Designing hexagonal close packed high entropy alloys using machine learning

Modelling Simul. Mater. Sci. Eng. **29** (2021) 085005 (15pp) https://doi.org/10.1088/1361-651X/ac2b37

Designing hexagonal close packed high entropy alloys using machine learning

Bejjipurapu Akhil, Anurag Bajpai, Nilesh P Gurao and Krishanu Biswas*

Department of Materials Science and Engineering, Indian Institute of Technology, Kanpur, Uttar Pradesh-208016, India

E-mail: kbiswas@iitk.ac.in

Received 25 June 2021, revised 19 August 2021
Accepted for publication 29 September 2021
Published 14 December 2021

Abstract

High entropy alloys (HEAs) have drawn significant interest in the materials research community owing to their remarkable physical and mechanical properties. These improved physicochemical properties manifest due to the formation of simple solid solution phases with unique microstructures. Though several pathbreaking HEAs have been reported, the field of alloy design, which has the potential to guide alloy screening, is still an open topic hindering the development of new HEA compositions, particularly ones with hexagonal close packed (hcp) crystal structure. In this work, an attempt has been made to develop an intelligent extra tree (ET) classification model based on the key thermodynamic and structural properties, to predict the phase evolution in HEAs. The results of correlation analysis suggest that all the selected thermodynamic and structural features are viable candidates for the descriptor dataset. Testing accuracy of above 90% along with excellent performance matrices for the ET classifier reveal the robustness of the model. The model can be employed to design novel hcp HEAs and as a valuable tool in the alloy design of HEAs in the future.

Keywords: high entropy alloys, hexagonal close packing, machine learning, extra tree classifier, correlation analysis

S Supplementary material for this article is available online

(Some figures may appear in colour only in the online journal)

*Author to whom any correspondence should be addressed.

1. Introduction

High entropy alloys (HEAs) [1, 2], since the early 21st century, have attracted high research interest in the scientific and engineering community. HEAs span vast compositional space due to multiple principal elements and occupy the central regions of the phase diagrams. This huge compositional space provides us with a plethora of opportunities and flexibility in alloy design for targeted properties. The exceptional properties exhibited by HEAs are attributed to the formation of simple solid solution phases over intermetallics [1, 3–5]. Hence, a robust and efficient platform is needed for accurate phase prediction in HEAs is essential to further their application potential.

In this direction, several empirical and computational design methodologies have been developed for predicting phase evolution in HEAs [3, 6–8]. The parametric approach is the earliest one which proposes considering parameters such as mixing enthalpy (ΔH_{mix}) [8], the difference in the atomic size of constituents (δ) [3], and valence electron concentration (VEC) [9], based on Hume–Rothery rules [10], for phase determination of HEAs. Albeit these parameters provide valuable insights into the phase formation in HEAs, the parametric approaches fail to provide a generalised alloy design space as their applicability to compositions outside the dataset used for formulating their empirical boundaries is unknown. With technological advancements to counter the deficiencies of the parametric approaches, several design strategies such as CALculation of PHAse Diagrams (CALPHAD) [6, 7], *ab initio* molecular dynamics (MD) simulations [11, 12] have also been employed in present times to predict phase formation in HEAs. CALPHAD [6, 7], a semi-empirical computational approach that has proved to be an efficient tool for alloy design. However, utilization of this technique for developing HEAs requires a legitimate thermodynamic database for the whole composition range, which is tedious and very difficult to build.

Due to the enormous number of conceivable HEAs, it is impractical to search for a viable alloy composition by experimenting with each alloy composition. Moreover, with the increase in the number of components in an alloy, it is not easy to construct the alloy phase diagram. Thus, besides the theoretical approach, computational methodologies such as density functional theory (DFT) [13] calculations and *ab initio* molecular dynamics (AIMD) [11, 12] have emerged to examine the phase formation in HEAs. MD simulations [11, 12] have helped circumvent several limitations of the previous approaches, but the limited availability of thermodynamic potentials in the literature restricts the applicability of this approach to the entire HEA composition space. Moreover, this approach requires both high computational power and time. Conversely, DFT [13, 14] calculations do not need experimental input, but they are computationally costly for HEAs. Thus, despite the evolution of such pragmatic alloy design strategies, a gap remains to be filled for a robust, ubiquitous, and high throughput alloy design approach with utmost urgency.

With recent advancements in material informatics, machine learning (ML) [15] has become a popular tool in the material science and engineering field to discover new alloys. ML is a state-of-the-art tool that builds a model to infer the relationships between the targeted property and a set of materials descriptors. ML has inherent advantages, including visualization and simple analysis of high-dimensional data space [16–18]. For the high-throughput screening and composition-property relationship determination of multicomponent alloys, supervised-learning classification algorithms such as kernel-based ones, tree-based ones, support vector machine (SVM), and artificial neural network models are in practical use. Several ML models have been developed to obtain HEA compositions with face centered cubic (fcc) and body centered cubic (bcc) phases [15, 19–25]. However, due to the paucity of data, the design of hexagonal close packed (hcp) compositions using ML is not at par with its counterparts. This

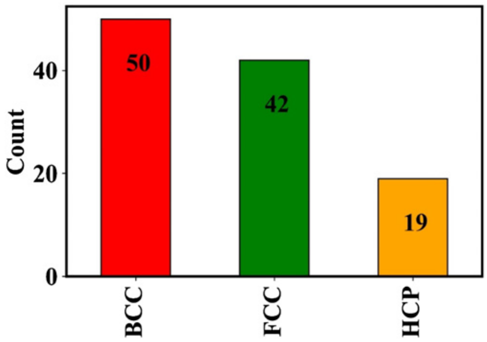

Figure 1. Distribution of input dataset based on the crystal structure.

can be attributed to the structural and property distinction of hcp alloys over fcc and bcc alloys. The hcp HEAs have complex mechanical properties than fcc and bcc HEAs because of the atomic closeness on each crystal plane changes with the axial *c*/*a* ratio. It has been observed that hcp HEAs consisting of transition metals promotes the activation of slip planes other than basal planes ({0 0 0 1} plane) due to change in the *c*/*a* ratio [26]. The change in the *c*/*a* ratio of HEAs is expected because of multiple principal elements. Because of the activation of several deformation systems, the ductility of these hcp HEAs is greatly enhanced. Moreover, multiple deformation twinning modes exist by default in hcp metals which can be exploited to achieve better strain hardening to design a new generation of HEAs. Similarly, a whole range of alloys can be constructed around the hcp single phase alloy like Ti and Zr alloys with alpha, beta, omega phase, and so on by alloying with other elements. Furthermore, rare-earth hcp HEAs have superior magnetic properties because of their small magnetic hysteresis and large refrigerant capacity [27, 28]. Attributable to all these entrancing properties and prospective applications, an effort is warranted to make up for the shortcoming left in the alloy design of hcp single-phase solid solutions.

With this background, the present work involves the development of an intelligent extra tree (ET) classification ML model to filter out HEA compositions on the basis of their phase evolution, with emphasis on hcp alloys. The study involves analyzing key thermodynamic and structural attributes that impact the evolution of phases in multicomponent alloys. The results indicate that the electronic characteristics of the constituents play a defining role in designing hcp HEAs. Moreover, the predicted compositions have several hcp forming alloys consisting of transition metals. Such compositional space has not been extensively explored to date and can bring new insights to the design and development of hcp HEAs.

2. Methodology

2.1. Designing the dataset

The dataset was built from the available literature on multi-principal element alloys (MPEAs). The input dataset consists of 111 unique MPEA compositions, out of which 50 are bcc, 42 are fcc, and the remaining 19 hcp single-phase solid solutions (see Table T1 in supporting information (https://stacks.iop.org/MSMS/29/085005/mmedia)) [29–31]. The chosen compositions have been illustrated in Figure 1.

The figure shows that the input dataset contains an uneven number of compositions in each class, known as an imbalanced dataset. Such a dataset poses the problem of downgrading the performance of the ML classifier. This manifests into the classifier showing a bias toward the majority class. The oversampling of minority class (hcp here) is done using the synthetic-minority oversampling technique (SMOTE) to overcome this impediment. SMOTE generates new samples of the minority class using the input dataset [32, 33]. The Euclidean distance between each sample (X) and other samples of the minority class is calculated to obtain its k-nearest neighbors. N samples (X_1, X_2, X_3, …, X_N) were selected according to the imbalanced proportion in the input dataset from these k-nearest neighbors. The following equation calculates the new synthetic samples:

$$X' = X + \mathrm{rand}(0, 1)^{*}|X - X_k|, \tag{1}$$

where rand (0, 1) represents a random number between 0 and 1 and k ranges from 1 to N.

After the oversampling process, the dataset is reconstructed and makes the counts of all classes equal. Now the input dataset is a balanced one. SMOTE analysis increases the data size and makes the decision region of the minority class more general. The reconstructed data contains a total of 150 samples, with each class having 150 data points.

2.2. Feature calculations

It is generally believed that phase selection in alloys is related to the fundamental atomic and thermodynamic properties of the alloy constituents. These include enthalpy of mixing (ΔH_{mix}) [4], configurational entropy (ΔS_{mix}) [3], atomic size difference (δ) [3], VEC [9] and electronegativity (χ) [34]. The parameters were considered as preliminary input features, and their numeric values were calculated using the following relations;

$$\delta = \sqrt{\sum_{i=1}^{n} c_i\left(1 - \left(\frac{r_i}{\bar{r}}\right)\right)^2} \times 100 \tag{2}$$

$$\Delta H_{\text{mix}} = \sum_{i=1, i<j}^{n} 4H_{ij}c_i c_j \tag{3}$$

$$\Delta S_{\text{mix}} = -R\sum_{i=1}^{n} c_i \ln c_i \tag{4}$$

$$\Delta\chi = \sqrt{\sum_{i=1}^{n} c_i(\chi_i - \overline{\chi})^2} \tag{5}$$

$$\mathrm{VEC} = \sum_{i=1}^{n} c_i \mathrm{VEC}_i, \tag{6}$$

where r_i and χ_i denote the atomic radius and Pauling electronegativity of the ith element, respectively. c_i refers to the atomic fraction of the ith element, VEC_i is its valence electron concentration, and n is the total number of constituents in an MPEA. H_{ij} is the enthalpy of binary atomic pairs of the ith and jth elements calculated through the Miedema method [35].

The average values of the parameters are computed as $\overline{X} = \sum_{i=1}^{n} c_i X_i$. The values of all features were normalised using the following relation:

$$X_{\text{new}} = \frac{X_i - X_{\min,i}}{X_{\max,i} - X_{\min,i}}. \tag{7}$$

X_{new} is the normalized feature value that lies in the range of 0 to 1, X_i ($i = 1, 2, 3, \ldots, 5$) is the original feature value, $X_{\max,i}$ and $X_{\min,i}$ maximum and minimum values of the respective feature. Post normalization, all input features become dimensionless, and their gradient descents converge quickly.

In order to understand the correlation between each pair of the input features, x and y, the Pearson correlation coefficient (P) [36] was calculated using,

$$P = \frac{1}{n-1}\left(\frac{\sum_{1=1}^{n}(x-\overline{x})(y-\overline{y})}{\sigma_x \sigma_y}\right), \tag{8}$$

where $\overline{x}$ and $\overline{y}$ are the mean values of the two features and σ_x and σ_y are the corresponding standard deviations of x and y. P lies between -1 and 1. $P = -1$ indicates that two input features have a perfectly negative correlation, whereas $P = 1$ indicates that two features have a perfectly positive linear correlation [36]. The mutual correlation features between $-0.7 < P < 0.7$ are considered powerful features for selection to descriptor space.

2.3. ML model (Extra Tree classifier)

The ET classifier, first proposed by Geurts *et al* [37], is an ensemble learning algorithm fundamentally based on decision trees (DT). ET classifier is introduced as an extension of the random forest algorithm to overcome the overfitting of data. It can be used for both regression and classification tasks. The basic idea involves combining the decisions of several models and making a decision centered around that combination, thereby improving the overall performance. The DT predicated ensemble algorithm can give better performance for independent base learners, which can be achieved through randomization. While growing the trees, a better tree diversity is needed for randomization, which facilitates truncating the correlation [38]. A robust and stable classifier (model) in supervised ML tasks with accurate predictions can be achieved utilizing an ensemble method to reduce the factors, i.e. variance, noise, and bias. Nevertheless, the main drawback that an ensemble learner faces is the imminent rise in computational time due to multiple individual classifiers training [38]. As a result, the ET algorithm has been highlighted, which works virtually akin to, yet more expeditious than the random forest algorithm.

The ET classifier comprises several DTs, each tree having a root node, child/split nodes, and leaf nodes, as illustrated in Figure 2.

The ET classifier picks up a split rule for a given dataset X subject to a random features' subset and a partially arbitrary cutpoint at the root node. The same process is reiterated until a leaf node is reached for every child node. Moreover, the most significant ET classifier parameters are the number of DTs employed in the ensemble (k), the number of randomly selected attributes/features (f), and the minimum number of samples/instances needed for splitting a node. In contrast to the random forest, the ET classifier uses an entire sample for training at each step instead of training data repeatedly with replacement. Since the ET classifier uses an entire sample to learn the individual DT, the bias is minimized. ET classifier diminishes the model's variance by employing more vigorous randomization techniques by randomly picking up decision boundaries rather than picking the best possible ones. This essentially allows the

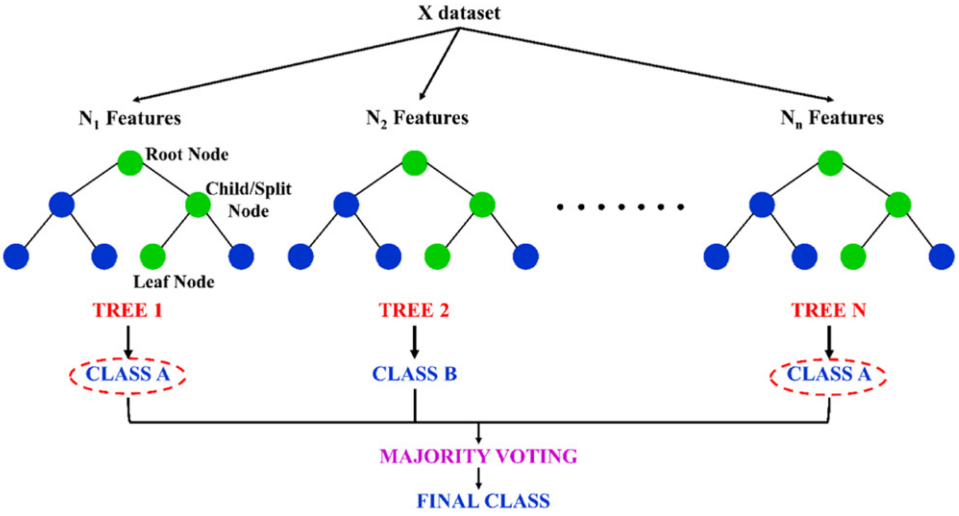

Figure 2. Illustration of the extra (extremely randomized) trees algorithm (an ensemble of DTs).

ET classifier to train faster than random forest. ET classifier outputs its classification result as an aggregation of multiple de-correlated DTs collected in a forest to output.

2.4. Evaluation of performance of the model

Since the present study involves a multi-class classification, considering the accuracy metric as a standalone metric is not suitable for judging the model performance. Accuracy can be the most promising metric only when there are equal false positives (FP) and false negatives (FN). Therefore, other evaluation metrics, such as precision, recall, and F1-score, are considered. The generally used terms in ML model evaluation are true positives (TP), which are the data points that are actually positive and are expressed as positive; true negatives (TN), which are the data points that are actually negative and are expressed as negative; FP, which are the data points that are actually negative but are expressed as positive; and FN, which are the data points that are actually positive but are expressed as negative.

The accuracy is the estimation of the correctly classified data points and is expressed as $\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}}$. Precision is the fraction of relevant data points obtained from the total dataset and is expressed as $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$. Recall corresponds to the fraction of relevant data points for a given class retrieved from the total set of relevant datapoints and is expressed as $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$. F1 Score is the harmonic mean of precision and recall, expressed as $\frac{2(\mathrm{Precision}\times\mathrm{Recall})}{\mathrm{Precision}+\mathrm{Recall}}$.

3. Results and discussion

3.1. Data analysis

Prior to training and testing any ML algorithm, it is imperative to extract the statistical information from the input dataset to design a new balanced dataset as well as a non-redundant descriptor space [39, 40]. A scatter matrix is plotted to understand the qualitative description of input features, as shown in Figure 3.

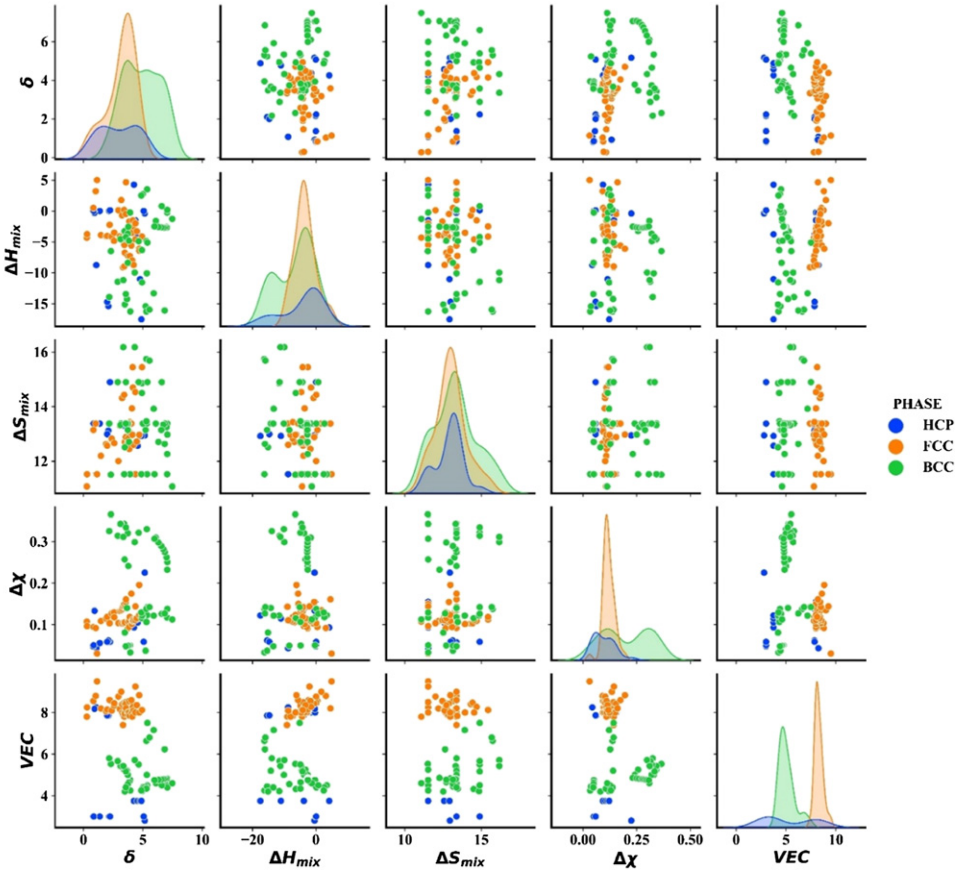

Figure 3. 5 × 5 Scatter matrix plot gives us the qualitative description of all our input descriptors. Diagonal subplots show bcc, fcc, and hcp phases distribution if only one feature is used. Off-diagonal subplots show the distribution of three phases between a pair of input features.

The overlapping nature of diagonal subplots in the scatter matrix indicate that identifying phases by employing a single empirical parameter is not adequate and desires to contemplate all parameter for an ML algorithm. In order to quantify the selection of crucial features, a Pearson correlation coefficient matrix is calculated and presented in Figure 4(a). It can be inferred that none of the considered features for the present dataset have values close to either 1 or −1, implying that none of them is correlated either positively or negatively. Moreover, a radar plot for the selected features considering the hcp compositions from the input dataset is illustrated in Figure 4(b). The radar plot gives the parameter variation in a two-dimensional space. Based on qualitative and quantitative descriptions, all five features were used as the descriptors for the ML model.

3.2. Model performance evaluation

The whole dataset was initially divided into training and testing datasets in the ratio of 0.65 to 0.35. Since the input dataset needs to be of smaller size to avoid overfitting, the ML model (ET classifier) accuracy is evaluated via repeated stratified k (= 5)-fold cross-validation, in which

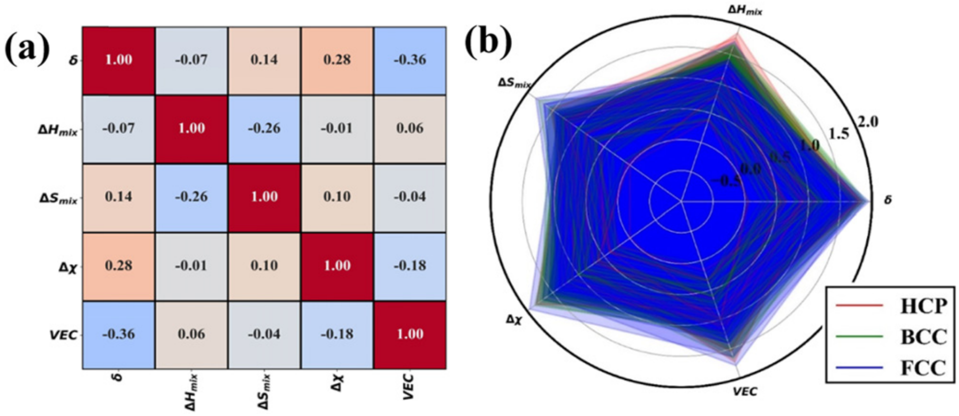

Figure 4. (a) Pearson correlation matrix for all the selected features; (b) Radar plot for the selected features considering the input dataset.

Table 1. Search space for all parameters and the best performing parameters for the ET classifier.

Model	Parameters	Search range	Best parameters
ET classifier	Max depth	[15, 20, 25, 30, 35]	30
	Number of estimators	[100, 125, 150, 200]	125
	Criterion	[‘Gini’, ‘entropy’]	Gini

the whole dataset is randomly divided into five smaller sets (folds). The model is trained on the four folds for each of the five folds and validated on the remaining one-fold. The entire procedure is repeated for different random splits three times, so fifteen different sets are used to estimate the model’s efficiency [39]. Grid search is used to search the best parameters (criterion, maximum depth, and a number of trees) of the present study’s ET classifiers. Grid search builds the model for possible combinations of each parameter and gives the best parameter. The searching space for each of the parameters and the best parameters for the ET classifier is presented in Table 1.

An average cross-validation accuracy of 91.6 ± 4.54% was achieved for the best parameter. To analyze how well the ET classifier can predict each class in the dataset, a confusion matrix is constructed on testing data that contains 53 samples [41] and is shown in Figure 5(a). It is clear from the confusion matrix that the ET classifier shows higher precision and recall values for all three classes, indicating an excellent validity of the model. Moreover, the accuracy value of the testing dataset is 92.45% which is an acceptable one. The difference in cross-validation accuracy and confusion matrix accuracy values is because cross-validation accuracy is the average of all validation accuracies, and the confusion matrix accuracy is reported directly instead. To assess the model extensively, alternative evaluation metrics such as receiver operator characteristic (ROC) curve are studied. Figure 5(b) shows the ROC curve. The ROC curve leads to another evaluation metric, the area under the curve (AUC), which calculates the relationship between FP and TP. AUC gives an idea of how well the model can distinguish the classes. The higher AUC for all classes in the present work indicates the model’s good predictability for the given dataset. In general, ROC curves will be curvy, but there are straight due to fewer data points in the testing dataset. Figure 5(c) shows the learning curve of the model. A similar

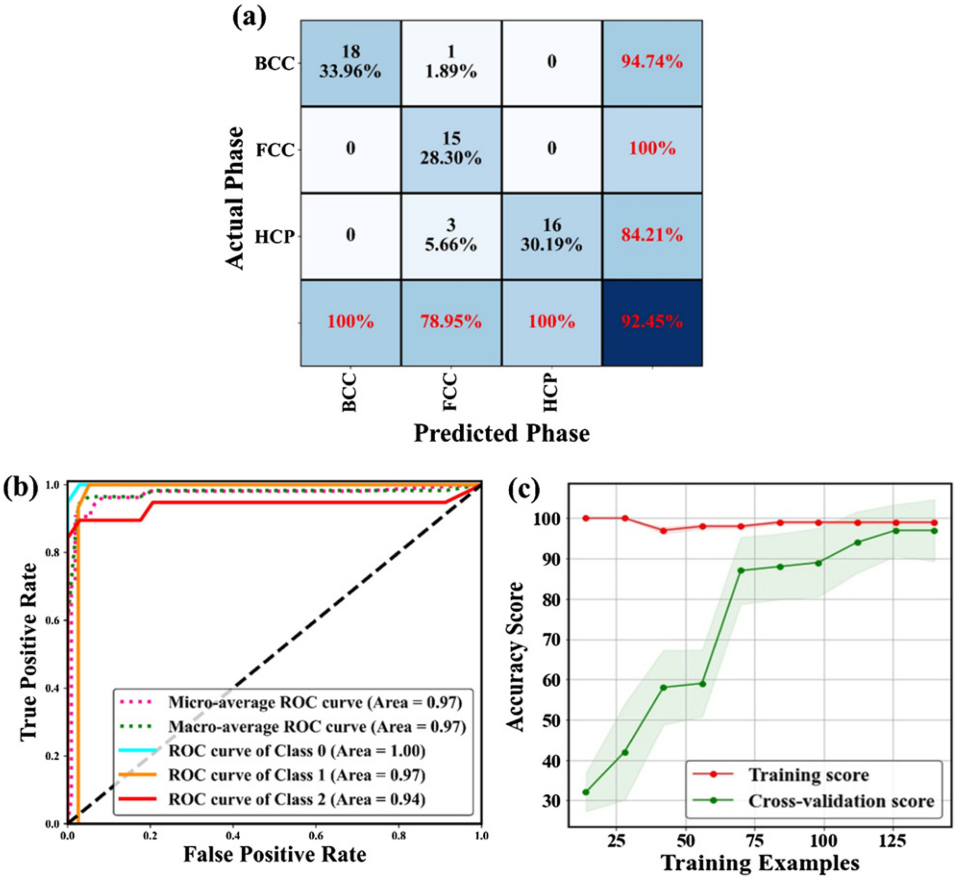

Figure 5. Evaluation metrics of ET classifier (a) confusion matrix (for testing data set); (b) receiver operating characteristic (ROC) curve (classes 0, 1, 2 correspond to bcc, fcc and hcp phases, respectively); (c) learning curve for ET classifier.

cross-validation process explained earlier has been applied to calculate the models' training and cross-validation accuracy scores. Learning curves explain how the model's accuracy changes with the increase in the dataset's size and provide insight into the model's variance and bias. The gap between the training and cross-validation scores is initially high, inferring that of higher variance. As the number of samples increases, the cross-validation score increases, and the training score has initially decreased, and then the curve is almost flattening. With the increase in the dataset size, there is a decrease in the gap between the two curves, indicating low variance, i.e. the model is not overfitted.

To compare the performance of the ET classifier with other models, the input data is trained using various models such as logistic regression (Log), Naïve Bayes (NB) classifier, random forest (RF), decision tree (D-tree), support vector machine (SVM), K-nearest neighbours (KNN), gradient boost classifier (GBC). All these models were compared with the ET classifier with both the default parameters and the tuned parameters. Tuned parameters were searched using the grid search method. A bar plot is shown in Figure 6(a), comparing the ET classifier with other ML algorithms on the basis of the accuracy of models. The ET classifier gives the best accuracy among all the ML models with the default parameters. It is observed that even

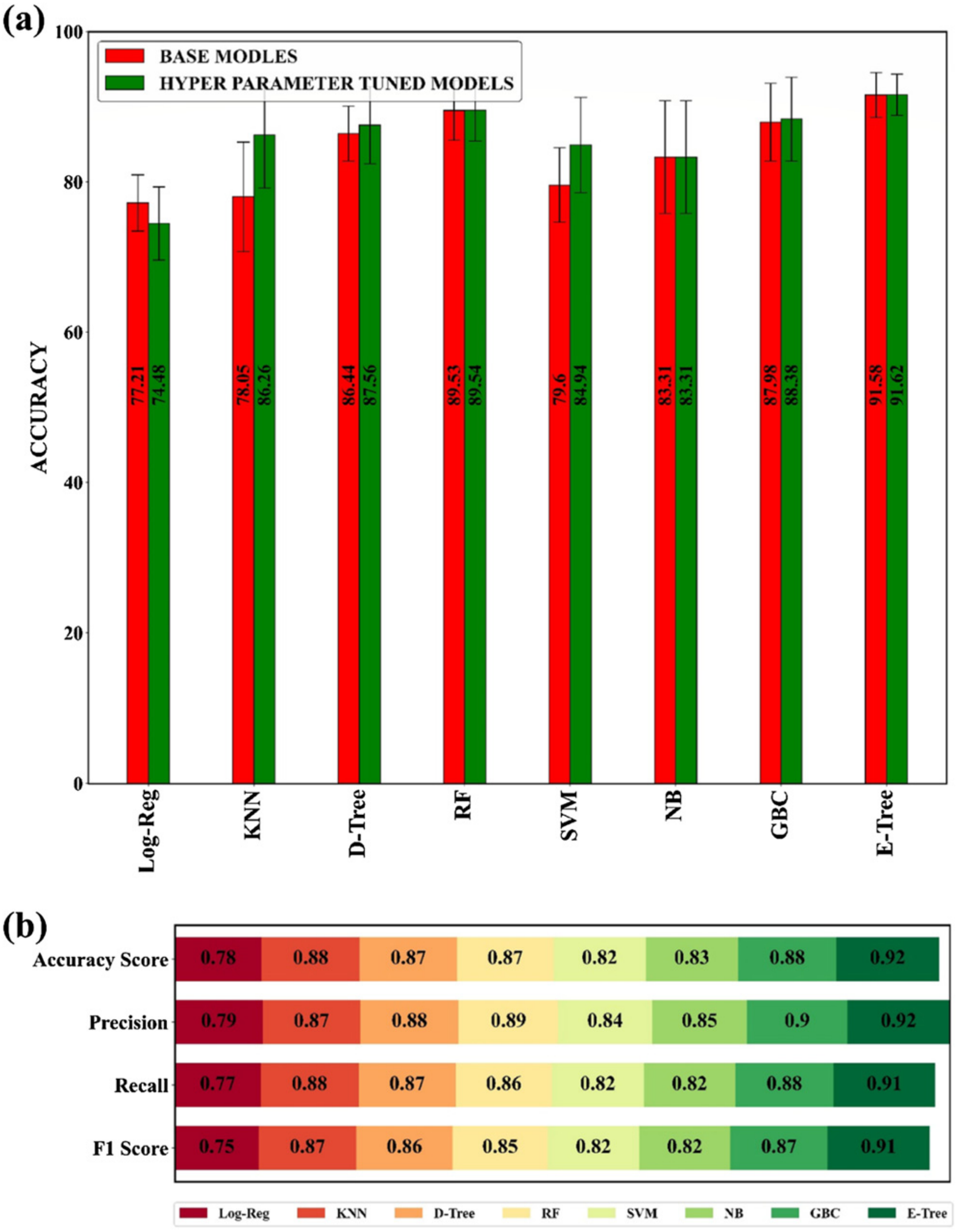

Figure 6. (a) Accuracies of different ML models; (b) evaluation metrics of different ML models. Log-reg is logistic regression, KNN is K-nearest neighbours, D-tree is decision tree, RF is random forest, SVM is support vector machine, NB Naïve Bayes classifier, GBC is gradient boost classifier, and E-tree is ET classifier.

after tuning the hyperparameters ET classifier is still the best model for the present classification problem. After tuning the hyperparameters, it is found that the accuracy of models such as KNN and SVM has notably increased.

Through Figure 6(b), a comparison of all evaluation metrics accuracy, F1 score, precision, and recall has been made. It is evident that the ET classifier has outperformed all other trained models in all metrics. The high F1-score of the ET classifier shows a right balance between the precision (a measure of a classifier's exactness) and recall (classifier's completeness) values. Therefore, a conclusion can be drawn that the ET classifier model can classify all three classes efficiently and act as a robust platform to predict new compositions.

3.3. Prediction of novel hcp HEAs

A new dataset of equiatomic quinary compositions is designed using a compositional space containing 33 elements (see Table T2 in supporting information) which have either an hcp structure at ambient temperature or are expected to form an hcp structure at high temperatures and/or high pressures were considered. This compositional space consists of both rare-earth and transition elements to design hcp HEAs. From the possible ($^{33}C_5 = 278\,256$), only those compositions which are having the difference in atomic radius (δ) $< 6.6\%$ and enthalpy of mixing ling between $-15 < \Delta H_{\text{mix}} < 5$ kJ mol^{-1} were considered for prediction purposes. This exercise was done to parametrically sort out the compositions forming solid solutions [3, 8]. It is found that only 5758 compositions follow these rules in unison. These were considered as the prediction input dataset, which was employed on the ET classifier model to predict the crystal structure. Out of 5758 compositions, 2391 compositions have an hcp crystal structure, 1604 have bcc crystal structure, and 1763 have fcc crystal structure. It is worthwhile to mention that although the literature on hcp alloys is scarce with transition metals as constituents, the presently developed compositional database contained several hcp alloy systems made up of transition metals too. Such a discovery can bring new insights to the alloy design of multicomponent hcp alloys as well as expand their existing compositional space.

3.4. Analysis of the feature space

Post the validation of the model, a '*feature importance plot*' was drawn to understand the influence of each feature on the model output and is shown in Figure 7. The overall significance of the input features is in the order of VEC $> \Delta\chi > \delta > \Delta H_{\text{mix}} > \Delta S_{\text{mix}}$. The feature importance trend is following the previously reported literature. However, it should be mentioned that the order may change as per the input dataset. VEC and $\Delta\chi$ represent the electronic properties of the constituent elements. VEC plays a vital role in the crystal structure classification since its calculation includes valence shell electrons that can alter the atomic size of constituent atoms and thereby control the net entropy of any alloy system. This can be understood by the postulation that the atomic radius of a constituent in an alloy depends on the electronic interaction between the outermost electron(s) or VEC and the nuclear charge(s). A larger VEC indicates more nuclear charge for atoms with the same number of electron shells, bringing a stronger binding force which constricts the outermost shell electrons in a smaller orbit and decreases the atomic radius [42]. An opposite phenomenon occurs for elements with a smaller VEC value. δ parameter, which represents the lattice strains due to the size difference between the constituent elements, appears to play a lesser significant role in the crystal structure evaluation using the presently trained ET classifier. However, this may change upon adding more compositions to the input dataset by discovering more hcp alloys. The feature importance analysis reveals that electronic properties are the most important features for the crystal structure classification than the thermodynamic properties for the dataset used in the present study.

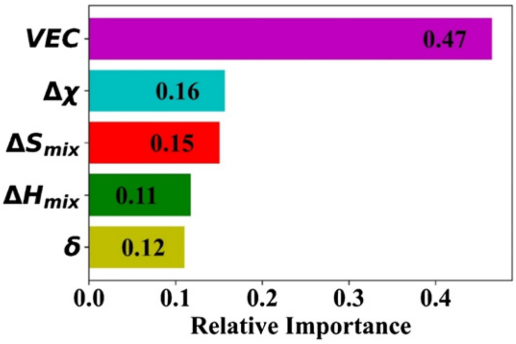

Figure 7. Relative Importance for the selected features in evaluating the model metrics.

Table 2. Phase prediction for several equiatomic HEAs using ET classifier and Thermo-Calc® database.

	Predicted Phase	
Alloy	ET Classifier	Thermo-Calc®
CoFeNiTiAl	BCC	BCC
CoFeNiTiCr	BCC	BCC
CoFeNiTiCu	FCC	FCC
CrCuFeNiZr	BCC	BCC
FeAlNiCu	HCP	BCC
ReAlNiCu	HCP	HCP
RuReFeAl	HCP	HCP
RuReFeIr	HCP	HCP
TiZrHfAl	HCP	HCP
ZnReAlNi	FCC	HCP
ZnReFeAl	HCP	HCP
ZnReIrAl	HCP	HCP

3.5. *Validity of the ET classifier*

In order to substantiate the prediction capability of the present ET classifier, Thermo-Calc® [43] software was used to validate the prediction outcomes of the model. Several quaternary and quinary equiatomic HEAs were designed based on the elements present in the TC-HEA3 [43] database of the Thermo-Calc® software. The designed compositions consist of both quaternary and quinary ones, covering the domain of both media as well as HEAs. These compositions were subsequently employed on the ET classifier for ML based phase prediction. 12 HEA compositions were randomly selected from this dataset for ET model validation. After that, the phase evolution in the selected HEAs was evaluated using Thermo-Calc® software. The Thermo-Calc® calculations of the prospective phases were made from 300 to 3000 K temperature and 1 atm pressure. The crystal structures for the Thermo-Calc® evaluated alloys shown in Table 2 were taken at ambient conditions (300 K, 1 atm). The ML predictions were calculated at the same conditions as that of the Thermo-Calc® calculations. Table 2 summarizes the predictive capability of the ET classifier for these alloys.

The phase diagrams developed via Thermo-Calc®, validating the ET classifier results, have been illustrated in Figure S1 (in supporting information). It is found that the ET classifier can accurately predict the phase for 10 of the 12 studied HEAs (inaccurate predictions are marked in red in Table 2) when evaluated in comparison to Thermo-Calc® predictions. Based on the above discussion, the presently developed supervised ET classification algorithm has the potential to accelerate the discovery of new HEA systems, particularly hcp HEAs, and expand the HEA compositional space.

4. Conclusion

In conclusion, an ET classifier, a supervised ensemble learning algorithm, is developed to predict the crystal structure of HEAs using a dataset of 111 multicomponent alloys. Several key thermodynamic and structural features were considered while designing the feature space. The selected input features showed independent nature to each other, and hence all were used as descriptors for the ML model. The model's performance was evaluated *via* different metrics, using k (=5)-fold repeated stratified cross-validation. The ET classifier achieved an average cross-validation accuracy of 91.6 ± 4.54% after hyperparameter tuning using grid search cross-validation. The learning curves indicate that the model has no overfitting. Moreover, the validity of the ET classifier was evaluated using the Thermo-calc® database, and the outcomes show good convergence. These results demonstrate that the developed ET classification model is quite efficient in classifying crystal structures and predicting new hcp compositions from the limited input dataset without prior experimentation.

Acknowledgments

The authors acknowledge the support offered by the Computer Centre, Indian Institute of Technology, Kanpur, by providing MATLAB 2020b software to carry out this study.

Author contributions

BA conceptualized the idea, performed calculations, developed the ML model, analyzed the results. BA and AB wrote the manuscript. NPG and KB supervised and reviewed the work.

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

ORCID iDs

Bejjipurapu Akhil https://orcid.org/0000-0002-6195-3230
Anurag Bajpai https://orcid.org/0000-0003-0456-1641
Nilesh P Gurao https://orcid.org/0000-0002-4502-2436
Krishanu Biswas https://orcid.org/0000-0001-5382-9195

References

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

[2] Sharma A, Yadav S, Biswas K and Basu B 2018 High-entropy alloys and metallic nanocomposites: processing challenges, microstructure development and property enhancement *Mater. Sci. Eng.* R 131

[3] Guo S and Liu C T 2011 Phase stability in high entropy alloys: formation of solid-solution phase or amorphous phase *Prog. Nat. Sci.: Mater. Int.* **21** 433–46

[4] Yang X and Zhang Y 2012 Prediction of high-entropy stabilized solid-solution in multicomponent alloys *Mater. Chem. Phys.* **132** 233–8

[5] Zhou Z, Zhou Y, He Q, Ding Z, Li F and Yang Y 2019 Machine learning guided appraisal and exploration of phase design for high entropy alloys *npj Comput. Mater.* **5** 128

[6] Tazuddin , Gurao N P and Biswas K 2017 In the quest of single phase multi-component multiprincipal high entropy alloys *J. Alloys Compd.* **697** 434–42

[7] Raturi A, Aditya C J, Gurao N P and Biswas K 2019 ICME approach to explore equiatomic and non-equiatomic single phase bcc refractory high entropy alloys *J. Alloys Compd.* **806** 587–95

[8] Zhang Y, Zhou Y J, Lin J P, Chen G L and Liaw P K 2008 Solid-solution phase formation rules for multi-component alloys *Adv. Eng. Mater.* **10** 534–8

[9] Zhang Y, Lu Z P, Ma S G, Liaw P K, Tang Z, Cheng Y Q and Gao M C 2014 Guidelines in predicting phase formation of high-entropy alloys *MRS Commun.* **4** 57–62

[10] Mizutani U 2012 Hume–Rothery rules for structurally complex alloy phases *MRS Bull.* **37** 169

[11] Li Z, Körmann F, Grabowski B, Neugebauer J and Raabe D 2017 *Ab initio* assisted design of quinary dual-phase high-entropy alloys with transformation-induced plasticity *Acta Mater.* **136** 262–70

[12] Konrad C, Kamil C, Piotr A, Paweł D, Tomasz K, Piotr B and Krzysztof M 2020 Assessment of utilization of *ab initio* and Calphad calculations for a design of high-entropy alloy for metal forming *Procedia Manuf.* **50** 677–83

[13] Ikeda Y, Grabowski B and Körmann F 2019 *Ab initio* phase stabilities and mechanical properties of multicomponent alloys: a comprehensive review for high entropy alloys and compositionally complex alloys *Mater. Charact.* **147** 464–511

[14] Sorkin V, Tan T L, Yu Z G and Zhang Y W 2021 High-throughput calculations based on the small set of ordered structures method for non-equimolar high entropy alloys *Comput. Mater. Sci.* **188** 110213

[15] Kaufmann K and Vecchio K S 2020 Searching for high entropy alloys: a machine learning approach *Acta Mater.* **198** 178–222

[16] Liu Z-K 2018 Ocean of data: integrating first-principles calculations and CALPHAD modeling with machine learning *J. Phase Equilib. Diffus.* **39** 635–49

[17] Butler K T, Davies D W, Cartwright H, Isayev O and Walsh A 2018 Machine learning for molecular and materials science *Nature* **559** 547–55

[18] Rahnama A, Clark S and Sridhar S 2018 Machine learning for predicting occurrence of interphase precipitation in HSLA steels *Comput. Mater. Sci.* **154** 169–77

[19] Krishna Y V, Jaiswal U K and Rahul M R 2021 Machine learning approach to predict new multiphase high entropy alloys *Scr. Mater.* **197** 113804

[20] Machaka R 2021 Machine learning-based prediction of phases in high-entropy alloys *Comput. Mater. Sci.* **188** 110244

[21] Lee S Y, Byeon S, Kim H S, Jin H and Lee S 2021 Deep learning-based phase prediction of high-entropy alloys: optimization, generation, and explanation *Mater. Des.* **197** 109260

[22] Zhang L, Chen H, Tao X, Cai H, Liu J, Ouyang Y, Peng Q and Du Y 2020 Machine learning reveals the importance of the formation enthalpy and atom-size difference in forming phases of high entropy alloys *Mater. Des.* **193** 108835

[23] Li R, Xie L, Wang W Y, Liaw P K and Zhang Y 2020 High-throughput calculations for high-entropy alloys: a brief review *Front. Mater.* **7** 290

[24] Pei Z, Yin J, Hawk J A, Alman D E and Gao M C 2020 Machine-learning informed prediction of high-entropy solid solution formation: beyond the Hume–Rothery rules *npj Comput. Mater.* **6** 50

[25] Risal S, Zhu W, Guillen P and Sun L 2021 Improving phase prediction accuracy for high entropy alloys with Machine learning *Comput. Mater. Sci.* **192** 110389

[26] Gao M C, Zhang B, Guo S M, Qiao J W and Hawk J A 2016 High-entropy alloys in hexagonal close-packed structure *Metall. Mater. Trans.* A **47** 3322–32

[27] Takeuchi A, Wada T and Kato H 2019 High-entropy alloys with hexagonal close-packed structure in $Ir_{26}Mo_{20}Rh_{22.5}Ru_{20}W_{11.5}$ and $Ir_{25.5}Mo_{20}Rh_{20}Ru_{25}W_{9.5}$ alloys designed by sandwich strategy for the valence electron concentration of constituent elements in the periodic chart *Mater. Trans.* **60** 1666–73

[28] Takeuchi A, Wada T and Kato H 2019 Solid solutions with bcc, hcp, and fcc structures formed in a composition line in multicomponent Ir–Rh–Ru–W–Mo system *Mater. Trans.* **60** 2267–76

[29] Li R-X, Qiao J-W, Liaw P K and Zhang Y 2020 Preternatural hexagonal high-entropy alloys: a review *Acta Metall. Sin. (Engl. Lett.)* **33** 1033–45

[30] Ye Y F, Wang Q, Lu J, Liu C T and Yang Y 2016 High-entropy alloy: challenges and prospects *Mater. Today* **19** 349–62

[31] Gao M C, Zhang C, Gao P, Zhang F, Ouyang L Z, Widom M and Hawk J A 2017 Thermodynamics of concentrated solid solution alloys *Curr. Opin. Solid State Mater. Sci.* **21** 238–51

[32] Himanen L, Geurts A, Foster A S and Rinke P 2019 Data-driven materials science: status, challenges, and perspectives *Adv. Sci.* **6** 1900808

[33] Reunanen J 2003 Overfitting in making comparisons between variable selection methods *J. Mach. Learn. Res.* **3** 1371–82

[34] Ji X 2015 Relative effect of electronegativity on formation of high entropy alloys *Int. J. Cast Met. Res.* **28** 229–33

[35] Zhang L, Wang R, Tao X, Guo H, Chen H and Ouyang Y 2015 Formation enthalpies of Al–Fe–Zr–Nd system calculated by using geometric and Miedema's models *Physica* B **463** 82–7

[36] Lee Rodgers J and Nicewander W A 1988 Thirteen ways to look at the correlation coefficient *Am. Stat.* **42** 59–66

[37] Geurts P, Ernst D and Wehenkel L 2006 Extremely randomized trees *Mach. Learn.* **63** 3–42

[38] Choudhury A, Konnur T, Chattopadhyay P P and Pal S 2019 Structure prediction of multi-principal element alloys using ensemble learning *Eng. Comput.* **37** 1003–22

[39] Huang W, Martin P and Zhuang H L 2019 Machine-learning phase prediction of high-entropy alloys *Acta Mater.* **169** 225–36

[40] Islam N, Huang W and Zhuang H L 2018 Machine learning for phase selection in multi-principal element alloys *Comput. Mater. Sci.* **150** 230–5

[41] Liu X, Li X, He Q, Liang D, Zhou Z, Ma J, Yang Y and Shen J 2020 Machine learning-based glass formation prediction in multicomponent alloys *Acta Mater.* **201** 182–90

[42] Hu Q, Guo S, Wang J M, Yan Y H, Chen S S, Lu D P, Liu K M, Zou J Z and Zeng X R 2017 Parametric study of amorphous high-entropy alloys formation from two new perspectives: atomic radius modification and crystalline structure of alloying elements *Sci. Rep.* **7** 39917

[43] Thermo Calc 2017 TCHEA3: TCS High Entropy Alloy Database https://www.engineering-eye.com/THERMOCALC/details/db/index.html