Available online at www.sciencedirect.com
ScienceDirect
Trans. Nonferrous Met. Soc. China 33(2023) 2110−2120
Transactions of Nonferrous Metals Society of China
www.tnmsc.cn

An end-to-end machine learning framework exploring phase formation for high entropy alloys

Hui-ran ZHANG[1,2,3], Rui HU[1], Xi LIU[1], Sheng-zhou LI[4],
Guang-jie ZHANG[1], Quan QIAN[1,2,3], Guang-tai DING[1], Dong-bo DAI[1]

1. School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China;
2. Materials Genome Institute, Shanghai University, Shanghai 200444, China;
3. Zhejiang Laboratory, Hangzhou 311100, China;
4. Department of Computer Science, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan

Received 1 March 2022; accepted 30 August 2022

Abstract: Exploring the rules of high entropy alloys (HEAs) phase formation has clear guiding significance for the design of new alloys. An end-to-end framework was proposed to select the feature subset and machine learning (ML) model from the feature pool and model pool, respectively. In this framework, each model in the pool is to determine its materials feature subset based on the feature importance. The final model was confirmed through the evaluation of the fitting result of every model and its feature subset. This method extracts important factors affecting the phase formation of HEAs. The results show that the chosen model could classify 430 HEAs into five phases, with test accuracy of 87.8%. And the model analysis suggests that the formation of single-phase solid solution is often inhibited when the atomic size difference is greater than 8.295%.

Key words: feature selection; high entropy alloys; machine learning; phase prediction; Hume−Rothery rules

1 Introduction

High entropy alloys (HEAs) have attracted considerable attentions and research interests owing to their different solid solution formation since it was reported by YEH et al [1] and CANTOR et al [2]. Commonly, the thermodynamic parametric approaches are used to conduct the phase selection of HEAs. WANG et al [3] have shown that the phase selection of HEAs is determined by parameters such as mixing enthalpy, atomic size difference, and valence electron concentration. TAKEUCHI and INOUE [4] proposed a thermodynamic model including three empirical rules to find new multicomponent alloys. Since then, more and more thermodynamic parameters have been proposed to study the rules of alloys phase formation [5,6]. However, it is still very difficult to find the condition parameters of alloys phase formation to guide alloys design from those quite a few parameters.

Recently, a number of researchers have applied efficient machine learning (ML) models to discover new materials [7], especially in the researches of the given alloys data of relevant properties [8,9]. Some researchers showed that ML models could assist to find the phase formation [10,11]. And some ML models with a good prediction accuracy were developed for predicting mechanical properties of HEAs [12,13]. Some efficient descriptors were constructed and then provided to ML model for phase prediction of HEAs [14−16]. But, the majority of the existing ML models are black-box ML models, which somewhat lack interpretability and hinder investigators from

Corresponding author: Dong-bo DAI, Tel: +86-13671705578, E-mail: dbdai@shu.edu.cn
DOI: 10.1016/S1003-6326(23)66247-8

acquiring further chemical insight from models [14]. The key of building an ML model is how to select the most relevant parameters as input features. Parameters or features, as the crucial factors, could help the researchers to clarify the phase formation rules of the alloys. Usually, it is necessary to rely on feature engineering to construct or filter the material descriptors if domain knowledge is insufficient [17]. The importance of features can be obtained by some models, such as the least absolute shrinkage and selection operator (LASSO) [18], random forest (RF), gradient tree boosting (GTB) and support vector machine (SVM) [19]. However, these material descriptors are valid only for a particular ML model, and if the ML model is inefficient for the prediction task, its material descriptors can be invalid. Dimensionality reduction algorithms such as principal component analysis (PCA) can map the high-dimensional features to low-dimensional space but lack the interpretability [14,20].

In order to solve the separation of ML model selection and feature selection and the concealment of knowledge obtained by HEAs phase prediction model, an end-to-end framework of HEAs phase prediction is proposed in this work, which is shown in Fig. 1. Based on the existing empirical parameters and ML models, we build materials descriptors pool and ML models pool respectively. Then, all ML models fitting all descriptors are evaluated and the most suitable model with the best material features is identified. Based on the above selected models and feature subset, we establish and validate the prediction model of HEAs phase formation.

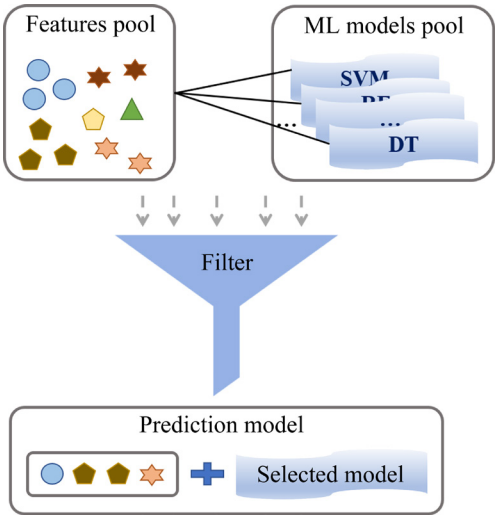


Fig. 1 Workflow of machine learning based HEAs phase prediction

2 Materials and methodologies

2.1 Data collection and preprocessing

The dataset, including 430 data, is collected from 155 articles by TAN et al [6], GAO et al [21], YE et al [22]. All entries are divided into five categories by their microstructures: FCC (face-centered cubic), BCC (body-centered cubic), HCP (hexagonal close-packed), MP (multi-phase), and AP (amorphous phase). Each entry of the dataset is composed of 14 empirical parameters and the phase structure of the alloy. The resulting subset contains 100 single solid solutions (48 FCC, 43 BCC, and 16 HCP), 237 multi-phase solid solutions, and 63 amorphous phase solid solutions. Figure 2 plots the proportion of these five phases of HEAs. Although HCP accounts for only 3.93% of the total data set, it still has a relatively large proportion (14%) in single-phase HEAs. Therefore, HCP is still an important microstructure of HEAs and is retained as an important classification target of HEAs phase classification.

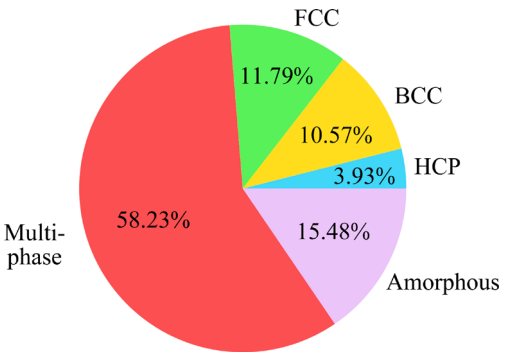


Fig. 2 Proportion of five phases data of HEAs

In the procedure of data preprocessing, there are two main operations: standardization of dataset and dataset partition. Most ML models make a priori assumption that the distribution of data obeys normal distribution (Gaussian with zero mean and unit variance), so standardization of datasets is a common requirement. The standard score of the sample x is calculated as

$$z=\frac{x-\mu}{\sigma} \tag{1}$$

where μ is the mean of training samples and σ is the standard deviation of the training samples.

Here, 20% of all data are held out as a test set and the rest as a training set. This data partition could avoid data leakage [23] and evaluate the generalization performance of the final phase classification model. A ML model could remember all training data perfectly, but it has no acceptable prediction ability for unprecedented data. This situation is called the overfitting problem of the model. Besides, five phases classification of the HEAs exhibits a large imbalance, seen from Fig. 2. To avoid the inconsistency between the distribution of fewer categories in the subset and the original category distribution, stratified sampling was used to divide the training set and the test set. Also, stratified 5-fold was used as the partition strategy of training set and verification set for subsequent cross-validation. In this work, the training set plays two important roles: the feature importance evaluation dataset of model-based feature selection, and the cross-validation set of model selection.

2.2 Machine learning classifiers

To fit small samples dataset and understand the feature importance in the process of fitting, eight ML classification models with higher interpretability are employed to build the ML model pool. They are SVM with the linear kernel (SVM), Naïve Bayes (NB), Decision Tree (DT), Extremely Randomized Trees (ExtraTree), Linear Discriminant Analysis (LDA), Logistic Regression (LR), Random Forest (RF), and Ridge Regression. All of them are implementations included in the scikit-learn package for python.

2.3 Feature selection

The performance of ML model depends very much on the features it uses, and the selection of features depends very much on specific target tasks [24]. Features with physical significance can also directly enhance the interpretability of the model and offer chemical insight for materials researchers. Overall, a total of 14 thermodynamic parameters are chosen as the initial features for ML training. Among these parameters, two of them are the two important factors of the Gibbs free energy of mixing ($\Delta G_{mix}=\Delta H_{mix}-T\Delta S_{mix}$), namely, enthalpy of mixing (ΔH_{mix}) and entropy of mixing (ΔS_{mix}). The thermodynamic model proposed by TAKEUCHI and INOUE [4] shows that the stability of solid solution phase is affected by mixing entropy. Meanwhile, the highest positive enthalpy of mixing ΔH_{mix}^{max} and the highest negative enthalpy of mixing ΔH_{mix}^{min} are also taken as the features. In addition, several descriptors measuring the variance of dimensionless enthalpy of mixing are included, $\sqrt{\Delta H_{mix}}$, $\sqrt{\Delta H_{mix}^{0}}$, $\sqrt{\Delta H_{mix}^{0+}}$, and $\sqrt{\Delta H_{mix}^{0-}}$ [6]. The superscripts "+" and "–" stand for the positive and negative enthalpy of mixing, respectively. Furthermore, the atomic size difference δr [25], atomic packing mismatch γ [25], the configurational entropy for the formation of an ideal solid-solution phase S_c [1], elastic residual strain root mean square (ε_{RMS}) [21], valence electron concentration (VEC) [3], and the parameter Φ [26] are also a part of these features.

Obviously, it is difficult to select these relevant features for phase classification in HEAs dataset. Too many (few) features can easily lead to overfitting (underfitting) of the model. In order to make a trade-off between the fitting ability and generalization ability of the model, an appropriate strategy needs to be used to select features that are related to phase formation. Based on the above considerations, we use recursive feature elimination (RFE) [27] to prove the necessity of feature reduction. Figure 3 shows the change trend of 5-fold cross-validation scores (F1 score) of each model with increasing the number of features. Figure 3 presents that classification performance of models would converge with the increase of the features. But, the subset of features used in ML models is different from each other. We can select the corresponding applicability characteristics for each model. And, the subsets of features combined with one ML model are determined by the ML model. Eight models fit the training dataset

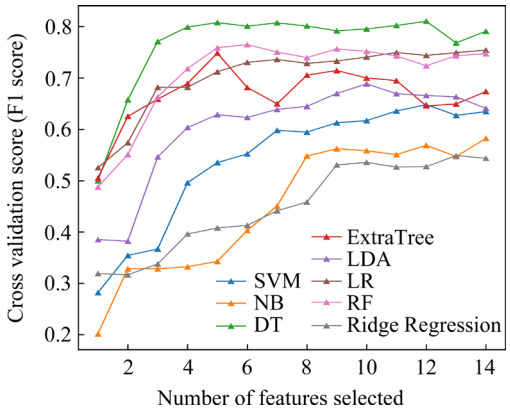


Fig. 3 Model performance with different feature numbers

respectively, and then coefficients or importance scores of features of each model can be obtained. To reduce features and hold back these important features, features whose feature importance scores exceed the average are selected to form feature subsets used to train the ML classification model. Through the above model-based feature selection methods, each model has a more appropriate and robust feature subset.

2.4 Multiclass classification evaluation criteria

In order to comprehensively evaluate the fitting ability and generalization ability of ML classifier, some statistical criteria are employed, such as the accuracy (A), kappa index (k), F1 score (F), and confusion matrix. The aforementioned statistics can be expressed by following equations:

$$A=\frac{O_{\mathrm{TP}}+O_{\mathrm{TN}}}{O_{\mathrm{TP}}+O_{\mathrm{TN}}+O_{\mathrm{FP}}+O_{\mathrm{FN}}} \tag{2}$$

$$k=\frac{l-c}{1-c} \tag{3}$$

$$p=\frac{O_{\mathrm{TP}}}{O_{\mathrm{TP}}+O_{\mathrm{FP}}} \tag{4}$$

$$r=\frac{O_{\mathrm{TP}}}{O_{\mathrm{TP}}+O_{\mathrm{FN}}} \tag{5}$$

$$F=2\frac{p\cdot r}{p+r} \tag{6}$$

where O_{TP}, O_{TN}, O_{FP}, and O_{FN} represent true positive, true negative, false positive, and false negative outcomes of classifiers respectively, p, r, l and c represent precision rate, recall rate, the observed level of agreement between raters and the hypothetical chance of agreement between raters.

3 Results and discussion

3.1 Feature selection and model selection

The feature coefficient or importance score of the fitted model can reveal the relation between the features and the target property of HEAs. Different ML models due to their different fitting principles could have different feature importance rankings. To obtain all feature importance scores of each model, the eight models fit the training set respectively, and then the feature coefficient or importance scores are extracted from the fitted model. The feature importance scores of each model are shown in Fig. 4.

From each subfigure of Fig. 4, it can be seen that there are not same importance scores of those features in the eight models, and the difference of feature importance is large. The features that help to improve the accuracy of phase classification of the model have a high feature importance score. After SVM maps the feature space to high-dimensional space, it can be found that $\varepsilon_{\mathrm{RMS}}$, $\sqrt{\Delta H_{\mathrm{mix}}}$, $\sqrt{\Delta H_{\mathrm{mix}}^{0}}$, $\sqrt{\Delta H_{\mathrm{mix}}^{0-}}$, γ and VEC are more favorable to distinguish the phases of HEAs. When DT is used to build HEAs phase classification model, δr, Φ, VEC and $\sqrt{\Delta H_{\mathrm{mix}}^{0-}}$, as nodes in the tree, can divide the data of different phase categories to the greatest extent. Due to the data from the models being very limited, not all the features are favorable for establishing the final phase prediction model. It is almost the features set with high importance scores that determine the upper limit of model fitting, while the features set with low importance scores are easy to result in over-fitting of the model. Thus, a feature selection strategy is carried out based on feature importance scores of the fitted model. In order to obtain a robust feature subset, the mean of all feature importance scores of the fitting model was used as the threshold to select features. In Fig. 4, the red line of each subfigure shows the mean of features, and all features more than the threshold (red line) were preserved. Then, those feature subset (Fig. 5(d)) was used to establish the phase prediction model.

Through the above process, the robust features of each model have been selected. The next work focuses on finding an optimal model to establish the phase prediction model. Each model combined with its unique feature subset is subject to 5-fold cross-validation on the training set, and the best model is determined by a variety of evaluation criteria.

Figures 5(a−c) show the result of eight ML model classification performances by three evaluation criteria on the 5-fold cross-validation dataset. From the results of the three indicators of the model, the ML models based on the tree (DT, RF, and ExtraTree) are better than the other five models. Especially in the training process, the three tree-based models can accurately fit the training data set. At the same time, the performance of the three tree-based models on the cross-validation test dataset has high accuracy (all more than 80%),

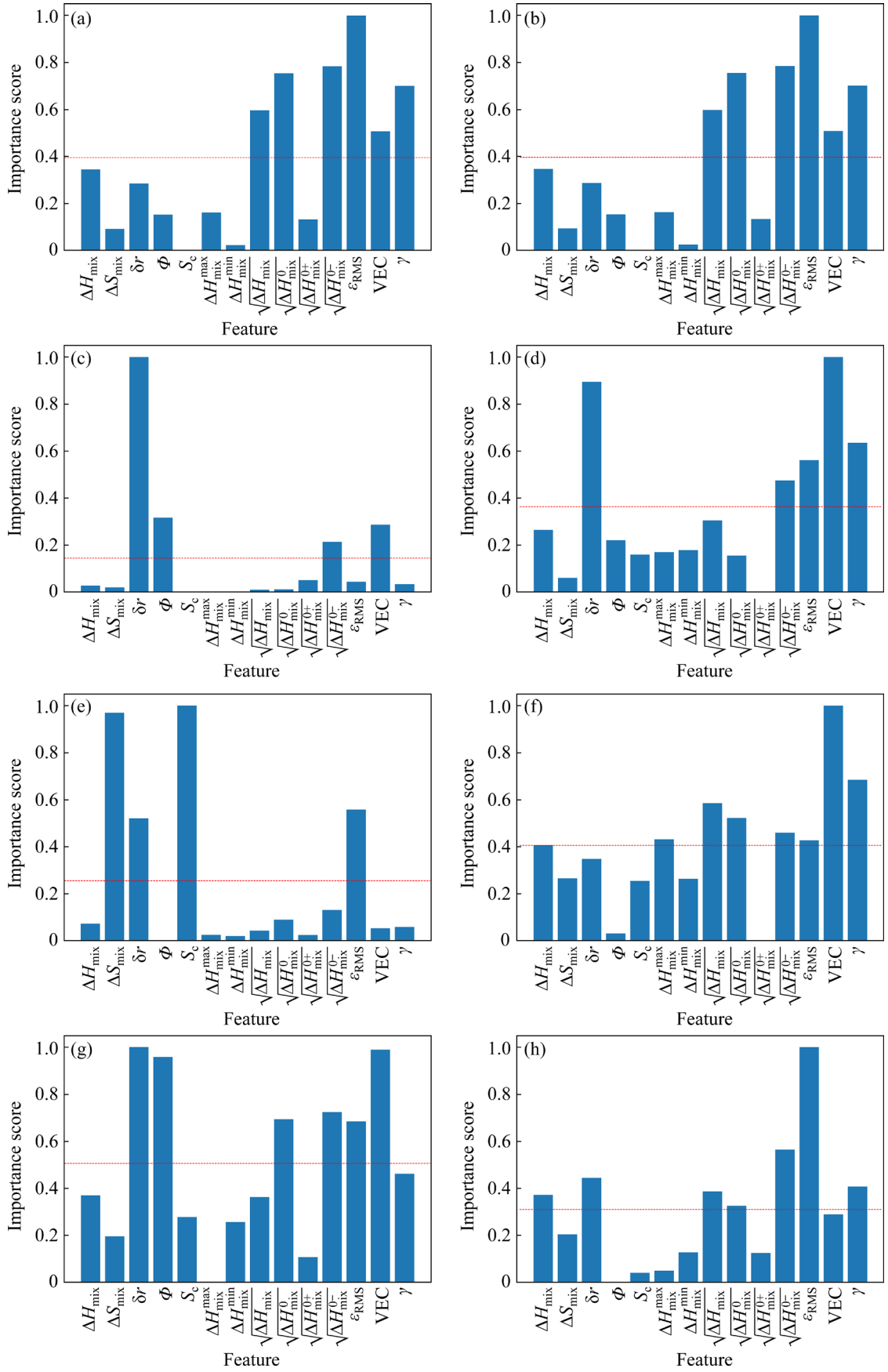
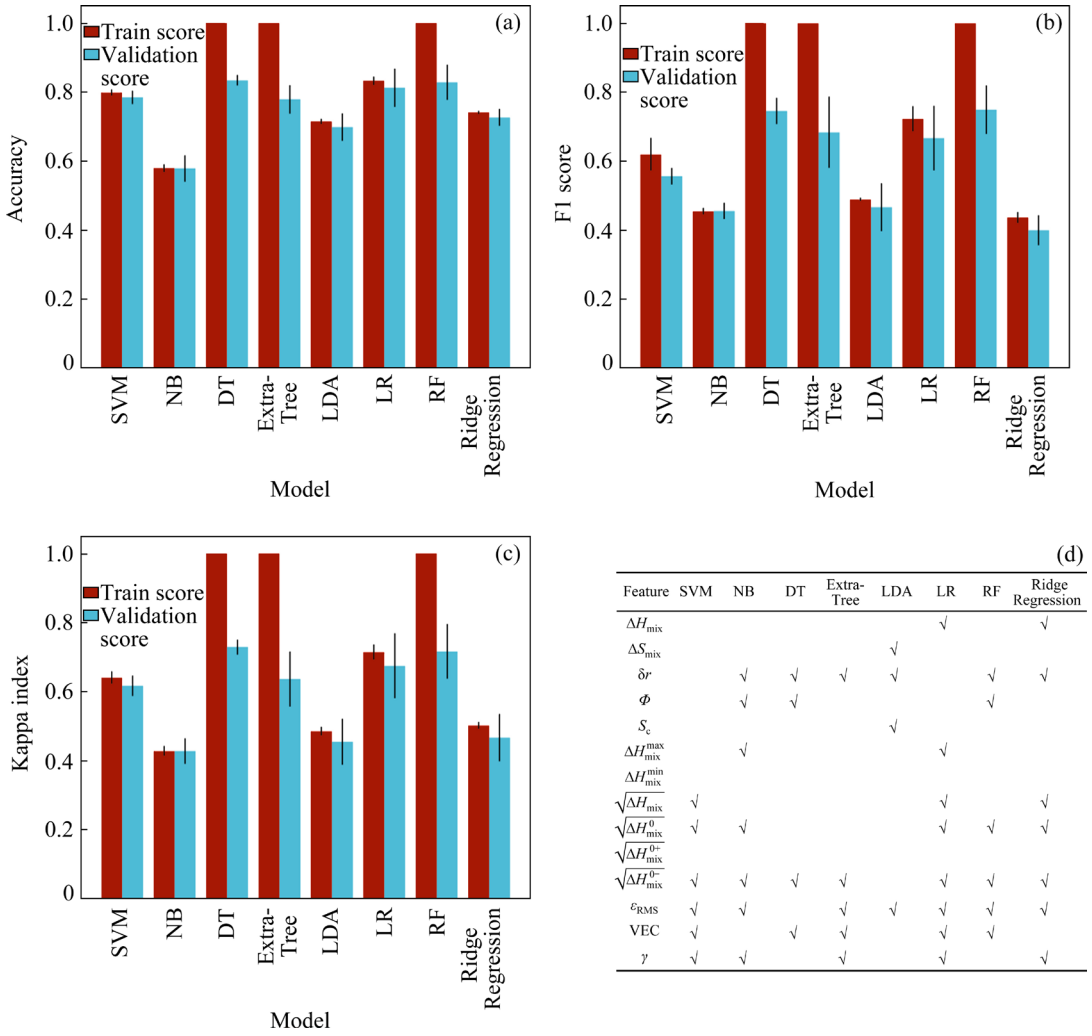


Fig. 4 Features importance scores of eight models: (a) SVM; (b) NB; (c) DT; (d) ExtraTree; (e) LDA; (f) LR; (g) RF; (h) Ridge Regression (The red line is the mean of all feature importance scores)



Feature	SVM	NB	DT	Extra-Tree	LDA	LR	RF	Ridge Regression
ΔH_{mix}						√		√
ΔS_{mix}					√			
δr		√	√	√	√		√	√
Φ		√	√				√	
S_c					√			
ΔH_{mix}^{max}		√				√		
ΔH_{mix}^{min}								
$\sqrt{\Delta H_{mix}}$	√					√		√
$\sqrt{\Delta H_{mix}^{0}}$	√	√				√	√	√
$\sqrt{\Delta H_{mix}^{0+}}$								
$\sqrt{\Delta H_{mix}^{0-}}$	√	√	√	√		√	√	√
ε_{RMS}	√	√		√	√	√	√	√
VEC	√		√	√		√	√	
γ	√	√		√		√		√

Fig. 5 Accuracy (a), F1 score (b), kappa index (c) of eight ML models classification on 5-fold cross-validation dataset, and features set used in each model (d)

which represents that the feature subset selected based on the model performs well in both fitting ability and generalization ability of the model. In a surprising result, the DT model with just four features (δr, Φ, VEC, and $\sqrt{\Delta H_{mix}^{0-}}$ in Fig. 4(d)) is better than the other two ensemble tree model (RF and ExtraTree). This shows that HEAs phase formation is affected by the four factors δr, Φ, VEC, and $\sqrt{\Delta H_{mix}^{0-}}$. Therefore, DT with its four features is selected as the final suitable phase prediction model.

The HEAs phase prediction model DT and its four features δr, Φ, VEC, and $\sqrt{\Delta H_{mix}^{0-}}$ are determined with the above features and models selection process. In this phase, a HEAs phase prediction model (DT) is established with the all data but without the hold-out test dataset for further improving the models, and the generalization of the model is tested on the hold-out test dataset unseen by the model.

3.2 Model performance

The test result of the DT model includes accuracy (0.87), F1 score (0.87), and kappa index (0.78) on hold-out test dataset. All test results are listed in Table S1 of Supplementary Information, and mispredicted samples are marked in bold. In order to better observe the classification accuracy of test samples, the confusion matrix of model test results is plotted in Fig. 6. It shows the classification results of the model on the five phases of HEAs. The yellow part gives the classification results of the DT model which shows that DT model has good classification performance for three single-phase of HEAs. The green part gives the

	BCC	FCC	HCP	Multi-phase	Amorphous
BCC	6 (100%)	0	0	2	0
FCC	0	9 (100%)	0	1	0
HCP	0	0	3 (100%)	0	0
Multi-phase	2	4	0	42 (88%)	0
Amorphous	0	0	0	1	12 (92%)

Fig. 6 Confusion matrix of HEAs phase prediction model test result

classification results of five phases of HEAs. The classification accuracy of each phase of the two classification targets is given in the brackets. In the matrix, it is found that several single-phase alloys are misclassified as multi-phase. The reason is that the change of elementary compositions or external factors may cause different structures. For example, in $Al_xCrCuFeNi_2$ alloy, the structure changes from FCC single-phase to FCC+BCC multi-phase with increasing Al molar fraction [28]. In HfNbTiZr refractory high entropy alloy, its crystal structure is different from that annealed at different temperatures, and the BCC+HCP mix phase appears at a lower temperature [29]. The $Al_{0.3}NbTa_{0.8}Ti_{1.4}V_{0.2}Zr_{1.3}$ HEA has a single-phase BCC structure. Two BCC mixed phases with different lattice parameters are often formed, when the content of its components is changed (e.g. $Al_{0.5}NbTa_{0.8}Ti_{1.5}V_{0.2}Zr$) [30].

Besides, by the DT model, a small amount of multi-phase HEAs were classified as single-phase alloys mistakenly and such single-phase recognized does exist in multi-phase. We guess that this has been caused by the proportion of phase: some multi-phase HEAs and single-phase HEAs may be misclassified as each other by DT model. The single phase in these mixed phases accounts for most of the grain interior, and the other single phase is often formed only in interdendritic region, such as $CrCu_2Fe_2MnNi_2$ [2], $Al_{0.3}CoCrFeMo_{0.1}Ni$ [31], $CrMo_{0.5}NbTa_{0.5}TiZr$ [32], and AlCrMoNbTi [33]. In addition, the phase formation of HEAs is sensitive to temperature. As the temperature decreases, the solid solution will undergo phase separation, which reduces the configurational entropy. The research of GAO et al [21] shows that the decrease of entropy could increase the nucleation driving force of intermetallic compounds. Precipitation of the σ phase in the single FCC phase CoCrFeMnNi alloy after annealing at 973 K is reported by OTTO et al [34]. In Fig. 7, all the misclassification HEAs samples in the test set are marked. The misclassification HEAs mainly appear in three areas (the areas circled by blue and gray), where the samples are mixed by HEAs of different phases. The area in the blue circle is the mixture of BCC phase HEAs and multi-phase HEAs. The two areas in the gray circle contain FCC phase HEAs and multi-phase alloy at the same time. These areas are not large, but the samples included have large quantity. This high mixing degree causes misclassification among these samples. In fact, in the case of relatively large intervals, the DT model can distinguish these HEAs of the two phases with good accuracy (the yellow circle, contains multi-phase HEAs and amorphous phase HEAs). The ZrHfTiCuCo [35] among the 13 amorphous phase alloys in the test set is wrongly divided into multi-phase in the confusion matrix. This also demonstrates the efficiency of the model for distinguishing the phases.

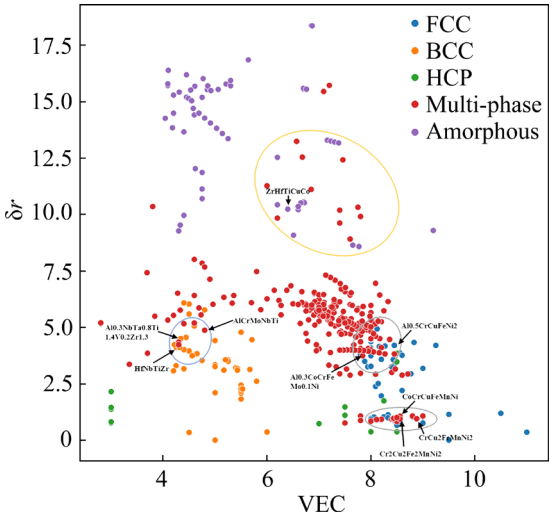


Fig. 7 Distribution visualization of HEAs dataset containing five phases on two-dimension δr−VEC

3.3 Exploration on formation rules of HEAs phase

Furthermore, in comparison to the previous studies, the DT model has higher interpretability for its explicit decision-making process. In order to analyze the formation rules of the HEAs phase and find the quantitative relationship of important factors affecting the formation of the HEAs phase, the internal decision-making process of the trained DT model is plotted in Fig. 8. The phase of each

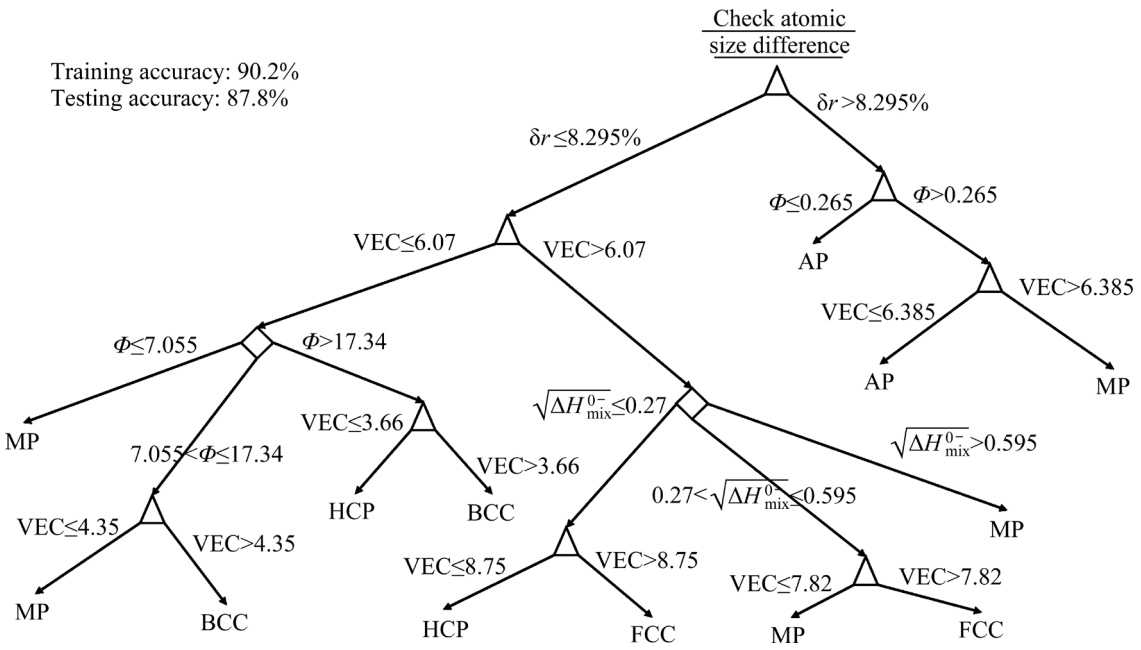


Fig. 8 DT of HEAs (The tree is the visualization of the DT model of trained HEAs phase prediction; Each alloy is assessed by DT)

alloy would be determined by the decision path of DT. The DT model selected four important features (δr, Φ, VEC, and $\sqrt{\Delta H_{\text{mix}}^{0-}}$) as significant factors during the formation of HEAs phase in Fig. 8. In Hume−Rothery rules, the size difference of the component atoms of alloys is the most important factor affecting the formation ability of the solid solution phase. When the atomic size difference is larger than 15%, the formation ability of the solid solution phase is commonly weak [36]. Hume−Rothery rules have been widely applied in binary alloy systems, and they are still at a primary stage in the study of solid-solution solubility for HEAs [19,37]. In the DT model, the atomic size difference (δr) was chosen as the first decision rule by the DT model. It proves that δr is the most important feature of the HEAs phase classification model (the root node of the DT). When δr is larger than 8.295%, the formation ability of single-phase solid solution is often inhibited in our phase prediction model of HEAs. Therefore, when designing a new type of solid solution HEAs, we prefer to take these elements whose size difference between atoms is less than 8.295% as the components. In the model, the atomic size difference not only provides qualitative structure−property relationship guidance but also provides quantitatively the structure−property relationship reference for HEAs design.

From Fig. 8, we can see that the VEC, one of the three factors of the Hume−Rothery rule, not only affects the stability of metallic phases of binary alloys but also affects the formation of different solid solution phases in HEAs. It was also found that when the atomic size differences were kept nearly identical, the VEC is the dominant factor controlling the phase stability in high-temperature structural alloys [28,38]. From the data distribution of HEAs we collected, the difference between FCC and BCC structures in the dimension of the atomic size difference is very small. But, FCC and BCC show the phenomenon of separation between classes in the dimension of VEC in Fig. 7. MIZUTANI [39] emphasized that the parameter VEC should be used in realistic electronic structure calculations to take into account the d-electron contribution. More theoretical work has been carried out to understand the physical basis for the mechanism behind the VEC rule on phase stability [40−43]. Our results prove that VEC also plays a decisive role in the stability of phases in the HEAs. In the DT model of HEAs phases classification, each alloy instance finally falls into the leaf node after VEC division in Fig. 8. Further, BCC phases are found to be stable at lower VEC (≤6.07) and instead FCC phases are stable at higher

VEC (>7.82) in this DT model. This phenomenon is also reflected in the two-dimensional δr−VEC phase diagram (Fig. 7). The VEC ranges acquired from this work are close to the result of GUO et al [28]. Therefore, the DT model could capture the quantitative structure attribution relationship of HEAs about VEC.

Due to the importance of enthalpy of mixing in designing HEAs, TAN et al [6] introduced several new parameters ($\sqrt{\Delta H_{\mathrm{mix}}}$, $\sqrt{\Delta H_{\mathrm{mix}}^{0}}$, $\sqrt{\Delta H_{\mathrm{mix}}^{0+}}$, and $\sqrt{\Delta H_{\mathrm{mix}}^{0-}}$) to reflect the variance of dimensionless enthalpy of mixing. The regions of HEAs with a single solid-solution phase, with multi-phases, and with an amorphous phase were separated qualitatively by using those parameters. In the process of model-based feature selection, seven models take parameter $\sqrt{\Delta H_{\mathrm{mix}}^{0-}}$ as an important feature of the corresponding model (in Fig. 4 and Fig. 5(d)). The parameter $\sqrt{\Delta H_{\mathrm{mix}}^{0-}}$ was also selected by our DT model in the feature selection. In Fig. 8, two critical values of $\sqrt{\Delta H_{\mathrm{mix}}^{0-}}$ was found: larger than 0.595 corresponds to HEAs with a multi-phase solid solution, and small than 0.27 corresponds to HEAs with a single-phase solid solution. In the DT model established, the parameter $\sqrt{\Delta H_{\mathrm{mix}}^{0-}}$ is a significantly important feature to distinguish single-phase (FCC and HCP) and multi-phase. As TAN et al [6] said, the difference between the enthalpy of mixing of component i, j (H_{mix}^{ij}) and 0 but not the average enthalpy of mixing (ΔH_{mix}) is more appropriate to characterize the resistance for the formation of a single solid-solution phase.

In the HEAs systems, entropy plays an important role in phase formation. Dimensionless parameter Φ proposed by YE et al [44] can be viewed as an "entropy" indicator for alloy ranking. YE et al [22,44] reported that the Φ parameter is an effective factor for distinguishing the single-phase HEAs from the others. The parameter Φ shows the same effect in our DT model, and it divides alloys instances into single-phase (BCC and HCP) and multi-phase. As seen in Fig. 8, a high entropy alloy tends to display a single-phase solid solution when Φ>17.34. The result is close to the critical value of Φ≈20 found by YE et al [44].

4 Conclusions

(1) The classification accuracy is up to 87% for the alloys not seen in the established model. This shows that the DT model we established has a strong generalization ability.

(2) Our method automatically extracts the important factors (δr, Φ, VEC, and $\sqrt{\Delta H_{\mathrm{mix}}^{0-}}$) affecting the formation of HEAs phase. Two (δr and VEC) of the four factors are in the Hume-Rothery rules, which shows that the Hume−Rothery rules still have extremely important guiding significance for HEAs design.

(3) The decision rules of the DT model can mine quantitative structure−property relationship in the process of HEAs phase formation to some extent. When the atomic size difference (δr) is greater than 8.295%, the formation ability of the single-phase solid solution is often inhibited.

Acknowledgments

This work was sponsored by the National Key Research and Development Program of China (No. 2018YFB0704400), Key Research Project of Zhejiang Laboratory, China (No.2021PE0AC02), Key Program of Science and Technology of Yunnan Province, China (Nos. 202002AB080001-2, 202102AB080019-3), and Key Project of Shanghai Zhangjiang National Independent Innovation Demonstration Zone, China (No. ZJ2021-ZD-006).

Supplementary Information

Supplementary Information in this paper can be found at: http://tnmsc.csu.edu.cn/download/13-p2110-2022-0237-Supplementary_Information.pdf.

References

[1] YEH J W, CHEN S K, LIN S J, GAN J Y, CHIN T S, SHUO T T, TSAU C H, CHANG S Y. Nanostructured high-entropy alloys with multiple principal elements: Novel alloy design concepts and outcomes [J]. Advanced Engineering Materials, 2004, 6(5): 299−303.

[2] CANTOR B, CHANG I T H, KNIGHT P, VINCENT A J B. Microstructural development in equiatomic multicomponent alloys [J]. Materials Science and Engineering A, 2004, 375/376/377: 213−218.

[3] WANG Zhi-jun, GUO Sheng, LIU C T. Phase selection in high-entropy alloys: From nonequilibrium to equilibrium [J]. JOM, 2014, 66(10): 1966−1972.

[4] TAKEUCHI A, INOUE A. Classification of bulk metallic glasses by atomic size difference, heat of mixing and period of constituent elements and its application to characterization of the main alloying element [J]. Materials Transactions, 2005, 46(12): 2817−2829.

[5] HE Quan-feng, DING Zhao-yi, YE Yi-fan, YANG Yan-cong. Design of high-entropy alloy: A perspective from nonideal mixing [J]. JOM, 2017, 69(11): 2092−2098.

[6] TAN Yi-ming, LI Jin-shan, TANG Zhao-wu, WANG Jun, KOU Hong-chao. Design of high-entropy alloys with a single solid-solution phase: Average properties vs their variances [J]. Journal of Alloys and Compounds, 2018, 742: 430−441.

[7] MEREDIG B, WOLVERTON C. A hybrid computational−experimental approach for automated crystal structure solution [J]. Nature Materials, 2013, 12(2): 123−127.

[8] RICKMAN J M, CHAN H M, HARMER M P, SMELTZER J A, MARVEL C J, ROY A, BALASUBRAMANIAN G. Materials informatics for the screening of multi-principal elements and high-entropy alloys [J]. Nature Communications, 2019, 10(1): 1−10.

[9] WEN Chen, ZHANG Yan, WANG Chang-xin, XUE De-zhen, BAI Yang, ANTONOV S, DAI Lan-hong, LOOKMAN T, SU Yan-jing. Machine learning assisted design of high entropy alloys with desired property [J]. Acta Materialia, 2019, 170: 109−117.

[10] PEI Zong-rui, YIN Jun-qi, HAWK J A, ALMAN D E, GAO M C. Machine-learning informed prediction of high-entropy solid solution formation: Beyond the Hume−Rothery rules [J]. NPJ Computational Materials, 2020, 6(1): 1−8.

[11] ZENG Ying-zhi, MAN Meng-ren, BAI Ke-wu, ZHANG Yong-wei. Revealing high-fidelity phase selection rules for high entropy alloys: A combined CALPHAD and machine learning study [J]. Materials & Design, 2021, 202: 109532.

[12] KIM G, DIAO H Y, LEE C, SAMAEI A T, PHAN T, JONG M, AN K, MA Dong, LIAW P K, CHEN W. First-principles and machine learning predictions of elasticity in severely lattice-distorted high-entropy alloys with experimental validation [J]. Acta Materialia, 2019, 181: 124−138.

[13] BHANDARI U, RAFI M R, ZHANG Cong-yan, YANG Shi-zhong. Yield strength prediction of high-entropy alloys using machine learning [J]. Materials Today Communications, 2021, 26: 101871.

[14] ZHANG Yan, WEN Cheng, WANG Chang-xin, ANTONOV S, XUE De-zhen, BAI Yang, SU Yan-jing. Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models [J]. Acta Materialia, 2020, 185: 528−539.

[15] FENG Shuo, FU Hua-dong, ZHOU Hui-yu, WU Yuan, LU Zhao-ping, DONG Hong-biao. A general and transferable deep learning framework for predicting phase formation in materials [J]. NPJ Computational Materials, 2021, 7(1): 1−10.

[16] YANG Chen, REN Chang, JIA Yue-fei, WANG Gang, LI Min-jie, LU Wen-cong. A machine learning-based alloy design system to facilitate the rational design of high entropy alloys with enhanced hardness [J]. Acta Materialia, 2022, 222: 117431.

[17] XU Qi-chen, LI Zhen-zhu, LIU Miao, YIN Wan-jian. Rationalizing perovskite data for machine learning and materials design [J]. Journal of Physical Chemistry Letters, 2018, 9(24): 6948−6954.

[18] GHIRINGHELLI L M, VYBIRAL J, LEVCHENKO S V, DRAXL C, SCHEFFLER M. Big data of materials science: Critical role of the descriptor [J]. Physical Review Letters, 2015, 114(10): 105503.

[19] LI Sheng-zhou, ZHANG Hui-ran, DAI Dong-bo, DING Guang-tai, WEI Xiao, GUO Yi-ke. Study on the factors affecting solid solubility in binary alloys: An exploration by machine learning [J]. Journal of Alloys and Compounds, 2019, 782: 110−118.

[20] BRODERICK S R, NOWERS J R, NARASIMHAN B, RAJAN K. Tracking chemical processing pathways in combinatorial polymer libraries via data mining [J]. Journal of Combinatorial Chemistry, 2009, 12(2): 270−277.

[21] GAO M C, ZHANG C, GAO P, ZHANG F, OUYANG L Z, WIDOM M, HAWK J A. Thermodynamics of concentrated solid solution alloys [J]. Current Opinion in Solid State and Materials Science, 2017, 21(5): 238−251.

[22] YE Yi-fan, WANG Qing, LU Jia-tian, LIU C T, YANG Yan-cong. High-entropy alloy: Challenges and prospects [J]. Materials Today, 2016, 19(6): 349−362.

[23] TOYAO T, MAENO Z, TAKAKUSAGI S, KAMACHI T, TAKIGAWA I, SHIMIZU K I. Machine learning for catalysis informatics: Recent applications and prospects [J]. ACS Catalysis, 2020, 10(3): 2260−2297.

[24] SCHIMIT J, MARQUES M R G, BOTTI S, MARQUES M A L. Recent advances and applications of machine learning in solid-state materials science [J]. NPJ Computational Materials, 2019, 5(1): 1−36.

[25] WANG Zhi-jun, HUANG Yun-hao, YANG Yong, WANG Jin-cheng, LIU C T. Atomic-size effect and solid solubility of multicomponent alloys [J]. Scripta Materialia, 2015, 94: 28−31.

[26] RATURI A, ADITYA C J, GURAO N P, BISWAS K. ICME approach to explore equiatomic and non-equiatomic single phase BCC refractory high entropy alloys [J]. Journal of Alloys and Compounds, 2019, 806: 587−595.

[27] GUYON I, ELISSEEFF A. An introduction to variable and feature selection [J]. Journal of Machine Learning Research, 2003, 3: 1157−1182.

[28] GUO Sheng, NG C, LU Jian, LIU C T. Effect of valence electron concentration on stability of fcc or bcc phase in high entropy alloys [J]. Journal of Applied Physics, 2011, 109(10): 103505.

[29] TU C H, LAI Y C, WU S K, LIN Y H. The effects of annealing on severely cold-rolled equiatomic HfNbTiZr high entropy alloy [J]. Materials Letters, 2021, 303: 130526.

[30] SENKOV O N, WOODWARD C, MIRACLE D B. Microstructure and properties of aluminum-containing refractory high-entropy alloys [J]. JOM, 2014, 66(10): 2030−2042.

[31] SHUN T T, HUNG C H, LEE C F. Formation of ordered/disordered nanoparticles in FCC high entropy alloys [J]. Journal of Alloys and Compounds, 2010, 493(1/2): 105−109.

[32] SENKOV O N, WOODWARD C F. Microstructure and properties of a refractory NbCrMo0.5Ta0.5TiZr alloy [J]. Materials Science and Engineering A, 2011, 529: 311−320.

[33] CHEN H, KAUFFMANN A, GORR B, SCHLIEPHAKE D, SEEMULLER C, WAGNER J N, CHRIST H J, HEILMAIER M. Microstructure and mechanical properties at elevated temperatures of a new Al-containing refractory high-entropy alloy Nb−Mo−Cr−Ti−Al [J]. Journal of Alloys and Compounds, 2016, 661: 206−215.

[34] OTTO F, DLOUHY A, PRADEEP K G, KUBENOVA M, RAABE D, EGGELER G, GEORGE E P. Decomposition of the single-phase high-entropy alloy CrMnFeCoNi after prolonged anneals at intermediate temperatures [J]. Acta Materialia, 2016, 112: 40−52.

[35] MA Li-qun, WANG Li-min, ZHANG Tao, INOUE A. Bulk glass formation of Ti−Zr−Hf−Cu−M (M=Fe, Co, Ni) alloys [J]. Materials Transactions, 2002, 43(2): 277−280.

[36] WANG Zhi-jun, HUANG Yun-hao, LIU C T, LI Jun-jie, WANG Jin-cheng. Atomic packing and size effect on the Hume−Rothery rule [J]. Intermetallics, 2019, 109: 139−144.

[37] JUAN Yong-fei, ZHANG Jiao, DAI Yong-bing, DONG Qing, HAN Yan-feng. Designing rules of laser-clad high-entropy alloy coatings with simple solid solution phases [J]. Acta Metallurgica Sinica, 2020, 33(8): 1064−1076.

[38] ZHU J H, LIAW P K, LIU C T. Effect of electron concentration on the phase stability of $NbCr_2$-based Laves phase alloys [J]. Materials Science and Engineering A, 1997, 239/240: 260−264.

[39] MIZUTANI U. Hume−Rothery rules for structurally complex alloy phases [J]. MRS Bulletin, 2012, 37(2): 169−169.

[40] TAKEUCHI A, WADA T, KATO H. High-entropy alloys with hexagonal close-packed structure in Ir26Mo20Rh22.5-Ru20W11.5 and Ir25.5Mo20Rh20Ru25W9.5 alloys designed by sandwich strategy for the valence electron concentration of constituent elements in the periodic chart [J]. Materials Transactions, 2019, 60(8): 1666−1673.

[41] LIU Min, XU Wen-yao, ZHANG Shi-dong, WANG Ze-min, WANG Zhan-yong, WANG Bin-jun, WANG Duo, LI Fang-jie. Microstructures and hardnesses of AlCoCr0.5-FexNi2.5 high entropy alloys with equal valence electron concentration [J]. Journal of Alloys and Compounds, 2020, 824: 153881.

[42] JIN Bing-qian, ZHANG Nan-nan, ZHANG Yue, LI De-yuan. Microstructure, phase composition and wear resistance of low valence electron concentration Al_xCoCrFeNiSi high-entropy alloys prepared by vacuum arc melting [J]. Journal of Iron and Steel Research International, 2021, 28(2): 181−189.

[43] LIU Bin, WU Ji-feng, CUI Yan-wei, ZHU Qin-qing, XIAO Guo-rui, WU Si-qi, CAO Guang-han, REN Zhi. Structural evolution and superconductivity tuned by valence electron concentration in the Nb−Mo−Re−Ru−Rh high-entropy alloys [J]. Journal of Materials Science & Technology, 2021, 85: 11−17.

[44] YE Yi-fan, WANG Qing, LU Jia-tian, LIU C T, YANG Yan-cong. Design of high entropy alloys: A single-parameter thermodynamic rule [J]. Scripta Materialia, 2015, 104: 53−55.

一种探索高熵合金相形成的端到端机器学习框架

张惠然[1,2,3]，胡 瑞[1]，刘 茜[1]，李盛洲[4]，张光捷[1]，钱 权[1,2,3]，丁广太[1]，戴东波[1]

1. 上海大学 计算机工程与科学学院，上海 200444；
2. 上海大学 材料基因研究院，上海 200444；
3. 之江实验室，杭州 311100；
4. Department of Computer Science, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan

摘 要：探索高熵合金(HEAs)的相形成规则对于新型合金的设计具有明确的指导意义。提出一种端到端的框架用来从特征池和模型池中分别选择特征子集和机器学习(ML)模型。在该框架中，模型池中的模型基于其获得的特征重要性来选择适合自身的特征子集；通过评估每个模型和其对应的特征子集的拟合结果，用于建立目标任务的预测模型；最终，获得影响 HEAs 相形成的重要因素。研究结果显示，建立的相预测模型可将 430 种 HEAs 分成 5 种相，测试准确度达到 87.8%，并且通过分析模型发现，当原子尺寸差异大于 8.295%时，HEAs 的单相固溶体的形成受到抑制。

关键词：特征选择；高熵合金；机器学习；相预测；Hume−Rothery 规则

(Edited by Bing YANG)