



# Phase Prediction in High Entropy Alloys by Various Machine Learning Modules Using Thermodynamic and Configurational Parameters

Pritam Mandal<sup>1</sup> · Amitava Choudhury<sup>2</sup> · Amitava Basu Mallick<sup>1</sup> · Manojit Ghosh<sup>1</sup>

Received: 2 February 2022 / Accepted: 8 April 2022 / Published online: 9 June 2022  
© The Author(s) under exclusive licence to The Korean Institute of Metals and Materials 2022

## Abstract

The purpose of this investigation is to predict the different phases present in various high entropy alloys and subsequently classify their crystal structure by various machine learning algorithms using five thermodynamic, configurational and electronic parameters, which are considered to be essential for the formation of high entropy alloy phases. Proper prediction of phases and crystal structures can eventually trace the properties of high entropy alloy, which is crucial for selecting the suitable elements for designing. The model has been developed by various machine learning (ML) algorithms using an experimental dataset consisting of 322 different HEAs, including 258 solid solution (SS), 31 intermetallic (IM), and 33 amorphous (AM) phases. The ML algorithms include (1) K-nearest neighbours (KNN), (2) support vector machines (SVM), and (3) logistic regression (LR), (4) decision tree (DT), (5) random Forest (RF), and (6) gaussian naive bayes classifier. Among them, both DT and SVM algorithms exhibited the highest accuracy of 93.84% for phase prediction. Crystal structure classification of SS phases was also done using a dataset consisting of 194 different HEAs data, including 76 body centered cubic (BCC), 61 face centered cubic (FCC) and 57 mixed body-centered and face-centered cubic (BCC + FCC) crystal structures and found that the SVM algorithm shows the highest accuracy of 84.32%. The effect of the parameters on determining the accuracy of the model was calculated and tracking the role of individual parameters in phase construction was also attempted.

**Keywords** High entropy alloy · Machine learning · Computational metallurgy · Phase prediction · Support vector machine

## Abbreviations

$\delta$	Atomic size difference	$R$	The gas constant
$\Delta\chi$	Electronegativity difference	$H_{ij}$	Enthalpy of atomic pairs of the $i$ -th and $j$ -th elements
$\Delta H_{\text{mix}}$	Mixing enthalpy	$Y$	The predicted output for logistic regression
$\Delta S_{\text{mix}}$	Mixing entropy	$b_0$	The bias or intercept term for logistic regression
VEC	Valence electron concentration	$b_1$	The coefficient for the single input value ( $x$ ) for logistic regression
$c_i$	Atomic concentrations of the $i$ th element	$\Omega$	Combining effects of mixing entropy and mixing enthalpy
$n$	Total number of species in a high entropy alloy		
$\chi_i$	Pauling electronegativity		
$r_i$	Radius of $i$ -th elements		
$\bar{\chi}$	The averaged Pauling electronegativity		
$\bar{r}$	Averaged atomic radius		

✉ Amitava Choudhury  
a.choudhury2013@gmail.com

✉ Manojit Ghosh  
mghosh.metal@faculty.iiests.ac.in

<sup>1</sup> Department of Metallurgy and Materials Engineering, Indian Institute of Engineering Science and Technology, Shibpur, Howrah, West Bengal 711103, India

<sup>2</sup> Department of Computer Science and Engineering, Pandit Deendayal Energy University, Gujrat, Gandhinagar, India

## 1 Introduction

Alloying has been conventionally used to provide desirable properties to materials. Traditional alloys entail the addition of a relatively small quantity of secondary elements to a principal matrix. For a couple of decades, a new alloying approach concerning the combinations of multiple principal elements in high concentration has been in trend to create new materials, called high-entropy alloys (HEAs) [1]. These alloys are generally defined using composition-based and entropy-based approaches. Composition-based HEAs are

classified as being comprised of five or more principal elements added in equimolar ratios, having the concentration of each element between 5 and 35 at%. [2]. The entropy-based approach, on the other hand, deals with the configurational entropy, which should be greater than  $1.5R$  ( $\Delta S_{\text{config}} > 1.5R$ ), where  $R$  is the universal gas constant [3]. HEAs have developed an enormous interest of the scientific community through the pioneering work by Cantor et al. [4], due to their exceptional mechanical and chemical properties, including high hardness and ductility [5], high-temperature strength [6], antioxidant capacity [7] and good wear resistance [8], those are absent in traditional metal alloys. Certain high-entropy alloys have already been proven to possess surprising properties, exceeding those of the conventional alloys [1]. Interestingly, the discovery of new outstanding high-entropy alloys is still in progress, making this subject a burning interest for researchers. Identifying the possible reasons for the exhibition of such extraordinary properties and simultaneously carrying this knowledge to set up strategies for developing new alloys with desired properties: remained the two most interesting issues to be addressed by the scientific community [7]. In order to accomplish those concerns, several efforts have been made utilizing density functional theory (DFT), thermodynamic modelling and, molecular dynamics (MD) calculations to guide the alloy design [9]. For example, Miracle et al. [10] used the CALPHAD approach to predict the formation of phases of HEAs and observed that the predicted number of stable stages is twice, or more, than those obtained via experiments. At the same time, DFT and MD calculations generally focus on phase stability, solidification behaviour, and crystallization kinetics of HEAs [9].

The work cited in the above literature is widely used by various researchers to simulate the atomistic structure of alloys. One of the major challenges of these tools is much more expensive in terms of computational time, complexity and efficacy. On the other hand, the current trends in materials research are inclined to machine learning (ML)-based algorithms, which are generally used in prediction and classification problems. ML is a broad class of data-driven algorithms used to make inferences and classifications from predetermined prior observed data. In materials research, the multipurpose learning and generalizability of ML algorithms offer a swift and cost-effective alternative approach to traditional design, analysis, and predictions. With adequate data from previous experiments, ML models can be constructed to discover new trends predict material characteristics such as promising alloy candidates, microstructures, physical and chemical properties [11]. Over the last few years, ML, one of the supreme data processing techniques, has offered adequate reliability for designing new materials and prediction of microstructure and behaviour of the materials. Gong et al. [12] sorted superheavy elements based on ML,

and established the relationship between atomic data with the class of elements. Huang and Islam et al. [13] have also predicted the phases of HEAs based on distinct ML methods such as artificial neural network, k-nearest neighbors, Support vector machines and concluded that the trained artificial neural network model gives the best accuracy and thus the most useful in predicting the formation of new HEAs. Islam et al. got an accuracy of above 80% for phase prediction of HEAs. Agarwal et al. [14] predicted the phase of HEAs by adaptive neuro fuzzy interface system (ANFIS) adopting two discrete approaches. In the first method, elements compositions were used as the input parameters and for the second method, six crucial thermodynamic and configurational parameters were used as input parameters and they obtained 84.21 and 80% prediction accuracy, respectively. Zhou et al. [15] have made a sensitivity assessment comparison between 13 non-identical design parameters on the basis of ANN model results. Various primary researches on materials design have been made by ML algorithms. Nevertheless, there is vast scope for improvement in the parameter selection for new alloy design and fabrication, the relationship established between parameter and property and the accuracy enhancement of models [16, 17]. The consciousness that real-world issues do not follow a fixed path to figure them yet they are unpredictable and random forms the core of Artificial Intelligence (AI). Implicitly, for each new problem, there must be a different approach to solve them as in the case of HEAs. The arbitrariness in the phase formation mechanism for HEAs does not obey a specific rule leading to the absence of any standard conventional function to reveal the whole phase formation mechanism. Fairly, the phase formation is quite dynamic in nature. Hence, a technique that can encapsulate the uncertainty of that complication will not be clarified by fixed rules. In lieu, a decision-making framework and ML tool seem to be compatible for probing the HEAs phase formation [14].

In recent years, numerous research investigations have been done to predict the phase formation of HEAs by using different ML algorithms [18–20] with the help of fundamental parameters of constituent elements based on theoretical and practical collected data. Nevertheless, the prospect of predicting the phase stability from the fundamental parameter of the constituent elements has still remained attractive. The computational data processing technology, which basically tries to discover complex predictive correspondence among different variables using antecedent data to track the properties of new desired materials, has already become a conspicuous route for material design. The key towards exhibiting superior properties or properties as per demand primarily depends on the phases of the HEA, namely amorphous (AM), solid solution (SS), and intermetallic (IM). For example, SS structure is regarded as the most important contributor to the superior qualities of HEAs [19].

On the other hand, the presence of a brittle intermetallic phase can degrade some quality of the alloy, like ductility properties [21]. At the same time, single-phase FCC HEAs are normally ductile but have low strength, whereas single-phase BCC HEAs have high strength, but it is typically brittle in nature [22]. However, dual-phase HEAs can have a nice combination of ductility and strength [23]. Therefore, an accurate phase prediction, beyond any doubt, is crucial for guiding the selection of a combination of elements to form a HEA with intended properties. The formation of phases depends on several factors, including composition, initial phases, fabrication route, history of heat treatment, etc. [24, 25] Among these controlling factors, composition and fabrication process are the most crucial ones deserving special attention. In this work, with the premise of ensuring the independence of parameters, six common ML algorithms, namely (1) K-nearest neighbors (KNN), (2) support vector machines (SVM), (3) logistic regression (LR), (4) decision tree (DT), (5) random Forest (RF), (6) gaussian naive bayes classifier were applied to predict the phase selection of the SS, IM, and AM phases and crystal structure determinations of SS phase in HEAs. The ultimate goal is to identify the best suited ML model for the discovery and future design of new HEAs.

## 2 Dataset and Modeling

### 2.1 Data Collection and Analysis

The dataset used in this work was established based on collected data from the literature. The data selection has been made arbitrarily based on some open literature sources [14, 18, 26–32]. No category division has been aimed in the selection process. Experimental data consisting of 322 different HEAs including 258 SS, 31 IM, and 33 AM phases was assembled from literature, which consists of five thermodynamic and configurational parameters  $\{VEC, \Delta\chi, \delta, \Delta H_{mix} \text{ (kJ}\cdot\text{mol}^{-1}), \Delta S_{mix} \text{ (J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1})\}$  of HEA along with their respective phases (AM, IM, SS). For crystal structure classification of SS phases, a different dataset was constructed consisting of 194 different HEAs data, including 76 BCC, 61 FCC and 57 BCC + FCC crystal structures with their respective thermodynamic and configurational parameters.

#### 2.1.1 Thermodynamic & Configurational Parameters Calculation

The following equations are used to obtain the numerical input values of the five features:

$$VEC = \sum_{i=1}^n c_i VEC_i \quad (1)$$

$$\Delta\chi = \sqrt{\sum_{i=1}^n c_i (\chi_i - \bar{\chi})^2} \quad (2)$$

$$\delta = 100 \times \sqrt{\sum_{i=1}^n c_i \left(1 - \frac{r_i}{\bar{r}}\right)^2} \quad (3)$$

$$\Delta S_{mix} = -R \sum_{i=1}^n c_i \ln c_i \quad (4)$$

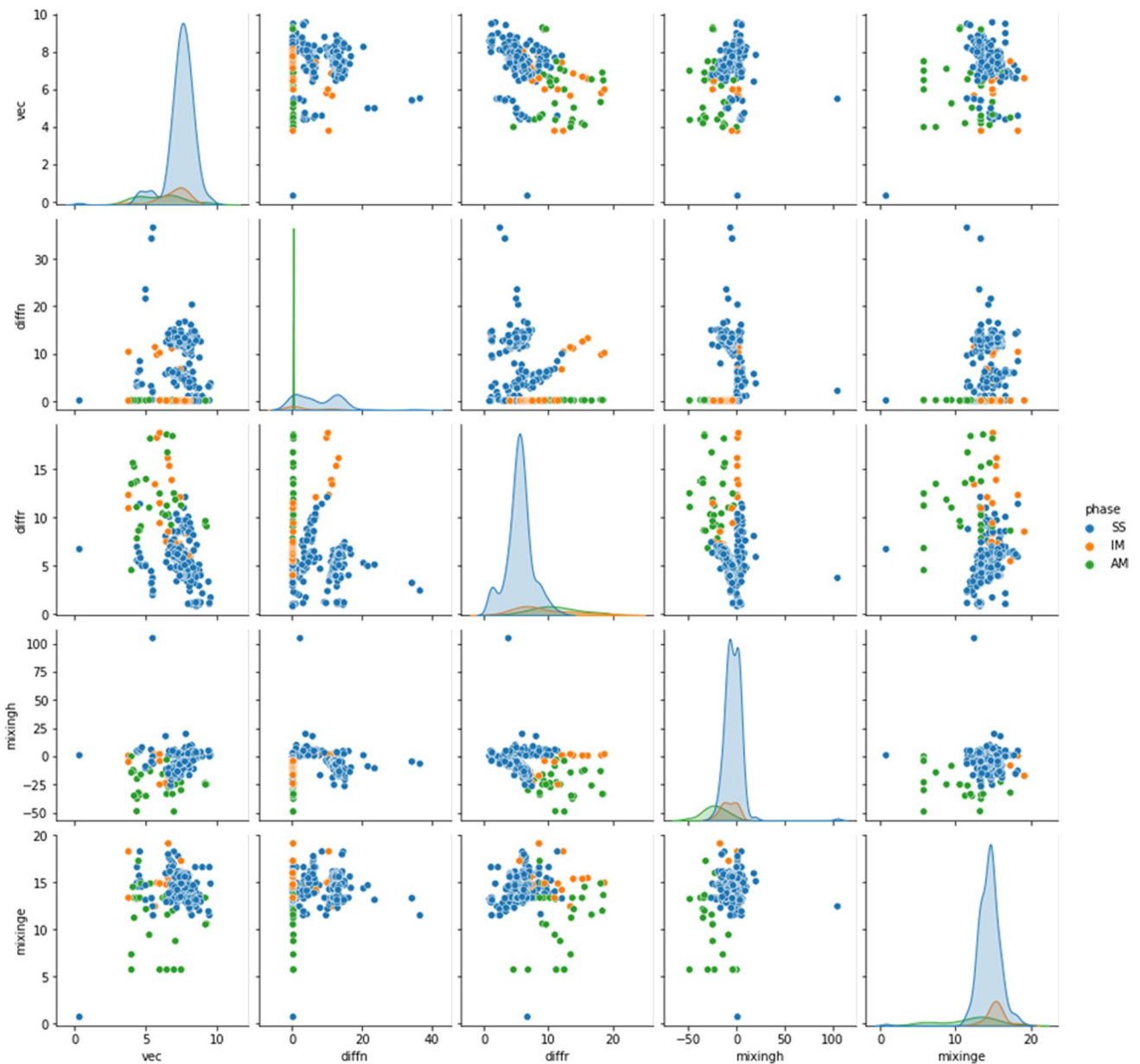
$$\Delta H_{mix} = \sum_{i=1, i < j}^n 4H_{ij} c_i c_j \quad (5)$$

where  $VEC_i$  is valence electron concentrations and  $c_i$  is atomic concentrations of the  $i$ -th element.  $n$  is the total number of species in a HEA.  $\chi_i$  signify the Pauling electronegativity of  $i$ -th element and  $r_i$  denotes the radius of  $i$ -th element. The averaged Pauling electronegativity  $\bar{\chi}$  can be calculated as  $\bar{\chi} = \sum_{i=1}^n c_i \chi_i$  and averaged atomic radius  $\bar{r}$  is calculated as  $\bar{r} = \sum_{i=1}^n c_i r_i$ .  $H_{ij}$  is the enthalpy of atomic pairs of the  $i$ -th and  $j$ -th elements evaluated by Takeuchi and Inoue with the Miedema method [7, 23].

#### 2.1.2 Correlation Among Features

Accessing meaningful information from a dataset is always challenging. Python has its own packages which use the panda library to get these essential details. Finding the correlation across all features is crucial to understanding the link between the input parameters. The graphical representation of such feature correlation is represented using a scatter matrix, generated from the python seaborn package.

Figure 1 depicts the link between the features for the dataset used in phase selection. Similarly, for crystal structure datasets, the correlation among various features is represented in Fig. 2. From both Figs. 1 and 2, it has been observed that the diagonal matrix projected inseparable features which reveals that all features are important to determine phase and crystal structure prediction. The dots indicate not just the values of individual data points, but also patterns when the data is viewed as a whole. Overplotting occurs when data points overlap to such an extent that no relationships between points and variables can be addressed. It might be difficult to understand how closely data points are packed when there are many of them together in a small space.



**Fig. 1** Scatter plot showing the correlation among various features for the phase selection

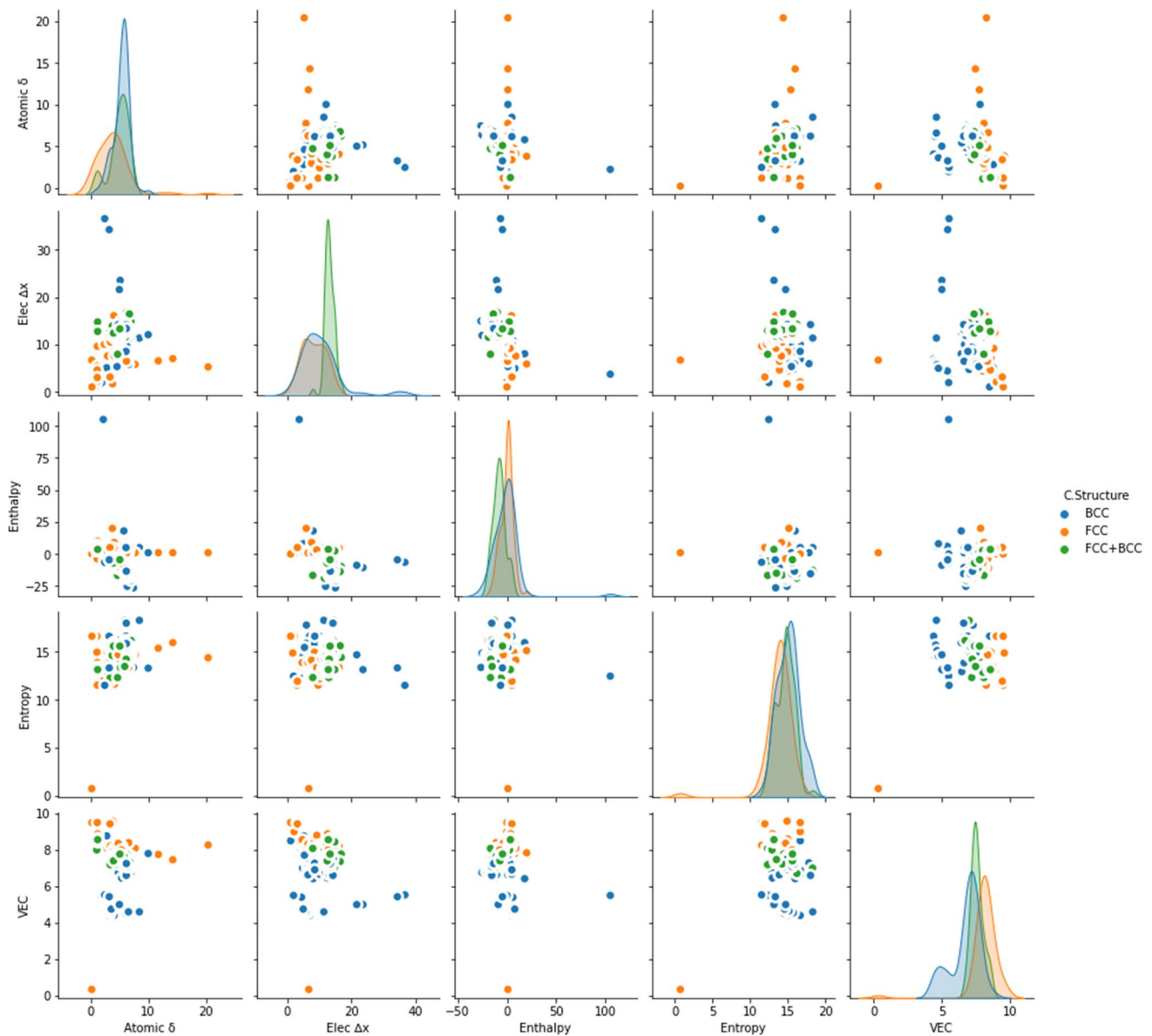
## 2.2 Architecture of Different ML Module

Scientific and technological progress has been a critical driver for uprisings in modelling, designing, and enhancing the prediction of properties of an alloy. Over the last few years, many ML methods have been brought forward in different fields of the research sector. As each ML algorithm has its own individual working principle and applicability, the divergence in prediction accuracy may be experienced. Basically, ML utilizes programmed algorithms which can acquire and analyse inserted data to predict the output values within a certain domain. Each time fresh data is fed into the

algorithms, they try to learn and optimize their functions to improve the performance, developing intelligence gradually [33]. Following are the descriptions of a few popular ML algorithms used for the present problem.

### 2.2.1 Logistic Regression (LR)

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. Logistic regression focuses on estimating the probability of an event occurring based on the previous data provided. Mathematically, it uses a binary dependent variable with



**Fig. 2** Scatter plot showing the correlation among various features for the crystal structure selection

only two values that are 0 and 1, to express the outcomes [34]. Logistic regression, uses the sigmoid function as a mathematical function used to map the predicted values to probabilities [35]. It is basically a S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits. When the sigmoid function returns a value from 0 to 1, generally, a threshold value is taken, such as 0.5. If the sigmoid function returns a value greater than or equal to 0.5, it is taken as 1, and if the sigmoid function returns a value less than 0.5, it is taken as 0. Input values ( $x$ ) can collaborate linearly, utilizing coefficient values or weights to predict an output value ( $y$ ). An example of a logistic regression equation is shown in Eq. 6

$$y = \frac{e^{(b_0+b_1x)}}{1 + e^{(b_0+b_1x)}} \quad (6)$$

where  $y$  is the predicted output,  $b_0$  is the bias or intercept term and  $b_1$  is the coefficient for the single input value ( $x$ ). Each column in your input data has an associated  $b$  coefficient (a constant real value) that must be learned from your training data. In general, logistic regression is used when the dependent variable is categorical. In the proposed work, multiclass classification is addressed using logistic regression using the one Vs Rest (OVR) method. In the OVR method, when work is done with a class, the class is denoted by 1 and the rest of the classes becomes 0.



### 2.2.2 K-Nearest Neighbors (KNN)

KNN algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. KNN algorithm anticipates how likely a data point is to be a member of one category or another. It fundamentally looks at the surrounding of a single data point to determine the actual category of that point. For instance, if one point is on a grid and the algorithm needs to decide the belonging category (Category A or category B, for example) for that particular data point, at first the algorithm will look at the data points surroundings to see in which group the majority of the points are in. Another way it works is by finding the distances between a query and all the examples in the data, selecting the specified number of examples (K) closest to the query, then voting for the most frequent label (in the case of classification) or averages the labels (in the case of regression). For example, suppose we have a new data point, and we need to put it in the required category, as shown in Fig. 3

To choose the number of neighbours the value of k equals to 5 has been considered followed by calculating the Euclidean distance among the data points (Fig. 4).

The nearest neighbours can be evaluated by calculating the Euclidean distance, as there are three nearest neighbours in category A and two nearest neighbours in category B, shown in Fig. 5. Involvement of more number of neighbours make category A as the most favourable one. Therefore, this new data point must belong to category A.

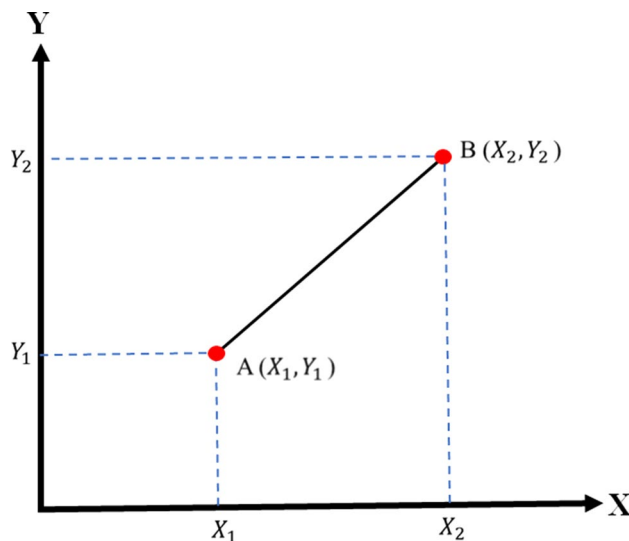


Fig. 4 Euclidean distance between the data points A and B  $(\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2})$

### 2.2.3 Support Vector Machines (SVM)

SVM algorithms are supervised learning models which use prior data for classification and regression analysis. The principle of SVM is to establish a classification hyper-plane as a decision surface to maximize the isolated edge between positive and negative examples but to find an optimal classification surface to minimize the error of all training samples from the optimal classification surface [36]. SVM chooses the extreme points or vectors that help in creating the hyper-plane. These extreme cases are called support vectors, and hence algorithm is termed as SVM [37]. Consider Fig. 6, in

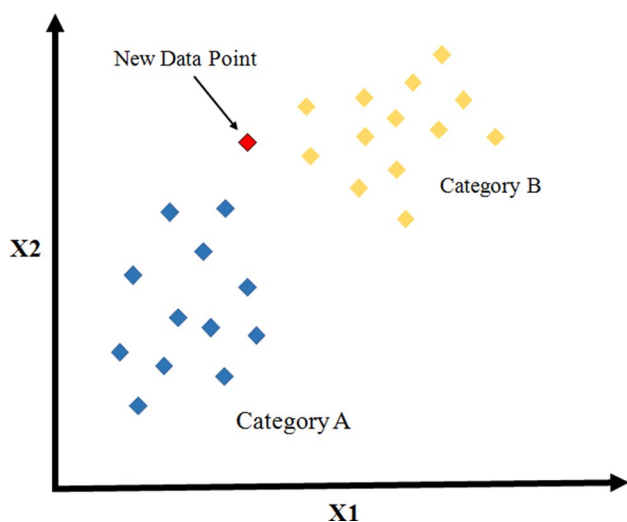


Fig. 3 A new data point needs to be categorized in its preferable category

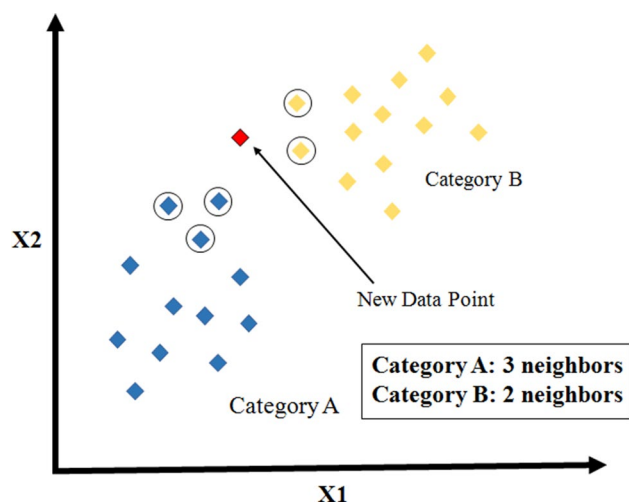


Fig. 5 The new data point has five closest neighbors from which three of them are from category A and the rest two are from category B

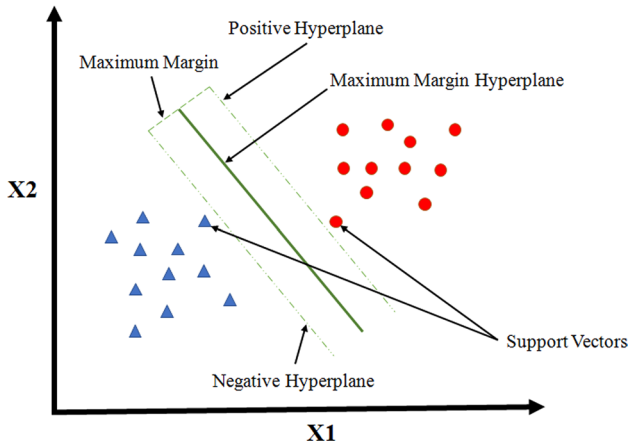


Fig. 6 Two different categories (red and blue) are classified using a decision boundary or hyper-plane in SVM [38]

which there are two different categories (red and blue) that are classified using a decision boundary or hyperplane. Let it be considered that there are two independent variables X1, X2 and one dependent variable, which is either a blue triangle or a red circle.

2.2.4 Decision Tree (DT)

The decision tree is a graph to represent choices and their consequences in order of a tree. The nodes in the graph indicate an event or preference and the edges of the graph specify the decision rules or conditions. Every tree comprises nodes and branches. Each node represents attributes in

a category that is to be classified and each branch represents a value that the node can take [39].

2.2.5 Random Forests (RF)

Random forests is also termed as ‘random decision forests, an ensemble learning method that uses a combination of multiple algorithms to generate better results for classification and regression. Each individual classifier is weak, but when combined with others, it can produce excellent results. The algorithm starts with a ‘decision tree’, and an input is entered at the top. It then travels down the tree, with data being segmented into smaller and smaller sets based on specific variables. A random forest structure based on a decision tree mechanism is shown below (Fig. 7):

2.2.6 Naïve Bayes Classifier

It is a Bayes theorem-based classification technique in which it predicts the output on the basis of the probability of an event [40]. To be specific, a Naïve Bayes Classifier assumes that the presence of a particular feature in a category is discrete from the presence of other features. It is used in text classification, which includes a high-dimensional training dataset. Naïve Bayes Classifier predominantly used for classification and clustering purposes depending on the occurring conditional probability. The probability of Y occurring for a selected dataset X can be expressed as follows in equation no.7, where Y considered as a random variable

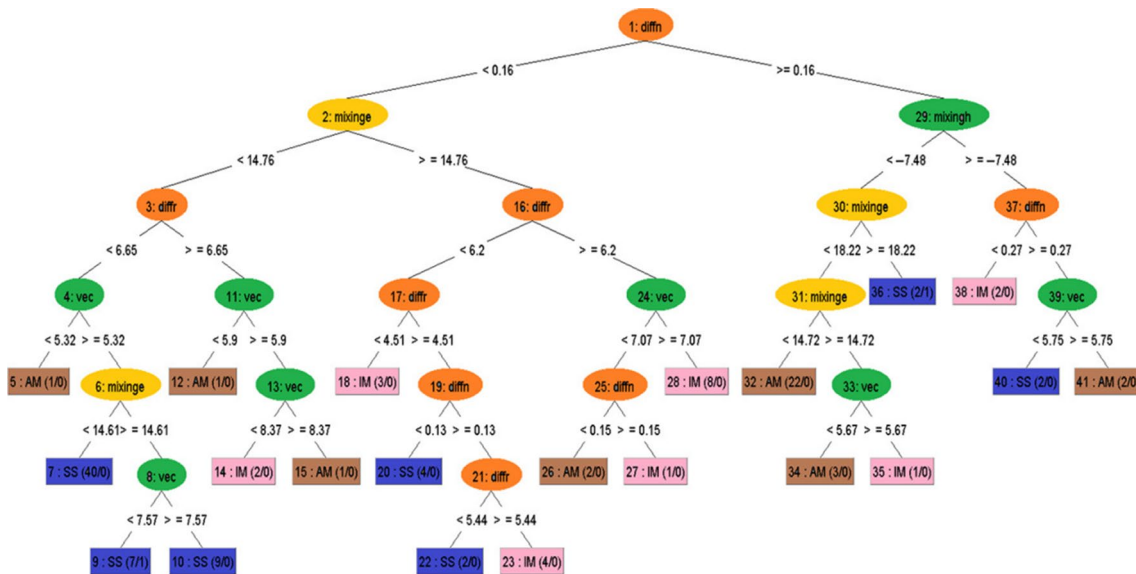


Fig. 7 A random forest structure based on a decision tree mechanism. Reprinted with permission, from [20]

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (7)$$

where  $P(Y|X)$  represents “posterior distribution” and  $P(Y)$  is the probability distribution of the variable  $Y$  in the absence of  $X$ , which is also indicated as the “prior distribution”.

---

If input:  $VEC = 8.27$ ,  $\Delta\chi = 0.11$ ,  $\delta = 4.17$ ,  $\Delta H_{mix} = -1.52$ ,  $\Delta S_{mix} = 14.63$   
Then Output = “Solid Solution (SS)” (9)

---

$P(X|Y)$  designate “likelihood function” that assess the probability of  $X$  occurring for a given  $Y$ . Generally,  $P(X)$  can be calculated by applying the law of total probabilities as follows in equation no.8 [41, 42]

$$P(X) = \int_Y P(X|Y)dY \quad (8)$$

### 2.3 Methodology

The first step to set up an ML framework is to make the input parameters available. In this work, the prediction of phases and crystal structure was accomplished by utilizing the thermodynamic and configurational parameters as input

---

If input:  $VEC = 7.2$ ,  $\Delta\chi = 12.06$ ,  $\delta = 5.44$ ,  $\Delta H_{mix} = -12.32$ ,  $\Delta S_{mix} = 13.38$ ;  
Then Output = “Body Centered Cubic (BCC)” (10)

---

parameters, which are atomic size difference, enthalpy of mixing, configurational entropy, electronegativity difference, and valence electronic concentration (VEC). These five parameters should ideally come up with amazing perceptiveness into different prospects of microstructure evolution in a HEA. This is evident from the facts that enthalpy of mixing governs the thermodynamic aspects of phase formation as atomic size difference, electronegativity difference and VEC considered for the behaviour of atoms on an atomic scale. Similarly, the configurational entropy describes the synergy between every probable combination of elemental pairs in the system [14]. Zhang et al. [43] proposed that atomic size difference and enthalpy of mixing have a crucial role in the phase development of HEAs.

#### 2.3.1 Prediction of Phases

For determination of phases in HEA by thermodynamic and configurational parameters, the value of those parameters should be known at the outset. The five input parameters ( $VEC = 8.27$ ,  $\Delta\chi = 0.11$ ,  $\delta = 4.17$ ,  $\Delta H_{mix} = -1.52$ ,

$\Delta S_{mix} = 14.63$ ) for FeCoNiCrCuAl0.5 system, when fed into the ML model, the model would interpret those values in the form of an equation (Eq. 9). Implementation of decision making by the ML algorithm follows, leading to the predicted output. The output for the FeCoNiCrCuAl0.5 system is SS and this has also been confirmed from the available dataset. Therefore,

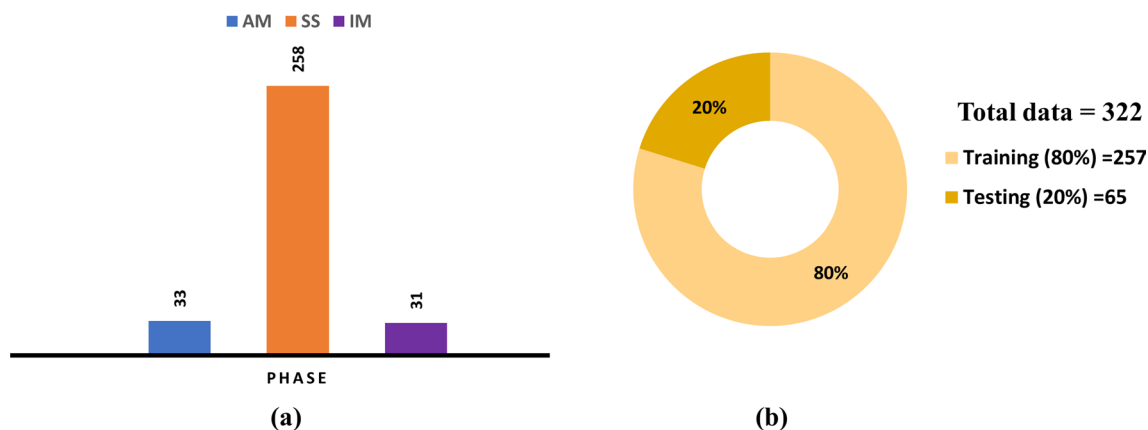
#### 2.3.2 Prediction of Crystal Structure

For the crystal structure determination of SS phases of respective HEA, the thermodynamic and configurational parameters of that respective alloy should be known first. For example, if previously known AlCoCrFeNi system is considered, whose thermodynamic and configurational parameter are  $VEC = 7.2$ ,  $\Delta\chi = 12.06$ ,  $\delta = 5.44$ ,  $\Delta H_{mix} = -12.32$ ,  $\Delta S_{mix} = 13.38$ . After putting those input values into the ML model, the model will interpret those values in the form of an equation (Eq. 10). Then the decisions making process is employed by the ML algorithm. The predicted output which model gives is a BCC crystal structure, which means, according to model algorithm, the crystal structure of AlCoCrFeNi system should be a solid BCC, which has been confirmed from the dataset. Therefore,

## 3 Results and Discussions

There are large numbers of possible HEAs with different phases and properties. It would be impractical to look for those desired alloys by the trial-and-error method by examining every alloy composition. For choosing the proper HEA with desired properties, it is essential to select the correct combinations of elements that influence phase formation. So, an accurate phase prediction is necessary to track the properties and behaviour of HEAs. Six common ML algorithms were applied in the current research, namely: (1) K-nearest neighbours (KNN), (2) support vector machines (SVM), (3) logistic regression (LR), (4) decision tree (DT), (5) random forest (RF), (6) gaussian naive bayes classifier to predict the phase selection (SS, IM, or AM phases) and crystal structure determination (BCC, FCC or BCC + FCC) in HEAs with the help of five thermodynamic and configurational parameters ( $VEC$ ,  $\Delta\chi$ ,  $\delta$ ,  $\Delta H_{mix}$ ,  $\Delta S_{mix}$ ). In addition, the impact of independent parameters on the accuracy of the model and phase formation was also investigated. Furthermore, the ultimate aim of this work is to sort out the most





**Fig. 8** **a** Distribution of various phases used in the current work, **b** the distribution of training and testing data within the dataset

**Table 1** ML algorithms along with their phase prediction accuracy (%)

Machine learning algorithm	Prediction accuracy
logistic Regression (LR)	90.76
K nearest neighbours (KNN)	92.30
Support Vector Machine (SVM)	93.84
Decision Tree (DT)	93.84
Random Forest (RF)	90.76
Gaussian naive bayes (NB)	90.76

suitable ML model for the future design and discovery of new HEAs.

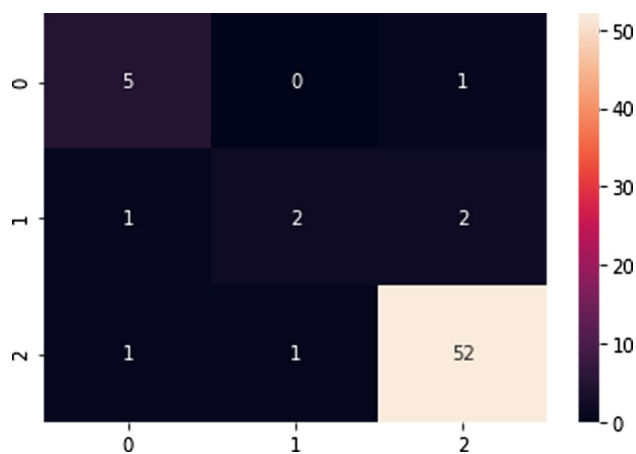
### 3.1 Results of Phase Prediction

A total dataset of 322 different alloy systems consisting of five thermodynamic and configurational parameters (VEC,  $\Delta\chi$ ,  $\delta$ ,  $\Delta H_{mix}$ ,  $\Delta S_{mix}$ ) along with their respective phases (258 SS, 33 AM and 31 IM phases) has been used in this work.

The entire dataset was split arbitrarily into training and testing dataset, maintaining a protocol mentioned in Fig. 8b. This distribution and a class representation of various phases have been represented in Fig. 8a for better visual understanding. Python programming was executed on Jupyter notebook by applying those ML algorithms to predict the phase selection and calculated the accuracy percentage. The accuracy percentages are listed below in Table 1

For this database, the SVM and decision tree ML algorithm show the highest accuracy of 93.84% for phase predictions even though the LR, KNN, and RF algorithm also record a decent amount of accuracy (above 90%).

In order to further compare the models, a gaussian naive bayes classifier algorithm was used to predict the phase



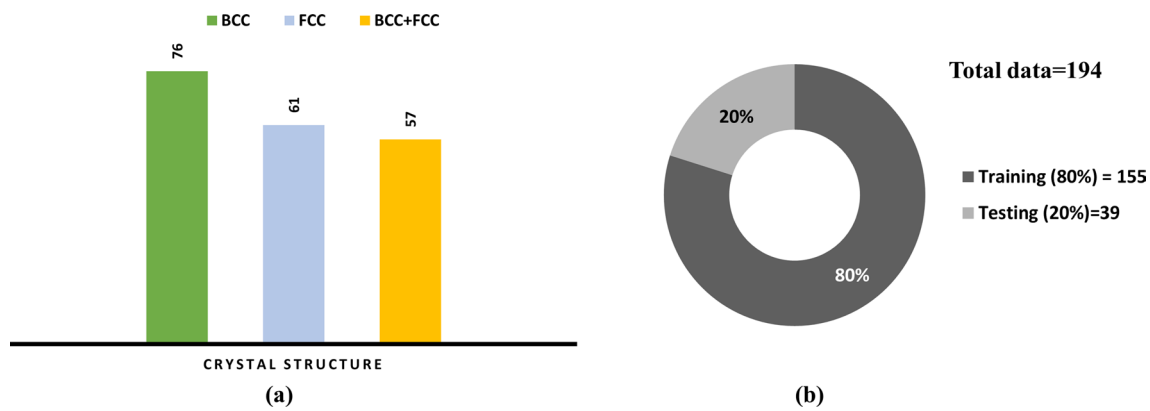
**Fig. 9** Confusion Matrix of gaussian naive bayes classifier for phase prediction

formation and accuracy percentage. It gives 90.76% accuracy for phase predictions. A confusion matrix, as shown in Fig. 9, was applied, which can summarize the number of true and false predictions made by a classifier.

Out of 322 data sets, 257 were used for training the model and the rest for testing. From Fig. 9, it can be seen from the confusion matrix that the accurately predicted regions of the matrix are matrix cell (0×0) for IM, (1×1) for AM and (2×2) for SS. So, IM = 5 (0×0), AM = 2 (1×1), SS = 52 (2×2). So, out of total test data of 65, The accurately Predicted data has been (5 + 2 + 52) 59, leading to an accuracy of (59/65) 90.76%.

### 3.2 Results of Crystal Structure Prediction (SS phase)

For crystal structure prediction, the complete dataset was arbitrarily divided into training and testing datasets



**Fig. 10** **a** Distribution of different crystal structures in SS phase mentioned in proposed work **b** The division of training and testing data within the dataset

**Table 2** ML algorithms and their crystal structure prediction accuracy (in %)

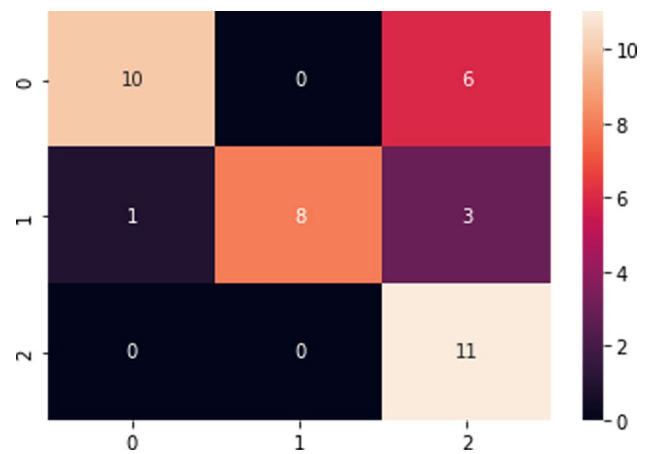
Machine learning algorithm	Prediction accuracy
Logistic regression (LR)	66.66
K nearest neighbours (KNN)	69.23
Support vector machine (SVM)	84.32
Decision tree (DT)	66.66
Random forest (RF)	71.79
Gaussian naive bayes (NB)	64.10

according to the strategy shown in Fig. 10b. This distribution and a class representation of different phases have been represented in Fig. 10a for better visual understanding.

The preciseness of crystal structure prediction was evaluated using python programming on Jupyter notebook by applying those ML algorithms. The results are listed below in Table 2.

For this database, the SVM algorithm shows the prediction accuracy of 84.32%, which is the best result for crystal structure determination of SS phases compared to all other ML algorithms. For further comparison of models, a gaussian naive bayes classifier algorithm was utilized. It gives 64.10% accuracy for crystal structure predictions. The confusion matrix is also applied, summarizing the number of true and false predictions made by a classifier.

The dataset contains 194 data of which 155 data was used for training and the rest 39 data for testing. From Fig. 11, it can be seen from the confusion matrix that the accurately predicted regions of the matrix are matrix cell (0×0) for BCC, (1×1) for FCC and (2×2) for BCC + FCC. So, BCC = 8 (0×0), FCC = 7 (1×1), BCC + FCC = 10 (2×2). So, the total test data becomes 39. As the accurately



**Fig. 11** Confusion Matrix of gaussian naive Bayes classifier algorithm for crystal structure prediction

Predicted data is (8 + 7 + 10) 25, the accuracy percent becomes (25/39) 64.10%.

Precision and recall are two extremely important model evaluation metrics. While precision refers to the percentage of results that are relevant, recall refers to the percentage of total relevant results correctly classified by the algorithm. Unfortunately, it is not possible to maximize both these metrics at the same time, as one comes at the cost of another. In addition, *F1* score is a harmonic mean of precision and recall. Both precision and recall are important, it is recommended to choose a model which maximizes the *F1* score. The performance of the proposed work is obtained by the *F1* score. The calculated *F1* score for each model is shown in Table 3. The following expression is used to calculate *F1*-score:

**Table 3** Recall, Precision, and F1 score value of their respective ML algorithms

ML model	For Phase	For Crystal Structure
Linear regression (LR)	Recall: 0.6827160493827161 Precision: 0.8099415204678362 f1: 0.7409068044189061 Accuracy: 0.908	Recall: 0.6799242424242425 Precision: 0.7023172905525846 f1: 0.6909393768424961 Accuracy: 0.667
K nearest neighbors (KNN)	recall: 0.7382716049382716 precision:0.9210526315789475 f1: 0.8195950973097859 Accuracy: 0.923	recall: 0.6988636363636364 precision: 0.7585784313725491 f1: 0.7274977067726647 Accuracy: 0.692
Support vector machine (SVM)	recall: 0.8049382716049384 precision: 0.8988095238095237 f1: 0.8492878893748437 Accuracy: 0.938	recall: 0.7916666666666666 precision: 0.8510526315789475 f1: 0.8197969543147208 Accuracy: 0.843
Decision tree (DT)	recall: 0.8049382716049384 precision: 0.848989898989899 f1: 0.8263774377302013 Accuracy: 0.938	recall: 0.6704545454545454 precision: 0.6846930846930848 f1: 0.6774990128916996 Accuracy: 0.667
Random forest (RF)	recall: 0.6827160493827161 precision: 0.7488095238095238 f1: 0.7142370201537972 Accuracy: 0.908	recall: 0.7146464646464645 precision: 0.7238095238095239 f1: 0.7191988095835213 Accuracy: 0.718
Gaussian naive bayes (NB)	recall: 0.7320987654320987 precision: 0.7754689754689754 f1: 0.753160026139064 Accuracy: 0.908	recall: 0.6641414141414143 precision: 0.7253968253968254 f1: 0.693418949870479 Accuracy: 0.641

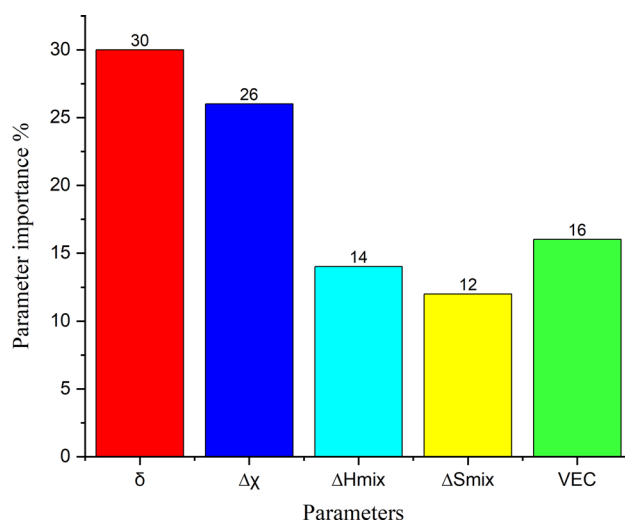
$$F1 - Score = 2 * \frac{Precision \times Recall}{Precision + Recall} \quad (11)$$

$$\text{where, Precision} = \frac{TP}{TP + FP} \text{ and } \text{Recall} = \frac{TP}{TP + FN}$$

$TP$  = True positive;  $FP$  = False Positive,  $FN$  = False Negative.

This section gives an overview of various ML techniques that have been used in the proposed work to achieve higher accuracy in order to predict the phases and the structure from HEAs. Though the discussed techniques are very popular, they having their own limitations due to which each model gives a different level of accuracy. LR is a parametric model that can work fine on a linear dataset but cannot handle outliers. Whereas KNN is non-parametric model and performs better than LR when the data has a very high signal-to-noise ratio SNR, but cannot handle outliers properly. On the other hand SVM handles outliers which makes the model more realistic and as in the presented work the number of input features are very less; it cannot be handled by LR or KNN, but higher accuracy can be achieved using SVM or DT model. Hence, in this paper, a model wise comparison has been studied to understand the performance of each model using the dataset. In this case, it is observed that the SVM gives the best performance over other ML algorithms. On a few occasions, accuracy may be the same for two or more models but the measured  $F1$  score value is not favorable except

for the support vector machine. Although higher accuracy and  $F1$  score have been achieved in this work, still there are some limitations of the proposed SVM based model. The first constraint was observed that when the data set contains more noise, such as overlapping target classes, the proposed model does not perform well. Secondly, the proposed model underperforms when the number of features for each data point exceeds the number of training data samples.

**Fig. 12** Parameter importance percentage in phase prediction accuracy

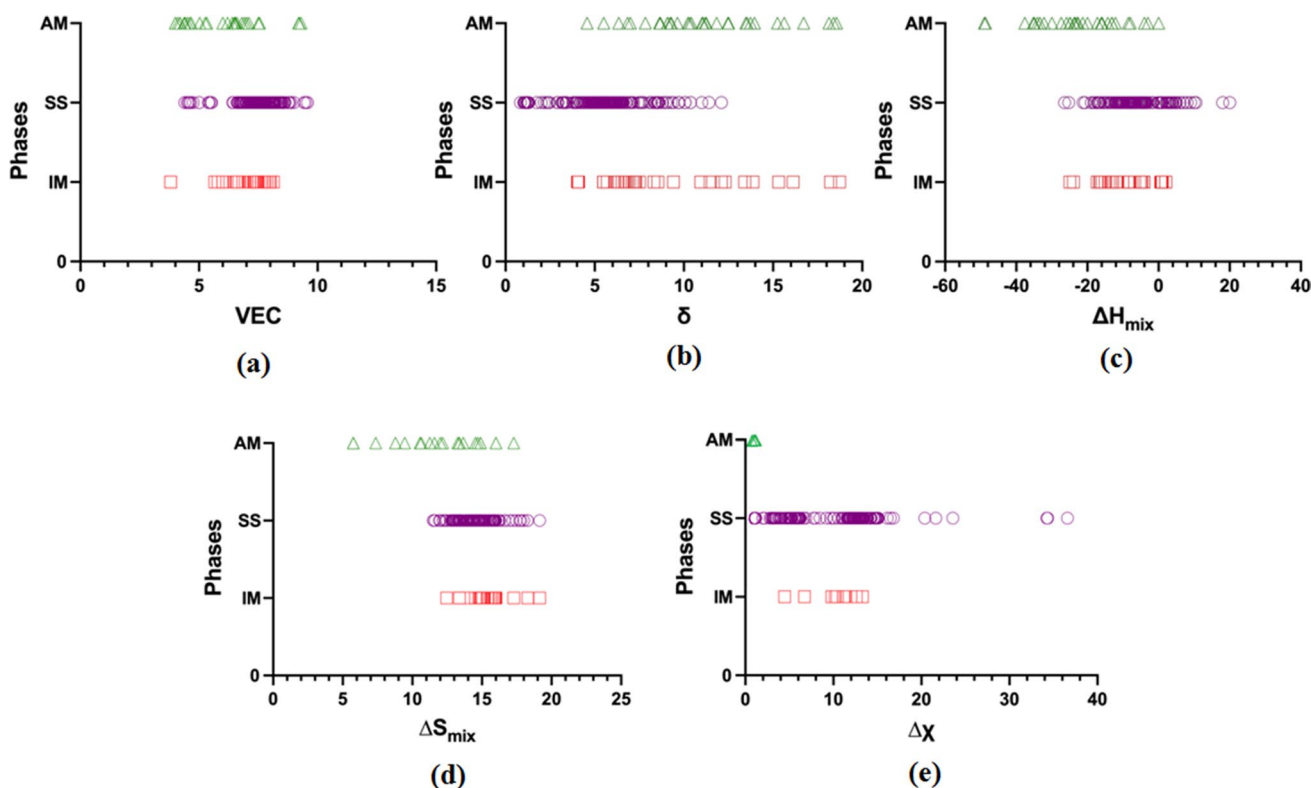
### 3.3 Influence of Parameters on the Accuracy Variance of the Model

As a standard practice, the parameter exhibiting the strongest influence on the model accuracy has been identified by removing each of the five parameters one at a time from the model. Then percentage decrease in the accuracy was calculated. Atomic size difference ( $\delta$ ) has exhibited the maximum influence on the accuracy variance of the model among the five parameters, shown to Fig. 12. The importance of input features is assessed using SVM, which has the maximum capacity for phase and crystal structure prediction in this study. From Fig. 10, it is clearly visible that removing any of these features can certainly affect the accuracy of the model, which again validates the selection of ideal features for this study. In addition, the decrease of accuracy follows the order:  $\delta > \Delta\chi > VEC > \Delta H_{mix} > \Delta S_{mix}$ . By contrast, the  $\delta$ ,  $\Delta\chi$ , and VEC features seem to play the most important role in determining the phase selection, providing guidance for designing HEAs by preferably considering these three features. Contradiction exists in reporting the most crucial parameter in controlling the accuracy of the model. VEC has been reported as the one by Chawdhury et al. [20] and Islam et al. [13] while  $\delta$  by Huang et al. [7].

### 3.4 Role of Parameters in Phase Formation

The selection of effective and suitable features of alloys is a crucial stage to achieve meaningful and accurate results. Here the features stand for the basic characteristics of alloy or the constituent elements of the system. The thermodynamic, configurational and electronic parameters such as  $\delta$ ,  $\Delta\chi$ ,  $\Delta H_{mix}$ ,  $\Delta S_{mix}$ , VEC has a crucial role in phase formation. The empirical rules for the solid solution phase formation are summarized as  $\Omega \geq 1.1$  and  $\delta \leq 6.6\%$ . In spite of that, the discrimination of phases is not good enough, especially for the mixture of different phases [44]. Attempts have been made to establish a relationship among each of the parameters to the output, four out of five parameters were kept fixed and the remaining one has been varied one at a time.

Figure 13 (a,b,c,d,e) show how the five thermodynamic and configurational parameters reflect the effect on the phase stability in HEAs, especially to reveal the rules governing the formation of AM, SS and IM. From Fig. 13, it is visible that SS phases form under the conditions of VEC is slightly positive ( $10 > VEC > 4$ ),  $\delta$  is small ( $1\% < \delta < 10\%$ ),  $\Delta S_{mix}$  is high ( $11 < \Delta S_{mix} < 18$ ), and  $\Delta\chi$  ranges between  $2 < \Delta\chi < 15$ , and  $\Delta H_{mix}$  is primarily negative yet positive in some cases ( $-20 < \Delta H_{mix} > 10$ ). Comparatively, AM phases generally form when VEC is slightly



**Fig. 13** Relationship established between each input parameters to the output phase **a** valence electron concentration (VEC), **b** atomic size difference ( $\delta$ ), **c** mixing enthalpy ( $\Delta H_{mix}$ ), **d** configurational entropy ( $\Delta S_{mix}$ ), **e** electronegativity difference ( $\Delta\chi$ )

positive ( $8 > \text{VEC} > 4$ ) and when  $\delta$  ranges between 4 and 20%,  $\Delta S_{\text{mix}}$  is high ( $5 < \Delta S_{\text{mix}} < 16$ ) and  $\Delta\chi$  is nearly 0 and  $\Delta H_{\text{mix}}$  is mostly negative ( $-35 < \Delta H_{\text{mix}} < 8$ ). In general, IM forms when valence electron concentration (VEC) is slightly positive ( $5.5 > \text{VEC} > 8.5$ ) and when atomic size difference ( $\delta$ ) is comparatively high ( $4\% < \delta < 20\%$ ), configurational entropy ( $\Delta S_{\text{mix}}$ ) is very high ( $13 < \Delta S_{\text{mix}} < 16$ ) and electronegativity difference ( $\Delta\chi$ ) also comparatively very high and mixing enthalpy ( $\Delta H_{\text{mix}}$ ) is mostly negative ( $-15 < \Delta H_{\text{mix}} < 0$ ).

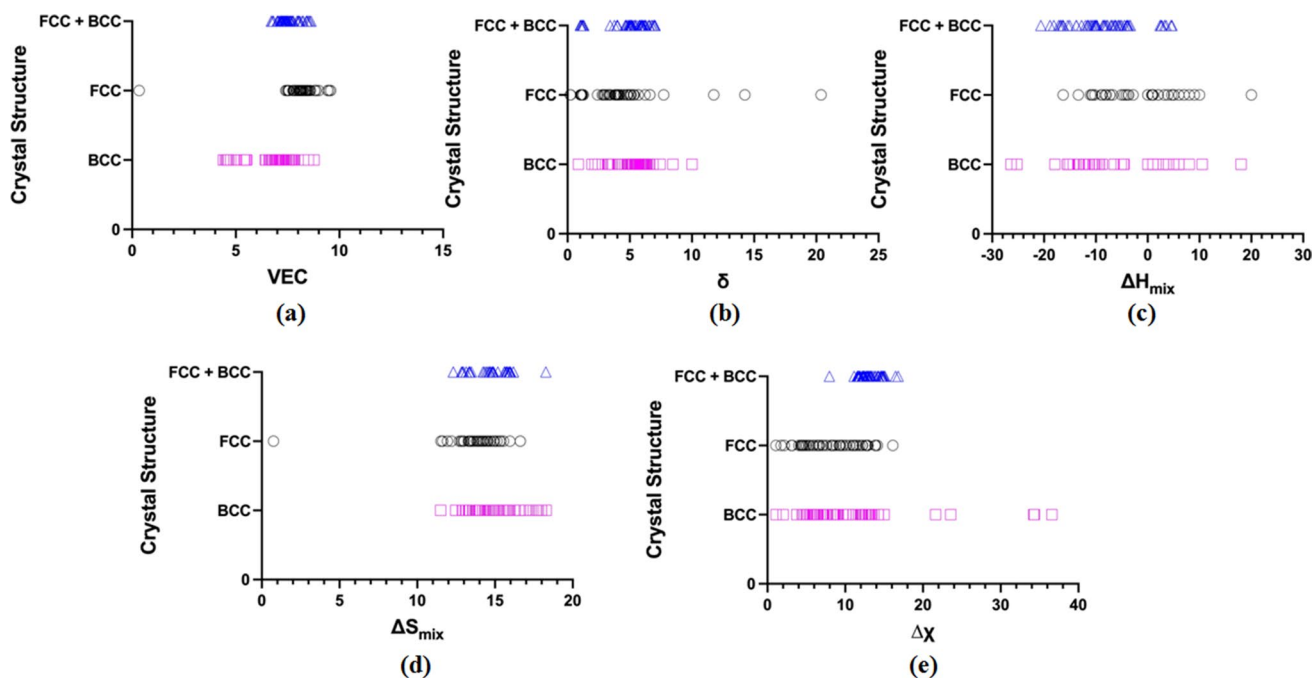
### 3.5 Role of Parameters in Crystal Structure Formation

An efficient way to identify the role of parameters in the formation of the crystal structure (BCC, FCC or FCC + BCC) of solid solution phase, a single parameter out of five has been varied at a time, keeping the remaining parameters constant.

Figure 14a–e show how the five thermodynamic and configurational parameters reflect the effect of the crystal structure stability in HEAs, especially to reveal the rules governing the formation of FCC, BCC and FCC + BCC solid solution phases. From Fig. 14, it can be seen that FCC crystal structure forms when VEC is slightly positive ( $10 > \text{VEC} > 7$ ),  $\delta$  is in the range of 5% to 15%,  $\Delta S_{\text{mix}}$  is high ( $12 < \Delta S_{\text{mix}} < 17$ ), electronegativity difference ( $\Delta\chi$ ) is less ( $2 < \Delta\chi < 15$ ), and  $\Delta H_{\text{mix}}$  has a range of values between -10 to 10. Comparatively, BCC crystal structure generally

forms when VEC is slightly positive ( $8 > \text{VEC} > 6$ ),  $\delta$  is in the range of 5%–13%,  $\Delta S_{\text{mix}}$  is high ( $13 < \Delta S_{\text{mix}} < 19$ ),  $\Delta\chi$  varies between 4 to 15 and,  $\Delta H_{\text{mix}}$  is typically negative ( $-15 < \Delta H_{\text{mix}} < 10$ ). In general, mixed FCC + BCC crystal structure forms when VEC is in the range between 5.5 to 8.5,  $\delta$  is comparatively high ( $12\% < \delta < 16\%$ ),  $\Delta S_{\text{mix}}$  has a range of values between 13 to 17,  $\Delta\chi$  also comparatively very high ( $11 < \Delta\chi < 15$ ), and  $\Delta H_{\text{mix}}$  is mostly negative ( $-20 < \Delta H_{\text{mix}} < 0$ ) in comparison with other phases.

The parameters under discussion have their impact on both phase and crystal structure formation. Below is summary of influence of those parameters used in the present work. The mixing entropy indicates the complexity of a system, and the higher the mixing entropy, the more difficult it is for the system to establish an ordered structure [45]. According to this viewpoint, the high mixing entropy promotes the formation of random SS or partially ordered SS. But it is also understandable from Figs. 13 and 14 that simply meeting the high mixing entropy criteria is not enough to create SS phases in high entropy alloys [32]. On the other hand, amorphous phase formation must be inhibited in order to form solid solution phases, and this is where other requirements, such as atomic size differences and mixing enthalpy come into play. SS phases are favored when atomic size differences are reduced, which is linked with the topological instability aspects [46]. The atomic size mismatch causes pressure on the atoms, resulting in local elastic strain.



**Fig. 14** Relationship established between each input parameters to the output crystal structure **a** valence electron concentration (VEC), **b** atomic size difference ( $\delta$ ), **c** mixing enthalpy ( $\Delta H_{\text{mix}}$ ), **d** configurational entropy ( $\Delta S_{\text{mix}}$ ), **e** electronegativity difference ( $\Delta\chi$ )



The system becomes topologically unstable over a certain critical volume strain, and glass transition may occur. The requirement for mixing enthalpy may be related to cluster formation, as a lower mixing enthalpy supports the creation of chemically ordered clusters and hence does not favor the development of solid solutions. However, the greater negative enthalpy must be combined with the considerable atomic size difference to generate a stable glassy phase; otherwise, intermetallic phases will develop [47]. About controlling the crystal structure VEC has been reported [20, 48] to be the most sensitive parameter. Larger VEC ( $\geq 8$ ) favored the formation of FCC type SS, while smaller VEC ( $< 6.87$ ) favored the formation of BCC-type solid solutions. In contrary, the present investigation highlights the role of  $\delta$  to be the most crucial one for determination of crystal structure, which appears more fundamental.

## 4 Conclusions

For choosing the proper HEA with desired properties, it is important to select the correct combination of elements that has an impact on phase formation. So, tracking the properties and behaviour of an alloy, it is crucial to know the actual phase present. Therefore, correct phase prediction is the crucial part for designing a new alloy. In this paper, phase prediction was carried out using a database consisting of 322 experimental alloy combinations with their thermodynamic and configurational parameters along with the respective phases. Primarily the phase is categorized into three sections that is AM, SS and IM. For better microstructure evolution, further crystal structure classification of SS phase (BCC, FCC, BCC + FCC) was attempted. Six common ML algorithm was implemented for phase and crystal structure prediction. Among them, the SVM and RT shows the highest accuracy (93.84%) for phase prediction and for crystal structure prediction SVM shows the highest accuracy (84.32%). A confusion matrix was used to validate the predicted results. In addition, determination of parameter influence on the model accuracy was carried out independently and tracking of parameter role for phase formation was also attempted. In the future, this work can be extended to implement on the larger dataset to more efficiently predict the correct phases for ease of designing the desired HEAs.

**Acknowledgements** Author Pritam Mandal acknowledges IEST, Shibpur for the support of his Institute PhD fellowship.

## Declarations

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.

## References

1. E.P. George, D. Raabe, R.O. Ritchie, *Nat. Rev. Mater.* **4**, 515 (2019). <https://doi.org/10.1038/s41578-019-0121-4>
2. M.-H. Tsai, J.-W. Yeh, *Mater. Res. Lett.* **2**, 107 (2014). <https://doi.org/10.1080/21663831.2014.912690>
3. S. Wang, *Entropy* **15**, 5536 (2013). <https://doi.org/10.3390/e15125536>
4. B. Cantor, I.T.H. Chang, P. Knight, A.J.B. Vincent, *Mater. Sci. Eng. A* **375–377**, 213 (2004). <https://doi.org/10.1016/j.msea.2003.10.257>
5. Y. Deng, C.C. Tasan, K.G. Pradeep, H. Springer, A. Kostka, D. Raabe, *Acta Mater.* **94**, 124 (2015). <https://doi.org/10.1016/j.actamat.2015.04.014>
6. B. Schuh, F. Mendez-Martin, B. Völker, E.P. George, H. Clemens, R. Pippan, A. Hohenwarter, *Acta Mater.* **96**, 258 (2015). <https://doi.org/10.1016/j.actamat.2015.06.025>
7. W. Huang, P. Martin, H.L. Zhuang, *Acta Mater.* **169**, 225 (2019). <https://doi.org/10.1016/j.actamat.2019.03.012>
8. C.-Y. Hsu, J.-W. Yeh, S.-K. Chen, T.-T. Shun, *Metall. Mater. Trans. A* **35**, 1465 (2004). <https://doi.org/10.1007/s11661-004-0254-x>
9. C. Wen, Y. Zhang, C. Wang, D. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman, Y. Su, *Acta Mater.* **170**, 109 (2019). <https://doi.org/10.1016/j.actamat.2019.03.010>
10. D.B. Miracle, O.N. Senkov, *Acta Mater.* **122**, 448 (2017). <https://doi.org/10.1016/j.actamat.2016.08.081>
11. R. Machaka, *Comp. Mater. Sci.* **188**, 110244 (2021). <https://doi.org/10.1016/j.commatsci.2020.110244>
12. S. Gong, W. Wu, F.Q. Wang, J. Liu, Y. Zhao, Y. Shen, S. Wang, Q. Sun, Q. Wang, *Phys. Rev. A* **99**, 022110 (2019). <https://doi.org/10.1103/PhysRevA.99.022110>
13. N. Islam, W. Huang, H.L. Zhuang, *Comp. Mater. Sci.* **150**, 230 (2018). <https://doi.org/10.1016/j.commatsci.2018.04.003>
14. A. Agarwal, A.K. Prasada Rao, *JOM* **71**, 3424 (2019). <https://doi.org/10.1007/s11837-019-03712-4>
15. Z. Zhou, Y. Zhou, Q. He, Z. Ding, F. Li, Y. Yang, *Npj Comput. Mater.* **5**, 128 (2019). <https://doi.org/10.1038/s41524-019-0265-1>
16. A. Choudhury, S. Pal, R. Naskar, A. Basumallick, *Eng. Computation.* **36**, 1913 (2019). <https://doi.org/10.1108/EC-11-2018-0498>
17. A. Choudhury, *Arch. Comput. Method. E.* **28**, 3361 (2021). <https://doi.org/10.1007/s11831-020-09503-4>
18. B. Chanda, P.P. Jana, J. Das, *Comp. Mater. Sci.* **197**, 110619 (2021). <https://doi.org/10.1016/j.commatsci.2021.110619>
19. Y. Zhang, C. Wen, C. Wang, S. Antonov, D. Xue, Y. Bai, Y. Su, *Acta Mater.* **185**, 528 (2020). <https://doi.org/10.1016/j.actamat.2019.11.067>
20. A. Choudhury, T. Konnur, P.P. Chattopadhyay, S. Pal, *Eng. Computation.* **37**, 1003 (2020). <https://doi.org/10.1108/EC-04-2019-0151>
21. D.J.M. King, S.C. Middleburgh, A.G. McGregor, M.B. Cortie, *Acta Mater.* **104**, 172 (2016). <https://doi.org/10.1016/j.actamat.2015.11.040>
22. O.N. Senkov, S.V. Senkova, C. Woodward, *Acta Mater.* **68**, 214 (2014). <https://doi.org/10.1016/j.actamat.2014.01.029>
23. S. Guo, *Mater. Sci. Technol.* **31**, 1223 (2015). <https://doi.org/10.1179/1743284715Y.0000000018>
24. Y.F. Juan, J. Li, Y.Q. Jiang, W.L. Jia, Z.J. Lu, *Appl. Surf. Sci.* **465**, 700 (2019). <https://doi.org/10.1016/j.apsusc.2018.08.264>
25. B.S. Murty, J.W. Yeh, S. Ranganathan, P.P. Bhattacharjee, in *High-Entropy Alloys*, 2nd edn (Elsevier, Amsterdam, 2019), pp. 103–117
26. J. Xiong, S.-Q. Shi, T.-Y. Zhang, *J. Mater. Sci. Technol.* **87**, 133 (2021). <https://doi.org/10.1016/j.jmst.2021.01.054>
27. M.G. Poletti, L. Battezzati, *Acta Mater.* **75**, 297 (2014). <https://doi.org/10.1016/j.actamat.2014.04.033>

28. X. Yang, Y. Zhang, *Mater. Chem. Phys.* **132**, 233 (2012). <https://doi.org/10.1016/j.matchemphys.2011.11.021>
29. A. Takeuchi, A. Inoue, *Mater. Trans.* **46**, 2817 (2005). <https://doi.org/10.2320/matertrans.46.2817>
30. Y.F. Ye, Q. Wang, J. Lu, C.T. Liu, Y. Yang, *Mater. Today* **19**, 349 (2016). <https://doi.org/10.1016/j.mattod.2015.11.026>
31. S. Gorsse, M.H. Nguyen, O.N. Senkov, D.B. Miracle, *Data Br.* **21**, 2664 (2018). <https://doi.org/10.1016/j.dib.2018.11.111>
32. S. Guo, C.T. Liu, *Prog. Nat. Sci. Mater. Int.* **21**, 433 (2011). [https://doi.org/10.1016/S1002-0071\(12\)60080-X](https://doi.org/10.1016/S1002-0071(12)60080-X)
33. L. Qiao, Y. Liu, J. Zhu, *J. Alloy. Compd.* **877**, 160295 (2021). <https://doi.org/10.1016/j.jallcom.2021.160295>
34. J. Tolles, W.J. Meurer, *JAMA* **316**, 533 (2016). <https://doi.org/10.1001/jama.2016.7653>
35. J. Han, C. Moraga, The influence of the sigmoid function parameters on the speed of backpropagation learning, in *From Natural to Artificial Neural Computation*, ed. by J. Mira, F. Sandoval. International Workshop on Artificial Neural Networks, Torremolinos, 7–9 June 1995. Lecture Notes in Computer Science, vol 930 (Springer-Verlag Berlin, Heidelberg, 1995), pp. 195–201
36. H. Dai, H. Zhang, W. Wang, G. Xue, *Comput.-Aided Civ. Eng.* **27**, 676 (2012). <https://doi.org/10.1111/j.1467-8667.2012.00767.x>
37. C. Cortes, V. Vapnik, *Mach. Learn.* **20**, 273 (1995). <https://doi.org/10.1007/BF00994018>
38. E. García-Gonzalo, Z. Fernández-Muñiz, P.J. Garcia Nieto, A.B. Sánchez, M.M. Fernández, *Materials* **9**, 531 (2016). <https://doi.org/10.3390/ma9070531>
39. B. Mahesh, *Int. J. Sci. Res.* **9**, 381 (2020)
40. S. Chen, G.I. Webb, L. Liu, X. Ma, *Knowl. Based Syst.* **192**, 105361 (2020). <https://doi.org/10.1016/j.knosys.2019.105361>
41. R.R.P.R. Purohit, T. Richeton, S. Berbenni, L. Germain, N. Gey, T. Connolley, O. Castelnau, *Acta Mater.* **208**, 116762 (2021). <https://doi.org/10.1016/j.actamat.2021.116762>
42. H. Rappel, L.A.A. Beex, S.P.A. Bordas, *Mech. Time Depend. Mater.* **22**, 221 (2018). <https://doi.org/10.1007/s11043-017-9361-0>
43. Y. Zhang, Y.J. Zhou, J.P. Lin, G.L. Chen, P.K. Liaw, *Adv. Eng. Mater.* **10**, 534 (2008). <https://doi.org/10.1002/adem.200700240>
44. L. Zhang, H. Chen, X. Tao, H. Cai, J. Liu, Y. Ouyang, Q. Peng, Y. Du, *Mater. Design* **193**, 108835 (2020). <https://doi.org/10.1016/j.matdes.2020.108835>
45. A.L. Greer, *Nature* **366**, 303 (1993). <https://doi.org/10.1038/366303a0>
46. T. Egami, Y. Waseda, *J. Non-Cryst. Solids* **64**, 113 (1984). [https://doi.org/10.1016/0022-3093\(84\)90210-2](https://doi.org/10.1016/0022-3093(84)90210-2)
47. B.S. Murty, J.W. Yeh, S. Ranganathan, in *High-Entropy Alloys* (Butterworth-Heinemann, Oxford, 2014), pp. 37–56
48. S. Guo, C. Ng, J. Lu, C.T. Liu, *J. Appl. Phys.* **109**, 103505 (2011). <https://doi.org/10.1063/1.3587228>

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