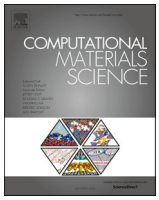
Computational Materials Science 223 (2023) 112140

Contents lists available at ScienceDirect

Computational Materials Science

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



Full Length Article

Improving machine learning based phase and hardness prediction of high-entropy alloys by using Gaussian noise augmented data

Yicong Ye [a,*], Yahao Li [a], Runlong Ouyang [b], Zhouran Zhang [a], Yu Tang [a], Shuxin Bai [a]

[a] Department of Materials Science and Engineering, College of Aerospace Science and Engineering, National University of Defense Technology, Changsha 410073, China
[b] College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China

ARTICLE INFO

Keywords:
Machine learning
High-entropy alloys
Data augmentation
Alloy prediction

ABSTRACT

Developing a machine learning (ML) based high-entropy alloys (HEA) prediction model is an advanced method to improve the traditional trial-and-error experiments with a long period and high cost. However, the ML model is highly dependent on data. This paper draws on the experience of the image data augmentation approaches, and proposes a simple material data augmentation method, which aims to solve the difficulties of small samples and large noise of material data, by adding Gaussian noise to the original data to generate more "pseudo samples". This work respectively starts by the HEA phase classification task and the hardness regression task, to study the effect of noise on data enhancement. It is found that the noise samples are different samples with new information. The noise samples enhanced data can significantly improve the test results of the models. Further testing results with the validation set that the models have never seen before, demonstrate that the regression model of medium noise samples enhanced has the best prediction accuracy ($R^2 = 0.954$). It turns out that the data enhancement method applied in this work helps the ML models to achieve a more efficient and accurate prediction of HEA phase and hardness. Besides, this work provides a reference method for improving the ML models and designing new HEA.

1. Introduction

Traditional alloys like iron and aluminum alloy have one principal element. Their properties can be improved by alloying other metal elements. Therefore, it is always constrained by the principal element. High-entropy alloys (HEA) proposed by Cantor [1] and Yeh [2], broke the design concept of traditional alloys. It is a kind of non-principal or multi-principal alloy. Each constituent element can be the principal element, making the atomic distribution chaotic and disordered. Its high disorder produces unexpected stability of HEA, making it more inclined to form a solid solution structure. By designing the principal elements of HEA, it is expected to break the strength-plastic contradiction of traditional structural materials [3-7]. Therefore, it has a wide application potential.

However, there are complex design patterns of HEA behind its excellent properties. The potential components of HEA are filled with the entire composition space, which brings a huge challenge to traditional trial-and-error experiments. Some researchers have tried to summarize an empirical law of the HEA formation from previous data [8-13]. For example, Zhang [12] et al. calculated the Ω parameter to distinguish whether a HEA formed solid solution structure. However, their empirical summarizations are usually limited by the scope of the chart visualization, considering no more than three factors at one time.

With the development of computer technology and the accumulation of material data, data-driven approaches represented by machine learning (ML) have turned to be a new pattern of materials science [14-16]. It is an effective means to process high-dimensional, difficult, and complex problems in designing HEA. In recent years, reports of ML assisted HEA design are gradually abundant [17-23]. The cost of a new HEA design could be greatly reduced through data-driven approaches. At present, ML models have been studied in predicting various properties of HEA [24-26]. For example, Yan [18] et al. successfully predicted and verified 10 new refractory HEA with solid solution structures through a gradient boosting (GB) model. Su [21] et al. used the ML method to study the γ' phase-enhanced Co-based superalloys. They combined a classification model of the γ and γ' phase and a regression model of the γ' phase solid solubility for the high-throughput screening of the composition space at the same time. Finally, they got a better result.

According to the "No Free Lunch Theorem", ML models are

* Corresponding author.
E-mail address: 18505993519@163.com (Y. Ye).

https://doi.org/10.1016/j.commatsci.2023.112140
Received 30 September 2022; Received in revised form 8 February 2023; Accepted 14 March 2023
Available online 21 March 2023

extremely data-dependent. However, material data are characterized by small samples, high dimensions, and large noise. ML algorithms can handle high-dimensional problems well with their powerful computing power, which those ML models reported have made good use of. ML models could incorporate a variety of parameters into the input features and establish high-dimensional deep relationships. Unfortunately, ML models are often plagued by the quantity and quality of data. The actual characteristics of small samples in materials science are difficult to make change. Hence, researchers need to fully explore how to utilize small samples to build high-precision ML models [20,27-30]. For example, Debnath [30] et al. generated new data based on generative adversarial network (GAN) to supplement origin data. Similarly, Zhao [27] et al. generated more pseudo data based on an ML model trained by raw data. Xu [28] et al. created a transfer learning model between a model based a large data and a model trained by smaller data. However, most of these methods of generating new samples are based on deep learning models, which are vitally complex and time-consuming for training.

Inspired by the image data augmentation in small sample learning, this paper proposes a simple data enhancement method to process the raw material data with Gaussian noise. In this method, the noise of material data is regarded as an intrinsic characteristic. And further applied it to produce twice "pseudo-data" containing noise at one time, taking the noise into models. In this way, the difficulty of small samples and large noise is solved at the same time. In addition, the influence of noise data on ML test results is studied, by discussing the two tasks of classification and regression respectively, which provides an orientation for popularizing this method in other ML models.

2. Methodology

2.1. Machine learning method

There are lots of articles since HEA was firstly reported, accumulating numerous data. Therefore, more and more ML based literature on the HEA phase and hardness prediction emerged in recent years. Refereeing the work of Senkov [31,32] et al., Machaka [17] et al., and Wen [25] et al., this paper screened out duplicate samples and a small part of erroneous data after integrating all data. A HEA phase dataset with 323 samples and a hardness dataset of AlCoCrCuFeNi HEA with 176 samples were obtained.

The HEA phase dataset includes many types of phase structures. This paper only considers whether one HEA contains intermetallic phases (IM). If it contains IM, it belongs to class "1″; and if not, it is class "0". In addition, since the phase dataset covers a large number of different elements, the elements of alloys are no longer taken into account. As shown in Table 1, 7 empirical parameters are adopted as input features: mixing entropy (ΔS_{mix})[11], mixing enthalpy (ΔH_{mix})[11], atomic size difference (δ)[11], electronegativity difference (Δχ)[10], valence electron concentration (VEC)[8], average melting point (T_m) and Ω [12]. These parameters have been considered to study the influence on the formation of HEA phases. The composition of the hardness dataset only contains 6 elements. Therefore, despite the empirical parameters, the content of components is selected as input features in the training hardness regression model.

Table 1
Features of HEAs considered in the phase prediction model.

Feature names	Equations
Mixing entropy (ΔS_{mix})	$\Delta Smix = -R\sum_{i=1}^{n} c_i \ln c_i$
Mixing enthalpy (ΔH_{mix})	$\Delta H_{mix} = \sum_{i=1,i<j}^{n} 4H_{ij}c_i c_j$
Atomic size difference (δ)	$\delta = 100 \times \sqrt{\sum_{i=1}^{n} c_i\left(1-\frac{r_i}{\bar{r}}\right)^2}$
Difference of the Pauling negativities ($\Delta\chi$)	$\Delta\chi = \sqrt{\sum_{i=1}^{n} c_i(\chi_i-\overline{\chi})^2}$
Valence electron concentration (VEC)	$VEC = \sum_{i=1}^{n} c_i(VEC)_i$
Average melting points (T_m)	$T_m = \sum_{i=1}^{n} c_i T_{mi}$
Ω	$\Omega = \frac{T_m \Delta S_{mix}}{\lvert\Delta H_{mix}\rvert}$

Before the training model, the dataset should be normalized. In this paper, the processing method of maximum normalization is carried out. As shown in equation(1), for each feature, Xi replaces one original value. X_i^{min} and X_i^{max} respectively replace the minimum and maximum values in this feature.

$$X_i^{norm} = \frac{X_i - X_i^{min}}{X_i^{max} - X_i^{min}} \tag{1}$$

In this work, support vector machine (SVM) is selected to train the HEA phase classification model and hardness regression model respectively. SVM is a classic classification algorithm that maps to higher dimensions to solve lower-dimensional classification problems. The algorithm is implemented by sklearn.SVC/SVR method in the scikit-learn library based on Python3.7. Considering that the dataset used in this paper is small, the SVM algorithm only selects the Radial Basis Function (rbf) kernel function. Other kernel functions will not be discussed in this work.

Classification and regression are two different tasks in ML. They have different evaluation criteria. When training phase classification model, accuracy (shown in equation (2)) is selected as the evaluation standard. In the formula, TP and TN are on behalf of the number of correct predictions of positive samples and negative samples. FP and FN represent the number of error predictions of positive samples and negative samples. Consequently, TP + TN represents the number of correct predictions, and the denominator represents all of the samples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Besides, the root mean square error (RMSE) and the coefficient of determination R^2 shown in equations (3) and (4) are regarded as the evaluation criteria for the regression model. In the formula, y_i is the true value in the dataset. $\bar{y}$ represents the average of all y in the dataset. And $\hat{y}_i$ is the predicted value of the model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$R2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{4}$$

In order to improve the generalization ability of the model and avoid overfitting, a 5-fold cross-validation method is applied to find the best hyperparameters. In addition, adding noise samples to the raw data may cause data leakage during training and testing ML models. Besides, the data leakage means that because part of the data in test set overlaps with the training set, the accuracy of test set becomes virtual-high. Here we take two steps to weaken this problem. First, the enhanced data will be randomly scrambled after the noise samples are integrated. Secondly, a part of the data set is divided into validation data, which could also be seen as experimental validation. In a total of 323 phase samples and 176 hardness samples, 26 (including 9 class "1″ and 17 class "0" samples) phase samples and 16 hardness samples are set aside respectively as the validation sets. Hence, the remaining 297 (including 109 class "1" and 188 class "0" samples) phase samples and 160 hardness samples were trained by the 5-fold cross-validation method.

2.2. A simple data enhancement method

Traditional sample enhancement methods for image data include copying, rotating, clipping, and adding noise. These little tricks could easily and effectively increase the number of pictures. Affected by the method of adding noise to the pictures, this article tries to enhance the material data by adding noise to samples. Gaussian noise is a common type of noise. In signal processing, the standard data is often made more

realistic by adding appropriate Gaussian noise. The material data have the same essence. The designed nominal component alloy will inevitably have deviations in the process of weighing and preparation. At this time, the actual component is different from the nominal composition. Therefore, the data set is actually full of nominal components. By adding some noise to the nominal component, the pseudo samples could be obtained. The way to add noise to raw samples is shown in Fig. 1. As it demonstrated in Fig. 1(a), the Gaussian noise is a random number whose probability density function obeys a Gaussian distribution. The noise value is closely related to σ. Higher σ makes the data more widely distributed. By adding the Gaussian noise, the raw data is slightly changed, which becomes the noise data. Therefore, as shown in Fig. 1 (b), by adding noise to the original component content values, the so-called "actual components", or noise samples are obtained. Similarly, the hardness experiments also have a certain deviation. The hardness in the raw data set is the average value of multiple measurements. Therefore, we also add noise to the hardness value to obtain a new hardness data set. In this way, the data sets are doubled by simply adding Gaussian noise to the samples.

3. Results and discussion

As it is known, a random number conformed to a Gaussian distribution has a 68% probability of being in the interval of $\mu \pm \sigma$, and a 95% probability of being in the interval of twice the standard deviation, where μ and σ are the mean and the standard deviation of the Gaussian distribution. Thus, although the noise is a random number, its magnitude is essentially determined by μ and σ. In order to maintain the randomness and symmetry of the noise, μ is usually set to 0 in the Gaussian distribution. In contrast, the value of σ needs more discussion, which is also the focus of this paper. Therefore, according to the empirical fluctuation of components, this paper considers $\sigma = 0.001$ as the low noise component, $\sigma = 0.003$ as the middle noise component, and the high noise component when $\sigma = 0.005$. In addition, after adding noise, the labels of samples remain unchanged in the classification task, while in the regression task, the hardness values of samples need to be processed with noise. According to the empirical fluctuation of HEA hardness, $\sigma = 2$ HV is considered as a low noise hardness. When $\sigma = 5$ HV, it is considered as a middle noise hardness, and when $\sigma = 8$ HV, it is considered as a high noise hardness. In Sections 3.1 and 3.2, this work discusses the research of noise-enhanced samples on two different tasks in ML.

3.1. Noise samples augmented for HEA phase classification

This section firstly explores how the model learns the noise samples differently from the raw samples respectively. Table 2 presents the accuracy of ML models based on different data, where the average training accuracy and test accuracy refers to the average accuracy under 5-fold cross-validation of the model. Validation accuracy represents the prediction accuracy using the validation data set as the test set.

Table 2
The accuracy of machine learning models based on different data.

Data based	Average training accuracy	Average test accuracy	Validation accuracy
Raw data	0.848	0.821	0.831
Low noise data	0.843	0.805	0.823
Middle noise data	0.846	0.795	0.823
High noise data	0.838	0.805	0.808

Firstly, by comparing the accuracy of training set and test set, the similar values indicate that the models are not overfitting. Besides, the accuracy between the raw data based model and noise data based models are quite different. It is found that the models based on noisy data usually get lower accuracy both in test set and validation set. However, it doesn't mean noisy data is useless. On the contrary, it confirms that the noisy data is not a simple repetition of the raw data. Therefore, it is expected that adding noise data will benefit for model training.

To achieve sample enhancement, this paper further adds different noise samples to the raw data and studies whether the new data set with noise samples enhancement could make the models better. The histogram in Fig. 2 shows the test (red bar) and validated (blue bar) accuracy of the models based on different augmented data sets. The data used in the figure includes the raw data and the raw data after adding the same raw samples, low noise samples, middle noise samples, and high noise samples. From the accuracy deviations shown in the Fig. 2, it is found that the accuracy of the models based on the enhanced data have improved to some extent compared to using the raw data only.

Compared with the 2 × raw data based model, the low noise samples enhanced model has a higher test accuracy. It reveals that the noise samples are not only an increase in the number of samples but also an increase in new information, which is conducive to improving the generalization ability of the model. The accuracy of the models obtained from the data enhanced by the medium and high noise samples are slightly lower than that of the sample repetition model, which may be related to the data leakage in the test set. However, it still higher than raw data. It can also demonstrate from the side that the noise data is different from the raw data, enabling the model to attain more information.

The validation accuracy could reflect the generalization ability of ML

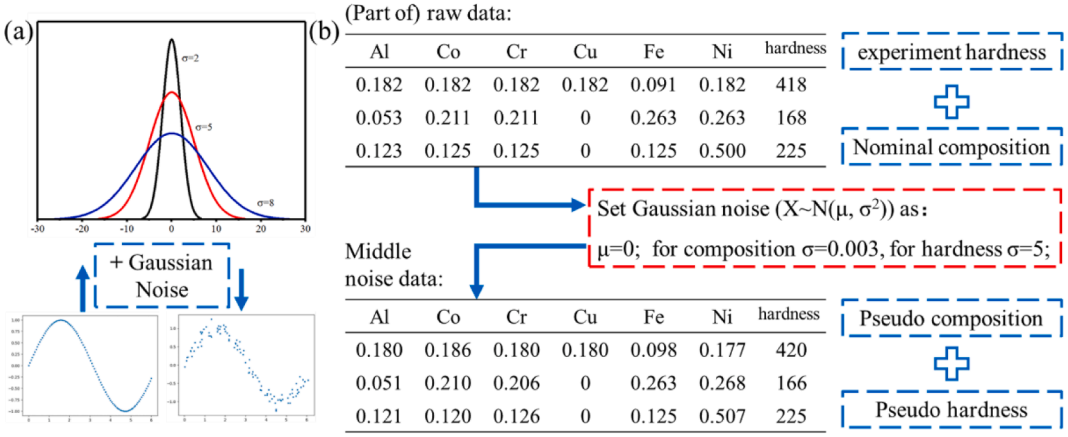


Fig. 1. The schematic diagram of (a) the comparison of raw samples and noise samples based on sine wave (b) adding middle Gaussian noise to hardness samples;

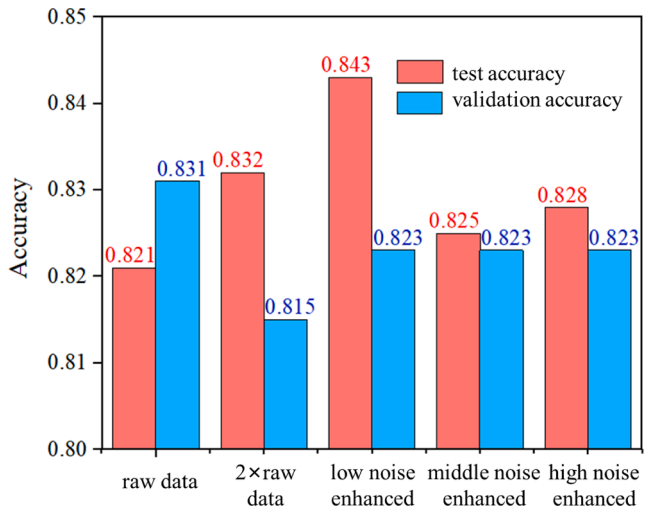


Fig. 2. The test and validation accuracy comparison based on different augmented data.

models. From the validation set accuracy in Fig. 2, it is found that these accuracy values are close and the model based on the raw data has the highest value, while the one with 2 × raw data has the lowest value. The lowest accuracy was possibly caused by overfitting. In addition, the models enhanced by noise data are slightly lower than the original results. According to the both results in Fig. 2, the addition of noise data improves the learning ability of the model itself to a certain extent, but the generalization ability is not improved. As a whole, the improvement has been limited.

At the same time, we wondered whether the limited improvement would be caused by overfitting. Considering that the ratio of the sample number of class "1" to class "0" mentioned in Section 2.1 is 109:186, it is easy to produce overfitting in such a biased data set. Therefore, confusion matrices of the 4 models were calculated as shown in Fig. 3(a)-(d). Since the ML models' training process were based on5-fold cross-validation, the confusion matrices calculated for 5 times are added to obtain the final confusion matrix.

Taking Fig. 3(a) as an example, the figure represents the sum of the confusion matrices of 5 cross-validations based on the 2 × raw data model. The ordinate is the actual label of the samples, and the abscissa is the predicted label. Therefore, "147" in the lower right corner represents the number of samples that are actually IM (1) and predicted to be IM, and "71" in the lower left represents the number of samples that are actually IM, while predicted another class (0). If there is an over-fitting occurred, the general model will tend to judge the samples as the ones with more samples. Thus, it really happens in Fig. 3(a), leading the actual IM class regarded as another class. Considering that over-fitting is caused by sample imbalance, low-noise samples of type "1" are augmented to the raw data to achieve targeted enhancement of under-samples. After a few random deletions of samples, the data ratio of class "1" to "0" is 200:186. As shown in Fig. 3(e), the confusion matrix was calculated based on the model after the under-samples enhancement. It can be found that the over-fitting is much reduced. The recall rate in the positive samples reaches 80%, while the recall rate is only about 67% in the first four over-fitting models, getting greatly improved. Although the recall rate of class 0 is also much lower, it maintains a similar accuracy to the other class, which means the model is not over-fitted. It reveals that the original models do have an over-fitting problem, which makes the recall rates of the two classes quite different. However, the over-fitting problem of the model can be reduced by enhancing the under-samples.

3.2. Noise samples augmented for HEA hardness regression

In the ML classification mission described in the previous section, only component noise needs to be added. Their labels do not change. Besides, since there are only two types of samples, the improvement of data enhancement is limited. In the regression task, in addition to the component noise, the hardness as the output also needs to add noise. And data changes will have more obvious effects because the regression task output is continuous results.

According to the method described in Section 2.2, we established three kinds of noise data, in which low-noise data enhancement refers to the case where both components and hardness are low noise. training results. Fig. 4(a) shows the training results of ML models based on different datasets. Two evaluation indicators, R^2 and RMSE, are presented in the figure, where the left axis is the average R^2 value in red, and the right axis is the average RMSE value in blue. The RMSE is more sensitive to values with larger errors. Therefore, taking RMSE as an example, the last four models after noise samples enhancement have lower RMSE than the first initial model. The RMSEs of the four enhanced models are relatively similar. On the one hand, the results in Fig. 4(a) confirm that the models after samples enhancement have a better

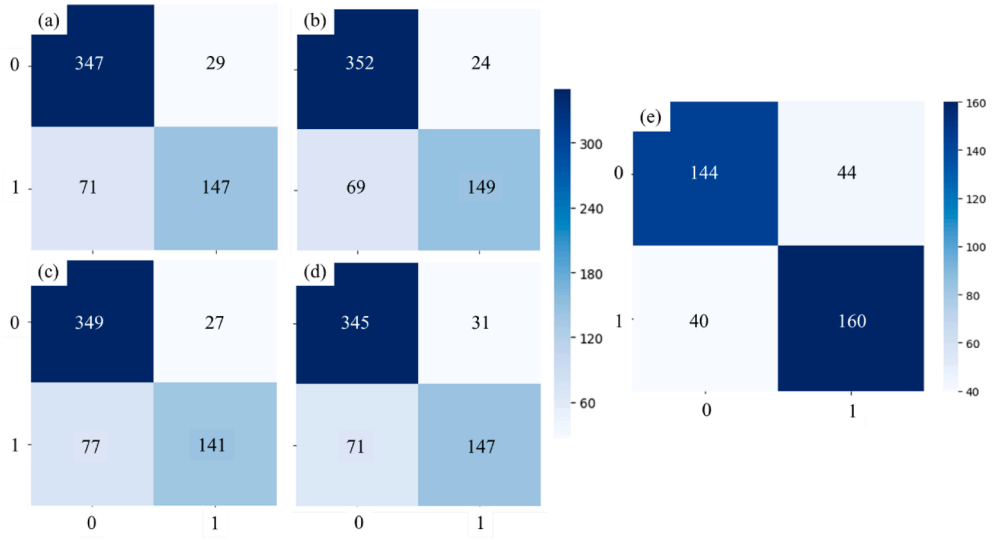


Fig. 3. Confusion matrices of models based on (a) 2 × raw data, (b) low noise enhanced data, (c) middle noise enhanced data, (d) high noise enhanced data, and (e) low noise of label 1 enhanced data.

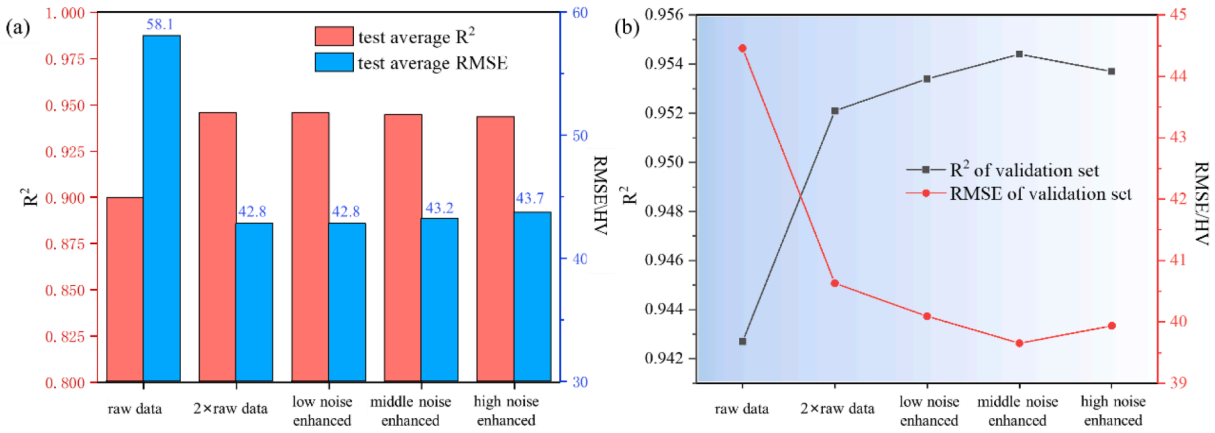


Fig. 4. (a) The test R^2 and RMSE of ML model based on different data; (b) R^2 and RMSE of the validation set based on different ML models.

prediction. On the other hand, models with duplicated data and low-noise augmented data have the same lowest RMSE, leading to the suspicion of data breaches. Although we have randomly shuffled the data before integration, the noisy data is actually not the same as the original data obtained in Section 3.1. Therefore, validation data that the model has never seen before is required.

Fig. 4(b) presents the R^2 and RMSE of the above models on the validation set. It demonstrates that with the addition of noise data, the RMSE on the validation set gets smaller. The middle-noise enhanced model has the highest R^2 (0.954) and the lowest RMSE (39.6 HV). The low-noise and high-noise models are also better than the original model and the model trained on repeated data. It reveals that the addition of noise allows the model to learn more information, so that it has stronger robustness and higher prediction accuracy for brand new data.

In order to further eliminate the tendency in data selection, we compared the distribution of the raw data and the validation data. As shown in Fig. 5(a), it is found that the hardness distribution of the validation set is almost similar to the raw data set, except for the missing 400–500 HV sample. Fig. 5(b) presents the comparison between the predicted hardness based on the medium-noise model and the actual hardness of the validation set. It is found that most of the values are very consistent, except a few samples are predicted slightly larger. Therefore, combining the results of Fig. 4 and Fig. 5, this paper initially obtained a generalized reference conclusion: the data enhanced by the middle noise samples are more suitable for the training of the regression model.

3.3. More noise samples augmented

The proposed method is simple and effective. This paper further tests the effect of more addition of noise augmented data on the model

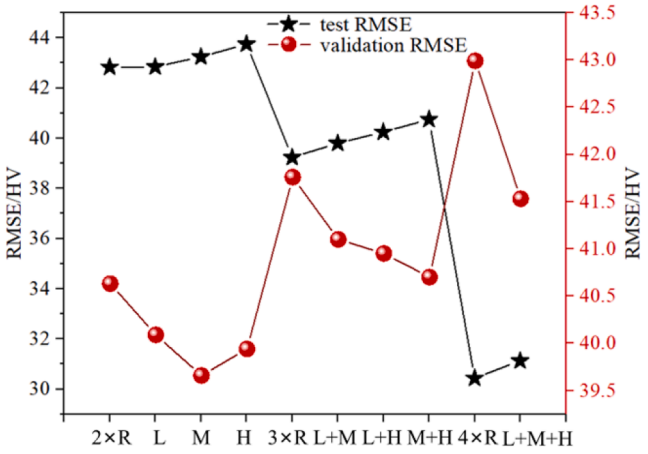


Fig. 6. The test rmse and validation rmse calculated by different models, where r represents raw data, l represents low noise data augmented, m represents middle noise data augmented, H represents high noise data augmented and L + M + H represents low noise data, plus middle noise data, plus high noise data augmented.

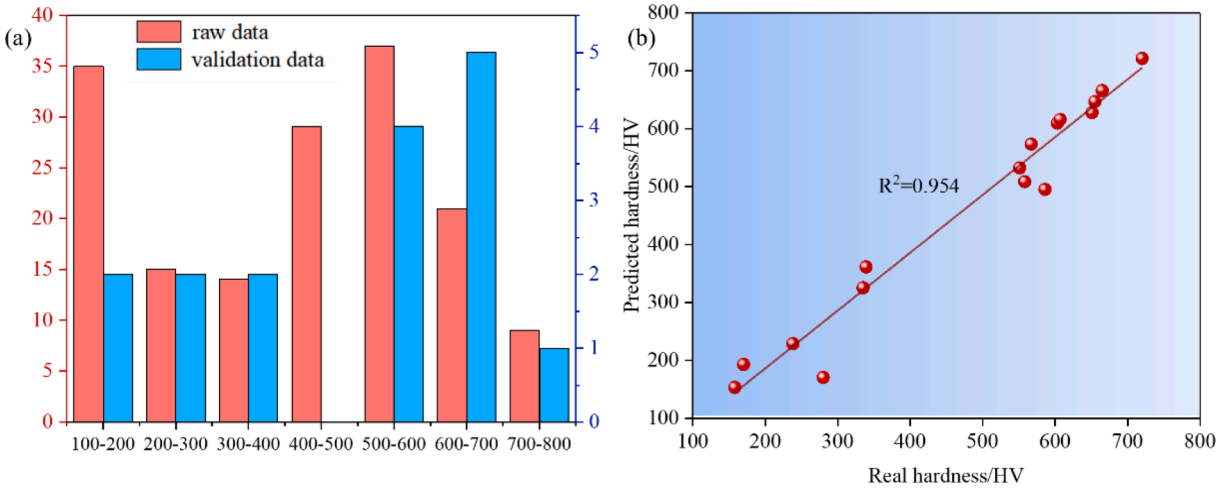


Fig. 5. (a) The hardness distribution of raw data and validation data; (b) the comparison of predictive values and real value.

prediction. Here we take the hardness prediction model as an example. Fig. 6 shows the test RMSE (black star) and validation RMSE (red ball) for models with different data. It contains the results of double raw data, triple raw data, and quadruple raw data.

Firstly, it is found that the test RMSE decreases as the data increases. It indicates that the models fit well. However, validation RMSE increase as data is added, which indicates that the generalization ability of the models is deteriorating. The high degree of fitting and low generalization ability of models prove that the models are overfitting. Since the generation of noisy data comes from the original data, too much addition inevitably leads to the problem of data leakage. The addition of more data can only improve the degree of model fitting, but have little influence on improving the generalization ability of the models. As a whole, noise samples enhancement could only be used once.

4. Conclusion and prospects

In this paper, a simple method of data enhancement based on noise samples is proposed to confront the high dependence on data of ML. A HEA phase classification model and a hardness regression model are created respectively to study the performance of noise samples enhanced data in classification tasks and regression tasks. The following conclusions are obtained:

(1) Whether the HEA phase classification task or hardness regression task, the method of noise sample enhancement could significantly improve the results of the model, indicating that noise samples help the model to learn the information of the data set better.
(2) In the HEA phase classification task, it is confirmed that the noise sample is a type of "new sample" different from the original sample through separate training on the raw sample and the noise sample, so that the model can learn more information. However, the noise enhancement model did not improve the accuracy of the validation set, demonstrating that the additional information provided by noise sample enhancement is limited.
(3) In the HEA hardness regression task, the model enhanced with middle noise samples has the highest R^2 and the lowest RMSE on the brand-new validation set. This result could be served as a reference for the generalization of this method in other regression tasks.
(4) More noise samples augmented could make models fit better, while the generalization ability of models gets worse.

This work studies the influence of noise samples enhancement on the ML model, which provides a case for the application of data enhancement methods based on noisy samples, demonstrating the method is simple and effective. However, there is still room for improvement in our proposed method, and we expect it to have wider applications.

Data availability statement

The raw data and codes required to reproduce these findings are available in the appendix files.

CRediT authorship contribution statement

Yicong Ye: Writing – review & editing, Funding acquisition. **Yahao Li:** Writing – original draft, Writing – review & editing. **Runlong Ouyang:** Conceptualization, Visualization. **Zhouran Zhang:** Supervision. **Yu Tang:** Project administration, Funding acquisition. **Shuxin Bai:** Funding acquisition, Supervision, Writing – review & editing.

Declaration of Competing Interest

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

Data availability

I have shared my data and codes in the Supplementary documents in the Attach files step.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.commatsci.2023.112140.

References

[1] B. Cantor, I.T.H. Chang, P. Knight, et al., 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.
[2] J.W. Yeh, S.K. Chen, S.J. Lin, et al., Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes, Adv. Eng. Mater. 6 (5) (2004) 299–303, https://doi.org/10.1002/adem.200300567.
[3] Y. Lu, Y. Dong, S. Guo, et al., A promising new class of high-temperature alloys: eutectic high-entropy alloys, Sci. Rep. 4 (1) (2015), https://doi.org/10.1038/srep06200.
[4] F. Otto, A. Dlouhý, C. Somsen, et al., The influences of temperature and microstructure on the tensile properties of a CoCrFeMnNi high-entropy alloy, Acta Mater. 61 (15) (2013) 5743–5755, https://doi.org/10.1016/j.actamat.2013.06.018.
[5] M. Naeem, H. He, F. Zhang, et al., Cooperative deformation in high-entropy alloys at ultralow temperatures, Sci. Adv. 6 (13) (2020), x4002, https://doi.org/10.1126/sciadv.aax4002.
[6] Z. Lei, X. Liu, Y. Wu, et al., Enhanced strength and ductility in a high-entropy alloy via ordered oxygen complexes, Nature 563 (7732) (2018) 546–550, https://doi.org/10.1038/s41586-018-0685-y.
[7] Z. Li, K.G. Pradeep, Y. Deng, et al., Metastable high-entropy dual-phase alloys overcome the strength–ductility trade-off, Nature 534 (7606) (2016) 227–230, https://doi.org/10.1038/nature17981.
[8] S. Guo, C. Ng, J. Lu, et al., Effect of valence electron concentration on stability of fcc or bcc phase in high entropy alloys, J. Appl. Phys. 109 (10) (2011), 103505, https://doi.org/10.1063/1.3587228.
[9] Y.F. Ye, Q. Wang, J. Lu, et al., The generalized thermodynamic rule for phase selection in multicomponent alloys, Intermetallics 59 (2015) 75–80, https://doi.org/10.1016/j.intermet.2014.12.011.
[10] Y. Dong, Y. Lu, L. Jiang, et al., Effects of electro-negativity on the stability of topologically close-packed phase in high entropy alloys, Intermetallics 52 (2014) 105–109, https://doi.org/10.1016/j.intermet.2014.04.001.
[11] Y. Zhang, Y.J. Zhou, J.P. Lin, et al., Solid-solution phase formation rules for multi-component alloys, Adv. Eng. Mater. 10 (6) (2008) 534–538, https://doi.org/10.1002/adem.200700240.
[12] X. Yang, Y. Zhang, Prediction of high-entropy stabilized solid-solution in multi-component alloys, Mater. Chem. Phys. 132 (2–3) (2012) 233–238, https://doi.org/10.1016/j.matchemphys.2011.11.021.
[13] Y.F. Ye, Q. Wang, J. Lu, et al., Design of high entropy alloys: a single-parameter thermodynamic rule, Scr. Mater. 104 (2015) 53–55, https://doi.org/10.1016/j.scriptamat.2015.03.023.
[14] J.X. Xie, Y.J. Su, D.J. Xue, et al., Machine learning for materials research and development, Acta Metall. Sin. 57 (11) (2021) 1343–1361, doi: 10.11900/0412.1961.2021.00357.
[15] Y. Cheng, T. Wang, G. Zhang, Artificial Intelligence for Materials Science, Springer, 2021.
[16] G.L.W. Hart T M C T. Machine learning for alloys, Nat. Rev. Mater.6 (2021), doi: 10.1038/s41578-021-00340-w.
[17] R. Machaka, Machine learning-based prediction of phases in high-entropy alloys, Comput. Mater. Sci 188 (2021), 110244, https://doi.org/10.1016/j.commatsci.2020.110244.
[18] Y. Yan, D. Lu, K. Wang, Accelerated discovery of single-phase refractory high entropy alloys assisted by machine learning, Comput. Mater. Sci 199 (2021), 110723, https://doi.org/10.1016/j.commatsci.2021.110723.
[19] 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.
[20] B. Steingrimsson, X. Fan, X. Yang, et al., Predicting temperature-dependent ultimate strengths of body-centered-cubic (BCC) high-entropy alloys, npj Comput. Mater. 7 (1) (2021) 1–10, doi: 10.1038/s41524-021-00623-4.
[21] P. Liu, H. Huang, S. Antonov, et al., Machine learning assisted design of γ′-strengthened Co-base superalloys with multi-performance optimization, npj Comput, Mater 62 (6) (2020), https://doi.org/10.1038/s41524-020-0334-5.
[22] Z. Zhou, Y. Zhou, Q. He, et al., Machine learning guided appraisal and exploration of phase design for high entropy alloys, npj Comput. Mater. 5 (1) (2019) 1–9, https://doi.org/10.1038/s41524-019-0265-1.
[23] J.M. Rickman, H.M. Chan, M.P. Harmer, et al., Materials informatics for the screening of multi-principal elements and high-entropy alloys, Nat. Commun. 10 (1) (2019), https://doi.org/10.1038/s41467-019-10533-1.

[24] L. Qiao, Z. Lai, Y. Liu, et al., Modelling and prediction of hardness in multi-component alloys: a combined machine learning, first principles and experimental study, J. Alloy. Compd. 853 (2021), 156959, https://doi.org/10.1016/j.jallcom.2020.156959.
[25] C. Wen, Y. Zhang, C. Wang, et al., 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.
[26] Z. Ding, E. Pascal, M. De Graef, Indexing of electron back-scatter diffraction patterns using a convolutional neural network, Acta Mater. 199 (2020) 370–382, https://doi.org/10.1016/j.actamat.2020.08.046.
[27] Z. Zhao, J. You, J. Zhang, et al., Data enhanced iterative few-sample learning algorithm-based inverse design of 2D programmable chiral metamaterials, Nanophotonics (Berlin, Germany). 11 (20) (2022) 4465–4478, https://doi.org/10.1515/nanoph-2022-0310.
[28] X. Wei, S. van der Zwaag, Z. Jia, et al., On the use of transfer modeling to design new steels with excellent rotating bending fatigue resistance even in the case of very small calibration datasets, Acta Mater. 235 (2022), 118103, https://doi.org/10.1016/j.actamat.2022.118103.
[29] H. Han, W. Li, S. Antonov, et al., Mapping the creep life of nickel-based SX superalloys in a large compositional space by a two-model linkage machine learning method, Comput. Mater. Sci 205 (2022), 111229, https://doi.org/10.1016/j.commatsci.2022.111229.
[30] A. Debnath, A.M. Krajewski, H. Sun, et al., Generative deep learning as a tool for inverse design of high entropy refractory alloys, J. Mater. Informatics (2021), doi: 10.20517/jmi.2021.05.
[31] J.P. Couzinié, O.N. Senkov, D.B. Miracle, et al., Comprehensive data compilation on the mechanical properties of refractory high-entropy alloys, Data Brief 21 (2018) 1622–1641, https://doi.org/10.1016/j.dib.2018.10.071.
[32] S. Gorsse, M.H. Nguyen, O.N. Senkov, et al., Database on the mechanical properties of high entropy alloys and complex concentrated alloys, Data Brief 21 (2018) 2664–2678, https://doi.org/10.1016/j.dib.2018.11.111.