scientific reports

Check for updates

OPEN

Phase prediction and experimental realisation of a new high entropy alloy using machine learning

Swati Singh[1], Nirmal Kumar Katiyar[2], Saurav Goel[1,2,3]✉ & Shrikrishna N. Joshi[1]✉

Nearly ~ 10^8 types of High entropy alloys (HEAs) can be developed from about 64 elements in the periodic table. A major challenge for materials scientists and metallurgists at this stage is to predict their crystal structure and, therefore, their mechanical properties to reduce experimental efforts, which are energy and time intensive. Through this paper, we show that it is possible to use machine learning (ML) in this arena for phase prediction to develop novel HEAs. We tested five robust algorithms namely, K-nearest neighbours (KNN), support vector machine (SVM), decision tree classifier (DTC), random forest classifier (RFC) and XGBoost (XGB) in their vanilla form (base models) on a large dataset screened specifically from experimental data concerning HEA fabrication using melting and casting manufacturing methods. This was necessary to avoid the discrepancy inherent with comparing HEAs obtained from different synthesis routes as it causes spurious effects while treating an imbalanced data—an erroneous practice we observed in the reported literature. We found that (i) RFC model predictions were more reliable in contrast to other models and (ii) the synthetic data augmentation is not a neat practice in materials science specially to develop HEAs, where it cannot assure phase information reliably. To substantiate our claim, we compared the vanilla RFC (V-RFC) model for original data (1200 datasets) with SMOTE-Tomek links augmented RFC (ST-RFC) model for the new datasets (1200 original + 192 generated = 1392 datasets). We found that although the ST-RFC model showed a higher average test accuracy of 92%, no significant breakthroughs were observed, when testing the number of correct and incorrect predictions using confusion matrix and ROC-AUC scores for individual phases. Based on our RFC model, we report the development of a new HEA ($Ni_{25}Cu_{18.75}Fe_{25}Co_{25}Al_{6.25}$) exhibiting an FCC phase proving the robustness of our predictions.

Abbreviations

AM	Amorphous phase
AUC	Area under the ROC curve
BCC	Body centered cubic
c_i	Concentration of the *i*th element
DTC	Decision tree classifier
FCC	Face centered cubic
FCC + BCC	Mixed FCC and BCC solid solutions
FP	False positives
FN	False negatives
FPR	False positive rate
HCP	Hexagonal close packed
IM	Intermetallic phase
KNN	K-nearest neighbours
n	Total number of metallic elements in a high entropy alloy
RFC	Random forest classifier
ROC	Receiver operating characteristic
r_i	Atomic radius of the *i*th element,
$\bar{r}$	Average atomic radius
SVM	Support vector machine

[1]Department of Mechanical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India. [2]School of Engineering, London South Bank University, 103 Borough Road, London SE1 0AA, UK. [3]University of Petroleum and Energy Studies, Dehradun 248007, India. ✉email: GoeLs@Lsbu.ac.uk; snj@iitg.ac.in

SS	Solid-solution
MIP	Mixture of intermetallic phases.
TP	True positives
TN	True negatives
TPR	True positive rate
VEC	Valence electron concentration
x_i	Pauling electronegativity
$\bar{x}$	Averaged Pauling electronegativity
XGB	XGBoost/extreme gradient boosting
$\Delta\chi$	Electronegativity difference
δ	Atomic size difference
ΔH_{mix}	Mixing enthalpy
ΔS_{mix}	Mixing entropy

To overcome twenty-first century grand engineering challenges, the investigation of unexplored central region of the ternary phase diagram is indispensable, which occupies the complex multi-component alloys or popularly known high-entropy alloys (HEAs)[1,2]. HEAs have ample compositional space and possess exceptional properties such as excellent mechanical performance at high temperatures, exceptional ductility, high fracture toughness at cryogenic temperatures, bio-compatibility, high conductivity, excellent catalytic and magnetic properties which means that one or more than one HEA can potentially offer a solution for most engineering problems concerning materials[3–5].

The combination theory suggests a large compositional space (nearly ~ 10^8 types of HEAs can be developed from about 64 elements in the periodic table) in the central region of the ternary phase diagram[6,7]. However, it is only since 2004 when HEAs were first discovered, and since then, the profound study is compelling to accelerate the pace of discovery of novel HEAs. The search for new HEAs is strenuous, because each element, its weight percentage, various synthesis routes (vacuum arc melting, powder metallurgy, selective laser melting, additive manufacturing and others) and their processing parameters (cooling rate, processing time, temperature, vacuum/gas) can affect the phase in which a high-entropy alloy stabilises[8,9]. The enormity of composition-processing-structure-performance space makes the searches based on the traditional trial-and-error approach extremely difficult and time-consuming.

Traditionally, new high-entropy alloys are recognised using empirical rules, for instance, a series of $Ti_xNbMoTaW$ (the molar ratio $x = 0, 0.25, 0.5, 0.75$ and 1) refractory high-entropy alloys were developed to find an alloy that can surpass the elevated temperature properties of Ni-based superalloys for further improvement of the turbine efficiency[8,10].

Computational tools can fast predict materials, which are enabling rapid advances in materials discovery and beyond through initiatives such as Materials-4.0[11,12]. Hitherto methods such as ab-initio calculations[13,14], Monte Carlo simulation[15], and CALPHAD[16,17] are used in the arena of material prediction for HEAs. Molecular dynamics (MD) & Density functional theory (DFT) methods are two other choices for studying the mechanical behaviour of materials. As with other techniques, these methods have limitations, for instance, DFT requires a large computational power and is limited to few atoms, while MD suffers limitations arising from force-field or inter-atomic potential function to capture the nature of atomic bonding (cocktail effect, lattice distortion, configurational entropy and sluggish diffusion) reported experimentally in HEAs[9].

Machine learning in particular is on the rise of prominence in the last decade as one can simply make use of the available dataset to discover a general trend[18]. Machine learning (ML) is a subset of artificial intelligence (a technique that enables machines to apply intelligence akin to a human brain), also known as a data-driven approach, which relies on pattern recognition from a given set of data[19,20]. The accuracy of results predicted by ML depends on the extent of data fed to the ML algorithm to train the system[21]. A significant surge in the use of ML in materials research is evidence of the promise this technique offers as explained by the other researchers[22]. Various ML algorithms such as the Artificial neural network, Convolutional neural network, Random Forest, Support vector machine, Decision trees, Gradient boosting, K-nearest neighbour, XGBoost, logistic regression and Naïve Bayes are employed over the past few years in predicting various phases of HEAs.

In all these studies, the known empirical rules for forming solid solution and phase determination have been applied, which include parameters such as atomic size difference (δ), electronegativity difference (Δχ), valence electron concentration (VEC), thermodynamic rule (mixing enthalpy (ΔH_{mix}) and mixing entropy (ΔS_{mix})), and others (Ω-parameter, ϕ-parameter, and γ-parameter)[23]. These terms describe the associated chemistry underlying the formation of HEAs and provide an insight into phase prediction, which can be mathematically stated as[24]:

$$\mathrm{VEC} = \sum_{i=1}^{n} (c_i \mathrm{VEC}_i) \tag{1}$$

$$\Delta \mathrm{S}_{\mathrm{mix}} = -R \sum_{i=1}^{n} (c_i \ln c_i) \tag{2}$$

$$\Delta \mathrm{H}_{\mathrm{mix}} = \sum_{i=1, i<j}^{n} \left(4H_{ij} c_i c_j\right) \tag{3}$$

$$\Delta\chi = \sqrt{\sum_{i=1}^{n} c_i (x_i - \overline{x})^2} \tag{4}$$

$$\delta = \sqrt{\sum_{i=1}^{n} c_i (1 - r_i/\overline{r})^2} \tag{5}$$

where, c_i is the atomic percentage of the ith element, n represents the total number of metallic elements in a high-entropy alloy. VEC_i is the valence electron concentration of the ith element, R is the gas constant, H_{ij} is the mixing enthalpy for the atomic pairs. x_i and $\bar{x}$ *a*re the Pauling electronegativity and averaged Pauling electronegativity, respectively, r_i is the atomic radius of the ith element, and $\overline{r}$ is the average atomic radius. Note that the actual value of δ was multiplied by numerical factor 100 for better clarity. Corroborating these parameters with the historical data have started to gain prominence[25,26] leading to the emergence of the use of ML to significantly identify, approximate and explain the structure–property relationships in HEAs in a cost-effective manner[12,27]. In the literature concerning phase prediction of HEAs using ML algorithms, no study can be seen that targets one particular synthesis route to extract the data reliably from experimental studies which can help avoid the spurious effect of an alternative synthesis routes on the resulting phase. For example, Bakr et al.[28] used neural network on 775 samples of HEAs synthesized from mixed manufacturing routes (Arc-melting, sintering, SLM, and others) and obtained 93.4% accuracy in predicting the existence of different phases (AM, BCC, FCC, and IM). Their study did not consider the effect of manufacturing method on the resulting phase of HEAs.

Furthermore, in an attempt to balance out the majority and minority class of an imbalanced dataset, various studies have exercised over-sampling and under-sampling methods. This has been done either by supplementing the synthetically generated data to remaining classes for making it equal to the majority class in case of over-sampling method or by subtracting the data from other classes for making it equal to the minority class in case of under-sampling method. Some studies have also utilised generative adversarial network (GAN) for generating synthetic data to avoid the biasness in the dataset. However, whether an alloy may be called as an HEA is controversial. This became the primary basis for our investigation as we believe that synthetic data is not comparable fully with the experimental data and cannot be considered prudent.

In this paper, we formulate the research objectives keeping in mind the current research gaps in the extant literature as below:

- Consolidate the scattered data on HEA synthesis obtained specifically through melting and casting routes such as: induction melting/induction levitation melting/ vacuum induction melting, arc melting/smelting and casting, arc melting + suction casting, electric/vacuum arc melting followed by suction casting techniques and to use this data as a fresh/new dataset for machine learning predictions. As opposed to previously published studies, the dataset used here included ternary, quaternary, quinary, and other alloys with more constituent elements making the algorithm ultra-robust, while targeting a synthesis route (melting + casting) that yields consistent phases during repeat experiments. This helped us avoid the spurious effect when combining the data from different synthesis routes on the resulting phase of a HEA. The dataset we tested was carefully screened from various experimental papers concerning synthesis of 3d-transition metals HEAs, refractory metals HEAs, HEA brasses and bronze, low-density HEAs, and some precious metal HEAs.
- To use a variety of available machine learning algorithms in their vanilla form (base models) such as K-nearest neighbours (V-KNN), support vector machine (V-SVM), decision tree classifier (V-DTC), random forest classifier (V-RFC), and XGBoost (V-XGB) to obtain phase prediction or to classify the phases of new HEAs into solid-solutions (FCC, BCC, FCC + BCC) or mixture of intermetallic phases (MIP) with the view to compare and contrast the robustness of each ML model based on various alternative evaluation metrics in case of imbalanced data, where accuracy percentage can be a misleading indicator.
- Whether synthetic data augmentation is reliable in predicting complex alloys such as HEAs? In testing this fact, we compared the vanilla RFC (V-RFC) model for 1200 original datasets with SMOTE-Tomek links augmented RFC (ST-RFC) model for 192 new datasets and in total 1392 datasets (1200 original + 192 generated = 1392).
- To synthesise HEA based on our ML predictions for proving the need to eliminate computationally/cost intensive approaches such as ThermoCalc, DFT and ab-initio methods in predicting phase of new HEAs.

Research methodology

Depending on the chemical nature of the constituting elements, HEAs can be classified into five main subfamilies: (i) 3d transition metal high-entropy alloys (3d TM HEAs) having Fe, Ni, Co, Mn, Ti and Cr typically exhibiting face-centered cubic (FCC) solid solutions; (ii) refractory high entropy alloys (RHEAs) constituted by elements of the groups IVB, VB and VIB exhibiting body-centered cubic (BCC) solid solutions; (iii) low-density high-entropy alloys, constituted of light elements like Al, Be, Li, Mg, Ti, Sc, typically presenting hexagonal closed-packed (HCP); and (iv) HEAs constituted by at least four of the lanthanide elements, also exhibiting HCP solid solutions; and (v) other HEAs, exhibiting the formation of multiple chemically disordered solid solutions (with FCC, BCC, or HCP lattice structures), ordered phases as B2 and L21, as well as different intermetallics (such as the σ, μ, C14, C15, and C36 Laves phases, among others)[29]. It suggests that a very scant number of HEAs have been discovered so far. It is timely to unearth the unexplored compositionally concentrated solid solution alloys at a faster pace to develop novel solutions for various engineering problems.

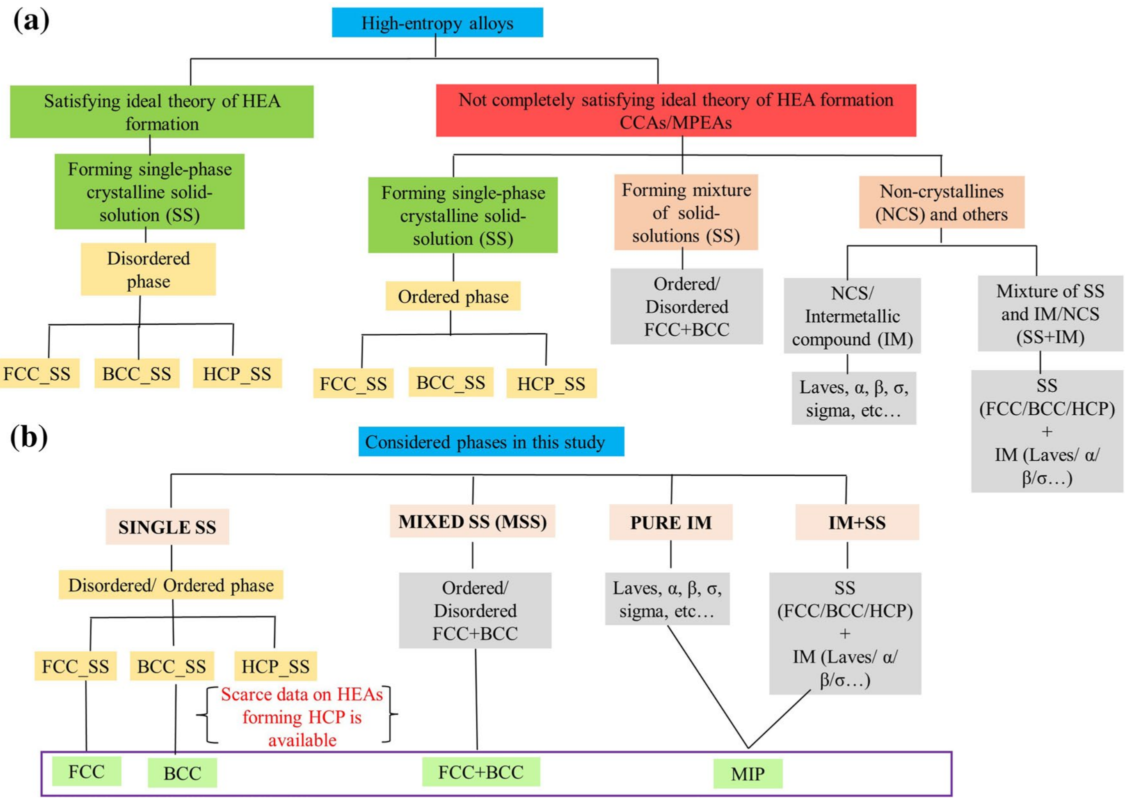

(a)
High-entropy alloys
Satisfying ideal theory of HEA formation
Not completely satisfying ideal theory of HEA formation CCAs/MPEAs
Forming single-phase crystalline solid-solution (SS)
Disordered phase
FCC_SS
BCC_SS
HCP_SS
Forming single-phase crystalline solid-solution (SS)
Ordered phase
FCC_SS
BCC_SS
HCP_SS
Forming mixture of solid-solutions (SS)
Ordered/ Disordered FCC+BCC
Non-crystallines (NCS) and others
NCS/ Intermetallic compound (IM)
Laves, α, β, σ, sigma, etc…
Mixture of SS and IM/NCS (SS+IM)
SS (FCC/BCC/HCP) + IM (Laves/ α/ β/σ…)
(b)
Considered phases in this study
SINGLE SS
Disordered/ Ordered phase
FCC_SS
BCC_SS
HCP_SS
Scarce data on HEAs forming HCP is available
MIXED SS (MSS)
Ordered/ Disordered FCC+BCC
PURE IM
Laves, α, β, σ, sigma, etc…
IM+SS
SS (FCC/BCC/HCP) + IM (Laves/ α/ β/σ…)
FCC
BCC
FCC+BCC
MIP

Figure 1. (**a**) Taxonomy of HEAs based on different definitions[30]. (**b**) Phases considered in this work to classify the data based on the existing literature.

HEAs emerged in about 2004 and currently a lot of work is ongoing on their developments. There are however open questions such as what constitutes an HEA. According to Miracle and Senkov[30], the term HEAs refers to a single-phase solid-solution prepared by controlling the configurational entropy, which limits the objective of exploring the vast compositional space of central region of hyper-dimensional phase diagram. On the other hand, terms such as compositionally complex alloys (CCAs) or multiprincipal element alloys (MPEAs) evokes the vastness of composition space, without concerning the types of phases present or the magnitude of configuration entropy. Figure 1a demonstrates the taxonomy of HEAs based on extant literature, which classifies compositions based on whether they satisfy the ideal theory of HEAs formation or not (called as MPEA/ CCAs).

Various phases in HEA known theoretically to date can be categorised as below:

(1) Ordered solid-solution (SS) phase: HEA residing in a singular crystalline phase such as B2 or β-ordered BCC phase
(2) Disordered solid-solution (SS) phase: BCC, FCC, HCP
(3) Mixed SS (Ordered + Disordered): FCC + BCC, BCC + B2, FCC + B2
(4) Pure Intermetallics (IM): α, β, σ, μ, L12, L21, C14, C15, and C36 Laves.
(5) IM + SS: BCC + C14 Laves, BCC1 + BCC2 + C15 Laves, BCC + β-ordered BCC, FCC + CoMo2Ni-type IM, FCC + IM and so on.

For the purpose of ML predictions, we clustered these phases together as for instance: (1)+(2) were considered as Single phase solid solution (SS), (3) was considered as Mixed solid solutions (MSS), and (4)+(5) were considered as mixture of intermetallic phases referred as 'MIP' as shown in Fig. 1.

Depending on the most-available phases procured from various literature, the current database used in this study contained four phases namely FCC, BCC, FCC + BCC, and MIP (mixture of intermetallic phases), as depicted in Fig. 1b. Due to scarcity of data belonging to HCP solid-solution phase, it was not considered in present study.

An open question in the literature is whether we can predict the type of phase (solid-solution, intermetallic, amorphous) for a given composition with known constituent elements, let's say: $Al_xCo_yHf_z$…… alloy, where *x, y, z* is the atomic weight percentage of each element. In this spirit, we demonstrate that ML strategy can be adopted to predict the phase of HEA merely using the reported experimental data by proper training, testing and validation of ML models which has been illustrated through the scheme shown in Fig. 2.

Data collection. Due to different stoichiometric ratios, distinct synthesis routes or processing conditions adopted by different researchers, the homogeneity in data collection on HEAs cannot be ensured, which makes

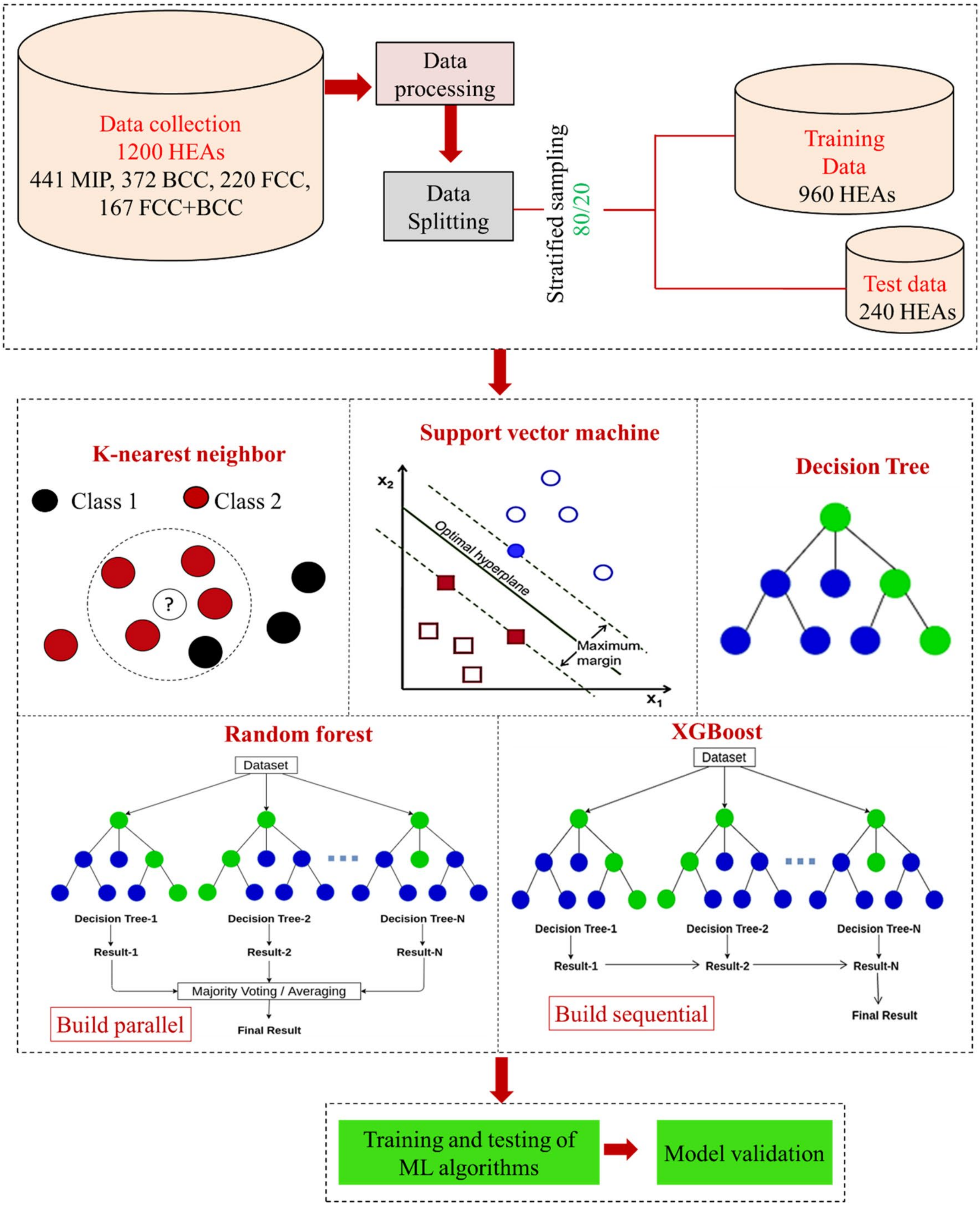


Figure 2. ML framework used in this work for phase prediction of HEAs as solid-solutions (FCC, BCC, FCC + BCC) or mixture of intermetallic phases (MIP).

it a challenging task to consolidate the data for comparison. This study extracted a dataset of 1200 unique compositions of HEAs experimentally synthesised from the melting and casting routes such as induction melting/induction levitation/vacuum induction melting + casting, arc melting/smelting + casting, arc melting + suction casting, or electric/vacuum arc melting followed by suction casting techniques, the corresponding reference to each HEA can be seen from the dataset provided and references[30–32]. The alloys prepared via other synthesis routes (powder metallurgy, selective laser melting, additive manufacturing and others) were not considered to avoid the effect of synthesis route[4]. The current dataset comprises 30 elements (Al, Co, Cr, Fe, Ni, Cu, Mn, Ti, V, Nb, Mo, Zr, Hf, Ta, W, C, Mg, Zn, Si, Re, N, Li, Sn, Be, B, Ag, Pt, Y, Pd, Au) and five physical parameters that are crucial for phase prediction of high-entropy alloys. The range of compositional and physical parameters (minimum, maximum, average and standard deviation values) are tabulated in Table 1. A detailed description of the complete dataset is provided as supplementary information [Table S1 and Fig. S1 in supplementary].

Variant	Minimum	Maximum	Average	Deviation
Al	10	100	33.55	24.86
Co	10	99.99	21.58	10.84
Cr	10	100	21.67	11.86
Fe	10	99.99	22.57	10.8
Ni	10	99.99	22	9.64
Cu	10	100	25.76	19.8
Mn	10	89.3	23.8	11.96
Ti	10	100	30.26	21.67
V	10	100	30.23	24.07
Nb	11.11	100	26.2	16
Mo	10	100	29.65	23.98
Zr	10	100	23.9	13.44
Hf	11	100	26.09	21.35
Ta	10	99	23.52	13.1
W	10	90	27.7	19.75
C	10	100	38.5	25.34
Mg	10	71.9	34.88	14.82
Zn	10	52.6	30.79	14.5
Si	10	100	34.63	26.6
Re	10	10	10	NaN
N	19	35	32.33	6.53
Li	10	50	24.75	15.27
Sn	10	99	37.51	25.33
Be	16.7	16.7	16.7	NaN
B	15.4	98.4	50.57	35.32
Ag	16.7	16.7	16.7	NaN
Pt	20	20	20	NaN
Y	11.8	20	16.4	4.01
Pd	20	40	30	14.14
Au	16.7	16.7	16.7	NaN
ΔH_{mix} (kJ/mol)	– 166.38	14.82	– 8.61	8.65
ΔS_{mix} (J/K mol)	1.609	19.05	12.99	1.905
δ ($\delta \times 100$)	0.056	69.93	5.48	3.29
$\Delta\chi$	0.015	3.92	0.171	0.253
VEC	2.0	10.4	6.74	1.54

Table 1. Range of composition (atomic weight %) and physical parameters used in this study.

Empirical relations observed in high entropy alloys suggest that an HEA (solid-solution phase) formation becomes plausible when $\delta < 6.6\%$ and $11.6 < \Delta H_{mix} < 3.2$ kJ/mol. When δ is large ($\delta > 6.6\%$) and ΔH_{mix} is noticeably negative ($\Delta H_{mix} = - 12.2$ kJ/mol)[24], it leads to an amorphous phase instead of a crystalline phase. Intermetallic compounds tend to form in the intermediate range in terms of δ and ΔH_{mix}, or it overlaps largely with those for solid solutions and amorphous phase. Furthermore, for the identification of crystal structure in various solid solution forming HEAs, the effect of VEC was formulated and the threshold value was found to be as:

- BCC when VEC < 6.87,
- FCC when VEC > 8.0 and
- Mixed phase (BCC + FCC) when VEC is in between 6.87 and 8.0.

A joint plot and swarm plot are shown for better visualisation [Fig. S2 in supplementary]. Zhang et al.[33] criterion were almost the same for δ ($\delta < 6.6\%$) but the range of ΔH_{mix} was slightly different ($-15 < \Delta H_{mix} < 5$ kJ/mol). Among all physical parameters (atomic size difference (δ), electronegativity difference ($\Delta\chi$), valence electron concentration (VEC), thermodynamic rule (mixing enthalpy (ΔH_{mix}) and mixing entropy (ΔS_{mix})), and others (Ω-parameter, ϕ-parameter, and γ-parameter)) proposed for guiding the design of stabilizing phases of HEAs, only five crucial parameters (ΔH_{mix}, ΔS_{mix}, δ, $\Delta\chi$, VEC) were considered for this study, as these are widely accepted and easy to compute. Also, the mere requirement of these five parameters which can easily be obtained theoretically, guiding to the development of a new alloy based on our methodology would ensure effortless development of HEAs in future. The ΔH_{mix} for available HEAs in the dataset were calculated using Miedema's rule[34], while ΔS_{mix}, δ, $\Delta\chi$, VEC were calculated by following Guo et al.[35]. Other parameters such as geometric parameter (γ) still awaits support from more experimental data. Accordingly, these five most influencing physical parameters being primarily responsible for a crystal structure in HEA were considered in the design of this study.

As a proof of concept for testing the cruciality of these five parameters, the heatmap shown in Fig. 3 was drawn using the seaborn library of python, which represents the Pearson correlation coefficient of five parameters governing the formation of HEA proposed by various researchers. This heatmap helps to visualize the correlation between features for sanity check of redundant features. Two features that are strongly positively correlated (when two features move in tandem) or negatively correlated (when two features are inversly related) leads to the problem of multicollinearity that significantly reduces the model performance and increases the standard error. Thus, it is suggested to eliminate one of the features that are strongly correlated[36,37]. No such strong positive or negative correlation between any two independent feature was observed; thus, all the five parameters were considered for further study without any elimination.

The complete dataset was labelled in four categories: FCC, BCC, FCC + BCC, and MIP. The alloys with a single-phase ordered/ disordered FCC or multiple FCC such as (FCC1 + FCC2) were considered in 'FCC' category. Similarly, alloys with a single-phase ordered/ disordered BCC or multiple BCC such as (BCC1 + BCC2) were considered in 'BCC' category and the mixture of FCC and BCC phases was considered in 'FCC + BCC' category. Compositions containing pure IM compounds (such as Laves, α, β, sigma etc.) or forming a mixture of SS + IM (such as FCC + IM, FCC + BCC + α, BCC + IM, FCC + α + β, BCC + Laves etc.) were considered in 'MIP' category, while the amorphous phase was not included in the analysis.

The 1200 datasets of HEAs used in this work contains 441 compositions of MIP phase, 372 compositions of BCC phase, 220 compositions of FCC phase and 167 compositions of FCC + BCC phase with no duplicated entry of any alloy. Depending on the number of instances belonging to each class, a dataset can typically be recognised as a balanced (when the number of instances available from each class is equal) or an imbalanced dataset (when the number of instances available from each class is different) for a classification problem. In case of an imbalanced dataset, the class with the highest and least number of instances is known as the majority and minority class, respectively.

It must be noted that the present study discusses the phase prediction of HEA as solid solution phases such as (FCC, BCC, or FCC + BCC) or MIP (pure IM or mixed IM + SS) phases for an imbalanced dataset by targeting only those HEAs that were developed via. melting and casting route. The effect that the imbalanced dataset makes on the performance of ML algorithms has been explicitly discussed in section "Results and discussions".

Data processing. Before feeding the data into the ML algorithms, some statistical processing steps were performed to make the predictions more meaningful[38,39]. The text data (phases) was converted into numeric values (MIP: 0, BCC: 1, FCC: 2, and FCC + BCC: 3), outlier detection was performed to remove the outliers from the dataset; various imputation methods such as simple imputer with different strategies (mean, median, and constant), KNN imputer and MICE imputer was employed to impute the missing values (NaN) in the dataset. Feature scaling was performed on each set of imputed data to normalise the data into a finite range, using robust scalar imported from scikit-learn library. The robust scaling formula can be expressed as[38]:

$$X_{\text{robust}} = \frac{X - X_{median}}{X_{75} - X_{25}} \tag{6}$$

where X is an input variable, X_{median} is the median of X, X_{75} is the 75th quantile and X_{25} is the 25th quantile of X. The difference between 75th quantile and 25th quantile is also known as interquartile range (IQR).

Brief description of the machine learning algorithms. *KNN algorithm.* The KNN algorithm searches for the nearest neighbours by measuring the distance between the two points[40,41] and is expressed as:

$$d(q, x_i) = \sum_{f \in F} w_f \,\delta\left(q_f, x_{i_f}\right) \tag{7}$$

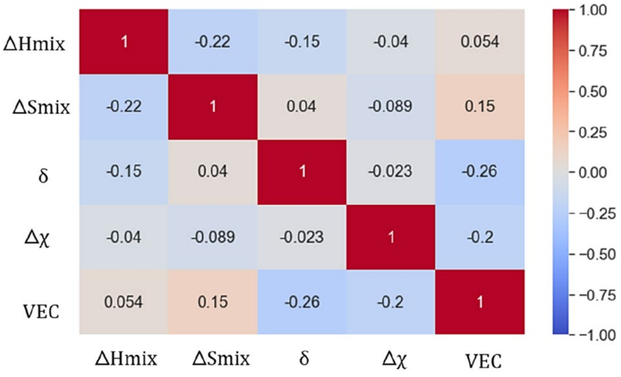


Figure 3. Heatmap showing the Pearson correlation coefficient among five HEA parameters.

For classifying an unknown input variable (q) one needs to know the existing input variable (x_i) in F and the weight factor ($\mathrm{w_f}$) for each feature. Based on this distance, the k nearest neighbours is selected, and the class of q is determined from the voting of the nearest neighbours as below:

$$\mathrm{Vote}\left(y_i\right) = \sum_{c=1}^{k} \frac{1}{d\left(q, x_c\right)\rho} 1\left(y_i, y_c\right) \tag{8}$$

This returns 1 if the class labels matches and 0 if does not match. The vote assigned to class y_i by neighbour x_c is the inverse of their distance, i.e., $1(y_i, y_c)$.

SVM algorithm. SVM classifier searches for the hyper plane that best separates different classes by maximising the margin (the distance between the nearest data points from different class sets) to avoid the local minima and to achieve the best separation of different classes[42,43]. The decision function is as below:

$$f(x) = w.x + b \tag{9}$$

$$\underset{\mathbf{w},\xi}{min}\left\{\frac{1}{2} \parallel \mathbf{w}\parallel^2 + C\sum_{i=1}^{N}\xi_i\right\} \tag{10}$$

$$\text{Subject to}: y_i(\mathbf{w}\cdot\mathbf{x_i}) \geq 1 - \xi_i, \xi_i \geq 0 \tag{11}$$

DTC algorithm. A decision tree classifier splits the dataset into root node, sub-node and leaf-node by calculating the information gain, i.e., change in entropy after dividing a dataset based on attributes, which helps to determine the order of features in various nodes of a decision tree (quality of splitting)[44,45]. Information gain is calculated as below:

$$H(Y|X) = H(X,Y) - H(X) \tag{12}$$

where $H(Y \mid X)$ is the conditional entropy, $H(X)$ is the entropy of random variable X, and $H(X, Y)$ is the joint entropy, calculated as follows:

$$H(X,Y) = -\sum_{i,j} p\left(x_i, y_j\right)\log_2 p\left(x_i, y_j\right) \tag{13}$$

RFC algorithm. In a random forest classifier, ensembles of various decision trees (base learners) are considered such as $h_1(x), h_2(x), \ldots, h_J(x)$. It takes majority of votes to calculate *f(x)* such that the loss function is minimised[46,47]. Loss function is expressed as below:

$$L(Y,f(x)) = I(Y \neq f(x)) = \begin{cases} 0, \text{ if } Y = f(x) \\ 1, \text{ otherwise} \end{cases} \tag{14}$$

$$\text{Voting is based on } f(x) = \mathrm{arg}max\sum_{j=1}^{J} I\left(y = h_j(x)\right) \tag{15}$$

XGBoost algorithm. XGBoost combines a set of weak classifiers to create a strong classifier[42,48]. The objective function is expressed as:

$$\mathrm{obj}(\theta) = \sum_{i}^{n} l\left(\widehat{y}_i, y_i\right) + \sum_{k}^{K}\Omega\left(f_k\right) \tag{16}$$

The term $l\left(\widehat{y}_i, y_i\right)$ represents the loss function, which measures the difference between predicted output and the actual output, where y_i is the actual output, and $\widehat{y}_i$ is the predicted output given by $\widehat{y}_i = \sum_k^K f_k(x_i), f_k \in F. x_i$ is the input variable and $\Omega\left(f_k\right)$ is regularisation term that helps to avoid overfitting by penalising the complexity of the model. XGBoost is trained additively, where one tree is optimised and added each at a time. Supplementary provides the description and proper visualisation of these algorithms [Table S2 in supplementary].

Results and discussions

For an imbalanced dataset problem such as the one tested in this work, careful treatment is essential or else the predictions can be out of order. Accuracy is well-accepted measurement for evaluating the performance of a classification problem. However, for an imbalanced dataset, the use of accuracy as an effective indicator has been questioned recently by various authors[49–53]. Therefore, alternative evaluation metrics for assessing the effectiveness of ML models for imbalanced dataset were explored, as accuracy alone is not trustworthy. Various other evaluation metrics such as ROC-AUC score, Precision, Recall, and F1-score available in scikit-learn version 1.1.1 module[54] in python version 3.9.12 are robust measures for imbalanced dataset classification[55].

The receiver operating characteristic (ROC) curve is a probability curve typically plotted for binary classification tasks at different classification threshold values[54,56].

This paper studies the phase prediction of HEAs as solid-solution phases such as FCC, BCC, and FCC + BCC, or MIP (which can be either pure IM or mixed IM + SS phase, as described in Fig. 1a), by targeting the multiclass classification of HEAs into four phases namely FCC, BCC, FCC + BCC and MIP using the real-world imbalanced dataset of HEAs.

The ROC curve can be extended to multiclass classification with 'one-vs-one' and 'one-vs-rest' strategies[54,57]. Here, 'one-vs-rest' strategy was employed, to compute the AUC score (by calculating the area under the ROC curve) for each class against the rest of the class and by subsequently taking its average. ROC_AUC score provides a summary of classifier's performance by measuring the area under the ROC curve, which is more likely to be true representative of model's performance. ROC_AUC score varies in between 0 to 1, where 1 denotes the perfect classifier, while 0 denotes a perfectly incorrect classifier.

Precision evaluates the fraction of predicted positives that were actually true positives (TP), Recall determines the ability of a model to predict the true positives (TP), and F1-score calculate the harmonic mean of Precision and Recall[58,59]. The detailed description of Precision, Recall and F1-score are given in supplementary [Table S3 in supplementary].

The effectiveness of five vanilla (base) models (V-KNN, V-SVM, V-DTC, V-RFC, and V-XGB) was tabulated and compared using all the above-mentioned evaluation metrics, for five different imputers (Simple imputer (SI) with strategy mean, median and constant; KNN imputer and MICE imputer), tabulated in supplementary [Table S4]. No significant difference in model's performance for these imputers were observed. Vanilla-RFC (V-RFC) performed best compared to other algorithms, with an average test accuracy of 84%, ROC-AUC score of 0.9649, tenfold cross-validation mean score of 0.9315 which is shown in Fig. 4a.

Furthermore, F1-score, Recall and Precision were also evaluated for all five vanilla models, where V-RFC obtained higher precision, recall and f1-score in contrast to other models (see Fig. 4b). Note that for each model, five iterations were performed and their average was considered. A bar-chart comparing the performance of all ML models tested in this work compares the three outcomes (Fig. 4a), namely, Accuracy (peach-coloured bars), ROC-AUC score (light green bars) and tenfold cross-validation score (light purple bars). We further explored hyper-parameter tuning of RFC model (HT-RFC) and noticed an increment of approximately 3% in average test accuracy (87.49%).

It is not surprising that many studies have reported higher accuracy from their ML predictions but the fact that these accuracies have come through the aid of synthetic data by mixing with the experimental data cast doubts on the reliability of these models. For instance, Risal et al.[60] obtained 92.31% accuracy with higher ROC-AUC, precision, recall and f1-scores but they used over-sampling/under-sampling method to balance out the majority and minority class data by augmenting it with synthetic data while acknowledging that the "ML algorithms usually do not perform well for imbalanced dataset".

In our considerations, augmenting or polluting the real-world data with synthetically generated data is not reliable for two reasons: first accuracy alone is not the most robust measure for assessing the performance of the ML model for an imbalanced data and second, the controversy on calling an alloy as HEA still exists, thus it cannot be assured that the generated samples are truly a high-entropy alloy.

Still, as per current vogue, we tried to resample our data using SMOTE-Tomek links method for V-RFC model (outperformed among other vanilla algorithms), which is quite different from other existing over-sampling and under-sampling methods. It generates synthetic data for minority class using SMOTE and removes the data from majority class that is closest to minority class using Tomek links[61]. An average of 92% accuracy was observed for augmented data (1200 + 192 = 1392) using SMOTE-Tomek links (ST-RFC), by generating 192 synthetic data.

We further evaluated the performance of V-RFC and ST-RFC with a confusion matrix to analyse it's prediction quality for each phase. A confusion matrix is an easy way to visualise classifier's performance, where $n \times n$ matrix is created (n is the number of classes) to provide better insights into the correctly and incorrectly classified instances. Two confusion matrices of 4 × 4 were created on test data of 240 HEA samples (93 MIP, 67 BCC, 46 FCC, and 34 FCC + BCC) for V-RFC model from the original dataset (1200), and 279 HEA samples (78 MIP, 67 BCC, 40 FCC, and 94 FCC + BCC) for ST-RFC model from augmented dataset (1200 + 192 = 1392 samples) to investigate the performance of the RFC model in predicting phases of HEAs, as shown in Fig. 5.

It can be noticed that the number of samples in the minority class increases by maintaining the stratified ratio between the classes for the ST-RFC model. Although the number of incorrect predictions becomes less in the case of ST-RFC model in contrast to the V-RFC model, it is still ineffective considering the uncertainty associated with using synthetic data which cannot guarantee a high-entropy alloy. Furthermore, the ROC curve and their AUC score for all five vanilla models trained on original data, and SMOTE-Tomek links models trained on augmented data were plotted as shown in Fig. 6a,c. The ROC-AUC score of all ST-models were higher than the vanilla models. We selected the best models i.e., V-RFC and ST-RFC models and evaluated AUC score for each phase (MIP, BCC, FCC, and FCC + BCC) for the test data of original dataset (test data = 240 HEAs) and augmeneted dataset (test data = 279 HEAs) which are depicted in Fig. 6b,d. It can be seen that although the ROC-AUC of ST-RFC model was approximately 3% higher than the vanilla model (V-RFC), still both models provided approximately similar AUC score for each phase except for the FCC + BCC phase. The reason of higher AUC score for FCC + BCC phase for ST-RFC model is the increased number of instances of minority class i.e., FCC + BCC (34 instances), which has now become the majority class (94 instances) by augmenting the data in case of ST-RFC model. Therefore, we reinforced our point by comparing the confusion matrices and ROC-AUC score for original and augmented dataset. As these matrices provided better insights and are considered as true

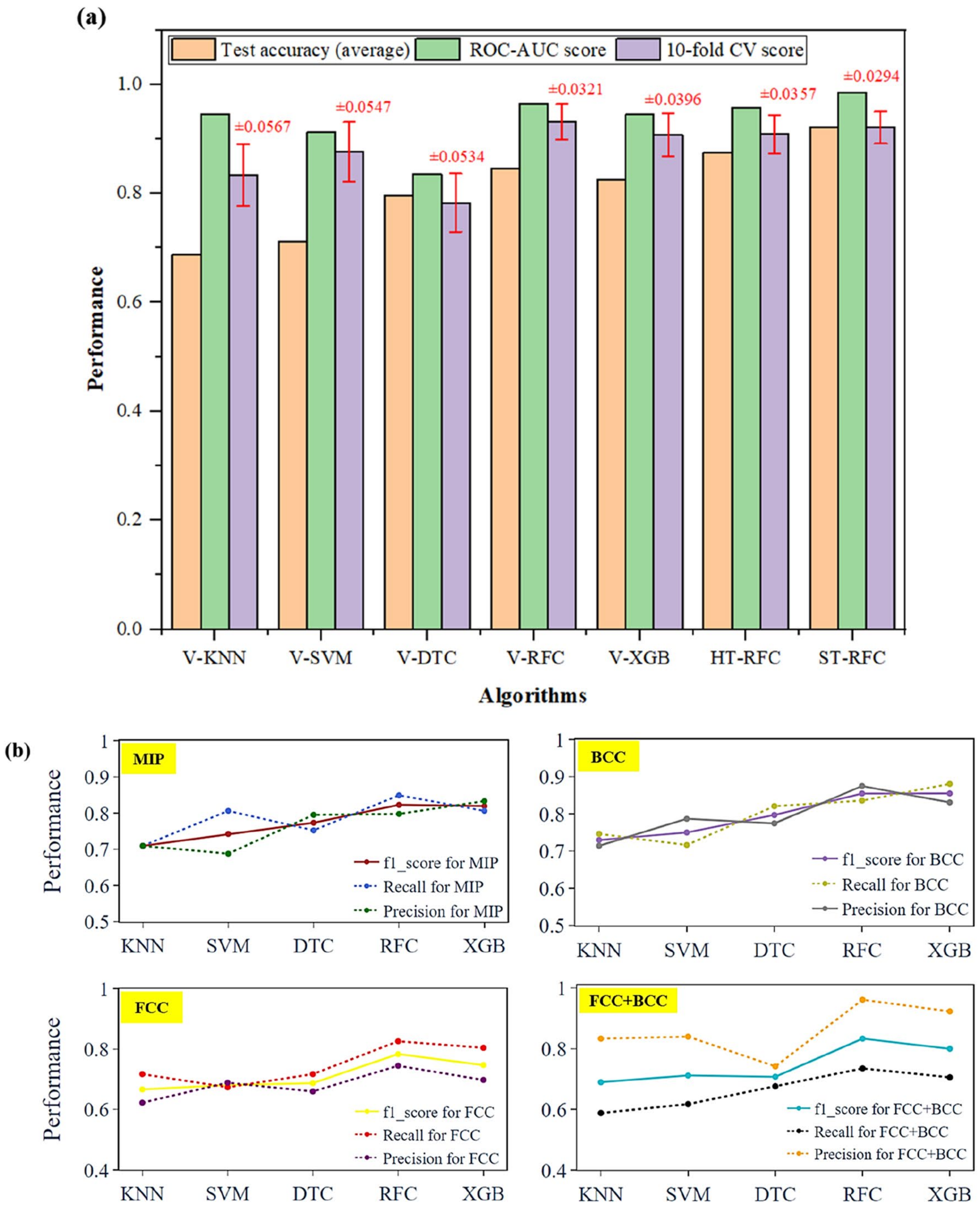


Figure 4. (**a**) Performance comparison of V-KNN, V-SVM, V-DTC, V-RFC, V-XGB, HT-RFC, and ST-RFC models using average test accuracy (multiply by 100 for % value), ROC_AUC score, tenfold cross-validation score and its standard deviation (values shown in red color), (**b**) F1-score, Recall and Precision for four distinct phases of HEAs for five vanilla models.

indicator of a classification model, we claim that augmenting data to increase model's accuracy is not a reliable practice. Therefore, this study is more pertinent considering the aforementioned issues.

Model validation

Validation based on the literature. The predictive capability of all five classifiers was further tested for alloys that were not considered for training or testing the dataset for sense check. Various phases of five alloys (2 refractory HEAs[10], one 3d-transition metal HEA[62] and 2 precious metal HEAs[63]) that are recently reported were taken as examples from experimental studies (literature) that are shown in Table 2. The physical parameters corresponding to these HEAs were calculated using the chemical formulae mentioned previously in earlier sections. The phase of HEA highlighted in bold fonts indicated wrong predictions (predictions does not match

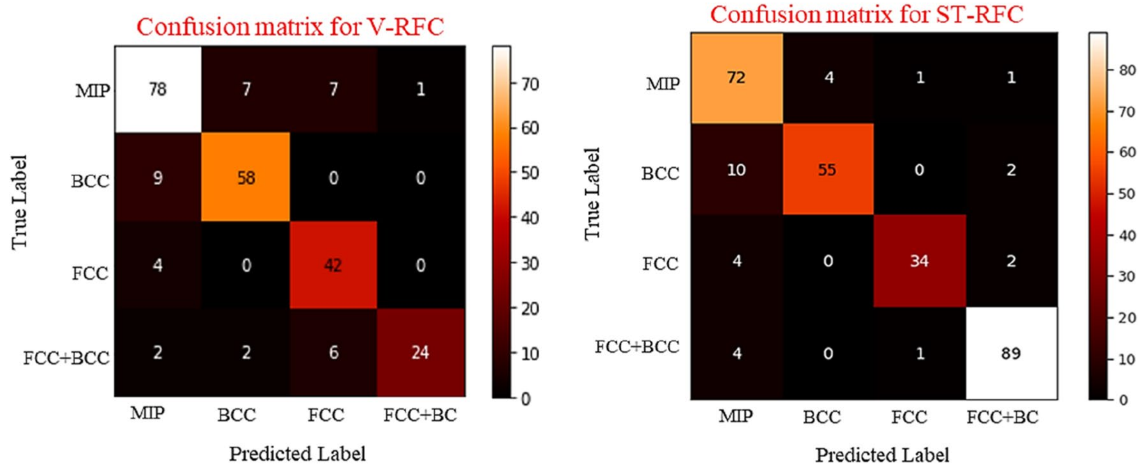


Figure 5. Confusion matrix comparing the performance of V-RFC, and ST-RFC for each distinct phase.

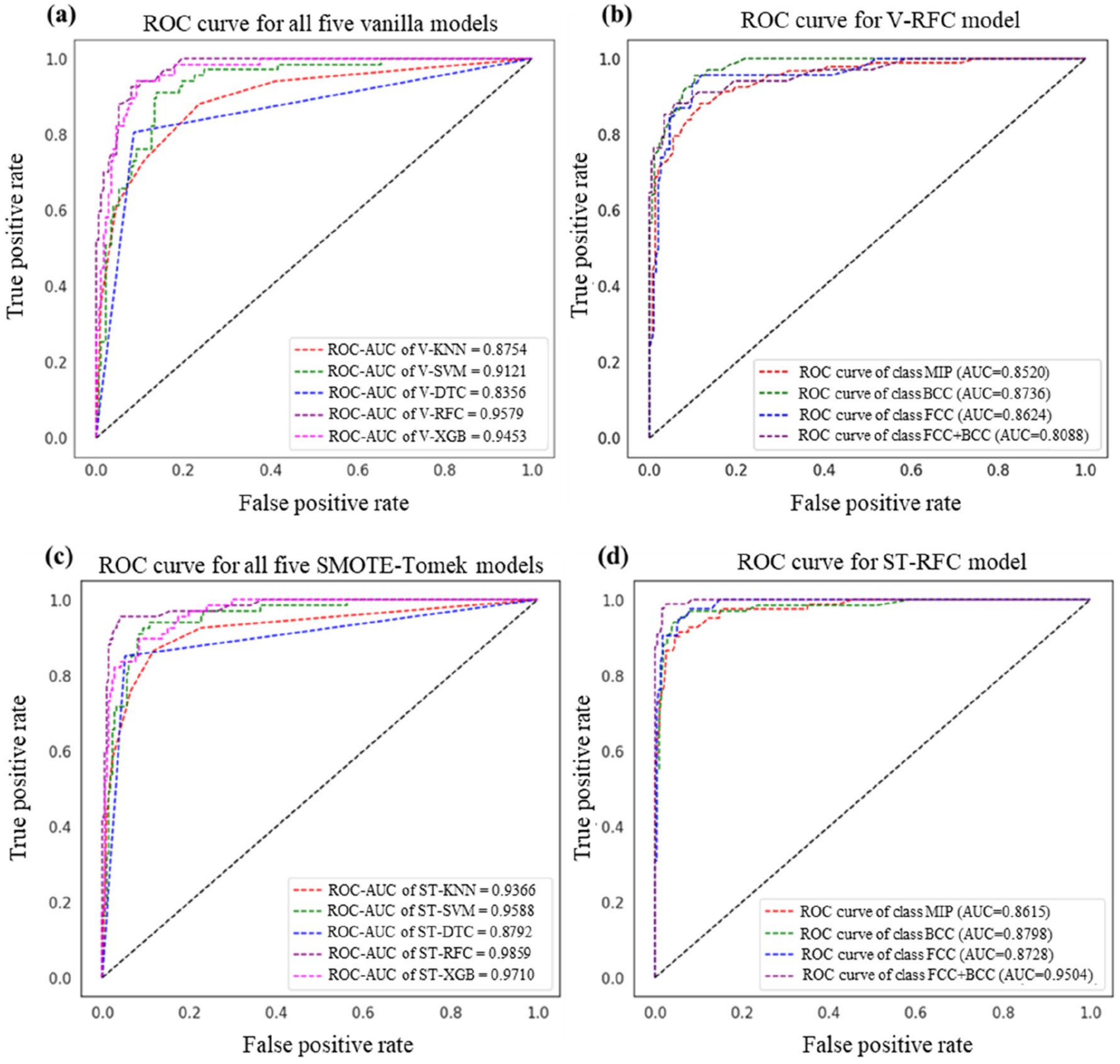


Figure 6. ROC-AUC scores for (**a**) five vanilla (base) models, (**b**) AUC score of each phase for the best vanilla model i.e., V-RFC model, (**c**) for all five SMOTE-Tomek links augmented model, (**d**) AUC score of each phase for the best SMOTE-Tomek links augmented model i.e., ST-RFC model.

	Physical parameters					Predicted phases			Actual phases		
Alloys	ΔH_{mix} (kJ/mol)	ΔS_{mix} (J/K.mol)	VEC	Δχ	δ	KNN	SVM	DTC	RFC	XGB	
Refractory HEAs[10]											
$Ti_{0.5}NbMoTaW$	− 3.06	13.15	5.33	0.361	2.6	BCC	**FCC**	BCC	BCC	BCC	BCC
TiNbMoTaW	− 3.04	13.38	5.2	0.356	2.75	BCC	BCC	BCC	BCC	BCC	BCC
3d transition metal HEAs[62]											
$Al_{0.5}CrCuNiV$	− 6.01	13.15	5.43	0.133	4.39	**BCC**	**BCC**	**BCC**	*MIP*	**BCC**	FCC + 2BCC + ordered B2 phase
Precious metal HEAs[63]											
PdPtRhIrCuNi	− 2.56	14.89	9.82	0.161	3.71	FCC	**BCC**	**FCC + BCC**	FCC	FCC	FCC
AuPdAgPtCuNi	− 2.22	14.89	10.49	0.236	5.39	**FCC + BCC**	FCC	**FCC + BCC**	FCC	FCC	FCC

Table 2. Validation of all vanilla (base) model's performance for unseen compositions (not used in training or test datasets). Phases shown in bold font indicate the wrong prediction, italic font shows an exceptional case, and the remaining (nonbold and nonitalic) indicate the correct prediction.

with experimentally characterised phase), the italicized phases show exceptional case (where the certainty of matching ML predictions and the actual phase is limited), and the remaining phases (nonbold and nonitalic) show the correct prediction revealing that the ML models corroborate with the experimentally reported phases.

It can further be noted that the RFC classifier predicted the phases correctly in most cases. In case of $Al_{0.5}CrCuNiV$ 3d-transition metal HEA[62], RFC model predicted MIP phase (in italic font), while the actual phase contains 1FCC + 2BCC + ordered B2 phase, which is a complex multiphase alloy. The ML models assumed that MIP can be either pure intermetallic compound (IM) or mixtures of intermetallic and solid solutions (IM + SS) which was discussed in section "Research methodology". Assuming that it can be inferred that the RFC model's prediction is correct for all new compositions taken from different experimental studies, that were not the part of either training or test dataset. However, it is limited in inferring the number and types of phases present in complex multiphase HEAs. A list of such complex multiphase HEAs (that were not the part of training or test set) is tabulated in Table S5 in supplementary as an additional information.

To strengthen the support to our claim further, we critically assessed the recently published literature based on various evaluation metrics, shown in Fig. 7. Scant literature was found to focus on alternative evaluation metrics such as ROC-AUC, precision, recall and F1-score. The proposed RFC model in the present work, revealing the ROC-AUC score, their tenfold cross-validation score, confusion matrix, F1-score, Recall, and Precision, showed satisfactory performance in predicting phases of HEAs as solid-solution phases (BCC, FCC, FCC + BCC) or MIP (denoting either pure intermetallic compounds (IM : such as α, β, σ , L12, C14, C15, C36 Laves, and others) or mixture IM + SS phases such as FCC + IM, FCC + BCC + α, BCC + IM, FCC + α + β, BCC + Laves, BCC1 + BCC2 + C15 Laves etc.)) for large imbalanced dataset of those HEAs that were synthesized via. melting and casting route only, without augmenting/polluting experimental data with generated ones.

Synthesis and characterisation of a new HEA ($Ni_{25}Cu_{18.75}Fe_{25}Co_{25}Al_{6.25}$).

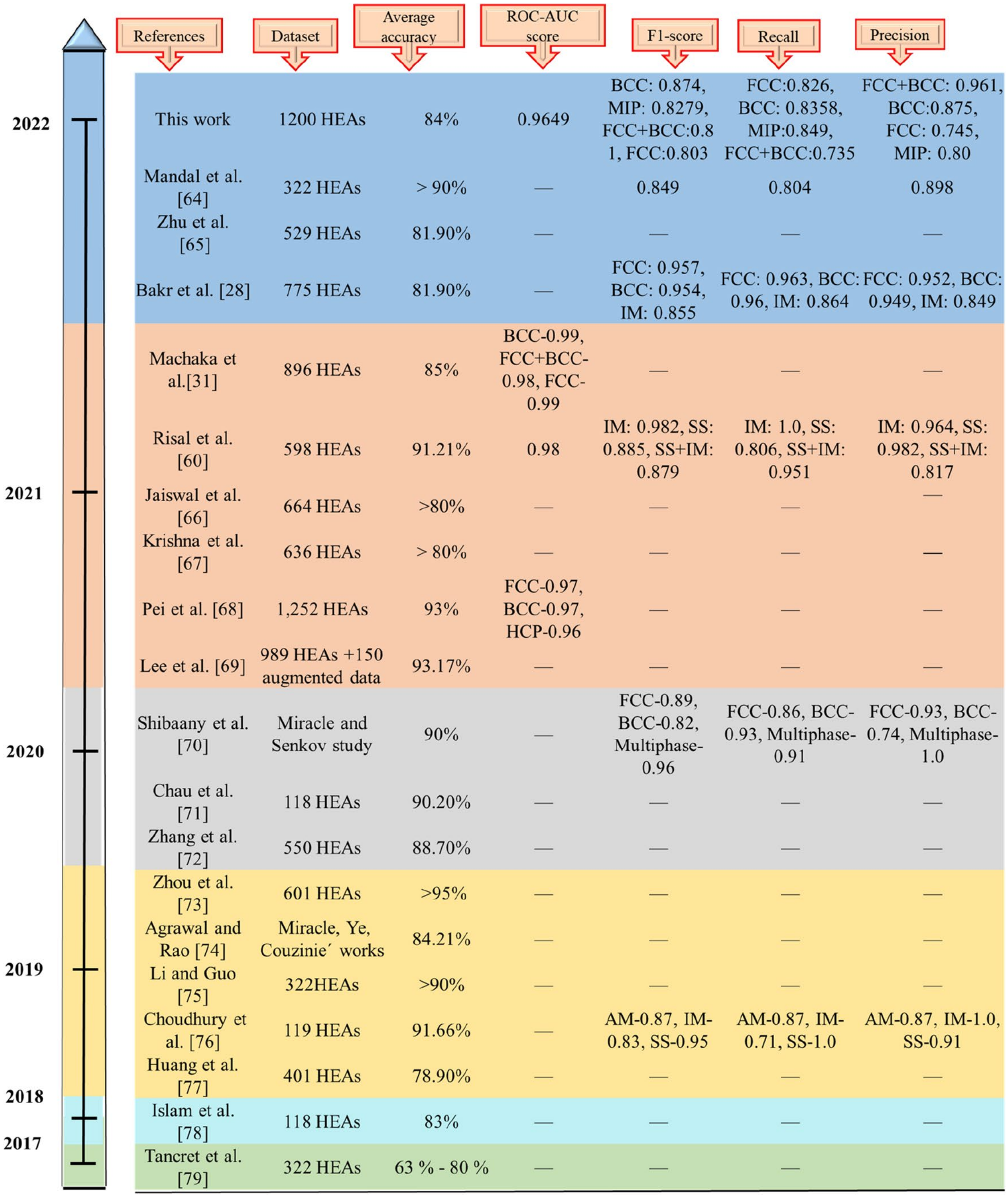
In accord with these learnings, a new high entropy alloy was synthesised based on the predictions obtained from the RFC algorithm. This alloy consists of Nickel, Copper, Iron, Cobalt and Aluminium with a composition of $Ni_{25}Cu_{18.75}Fe_{25}Co_{25}Al_{6.25}$.

To begin the experimental synthesis, the metal buttons were procured. Various metal buttons of Ni, Cu, Fe, Co and Al elements (purities > 99.99%) were purchased from Thermofisher Scientific®. All elemental metals were melted together by vacuum arc melting under inert gas (high purity Ar) environment. The ingot formed in the process was melted and solidified multiple times to ensure chemical homogeneity, and then the HEA button was vacuum sealed in a quartz tube, homogenised at 1000 °C for 10 h, and then quenched into water for stabilising high-temperature phase. The detailed description of the newly synthesised high-entropy alloy is specified in Table 3.

X-ray diffraction (XRD, Broker D8) was used to identify the phase of $Ni_{25}Cu_{18.75}Fe_{25}Co_{25}Al_{6.25}$ alloy, with the wavelength Cu Kα (λ = 1.54056 Å) at a step size of 0.02° recorded with angles *(2θ)* in the range of 20°–100° (see Fig. 8). The Bragg's peaks (111), (200), (220), (311), and (222), belong to the lattice planes of FCC phase, while no other peaks corresponding to ordered structure were detected, indicating that this new HEA resides in a crystalline FCC structure.

Figure 9 compares various HEAs for the test datasets (240 alloys) from original data with the newly developed and synthesised HEA composition $Ni_{25}Cu_{18.75}Fe_{25}Co_{25}Al_{6.25}$. The orange dot represents the reported experimental phase of the HEA for the test data while the blue triangles represent the RFC prediction, and the red asterisk represents the new composition of the $Ni_{25}Cu_{18.75}Fe_{25}Co_{25}Al_{6.25}$. RFC algorithm indicated that this new HEA would stabilise as FCC phase at room temperature.

A remarkable agreement between RFC model prediction and experiment can be seen for this new composition of HEA. It can be inferred that the V-RFC model is reliable and robust in predicting phases of novel compositions of HEAs as simple solid solution (FCC, BCC, FCC + BCC) and MIP (Mixture of intermetallic phases) with higher reliability of phase prediction, where MIP denotes the presence of either pure IM compounds such as α, β, σ, L12, L21, C15, C15, C36 Laves or mixture of IM + SS phases (FCC + IM, FCC + BCC + α, BCC + IM, FCC + α + β, BCC + Laves, BCC1 + BCC2 + C15 Laves, BCC + β-ordered BCC, FCC + CoMo2Ni-type IM, FCC + IM etc.). However, this method is limited in exactly interpreting the number and types of phases present in a complex multiphase HEA, which it usually predicts as MIP phase, but it is robust for predicting solid-solution phases.



2022
2021
2020
2019
2018
2017

References	Dataset	Average accuracy	ROC-AUC score	F1-score	Recall	Precision
This work	1200 HEAs	84%	0.9649	BCC: 0.874, MIP: 0.8279, FCC+BCC:0.81, FCC:0.803	FCC:0.826, BCC: 0.8358, MIP:0.849, FCC+BCC:0.735	FCC+BCC: 0.961, BCC:0.875, FCC: 0.745, MIP: 0.80
Mandal et al. [64]	322 HEAs	> 90%	—	0.849	0.804	0.898
Zhu et al. [65]	529 HEAs	81.90%	—	—	—	—
Bakr et al. [28]	775 HEAs	81.90%	—	FCC: 0.957, BCC: 0.954, IM: 0.855	FCC: 0.963, BCC: 0.96, IM: 0.864	FCC: 0.952, BCC: 0.949, IM: 0.849
Machaka et al.[31]	896 HEAs	85%	BCC-0.99, FCC+BCC-0.98, FCC-0.99	—	—	—
Risal et al. [60]	598 HEAs	91.21%	0.98	IM: 0.982, SS: 0.885, SS+IM: 0.879	IM: 1.0, SS: 0.806, SS+IM: 0.951	IM: 0.964, SS: 0.982, SS+IM: 0.817
Jaiswal et al. [66]	664 HEAs	>80%	—	—	—	—
Krishna et al. [67]	636 HEAs	> 80%	—	—	—	—
Pei et al. [68]	1,252 HEAs	93%	FCC-0.97, BCC-0.97, HCP-0.96	—	—	—
Lee et al. [69]	989 HEAs +150 augmented data	93.17%	—	—	—	—
Shibaany et al. [70]	Miracle and Senkov study	90%	—	FCC-0.89, BCC-0.82, Multiphase-0.96	FCC-0.86, BCC-0.93, Multiphase-0.91	FCC-0.93, BCC-0.74, Multiphase-1.0
Chau et al. [71]	118 HEAs	90.20%	—	—	—	—
Zhang et al. [72]	550 HEAs	88.70%	—	—	—	—
Zhou et al. [73]	601 HEAs	>95%	—	—	—	—
Agrawal and Rao [74]	Miracle, Ye, Couzinie´ works	84.21%	—	—	—	—
Li and Guo [75]	322HEAs	>90%	—	—	—	—
Choudhury et al. [76]	119 HEAs	91.66%	—	AM-0.87, IM-0.83, SS-0.95	AM-0.87, IM-0.71, SS-1.0	AM-0.87, IM-1.0, SS-0.91
Huang et al. [77]	401 HEAs	78.90%	—	—	—	—
Islam et al. [78]	118 HEAs	83%	—	—	—	—
Tancret et al. [79]	322 HEAs	63 % - 80 %	—	—	—	—

Figure 7. Performance comparison with existing literature[28,31,60,64–79].

Novel high-entropy alloy ($Ni_{25}Cu_{18.75}Fe_{25}Co_{25}Al_{6.25}$)				
Ni	**Cu**	**Fe**	**Co**	**Al**
Chemical composition (wt %)				
0.25	0.1875	0.25	0.25	0.0625
Novel high-entropy alloy ($Ni_{25}Cu_{18.75}Fe_{25}Co_{25}Al_{6.25}$)				
ΔHmix (kJ/mol)	**ΔSmix (J/K mol)**	**VEC**	**Δχ**	**δ**
Calculated physical parameters				
0.2656	12.689	9	0.07178	3.5973

Table 3. Detailed description of newly synthesized high-entropy alloy.

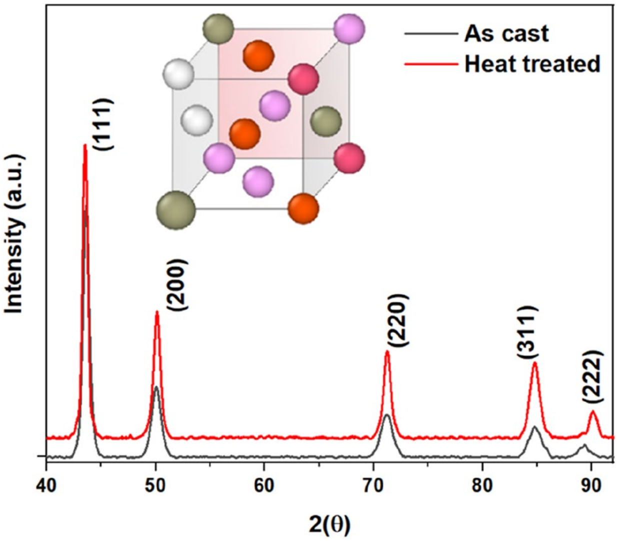


Figure 8. XRD analysis of newly synthesized HEA ($Ni_{25}Cu_{18.75}Fe_{25}Co_{25}Al_{6.25}$) for as-cast and heat-treated sample. Peaks (111), (200), (220), (311) and (222) correspond to the FCC structure.

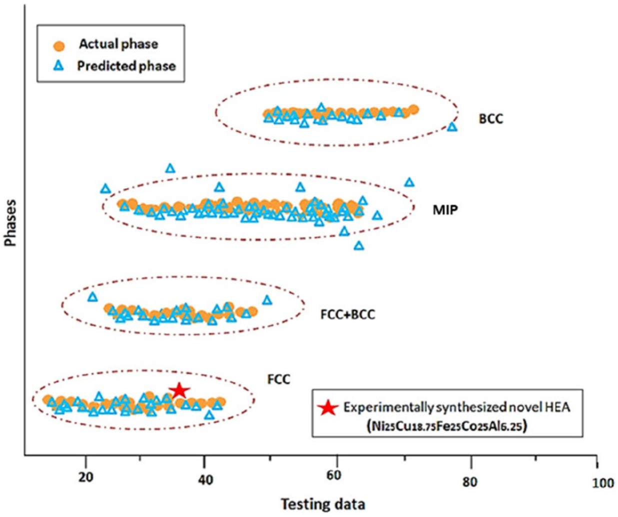


Figure 9. Phase prediction of novel HEA composition $Ni_{25}Cu_{18.75}Fe_{25}Co_{25}Al_{6.25}$ (shown in red asterisk), along with 240 test data.

A follow-on work of this study will be to test and predicting the mechanical properties of HEAs which will require atomistic studies on HEA. On surveying the wealth of literature in the arena of molecular dynamics simulation, the EAM potential currently available for 16 elements namely, Cu, Ag, Au, Ni, Pd, Pt, Al, Pb, Fe, Mo, Ta, W, Mg, Co, Ti, Zr was tested by the authors and found robust to predict HEAs mechanical properties reliably[80]. In keeping this momentum for the purpose of traceability, we have taken the same alloy used in this MD study for the purpose of ML validation in this work as well ($Ni_{25}Cu_{18.75}Fe_{25}Co_{25}Al_{6.25}$).

Conclusion

This study attempts to develop a novel high entropy alloy through comparison of prior literature using robust machine learning algorithms. An effort in this area will strengthen the materials discovery research to guide the initiatives at the forefront of Materials-4.0. Major conclusions from this study can be summarised as:

a. An imbalanced dataset involving synthetic data merged into the experimental data can lead to spurious outcomes when feeding to the machine learning algorithms. An attempt like this (which has been routinely done in literature) can although help achieve higher accuracy from the model, but it can compromise the quality of prediction, particularly, while inferring complex phases of high entropy alloys (HEAs). Our novel work using machine learning revealed that it is possible to make reliable predictions to infer phase information of an HEA merely by using five crucial parameters (Valence electron concentration (VEC), Electronegativity difference ($\Delta\chi$), Mixing entropy (ΔS_{mix}), Atomic size difference (δ), and Mixing enthalpy (ΔH_{mix})). One must however be cautious of using selectively screened input experimental data to feed the ML algorithm.

b. The performance of ML models was assessed using accuracy, precision, recall, f1-score, ROC-AUC score and tenfold cross-validation scores. Across, K-nearest neighbours (V-KNN), support vector machine (V-SVM), decision tree classifier (V-DTC), random forest classifier (V-RFC) and XGBoost (V-XGB), Random Forest Classifier (V-RFC) model performed the best in correctly predicting the phase of an alloy as solid-solutions (FCC, BCC, FCC + BCC) or MIP which denotes the presence of either pure IM compounds (such as α, β, σ, L12, L21, C15, C15, C36 Laves) or mixture of IM + SS phases (such as FCC + IM, FCC + BCC + α, BCC + IM, FCC + α + β, BCC + Laves, BCC1 + BCC2 + C15 Laves, BCC + β-ordered BCC, FCC + CoMo2Ni-type IM, FCC + IM etc.), with an average accuracy of 84%, ROC-AUC score of 0.9649, tenfold cross-validation mean score of 0.9315. Thus, V-RFC model can be used for predicting phases of new HEAs as solid-solution (FCC, BCC, FCC + BCC) or MIP (Mixture of Intermetallics phases). This claim was reinforced by comparing the V-RFC predicted phases with experimental phases reported recently for the newly developed HEAs, where V-RFC correctly predicted solid solution phases (BCC and FCC) for 2 refractory HEAs[10], and 2 precious metal HEAs[63] respectively. The phase of 3d-transition metal HEA ($Al_{0.5}CrCuNiV$)[62] was also correctly predicted as MIP, as per the considered assumption, however the actual phase contained 1FCC + 2BCC + ordered B2. Note that our algorithm worked robustly in predicting solid-solution phases, and complex multiphase HEAs as MIP, but it was found limited in interpreting the number and types of phases present in a complex multiphase HEA.

c. Although there are few studies reporting higher accuracy from the models using synthetic data, we showed that this can lead to inaccurate predictions. For instance, care must be taken while extracting the data from mixed manufacturing routes and tackling an imbalanced dataset. This becomes clear from the fact that although the ANN model used in Bakr et al.[28] study achieved an accuracy of 93.4% but could not correctly predicted the existence of amorphous phase. Hence, proving the fact that even after achieving 93.4% of accuracy, their model resulted in erroneous predictions while treating the imbalanced data. Also. the recall, precision, and F1-score for amorphous (AM) phase were not defined clearly. It was also acknowledged by Risal et al.[51] that the "ML algorithms usually do not perform well for imbalanced dataset" and reported 92.31% accuracy by using oversampling method to balance out the minority class data by polluting it with synthetic data. Accordingly, we explored the use of SMOTE-Tomek link to resample our dataset in support of testing our claim, using RFC model (ST-RFC). An average accuracy of 92% was observed on augmented data of 1392 instances (1200 + 192) for ST-RFC model. Although, a great increment in accuracy was observed, but it could not yield better phase predictions.

d. Using the robust RFC algorithm developed in this work, we report the development of a novel HEA with its composition $Ni_{25}Cu_{18.75}Fe_{25}Co_{25}Al_{6.25}$. The peaks from X-ray diffraction revealed an FCC structure in corroboration with the ML predictions.

Data availability

The data is available from https://gitfront.io/r/user-6296136/ErgmsuZSHXiG/Phase-prediction-of-HEAs-private-share/blob/HEA%20dataset-1200%20instances-NEW.xlsx. The source of data was from the literature[30–32].

Code availability

The code used for training and testing various machine learning algorithms: K-nearest neighbour, Support Vector Machine, Decision Tree Classifier, Random Forest Classifier, and XGBoost are available from https://gitfront.io/r/user-6296136/ErgmsuZSHXiG/Phase-prediction-of-HEAs-private-share/ .The code for model interrogation (that supports the findings of this study) can be made available upon reasonable request.

Received: 2 November 2022; Accepted: 13 March 2023
Published online: 23 March 2023

References

1. George, E. P., Raabe, D. & Ritchie, R. O. High-entropy alloys. *Nat. Rev. Mater.* **4**(8), 515–534 (2019).
2. Ye, Y. *et al.* High-entropy alloy: Challenges and prospects. *Mater. Today* **19**(6), 349–362 (2016).
3. Pickering, E. J. & Jones, N. G. High-entropy alloys: A critical assessment of their founding principles and future prospects. *Int. Mater. Rev.* **61**(3), 183–202 (2016).
4. Katiyar, N. K. *et al.* A perspective on the catalysis using the high entropy alloys. *Nano Energy* **88**, 106261 (2021).
5. Li, Z. *et al.* Metastable high-entropy dual-phase alloys overcome the strength–ductility trade-off. *Nature* **534**(7606), 227–230 (2016).
6. Cantor, B. Multicomponent high-entropy Cantor alloys. *Prog. Mater Sci.* **120**, 100754 (2021).
7. Cantor, B. Multicomponent and high entropy alloys. *Entropy* **16**(9), 4749–4768 (2014).
8. Murty, B.S., et al., *High-entropy alloys* (Elsevier, 2019).
9. Katiyar, N. K., Goel, G. & Goel, S. Emergence of machine learning in the development of high entropy alloy and their prospects in advanced engineering applications. *Emerg. Mater.* **4**(6), 1635–1648 (2021).
10. Han, Z. *et al.* Microstructures and mechanical properties of TixNbMoTaW refractory high-entropy alloys. *Mater. Sci. Eng., A* **712**, 380–385 (2018).
11. Pan, Y. *et al.* New insights into the methods for predicting ground surface roughness in the age of digitalisation. *Precis. Eng.* **67**, 393–418 (2021).
12. Jose, R. & Ramakrishna, S. Materials 4.0: Materials big data enabled materials discovery. *Appl. Mater. Today* **10**, 127–132 (2018).
13. Lederer, Y. *et al.* The search for high entropy alloys: A high-throughput ab-initio approach. *Acta Mater.* **159**, 364–383 (2018).

14. Sun, X. *et al.* Phase selection rule for Al-doped CrMnFeCoNi high-entropy alloys from first-principles. *Acta Mater.* **140**, 366–374 (2017).
15. Liu, X. *et al.* Monte Carlo simulation of order-disorder transition in refractory high entropy alloys: A data-driven approach. *Comput. Mater. Sci.* **187**, 110135 (2021).
16. Gao, M. C. *et al.* Computational modeling of high-entropy alloys: Structures, thermodynamics and elasticity. *J. Mater. Res.* **32**(19), 3627–3641 (2017).
17. Wu, M. *et al.* CALPHAD aided eutectic high-entropy alloy design. *Mater. Lett.* **262**, 127175 (2020).
18. Pyzer-Knapp, E. O. *et al.* Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *NPJ Comput. Mater.* **8**(1), 1–9 (2022).
19. Schmidt, J. *et al.* Recent advances and applications of machine learning in solid-state materials science. *NPJ Comput. Mater.* **5**(1), 83 (2019).
20. Ourmazd, A. Science in the age of machine learning. *Nat. Rev. Phys.* **2**(7), 342–343 (2020).
21. Kailkhura, B. *et al.* Reliable and explainable machine-learning methods for accelerated material discovery. *NPJ Comput. Mater.* **5**(1), 1–9 (2019).
22. Cai, J. *et al.* Machine learning-driven new material discovery. *Nanoscale Adv.* **2**(8), 3115–3130 (2020).
23. Jiang, L. *et al.* Formation rules of single phase solid solution in high entropy alloys. *Mater. Sci. Technol.* **32**(6), 588–592 (2016).
24. Guo, S. Phase selection rules for cast high entropy alloys: an overview. *Mater. Sci. Technol.* **31**(10), 1223–1230 (2015).
25. Borg, C. K. H. *et al.* Expanded dataset of mechanical properties and observed phases of multi-principal element alloys. *Sci. Data* **7**(1), 430 (2020).
26. Gorsse, S. *et al.* Database on the mechanical properties of high entropy alloys and complex concentrated alloys. *Data Brief* **21**, 2664–2678 (2018).
27. Himanen, L. *et al.* Data-driven materials science: Status. *Challenges, Perspect.* **6**(21), 1900808 (2019).
28. Bakr, M., Syarif, J. & Hashem, I. A. T. Prediction of phase and hardness of HEAs based on constituent elements using machine learning models. *Mater. Today Commun.* **31**, 103407 (2022).
29. Martin, P. *et al.* HEAPS: A user-friendly tool for the design and exploration of high-entropy alloys based on semi-empirical parameters. *Comput. Phys. Commun.* **278**, 108398 (2022).
30. Miracle, D. B. & Senkov, O. N. A critical review of high entropy alloys and related concepts. *Acta Mater.* **122**, 448–511 (2017).
31. Machaka, R. *et al.* Machine learning-based prediction of phases in high-entropy alloys: A data article. *Data Brief* **38**, 1 (2021).
32. Precker, C.E.G.C., Andrea, Landín, M., *Materials for design open repository. high entropy alloys* (2021).
33. Zhang, Y. *et al. Solid-solution phase formation rules for multi-component alloys.* **10**(6), 534–538 (2008).
34. Takeuchi, A. & Inoue, A. Quantitative evaluation of critical cooling rate for metallic glasses. *Mater. Sci. Eng., A* **304**, 446–451 (2001).
35. Sheng, G. & Liu, C. T. Phase stability in high entropy alloys: Formation of solid-solution phase or amorphous phase. *Prog. Nat. Sci. Mater. Int.* **21**(6), 433–446 (2011).
36. Katrutsa, A. & Strijov, V. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Syst. Appl.* **76**, 1–11 (2017).
37. Cuartas, M. *et al.* Machine learning algorithms for the prediction of non-metallic inclusions in steel wires for tire reinforcement. *J. Intell. Manuf.* **32**(6), 1739–1751 (2021).
38. Fan, C. *et al.* A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Front. Energy Res.* **9**, 1 (2021).
39. Famili, A. *et al.* Data preprocessing and intelligent data analysis. *Intell. Data Anal.* **1**(1), 3–23 (1997).
40. Cunningham, P. & Delany, S. J. k-Nearest Neighbour Classifiers—A Tutorial. *ACM Comput. Surv.* **54**(6), 128 (2021).
41. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**(1), 21–27 (1967).
42. Salam Patrous, Z. *Evaluating XGBoost for User Classification by using Behavioral Features Extracted from Smartphone Sensors*, in *TRITA-EECS-EX* (2018).
43. Jakkula, V. *Tutorial on support vector machine (svm)*. School of EECS, Washington State University. **37**(2.5), 3 (2006).
44. Song, Y. Y. & Lu, Y. Decision tree methods: Applications for classification and prediction. *Shanghai Arch Psychiatry* **27**(2), 130–135 (2015).
45. Izza, Y., Ignatiev, A., Marques-Silva, J. On explaining decision trees. arXiv preprint arXiv:2010.11034 (2020).
46. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32 (2001).
47. Biau, G. Analysis of a random forests model. *J. Mach. Learn. Res.* **13**(1), 1063–1095 (2012).
48. Chen, T., *et al.* Xgboost: extreme gradient boosting. *R package version* 0.4–2. **1**(4), 1–4 (2015).
49. Akosa, J. *Predictive accuracy: A misleading performance measure for highly imbalanced data*. in *Proceedings of the SAS global forum* (2017).
50. Luque, A. *et al.* The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recogn.* **91**, 216–231 (2019).
51. Gu, Q., Zhu, L., & Cai, Z. *Evaluation measures of the classification performance of imbalanced data sets*. In *International symposium on intelligence computation and applications* (Springer, 2009).
52. Kulkarni, A., Chong, D. & Batarseh, F. A. Foundations of data imbalance and solutions for a data democracy. In *data democracy* 83–106 (Elsevier, 2020).
53. Thölke, P., *et al.* Class imbalance should not throw you off balance: Choosing classifiers and performance metrics for brain decoding with imbalanced data. bioRxiv (2022).
54. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
55. Hossin, M. & Sulaiman, M. N. A review on evaluation metrics for data classification evaluations. *Int. J. Data Min. Knowl. Manag. Process* **5**(2), 1 (2015).
56. Bewick, V., Cheek, L. & Ball, J. Statistics review 13: Receiver operating characteristic curves. *Crit. Care* **8**(6), 1–5 (2004).
57. Varpa, K. *et al.* Applying one-vs-one and one-vs-all classifiers in k-nearest neighbour method and support vector machines to an otoneurological multi-class problem. In *User Centred Networked Health Care* 579–583 (IOS Press, 2011).
58. Yacouby, R., & Axman, D. *Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models*. In *Proceedings of the first workshop on evaluation and comparison of NLP systems* (2020).
59. Goutte, C., & Gaussier, E. *A probabilistic interpretation of precision, recall and F-score, with implication for evaluation*. in *European conference on information retrieval* (Springer, 2005).
60. Risal, S. *et al.* Improving phase prediction accuracy for high entropy alloys with machine learning. *Comput. Mater. Sci.* **192**, 110389 (2021).
61. Batista, G.E., Bazzan, A.L., & Monard, M.C. *Balancing Training Data for Automated Annotation of Keywords: a Case Study*. in *WOB* (2003).
62. Yi, J. *et al.* A novel Al0 5CrCuNiV 3d transition metal high-entropy alloy: Phase analysis, microstructure and compressive properties. *J. Alloys Compounds* **846**, 156466 (2020).
63. Sohn, S. *et al.* Noble metal high entropy alloys. *Script. Mater.* **126**, 29–32 (2017).
64. Mandal, P. *et al.* Phase prediction in high entropy alloys by various machine learning modules using thermodynamic and configurational parameters. *Metals Mater. Int.* **1**, 1–15 (2022).
65. Zhu, W. *et al.* Phase formation prediction of high-entropy alloys: a deep learning study. *J. Market. Res.* **18**, 800–809 (2022).

66. Jaiswal, U. K. *et al.* Machine learning-enabled identification of new medium to high entropy alloys with solid solution phases. *Comput. Mater. Sci.* **197**, 110623 (2021).
67. Krishna, Y. V., Jaiswal, U. K. & Rahul, M. Machine learning approach to predict new multiphase high entropy alloys. *Script. Mater.* **197**, 113804 (2021).
68. Pei, Z. *et al.* Machine-learning informed prediction of high-entropy solid solution formation: Beyond the Hume-Rothery rules. *NPJ Comput. Mater.* **6**(1), 1–8 (2020).
69. Lee, S. Y. *et al.* Deep learning-based phase prediction of high-entropy alloys: Optimization, generation, and explanation. *Mater. Des.* **197**, 109260 (2021).
70. Al-Shibaany, Z.Y.A., *et al. Deep learning-based phase prediction of high-entropy alloys*. In *IOP Conference Series: Materials Science and Engineering*. 2020. IOP Publishing.
71. Chau, N. H. *et al. Phase prediction of multi-principal element alloys using support vector machine and bayesian optimization* (Springer International Publishing, 2021).
72. 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).
73. Zhou, Z. *et al.* Machine learning guided appraisal and exploration of phase design for high entropy alloys. *NPJ Comput. Mater.* **5**(1), 1–9 (2019).
74. Agarwal, A., & Prasada Rao, A. Artificial intelligence predicts body-centered-cubic and face-centered-cubic phases in high-entropy alloys. *Jom*. **71**(10), 3424–3432 (2019).
75. Li, Y. & Guo, W. Machine-learning model for predicting phase formations of high-entropy alloys. *Phys. Rev. Mater.* **3**(9), 095005 (2019).
76. Choudhury, A., *et al.* Structure prediction of multi-principal element alloys using ensemble learning. *Eng. Comput.* (2019).
77. Huang, W., Martin, P. & Zhuang, H. L. Machine-learning phase prediction of high-entropy alloys. *Acta Mater.* **169**, 225–236 (2019).
78. Islam, N., Huang, W. & Zhuang, H. L. Machine learning for phase selection in multi-principal element alloys. *Comput. Mater. Sci.* **150**, 230–235 (2018).
79. Tancret, F. *et al.* Designing high entropy alloys employing thermodynamics and Gaussian process statistical analysis. *Mater. Des.* **115**, 486–497 (2017).
80. Fan, P. *et al. Uniaxial pulling and nano-scratching of a newly synthesized high entropy alloy.* **10**(11), 111118 (2022).

Acknowledgements

SNJ would like to thank the Ministry of Education, Government of India. NKK would like to thank the Royal Society UK for the Newton Fellowship award (NIF/ R1\191571). SG would like to acknowledge the financial support from the UKRI via Grants No. EP/S036180/1 and EP/T024607/1, feasibility study awards to LSBU from the UKRI National Interdisciplinary Circular Economy Hub (EP/V029746/1) and Transforming the Foundation Industries: a Network+ (EP/V026402/1) and Transforming the Partnership award from the Royal Academy of Engineering (TSP1332).

Author contributions

S.S.: writing original draft, data extraction and machine learning model development. N.K.: data analysis. S.G.: supervision, reviewing and editing. S.N.J.: supervision, reviewing and resources.

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/s41598-023-31461-7.

Correspondence and requests for materials should be addressed to S.G. or S.N.J.