Materials and Design 193 (2020) 108835

Contents lists available at ScienceDirect

Materials and Design

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

Machine learning reveals the importance of the formation enthalpy and atom-size difference in forming phases of high entropy alloys

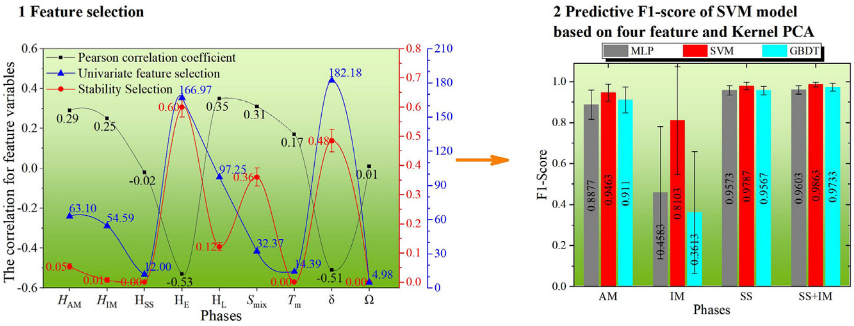
Lei Zhang [a,d], Hongmei Chen [b], Xiaoma Tao [b], Hongguo Cai [a,d], Jingneng Liu [c], Yifang Ouyang [b,*], Qing Peng [e], Yong Du [f]

[a] School of Mathematics and Information Science, Guangxi College of Education, Nanning 530023, China,
[b] School of Physical Science and Technology, Guangxi University, Nanning 530004, China,
[c] Maritime College, Beibu Gulf University, Qinzhou 535011, China,
[d] Institute for Intelligent Computing and Simulation Research, Guangxi College of Education, Nanning 530023, China
[e] Physics Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia
[f] State Key Laboratory of Powder Metallurgy, Central South University, Changsha 410083, China

HIGHLIGHTS

- The accurate thermodynamic properties of alloys can be calculated quickly by extended Miedema theory
- The feature variables are optimized by Kernel Principal Component Analysis.
- The phases of multi-principal element alloys are distinguished well by support vector machine of machine learning model.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 1 April 2020
Received in revised form 14 May 2020
Accepted 28 May 2020
Available online 30 May 2020

Keywords:
Multi-principal element alloys
Miedema theory
Machine learning
Feature selection
Alloy phase prediction

ABSTRACT

Despite outstanding and unique properties, the structure-property relationship of high entropy alloys (HEAs) is not well established. The machine learning (ML) is used to scrutinize the effect of nine physical quantities on four phases. The nine parameters include formation enthalpies determined by the extended Miedema theory, and mixing entropy. They are highly related to the phase formation, common ML methods cannot distinguish accurately. In this paper, feature selection and feature variable transformation based on Kernel Principal Component Analysis (KPCA) are proposed, the feature variables are optimized, the distinction of phases is carried out by Support vector machine (SVM) model. The results indicate that elastic energy and atom-size difference contribute significantly in the formation of different phases. The accuracy of testing set predicted by SVM based on four feature variables and KPCA (4V-KPCA) is 0.9743. The F1-scores predicted detailedly by SVM based on 4V-KPCA for the considered alloy phases are 0.9787, 0.9463, 0.9863 and 0.8103, corresponding to solid solution, amorphous, the mixture of solid solution and intermetallic, and intermetallic respectively. The extended Miedema theory provides accurate thermodynamic properties for the design of HEAs, and ML methods (especially SVM combined KPCA) are powerful in the prediction of alloy phases.

1. Introduction

The conventional alloys were designed primarily on the basis of one or two principal constituent elements and a few other alloying elements

* Corresponding author.
E-mail address: ouyangyf@gxu.edu.cn (Y. Ouyang).

https://doi.org/10.1016/j.matdes.2020.108835

to adjust their microstructures and properties. However, the addition of alloying elements might result in the formation of some brittle intermetallic compounds, and the mitigation of the mechanical properties. It is a desirable and continuous effort to discover the relationships among the alloying elements, the composition of alloying elements, and the performance of alloys. The alloys composed of several constituent elements with equal atomic composition are named multi-principal element alloys (MPEAs). The seminal papers of Yeh et al. [1] and Cantor et al. [2] proposed a new class of materials and increased the awareness of the understanding for alloy design. The alloys such as high-entropy alloys (HEAs), which were subject to MPEAs, typically comprise of five or more constituent elements (for brevity, the term HEA is used for both of HEA and MPEA in this paper). They possess a single phase with a face-centered cubic (FCC), body-centered cubic (BCC), or hexagonal close-packed (HCP) structure [3]. Since then, HEAs have attracted considerable attentions and research interests [4–7]. Due to the high entropy effect, lattice distortion effect, hysteretic effect among alloying elements, HEAs show excellent characteristics different from conventional alloys in mechanical properties [8–10], high temperature properties [11–13], corrosion resistance properties [14–16], and magnetic properties [17–19].

A few of theoretical methods [20] have been used to design HEAs, such as CALPHAD method [21,22], ab-initio calculations [23,24], and Monte Carlo Simulation [25,26]. These methods have reliable theoretical basis for the design of HEAs. However, they are limited to simple cases due to complexity, time-consuming process, and/or low efficiency. Alternatively, Zhang et al. [27,28] and Senkov et al. [29] proposed an empirical method to predict the formation of phases in HEAs. Poletti et al. [30] proposed Electronic parameters for alloys (e.g., electronegativity, valence electron concentration (VEC), itinerant electron concentration) to improve the formulation of HEAs. However, the accuracy for distinguish of different phases is far below satisfactory. Experimentally, the formation of different phases depends on the preparation process. The preparation methods [31] for HEAs include the melt-cast process, powder metallurgy, melt spinning, and deposition techniques [32,33]. In the procedure of alloy prepared, the cost, processing ability, and the experimental complexity need to be considered. Despite these difficulties, quite a few of meaningful data for HEAs have been obtained by theoretical and/or experimental methods.

Computer simulation technology has been widely applied to the design of complex material systems. The long-term accumulation of the high-throughput calculations and experiments provides a meaningful material database. The new computer data processing method that collates the existing data and discovers complex predictive relationship among multiple variables to evaluate the properties of new materials has already become an important new path for material design. In the past few years, machine learning (ML), one of the data processing methods, has been used to design new materials and predict various performances of materials [34–40]. Pilania et al. [41] demonstrated a systematic feature-engineering approach and a robust learning framework based on ML for accurate predictions of electronic bandgaps of double perovskites. Ubaru et al. [42] used ML methods such as sensitivity analyze, least absolute shrinkage and selection operator (LASSO) based methods, and Support Vector Machine to predict the formation enthalpies of binary intermetallic compounds. Choudhury et al. [43] classified the HEAs based on several ML algorithms such as K-nearest neighbor (KNN), support vector machine (SVM), logistic regression, naïvebased approach, decision tree and neural network. Gong et al. [44] classified superheavy elements based on ML, and found the relationship between atomic data and classification of elements. Huang and Islam et al., and Zhuang's group [45,46] have predicted the phases of HEAs based on ML methods including KNN, SVM, and artificial neural network (ANN). They concluded that the trained ANN model is the best and thus the most useful in predicting the formation of new HEAs. Zhou et al. [47] have compared the sensitivity measures of the 13 design parameters based on the result of the ANN model. As aforementioned, many initiate researches on materials design by ML have been made. However, there is ample room for improvement in the construction of data samples for alloy systems, the generalization ability, the learning effectively, and the accuracy for models.

In the implementation of ML methods, an important question is how to select relevant and effective features of alloys. The feature represents the basic attributes for alloy or constituent elements of alloy system. The properties include the thermodynamic properties (e.g., enthalpy, entropy), the atomic radius, VEC [48], parameter Ω, and atom size difference δ [28] etc. In the empirical prediction of alloy phases, the empirical rules are summarized as $\Omega \geq 1.1$ and $\delta \leq 6.6\%$ for the solid solution phase. However, the discrimination of phases is not good enough, especially for the mixture of different phases. In this work, the thermodynamic properties of HEAs calculated with Miedema theory and atomic attributes were used to establish a dataset for HEAs. We considered nine parameters, including mixing enthalpy of amorphous phase (H_{AM}), formation enthalpy of intermetallic compound phase (H_{IM}), formation enthalpy of solid solution phase (H_{SS}), elastic energy of alloy (H_{E}), mixing enthalpy of liquid phase (H_{L}), mixing entropy of alloy (S_{mix}), weighted melting temperature of alloy (T_{m}), atomic size difference δ and parameter Ω. In addition, several ML algorithms were applied to select feature, train data, model and predict different phases for HEAs.

2. Methods

2.1. Establishing the dataset

The dataset is firstly built up from Refs [15, 28, 29, 49–53] containing 556 entries. After the removal of the duplicated data, the new dataset is composed of 407 HEAs, consisting of 215 solid solutions (SS), 12 intermetallic compounds (IM), 142 mixtures of solid solutions and intermetallic compounds (SS + IM), and 38 amorphous alloys (AM). The nine corresponding properties for HEA were used for feature variables in ML.

Ouyang's model [54] based on Miedema theory [55], which has a good prediction [56–59] of the formation enthalpies for multicomponent alloys, was used to predict the thermodynamic properties, such as H_{AM}, H_{IM}, H_{SS} and H_{L}. The formation or mixing enthalpy for binary alloy is calculated by Miedema theory as:

$$\Delta H_{ij}^{C}\left(y_i, y_j\right) = y_i y_j\left(f_j^i \Delta H_{i\ in\ j} + f_i^j \Delta H_{j\ in\ i}\right), \tag{1}$$

$$f_j^i = y_j^s\left[1 + \gamma\left(y_i^s y_j^s\right)^2\right], \tag{2}$$

$$y_i^s = \frac{y_i V_i^{2/3}}{y_i V_i^{2/3}} + y_j V_j^{2/3}, \tag{3}$$

$$\Delta H_{i\ in\ j}^{C} = \frac{P V_i^{2/3}}{2\left(n_i^{-1/3} + n_j^{-1/3}\right)}\left[-(\Delta\varphi)^2 + \frac{Q}{P}\left(\Delta n^{1/3}\right)^2 - \alpha\frac{R}{P}\right], \tag{4}$$

$$V_i^{2/3}(alloy) = V_i^{2/3}(pure)\left[1 + a_i f_j^i\left(\varphi_i - \varphi_j\right)\right], \tag{5}$$

where V, φ, and n are mole volume, electron chemical potential and electronic density at the Wigner-Seitz cell boundary, respectively. P, Q, R, α, γ and a are empirical parameters, in which $Q/P = 9.4$, $\alpha = 0.73$ for a liquid alloy, $\alpha = 1$ for a solid alloy. $\gamma = 0$ for random status (i.e. liquid and solid solution phase), $\gamma = 5$ for amorphous phase, $\gamma = 8$ for intermetallic phase, respectively. The description of all abovementioned parameters referred to Ref. [55].

As for the binary alloy, the elastic energy was estimated by the following formulas

$$\Delta H^{e}_{i\ \mathrm{in}\ j} = \frac{2B_iG_j\left(V_i - V_j\right)^2}{3V_jB_i + 4V_iG_j}, \tag{6}$$

$$\Delta H^{e}_{ij} = \frac{y_iy_j\Delta H_{i\ \mathrm{in}\ j}{}^{e}\Delta H_{j\ \mathrm{in}\ i}{}^{ee}}{y_i\Delta H_{i\ \mathrm{in}\ j}} + y_j\Delta H^{e}_{j\ \mathrm{in}\ i}, \tag{7}$$

where B and G are the bulk modulus and shear modulus, respectively.

On the basis of the properties for binary alloys, the thermodynamic properties for the HEAs were calculated by the extended geometric model [54].

The features S_{mix}, T_{m}, δ and Ω can be calculated as following.

$$S_{\mathrm{mix}} = -R\sum_{i=1}^{n} x_i\ \ln x_i, \tag{8}$$

$$T_{\mathrm{m}} = \sum_{i=1}^{n} x_i(T_{\mathrm{m}})_i, \tag{9}$$

$$\delta = 100 \times \sqrt{\sum_{i=1}^{n} x_i\left(1 - r_i/\sum_{i=1}^{n} x_ir_i\right)^2}, \tag{10}$$

$$\Omega = \frac{T_{\mathrm{m}}S_{\mathrm{mix}}}{\mid H_{\mathrm{SS}} \mid}, \tag{11}$$

where x_i, $(T_m)_i$, and r_i refer to the atomic concentration, melting temperature and atomic radius of the ith element, respectively. R in Eq. (8) denotes the gas constant. H_{SS} is the formation enthalpy of solid solution phase for HEAs.

2.2. Feature selection

The feature selection is often used to reduce feature space dimensionality and remove noisy and redundant features [60–62]. It aims to select a small subset of original features that minimize redundancy and maximize relevance, and it is superior in terms of better readability and interpretability.

Pearson correlation coefficient (PCC) [63] belonging to statistical index is expressed as:

$$\frac{r_{x,y} = \sum_{i=1}^{n}(x_i - \overline{x})\sum_{i=1}^{n}(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2},} \tag{12}$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \overline{y} = \frac{1}{n}\sum_{i=1}^{n}y_i$.The coefficient $r_{x,y}$ ranges from -1 to 1 and is constant for linear transformation of either variables. Pearson correlation coefficient represents the strength of the linear relationship between two random variables x and y. The positive (negative) sign of the correlation coefficient corresponds that these variables correlate directly (inversely), otherwise $r_{x,y} = 0$, meaning they are uncorrelated. The closer the value of $|r_{x,y}|$ is to 1, the stronger the measure is to the linear relationship. This is because correlation measures reflect trends in the expression levels of each pair in the two profiles.

Univariate feature selection [62] helps to determine the strength of the relationship between each feature and the target variable through some statistical methods or ML algorithms such as Chi square, F-test, and Mutual Information. The features are ranked and extracted according to the strength of relationship. The screened feature variables should be used for the training, testing and verifying for ML models. Therefore, univariate feature selection is often used as a preprocessor before applying the estimation model to the dataset. Compared to Pearson correlation coefficient, univariate feature selection has better performance for the discrete data.

Stability selection [64], which is based on subsampling in combination with selection algorithms (e.g., regression, SVM), is a relatively novel method for feature selection. The high-level idea is to apply the feature selection algorithm to different subsets of data and different subsets of features. After repeating the process several times, the selection results can be aggregated. The strong features will have scores close to 100%; the weak relevant features will have non-zero scores; and the irrelevant features will have scores (close to) zero. In this paper, the randomized Lasso algorithm was used to estimate the stability.

2.3. Machine learning algorithms

To predict phases of HEAs, the key topic is to seek the relationship between some properties of alloys and the corresponding phases, and then different alloy phases can be distinguished. In fact, this is a classification problem. The relationship may be explicit (e.g., functional expression) or implicit (e.g., mapping matrix). The aforementioned empirical methods belong to explicit functions. By contrast, the ML algorithm is able to get the implicit mapping matrix. The classification algorithms such as Multilayer perceptron (MLP), SVM and Gradient boosting decision tree (GBDT) were used for the predictions.

2.3.1. Multilayer perceptron

Multilayer perceptron [65] is considered feed-forward neural networks because all data flow in only one direction, from input to output units. MLP viewed as a universal approximator is very fast and easy to use. MLP is given as follows:

$$y = \varphi(x)\left(\sum_{i=1}^{n} w_ix_i - \theta\right), \tag{13}$$

where x_i, for $i = 1, 2, \ldots, n$, are inputs, y is output, w_i is the weights with the ith input, and θ is a threshold. Most often $\varphi(x)$ is $(1 + e^{-x})^{-1}$.

This neural network consists of multiple processing units. Each unit performs a biased weighted sum of its inputs and passes this activation level through a transfer function to generate output, and the units are arranged in a layered feed-forward topology. The learning of MLP is accomplished by adjusting the weights of connections between neurons. MLP is a nonlinear classifier and is suitable for handling discrete data.

2.3.2. Support vector machine

Support vector machine [66] is one of the binary classifiers based on maximum margin strategy, which is a concise and effective classification method based upon statistical learning theory. SVM maps input vectors into a high dimensional feature space to obtain the optimal separation hyperplane. SVM was originally used for linear classification with margin, and was extended to nonlinear examples until the nonlinear separation problem was transformed into a high dimensional feature space. The separation hyperplane is determined by the support vector, so it has strong robustness to outliers and is more suitable than other classification algorithms for dealing with unbalanced class problems.

In the linearly separating case, the decision surface equation of the separating hyperplane can be written [67] as

$$\mathbf{w}^T\mathbf{x} + b = 0, \tag{14}$$

where $\mathbf{x}$ is input vector, $\mathbf{w}$ is an adjustable weight vector and b is bias.

In the nonlinearly separating case, the decision surface equation of the separating hyperplane can be written [67] as

$$\sum_{i=1}^{N_s} \alpha_id_ik(\mathbf{x}, \mathbf{x}_i) = 0, \tag{15}$$

where α_i is a Lagrangian multiplier, d_i is expectation response, and $k(\mathbf{x}, \mathbf{x}_i)$ is called kernel function. Different kernel functions, including linear kernel ($\mathbf{x}^T\mathbf{x}'$), polynomial kernel ($(\mathbf{x}^T\mathbf{x}' + 1)^d$), RBF kernel ($\exp(-\gamma||\mathbf{x}\text{-}\mathbf{x}'||^2)$) and sigmoidal kernel ($\tanh(\gamma\mathbf{x}\mathbf{x}' + C)$) can be used in SVMs for the nonlinear problem.

2.3.3. Gradient boosting decision tree

Gradient boosting decision tree [68,69] is an ensemble model of decision trees, which is trained in sequence and learned by fitting the negative gradient. GBDT is an enhancement algorithm originally used for regression task. GBDT can also be used for classification tasks by using suitable loss functions. In order to avoid over-fitting, it is very important to choose the correct number of iterations in the gradient boosted forest. Setting it too high may result in overfitting, and setting it too low may result in under-fitting. GBDT over-fitting can also be greatly reduced through random sampling training.

With the training dataset $\{\mathbf{x}_i, y_i\}$, the approximation function can be expressed as [69]:

$$F_k(\mathbf{x}) = F_{k-1}(\mathbf{x}) + \gamma_k h_k, \tag{16}$$

where the corresponding training dataset of decision tree h_k is $\{\mathbf{x}_i, -\frac{\partial L}{\partial \hat{y}_i}[F_{k-1}(\mathbf{x}_i), y_i]\}$, and γ_k is $\arg\min_\gamma \sum_{i=1}^{n} L(y_i, F_{k-1}(\mathbf{x}_i) + \gamma h_k)$. It indicates the update rate for GBDT. With its increasing depth the decision tree constantly corrects the errors left by the previous model, thus it improves the prediction effect of GBDT.

2.3.4. Kernel principal component analysis

Kernel Principal Component Analysis integrates kernel function on traditional linear Principal Component Analysis (PCA), and it is helpful to solve the nonlinear problems. PCA is a powerful technique for extracting a structure from potentially high-dimensional datasets, and KPCA [70,71] calculates principal components in a high-dimensional feature space that is nonlinearly related to the input space. By adjusting parameters of KPCA, KPCA can achieve dimensionality reduction and expanding dimensions of input data. The kernel function in KPCA is similar to the kernel function used in SVM. In this paper, the poly kernel function was used to predict the classification.

3. Results and discussion

3.1. Feature selection

PCCs have been calculated to analyze the correlation between the properties (i.e. H_{AM}, H_{IM}, H_{SS}, H_E, H_L, S_{mix}, T_m, δ, Ω) for HEAs and different phases in Fig. 1. From Fig. 1, the PCCs for H_{SS} and Ω are close to zero in the first column. This indicates that H_{SS} and Ω are irrelevant with phases. The PCC for T_m is very small, and it is also irrelevant. The PCCs for H_{AM}, H_{IM}, H_L and S_{mix} are around 0.3, which indicates there is a certain degree of relevance. The absolute value of PCCs for H_E and δ are larger than 0.5, which shows H_E and δ are strongly relevant.

On the other hand, the values of mutual PCCs for H_{AM}, H_{IM}, H_{SS} and H_L are very close to 1, indicating that they are strongly relevant. The four formation enthalpies have been calculated on the basis of Miedema theory (Eqs. (1)–(5)) and extended geometric model. In these equations, the expressions are similar and just several parameters (i.e. α and γ) should be changed in different phases. Even so, the effect of the changes for parameters (i.e. α and γ) is still small, and the expressions show strong relevance in the mathematical sense. It can be seen from the above correlation analysis for the PCCs of nine feature variables that the parameters H_{SS}, T_m and Ω are redundant variables.

However, the PCC can only describe the linear dependency between variables. If there is a nonlinear correlation between variables, the result for PCC is poor. Therefore, other methods should be used to further evaluate the correlation for feature variables.

The correlation of phases between feature variables is illustrated in Fig. 2. These correlations have been evaluated by PCC, univariate feature selection, and stability selection. The larger the values for univariate feature selection and stability selection are, the stronger the correlation is. For the univariate feature selection, the values of H_{SS}, T_m and Ω are significantly smaller than the rest, which indicates that these feature

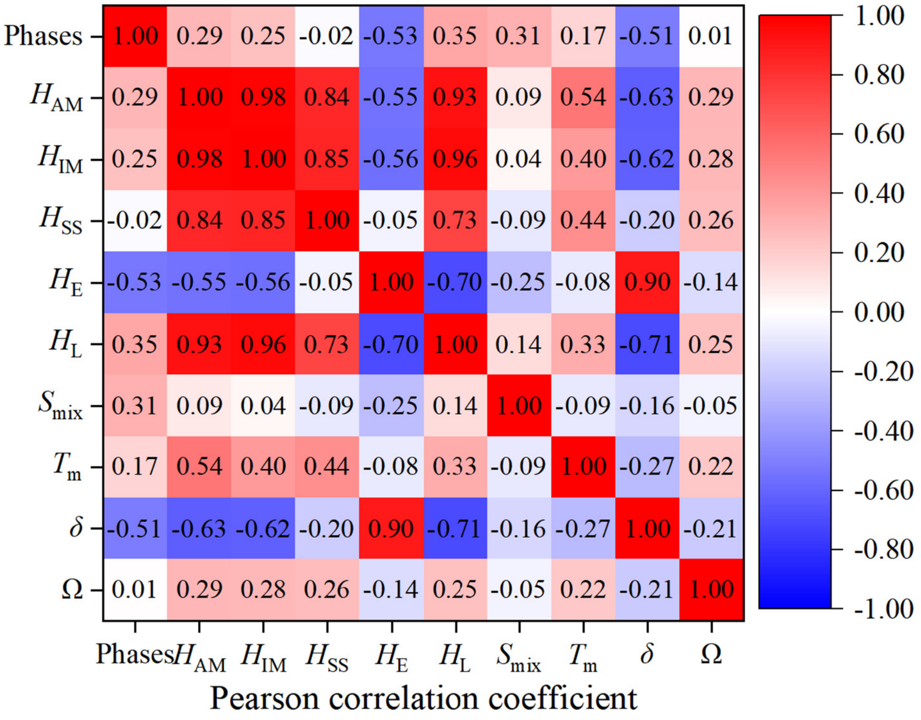

	Phases	H_{AM}	H_{IM}	H_{SS}	H_E	H_L	S_{mix}	T_m	δ	Ω
Phases	1.00	0.29	0.25	-0.02	-0.53	0.35	0.31	0.17	-0.51	0.01
H_{AM}	0.29	1.00	0.98	0.84	-0.55	0.93	0.09	0.54	-0.63	0.29
H_{IM}	0.25	0.98	1.00	0.85	-0.56	0.96	0.04	0.40	-0.62	0.28
H_{SS}	-0.02	0.84	0.85	1.00	-0.05	0.73	-0.09	0.44	-0.20	0.26
H_E	-0.53	-0.55	-0.56	-0.05	1.00	-0.70	-0.25	-0.08	0.90	-0.14
H_L	0.35	0.93	0.96	0.73	-0.70	1.00	0.14	0.33	-0.71	0.25
S_{mix}	0.31	0.09	0.04	-0.09	-0.25	0.14	1.00	-0.09	-0.16	-0.05
T_m	0.17	0.54	0.40	0.44	-0.08	0.33	-0.09	1.00	-0.27	0.22
δ	-0.51	-0.63	-0.62	-0.20	0.90	-0.71	-0.16	-0.27	1.00	-0.21
Ω	0.01	0.29	0.28	0.26	-0.14	0.25	-0.05	0.22	-0.21	1.00

Fig. 1. Pearson correlation coefficients with the nine thermodynamics variables. The nine parameters examined by the extended Miedema theory are mixing enthalpy of amorphous phase (H_{AM}), formation enthalpy of intermetallic compound phase (H_{IM}), formation enthalpy of solid solution phase (H_{SS}), elastic energy of alloy (H_E), mixing enthalpy of liquid phase (H_L), mixing entropy of alloy (S_{mix}), weighted melting temperature of alloy (T_m), atom-size difference (δ), and parameter Ω.

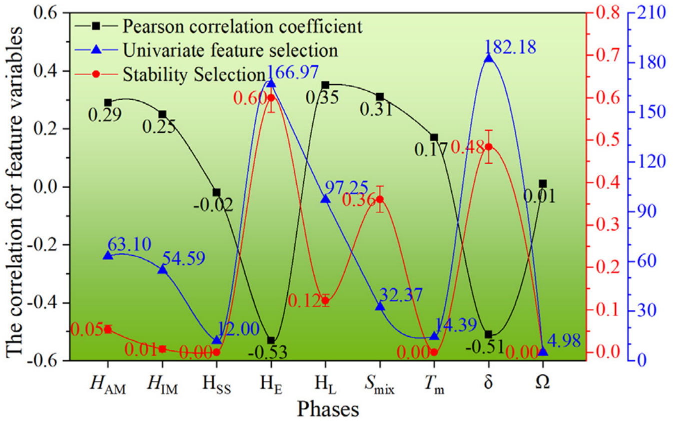


Fig. 2. The correlation for feature variables evaluated by Pearson correlation coefficient, univariate feature selection and stability selection.

variables are strongly relevant with phases. For the stability selection, the value of H_{SS}, T_m and Ω are almost zero, and the value of H_{IM} is very close to zero. This indicates the corresponding feature variables are irrelevant with phases, and the remaining variables are strongly relevant with phases.

From the comparison of the above three methods for feature selection, the results of H_E and δ are consistently strong relevance. According to the Hume-Rothery rules [72] for the solubility in binary alloy systems, the atomic size and the formation enthalpy will affect the formation of the solid solution phase. First, if the atomic size difference of the constituent elements of alloy is >15%, it is the most improbable to form solid solution. Second, compared with solid solution, the more negative formation enthalpy is, the more likely the alloys form intermetallic compound. On one hand, the parameter δ proposed by Zhang et al. [27,28] indicates the size effects for the component of alloys, with $\delta \le 6.6\%$. Takeuchi and Inoue [73,74] also proposed similar criterion. The lager difference of atomic size for component in alloys can result in disordered arrangement of atoms, and benefit on amorphous forming. On the other hand, the elastic energy is positive, and the formation enthalpy is the sum of chemical enthalpy and elastic energy. Thus the elastic energy can shift the formation enthalpy towards positive. The formation enthalpy of small magnitude benefits to form solid solution. In view of this, we classify different phases by two variables with H_E and δ, as follows.

The relationship between H_E and δ with different alloy phases is displayed in Fig. 3. Surprisingly, both 3D scatter plot and the projection drawing of H_E and δ are disable to classify phases. The solid solution phase (i.e. HEA) and the mixture of solid solution and intermetallic phase overlap each other, suggesting that the parameters of H_E and δ are not enough to distinguish different phases. Some other properties also contribute to the formation of phase. It is important that a number of properties for HEAs must be used to establish efficient dataset from feature variables.

Furthermore, the feature variables of H_{AM}, H_E, H_L, S_{mix} and δ are relevant with phases, these thermodynamic properties are efficient feature variables, and the value of variance for δ is small and indicates high consistency. H_{SS}, T_m and Ω are irrelevant with phases, and these properties are redundant. This result is contrary to Zhang's point of view [27,28] where the parameter Ω is important for the forming of solid solution, and the larger Ω would facilitate the formation of solid solution. This discripancy may be attribute to the fact that the parameter Ω calculated indirectly from T_m, S_{mix} and H_{SS}, but T_m and H_{SS} are redundant, and S_{mix} is not strongly relevant with the phases, resulting in that the parameter Ω is irrelevant. The results of H_{IM} in different methods are discrepant, so it is impossible to be eliminated. Therefore, the efficient subset can consist of the feature variables with H_{AM}, H_{IM}, H_E, H_L, S_{mix} and δ.

3.2. Classification by machine learning algorithms

The models of MLP, SVM and GBDT were used to predict the phase with the new subset, which consists of six feature variables. The learning curve for *k*-fold cross validation was used to select models and evaluate the performance of fitted models. 30% of the dataset was extracted randomly as the testing set (hereinafter the proportion of the test set is 30%), and the numbers of *k*-fold are 10 (hereinafter the value of *k*-fold for cross validation is 10). The learning curves based on six feature variables (6 V) predicted by adjusting the parameters with different models are shown in Fig. 4(a).

From Fig. 4(a), all learning curves are convergent, and there is slight overfitting in the models. However, the curves are still steep, indicating that the models have a fast learning rate. The evaluated result of prediction model for SVM is better than those for MLP and GBDT. The learning curve for MLP is not stable, and that for GBDT is slow in converging. The accuracy of SVM with its faster convergence rate and higher stability is about 0.75 for the testing data. In addition, the dataset only has 407 samples, which leads to instability of the learning curve for MLP, but the models for SVM and GBDT are not sensitive to the number of samples.

The accuracy 0.75 is obviously not satisfactory. It is challenging to improve the accuracy of prediction. Previously, the method of reducing the feature variables has been used to optimize the data structure. The low evaluation accuracy is not enough to predict the different phases of HEA. Conversely, does increasing the feature variable improve prediction accuracy? Zhou et al. [47] used thirteen parameters to build subset and train the ML model. Tancret et al. [75] trained ML model based on nine physical parameters, Zhang et al. [76] used fourteen empirical materials descriptors to train the ML model. The more feature variables, the more information can be provided. But the redundant feature variables still cannot be increased. Kernel Principal Component Analysis (KPCA) can be used to expand dimension and increase feature variables. The KPCA model was optimized by adjusting parameters, and the feature variables with 6 dimensions is expanded to 11 dimensions. The preprocessed data can be used to train the prediction model. The learning curves for prediction model based on six feature variables and KPCA (6V-KPCA) are displayed in Fig. 4(b). Compared to Fig. 4(a), the values of predictive accuracy for train and test increase substantially. The

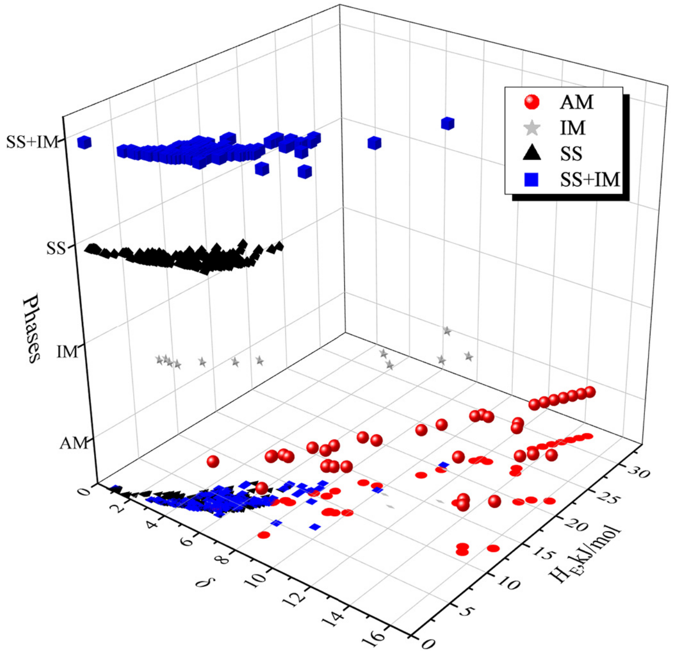


Fig. 3. The relationship between H_E and δ with different alloy phases. The H_E indicates the elastic energy of HEA, and δ indicates the atom-size difference.

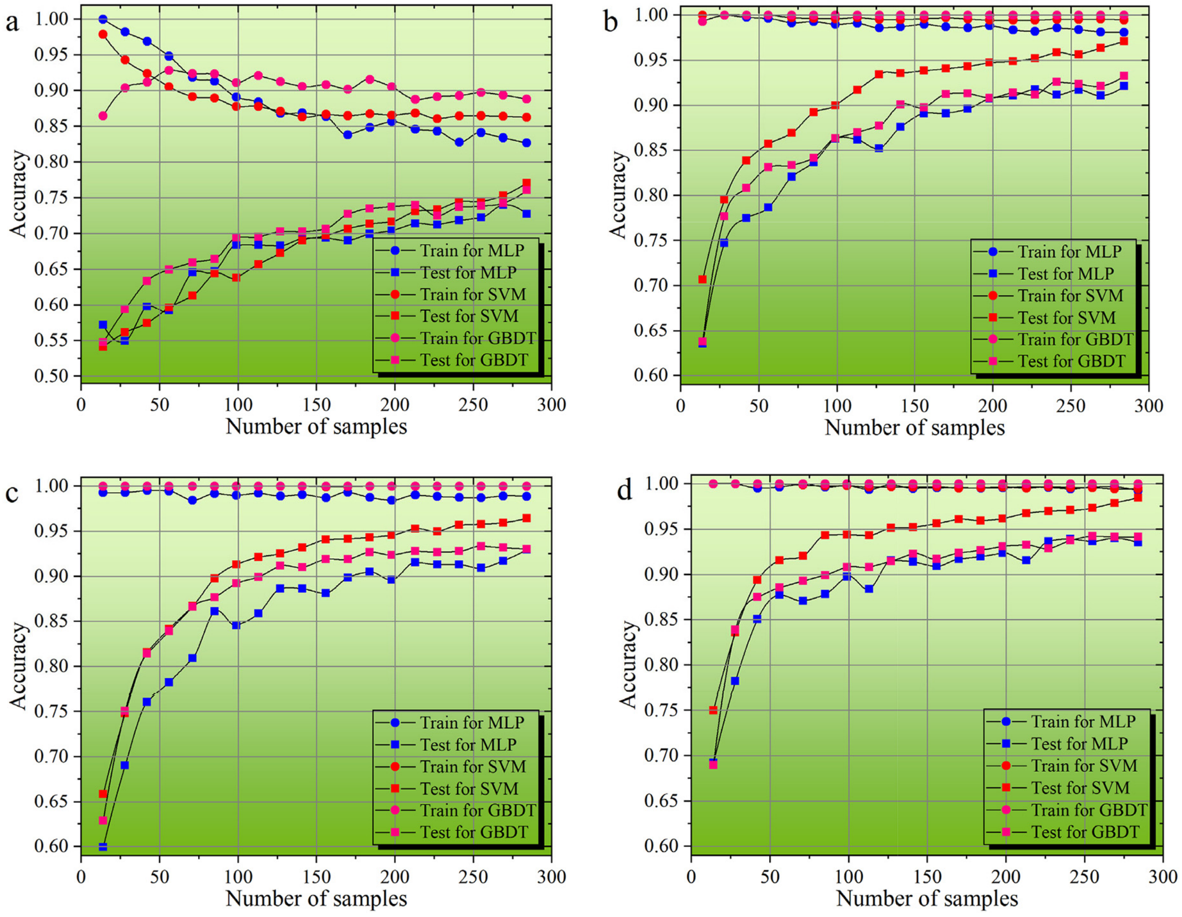


Fig. 4. The learning curves for different predictive models: (a) The model based on six feature variables (6V); (b) The model based on six feature variables and KPCA (6V-KPCA); (c) The model based on nine feature variables and KPCA (9V-KPCA); (d) The model based on four feature variables and KPCA (4V-KPCA).

accuracies of training set and testing set converge to a higher value. The accuracy of testing set for SVM model is close to 0.975. Thus, expanding dimension benefits to distinguish different phases. The preprocessing of KPCA provides more valuable information for the classification in ML models.

Our aforementioned results suggest that expanding dimensions can improve the predictive accuracy. It is equivalent to increasing the number of variables from a certain perspective. Does that mean the more variables will lead to better predictive results by expanding dimensions? In the following, we make an attempt to predict phases useing different ML models based on nine feature variables and KPCA (9V-KPCA). The KPCA model was optimized by adjusting parameters. The feature variables with 9 dimensions is expanded to 20 dimensions. In Fig. 4(c), the learning curves perform well, but the results are still not as good as those in Fig. 4(b). The predictive accuracy of SVM model is <0.975. Therefore, the feature selection is necessary. The redundant feature variable can interfere with the ML model and degrade the predictive accuracy.

The strongly relevant feature variable has great influence on prediction. Furthermore, we analyze the above six feature variables in Fig. 2. The variables H_E and δ are relevant consistently in the different feature selection models. H_L and S_{mix} must be preserved because of relevance. In contrast, H_{AM} and H_{IM} show a certain relevant in PCCs. The rest show little or even no correlation in other evaluation matrixes. The PCC performs better in the evaluation of linear correlation. For the present nonlinear problem, the error of the evaluation by univariate feature selection and stability selection might be smaller. Thus the efficient

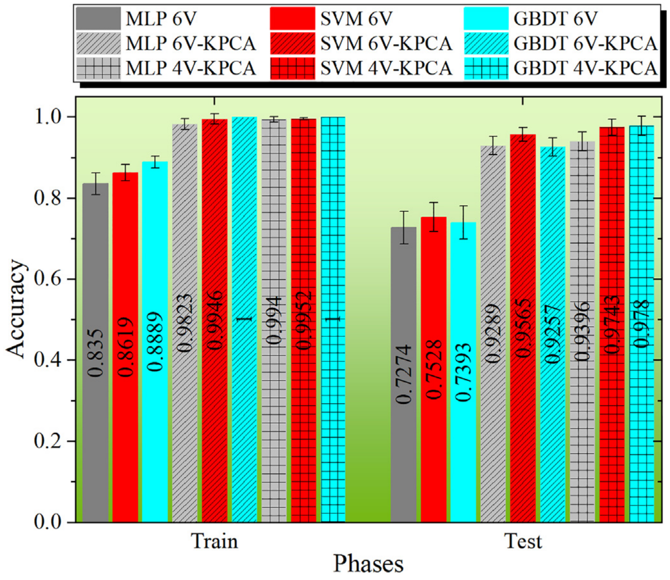


Fig. 5. The predictive accuracies of different preprocessing methods and different models. 6V indicates the six feature variables are unprocessed; 6V-KPCA indicates that the six feature variables are processed by Kernel PCA; 4 V-KPCA indicates that the four feature variables are processed by Kernel PCA. MLP indicates multilayer perceptron model; SVM indicates support vector machine model; GBDT indicates gradient boosting decision tree model.

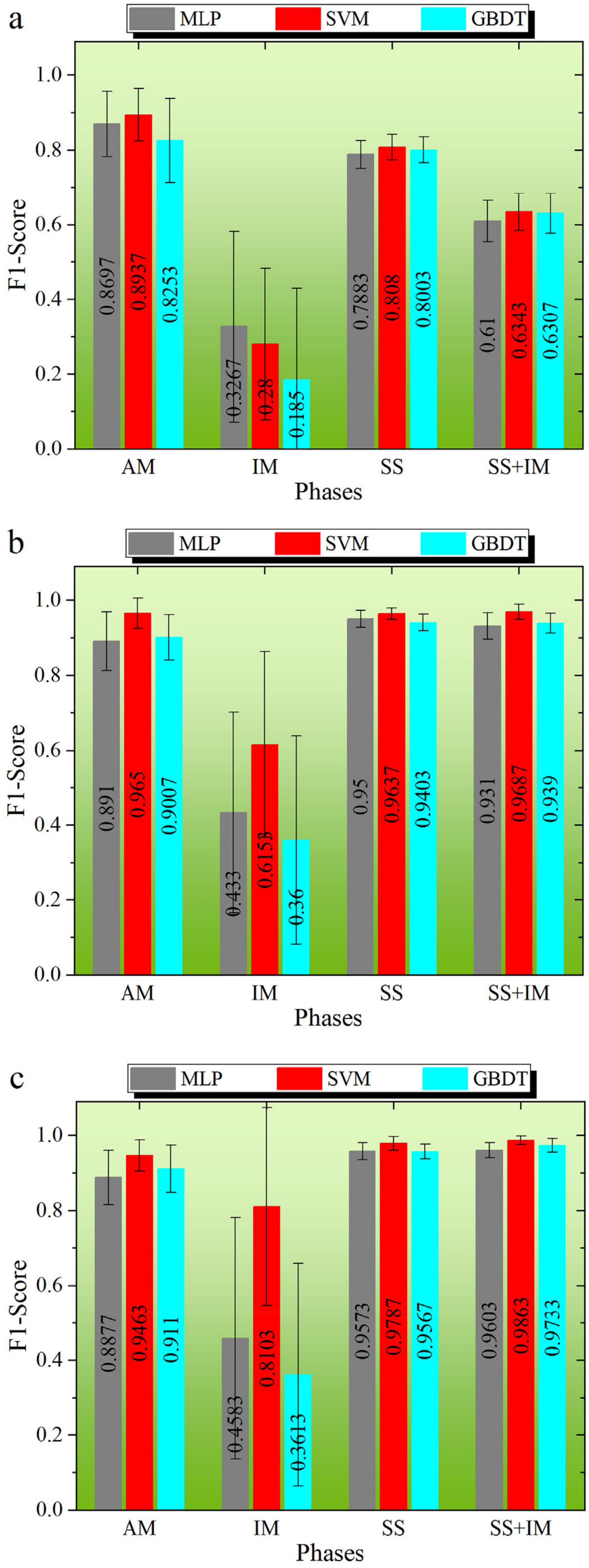


Fig. 6. The predictive F1-scores of different preprocessing methods and different models. (a) F1-scores were predicted based on the unprocessed six feature variables (6 V); (b) F1-scores were predicted based on the six feature variables processed by Kernel PCA (6 V-KPCA); (c) F1-scores were predicted based on the four feature variables processed by Kernel PCA (4 V-KPCA). MLP indicates multilayer perceptron model; SVM indicates support vector machine model; GBDT indicates gradient boosting decision tree model.

subset is further refined and consist of the feature variables with H_E, H_L, S_{mix} and δ.

The KPCA model was optimized by adjusting parameters, and the feature variables with 4 dimensions is expanded to 13 dimensions. The learning curves for prediction model based on four feature variables and KPCA (4V-KPCA) are illustrated in Fig. 4(a). Compared with Fig. 4(b) and (c), the learning curve of 4V-KPCA performs best in both predictive accuracies, convergence and learning rate. In particular, the predictive accuracy of SVM model attains as much as 0.98.

In order to reduce the random error, the learning processes were carried out 30 times and the results were averaged. The predictive accuracy of different phases is illustrated in Fig. 5 with the models of MLP, SVM and GBDT. The overall performances are consistent with the above mentioned methodologies. The results for SVM and GBDT are better than that for MLP. Among the preprocessing methods, 4 V-KPCA performs best by the optimized feature variables and KPCA. The worst predictive results were obtained by six unprocessed feature variables. The accuracies of testing set predicted by SVM, GBDT and MLP for 4 V-KPCA are 0.9743, 0.9780 and 0.9396,and those for training set are 0.9952, 1 and 0.994. The worst fitting for MLP may be caused by its instability. The difference of the accuracies predicted by GBDT based on 4 V-KPCA between training set and testing set is 0.022, and is bigger than that for SVM. The model of SVM shows better convergence. From the variance perspective, the smaller variance means the more stablilty of the model. The results of variance predicted by MLP based on 4 V-KPCA for both training set and testing set are not good enough.

From Figs. 4-5, the results reflect that the increment of the number of precise feature variable improves the predictive accuracy. The negative effect of redundant feature variable will be further magnified by expanding dimension of KPCA. The more dimensions KPCA expands, the more complex the relationships between them become. A model with too many variables ends up degrading the results. Therefore, it is essential to select the most important feature variables and appropriately expand dimensions.

The predictive accuracies of SVM and GBDT in Fig. 4(d) and 5 are slightly different. It is convenient and fast to evaluate the predictive effect of the ML models using accuracy, which, however, is too simple and rough. For imbalanced data, the evaluation of F1-score [77] which comprehensively considered the precision and recall will be more sensitive. The evaluation results of the classification prediction for each phase with F1-score predicted by 6 V, 6 V-KPCA and 4 V-KPCA methods are given in Fig. 6.

There are remarkable differences in the predictive F1-scores for different phases and methods. Compared with the predictive accuracies, the results of F1-score for 4 V-KPCA still perform the best. We analyze Fig. 6(c) in detail. The F1-score of IM phase is the lowest. The value of variance for IM is larger than those of the rest. The predictive F1-score for amorphous phase ranges from 0.8877 to 0.9463, and that for solid solution phase ranges from 0.9567 to 0.9787. The F1-score for solid solution and intermetallic phase performs best and ranges from 0.9603 to 0.9863. This result indicates that the amorphous phase, solid solution phase, and solid solution and intermetallic phase can be separated by ML models. The predictive F1-score for intermetallic compound ranges from 0.3613 to 0.8103. The F1-score is significantly lower than those obtained in other phases. The main reason is probably the following. On one hand, the number of samples for intermetallic compound is too small (i.e. 12). This is called imbalance of samples in ML. Too little sample data leads to underfitting of the model. The largest variance also confirms the imbalance of samples. On the other hand, from Fig. 4(d), the model has low predictive accuracy when the number of samples is small. It is noteworthy that the model for SVM still fits better than the other two when the number of samples is very small, this is inseparable from the preprocessing feature variables by KPCA. For intermetallic phase, the intermetallic is often precipitated in small amounts in solid solution matrix. Then the present model does not fit well. However, the predictive accuracy of 0.8103 for SVM is still significant for the prediction of mixtures of solid solution and intermetallic phase.

From the thermodynamic point of view, formation enthalpy plays an important role in formation of phase. According to Miedema theory, the calculation of enthalpy for different phases is different. But the enthalpy of H_{AM} and H_{IM} is mainly composed of chemical enthalpy (i.e. H_L). The chemical enthalpy represents the combined effect of interatomic interactions during alloying under different atoms and structures in the alloy. However, H_{AM} contains chemical enthalpy and topological energy, the express of H_{IM} is very similar to that of H_L. The difference between them is small, indicating that the interatomic interactions in different phase are similar. The alternative is chemical enthalpy, where it contributes significantly to the formation of phase. So it is important that H_L is retained in the feature selection. The elastic energy H_E and atomic size difference δ are closely related to the atomic radius and have the same role. H_E involves the effect of atomic size, and also involves interactions between atoms of the same structure. Both H_E and δ are important. S_{mix} is a basic thermodynamic property (chemical mixing entropy) for HEA, thus retained in the feature selection.

Takeuchi and Inoue [73] proposed empirical rule including the chemical enthalpy ΔH^C and the normalized mismatch entropy S_σ/k_B. Bhatt et al. [78] and Rao et al. [79] further developed empirical criterion including the chemical enthalpy ΔH^C, normalized mismatch entropy S_σ/k_B, the configurational entropy (S_c/R) and $\Delta H^C(S_\sigma/k_B)(S_c/R)$. In our previous work [57], $\Delta H^C(S_\sigma/k_B)(S_c/R)$ has been used to predict the amorphous forming composition ranges of Al-Fe-Nd-Zr system. Furthermore, the chemical enthalpy is actually the mixing enthalpy of liquid phase (H_L). The mismatch entropy S_σ can be calculated by the equation proposed by Mansoori [80]. In other word, S_σ can be calculated by atomic radius and has a strong correlation with δ. S_c is actually S_{mix}. Therefore, the thermodynamic properties of chemical enthalpy ΔH^C, mismatch entropy S_σ and configurational entropy S_c can be used to distinguish different phases.

Besides in distinguishing between crystalline and amorphous states, the ΔH^C, S_σ and S_c play important roles in phase formation. This result is consistent with the previously optimized feature variables of H_E, H_L, S_{mix} and δ. However, it's worth noting the empirical criterions from Refs [73, 78, 79] could not predict the different phases well. The functions of empirical criterions are still simple and can not to distinguish more complex phases. The deviation of the results in Refs [57,79] indicates that the empirical criterion is not robust. A better criterion should be developed to predict different phases of HEA.

4. Conclusions

The relationship between the phases and the nine thermodynamics properties of HEAs is examined using machine-learning method accompany with extended Miedema theory. The thermodynamic properties of HEAs were calculated by Miedema theory and geometric model. These data were used as feature variables to establish feature dataset for ML. The relative importance of the nine feature properties was evaluated by the feature selection with Pearson correlation coefficient, univariate feature selection, and stability selection. The parameters of H_E and δ have a strong relevance with phases. However, it need more parameters to distinguish different phases. After removing irrelevant feature variables, the new subset consists of H_{AM}, H_{IM}, H_E, H_L, S_{mix} and δ.

The ML models of MLP, SVM and GBDT were used to build implicit mapping matrix and classify the dataset. The model of RBF kernel SVM evaluated by learning curve with *k*-fold cross validation has the best fitting. The predictive accuracy based on the six feature variables is 0.7528, which is below satisfactory. The four feature variables of H_E, H_L, S_{mix} and δ has been optimized and preprocessed with expanding dimensions by Kernel PCA. The different preprocessed method of KPCA and the different models of MLP, SVM and GBDT were compared. The model of SVM for 4 V-KPCA is the best overall due to higher stability and convergence.

The imbalance of sample leads to the worst fitting of intermetallic phase for the various models. The predictive accuracy and F1-score of the model could be improved by increasing the number of samples and the effective relevant feature variables. The expanding dimensions by Kernel PCA improves the predictive results. The predictive accuracies of SVM and GBDT for 4 V-KPCA are over 0.97. The F1-score of HEA (i.e. SS) for 4 V-KPCA is 0.9787, and that for amorphous phase, mixture of solid solution and intermetallic phase and intermetallic phase is 0.9463, 0.9863 and 0.8103, respectively. All of them are higher than those from using MLP and GBDT. Therefore, the model of SVM combined KPCA is the best ML model for the phase selection of HEAs in present dataset. H_E, H_L, S_{mix} and δ are the effective and relevant variables. The present ML model is helpful to distinguish different phases in HEAs, and beneficial to the discovery of the new HEAs.

CRediT authorship contribution statement

Lei Zhang: Conceptualization, Investigation, Methodology, Data curation, Formal analysis, Writing - original draft.**Hongmei Chen:**Supervision, Writing - review & editing.**Xiaoma Tao:**Supervision, Writing - review & editing, Funding acquisition.**Hongguo Cai:**Methodology, Data curation.**Jingneng Liu:**Methodology, Data curation.**Yifang Ouyang:**Supervision, Writing - review & editing, Funding acquisition. **Qing Peng:**Supervision, Writing - review & editing.**Yong Du:**Supervision, Writing - review & editing.

Declaration of competing interest

The authors declare no competing financial interest.

Acknowledgements

This work is financially supported by the National Natural Science Foundation of China (Grant Nos. 51531009, 61962006 and 51961007), the Guangxi Natural Science Foundation (Grant Nos. 2016GXNSFBA380166 and 2018GXNSFAA281291), “BAGUI Scholar” Program of Guangxi Zhuang Autonomous Region of China (201979) and the Science Foundation of Guangxi Education Department (Grant Nos. 2017KY1472).

References

[1] J.W. Yeh, S.K. Chen, S.J. Lin, J.Y. Gan, T.S. Chin, T.T. Shun, C.H. Tsau, S.Y. Chang, Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes, Adv. Eng. Mater. 6 (2004) 299–303, https://doi.org/10.1002/adem.200300567.

[2] B. Cantor, I.T.H. Chang, P. Knight, A.J.B. Vincent, Microstructural development in equiatomic multicomponent alloys, Mater. Sci. Eng. A 375-377 (2004) 213–218, https://doi.org/10.1016/j.msea.2003.10.257.

[3] S.H. Joo, H. Kato, M.J. Jang, J. Moon, E.B. Kim, S.J. Hong, H.S. Kim, Structure and properties of ultrafine-grained CoCrFeMnNi high-entropy alloys produced by mechanical alloying and spark plasma sintering, J. Alloy. Compd. 698 (2017) 591–604, https://doi.org/10.1016/j.jallcom.2016.12.010.

[4] W. Zhang, P.K. Liaw, Y. Zhang, Science and technology in high-entropy alloys, Sci. China Mater. 61 (2018) 2–22, https://doi.org/10.1007/s40843-017-9195-8.

[5] Z. Lei, X. Liu, H. Wang, Y. Wu, S. Jiang, Z. Lu, Development of advanced materials via entropy engineering, Scripta Mater 165 (2019) 164–169, https://doi.org/10.1016/j.scriptamat.2019.02.015.

[6] M. Vaidya, G.M. Muralikrishna, B.S. Murty, High-entropy alloys by mechanical alloying: a review, J. Mater. Res. 34 (2019) 664–686, https://doi.org/10.1557/jmr.2019.37.

[7] E.P. George, D. Raabe, R.O. Ritchie, High-entropy alloys, Nat. Rev. Mater. 4 (2019) 515–534, https://doi.org/10.1038/s41578-019-0121-4.

[8] Y. Zhang, Z.P. Lu, S.G. Ma, P.K. Liaw, Z. Tang, Y.Q. Cheng, M.C. Gao, Guidelines in predicting phase formation of high-entropy alloys, MRS Commun 4 (2014) 57–62, https://doi.org/10.1557/mrc.2014.11.

[9] T. Yang, Y.L. Zhao, Y. Tong, Z.B. Jiao, J. Wei, J.X. Cai, X.D. Han, D. Chen, A. Hu, J.J. Kai, K. Lu, Y. Liu, C.T. Liu, Multicomponent intermetallic nanoparticles and superb mechanical behaviors of complex alloys, Science 362 (2018) 933–937, https://doi.org/10.1126/science.aas8815.

[10] L. Lilensten, J.P. Couzinie, L. Perriere, A. Hocini, C. Keller, G. Dirras, I. Guillot, Study of a bcc multi-principal element alloy: tensile and simple shear properties and underlying deformation mechanisms, Acta Mater. 142 (2018) 131–141, https://doi.org/10.1016/j.actamat.2017.09.062.

[11] O.N. Senkov, G.B. Wilks, J.M. Scott, D.B. Miracle, Mechanical properties of $Nb_{25}Mo_{25}Ta_{25}W_{25}$ and $V_{20}Nb_{20}Mo_{20}Ta_{20}W_{20}$ refractory high entropy alloys, Intermetallics 19 (2011) 698–706, https://doi.org/10.1016/j.intermet.2011.01.004.

[12] B. Gludovatz, A. Hohenwarter, D. Catoor, E.H. Chang, E.P. George, R.O. Ritchie, A fracture-resistant high-entropy alloy for cryogenic applications, Science 345 (2014) 1153–1158, https://doi.org/10.1126/science.1254581.

[13] V. Shivam, Y. Shadangi, J. Basu, N.K. Mukhopadhyay, Alloying behavior and thermal stability of mechanically alloyed nano AlCoCrFeNiTi high-entropy alloy, J. Mater. Res. 34 (2019) 787–795, https://doi.org/10.1557/jmr.2019.5.

[14] Y.L. Chou, J.W. Yeh, H.C. Shih, The effect of molybdenum on the corrosion behaviour of the high-entropy alloys $Co_{1.5}CrFeNi_{1.5}Ti_{0.5}Mo_x$ in aqueous environments, Corros. Sci. 52 (2010) 2571–2581, https://doi.org/10.1016/j.corsci.2010.04.004.

[15] Y. Shi, B. Yang, P.K. Liaw, Corrosion-resistant high-entropy alloys: a review, Metals 7 (2017) 43, https://doi.org/10.3390/met7020043.

[16] R.K. Mishra, P.P. Sahay, R.R. Shahi, Alloying, magnetic and corrosion behavior of AlCrFeMnNiTi high entropy alloy, J. Mater. Sci. 54 (2019) 4433–4443, https://doi.org/10.1007/s10853-018-3153-z.

[17] Y. Zhang, T.T. Zuo, Y.Q. Cheng, P.K. Liaw, High-entropy alloys with high saturation magnetization, electrical resistivity, and malleability, Sci. Rep. 3 (2013) 1455, https://doi.org/10.1038/srep01455.

[18] U. Roy, H. Roy, H. Daoud, U. Glatzel, K.K. Ray, Fracture toughness and fracture micromechanism in a cast AlCoCrCuFeNi high entropy alloy system, Mater. Lett. 132 (2014) 186–189, https://doi.org/10.1016/j.matlet.2014.06.067.

[19] O. Schneeweiss, M. Friák, M. Dudová, D. Holec, M. Šob, D. Kriegner, V. Holý, P. Beran, E.P. George, J. Neugebauer, A. Dlouhý, Magnetic properties of the CrMnFeCoNi high-entropy alloy, Phys. Rev. B 96 (2017) 014437, https://doi.org/10.1103/PhysRevB.96.014437.

[20] M.C. Gao, J.W. Yeh, P.K. Liaw, Y. Zhang, High-Entropy Alloys: Fundamentals and Applications, Springer Press, Cham, 2016https://doi.org/10.1007/978-3-319-27013-5.

[21] W.M. Choi, S. Jung, Y.H. Jo, S. Lee, B.J. Lee, Design of new face-centered cubic high entropy alloys by thermodynamic calculation, Met. Mater. Int. 23 (2017) 839–847, https://doi.org/10.1007/s12540-017-6701-1.

[22] J.E. Saal, I.S. Berglund, J.T. Sebastian, P.K. Liaw, Equilibrium high entropy alloy phase stability from experiments and thermodynamic modeling, Scripta Mater 146 (2018) 5–8, https://doi.org/10.1016/j.scriptamat.2017.10.027.

[23] C. Jiang, B.P. Uberuaga, Efficient ab initio modeling of random multicomponent alloys, Phys. Rev. Lett. 116 (2016), 105501. https://doi.org/10.1103/PhysRevLett.116.105501.

[24] Y. Lederer, C. Toher, K.S. Vecchio, S. Curtarolo, The search for high entropy alloys: a high-throughput ab-initio approach, Acta Mater. 159 (2018) 364–383, https://doi.org/10.1016/j.actamat.2018.07.042.

[25] Z. Liu, Y. Lei, C. Gray, G. Wang, Examination of solid-solution phase formation rules for high entropy alloys from atomistic Monte Carlo simulations, JOM 67 (2015) 2364–2374, https://doi.org/10.1007/s11837-015-1508-3.

[26] C. Niu, W. Windl, M. Ghazisaeidi, Multi-cell Monte Carlo relaxation method for predicting phase stability of alloys, Scripta Mater 132 (2017) 9–12, https://doi.org/10.1016/j.scriptamat.2017.01.001.

[27] Y. Zhang, Y.J. Zhou, J.P. Lin, G.L. Chen, P.K. Liaw, Solid-solution phase formation rules for multi-component alloys, Adv. Eng. Mater. 10 (2008) 534–538, https://doi.org/10.1002/adem.200700240.

[28] X. Yang, Y. Zhang, Prediction of high-entropy stabilized solid-solution in multicomponent alloys, Mater. Chem. Phys. 132 (2012) 233–238, https://doi.org/10.1016/j.matchemphys.2011.11.021.

[29] O.N. Senkov, D.B. Miracle, A new thermodynamic parameter to predict formation of solid solution or intermetallic phases in high entropy alloys, J. Alloy. Compd. 658 (2016) 603–607, https://doi.org/10.1016/j.jallcom.2015.10.279.

[30] M.G. Poletti, L. Battezzati, Electronic and thermodynamic criteria for the occurrence of high entropy alloys in metallic systems, Acta Mater. 75 (2014) 297–306, https://doi.org/10.1016/j.actamat.2014.04.033.

[31] Y. Zhang, High-Entropy Materials: A Brief Introduction, Springer Press, 2019https://doi.org/10.1007/978-981-13-8526-1.

[32] F. Zhang, H. Lou, B. Cheng, Z. Zeng, Q. Zeng, High-pressure induced phase transitions in high-entropy alloys: a review, Entropy 21 (2019) 239, https://doi.org/10.3390/e21030239.

[33] Y.J. An, L. Zhu, S.H. Jin, J.J. Lu, X.Y. Liu, Laser-ignited self-propagating sintering of AlCrFeNiSi high-entropy alloys: an improved technique for preparing high-entropy alloys, Metals 9 (2019) 438, https://doi.org/10.3390/met9040438.

[34] P. Raccuglia, K.C. Elbert, P.D.F. 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, https://doi.org/10.1038/nature17439.

[35] Z.K. Liu, Ocean of data: integrating first-principles calculations and CALPHAD modeling with machine learning, J. Phase Equilib. Diff. 39 (2018) 635–649, https://doi.org/10.1007/s11669-018-0654-z.

[36] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature 559 (2018) 547, https://doi.org/10.1038/s41586-018-0337-2.

[37] J.E. Gubernatis, T. Lookman, Machine learning in materials design and discovery: examples from the present and suggestions for the future, Phys. Rev. Mater. 2 (2018) 120301, https://doi.org/10.1103/PhysRevMaterials.2.120301.

[38] C. Wen, Y. Zhang, C. Wang, D. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman, Y. Su, Machine learning assisted design of high entropy alloys with desired property, Acta Mater. 170 (2019) 109–117, https://doi.org/10.1016/j.actamat.2019.03.010.

[39] S.P. Ong, Accelerating materials science with high-throughput computations and machine learning, Comput. Mater. Sci. 161 (2019) 143–150, https://doi.org/10.1016/j.commatsci.2019.01.013.

[40] L. Himanen, A. Geurts, A.S. Foster, P. Rinke, Data-driven materials science: status, challenges, and perspectives, Adv. Sci. 6 (2019) 1900808, https://doi.org/10.1002/advs.201900808.

[41] G. Pilania, A. Mannodi-Kanakkithodi, B.P. Uberuaga, R. Ramprasad, J.E. Gubernatis, T. Lookman, Machine learning bandgaps of double perovskites, Sci. Rep. 6 (2016), 19375. https://doi.org/10.1038/srep19375.

[42] S. Ubaru, A. Międlar, Y. Saad, J.R. Chelikowsky, Formation enthalpies for transition metal alloys using machine learning, Phys. Rev. B 95 (2017), 214102. https://doi.org/10.1103/PhysRevB.95.214102.

[43] A. Choudhury, T. Konnur, P.P. Chattopadhyay, S. Pal, Structure prediction of multiprincipal element alloys using ensemble learning, Eng. Comput. 37 (2019) 1003–1022, https://doi.org/10.1108/EC-04-2019-0151.

[44] S. Gong, W. Wu, F.Q. Wang, J. Liu, Y. Zhao, Y. Shen, S. Wang, Q. Sun, Q. Wang, Classifying superheavy elements by machine learning, Phys. Rev. A 99 (2019), 022110. https://doi.org/10.1103/PhysRevA.99.022110.

[45] N. Islam, W. Huang, H.L. Zhuang, Machine learning for phase selection in multiprincipal element alloys, Comput. Mater. Sci. 150 (2018) 230–235, https://doi.org/10.1016/j.commatsci.2018.04.003.

[46] W. Huang, P. Martin, H.L. Zhuang, Machine-learning phase prediction of high-entropy alloys, Acta Mater. 169 (2019) 225–236, https://doi.org/10.1016/j.actamat.2019.03.012.

[47] Z.Q. Zhou, Y. Zhou, Q. He, Z. Ding, F. Li, Y. Yang, Machine learning guided appraisal and exploration of phase design for high entropy alloys, npj, Computational Materials 5 (2019) 1–9, https://doi.org/10.1038/s41524-019-0265-1.

[48] S. Guo, C. Ng, J. Lu, C.T. Liu, Effect of valence electron concentration on stability of fcc or bcc phase in high entropy alloys, J. Appl. Phys. 109 (2011), 103505. https://doi.org/10.1063/1.3587228.

[49] Y.F. Ye, Q. Wang, J. Lu, C.T. Liu, Y. Yang, High-entropy alloy: challenges and prospects, Mater. Today 19 (2016) 349–362, https://doi.org/10.1016/j.mattod.2015.11.026.

[50] D.B. Miracle, O.N. Senkov, A critical review of high entropy alloys and related concepts, Acta Mater. 122 (2017) 448–511, https://doi.org/10.1016/j.actamat.2016.08.081.

[51] O.N. Senkov, D.B. Miracle, K.J. Chaput, J.P. Couzinie, Development and exploration of refractory high entropy alloys-a review, J. Mater. Res. 33 (2018) 3092–3128, https://doi.org/10.1557/jmr.2018.153.

[52] J. Chen, X. Zhou, W. Wang, B. Liu, Y. Lv, W. Yang, D. Xu, Y. Liu, A review on fundamental of high entropy alloys with promising high-temperature properties, J. Alloy. Compd. 760 (2018) 15–30, https://doi.org/10.1016/j.jallcom.2018.05.067.

[53] F. He, Z. Wang, C. Ai, J. Li, J. Wang, J.J. Kai, Grouping strategy in eutectic multiprincipal-component alloys, Mater. Chem. Phys. 221 (2019) 138–143, https://doi.org/10.1016/j.matchemphys.2018.09.044.

[54] Y.F. Ouyang, X.P. Zhong, Y. Du, Y.P. Feng, Y.H. He, Enthalpies of formation for the Al-Cu-Ni-Zr quaternary alloys calculated via a combined approach of geometric model and Miedema theory, J. Alloy. Compd. 420 (2016) 175–181, https://doi.org/10.1016/j.jallcom.2005.10.047.

[55] F.R. De Boer, W.C.M. Mattens, R. Boom, A.R. Miedema, A.K. Niessen, Cohesion in Metals, North-Holland, Amsterdam, 1988.

[56] Z. Śniadecki, J.W. Narojczyk, B. Idzikowski, Calculation of glass forming ranges in the ternary Y-Cu-Al system and its sub-binaries based on geometric and Miedema's models, Intermetallics 26 (2012) 72–77, https://doi.org/10.1016/j.intermet.2012.03.003.

[57] L. Zhang, H.M. Chen, Y.F. Ouyang, Y. Du, Amorphous forming ranges of Al-Fe-Nd-Zr system predicted by Miedema and geometrical models, J. Rare Earth. 32 (2014) 343–351, https://doi.org/10.1016/S1002-0721(14)60077-6.

[58] L. Zhang, R.C. Wang, X.M. Tao, H. Guo, H.M. Chen, Y.F. Ouyang, Formation enthalpies of Al-Fe-Zr-Nd system calculated by using geometric and Miedema's models, Physica B 463 (2015) 82–87, https://doi.org/10.1016/j.physb.2015.01.023.

[59] L. Zhang, H.M. Chen, X.M. Tao, Y.F. Ouyang, Thermodynamics study of Al-based high entropy quinary alloys, Chin. J. Nonferr. Met. 29 (2019) 2601–2608, https://doi.org/10.19476/j.ysxb.1004.0609.2019.11.17.

[60] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artif. Intell. 97 (1997) 245–271, https://doi.org/10.1016/S0004-3702(97)00063-5.

[61] J. Reunanen, Overfitting in making comparisons between variable selection methods, J. Mach. Learn. Res. 3 (2003) 1371–1382.

[62] A. Jović, K. Brkić, N. Bogunović, A review of feature selection methods with applications, 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE 2015, pp. 1200–1205, https://doi.org/10.1109/MIPRO.2015.7160458.

[63] J. Lee Rodgers, W.A. Nicewander, Thirteen ways to look at the correlation coefficient, Am. Stat. 42 (1988) 59–66, https://doi.org/10.2307/2685263.

[64] N. Meinshausen, P. Bühlmann, Stability selection, J. R. Stat. Soc. B 72 (2010) 417–473https://doi-org.proxy.lib.utk.edu/10.1111/j.1467-9868.2010.00740.x.

[65] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, Readings in Cognitive Science, A Perspective from Psychology and

Artificial Intelligence 1988, pp. 399–421, https://doi.org/10.1016/B978-1-4832-1446-7.50035-2.

[66] V. Vapnik, Estimation of Dependences Based on Empirical Data, Springer, New York, 2006https://doi.org/10.1007/0-387-34239-7.

[67] S.S. Haykin, Neural Networks and Learning Machines, Prentice Hall, New Jersey, 2009.

[68] J.H. Friedman, Stochastic gradient boosting, Comp. Stat. Data Anal. 38 (2002) 367–378, https://doi.org/10.1016/S0167-9473(01)00065-2.

[69] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (2001) 1189–1232, https://doi.org/10.1214/aos/1013203451.

[70] B. Schölkopf, A. Smola, K.R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319, https://doi.org/10.1162/089976698300017467.

[71] K.R. Muller, S. Mika, G. Ratsch, K. Tsuda, B. Scholkopf, An introduction to kernel-based learning algorithms, IEEE T. Neural Networ. 12 (2001) 181–201, https://doi.org/10.1109/72.914517.

[72] R. Abbaschian, R.E. Reed-Hill, Physical Metallurgy Principles, PWS Publishing Company, Boston, 1994.

[73] A. Takeuchi, A. Inoue, Calculations of mixing enthalpy and mismatch entropy for ternary amorphous alloys, Mater. Trans. JIM 41 (2000) 1372–1378, https://doi.org/10.2320/matertrans1989.41.1372.

[74] A. Takeuchi, A. Inoue, 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, Mater. Trans. JIM 46 (2005) 2817–2829, https://doi.org/10.2320/matertrans.46.2817.

[75] F. Tancret, I. Toda-Caraballo, E. Menou, P.E.J.R. Díaz-Del, Designing high entropy alloys employing thermodynamics and Gaussian process statistical analysis, Mater. Design 115 (2017) 486–497, https://doi.org/10.1016/j.matdes.2016.11.049.

[76] Y. Zhang, C. Wen, C. Wang, S. Antonov, D. Xue, Y. Bai, Y. Su, Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models, Acta Mater. 185 (2020) 528–539, https://doi.org/10.1016/j.actamat.2019.11.067.

[77] C.J. Van Rijsbergen, Information Retrieval, Butterworth-Heinemann, London, 1979.

[78] J. Bhatt, W. Jiang, X. Junhai, W. Qing, C. Dong, B.S. Murty, Optimization of bulk metallic glass forming compositions in Zr-Cu-Al system by thermodynamic modeling, Intermetallics 15 (2007) 716–721, https://doi.org/10.1016/j.intermet.2006.10.018.

[79] B.R. Rao, M. Srinivas, A.K. Shah, A.S. Gandhi, B.S. Murty, A new thermodynamic parameter to predict glass forming ability in iron based multi-component systems containing zirconium, Intermetallics 35 (2013) 73–81, https://doi.org/10.1016/j.intermet.2012.11.020.

[80] G.A. Mansoori, N.F. Carnahan, K.E. Starling, T.W. Leland Jr, Equilibrium thermodynamic properties of the mixture of hard spheres, J. Chem. Phys. 54 (1971) 1523–1525, doi:https://doi.org/10.1063/1.1675048.