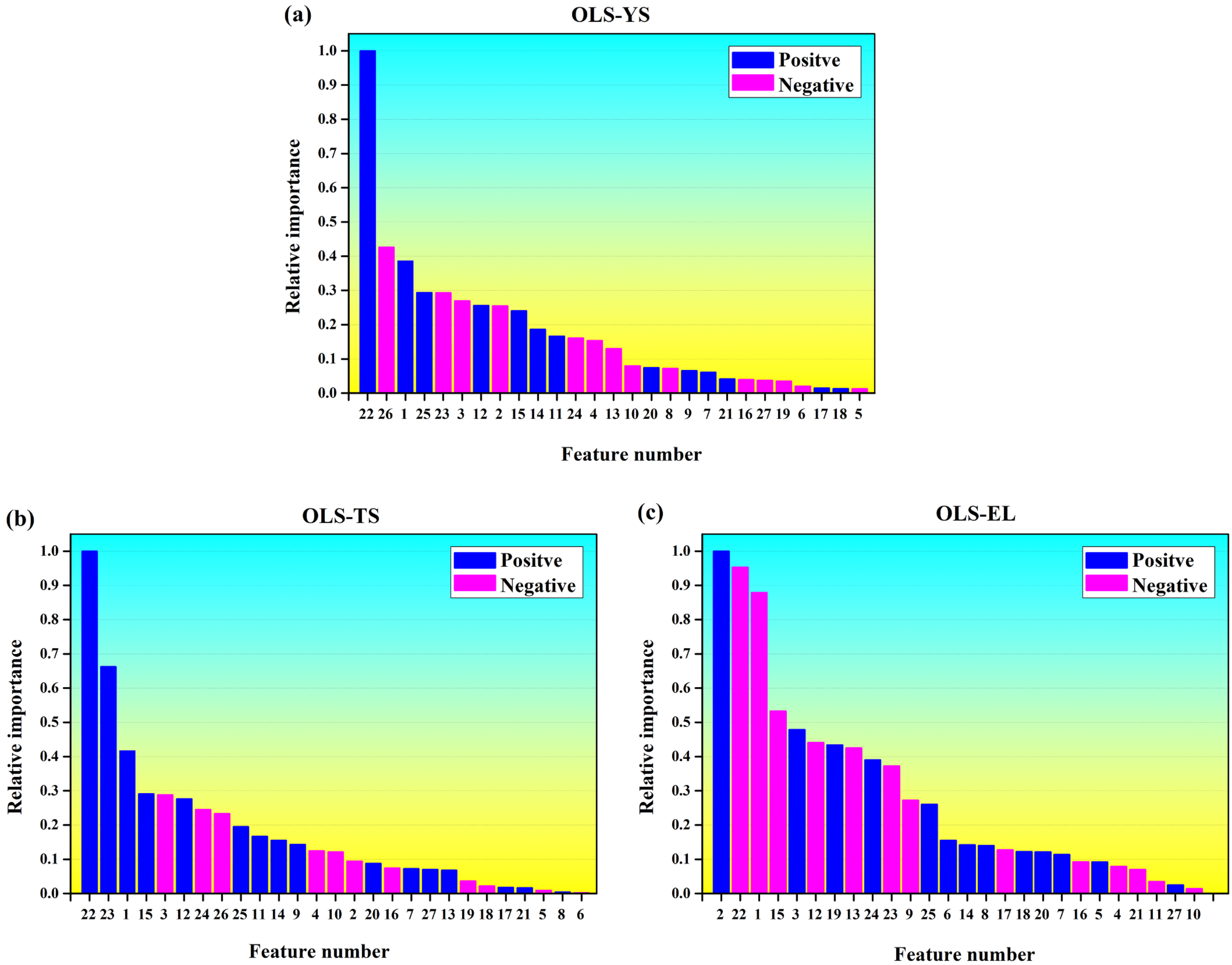

Fig. 2. Relative importance of the features on different properties based on the OLS model: (a) YS, (b) TS, (c) EL. Positive influence is blue and negative influence is purple.

of the ways that improving the strength of steel would lead to the decrease of the plasticity. Therefore, it is important to optimize both of them at the same time, that is, obtaining the upper boundaries of them under the restriction. Our model can also be viewed as the equation as follows:

$$(X, \varepsilon) = f_m(Y_1, Y_2, Y_3, \alpha) \tag{13}$$

where X is the features (the solution of the model), ε is the error of X, Y_1, Y_2 and Y_3 are YS, TS and EL respectively, and α is the alloy content.

Firstly, we set $Y_3 = y_3$, $\alpha = a$ and fixed the value. Then, given $Y_2 = y_2$, $Y_1 = 0.8 * y_2$, the X and ε could be obtained with these settings. If ε is smaller than the error limit, the values of Y_1 and Y_2 would be increased and induced into Eq. (13) to calculate new X and ε. This process would repeat until ε is larger or equal to the error limit, which also means that the upper boundaries of Y_1 and Y_2 had been found under the constraints ($Y_3 = y_3$, $\alpha = a$). Similarly, with the different settings of Y_3 and α, the corresponding upper boundaries of Y_1 and Y_2 would be obtained. In this way, we plotted Fig. 5, which reflected the predicted upper boundaries of the properties of steel with different alloy contents.

As it is observed from Fig. 5, in general, with the increment of alloy content, the upper boundaries of the properties of steel would be gradually raised. Meanwhile, as the value of EL exceeds the critical point, the upper boundaries of the YS and the TS would decrease with the increase of the EL. These results agree with the engineering experience and the materials knowledge well.

The chemical compositions and the process parameters can be designed more efficiently with the help of Fig. 5. For example, to design a new kind of steel with 35% EL (Line 1) in Fig. 5, the influence of the

Table 5
Influence directions of the features on three properties of steel based on the OLS model.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
YS	+	−	−	−	−	+	+	−	+	−	+	+	−	+	+	−	+	+	−	+	+	+	−	−	+	−	−
TS	+	−	−	−	−	+	+	+	+	−	+	+	+	+	+	−	+	−	−	+	+	+	+	−	+	−	+
EL	−	+	+	−	+	−	+	+	−	+	−	−	−	+	−	−	−	+	+	+	−	−	−	+	+	−	+

"+" means the relative importance of the feature is positive while "−" means negative.

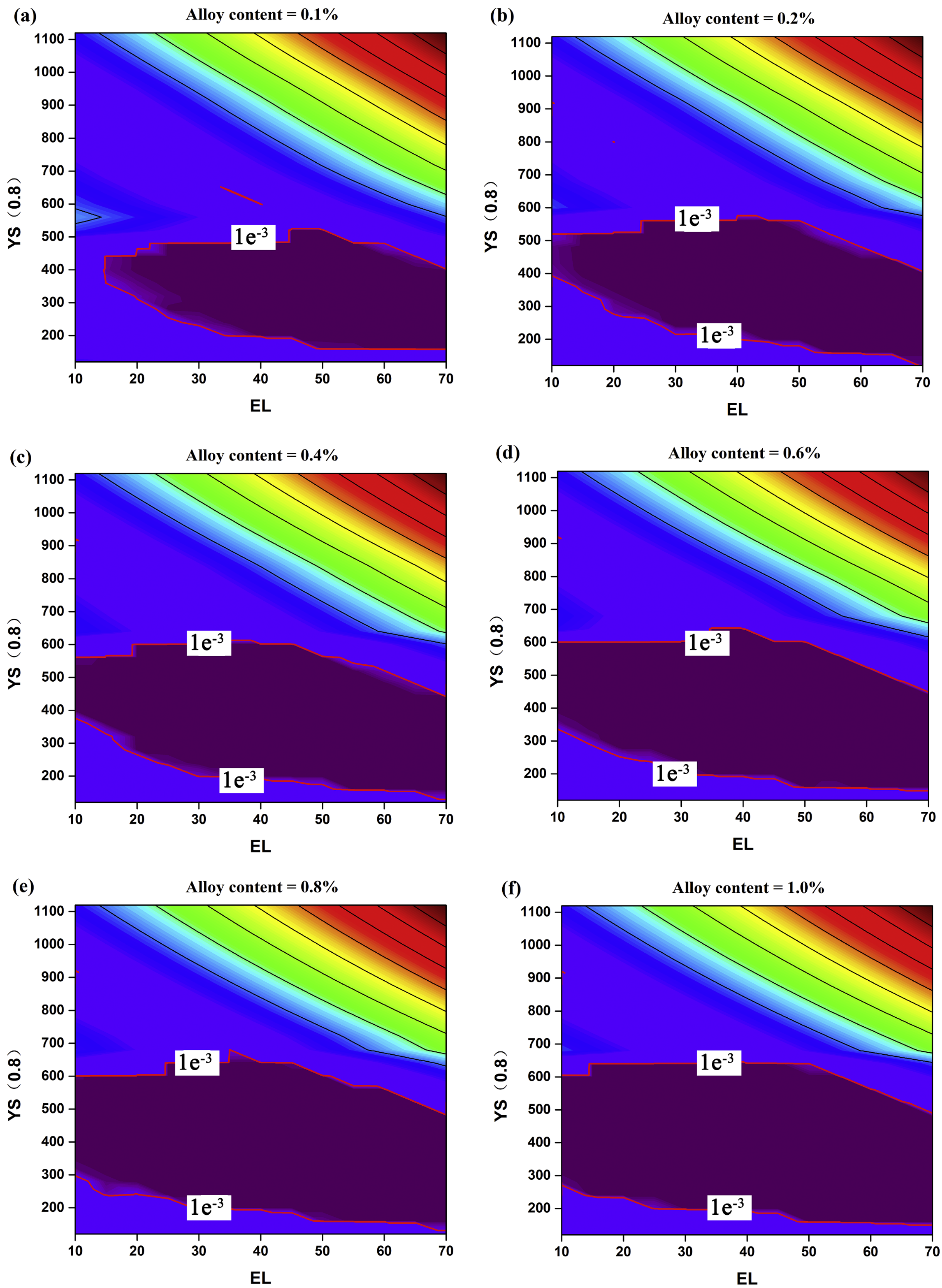

Fig. 3. The predicted boundaries of the three properties of steel with different alloy contents. The error limit that based on the $\sqrt{\varphi(X)} \leq 1e^{-3}$ is marked in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

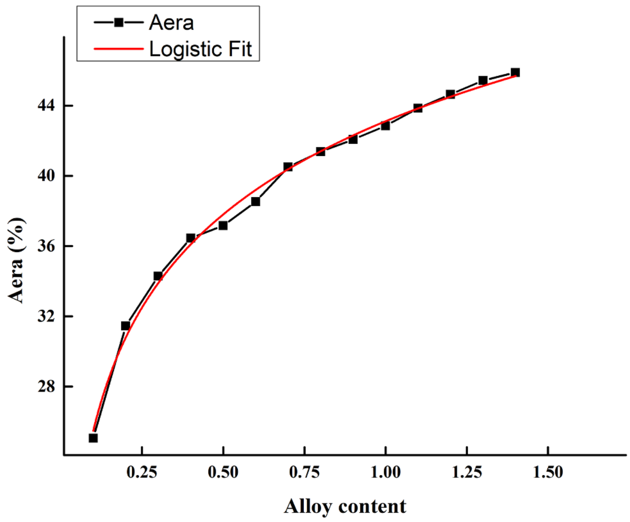

Fig. 4. The proportion of the area within the error limit. The gradient of the curve drops and tends to be gentle with the increase of the alloy content, which indicates that the effect of the alloy content on the properties becomes weaker.

alloy content on the strength is extracted (see Fig. 6(a)). A suitable alloy content could be chosen to fit the requirement of the application environment. We also studied the line 2 in Fig. 5 and plotted Fig. 6(b), which reflected the trend of the EL with different alloy contents under the constraints (YS = 600 MPa and YS/TS = 0.8). It can be observed that, based on the constraint conditions, as the alloy content ranges from 0.36% to 0.8%, the EL may have two values. The reason may be due to the existence of the critical point of the EL. That is, with the same alloy content, the upper boundaries of the YS and the TS may increase with the EL before the critical point, but decrease after that. To this end, even though the alloy content, the YS and the TS are the same, the EL could be different. From the aspect of our model, it can be interpreted as that given the same values of Y_1, Y_2 and α but different values of Y_3, two X would be obtained according to Eq. (13), where both of them satisfy the error limit. In this regard, the EL of steel could be improved with the appreciate X.

Carbon content is one of the most important influence factors on the overall properties of steel, so it is necessary to study its specific impact. As mentioned before, the main constitutes of Ceq is carbon. Therefore, based on our model, the relationships between upper boundaries of steel strength (YS/TS was set to be 0.8) and Ceq content with different alloy contents were calculated and plotted in Fig. 7. Based on the previous study [49], 1% carbon would increase 780 MPa YS, where it is predicted to be 800 MPa by our model. This result also verifies the reliability of the model.

4. Conclusions

To accelerate the development of new and cost-effective materials, it is important to mine the information about the relationships between the chemical compositions, the processing parameters and the overall properties of the materials. Most previous studies focus on constructing a relation model with one single property of the materials that best fitting the data, where it could be considered as a regression problem. However, the design of the materials requires considering multiple properties of the materials. Moreover, the improvement of one property would decrease another property at most time. In this regard, additional model that can balance the requirement of different properties is required. In this study, we transformed the problem into a nonlinear programming problem, where most multiple properties of materials were treated as the constraint conditions in our model.

To demonstrate the utility of our method, we applied our model on the steel production data with 27 features, where the effects of the related factors on the three properties of steel were studied. Firstly, three OLS regression models corresponding to the three properties were constructed, where all of them can achieve relatively good prediction performance. Then, based on them, our model was built using nonlinear programming. With the help of the model, the possible boundaries of the three properties under different alloy contents were evaluated. The results reflected the trends of three properties with the alloy content, which broadly in line with the engineering experience and materials knowledge. Specifically, the quantitative impact of carbon content on the strength of steel that studied from our model was quite consistent with the result of the previous study, which indicated the reliability and practicability of our model.

The error of our model major depends on the functions that chosen for each property of the materials, thus how to determine the appropriate mapping function for the corresponding property of the materials and apply our method to other industrial data would be another avenue

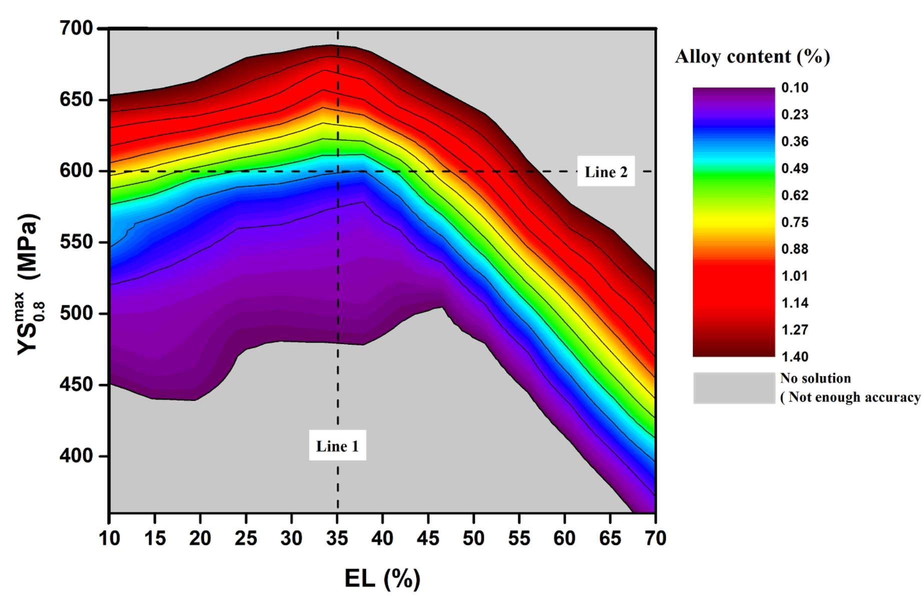

Fig. 5. The predicted upper boundaries of the properties of steel with different ally contents. The chemical compositions and the process parameters can be designed more efficiently with the help of this map.

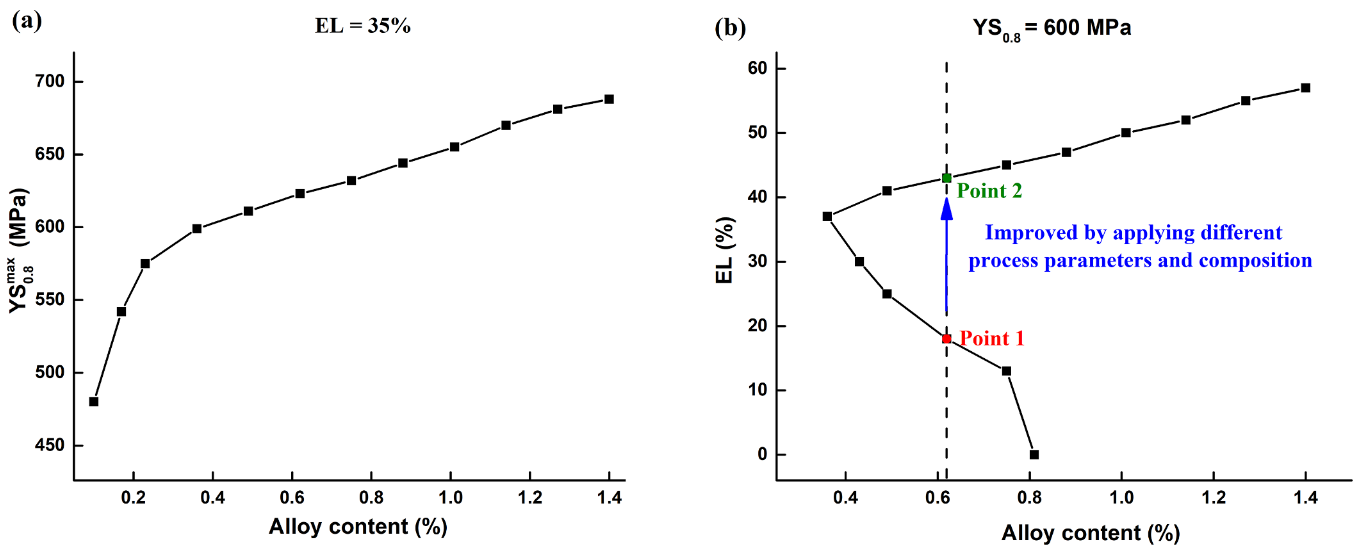

Fig. 6. Line 1 and Line 2 in Fig. 5, which is the steel with 35% EL (Line 1) and the steel with 600 MPa, separately.

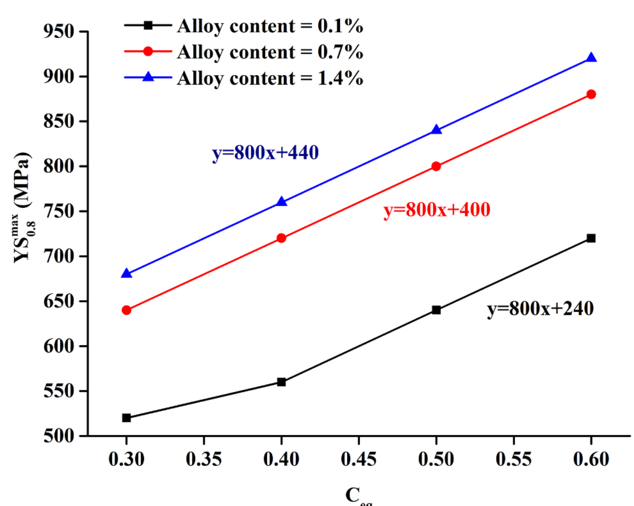

Fig. 7. The predicted influence of the Ceq on the strength of steel with different alloy contents. 1% carbon would increase 800 MPa according to our model.

to explore in our further study.

5. Data availability

The raw data rand the processed data required to reproduce these findings have been uploaded to https://data.mendeley.com/datasets/msf6jzm52g/draft?a=95752b76-3d1b-4568-97e9-cb5d6ff20bd5.

CRediT authorship contribution statement

Shun Guo: Formal analysis, Methodology, Software, Writing - original draft, Writing - review & editing. **Jinxin Yu:** Data curation, Formal analysis, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing. **Xingjun Liu:** Conceptualization, Investigation, Project administration, Resources, Supervision. **Qingshan Jiang:** Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision.

Acknowledgement

This research work was supported by Shenzhen Discipline Construction Project for Urban Computing and Data Intelligence, Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, China Postdoctoral Science Foundation Grant (No. 2018M633187), National Key R&D Program of China (No. 2017YFB0702901).

References

[1] P. Raccuglia, K.C. Elbert, P.D. Adler, C. Falk, M.B. Wenny, A. Mollo, M. Zeller, S.A. Friedler, J. Schrier, A.J. Norquist, Machine-learning-assisted materials discovery using failed experiments, Nature 533 (2016) 73–76.
[2] N. Nosengo, Can artificial intelligence create the next wonder material? Nature 533 (2016) 22–25.
[3] W.D. Callister Jr, Materials science and engineering - an introduction, Anti-Corros. Methods Mater. (2000).
[4] K.L. Reifsnider, V. Tamuzs, S. Ogihara, On nonlinear behavior in brittle heterogeneous materials, Compos. Sci. Technol. 66 (2006) 2473–2478.
[5] I.R. Peterson, Organic materials for nonlinear optics, Angew. Chem. Int. Ed. 100 (2010) 1257–1258.
[6] C.C. Fischer, K.J. Tibbetts, D. Morgan, G. Ceder, Predicting crystal structure by merging data mining with quantum mechanics, Nat. Mater. 5 (2006) 641–646.
[7] A.R. Oganov, C.W. Glass, Crystal structure prediction using ab initio evolutionary techniques: Principles and applications, J. Chem. Phys. 124 (2006) 201–419.
[8] L.P. Yu, A. Zunger, Identification of potential photovoltaic absorbers based on first-principles spectroscopic screening of materials, Phys. Rev. Lett. 108 (2012) 068701.
[9] I.E. Castelli, T. Olsen, S. Datta, D.D. Landis, S. Dahl, K.S. Thygesen, K.W. Jacobsen, Computational screening of perovskite metal oxides for optimal solar light capture, Energy Environ. Sci. 5 (2012) 5814–5819.
[10] L.C. Lin, A.H. Berger, R.L. Martin, J. Kim, J.A. Swisher, K. Jariwala, C.H. Rycroft, A.S. Bhown, M.W. Deem, M. Haranczyk, B. Smith, In silico screening of carbon-capture materials, Nat. Mater. 11 (2012) 633–641.
[11] C. Ortiz, O. Eriksson, M. Klintenberg, Data mining and accelerated electronic structure theory as a tool in the search for new functional materials, Comput. Mater. Sci. 44 (2009) 1042–1049.
[12] H. Lin, L.A. Wray, Y. Xia, S. Xu, S. Jia, R.J. Cava, A. Bansil, M.Z. Hasan, Half-heusler ternary compounds as new multifunctional experimental platforms for topological quantum phenomena, Nat. Mater. 9 (2010) 546-540.
[13] G.D. Mahan, J.O. Sofo, The best thermoelectric, Proc. Natl. Acad. Sci. USA 93 (1996) 7436–7439.
[14] W.J. Lackey, Structural Materials, Wiley-VCH Verlag GmbH, 2007.
[15] J.C. Williams, E.A. Starke Jr., Progress in structural materials for aerospace systems, Acta Mater. 51 (2003) 5775–5799.
[16] W. Gowland, The metallurgy of iron and steel, Nature 52 (1923) 613–614.
[17] J.W. Simmons, Overview: high-nitrogen alloying of stainless steels, Mater. Sci. Eng. A 207 (1996) 159–169.
[18] J.B. Leblond, G. Mottet, J.C. Devaux, A theoretical and numerical approach to the plastic behaviour of steels during phase transformations—ii, J. Mech. Phys. Solids 34 (1986) 411–432.
[19] R.J. Emmerling, Steel Metallurgy - Properties, Specifications and Applications, McGraw-Hill, 2015.
[20] C. Wiklund, M. Helle, T. Kohl, M. Järvinen, H. Saxén, Feasibility study of woody-biomass use in a steel plant through process integration, J. Clean. Prod. 142 (2017) 4127–4141.
[21] J.L. Liang, G.D. Luo, H.X. Zhang, Effects of rollling temperature and carbon

equivalent on mechanical property of grade b ship plates, Iron Steel 5 (1985) 41–44.
[22] S.B. Singh, H.K.D.H. Bhadeshia, D.J.C. Mackay, H. Carey, I. Martin, Neural network analysis of steel plate processing, Ironmaking Steelmaking 25 (1998) 355–365.
[23] F. Pettersson, N. Chakraborti, S.B. Singh, Neural networks analysis of steel plate processing augmented by multi-objective genetic algorithms, Steel Res. Int. 78 (2007) 890–898.
[24] F. Pettersson, N. Chakraborti, H. Saxén, A genetic algorithms based multi-objective neural net applied to noisy blast furnace data, Appl. Soft Comput. 7 (2007) 387–397.
[25] B.K. Giri, F.S. Pettersson, H. Saxén, N. Chakraborti, Genetic programming evolved through bi-objective genetic algorithms applied to a blast furnace, Mater. Manuf. Processes 28 (2013) 776–782.
[26] B. Debanjana, P.P. Ranjan, D.P. Kumar, H. Chadan, P. Snehansju, Data-driven bi-objective genetic algorithms evonn applied to optimize dephosphorization process during secondary steel making operation for producing lpg (liquid petroleum gas cylinder) grade of steel, Steel Res. Int. 89 (2018) 1800095.
[27] S. Pal, C. Halder, Optimization of phosphorous in steel produced by basic oxygen steel making process using multi-objective evolutionary and genetic algorithms, Steel Res. Int. 88 (2017) 1600193.
[28] T. Chugh, N. Chakraborti, K. Sindhya, Y. Jin, A data-driven surrogate-assisted evolutionary algorithm applied to a many-objective blast furnace optimization problem, Mater. Manuf. Processes 32 (2017) 1172–1178.
[29] B.K. Mahanta, N. Chakraborti, Evolutionary data driven modeling and multi objective optimization of noisy data set in blast furnace iron making process, Steel Res. Int. 89 (2018) 1800121.
[30] H. Peters, A. Ebel, M. Holzknecht, N. Link, J. Häckmann, T. Heckenthaler, F. Lücking, M. Pander, In Industrial data mining in steel industry, Journees Siderurgiques Internationales 7 (2012) 53–68.
[31] W.N.L. Browne, The development of an industrial learning classifier system for data-mining in a steel hot strip mill, Applications of Learning Classifier Systems, vol. 150, 2004, pp. 223–259.
[32] A. Agrawal, P.D. Deshpande, A. Cecen, G.P. Basavarsu, A.N. Choudhary, S.R. Kalidindi, Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters, Integrating Mater. Manuf. Innovation 3 (2014) 8.
[33] R. Jha, N. Chakraborti, D.R. Diercks, A.P. Stebner, C.V. Ciobanu, Optimal mean radius and volume fraction of the nanocrsytalline phase in softmagnetic alloys: a combined machine learning and calphad approach, Comput. Mater. Sci. 150 (2018) 202–211.
[34] J.F. Wen, Optimization of ceq, specification and properties of 16mn steel grade, Wide Heavy Plate 1 (1995) 17–21.
[35] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, An interior-point method for large-scale -regularized least squares, IEEE J. Sel. Top. Signal Process. 1 (2007) 606–617.
[36] R.H. Byrd, J.C. Gilbert, J. Nocedal, A trust region method based on interior point techniques for nonlinear programming, Math. Program. 89 (2000) 149–185.
[37] R.A. Waltz, J.L. Morales, J. Nocedal, D. Orban, An interior algorithm for nonlinear optimization that combines line search and trust region steps, Math. Program. 107 (2006) 391–408.
[38] R.H. Byrd, M.E. Hribar, J. Nocedal, An interior point algorithm for large-scale nonlinear programming, SIAM J. Optim. 9 (1999) 877–900.
[39] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[40] A.L. Edwards, An introduction to linear regression and correlation, Math. Gazette 69 (1984) 1–17.
[41] T.F. Coleman, A. Verma, A preconditioned conjugate gradient approach to linear equality constrained minimization, Comput. Optimization Appl. 20 (2001) 61–72.
[42] O. Axelsson, L.Y. Kolotilina, Preconditioned Conjugate Gradient Methods, Springer-Verlag, 1990.
[43] R.G. Brereton, G.R. Lloyd, GR, Support vector machines for classification and regression, Analyst 135 (2010) 230–267.
[44] B.S. Everitt, Classification and regression trees, Encyclopedia of Statistics in Behavioral Science, (2005).
[45] L.Å. Norström, The influence of nitrogen and grain size on yield strength in type AISI 316l austenitic stainless steel, Met. Sci. J. 11 (2013) 208–212.
[46] L.C. An, T.S. Liu, B. Lu, Y.T. Yang, Effect of Mo and Nb on microstructures and mechanical properties of the medium-carbon low-alloyed cast steel, Foundry Technol. (2015) 847–850.
[47] K.C. Hwang, S. Lee, C.L. Hui, Effects of alloying elements on microstructure and fracture properties of cast high speed steel rolls. Part i: Microstructural analysis, Mater. Sci. Eng. A 254 (1998) 296–304.
[48] S.C. Wang, P.W. Kao, The effect of alloying elements on the structure and mechanical properties of ultra low carbon bainitic steels, J. Mater. Sci. 28 (1993) 5169–5175.
[49] Q.B. Yu, Y. Sun, Effect of carbon content and microstructure on the yield-strength ratio of steel, J. Plasticity Eng. 16 (2009) 119–126.