

<https://doi.org/10.1038/s41524-025-01873-2>

# Machine learning for phase prediction of high entropy carbide ceramics from imbalanced data



Xuemeng Zhang<sup>1</sup>, Jia Sun<sup>1</sup> ✉, Yuyu Zhang<sup>1</sup>, Kaifei Fan<sup>1</sup>, Zhixiang Zhang<sup>1</sup>, Yujia Zhang<sup>1</sup>, Keke Wu<sup>1</sup>, Laura Feldmann<sup>2</sup>, Lianwei Wu<sup>3</sup>, Ralf Riedel<sup>2</sup> & Hejun Li<sup>1</sup> ✉

High-entropy carbide ceramics (HECCs) possess promising properties for extreme high-temperature applications. Machine learning offers an effective pathway to accelerate the discovery of novel HECCs, but data imbalance poses challenges for predictive performance. Here, we integrate the Borderline-SMOTE with machine learning algorithms to address this issue. A dataset containing 251 samples was established from literature, experimental synthesis, and synthetic oversampling. Key features influencing phase formation were selected via a four-step feature selection strategy. Ten common machine learning models were trained and optimized, with the random forest (RF) model identified as the most suitable for predicting HECCs phase formation ability. Eight HECCs compositions with high uncertainty were experimentally validated, and the results were incorporated back into the dataset to iteratively improve model accuracy. This work provides an efficient strategy for predicting phase formation in HECCs, particularly for small or imbalanced datasets, facilitating the accelerated design and reliable prediction of new HECCs.

The concept of “high-entropy” was introduced into the field of ceramics in 2015, when Rost et al.<sup>1</sup> first reported a rock-salt-structured high-entropy oxide ceramic (MgNiCoCuZn)O, marking a pioneering advancement in high-entropy ceramics. Whereafter, various types of high-entropy materials have been developed<sup>2,3</sup>. In addition to high-entropy oxides, a battery of non-oxide ceramic systems has also emerged, including high-entropy carbides<sup>4</sup>, borides<sup>5</sup>, nitrides<sup>6</sup>, and silicides<sup>7</sup>. Among them, high-entropy carbide ceramics (HECCs), developed on the basis of traditional binary ultra-high temperature ceramics, exhibit advantages over the corresponding binary carbides, including higher melting points, superior high-temperature strength, and better high-temperature oxidation and ablation resistance, which make them promising materials for applications in ultra-high temperature extreme environments<sup>8–13</sup>.

Although the field of high-entropy ceramics has flourished in recent years, the vast compositional space and complex atomic structures have hindered comprehensive exploration<sup>14</sup>. Traditional new material synthesis route often relies on extensive “trial and error” experiments, resulting in long developing periods and huge workloads. It is difficult to efficiently screen the optimal combination of components and concentrations of single-phase high-entropy ceramic materials that meet specific demands through large-scale experiments<sup>15</sup>. To address this, researchers have sought

to establish criteria for the phase formation ability of high entropy ceramics, aiming to guide the design of new high-entropy ceramics more reasonably and efficiently<sup>16</sup>. Thermodynamic and empirical criteria have played a crucial role in the early stages of developing novel high-entropy materials. Based on the classic Hume-Rothery rule<sup>17–21</sup>, several empirical parameters, such as atomic size difference ( $\sigma$ ), electronegativity ( $\chi$ ), and valence electron concentration (VEC), have been proposed for phase prediction, but their predictive accuracy remains unsatisfactory<sup>22</sup>. With the rapid development of computational simulations within materials science, high-throughput density functional theory (DFT)<sup>23</sup>, molecular dynamics (MD)<sup>24</sup> and calculation of the phase diagram (CALPHAD)<sup>25</sup> have also been applied to phase prediction of high-entropy ceramics. Although descriptors such as entropy forming ability (EFA) and mixing enthalpy ( $\Delta H_{\text{mix}}$ ) can predict the phase formation ability of HECCs under most conditions, achieving reliable accuracy typically requires thousands of computationally expensive DFT calculations for each candidate composition. The high computational cost significantly limits the applicability of simulation calculations for the exploration and prediction of novel high-entropy ceramics in a wide range of composition space<sup>15</sup>.

In recent years, machine learning (ML) has emerged as a revolutionary tool in the field of materials science. It not only enhances material discovery

<sup>1</sup>Science and Technology on Thermostructural Composite Materials Laboratory, Shaanxi Key Laboratory of Fiber Reinforced Light Composite Materials, Northwestern Polytechnical University, Xi'an, PR China. <sup>2</sup>Technische Universität Darmstadt, Institut für Materialwissenschaft, Darmstadt, Germany. <sup>3</sup>School of Computer Science, Northwestern Polytechnical University, Xi'an, PR China. ✉e-mail: [j.sun@nwpu.edu.cn](mailto:j.sun@nwpu.edu.cn); [lihejun@nwpu.edu.cn](mailto:lihejun@nwpu.edu.cn)

and design strategies but also predicts the performance of unknown materials, thereby accelerating the development and application of new materials<sup>26–29</sup>. In the domain of high entropy ceramics, data driven machine learning has made remarkable advances in accelerating the rapid designing and high-throughput prediction of novel high-entropy ceramics<sup>15</sup>. ML algorithms such as artificial neural networks (ANN)<sup>30</sup>, support vector machines (SVM)<sup>30</sup>, random forests (RF)<sup>25</sup>, Gaussian naive bayes (GNB)<sup>31</sup>, and logistic regression (LOG)<sup>32</sup> have been successfully applied to predict the phase formation ability of HECCs. However, phase prediction of HECCs remains constrained by the low amount of available data (<100 samples) and data imbalance, both of which adversely affect the generalization ability and prediction performance of ML models<sup>33</sup>. Specifically, most HECCs reported in the current literature are concentrated in the IV<sub>B</sub>, V<sub>B</sub>, VI<sub>B</sub> subgroups, whereas studies on HECCs containing rare-earth elements from the III<sub>B</sub> subgroup remains limited. Under such context, classifiers often sacrifice the prediction accuracy of the minority class to prioritize learning the majority class, which easily leads to the misclassification of minority samples<sup>33</sup>.

To address the challenges posed by small and imbalanced datasets in existing models for predicting the phase formation ability of HECCs, researchers often overcome this limitation by extensively collecting literature data and high-throughput experimental synthesis<sup>15,32</sup>. In general, there exist two kinds of imbalances, one is the imbalance between classes, and another is the imbalance within classes<sup>34</sup>. To address the issue of data imbalance, researchers have made significant advancements at both the data level and the algorithm level<sup>35</sup>. The former focuses on adjusting the class ratios within the input dataset to balance class distribution, while the latter modifies the learning algorithm or classifier specifically with respect to the minority class, without amending the input data distribution<sup>35</sup>. Among data-level methods, data preprocessing techniques, particularly resampling methods, have been widely employed in the field of computer science<sup>36</sup>. One of the most popular oversampling techniques is the synthetic minority oversampling technique (SMOTE), which generates synthetic samples by interpolating between existing minority class samples to balance the majority and minority classes<sup>37</sup>. However, SMOTE may also bring new challenges, such as the generation of noisy or irrelevant samples. To address these shortcomings, a series of advanced oversampling methods have been developed, including K-means SMOTE<sup>38</sup>, safe-level-SMOTE<sup>39</sup>, Borderline-SMOTE<sup>34</sup>, and adaptive synthetic sampling (ADASYN)<sup>40</sup>, etc. In contrast to ADASYN, the Borderline-SMOTE method pays more attention to the synthetic samples near difficult-to-distinguish decision boundaries between classes, which effectively minimizes the creation of noisy samples<sup>34</sup>. At the algorithm level, common strategies for addressing data imbalance include cost-sensitive learning and ensemble learning. These methods enhance the model's discriminative ability and increase its focus on minority samples by modifying the algorithm or decision threshold<sup>35</sup>. Classic ensemble learning methods, such as random forest, bagging,

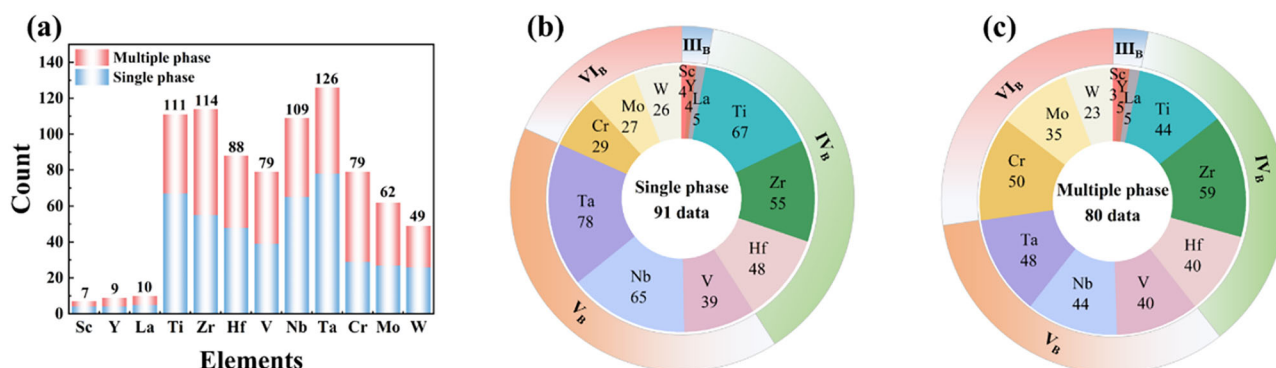
XGBoost, and AdaBoost, can be further improved by combining them with oversampling techniques to better handle data imbalance<sup>41</sup>.

In this work, we integrate the Borderline-SMOTE method with machine learning algorithms to address the challenges of small dataset and the data imbalance, aiming to establish an accurate and effective ML model for predicting the phase formation ability of HECCs. First, a dataset comprising 171 HECCs was constructed through experimental synthesis and literature collection. To address the scarcity of HECCs containing the III<sub>B</sub> elements (Sc, Y, and La) in the dataset, 80 new HECCs samples containing Sc, Y, and La elements were generated by the Borderline-SMOTE method, expanding the dataset to 251 samples. Then, redundant and unimportant features were removed through a combination of Pearson correlation coefficient analysis, recursive feature elimination, and exhaustive enumeration. As a result, an optimal feature subset containing six features was screened. Subsequently, ten commonly used ML algorithms were well trained and their prediction performances were compared to select the most suitable model. Finally, the optimized ML model was employed to predict the phase formation ability of HECCs composed of the III<sub>B</sub>, IV<sub>B</sub>, V<sub>B</sub> elements. The strong agreement between the experimental and predicted results demonstrates that the Borderline-SMOTE assisted ML method is a promising strategy for developing novel HECCs materials.

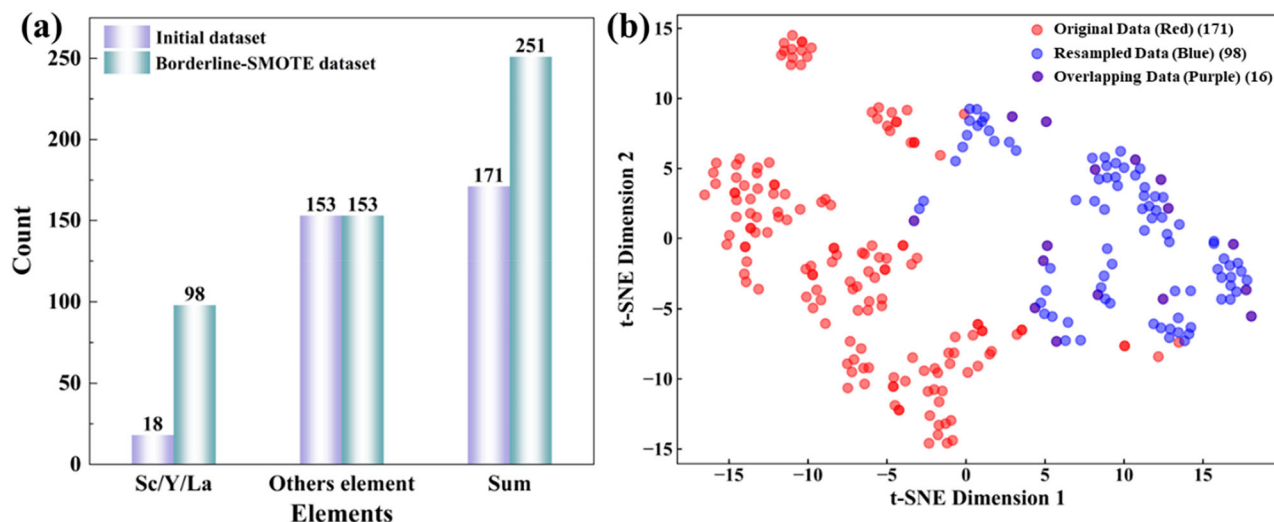
## Results

### Construction of the initial dataset and Borderline-SMOTE

Since the relevant research on HECCs containing rare-earth elements is still in its infancy, the number of HECCs samples containing group III<sub>B</sub> elements (Sc, Y, and La) available in the literature remains very limited, resulting in severe data imbalance in the dataset. In this case, minority class samples are often poorly represented and may be treated as meaningless or noisy data, resulting in overfitting and poor generalization performance of ML models<sup>42,43</sup>. Therefore, in addition to the 156 HECCs samples collected from the literature (Table S1), 15 new HECCs samples containing the rare-earth elements Sc, Y, or La were experimentally synthesized, and their phase compositions were identified by XRD. The corresponding characterization results are shown in Fig. S1 and Table S2. Figure 1 illustrates the elemental distributions of the dataset, which combines samples from both literature and experimental supplementation. As can be seen from Fig. 1(a), current studies on HECCs predominantly focus on ultra-high-temperature transition metal elements from the IV<sub>B</sub> and V<sub>B</sub> groups. Despite the addition of 15 samples containing rare-earth elements through experimental synthesis, the representation of the III<sub>B</sub> group elements remains significantly lower than that of other groups. Figure 1(b, c) exhibits a detailed breakdown of the quantities of each element in single-phase and multiple-phase carbides, respectively. It is evident that the dataset suffers from a severe data imbalance issue, which could negatively impact the prediction performance of ML models<sup>33</sup>.



**Fig. 1 | Element distribution of the initial dataset.** **a** All element distribution, **b** element distribution of single-phase HECCs, **c** element distribution of multiple-phase carbides.



**Fig. 2 | Structure of the dataset after Borderline-SMOTE treatment. a** Data distribution before and after Borderline-SMOTE, **b** data visualization of the original data and generated data using t-SNE.

**Table 1 | A list of 20 input features used in ML models for predicting HECCs phase formation ability**

Features	Abbreviations	Features	Abbreviations
Average valence electron concentration	$\overline{VEC}$	Valence electron concentration difference	$\sigma_{VEC}$
Average Pauling electronegativity	$\bar{\chi}_p$	Pauling electronegativity difference	$\sigma_{\chi_p}$
Average Mulliken electronegativity	$\bar{\chi}_m$	Mulliken electronegativity difference	$\sigma_{\chi_m}$
Average density	$\bar{\rho}$	Density difference	$\sigma_{\rho}$
Average mass	$\bar{m}$	Mass difference	$\sigma_m$
Average lattice size	$\bar{l}$	Lattice size difference	$\sigma_l$
Average metallic radius	$\bar{r}_{Me}$	Metallic radius difference	$\sigma_{r_{Me}}$
Average first ionization energy	$\bar{I}_1$	First ionization energy difference	$\sigma_{I_1}$
Average effective nuclear charge	$\bar{Z}^*$	Effective nuclear charge difference	$\sigma_{Z^*}$
configurational entropy	$\Delta S_{conf}$	Geometrical parameter	$\Lambda$

The Borderline-SMOTE method is applied to address data imbalance in the dataset. Unlike traditional SMOTE method that randomly selects minority samples, Borderline-SMOTE pays close attention to boundary samples that are more prone to misclassification by the classifier (i.e., those minority samples that are closer to the majority samples). Figure 2 exhibits the dataset structure after applying Borderline-SMOTE. As shown in Fig. 2(a), the number of samples containing rare-earth elements (Sc, Y, and La) increased significantly from 18 to 98 after Borderline-SMOTE. Figure 2(b) compares the data distribution of the original dataset (red dots) with the synthetic samples generated by Borderline-SMOTE for HECCs containing elements Sc, Y, and La (blue dots). The purple dots in Fig. 2(b) represent the original HECCs samples containing rare-earth elements that are located in the boundary region between the majority and minority class samples. Specifically, the red dots in Fig. 2(b) represent all 171 samples in the original dataset, while the blue dots denote 80 newly generated synthetic HECCs samples containing Sc, Y, and La elements, produced by Borderline-SMOTE based on the 18 original HECCs samples containing Sc, Y, and La elements. Under the superposition of red and blue colors, 16 HECCs samples containing Sc, Y, and La elements, which are situated in the boundary area between the majority and minority class samples in the original dataset are displayed in purple. Comparing the distribution of red and blue dots in Fig. 2(b), it can be observed that the distribution of HECCs samples in the dataset shows a certain regularity, that is, similar categories show a clustered distribution. Specifically, HECCs without rare-earth

elements are mainly distributed on the left side of Fig. 2(b), whereas those HECCs containing rare-earth elements are more concentrated on the right side of Fig. 2(b). The Borderline-SMOTE method, emphasizing the boundary samples, reduces the generation of synthetic samples in the safe area (regions with minimal impact on classification boundaries) and avoids the generation of noisy samples. By focusing on the boundary regions, Borderline-SMOTE improves both the quality and effectiveness of the synthetic samples, thereby enhancing the accuracy of ML model to classify the minority class samples<sup>34</sup>.

### Feature selection and model construction

Table 1 presents an overview of the 20 input features used to establish ML models for predicting the phase formation ability of HECCs. These features are selected based on the following three considerations<sup>44,45</sup>. The first category pertains to the fundamental thermodynamic properties of HECCs, such as configurational entropy ( $\Delta S_{conf}$ ). The second category includes parameters derived from the Hume-Rothery rules of the constituent binary carbides, such as valence electron concentration (VEC), metallic radius ( $r_{Me}$ ), lattice size ( $l$ ), Pauling electronegativity ( $\chi_p$ ), Mulliken electronegativity ( $\chi_m$ ), and the geometric parameters ( $\Lambda$ ). The third category encompasses elemental properties of the constituent binary carbides, including atomic mass ( $m$ ), density ( $\rho$ ), first ionization energy ( $I_1$ ), and effective nuclear charge ( $Z^*$ ). In order to eliminate highly correlated redundant features and mitigate overfitting in subsequent ML models,

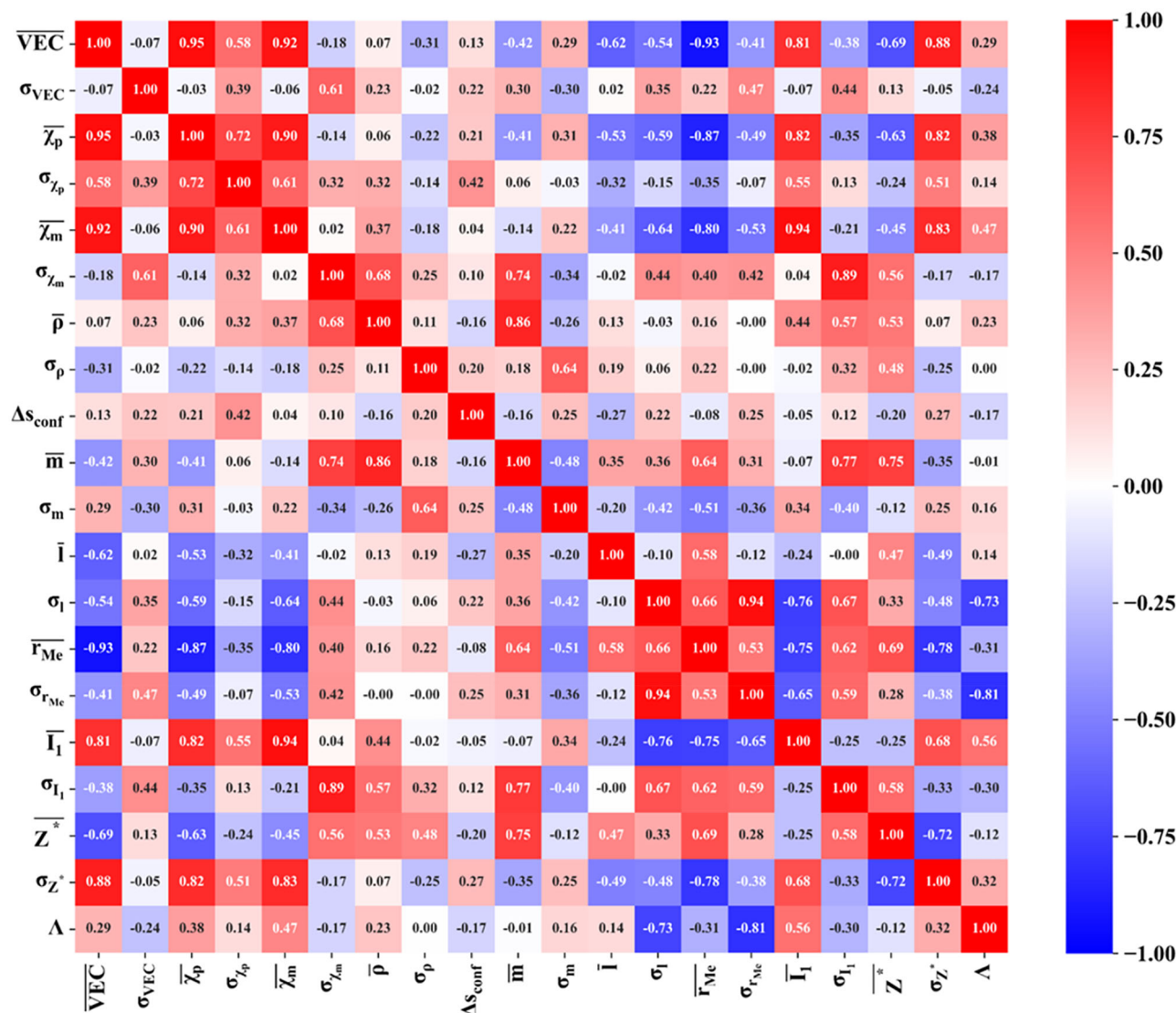


Fig. 3 | Pearson correlation coefficient values of all feature pairs.

Pearson correlation coefficients (PCC) analysis is performed to evaluate the correlations among different features. Figure 3 shows the PCC values between all feature pairs. Features with an absolute PCC value greater than 0.90 are considered highly correlated, and only one feature from each correlated pair is retained. As summarized in Table S3, six pairs of features exhibit absolute PCC values exceeding 0.90, which easily lead to overfitting during ML models training. Therefore, reducing these redundant features is essential for improving the robustness and prediction accuracy of the ML models<sup>30</sup>.

When removing redundant features from the six pairs of highly correlated features, about ten distinct feature deletion schemes are generated, as shown in Table S4. Based on these schemes, a total of 100 ML models are constructed and trained, corresponding to ten different ML algorithms for each feature deletion scheme. The AUC values of each model are evaluated through 5-fold repeated cross-validation to assess the effects of both feature deletion schemes and ML algorithms on model performance. Figure 4 reveals the AUC values of the 100 individual ML models constructed using ten different ML algorithms across ten feature deletion schemes. As can be seen from Fig. 4, among the ten feature deletion schemes, scheme 5 exhibits superior overall performance across the ten ML algorithms. Notably, RF algorithm achieves the highest AUC value of 87.49%, making it the best-

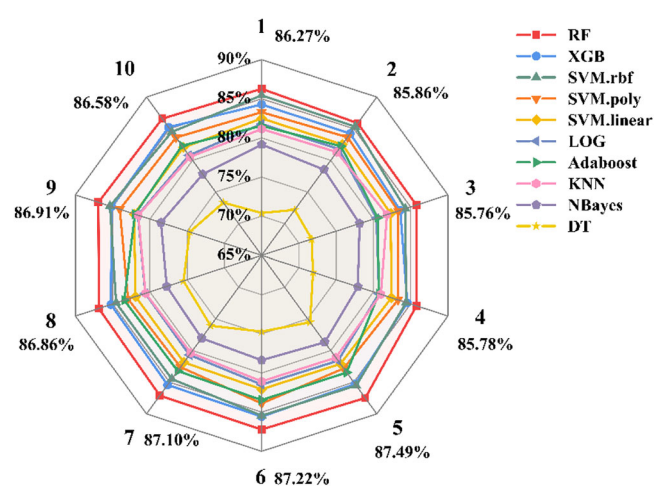


Fig. 4 | AUC values of 100 individual ML models built by ten different ML algorithms for ten feature deletion schemes. The specific representatives of such ten schemes are shown in detail in Table S4 in supplementary document.

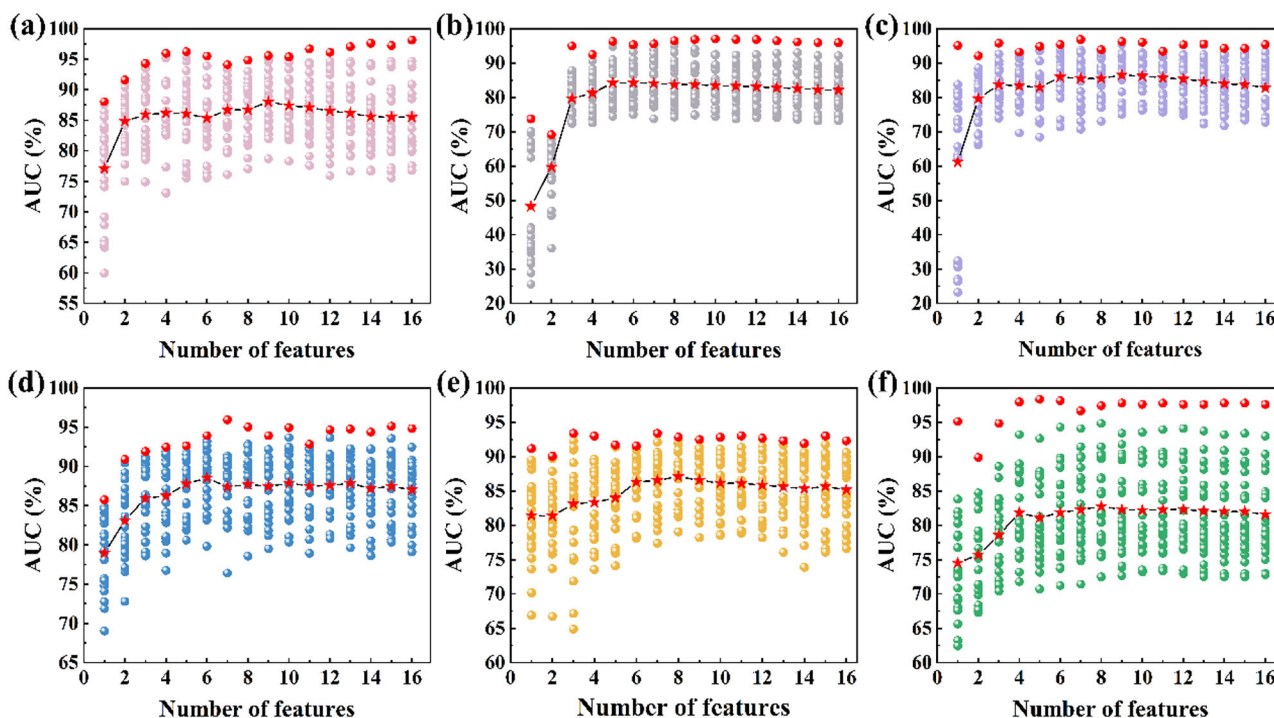


Fig. 5 | Recursive feature elimination for six ML models. a SVM.rbf, b SVM.linear, c SVM.poly, d RF, e XGB, f LOG.

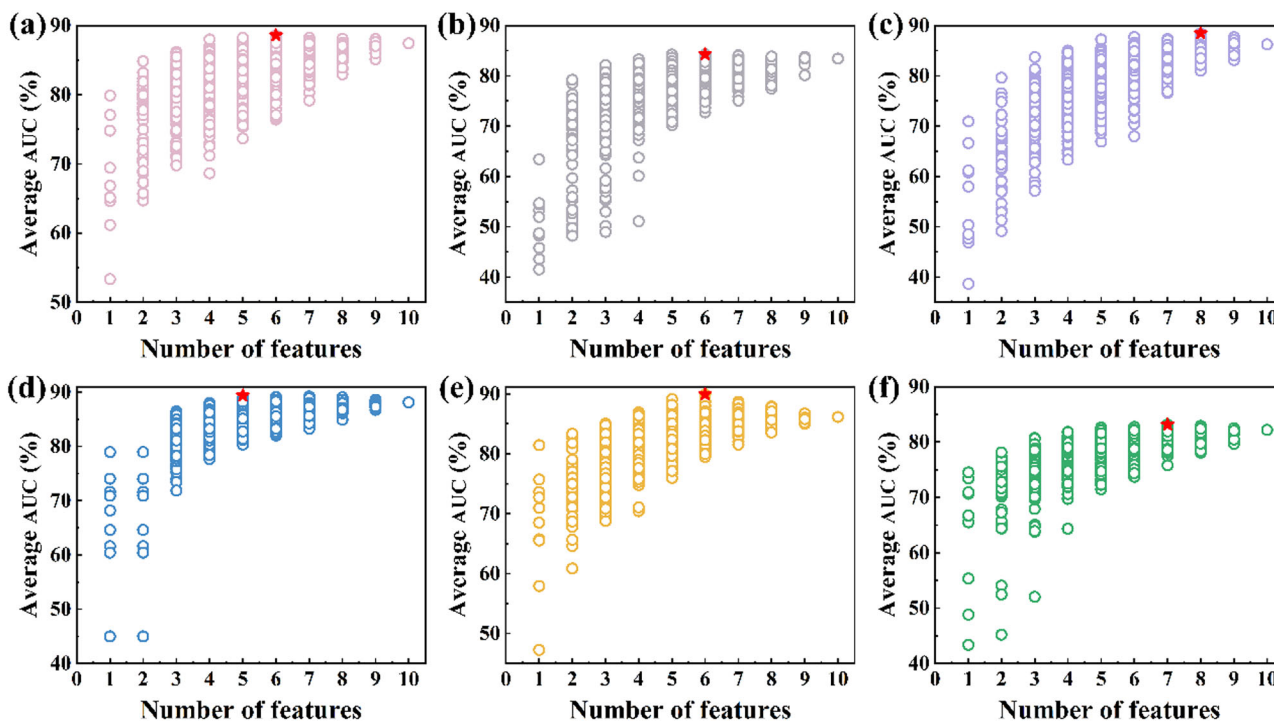


Fig. 6 | Exhaustive screening for six ML models. a SVM.rbf, b SVM.linear, c SVM.poly, d RF, e XGB, f LOG.

performing model among all 100 ML models. Accordingly, 16 features are retained as input features for subsequent ML modeling to predict the phase formation ability of HECCs. These features specifically include  $\sigma_{VEC}$ ,  $\bar{\chi}_P$ ,  $\sigma_{\chi_m}$ ,  $\sigma_{\chi_m}$ ,  $\bar{p}$ ,  $\sigma_p$ ,  $\bar{m}$ ,  $\sigma_m$ ,  $\bar{l}$ ,  $r_{Me}^-$ ,  $\sigma_{r_{Me}}$ ,  $\sigma_{l_1}$ ,  $z^*$ ,  $\sigma_{z^*}$ ,  $\Delta Sconf$ , and  $\Lambda$ . Among the ten ML algorithms evaluated, RF, XGB, SVM.rbf, SVM.poly, SVM.linear and LOG algorithms show better model performances than other algorithms. As a result, these six ML algorithms are employed for further feature engineering

and hyperparameter optimization to search for the optimal feature subset and model parameters for improving prediction performance.

Recursive feature elimination is a feature selection process that iteratively remove the least important features by training an initial model and ranking feature importance until only the most significant features remain<sup>46</sup>. In comparison to recursive feature elimination, exhaustive feature combination evaluates the performance of each feature subset by traversing all

possible combinations<sup>47</sup>. To enhance the efficiency of feature screening and identify a more globally optimal feature subset, both recursive feature elimination and exhaustive feature combination are employed in this study. Figures 5 and 6 illustrate the whole process of recursive feature elimination and exhaustive feature combination applied to six ML models. In Fig. 5, each dot represents the score of an individual fold in the 5-fold repeated cross-validation for different feature numbers, while the red dots indicate the highest scores among all folds for the corresponding feature number. The red stars in Fig. 5 correspond to the average scores of the six ML models, calculated from 5-fold repeated cross-validation. As shown in Fig. 5(a-f), as the number of features gradually decreases, the average scores of the six ML models exhibit a trend of initially increasing and subsequently decreasing. In

the early stages of feature elimination, as the number of features decreases, removing unimportant features reduces noise and redundancy, thereby improving model performance. However, when the number of retained features is further reduced, some important features may be excluded, preventing the models from fully capturing key information in the dataset. This leads to a significant decline in model performance. After recursive feature elimination, the optimal numbers of features with the highest AUC values corresponding to SVM.rbf, SVM.linear, SVM.poly, RF, XGB and LOG models are 9, 5, 6, 6, 8, and 4, respectively.

Figure 6 presents an exhaustive enumeration of all feature combinations derived from the top ten most important features identified by the recursive feature elimination results shown in Fig. 5. From Fig. 6(a-f), it can be seen that all six ML models present a similar trend with respect to recursive elimination, that is, as the number of features increases, the ML models performance initially increases and then decreases. Considering the trade-off between higher predictive performance and fewer input features, the SVM.rbf, RF and XGB algorithms outperform the SVM.linear, SVM.poly and LOG algorithms. Table 2 specifically summarizes the optimal number of features, corresponding feature subsets, and model performances for six ML algorithms. It can be seen that the optimal feature subsets are primarily divided into two categories: those screened by SVM-based and LOG models, and those filtered by RF and XGB models. This distinction arises from the differing mechanisms used by these algorithms to evaluate feature importance. Both SVM and LOG evaluate feature importance by optimizing the weight coefficients of the models and automatically selecting key features by regularization method. For RF and XGB models, both of which rely on decision tree as base learners, determine feature importance based on the contribution of features within the tree structures<sup>28</sup>. Owing to their ensemble and regularization mechanisms, RF and XGB algorithms effectively mitigate overfitting, resulting in a high degree of overlap in the selected feature subsets.

**Table 2 | Optimal number of features, corresponding feature subsets and model performances for six ML models obtained by exhaustive screening**

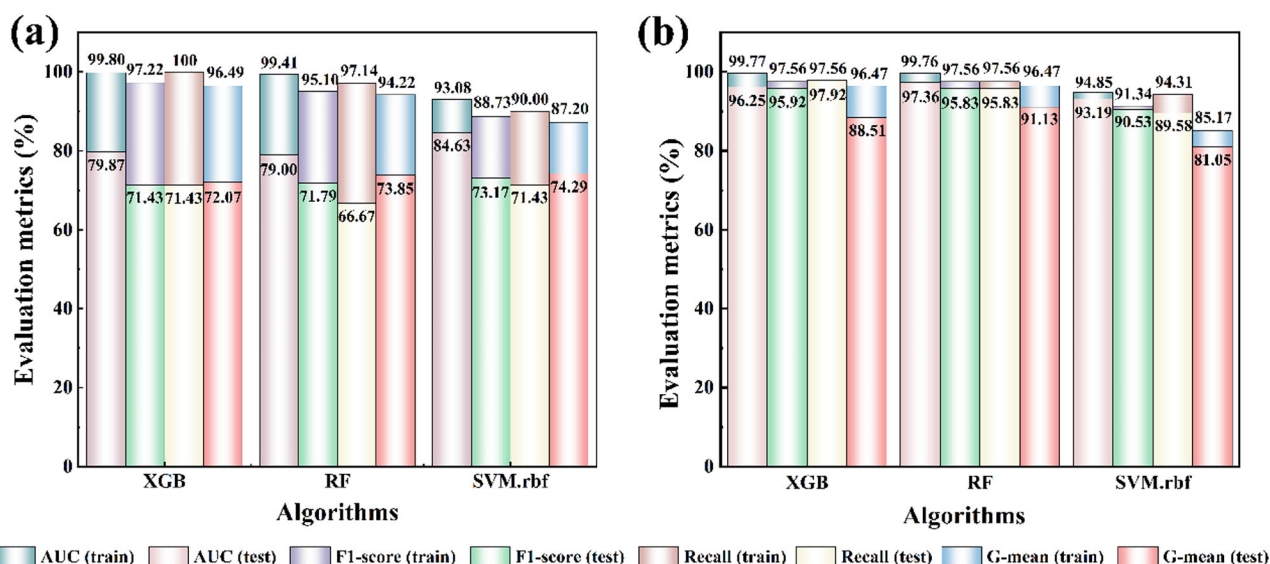
Algorithm	Feature number	Best feature subset	Best AUC value
SVM.rbf	6	$\sigma_{VEC}, \sigma_{\chi_p}, \Delta Sconf, \overline{r_{Me}}, \sigma_{r_{Me}}, \overline{Z^x}$	88.58%
SVM.linear	8	$\sigma_{\chi_p}, \sigma_{\chi_m}, \Delta Sconf, \overline{r_{Me}}, \sigma_{r_{Me}}, \sigma_{Z^*}$	84.28%
SVM.poly	6	$\sigma_{VEC}, \sigma_{\chi_p}, \sigma_{\chi_m}, \Delta Sconf, \overline{r_{Me}}, \sigma_{r_{Me}}, \overline{Z^x}$	88.46%
RF	5	$\sigma_{VEC}, \overline{\chi_p}, \sigma_{I_1}, \overline{Z^x}, \Lambda$	90.22%
XGB	6	$\overline{r_{Me}}, \Lambda, \overline{\chi_p}, \sigma_{I_1}, \sigma_{VEC}, \overline{Z^x}$	89.92%
LOG	7	$\sigma_{VEC}, \sigma_{\chi_p}, \sigma_{\chi_m}, \Delta Sconf, \overline{r_{Me}}, \sigma_{r_{Me}}, \sigma_{I_1}$	83.15%

**Table 3 | Best hyperparameters for SVM.rbf, RF, XGB models**

Algorithm	Best hyperparameter
SVM.rbf	C = 1.92, gamma=0.45
RF	n_estimators=419, max_depth=19, max_features = 'log2'
XGB	n_estimators=80, max_depth=6, learning_rate=0.3

**Hyperparameter optimization and model evaluation**

In order to prevent overfitting or underfitting and improve the generalization ability of ML models, hyperparameter optimization is carried out for the SVM.rbf, RF and XGB models, which show the highest AUC values during the exhaustive screening process. The hyperparameter search spaces explored for these models are summarized in the supplementary information (Table S5). As shown in Table 3, the optimal hyperparameters identified for the SVM.rbf, RF, and XGB models are obtained using the RandomizedSearchCV method. Taking the RF model as an example,



**Fig. 7 | Performance evaluation of three ML models on training and test datasets using metrics AUC, F1-score, Recall, and G-mean. a Without Borderline-SMOTE; b with Borderline-SMOTE.**

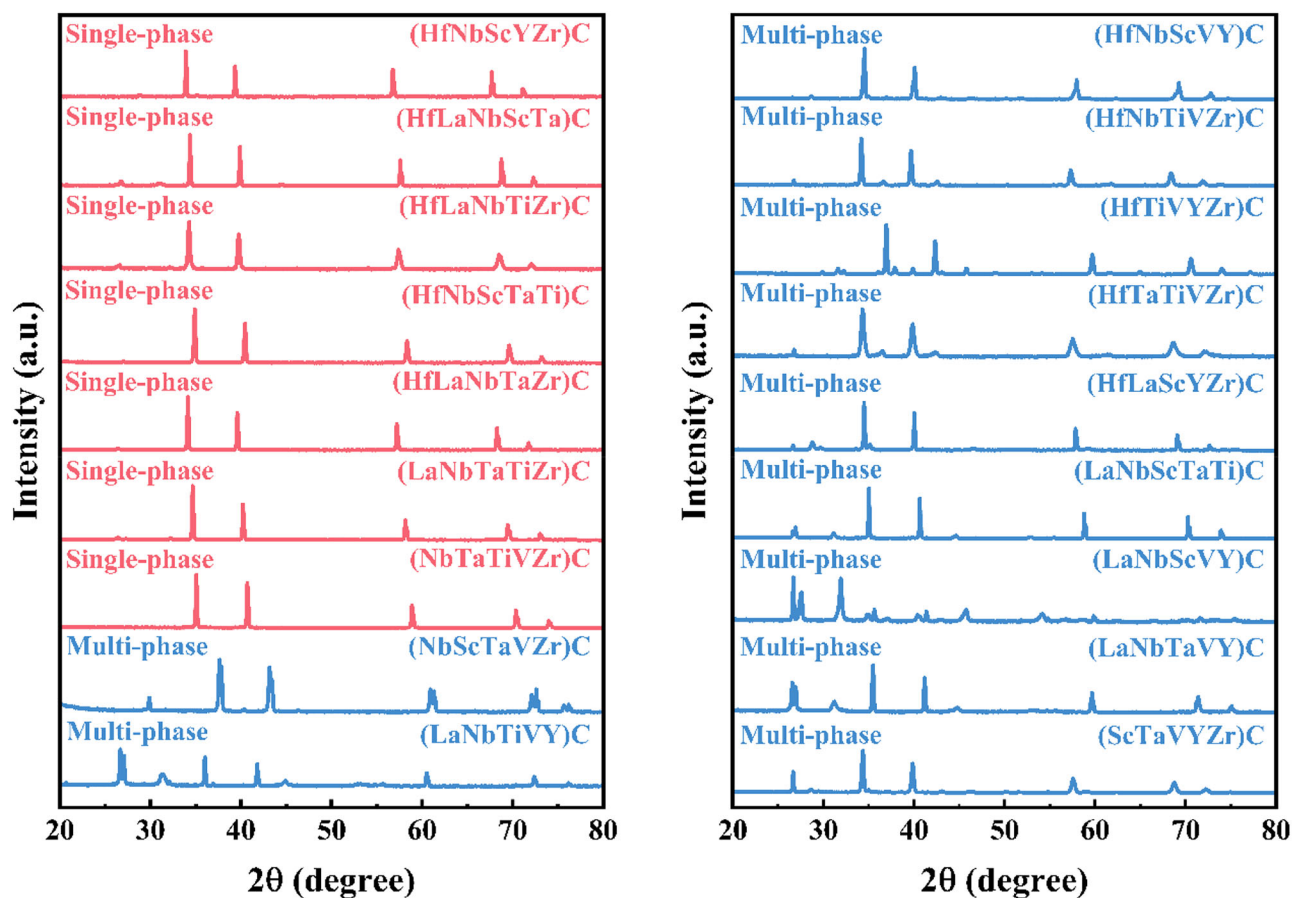


Fig. 8 | XRD patterns of 18 quinary HECCs verified by experiment.

RandomSearchCV is applied to explore all the combinations of  $n_{\text{estimators}}$  in the range of [100, 1000],  $\text{max\_depth}$  in the range of  $2^{20}$ , and  $\text{max\_features}$  with options ['auto', 'sqrt', 'log2']. The results indicate that the RF model achieves the best AUC performance with  $n_{\text{estimators}}=419$ ,  $\text{max\_depth}=19$ , and  $\text{max\_features} = \text{'log2'}$ , respectively.

Figure 7 compares the prediction performance of three ML models on the training and test dataset, both without and with Borderline-SMOTE. The evaluation metrics selected include AUC, F1-score, Recall and G-mean, which are generally used to evaluate the model performance during data imbalance. Among these evaluation metrics, F1-score and Recall mainly focus on the prediction performance for the positive samples, that is, single-phase HECCs, while AUC and G-mean reflect a more comprehensively assessment of the predictive performances for both positive (single-phase HECCs) and negative (multiple-phase carbides) samples at the same time<sup>48</sup>. As can be seen from Fig. 7(a), without Borderline-SMOTE, all three models perform well on the training dataset. In particular, the Recall value of the XGB model on the training dataset reaches 100%, which indicates that the model can correctly identify all single-phase HECCs samples in the training data. Nevertheless, the performance of the three models on the test dataset is poor, which demonstrates that data imbalance exacerbates the overfitting of the model. In Fig. 7(b), after applying Borderline-SMOTE, all three models show good predictive performance on both the training and test datasets. This indicates that the Borderline-SMOTE-assisted machine learning method contributes to tackling the overfitting and effectively improves the generalization ability of the model. A comprehensive comparison of the three models shows that RF is the most suitable model for predicting the phase formation ability of HECCs in the current investigation.

### Model prediction and experimental validation

The well-trained RF model is employed to predict the phase formation ability of novel HECCs within the compositional space consisting of Sc, Y, La, Zr, Ti, Hf, V, Nb, and Ta elements. The detailed prediction results are summarized in Supplementary information. The classification threshold is set at 0.5, where prediction probabilities closer to 0 indicate a higher likelihood of forming multiple-phase carbides, and probabilities closer to 1 suggest a higher likelihood of forming single-phase HECCs. Based on the predicted results, eight HECCs samples with relatively high uncertainty and ten HECCs samples with low uncertainty are selected for experimental verification and characterization (Table S6). High uncertainty refers to samples with predicted probabilities in the range of 0.3 to 0.7, where the predicted labels of the classifier may be ambiguous and prone to misclassification. In contrast, low uncertainty refers to samples with predicted probabilities in the ranges of 0–0.3 or 0.7–1, which are expected to show good consistency with experimental results. The phase compositions of these above 18 HECCs samples were analyzed by XRD, and their XRD patterns are shown in Fig. 8. Seven samples, including (HfNbScYZr)C, (HfLaNbScTa)C, (HfLaNbTiZr)C, (HfNbScTaTi)C, (HfLaNbTaZr)C, (LaNbTaTiZr)C, and (NbTaTiVZr)C, exhibit single-phase with face-centered cubic structures. The remaining 11 samples including (NbScTaVZr)C, (LaNbTiVY)C, (HfNbScVY)C, (HfNbTiVZr)C, (HfTiVYZr)C, (HfTaTiVZr)C, (HfLaScYZr)C, (LaNbScTaTi)C, (LaNbScVY)C, (LaNbTaVY)C, and (ScTaVYZr)C are identified as multiple-phase carbides. Comparing the experimental results with RF model predicted results reveals that, among 18 experimental results, except for (LaNbScTaTi)C, (HfTaTiVZr)C, (NbTaTiVZr)C, and (LaNbScVY)C, the experimental results of the remaining 14 samples are consistent with RF model prediction results, which implies that the prediction accuracy of the model is 77.78%.

**Table 4 | Comparison of prediction and actual results of RF model**

Composition	Predicted results	Actual results	Composition	Predicted results	Actual results
(HfNbScYZr)C	0.82	Single	(HfNbScVY)C	0.21	Multiple
(HfLaNbScTa)C	0.53	Single	(HfNbTiVZr)C	0.35	Multiple
(HfLaNbTiZr)C	0.58	Single	(HfTiVYZr)C	0.40	Multiple
(HfNbScTaTi)C	0.72	Single	(HfTaTiVZr)C	0.86	Multiple
(HfLaNbTaZr)C	0.53	Single	(HfLaScYZr)C	0.12	Multiple
(LaNbTaTiZr)C	0.40	Single	(LaNbScTaTi)C	0.36	Multiple
(NbTaTiVZr)C	0.55	Single	(LaNbScVY)C	0.24	Multiple
(NbScTaVZr)C	0.37	Multiple	(LaNbTaVY)C	0.17	Multiple
(LaNbTiVY)C	0.25	Multiple	(ScTaVYZr)C	0.36	Multiple

To further improve the robustness and reliability of the prediction model, we employed an active learning strategy to intentionally incorporate additional valuable HECCs samples. Specifically, the above eight HECCs samples with relatively high uncertainty are sequentially added to the dataset, and the RF model is retrained and optimized over multiple iterations to further enhance prediction performance. The results of each iteration are summarized in the supplementary information. It can be seen that, after five iterations, the prediction accuracy of the optimized RF model tends to be stable, maintaining an experimental validation accuracy of 88.89%. The specific prediction results of the 18 experimental samples after five iterations are shown in Table 4. Comparing the experimental results in Fig. 8 and the supplementary information (Figs. S2–S19), except for (LaNbTaTiZr)C and (HfTaTiVZr)C, the phase formation ability of the remaining 16 HECCs samples is successfully predicted. This strong agreement between model predictions and experimental validation demonstrates the good prediction ability of the proposed ML model.

Based on the above results, two samples with model prediction deviations, (LaNbTaTiZr)C and (HfTaTiVZr)C, are selected for TEM characterization. As shown in the dark-field images, high-angle annular dark-field (HAADF) images, and EDS elemental mapping results in Fig. 9(a, b, e), (LaNbTaTiZr)C exhibits ideal elemental homogeneity at the nanoscale, with no significant elemental segregation detected. Notably, the results in Fig. 9(f, g, j) confirm the presence of elemental segregation in (HfTaTiVZr)C, indicating that the VC phase fails to fully dissolve into the matrix lattice, leading to the formation of V-rich regions. Further analysis of the high-resolution transmission electron microscopy (HRTEM) images in Fig. 9(c, h, k) reveals that both samples maintain well-ordered periodic lattice structure of metal carbides. The measured lattice spacings of (LaNbTaTiZr)C are 2.50 Å, slightly smaller than that of the main phase in the multiphase (HfTaTiVZr)C, measured at 2.55 Å. This variation in lattice parameters primarily originates from the atomic size effect and lattice distortion in HECCs. According to Vegard's law<sup>49</sup>, the average atomic radius differences ( $\Delta r$ ) of the (LaNbTaTiZr)C systems is 9.8%, exceeding the critical threshold for solid solution formation ( $\Delta r_{\max}=6.5\%$ ), leading to significant lattice distortion. Additionally, the selected area electron diffraction (SAED) patterns in Fig. 9(d, i, l) confirm that both samples exhibit a typical face-centered cubic (FCC) structure with space group  $Fm\bar{3}m$ . In summary, the TEM results align well with the XRD results, clearly demonstrating the effectiveness of incorporating HECC samples with high uncertainty into the dataset and iteratively improving the model.

Figure 10 illustrates the predicted single-phase formation probability of several HECCs systems, including (HfZrTaScY)C, (HfZrTaScLa)C, and (HfZrTaYLa)C, within the non-equimolar composition space using the established RF model. By analyzing the relationship between compositional variations and single-phase formation ability across these three systems, Fig. 10(a–c) highlights the significant impact of element ratios on the formation of single-phase HECCs. When the content of (HfZrTa)C is less than 0.25, corresponding to a relatively high concentration of rare-earth elements, all three systems tend to form single-phase HECCs in a relatively broad concentrational range. Conversely, when the (HfZrTa)C content exceeds 0.85,

corresponding to a lower concentration of rare-earth elements, only a small compositional region in the lower-right corner of the predicted phase diagrams exhibits a high likelihood of forming single-phase HECCs. In addition, comparison of the three systems reveals that the (HfZrTaScLa)C system possesses fewer regions favorable for single-phase formation compared to the (HfZrTaScY)C and (HfZrTaYLa)C systems. Furthermore, as the Sc content increases, the formation of single-phase HECCs becomes progressively more challenging. These findings serve as a valuable reference for experimental efforts aimed at tailoring compositions to achieve the desired phase stability in the (HfZrTaScY)C, (HfZrTaScLa)C, and (HfZrTaYLa)C systems. Overall, the predicted phase diagrams provide critical insights for the rational design of HECCs systems, offering intuitive guidance for optimizing elemental ratios and exploring potential single-phase HECCs materials.

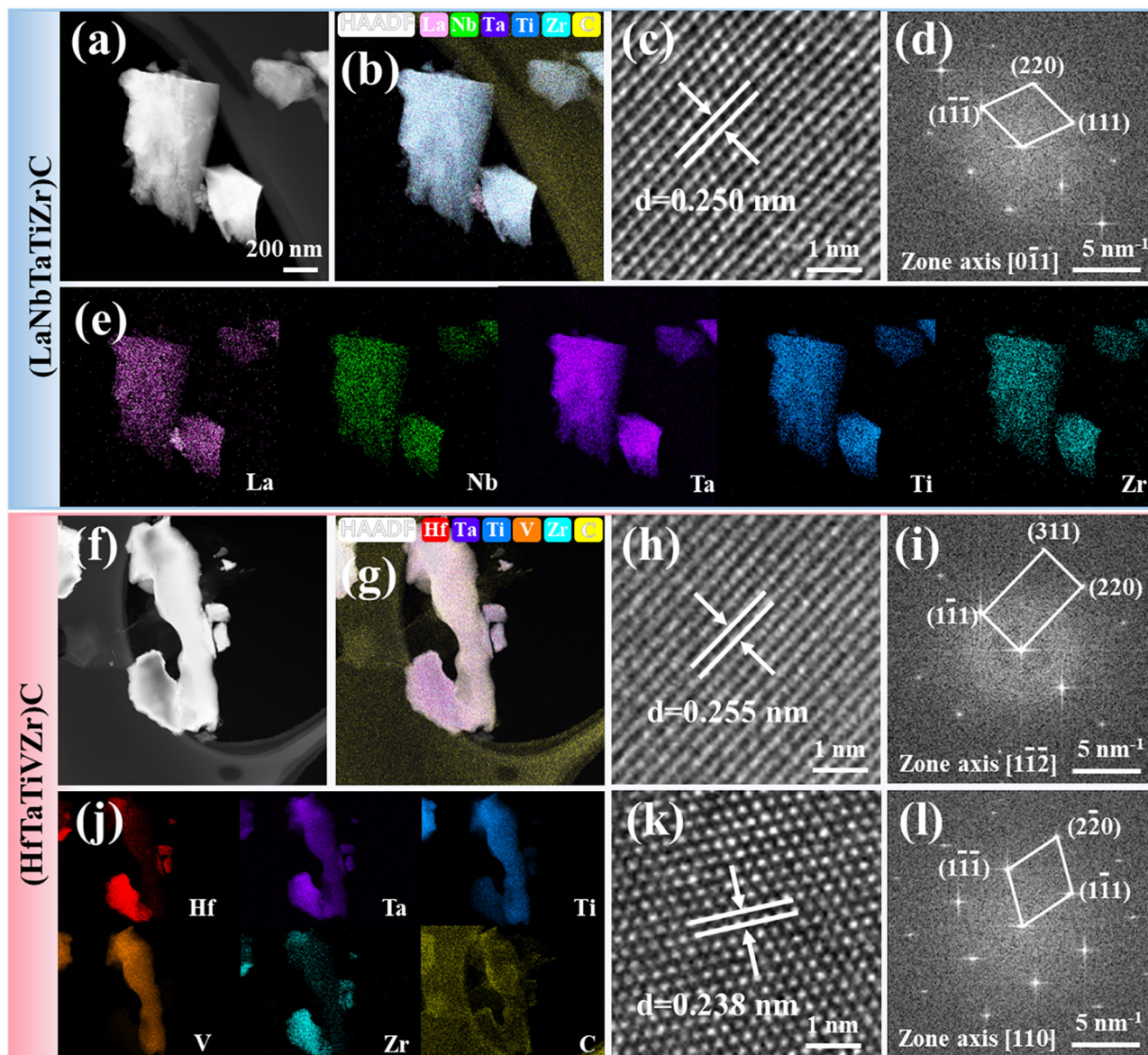
## Discussion

In this work, we developed a machine learning framework to predict the phase formation ability of HECCs containing rare-earth elements, addressing the challenge of data imbalance arising from the limited availability of experimentally reported compositions. To overcome this limitation, Borderline-SMOTE was employed to expand the dataset of HECCs containing rare-earth elements from 18 samples (3 from literature and 15 from experimental synthesis) to 98 samples, thereby enabling a more balanced and representative dataset for model training. This strategy not only alleviated the negative effects of data scarcity and imbalance, but also significantly improved the reliability of subsequent model training.

Furthermore, a comprehensive four-step feature selection strategy was employed, integrating Pearson correlation coefficients, recursive feature elimination, exhaustive feature combinations, and hyperparameter optimization. This systematic approach enabled the identification of key features governing phase stability while minimizing redundancy and overfitting risks. The performance of the machine learning models was rigorously assessed using multiple metrics, including AUC, F1-score, Recall, and G-mean. Among the evaluated models, the Random Forest (RF) model achieved the best overall balance between accuracy and generalization, with AUC, F1-score, Recall, and G-mean values of 97.36%, 95.83%, 95.83%, and 91.13% on the test set, respectively.

The well-trained RF model was further applied to predict the phase formation ability of equimolar HECCs. Eight compositions with relatively high predictive uncertainty were selected for experimental synthesis and phase identification, and their outcomes were iteratively incorporated into the dataset. This feedback loop between prediction and experiment improved model reliability, ultimately achieving an experimental validation accuracy of 88.89%. Importantly, the optimized RF model was extended to explore the broader non-equimolar composition space of the (HfZrTaScY), (HfZrTaScLa), and (HfZrTaYLa) systems, providing guidance for designing novel HECCs and laying the foundation for discovering materials with tailored properties.

While Borderline-SMOTE effectively improves class balance, it introduces synthetic data that may not fully capture the complexity of real



**Fig. 9** | TEM results of two experimentally synthesized (LaNbTaTiZr)C and (HfTaTiVZr)C samples. **a, b, f, g** HAADF images; **c, h, k** HRTEM images; **d, i, l** SAED patterns; **e, j** EDS elemental mapping images.

experimental systems. Future work will focus on further expanding the dataset through high-throughput synthesis, integrating domain-knowledge-based features, and exploring advanced machine learning strategies to further enhance predictive accuracy and support the rational design of HECCs with tailored properties.

## Methods

Figure 11 shows the flowchart of phase prediction for HECCs based on Borderline-SMOTE-assisted machine learning. It mainly consists of four parts: data preparation, feature engineering, ML model building, and model evaluation.

### Data preparation

To construct a comprehensive dataset, we collected a wide range of HECCs from relevant literature, spanning ternary to novenary compositions. This dataset encompasses 84 single-phase HECCs and 72 multiple-phase carbides, totaling 156 samples, as summarized in Table S1. These HECCs samples are composed of elements from group III<sub>B</sub> (Sc, Y, and La), IV<sub>B</sub> group (Ti, Zr, and Hf), V<sub>B</sub> group (V, Nb, and Ta) and VI<sub>B</sub> group (Cr, Mo, and W) in the periodic table. Notably, only three HECCs samples

containing Sc, Y, or La were collected, highlighting a significant data imbalance in this dataset. For ML model training, HECCs with single-phases were labeled as “1”, whereas those forming multiple-phases carbides were labeled as “0”.

To establish the input features for the ML models, relevant properties of HECCs and their constituent carbides were collected, resulting in a total of 20 features (Table 1). These features specifically include valence electron concentration (VEC), Pauling electronegativity ( $\chi_p$ ), Mulliken electronegativity ( $\chi_m$ ), density ( $\rho$ ), mass ( $m$ ), lattice size ( $l$ ), metallic radius ( $r_{Me}$ ), first ionization energy ( $I_1$ ), effective nuclear charge ( $Z^*$ ), configurational entropy ( $\Delta S_{conf}$ ) and geometric parameters ( $\Lambda$ ). The  $\Delta S_{conf}$  of HECCs was calculated based on the molar fraction of each element ( $X_i$ ) by the formula (1):

$$\Delta S_{conf} = -R \sum_{i=1}^N X_i \ln(X_i) \quad (1)$$

where  $N$  is the number of elements in HECCs, and  $R$  represents the ideal gas constant ( $R = 8.314 \text{ J}/(\text{K mol})$ ). The average ( $\overline{\text{prop}}$ ) and standard deviation ( $\sigma_{\text{prop}}$ ) values of the relevant properties, including VEC,  $\chi_p$ ,  $\chi_m$ ,  $\rho$ ,  $m$ ,  $l$ ,  $r_{Me}$ ,  $I_1$ ,

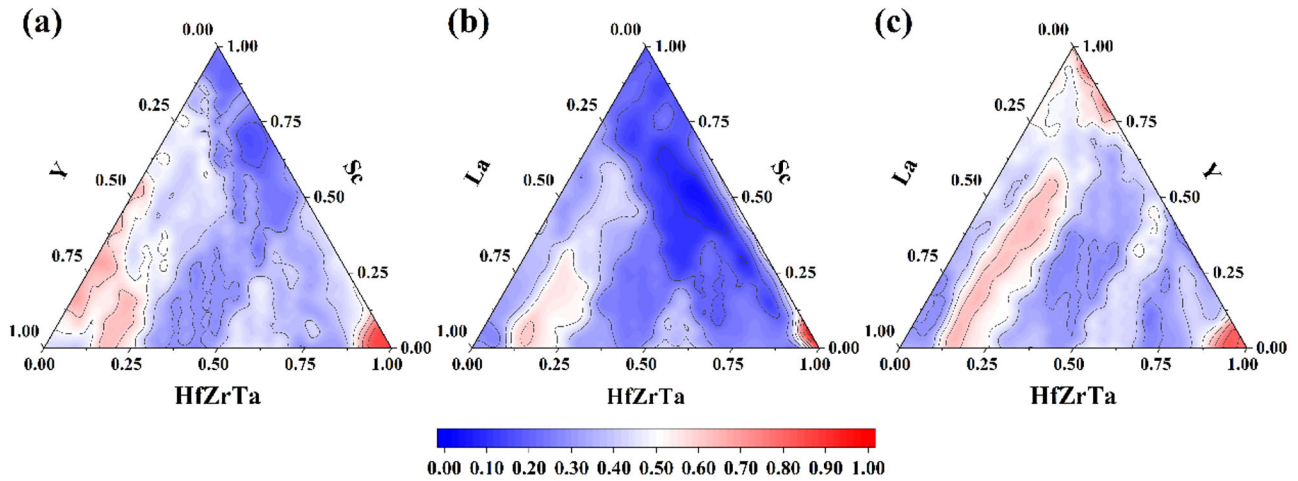


Fig. 10 | Phase formation diagrams of quinary HECCs systems predicted by the well-trained RF model. a (HfZrTaScY)C; b (HfZrTaScLa)C; c (HfZrTaYLa)C.

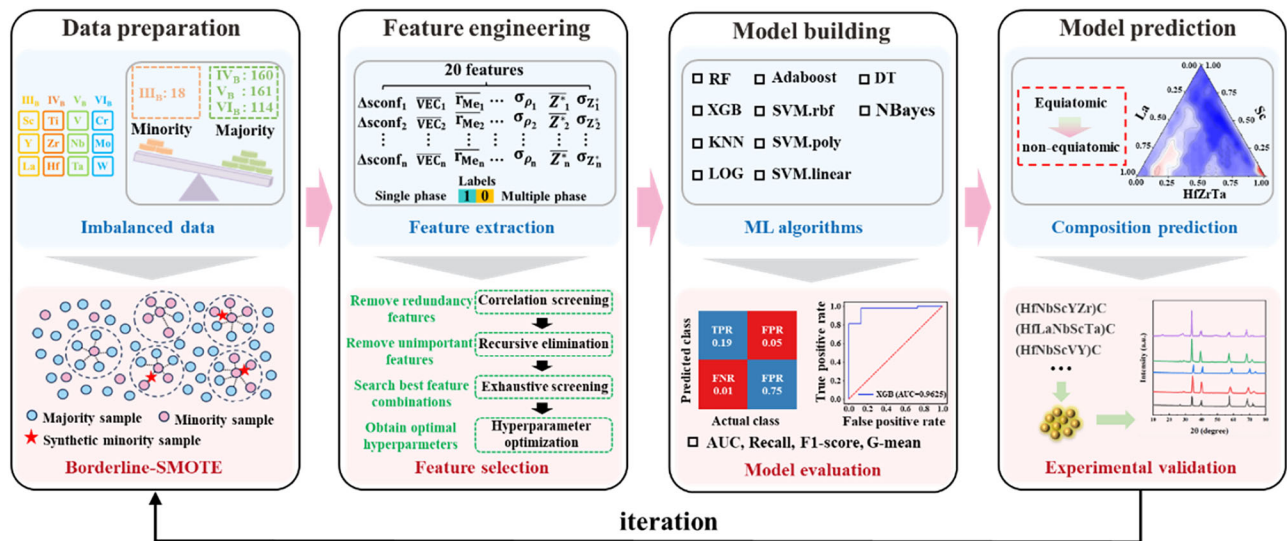


Fig. 11 | Technique flowchart of phase prediction for HECCs based on Borderline-SMOTE-assisted machine learning.

and  $Z^*$ , were calculated by the following formulas:

$$\overline{\text{prop}} = \sum_{i=1}^N X_i(\text{prop})_i \quad (2)$$

$$\sigma_{\text{prop}} = \sqrt{\sum_{i=1}^N X_i \left(1 - \frac{\text{prop}_i}{\overline{\text{prop}}}\right)^2} \quad (3)$$

where  $X_i$  is the molar fraction of each element in HECCs, and  $\text{prop}_i$  refers to a certain property of the  $i^{\text{th}}$  constituent binary carbide of HECCs. The geometric parameter ( $\Lambda$ ) was defined as follows:

$$\Lambda = \frac{\Delta\text{Sconf}}{(\sigma_1)^2} \quad (4)$$

### Borderline-SMOTE method

The Borderline-SMOTE method is an oversampling algorithm that develops on the basis of traditional SMOTE technique, which posits that samples located near the classification boundary are often crucial for determining classification results. Borderline-SMOTE adopts a more targeted strategy, which divides the minority class samples into three non-overlapping

categories: (1) safety samples: minority class samples where the majority of neighbors also belong to the minority class, indicating they are located well within the decision region of the minority class samples. (2) noise samples: minority class samples whose nearest neighbors are entirely composed of the majority class samples, suggesting they are situated within the decision region of the majority class samples. (3) danger samples: minority class samples that have a mix of minority and majority neighbors, which are often located near the decision boundary of majority and minority class samples. After that, Borderline-SMOTE focuses exclusively on the danger samples and applies the SMOTE technique to generate new synthetic minority class samples. This selective oversampling strategy ensures that the newly generated samples are positioned near the decision boundaries that are prone to misclassification, which is beneficial to improve the prediction performance of ML models.

### Feature engineering

Feature engineering plays a crucial role in determining the ultimate performance of ML models. Since there is a lack of prior knowledge that certain features are inherently superior to others, it is essential to accumulate as many potentially relevant features as possible when constructing the initial dataset<sup>44</sup>. In this study, 20 features related to phase formation were utilized as input features for subsequent ML models. However, it is worth noting that material features often vary in terms of magnitudes, units and value ranges,

which have an enormous impact on the predicted accuracy of ML models. To address this issue, it is necessary to normalize the initial dataset, ensuring that all features were unified to the same magnitude level. This normalization process prevents features with smaller magnitudes from being overlooked and avoids disproportionately assigning greater weights to features with larger magnitudes<sup>50</sup>. The normalization function was defined by formula (5):

$$X_i^{\text{norm}} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \quad (5)$$

where  $X_i$  and  $X_i^{\text{norm}}$  represent the  $i$ -th values of the input feature  $X$  before and after normalization, respectively.  $X_{\max}$  and  $X_{\min}$  denote the maximum and minimum values of the input feature  $X$ , respectively. After normalization, the dataset samples were randomly shuffled to eliminate any potential ordering bias and to ensure robust performance during cross-validation and model training. This procedure enhances the statistical representativeness of the training and validation subsets, ultimately supporting better model generalization.

A four-step feature selection strategy was executed on the shuffled feature set to eliminate redundant features and screen the optimal feature subset<sup>46,51</sup>. The first step involved correlation screening. Pearson correlation coefficients (PCC) between different features were calculated using formula (6) as follows:

$$\text{PCC} = \frac{\sum_{i=1}^N (F_{ki} - \bar{F}_k)(F_{li} - \bar{F}_l)}{\sqrt{\sum_{i=1}^N (F_{ki} - \bar{F}_k)^2} \sqrt{\sum_{i=1}^N (F_{li} - \bar{F}_l)^2}} \quad (6)$$

where  $N$  is the total number of samples in the dataset,  $F_{ki}$  and  $F_{li}$  are the values of two different features for the  $i^{\text{th}}$  sample,  $\bar{F}_k$  and  $\bar{F}_l$  represent their corresponding average values. Feature pairs with an absolute PCC value greater than 0.90 was considered to be strongly linear correlated, and one feature from each such pair was removed. When highly correlated features were removed, corresponding ML models were then constructed based on different feature deletion combinations, and the combination yielding higher prediction accuracy was remained.

The second step was recursive feature elimination. The remaining features after correlation filtering were subjected to recursive elimination process, in which the least important feature was successively removed. For an  $m$ -dimensional feature set, each  $m-1$  dimensional feature subset was separately utilized to train ML models, and the subset with the highest prediction accuracy was retained for the following recursion. This process was repeated until only one feature remained. After the whole recursive elimination, the subset with the highest prediction accuracy was preserved.

The third step involved exhaustive screening. All possible feature subsets generated from the remaining features after recursive elimination were evaluated, and then the feature subset with the highest prediction accuracy of ML models was selected as the optimal feature combination.

The final step was hyperparameter optimization. The RandomizedSearchCV method was employed to optimize the hyperparameters of ML models, thereby further improving the prediction accuracy<sup>52</sup>.

### Machine learning algorithms

According to the “no free lunch” theorems proposed by Wolpert et al., there is no universally suitable ML algorithm applicable to all materials problems<sup>53</sup>. Therefore, it is necessary to select the most suitable ML algorithm for a specific material problem. For predicting the phase formation ability of HECCs, ten commonly used ML algorithms were applied to the modeling for comparison, including random forest (RF)<sup>54</sup>, Xgboost (XGB)<sup>55</sup>, K-nearest neighbor (KNN)<sup>56</sup>, logistic regression (LOG)<sup>57</sup>, AdaBoosting (Adaboost)<sup>52</sup>, support vector machine using radial basis function, polynomial kernel, and linear kernel (SVM.rbf, SVM.poly, and SVM.linear)<sup>58</sup>, decision tree (DT)<sup>59</sup>, and naive Bayes (NBayes)<sup>60</sup>. More specifically, RF, Adaboost, XGB, and DT are tree-based algorithms that can

capture complex non-linear relationships. KNN is a distance-based algorithm that predicts new data by estimating the proximity to their  $K$  nearest neighbors. LOG is a generalized linear algorithm used for classification problems, which predicts the probability of data belonging to a certain class by combining features linearly. SVM is a kernel-based algorithm whose goal is to identify the optimal hyperplane in the feature space that separates data points of different classes while maximizing the classification boundary margin. NBayes is a classification algorithm based on Bayes’ theorem and the assumption of conditional independence among features. It calculates the posterior probability of each category under a given data sample, and selects the category with the highest posterior probability as the prediction.

During modeling process, the shuffled dataset was randomly divided into training and test subsets, with 75% allocated for training and 25% for testing. To evaluate ML model’s performance, 5-fold repeated cross-validation was adopted, which develops from traditional 5-fold cross-validation process. Specifically, 5-fold cross-validation typically divides the dataset into five non-overlapping subsets, each subset has an opportunity to be used as a test set, while all the remaining subsets serve as the training set. ML models are trained and tested for each fold, and calculate the average performance of all folds. On this basis, 5-fold repeated cross-validation is performed by repeating 5-fold cross-validation. The dataset is randomly redivided in each repetition, and the mean and standard deviation of the results across all repetitions are taken to reduce the variance in the evaluation results and obtain a more stable and reliable model performance estimation.

### Model evaluation

For imbalanced datasets, traditional evaluation metrics such as accuracy can be misleading, as they fail to effectively reflect the model’s performance on minority class samples. The receiver operating characteristic (ROC) curve is often used to summarize the classifier performance under a range of trade-offs between the true positive rate and the false positive rate. The area under the ROC curve (AUC) is a well-known performance metric of the ROC curve, which makes AUC often used to evaluate the performance of imbalanced data<sup>37</sup>. Moreover, metrics like F1-score, Recall, and G-mean have been increasingly employed to comprehensively assess the ML models’ performance<sup>61</sup>. These metrics can provide a more integrated assessment, particularly in scenarios involving imbalanced data. These metrics are defined as follows:

$$\text{TPR} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{FPR} = \frac{FP}{FP + TN} \quad (8)$$

$$\text{AUC} = \int_0^1 \text{TPR} \, d(\text{FPR}) \quad (9)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (11)$$

$$\text{F1 - score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

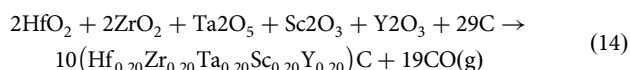
$$\text{G - mean} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \quad (13)$$

where TP, FN, FP, and TN are defined based on the relationship between the ML model-predicted results and the real results. In detail, TP, FN, FP, and TN denote true positive, false negative, false positive, and true negative, respectively.

## Experiments

HECCs samples for expanding the initial dataset and validating the ML model were synthesized by carbothermal reduction route. The compositions were randomly selected from elements in the III<sub>B</sub> (Sc, Y, and La), IV<sub>B</sub> (Ti, Zr, and Hf), V<sub>B</sub> (V, Nb, and Ta) groups in the periodic table. Commercially available raw materials of Sc<sub>2</sub>O<sub>3</sub> (>99.9%, 1–3 μm), Y<sub>2</sub>O<sub>3</sub> (>99.9%, 1–3 μm), La<sub>2</sub>O<sub>3</sub> (>99.9%, 1–3 μm), TiO<sub>2</sub> (>99.9%, 1–3 μm), ZrO<sub>2</sub> (>99.9%, 1–3 μm), HfO<sub>2</sub> (>99.9%, 1–3 μm), V<sub>2</sub>O<sub>5</sub> (>99.9%, 1–3 μm), Nb<sub>2</sub>O<sub>5</sub> (>99.9%, 1–3 μm), Ta<sub>2</sub>O<sub>5</sub> (>99.9%, 1–3 μm) and carbon powder (>99.9%, 300–500 nm) were purchased from Eno Material Development Co. Ltd (Qinhuangdao, China).

Taking (Hf<sub>0.20</sub>Zr<sub>0.20</sub>Ta<sub>0.20</sub>Sc<sub>0.20</sub>Y<sub>0.20</sub>)C as an example, the synthesis mechanism was dominated by the following reaction:



First, the raw material powders were weighed according to the molar ratio of HfO<sub>2</sub>:ZrO<sub>2</sub>:Ta<sub>2</sub>O<sub>5</sub>:Sc<sub>2</sub>O<sub>3</sub>:Y<sub>2</sub>O<sub>3</sub>:C = 2:2:1:1:1:29, using ethanol as the medium. The mixture was stirred for 4 h, and the mixed slurry was dried in an oven at 100 °C for 8 h. Next, the mixture powders were wrapped by graphite paper and placed into a graphite crucible. This crucible was transferred to an induction heating furnace with graphite as the heating component. The furnace temperature was gradually raised to 2100 °C at a rate of 5 °C/min, followed by 2 h isothermal hold and natural cooling under a flowing argon atmosphere. The phase compositions of the synthesized HECCs powders were characterized by X-ray diffraction (XRD; X'Pert Pro-MPD, PANalytical, Netherlands) with Cu Kα radiation. The morphologies and element distributions of the experimentally synthesized HECCs powders were analyzed by electron microscopy using a scanning electron microscope (SEM, ZEISS Gemini 500, Germany) equipped with energy dispersive spectroscopy (EDS, Oxford INCA, Holland). The morphologies and element distributions of the experimentally synthesized (LaNbTaTiZr)C and (HfTaTiVZr)C powders were investigated by transmission electron microscopy (TEM, FEI TECNAI G20, Hillsboro, OR, USA) equipped with the EDS.

## Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files. The underlying code for this study may be made available to qualified researchers on reasonable request from the corresponding author.

## Code availability

The underlying code for this study may be made available to qualified researchers on reasonable request from the corresponding author.

Received: 7 April 2025; Accepted: 17 November 2025;

Published online: 10 January 2026

## References

- Rost, C. M. et al. Entropy-stabilized oxides. *Nat. Commun.* **6**, 8485 (2015).
- Oses, C., Toher, C. & Curtarolo, S. High-entropy ceramics. *Nat. Rev. Mater.* **5**, 295–309 (2020).
- Cui, D. et al. Unraveling microstructure and mechanical response of an additively manufactured refractory TiVHfNbMo high-entropy alloy. *Addit. Manuf.* **84**, 104126 (2024).
- Harrington, T. J. et al. Phase stability and mechanical properties of novel high entropy transition metal carbides. *Acta Mater.* **166**, 271–280 (2019).
- Gild, J. et al. High-entropy metal diborides: a new class of high-entropy materials and a new type of ultrahigh temperature ceramics. *Sci. Rep.* **6**, 37946 (2016).
- Jin, T. et al. Mechanochemical-assisted synthesis of high-entropy metal nitride via a soft urea strategy. *Adv. Mater.* **30**, 1707512 (2018).
- Gild, J. et al. A high-entropy silicide: (Mo<sub>0.2</sub>Nb<sub>0.2</sub>Ta<sub>0.2</sub>Ti<sub>0.2</sub>W<sub>0.2</sub>)Si<sub>2</sub>. *J. Mater.* **5**, 337–343 (2019).
- Wang, Y., Zhang, B., Zhang, C., Yin, J. & Reece, M. J. Ablation behaviour of (Hf-Ta-Zr-Nb)C high entropy carbide ceramic at temperatures above 2100 °C. *J. Mater. Sci. Technol.* **113**, 40–47 (2022).
- Ma, M. et al. Nanocrystalline high-entropy carbide ceramics with improved mechanical properties. *J. Am. Ceram. Soc.* **105**, 606–613 (2021).
- Zhang, Y. et al. Heat dissipation of carbon shell in ZrC-SiC/TaC coating to improve protective ability against ultrahigh temperature ablation. *J. Adv. Ceram.* **13**, 1080–1091 (2024).
- Xu, Y. et al. Composition design of oxidation-resistant non-equimolar high-entropy ceramic materials: An example of (Zr-Hf-Ta-Ti)B<sub>2</sub> ultra-high temperature ceramics. *J. Adv. Ceram.* **13**, 2087–2100 (2024).
- Guo, L., Huang, S., Li, W., Lv, J. & Sun, J. Theoretical design and experimental verification of high-entropy carbide ablative resistant coating. *Adv. Powder Mater.* **3**, 100213 (2024).
- Dong, S., Ni, D. & Cai, F. Research progress of high-entropy carbide ultra-high temperature ceramics. *J. Inorg. Mater.* **39**, 591–608 (2024).
- Dube, T. C. & Zhang, J. Underpinning the relationship between synthesis and properties of high entropy ceramics: A comprehensive review on borides, carbides and oxides. *J. Eur. Ceram. Soc.* **44**, 1335–1350 (2024).
- Zhang, J. et al. Rational design of high-entropy ceramics based on machine learning—a critical review. *Curr. Opin. Solid State Mater. Sci.* **27**, 101057 (2023).
- Divilov, S. et al. Disordered enthalpy-entropy descriptor for high-entropy ceramics discovery. *Nature* **625**, 66–73 (2024).
- Hume-Rothery, W. *Atomic Theory for Students of Metallurgy* (The Institute of Metals, 1952).
- Pei, Z., Yin, J., Hawk, J. A., Alman, D. E. & Gao, M. C. Machine-learning informed prediction of high-entropy solid solution formation: beyond the Hume-Rothery rules. *npj Comput. Mater.* **6**, 50 (2020).
- Hume-Rothery, W. & Powell, H. M. On the theory of super-lattice structures in alloys. *Z. Kristallographie-Cryst. Mater.* **91**, 23–47 (1935).
- Mizutani, U., Sato, H. & Massalski, T. B. The original concepts of the Hume-Rothery rule extended to alloys and compounds whose bonding is metallic, ionic, or covalent, or a changing mixture of these. *Prog. Mater. Sci.* **120**, 100719 (2021).
- Hume-Rothery, R. W. S. W., Haworth, C. W. *The Structure of Metals and Alloys* (The Institute of Metals, 1936).
- Poletti, M. G. & Battezzati, L. Electronic and thermodynamic criteria for the occurrence of high entropy alloys in metallic systems. *Acta Mater.* **75**, 297–306 (2014).
- Liu, S. et al. Stability and mechanical properties of single-phase quinary high-entropy metal carbides: first-principles theory and thermodynamics. *J. Eur. Ceram. Soc.* **42**, 3089–3098 (2022).
- Liu, Y. et al. Lattice distortion enhanced hardness in high-entropy borides. *Adv. Funct. Mater.* **35**, 2416992 (2024).
- Kaufmann, K. et al. Discovery of high-entropy ceramics via machine learning. *npj Comput. Mater.* **6**, 42 (2020).
- Mobarak, M. H. et al. Scope of machine learning in materials research—a review. *Appl. Surf. Sci. Adv.* **18**, 100523 (2023).
- Yan, Y. et al. Recent machine learning-driven investigations into high entropy alloys: a comprehensive review. *J. Alloy. Compd.* **1010**, 177823 (2025).
- Wei, J. et al. Machine learning in materials science. *InfoMat* **1**, 338–358 (2019).
- Jaafreh, R., Kang, Y. S., Kim, J. & Hamad, K. Machine learning guided discovery of super-hard high entropy ceramics. *Mater. Lett.* **306**, 130899 (2022).

30. Zhang, J. et al. Design high-entropy carbide ceramics from machine learning. *npj Comput. Mater.* **8**, 5 (2022).
31. Mitra, R., Bajpai, A. & Biswas, K. ADASYN-assisted machine learning for phase prediction of high entropy carbides. *Comput. Mater. Sci.* **223**, 112142 (2023).
32. Meng, H., Yu, R., Tang, Z., Wen, Z. & Chu, Y. Formation ability descriptors for high-entropy carbides established through high-throughput methods and machine learning. *Cell Rep. Phys. Sci.* **4**, 101512 (2023).
33. Zhang, L., Oh, C. & Choi, Y. S. Improved phase prediction of high-entropy alloys assisted by imbalance learning. *Mater. Des.* **246**, 113310 (2024).
34. Han, H., Wang, W. Y. & Mao, B. H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Proc. International Conference on Intelligent Computing*, 878–887 (Springer, 2005).
35. Thabtah, F., Hammoud, S., Kamalov, F. & Gonsalves, A. Data imbalance in classification: Experimental evaluation. *Inf. Sci.* **513**, 429–441 (2020).
36. Bagui, S. & Li, K. Resampling imbalanced data for network intrusion detection datasets. *J. Big Data* **8**, 6 (2021).
37. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
38. Douzas, G., Bacao, F. & Last, F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inf. Sci.* **465**, 1–20 (2018).
39. C. Bunkhumpompat, K. Sinapiromsaran, C. Lursinsap, Safe-level-SMOTE: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem, In: T. Theeramunkong, B. Kijisirikul, N. Cercone, T.-B. Ho (Eds.) *Advances in Knowledge Discovery and Data Mining* 475–482 (Springer Berlin Heidelberg, 2009) .
40. Haibo, H. Yang, H. B. Garcia, E. A. Shutao, L. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: *Proc. IEEE World International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* 1322–1328 (IEEE, 2008).
41. Wang, S., Dai, Y., Shen, J. & Xuan, J. Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Sci. Rep.* **11**, 24039 (2021).
42. Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Prog. Artif. Intell.* **5**, 221–232 (2016).
43. Basha, S.J., Madala, S.R., Vivek, K., Kumar, E.S., Ammannamma, T. A review on imbalanced data classification techniques. In *Proc. International Conference on Advanced Computing Technologies and Applications (ICACTA)*, 2022, 1–6.
44. Zhang, Y. et al. Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models. *Acta Mater.* **185**, 528–539 (2020).
45. Roy, A. & Balasubramanian, G. Predictive descriptors in machine learning and data-enabled explorations of high-entropy alloys. *Comput. Mater. Sci.* **193**, 110381 (2021).
46. Zhang, H. et al. Dramatically enhanced combination of ultimate tensile strength and electric conductivity of alloys via machine learning screening. *Acta Mater.* **200**, 803–810 (2020).
47. Yang, C. et al. A machine learning-based alloy design system to facilitate the rational design of high entropy alloys with enhanced hardness. *Acta Mater.* **222**, 117431 (2022).
48. Araf, I., Idri, A. & Chair, I. Cost-sensitive learning for imbalanced medical data: a review. *Artif. Intell. Rev.* **57**, 80 (2024).
49. Denton, A. R. & Ashcroft, N. W. Vegard's law. *Phys. Rev. A* **43**, 3161–3164 (1991).
50. Meng, H. et al. Formation ability descriptors for high-entropy diborides established through high-throughput experiments and machine learning. *Acta Mater.* **256**, 119132 (2023).
51. Huang, X., Zheng, L., Xu, H. & Fu, H. Predicting and understanding the ductility of BCC high entropy alloys via knowledge-integrated machine learning. *Mater. Des.* **239**, 112797 (2024).
52. Takkala, H. R. et al., Kyphosis disease prediction with help of Randomized Search CV and AdaBoosting. In *Proc. 13th International Conference on Computing Communication and Networking Technologies, ICCCNT 2022*, (IEEE, 2022).
53. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Trans. Evolut. Comput.* **1**, 67–82 (1997).
54. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
55. T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (ACM, 2016).
56. Zhang, M. L. & Zhou, Z. H. ML-KNN: a lazy learning approach to multi-label learning. *Pattern Recognit.* **40**, 2038–2048 (2007).
57. Vasantha, N., Surendran, R., Saravanan, M.S., Madhusundar, N. Prediction of defective products using logistic regression algorithm against linear regression algorithm for better accuracy. In *Proc. International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)* 161–166 (IEEE, 2022).
58. Jordan, M. I. & Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
59. Quinlan, J. R. Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986).
60. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
61. Rezvani, S. & Wang, X. A broad review on class imbalance learning techniques. *Appl. Soft Comput.* **143**, 110415 (2023).

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (52472107), National Key R&D Program of China (2021YFA0715802), Key Research and Development Projects of Shaanxi Province (2024GH-ZDXM-14), National Defense Basic Scientific Research Program of China (JCKY2022607C007, JCKYS20250106), National Science and Technology Major Project (J2022-VI-0011-0042), Sino-German DFG project under the grant number RI 510/79-1 and the financial support from the DFG-funded Research Training Group RTG 2561.

## Author contributions

H.J.L. supervised the project and conceived the idea; X.M.Z. conceptualized, collected data, and wrote the manuscript with help and guidance from J.S. and L.W.W.; Y.Y.Z. and K.F.F. assisted the experiments; Z.X.Z. and Y.J.Z. assisted in ML models building; K.K.W., L.F. and R.R. conducted the writing review and editing.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41524-025-01873-2>.

**Correspondence** and requests for materials should be addressed to Jia Sun or Hejun Li.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026