



Available online at www.sciencedirect.com
jmr&t
 Journal of Materials Research and Technology
 journal homepage: www.elsevier.com/locate/jmrt



Molecular dynamics simulation and machine learning-based analysis for predicting tensile properties of high-entropy FeNiCrCoCu alloys

Omarelfarouq Elgack ^{a,b,c}, Belal Almomani ^c, Junaidi Syarif ^{a,*},
 Mohamed Elazab ^a, Mohammad Irshaid ^a, Mohammad Al-Shabi ^a

^a Department of Mechanical & Nuclear Engineering, University of Sharjah, 27272, Sharjah, United Arab Emirates

^b Mechanical Engineering Program, University of Sharjah, 27272, Sharjah, United Arab Emirates

^c Nuclear Energy System Simulation and Safety Research Group, Sustainable Energy & Power Systems Research Centre, RISE, University of Sharjah, 27272, Sharjah, United Arab Emirates

ARTICLE INFO

Article history:

Received 22 May 2023

Accepted 3 July 2023

Available online 5 July 2023

Keywords:

ANN

High entropy alloy

Machine learning

Molecular dynamics

Polycrystalline

Tensile properties

ABSTRACT

High entropy alloys (HEAs) attract many researchers due to their unique and desirable properties in comparison to conventional alloys, and their potential for advanced applications. Because of the complexity of designing HEAs, several attempts have been conducted to integrate experimental and computational studies with machine learning (ML) algorithms to predict their mechanical properties. Yet, few studies have considered a set of input parameters including atomic concentrations, grain size, operating temperature, and strain rate. Therefore, this study considers these combined predictors to forecast the tensile properties of FeNiCrCoCu HEAs, including Young's modulus, yield strength, and ultimate tensile strength based on molecular dynamics (MD) and ML algorithms. 918 datasets of polycrystalline HEAs were generated by MD simulations. Some of the MD datasets were selected as representative samples and assessed by checking the isotropy of mechanical properties. Also, the MD simulations provided data that reasonably agreed with previously published results. All the generated datasets were used afterward to train Artificial neural networks (ANN), support vector machine, and Gaussian process regression models. The proposed ANN models revealed the most accurate predictions among the other ML models, and their performances were evaluated on new datasets containing different predictor variables' values that were not used to build the models. It was found that the ANN models were most sensitive to the strain rate predictor variable. The proposed ANN models can assist in guiding the experimental work to optimize the search for HEAs with desired tensile properties.

© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

* Corresponding author.

E-mail address: sjunaidi@sharjah.ac.ae (J. Syarif).

<https://doi.org/10.1016/j.jmrt.2023.07.023>

2238-7854/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

High entropy alloys (HEAs) have been considered attractive materials owing to their superior properties to conventional materials in various aspects and their promising potential in developing sophisticated materials with unique characteristics [1–3]. For instance, $\text{Cu}_{0.5}\text{NiAlCoCrFeSi}$ HEAs have better corrosion resistance than 304-stainless steel [4]. Moreover, the high entropy Cantor alloys that exhibit face-centered cubic (fcc) crystal structures showed the highest yield strength and fracture toughness compared to nickel superalloys, high-strength steels, metallic glasses, and brittle ceramics [5]. The HEAs were initially defined as an alloying strategy in which no primary element exists, but five or more multi-principal elements of varying atomic concentrations between 5 at.% and 35 at.% are mixed to form a single-phase solid solution alloy. Furthermore, the term “high entropy” is assigned to alloys with substantially high configurational entropy of mixing that is affected by the number of elements in the HEA and their respective molar fractions. Moreover, other factors such as the enthalpy of mixing, valence electron concentration, and melting temperature affect the alloys' phase stability and play a role in predicting the phase of the HEA [6–8]. It is well-known that HEAs offer an intricate design space and a vast range of materials. In that regard, extensive experimental works have been implemented to investigate the phase formation, microstructure, and properties of such materials [9–15]. However, experiment-based analyses usually require a trial-and-error approach, which consumes time and resources. Nowadays, machine learning (ML) algorithms are widely applied to lessen the number of experiments and boost the process of searching for desired material properties in a broad design space. Bundela and Rahul [16] developed eight different ML models for microhardness prediction of HEAs with and without the implementation of principal component analysis followed by experimental verification. Bhandari et al. [17] trained random forest regressors to predict the yield strength of refractory HEAs at elevated temperatures. Wen et al. [18] fabricated HEAs with a hardness 10% higher than the highest value in the primary dataset after they built several predictive models. Additionally, Baker et al. [19,20] reported the establishment of several ML algorithms, including artificial neural networks (ANN) for phase prediction as well as the hardness of HEAs with good accuracy. Other studies were done to predict the phase of HEAs using ML since their properties could be directly impacted by the phase structure [21–23]. Those previous studies relied on experimental data from various resources to train their ML models. One issue stumbling the progress of such studies is related to the deficiency in systematic documentation of HEAs data [24]. For example, two different values of the same property for the same HEA can be reported in different studies. Thereby, much data searching and cleaning or averaging is required to assure that the models will be well-trained on accurate data. On the other hand, combining computational material science with ML has been proven to open new avenues for discovering a vast extent of materials [25]. In that regard, Zhang et al. used molecular dynamic

(MD) with ML to predict the tensile properties of the single-crystal FeCrNiCoMn alloys, i.e. Cantor alloy [26], and CuFeNiCrCo single-crystal HEAs [27]. In both studies, the best-performing model was also capable of accurately predicting the properties of polycrystalline HEAs. Most studies have considered the input features to the ML models to be centered on the atomic concentrations, Hume-Rothery rule variables such as atomic size difference and electronegativity difference, and other thermodynamic parameters. Unlike the atomic concentrations, utilizing Hume-Rothery variables or other calculated thermodynamic parameters may cause inconsistency in the datasets since they require many estimations and averaging which is not appropriate for the ML models [24]. Many experimental and computational studies have attempted to investigate the effect of other parameters such as grain size, operating temperature, and strain rate on the properties of HEAs [28–30]. However, building a high-performance ML model using the latter parameters, besides the composition of HEA as predictors to forecast the tensile properties has not yet been made. The advantage of using such parameters as predictors is that they do not require estimations nor averaging to be calculated (except the average grain size), which will assure consistency and a better learning process for the models. Moreover, it opens a room for analyzing the problem from a different perspective.

In the present work, several ML models were applied including ANN, support vector machine (SVM), and Gaussian process regression (GPR) to predict the basic mechanical properties of the FeNiCrCoCu HEA; Young's modulus (E), yield strength (S_y), and ultimate tensile strength (UTS). **The dataset was generated via MD simulations to build the ML models by various input parameters:** atomic concentrations of constituent elements (C), average grain size (G_s), operating temperature (T), and strain rate (S_r). The ML models could be used before the actual experimentation to freely search for materials with the desired properties; avoiding resources and time waste associated with prototyping. In other words, the ML models offer an easy and fast way to guide and optimize the unguided experimental trial-and-error approach. The best-performing ML model of each output variable was investigated further by checking whether it can accurately extend the predictions to unseen datasets including new predictor variables' values that were not used to build the models.

2. Methodology

2.1. Analysis flow

Fig. 1 shows the analysis flow for the present work. The data were generated using MD simulation that map the inputs C , G_s , T , and S_r variables to the tensile properties E , S_y , and UTS. The tensile properties were extracted from the stress-strain curve. The different ML models were trained and tested afterward based on the produced dataset. A final evaluation of the best model for each response variable was done by supplying the model with new datasets containing predictor variables' values that were not used in the primary data.

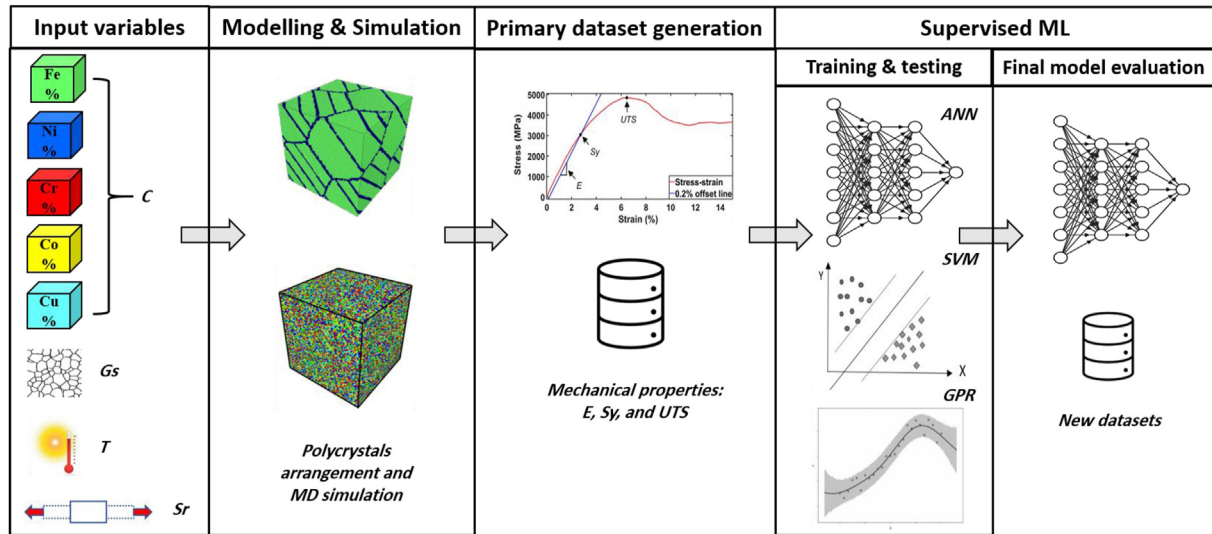


Fig. 1 – Schematic diagram showing the analysis flow in the present work.

2.2. Polycrystals arrangement

The 3D polycrystalline structures of the material were constructed using Atomsk software based on Voronoi tessellation technique [31]. The software is sufficient to arrange 3D and 2D polycrystals comprised of a specified number of grains that are randomly or selectively oriented in the structure. Three fcc pure iron polycrystalline samples containing 3, 5, and 7 randomly oriented grains, corresponding to average grain

sizes of 17.2 nm, 14.3 nm, and 12.6 nm, respectively, were constructed to investigate the effect of grain size on the tensile property of the HEA. The remaining four elements; Ni, Cr, Co, and Cu were added later to the structure during the MD simulation. Moreover, high-angle grain boundaries were ensured for all adjacent grains in these three samples i.e., misorientation between adjacent grains is equal to or greater than 15° . The step of ensuring high-angle boundaries was vital to assure that the grains would not merge later during the

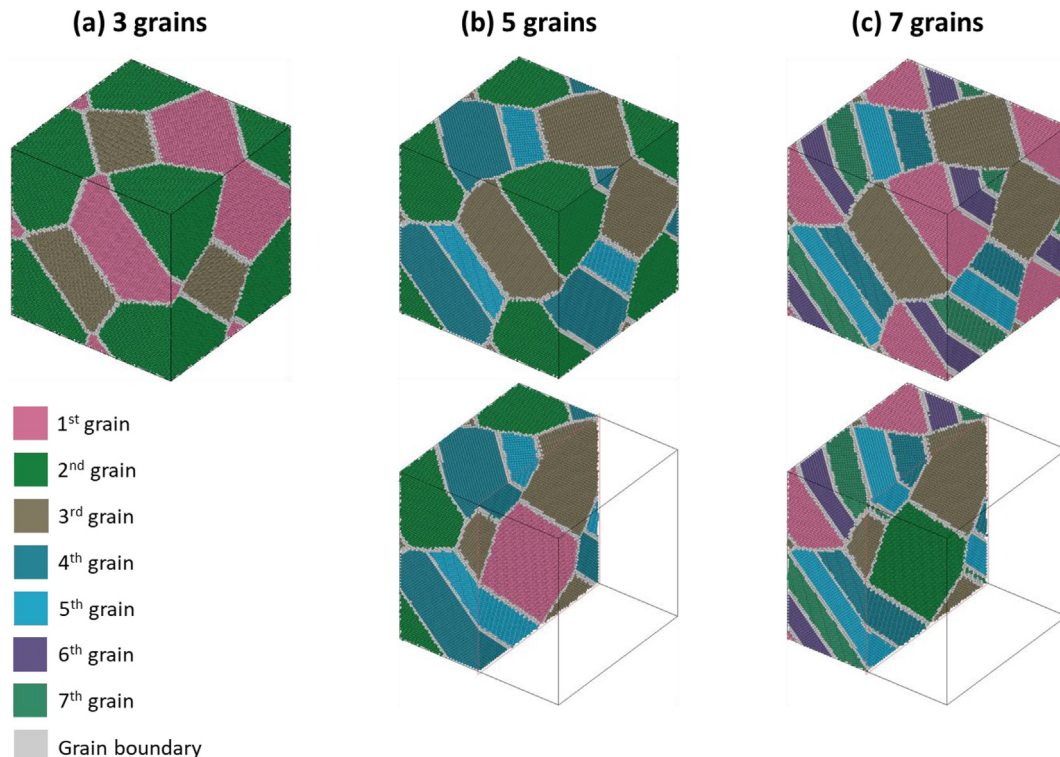


Fig. 2 – The three selected polycrystalline samples containing different number of grains. The different grains and the grain boundaries were distinguished using the grain segmentation tool and the PTM analysis.

relaxation process in the MD simulation, forming subgrains that would change the value of the average grain size in the structure. Fig. 2 illustrates the three selected polycrystalline models using the visualization tool Ovito [32]. The different grains and the grain boundaries were detected with the help of the grain segmentation and the polyhedral template matching (PTM) modules [33]. Polycrystals with more grains were excluded since they did not meet the condition of high-angle grain boundaries. Furthermore, the dimensions of the polycrystalline samples were chosen to be 20 nm along the x, y, and z axes, which resulted in a total of 700,968 atoms inside the simulation box. The box size provided a reasonable simulation time and maintained the crystallinity of the model during simulation.

2.3. MD simulations

The Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) was used in this study for MD simulations [34]. The embedded atom method (EAM) interatomic potential developed by Farkas and Caro [35] was implemented to describe the atomic interaction between the five constituent elements in the fcc phase. The EAM potential reproduced the basic equilibrium properties of the HEA including elastic constants, lattice constant, and cohesive energy; and it was shown to be suitable for the general investigation of the mechanical response [27,28]. The atoms of the remaining four elements which are Ni, Cr, Co, and Cu were replaced with some random iron atoms in the polycrystalline samples to form the HEA. Then, the total potential energy of the system was minimized using the conjugate gradient method available in LAMMPS. Subsequently, the HEA sample was heated up to the desired T level and later relaxed in the isobaric-isothermal (NPT) ensemble to release the residual stress while maintaining the desired T value. The NPT ensemble was used to control both the pressure and T during relaxation. Lastly, a uniaxial tensile test was carried out in the NPT ensemble on the sample along the x-direction. T was held constant and the stresses along the y and z directions were zeroed out during the deformation. Fig. 3 summarizes the MD simulation's modeling steps of the HEA sample. Throughout all the simulations, periodic boundary conditions were implemented in all three directions, and a time-step of 1 fs was set. For obtaining various compositions of the HEA, the concentration of elements was changed one at a time from 10 at.% to 30 at.% with a 2 at.% increment, while maintaining the concentrations of other elements equally. Accordingly, 51 different C were made. T values of 200 K, 300 K, and 400 K were also selected to perform the tensile deformation. Furthermore, the deformation was carried out with two $\dot{\epsilon}$: $4 \times 10^{-3} \text{ ps}^{-1}$ and $4 \times 10^{-2} \text{ ps}^{-1}$. Hence, 918 datasets were generated that encompass $51 (\text{C}) \times 3 (\text{Gs}) \times 3 (\text{T}) \times 2 (\text{Sr})$. The stress-strain data were produced via the tensile test implemented in the MD simulations. The response variables that are E , S_y , and UTS were extracted from the stress-strain curves. A 0.2% offset method was applied to obtain S_y . MD results were evaluated afterward by examining whether the three tensile properties were isotropic or not. Moreover, the tensile properties were compared to previously published datasets as will be seen later in section 3.

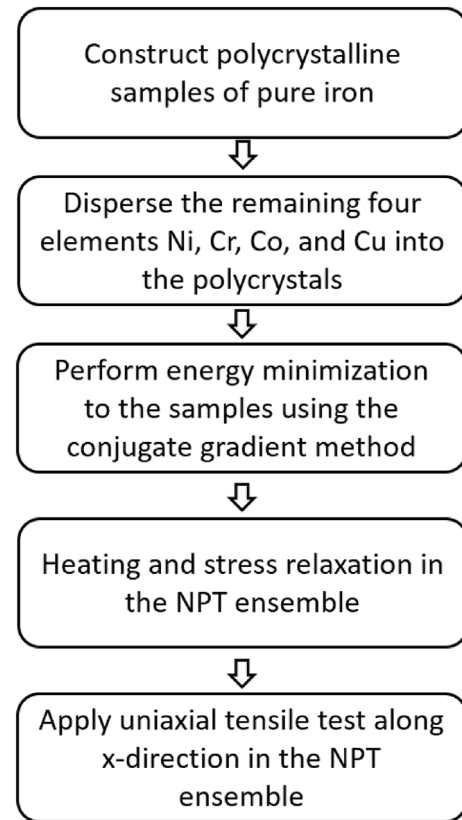


Fig. 3 – Flowchart of the FeNiCrCoCu HEA sample preparation/modeling.

2.4. ML prediction models

Supervised ML regression models were used to predict E , S_y , and UTS of FeNiCrCoCu HEA corresponding to C , G_s , T , and S_r . The ANN was nominated as the main regression model owing to its ability to detect complex relations between the predictors and the responses. Fig. 4 shows two different architectures of the ANN model. The architecture of a conventional ANN model includes an input layer, hidden layers, and an output layer that are connected based on updated weights and biases in each iteration to achieve an acceptable estimation. In addition, activation functions are usually implemented to account for nonlinear behavior in the training dataset. In this study, minimum-maximum normalization was used within the range of $[-1 \ 1]$ as given below:

$$x_{\text{normalized}} = 2 \times \frac{(x - x_{\min})}{x_{\max} - x_{\min}} - 1 \quad (1)$$

where, x can represent any of the predictor variables (C , G_s , T , and S_r), and x_{\min} and x_{\max} are the minimum and maximum values in each predictor variable, respectively. This normalization helps to avoid local minima and guarantees faster convergence to optimal values of weights and biases. Furthermore, the total data were divided into 70% training dataset and 30% testing dataset. The training set was used to optimize the values of the weights and biases such that the predicted values of the response variables in the training set approach the actual values as much as possible. The testing

set was used to assess the performance of the model on unseen data after the training phase was over. Moreover, the typical k -fold cross-validation (CV) sampling was implemented where k was set to 10. The k -fold CV indicates that the total data were sampled into k subsets called folds where $(k - 1)$ folds were used for training and 1-fold was allocated for testing. This process was repeated until each fold was used as a testing set once. Accordingly, the model was assessed based on an average performance metric over all the k testing folds. The k -fold CV is a more robust method for the model's evaluation since the model might show some variance in the predictions when trained and tested using different samplings. In addition, the hyperbolic tangent function was used as an activation function in the hidden layers. Three optimized single-output ANN models were established. Each one was built to predict one of the three tensile properties of the HEA. The optimization procedures can be summarized as follows:

- A trial-and-error approach was implemented to tune the number of neurons in the hidden layer for the three single-output models. 10 ANN architectures having one hidden layer were established starting from one up to ten neurons. Another 100 ANN architectures were made including two hidden layers starting from one to ten neurons in each hidden layer. The total number of examined architectures was 110. Fig. 4 depicts (a) the initial and (b) the final architectures of the ANN models.
- The k different data samplings (each contains training folds and a distinct testing fold) were used to train and test each of the 110 different architectures.
- As the ANN architectures are assigned with random initial weights and biases in the training phase, the previous procedure was repeated for 100 trials. The one trial involves the whole process of optimizing the randomly assigned initial values of the weights and biases using the training folds followed by testing using the one testing fold. This approach reduces the effect of the random initiation of the weights and biases on the models' performance.

- The root mean squared errors of the 10 testing folds that correspond to each trial ($RMSE_f(j)$) was calculated as given by Eq. (2), which provides an average indication of how far the predicted values are from the actual values:

$$RMSE_f(j) = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad f = 1, 2, \dots, 10; j = 1, 2, \dots, 100 \quad (2)$$

where f refers to a testing fold from the 10 different testing folds, j is the trial index for each testing fold, i indicates the observation, y_i is the predicted value of any response variable (E , S_y , and UTS), \hat{y}_i is the actual response value, and N is the total number of observations in each testing fold. $RMSE_f$ is an error vector of size 100 for each testing fold since 100 trials were performed.

- The minimum $RMSE_f$ value from the 100 for each testing fold was used to calculate the average $RMSE_f$ value as expressed in Eq. (3), which is the criterion to evaluate the 110 architectures and choose the best one.

$$\mu_{RMSE_f} = \frac{1}{10} \sum_{f=1}^{k=10} \min[RMSE_f(j)] \quad (3)$$

where $\min[RMSE_f(j)]$ indicates the best testing trail among the 100 trials for each f testing fold.

The coefficient of determination (R^2) was also found for the best testing trail for all k testing folds. R^2 gives an indication of how far the predicted and the actual values of the response variable are correlated to each other. The optimal scenario is when the predicted and the actual values are exactly matching, and the R^2 will have a maximum value of unity accordingly. Other regression models were also considered as a further step to compare their performance with the optimized ANN models including different types of SVMs and GPRs. The data were standardized (z-normalized) to have a mean of zero and a standard deviation of unity using Eq. (4) as follows:

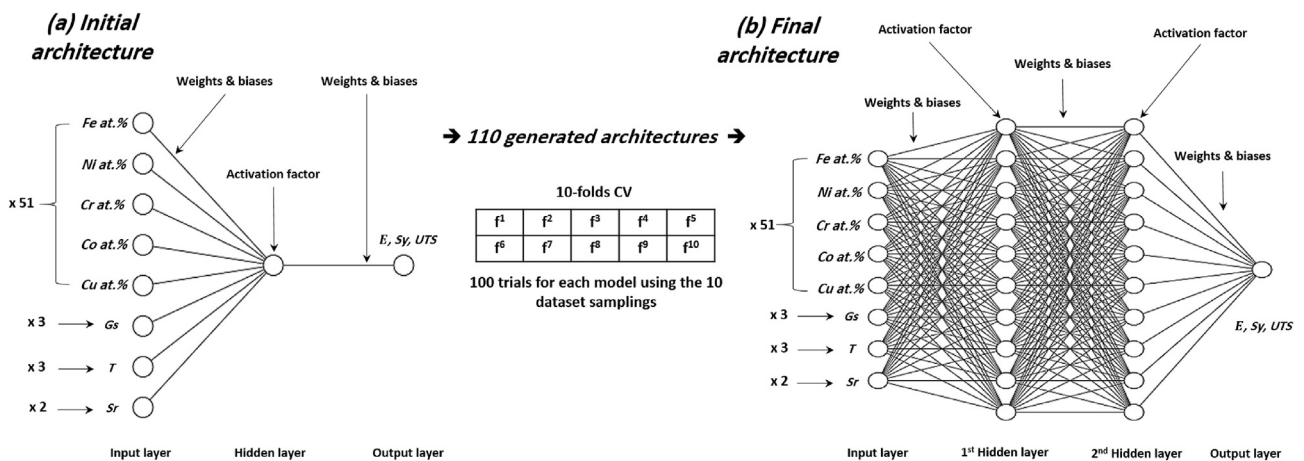


Fig. 4 – The (a) initial and (b) final architectures of ANN models for predicting the E , S_y , and UTS of FeNiCrCoCu HEA. A 10-fold CV was implemented to reduce the variances of the models. Each model was trained and tested 100 times using the different 10 dataset samplings to reduce the effect of the random initiations of the weights and biases.

$$z_{\text{normalized}} = \frac{z - \mu_z}{\sigma_z} \quad (4)$$

where z represents each predictor variable to be standardized (C, Gs, T, and Sr), and μ_z and σ_z are the mean and the standard deviation of the predictor variable, respectively. The 10-fold CV was also used to sample the dataset. It should be noted that default training settings and hyperparameters were implemented for these models without considering any further optimization. It should be emphasized that the main focus is on ANN models, while the SVMs and GPRs models were not optimized because they were just considered as benchmarks to compare them to the optimized ANN models and verify their performances. Unseen datasets of new C, Gs, T, and Sr values, that are not in the primary 918 datasets, were used to further evaluate the best ML model for each response variable.

3. Results and discussion

3.1. MD results

In experiments, the orientations of the grains with respect to the tensile direction are not necessarily similar or even close to that in the HEA polycrystalline models. Since the main objective of this study is to guide the experimental work, it is essential to examine the isotropy of mechanical properties in the polycrystalline models to make sure that there is no significant dependence between the tensile properties and the tensile direction. Although polycrystals are generally isotropic, the number of grains may not be sufficient to average the directional properties in the simulated HEA polycrystals. Therefore, 102 representative samples were chosen from the data and then subjected to two uniaxial tensile tests along the y and z axes. The same MD simulation procedures such as energy minimization and stress relaxation were implemented again before applying uniaxial tensile tests in the y and z directions. All the possible 51 different C were chosen from the primary 918 data. The three values of T and Gs were randomly assigned for each HEA composition from the 51. Finally, the 51 HEAs were deformed with the two values of Sr selected previously. This produces a total of 102 different samples in which each sample has its newly obtained E, S_y , and UTS along the y and z directions. The coefficient of variation (cov) for each sample was calculated given by Eq. (5) to evaluate the isotropy of the polycrystalline HEAs.

$$\text{cov} = \frac{\sigma}{\mu} \quad (5)$$

where

$$\sigma = \sqrt{\frac{(P_x - \mu)^2 + (P_y - \mu)^2 + (P_z - \mu)^2}{3}} \quad (6)$$

$$\mu = \frac{P_x + P_y + P_z}{3} \quad (7)$$

The variable P refers to any tensile properties: E, S_y , and UTS. σ is the standard deviation in the tensile property (P) of a given sample in the three directions (P_x , P_y , and P_z) and μ is the mean value of the property in the three directions. The calculated cov for each sample is a useful statistical quantity that reflects the extent of variation in the data. The optimal case is when P_x , P_y , and P_z show no variation. Accordingly, σ and cov will be zeros, and the sample is said to be perfectly isotropic. To obtain an average estimation of the isotropy of each tensile property over all 102 representative samples, the average cov value was computed such that:

$$\mu_{\text{cov}} = \frac{\sum_{n=1}^{102} \text{cov}_n}{102} \quad (8)$$

where n is an index referring to the sample. Fig. 5 presents the histogram of cov for each tensile property. A normal distribution was fitted to the histogram to visualize μ_{cov} graphically which is the centerline value of the normal distribution. As shown in Fig. 5, μ_{cov} was found to be equal to 4.17%, 6.55%, and 2.44% for E, S_y , and UTS, respectively. Hence, it can be concluded that the 102 HEA polycrystalline samples are nearly isotropic. It is valid to assume that the remaining 816 samples from the 918 datasets, which were only deformed in the x-direction, are generally isotropic since the 102 representative samples contain the same range of the predictor variables as in the 816 remaining samples.

Moreover, Zhang et al. [27] constructed 11 polycrystalline samples of the same FeNiCrCoCu elements system having 16 random-oriented grains for investigating E and peak stress. The same 11 samples were constructed and simulated to be compared with the reported results by Zhang et al. for verification. As shown in Fig. 6, E and peak stress results of the present study for the 11 representative samples agreed well with Zhang et al. results within a difference equal to and less than 6.03% and 11.11%, respectively. Therefore, it is suggested that the current analysis procedure based on MD simulation is

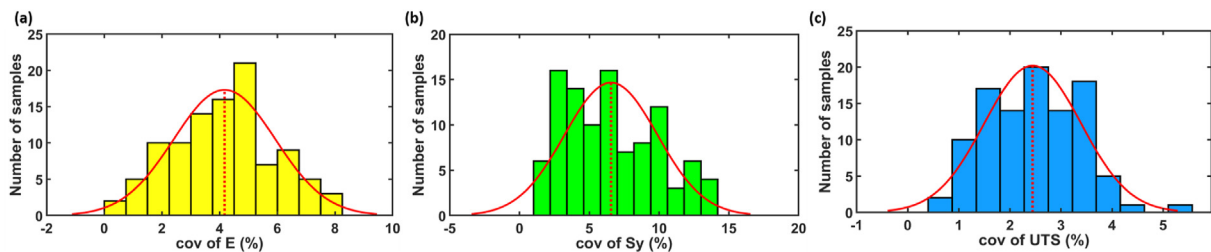


Fig. 5 – The cov distribution (histogram) of (a) E, (b) S_y , and (c) UTS and their corresponding μ_{cov} that is the centerline value of the fitted normal distribution.

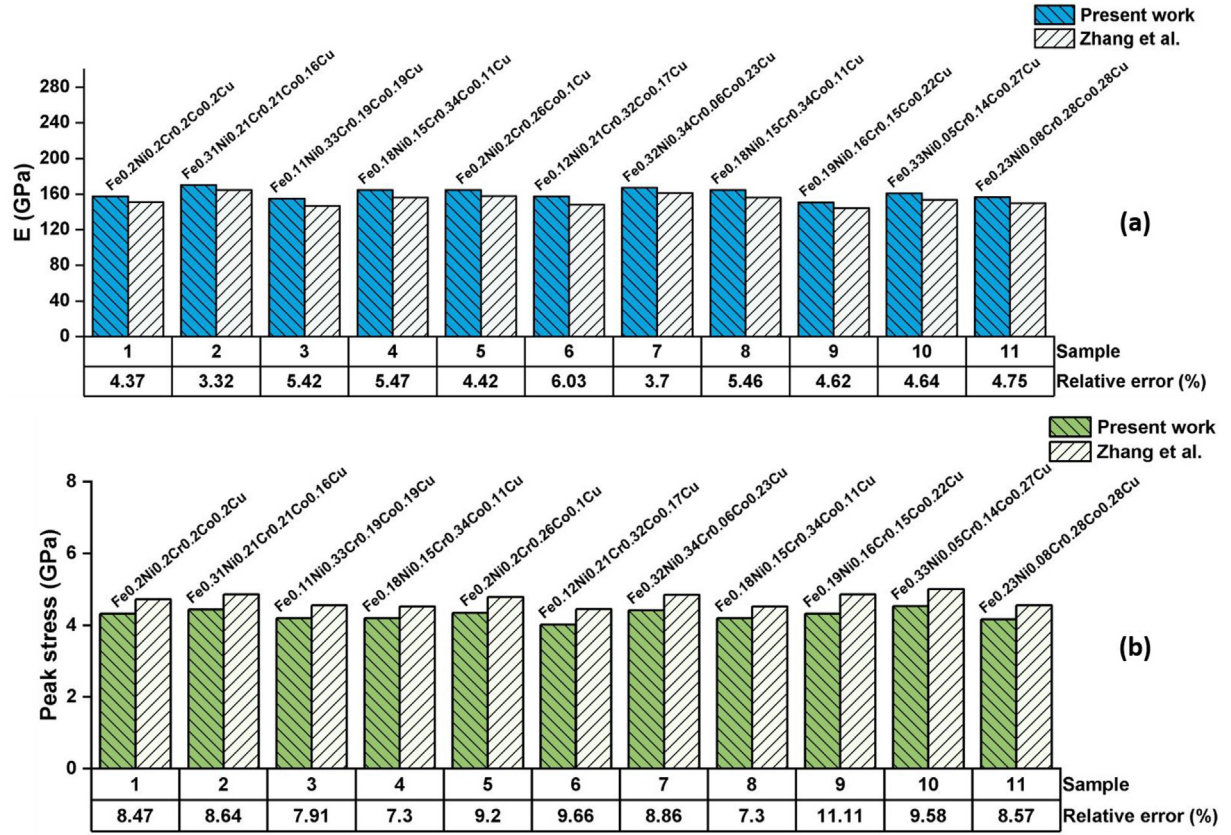


Fig. 6 – Comparison between Zhang's outcomes and the acquired MD results in the present work for 11 representative samples for (a) E and (b) peak stress.

appropriate to perform a tensile test on polycrystalline FeNiCrCoCu HEAs and find the tensile properties.

3.2. ML results

3.2.1. Evaluation of ML models' performance

Several ML models were established using the 918 datasets that were created by performing tensile test simulations in the MD environment and obtaining the corresponding properties. Subsequently, the best ANN models for predicting E turned out to contain 8 neurons in the first hidden layer and 2 neurons in the second hidden layer. As for S_y , the best ANN model had 6 and 8 neurons. On the other hand, the best prediction of UTS was obtained by an ANN model holding 6 and 9 neurons. The performances of the ML regression models were evaluated by μ_{RMSE_f} . Table 1 shows μ_{RMSE_f} for the best ANN models as well as the other ML regression models. It can be seen that the best ANN models for E, S_y , and UTS resulted in μ_{RMSE_f} values of 1.837 GPa, 0.1226 GPa, and 0.0609 GPa, respectively. Note that the ANN models outperform the other ML models. Fig. 7 depicts a scatter plot between the actual values of the tensile property versus the predicted values by the best ANN model, i.e., regression plot, for each f testing fold. As shown in the figure, all the predictions for the different testing folds are colored distinctively. Besides, the R^2 values of all testing folds are shown in brackets. Rather high R^2 values (nearly unity) were acquired for the three response variables. In addition, all

E and UTS predictions fell within a 5% error which is fairly good. As for S_y , most of the predictions lied within a 10% error. Furthermore, the root mean squared error and the average root mean squared error results of the training folds, $RMSE_t(j)$ and μ_{RMSE_t} , in which the subscript t indicates each $k - 1$ set of training folds, were evaluated in a similar manner to $RMSE_f(j)$ and μ_{RMSE_f} in Eq. (2) and Eq. (3), respectively. The differences

Table 1 – Summary of the obtained μ_{RMSE_f} for each ML model.

| ML model | μ_{RMSE_f} (GPa) | | |
|---------------------|----------------------|--------|--------|
| | E | S_y | UTS |
| ANN | 1.837 | 0.1226 | 0.0609 |
| SVM _L | 5.469 | 0.2120 | 0.2085 |
| SVM _Q | 2.056 | 0.1376 | 0.1801 |
| SVM _C | 2.176 | 0.1406 | 0.1873 |
| SVM _{Fg} | 9.092 | 0.3782 | 0.9677 |
| SVM _{Mg} | 2.657 | 0.1498 | 0.2219 |
| SVM _{Cg} | 4.054 | 0.1702 | 0.1880 |
| GPR _{Rq} | 2.048 | 0.1363 | 0.0677 |
| GPR _E | 2.281 | 0.1435 | 0.0840 |
| GPR _{Se} | 2.048 | 0.1362 | 0.0691 |
| GPR _{M5/2} | 2.039 | 0.1361 | 0.0676 |

L: linear; Q: quadratic; C: cube; Fg: fine Gaussian; Mg: medium Gaussian; Cg: coarse Gaussian; Rq: rational quadratic; E: exponential; Se: squared exponential; M5/2: matern 5/2.

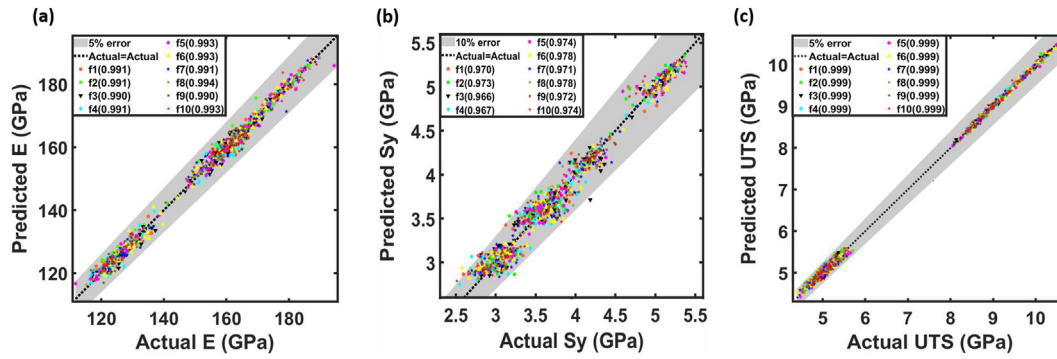


Fig. 7 – Best ANN model's regression plot for each output variable (a) E, (b) Sy, and (c) UTS for each f testing fold. All the predictions that lie within a certain fold are given a distinct colour. R^2 value of each f testing folds is shown in bracket.

are: (1) to evaluate Eq. (2) using all the observations in the $k-1$ training folds after settling to the final weights and biases instead of using the testing folds observations and (2) the $RMSE_t$ value from the 100 which has a trial index j that corresponds to $\min[RMSE_f(j)]$ in Eq. (3) was used to calculate μ_{RMSE_t} as in Eq. (9) below.

$$\mu_{RMSE_t} = \frac{1}{10} \sum_{t=1}^{k=10} RMSE_t(j \text{ of } \min[RMSE_f(j)]) \quad (9)$$

The μ_{RMSE_t} results of the training folds turned out to be equal to 1.872 GPa for E, 0.1234 GPa for Sy, and 0.0615 GPa for UTS. Those values are low and pretty much close to the reported μ_{RMSE_f} for the best ANN models, implying that there was no sign of overfitting [20].

3.2.2. New C, Gs, T, and Sr datasets

For experiments, the concentrations of the constituent elements may differ from those used to train the ML models (i.e. one element varied between 10 at.% to 30 at.% while other elements equally concentrated). In addition, the values of Gs, T, and Sr can be different from the ones that were utilized for the training of the ANN models in this study. This leads to the following question: Can the ANN models generalize effectively on unseen datasets containing new values of C, Gs, T, and Sr? To answer that, another 918 samples of completely randomized C values between 10 at.% to 30 at.% were generated and supplied to the best ANN model for each response variable. Moreover, 102 datasets of new Gs values, 102 datasets of new T values, and 102 datasets of new Sr values were produced and fed to the models. Table 2 shows the four new datasets, the number of observations in each dataset, and the new values of the predictor variable. The original values of the other predictor variables in the primary 918 datasets were used again in each of the four new datasets. For example,

dataset 1 comprises Gs values of 17.2 nm, 14.3 nm, and 12.6 nm, T values of 200 K, 300 K, and 400 K, as well as Sr values of 0.004 ps⁻¹ and 0.04 ps⁻¹. The root mean squared error (RMSE) and R^2 values were calculated for dataset 1. RMSE was simply calculated by initially finding the squared differences between the actual values of the response variable in dataset 1 and the predicted value by the ANN model, getting the mean value of the squared differences, and finally the square root. Favorably, an outstanding performance was also manifested by the three ANN models where the RMSE values are 2.158 GPa, 0.1367 GPa, and 0.0867 GPa for E, Sy, and UTS, respectively. These values are slightly higher than the earlier acquired μ_{RMSE_f} . This implies that the primary datasets contained sufficient information for the models to capture the relationship between C and the response variables. Fig. 8 also presents the regression plots of the models and the corresponding R^2 values. It is found that the R^2 values are quite high: 0.989 for E, 0.967 for Sy, and 0.999 for UTS. Furthermore, the error limits in the three plots depict that: (1) all predicted values of E were within 6% error, (2) the majority of Sy predictions lied within 10% error, and (3) all the predicted UTS results fell within 7% error. This reinforces the assertion that the ANN models were effectively generalized on the unseen datasets containing randomized values of C. For datasets 2, 3, and 4, boxplots of the relative error between the predicted values and the actual ones were plotted as shown in Fig. 9. The results in Fig. 9 (a, c) demonstrate that the models yielded a satisfactory level of error when predicting E and UTS for the new Gs dataset. Nevertheless, for Sy predictions, the maximum error increased significantly to 18.79% and 22.36% for Gs of 13.7 nm (2 grains) and 19.7 nm (6 grains), respectively, as illustrated in Fig. 9 (b). Concerning dataset 3 of the new T values, the relative errors of the three response variables' predictions are quite low even for the outliers as shown in Fig. 9 (e, f, g). Finally, yet importantly, it was found that the

Table 2 – The new four generated datasets to further evaluate the best ML models.

| Dataset No. | Predictor variable | No. of observations | Values of predictor variable |
|-------------|--------------------|---------------------|---|
| 1 | C | 918 | Random concentrations in the range {0.1–0.3} at. % |
| 2 | Gs | 102 | {13.7 - 15.6 - 19.7} nm |
| 3 | T | 102 | {100 - 225 - 350 - 385 - 435 - 500} K |
| 4 | Sr | 102 | {0.0004 - 0.001 - 0.007 - 0.01 - 0.07} ps ⁻¹ |

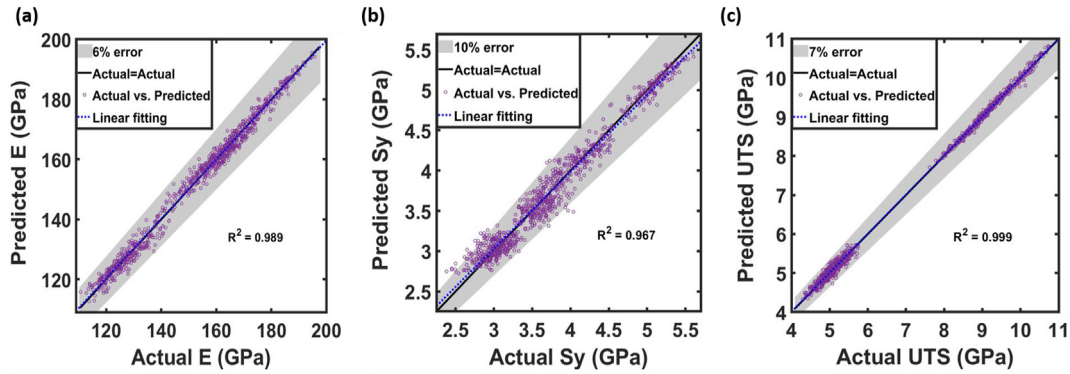


Fig. 8 – Regression plot showing R^2 value as well as the predictions of the best ANN model for each response variable (a) E , (b) S_y , and (c) UTS for the new random C dataset.

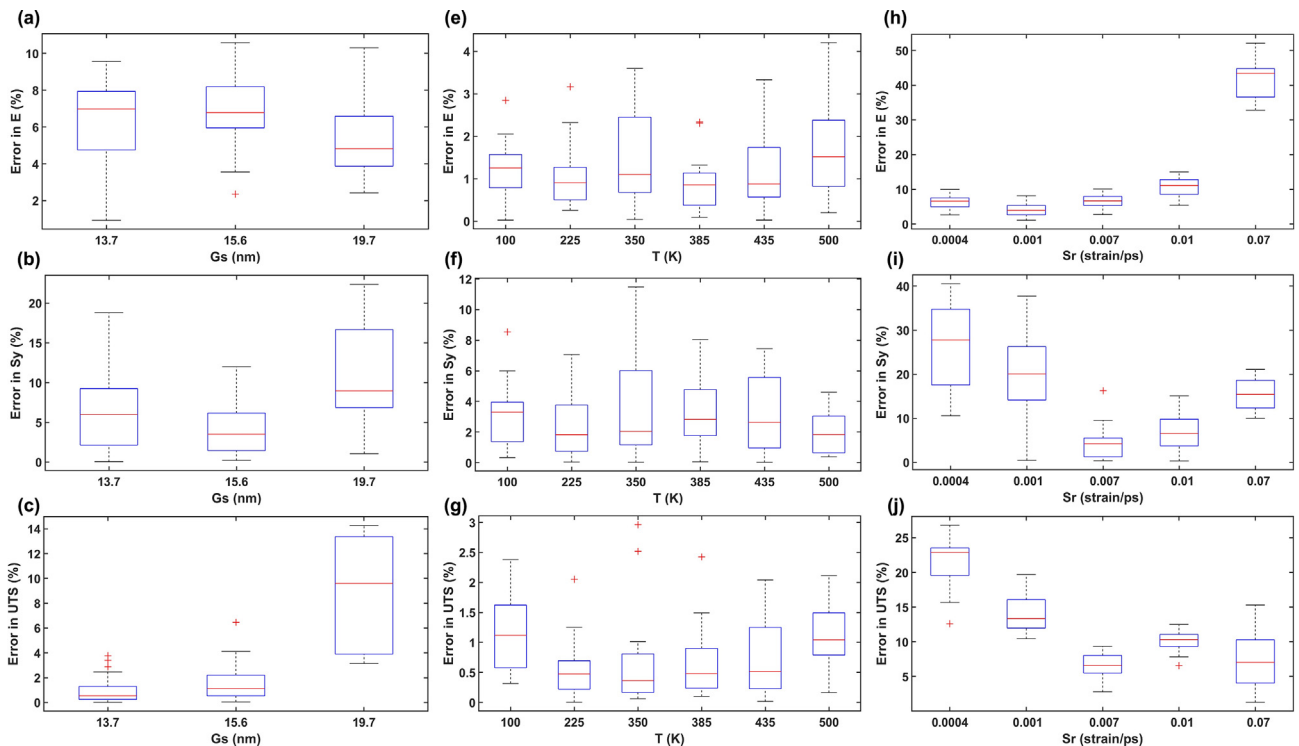


Fig. 9 – Boxplots of the error associated with (a) E , (b) S_y , and (c) UTS predictions for the new G_s dataset, (e) E , (f) S_y , and (g) UTS predictions for the new T dataset, and (h) E , (i) S_y , and (j) UTS predictions for the new S_r dataset. The plus sign indicates the outliers.

ANN models were not very capable of interpolating and extrapolating well to foresee the mechanical properties of the new S_r datasets as illustrated in Fig. 9 (h, i, j). The ANN model for E tends to maintain good predictions for observations containing S_r values of an order less than or equal to 10^{-2} . On the other hand, S_y and UTS models provided the most accurate predictions with an error less than or equal to around 10% for 0.007 ps^{-1} S_r observations. Therefore, it could be inferred that the tensile properties were more sensitive to changes in S_r than the other predictor variables, highlighting the need for additional S_r datasets to enhance the models' ability to capture more accurate correlations between S_r and the tensile

properties. More G_s datasets might also improve the predictions, especially for S_y since the error was observed to be high for the highest and lowest G_s values in dataset 2.

4. Conclusion

In this study, MD simulation coupled with ML was implemented to predict the tensile properties, namely: E , S_y , and UTS of FeNiCrCoCu HEA using C , G_s , T , and S_r as input parameters. The primary datasets of 918 polycrystalline samples were generated from MD simulations by performing tensile

tests and extracting the mechanical properties accordingly from the stress-strain curve. This work may offer a new way of investigating the wide possible range of HEAs properties via ML models for discovering their underlying treasures and guiding intensive experimental efforts. The main findings of the study can be summarized in the following points:

- For the MD results, it was found that the tensile properties of the polycrystalline HEAs were nearly isotropic. Besides, E and the peak stress results for the 11 representative samples agreed well with Zhang et al. results with errors equal to and less than 6.03% and 11.11%, respectively.
- The ANN was found to be the best model among other ML regression models in terms of predicting the three tensile properties of the HEA according to the comparison of the calculated μ_{RMSE_f} values. All E and UTS predictions lied within a 5% error, while most of the predictions for Sy were within a 10% error.
- The values of μ_{RMSE_f} seemed to be relatively low and very close to μ_{RMSE_f} , which indicated that overfitting issues were evaded.
- The best ANN models were generalized effectively on unseen datasets of random values of C . Additionally, these models provided careful predictions for the new T datasets. However, the Sy model showed relatively high maximum errors for some cases of Gs , namely when Gs was equal to 13.7 nm and 19.7 nm. In regard to the new Sr dataset, the ANN models were unable to accurately interpolate and extrapolate the tensile properties, which resulted in predictions with excessive errors. It might be necessary to incorporate additional Sr datasets to train the models to better identify the accurate relationships between Sr and the tensile properties.

To facilitate and expedite the discovery of HEAs and their potential applications, future research could integrate computational material science with ML techniques. Such research could explore radiation damage, wear, and erosion under harsh conditions, revealing hidden properties of HEAs that could be utilized in the aerospace and nuclear industries.

CRedit authorship contribution statement

Omarelfarouq Elgack: Methodology, Software, Formal analysis, Data curation, Investigation, Writing-original draft, Visualization. **Belal Almomani:** Validation, Writing-review & editing. **Junaidi Syarif:** Conceptualization, Methodology, Resources, Writing-review & editing, Supervision, Project administration. **Mohamed Elazab:** Investigation, Visualization, Data curation. **Mohammad Irshaid:** Investigation, Formal analysis, Data curation. **Mohammad Al-Shabi:** Validation, Writing-review & editing.

Data availability

The processed data required to reproduce these findings are available to download from: <https://data.mendeley.com/datasets/hthwb4nh83>.

Declaration of competing interest

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

Acknowledgment

The authors are grateful for financial support from the University of Sharjah [Grant No. 22020408199]. The authors of this paper would like to thank the Center for Advanced Materials Research at the University of Sharjah for their generous contribution and for the amount of computing time.

REFERENCES

- [1] Cantor B. Multicomponent high-entropy Cantor alloys. *Prog Mater Sci Jul.* 2021;120. <https://doi.org/10.1016/j.pmatsci.2020.100754>.
- [2] Li W, Xie D, Li D, Zhang Y, Gao Y, Liaw PK. Mechanical behavior of high-entropy alloys. *Prog Mater Sci May* 2021;118. <https://doi.org/10.1016/j.pmatsci.2021.100777>.
- [3] Liu C, Yang C, Liu J, Tang Y, Lin Z, Li L, et al. Medical high-entropy alloy: outstanding mechanical properties and superb biological compatibility. *Front Bioeng Biotechnol Aug.* 2022. <https://doi.org/10.3389/fbioe.2022.952536>.
- [4] Zhang Y, Zuo TT, Tang Z, Gao MC, Dahmen KA, Liaw PK, et al. Microstructures and properties of high-entropy alloys. *Prog Mater Sci Apr.* 2014;61:1–93. <https://doi.org/10.1016/j.pmatsci.2013.10.001>.
- [5] George EP, Curtin WA, Tasan CC. High entropy alloys: a focused review of mechanical properties and deformation mechanisms. *Acta Mater Apr.* 15, 2020;188:435–74. <https://doi.org/10.1016/j.actamat.2019.12.015>. Acta Materialia Inc.
- [6] Kumar A, Dhekne P, Swarnakar AK, Chopkar MK. Analysis of Si addition on phase formation in AlCoCrCuFeNiSix high entropy alloys. *Mater Lett Feb.* 2017;188:73–6. <https://doi.org/10.1016/j.matlet.2016.10.099>.
- [7] Torrento JE, Sousa TSP, Cruz NC, Almeida GS, Zambuzzi WF, Grandini CR, et al. Development of non-equiatomic Bio-HEAs based on TiZrNbTa-(Mo and Mn). *Appl Mater Aug.* 2022;10(8):81113. <https://doi.org/10.1063/5.0100465>.
- [8] Martin P, Madrid-Cortes CE, Cáceres C, Araya N, Aguilar C, Cabrera JM. HEAPS: a user-friendly tool for the design and exploration of high-entropy alloys based on semi-empirical parameters. *Comput Phys Commun Sep.* 2022;278. <https://doi.org/10.1016/j.cpc.2022.108398>.
- [9] Gao X, Wang L, Guo N, Luo L, Zhu G, Shi C, et al. Microstructure and mechanical properties of multi-phase reinforced Hf-Mo-Nb-Ti-Zr refractory high-entropy alloys. *Int J Refract Metals Hard Mater Jan.* 2022;102. <https://doi.org/10.1016/j.jrmhm.2021.105723>.
- [10] Yang Y, Ren Y, Tian Y, Li K, Zhang W, Shan Q, et al. Microstructure and properties of FeCoCrNiMoSix high-entropy alloys fabricated by spark plasma sintering. *J Alloys Compd Dec.* 2021;884. <https://doi.org/10.1016/j.jallcom.2021.161070>.
- [11] Alijani F, Reihanian M, Gheisari K. Study on phase formation in magnetic FeCoNiMnV high entropy alloy produced by mechanical alloying. *J Alloys Compd Jan.* 2019;773:623–30. <https://doi.org/10.1016/j.jallcom.2018.09.204>.
- [12] Manta E, Lucaci M, Vasile E, Lungu MV, Tâlpeanu D, Stancu N, et al. Influence of processing route on

- microstructure and properties of Al_{13.45}FeCrNiCo high entropy alloys. *J Alloys Compd* Jun. 2022;907. <https://doi.org/10.1016/j.jallcom.2022.164457>.
- [13] Bhardwaj V, Zhou Q, Zhang F, Han W, Du Y, Hua K, et al. Effect of Al addition on the microstructure, mechanical and wear properties of TiZrNbHf refractory high entropy alloys. *Tribol Int* Aug. 2021;160. <https://doi.org/10.1016/j.triboint.2021.107031>.
- [14] Akmal M, Hussain A, Afzal M, Lee YI, Ryu HJ. Systematic study of (MoTa)_xNbTiZr medium- and high-entropy alloys for biomedical implants- in vivo biocompatibility examination. *J Mater Sci Technol* Jul. 2021;78:183–91. <https://doi.org/10.1016/j.jmst.2020.10.049>.
- [15] Lee C, Song G, Gao MC, Ouyang L, An K, Fensin SJ, et al. Effects of Zr addition on lattice strains and electronic structures of NbTaTiV high-entropy alloy. *Mater Sci Eng Jan.* 2022;831. <https://doi.org/10.1016/j.msea.2021.142293>.
- [16] Bundela AS, Rahul MR. Machine learning-enabled framework for the prediction of mechanical properties in new high entropy alloys. *J Alloys Compd* Jul. 2022;908. <https://doi.org/10.1016/j.jallcom.2022.164578>.
- [17] Bhandari U, Rafi MR, Zhang C, Yang S. Yield strength prediction of high-entropy alloys using machine learning. *Mater Today Commun* Mar. 2021;26. <https://doi.org/10.1016/j.mtcomm.2020.101871>.
- [18] Wen C, Zhang Y, Wang C, Xue D, Bai Y, Antonov S, et al. Machine learning assisted design of high entropy alloys with desired property. *Acta Mater* May 2019;170:109–17. <https://doi.org/10.1016/j.actamat.2019.03.010>.
- [19] Syarif J, Elbeltagy MB, Nassif AB. A machine learning framework for discovering high entropy alloys phase formation drivers. *Heliyon* Jan. 2023;9(1). <https://doi.org/10.1016/j.heliyon.2023.e12859>.
- [20] Bakr M, Syarif J, Hashem IAT. Prediction of phase and hardness of HEAs based on constituent elements using machine learning models. *Mater Today Commun* Jun. 2022;31. <https://doi.org/10.1016/j.mtcomm.2022.103407>.
- [21] Chang H, Tao Y, Liaw PK, Ren J. Phase prediction and effect of intrinsic residual strain on phase stability in high-entropy alloys with machine learning. *J Alloys Compd* Nov. 2022;921. <https://doi.org/10.1016/j.jallcom.2022.166149>.
- [22] Mishra A, Kompella L, Sanagavarapu LM, Varam S. Ensemble-based machine learning models for phase prediction in high entropy alloys. *Comput Mater Sci* Jul. 2022;210. <https://doi.org/10.1016/j.commatsci.2021.111025>.
- [23] Qu N, Liu Y, Zhang Y, Yang D, Han T, Liao M, et al. Machine learning guided phase formation prediction of high entropy alloys. *Mater Today Commun* Aug. 2022;32. <https://doi.org/10.1016/j.mtcomm.2022.104146>.
- [24] Nassar AE, Mullis AM. Rapid screening of high-entropy alloys using neural networks and constituent elements. *Comput Mater Sci* Nov. 2021;199. <https://doi.org/10.1016/j.commatsci.2021.110755>.
- [25] Huang E, Lee W, Singh SS, Kumar P, Lee C, Lam T, et al. Machine-learning and high-throughput studies for high-entropy materials. *Mater Sci Eng Report* Jan. 2022;147. <https://doi.org/10.1016/j.mser.2021.100645>.
- [26] Zhang L, Qian K, Huang J, Liu M, Shibuta Y. Molecular dynamics simulation and machine learning of mechanical response in non-equiatomic FeCrNiCoMn high-entropy alloy. *J Mater Res Technol* Jul. 2021;13:2043–54. <https://doi.org/10.1016/j.jmrt.2021.06.021>.
- [27] Zhang L, Qian K, Schuller BW, Shibuta Y. Prediction on mechanical properties of non-equiatomic high-entropy alloy by atomistic simulation and machine learning. *Metals* Jun. 2021;11(6). <https://doi.org/10.3390/met11060922>.
- [28] Zhang L, Shibuta Y. Inverse Hall-Petch relationship of high-entropy alloy by atomistic simulation. *Mater Lett Sep.* 2020;274. <https://doi.org/10.1016/j.matlet.2020.128024>.
- [29] Verma A, Tarate P, Abhyankar AC, Mohape MR, Gowtam DS, Deshmukh VP, et al. High temperature wear in CoCrFeNiCu_x high entropy alloys: the role of Cu. *Scripta Mater Mar.* 2019;161:28–31. <https://doi.org/10.1016/j.scriptamat.2018.10.007>.
- [30] Wang CT, He Y, Guo Z, Huang X, Chen Y, Zhang H, et al. Strain rate effects on the mechanical properties of an AlCoCrFeNi high-entropy alloy. *Met Mater Int* Jul. 2021;27(7):2310–8. <https://doi.org/10.1007/s12540-020-00920-5>.
- [31] Hirel P. AtomsK: a tool for manipulating and converting atomic data files. *Comput Phys Commun* Dec. 2015;197:212–9. <https://doi.org/10.1016/j.cpc.2015.07.012>.
- [32] Stukowski A. Visualization and analysis of atomistic simulation data with OVITO—the Open Visualization Tool. *Model Simulat Mater Sci Eng* Jan. 2010;18(1):015012. <https://doi.org/10.1088/0965-0393/18/1/015012>.
- [33] Larsen PM, Schmidt S, Schiøtz J. Robust structural identification via polyhedral template matching. *Model Simulat Mater Sci Eng* May 2016;24(5). <https://doi.org/10.1088/0965-0393/24/5/055007>.
- [34] Thompson AP, Aktulga HM, Berger R, Bolintineanu DS, Brown WM, Crozier PS, et al. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput Phys Commun* Feb. 2022;271. <https://doi.org/10.1016/j.cpc.2021.108171>.
- [35] Farkas D, Caro A. Model interatomic potentials and lattice strain in a high-entropy alloy. *J Mater Res* Oct. 2018;33(19):3218–25. <https://doi.org/10.1557/jmr.2018.245>.