2022

A feasibility study of machine learning-assisted alloy design using wrought aluminum alloys as an example

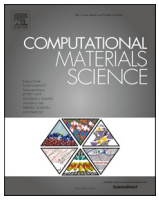
Yasaman J Soofi, *Missouri University of Science and Technology*
Md Asad Rahman, *Missouri University of Science and Technology*
Yijia Gu, *Missouri University of Science and Technology*
Jinling Liu, *Missouri University of Science and Technology*

Computational Materials Science 215 (2022) 111783

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci



Full Length Article

A feasibility study of machine learning-assisted alloy design using wrought aluminum alloys as an example

Yasaman J. Soofi [a], Md Asad Rahman [a], Yijia Gu [b,*], Jinling Liu [a,c,*]

[a] Department of Engineering Management and Systems Engineering, Missouri University of Science and Technology, Rolla, MO 65401, USA
[b] Department of Materials Science and Engineering, Missouri University of Science and Technology, Rolla, MO 65401, USA
[c] Department of Biological Sciences, Missouri University of Science and Technology, Rolla, MO 65401, USA

ARTICLE INFO

Keywords:
Alloy design
Aluminum alloy
Material property
Small dataset
Machine learning
Data mining

ABSTRACT

Machine learning (ML) often requires large datasets for reliable predictions, which may not be feasible for most commercial alloy systems. Also, the alloy development requires a full set of balanced properties, many of which have not been thoroughly investigated by ML. In this study, we focused on the practicality and reliability of ML in estimating alloy properties with a realistic small dataset of commercial wrought aluminum alloys as an example. We have compiled a small but comprehensive dataset that contains 236 entries with 6 mechanical properties and 9 technological properties. We first performed statistical analysis to understand the encoded correlation among compositions, mechanical and technological properties. Then, we systematically evaluated the predictive performance of several popular ML models with a focus on the bias-variance trade-off, a central problem in training supervised ML models. Moreover, we looked into the prospect of improving ML models by engineering the feature space. Finally, our feature importance analysis suggested the soundness of the developed models and revealed new insights on the underlying composition/processing-property relations. This study demonstrated that alloy design may be aided by using machine learning and data mining techniques on realistic small datasets.

1. Introduction

Material properties are dictated by the microstructures which are further defined by the material compositions and processing. The precise control of material microstructures at several length scales simultaneously is required for achieving desired material properties, which poses a grand challenge for alloy development. Recently, more and more integrated multi-scale multi-physics models (ICME) emerge to accelerate the alloy development process [1]. These physics-based models typically need the microstructure as the bridge to link composition and processing to properties, but much information in the microstructure cannot be described quantitively [2,3]. Therefore, those physics-based models often adopt certain assumptions, which may introduce uncertainties and errors and cause error propagation when integrated into an ICME framework [4]. Furthermore, generic and reliable material models for many specific composition/processing-microstructure and microstructure-property relationships are still lacking, especially for those properties controlled by complex physical mechanisms, such as corrosion and hot tearing. Therefore, it is still challenging to develop alloys to meet a whole set of desired properties even assisted with the ICME framework. A method that directly connects composition and processing to properties may avoid the potentially propagated errors and reduce the complicated modeling efforts.

Machine learning (ML) is a method of data analysis which trains computers to extract useful information from the data without being explicitly programmed. Supervised machine learning models learn the statistical patterns from available data and predict the pattern for new and unseen data samples. ML models excel in dealing with complex problems by quantifying the relationships among variables. Also, ML can directly model the relationship between composition/processing and material properties, rather than relying on the microstructure as a bridge like ICME does. Therefore, machine learning promises another viable computational route for material discovery problems. ML algorithms have recently been used to solve materials science problems such as material design and alloy development [5,6]. To design new alloys with desirable properties, various studies used ML techniques to investigate the relationship between the physical properties of the alloys and the composition and processing parameters. As an instance, Wen et al

* Corresponding authors at: Department of Materials Science and Engineering, Missouri University of Science and Technology, Rolla, MO 65401, USA (Y. Gu). Department of Engineering Management and Systems Engineering, Missouri University of Science and Technology, Rolla, MO 65401, USA (J. Liu).
E-mail addresses: yijia.gu@mst.edu (Y. Gu), jinling.liu@mst.edu (J. Liu).

https://doi.org/10.1016/j.commatsci.2022.111783
Received 5 May 2022; Received in revised form 12 August 2022; Accepted 3 September 2022
Available online 15 September 2022

compared the performance of various ML techniques in designing the composition of high entropy alloys with desirable hardness [7]. Furthermore, ML techniques, mostly artificial neural networks, have been used to model the relationships between the processing parameters and tensile properties of maraging steel, titanium, copper, and aluminum alloys including ultimate tensile strength (UTS), yield strength (YTS), and elongation (El) [8–12].

Although there are existing ML applications to material design [5,6], two fundamental questions need to be fully addressed before it can be considered a reliable and practical method to assist alloy development for industry.

1. For an engineering material, the complete property set includes not only mechanical properties including yield strength, ultimate tensile strength, elongation, and so forth, but also technological properties such as machinability, weldability, and cold workability. For many alloys applications, the corrosion resistance and resistance to stress corrosion are also critical. Is machine learning capable of predicting those properties as well?
2. Even for a well-developed commercial alloy system, the available data including compositions, processing, and a full set of properties is limited, maybe only a few hundred. Is that enough for reliable machine learning?

In this work, we use ML to assist alloy design using a small dataset (236), which is typical for commercial alloy systems. We use wrought Al alloy as an example to show that ML is capable of predicting the full set of alloy properties including both mechanical and technological properties. We chose wrought Al alloys for this study for the following reasons. Al alloy is a well-developed commercial alloy family, which plays important role in automotive, aerospace, and defense applications thanks to their high specific strength and high specific stiffness. The wrought Al alloys, which generally have better mechanical properties than casting Al alloys, have higher requirements in technological properties for downstream manufacturing processes. We thus selected wrought Al alloys for this study considering it is representative and more challenging to model with the requirement of a full set of balanced properties. Several studies have applied ML to predict tensile properties of Al alloys in general [12–16], but none of those have thoroughly studied the technological properties for wrought Al alloys or the bias-variance trade-off of the developed ML models for a small dataset.

In this work, we studied a complete set of properties required for wrought Al alloys using a small commercial alloy dataset: tensile strength, yield strength, elongation, hardness, fatigue, corrosion resistance, stress corrosion cracking resistance, cold workability, machinability, and weldability. We first examined the data and performed a correlation analysis. Then, several popular ML methods were evaluated for such a small dataset, and the random forest method with both good performance and easy implementation was selected for further modeling. Then, we built and developed random forest models for six mechanical properties and nine technological properties with good accuracy. In addition, we were able to further improve the ML model performance by engineering the feature space with certain knowledge in material science. Importantly, our randomness study in understanding the bias-variance tradeoff suggested these ML models developed from such a small dataset were robust and reliable. Lastly, we carried out feature importance analysis and obtained useful insights on the composition/processing-microstructure-property relationship. Overall, we have demonstrated a practical example of how to utilize ML and data mining for alloy design from a realistic small dataset.

2. Methods

2.1. Data preparation and Spearman correlation

The final dataset has 236 samples which were collected and cross-validated from different references including literature and websites [17–21] (see supplemental materials). Collected elongations were from different measurement methods such as A5 and A50. A5 is the proportion of original specimen gauge length to diameter ratio of 5, and A50 is the ratio of original gauge length to diameter ratio of 50 mm. All A50 measurements are approximated to A5 by increasing the measurements by 10 % [22]. Duplicated samples with the same compositions and temper types had been removed from the original dataset. Thus, the 236 samples within the final dataset represent unique combinations of elemental composition and temper for various aluminum alloys. The aluminum alloys are designated and categorized into different series according to their principal alloying element. There are 5 samples of series 1xxx, 44 of series 2xxx, 31 of series 3xxx, 2 of series 4xxx, 65 of series 5xxx, 49 of series 6xxx, 37 of series 7xxx, and 3 of series 8xxx within the dataset.

Elemental compositions and temper types are provided for each sample within the data set. The dataset used in this study is mainly collected from Kaufman Aluminum Alloy Database [18]. In this database, only 9 elements, which are the exact 9 elements investigated in this work, are included. According to the Introduction of this book, those 9 elements are intentionally added, which means there exist other elements that are treated as insignificant. The Processing parameters are described as 66 temper types by linguistic terms (e.g., O, T4 and T6). We encoded the temper types to binary values using one-hot-encoding, a popular and easy-to-implement method for preprocessing non-ordinal categorical variables for ML models. One-hot-encoding creates one column for each temper type and assigns 1 to the corresponding sample's temper type and 0 to the other temper types. Thus, 66 columns were added to the chemical elements columns to make up the entire training features. Thus, this one-hot-encoding scheme also allows us to explore the influence of each of the temper types on the properties.

The dataset consists of a comprehensive list of mechanical and technological properties for aluminum alloys that were considered the target variables by the ML models. The mechanical properties are continuous variables while the technological properties are ordinary variables that were further coded as integers. Spearman correlation [23] was performed to understand the rank correlation among compositions, mechanical properties, and technological properties.

2.2. ML model development and evaluation

Machine learning models were developed to estimate the mechanical and technological properties of the alloys. Mechanical properties, expressed as numerical values were estimated by regression models, while technological properties, described as categorical values, were classified by classification models. In this study, popular supervised ML algorithms including ANN, SVM, RF, and XGBoost were investigated as ML model candidates. Models were trained and developed by 80 % of the data and tested by 20 % of the data samples which have never been seen by the model. Extensive hyperparameter tuning was done for each ML model developed for each property with grid search following random search. Our random search, via scikit-learn RandomizedSearchCV, defined a larger search space of hyperparameter values and randomly sampled different hyperparameter sets in this search space. This random search provided insights into a good range or domain of the hyperparameters and helped us to define a narrower grid of hyperparameter values. We then used grid search via scikit-learn GridSearchCV to evaluate each set of hyperparameters within this grid to identify the best set of hyperparameters. 5-fold cross-validation within the training set was used to evaluate and select the best hyperparameter set. The hyperparameter tuning significantly improved the model performance on unseen test data for all the above models except RF. For RF, its default set of hyperparameters performed as well as the tuned hyperparameters, suggesting its performance is comparatively robust to hyperparameter changes compared to the other models evaluated here. Considering its good performance and easy implementation

Table 1
Statistical information of the compositional elements in the dataset.

	Si	Fe	Cu	Mn	Mg	Cr	Ni	Zn	Ti
Min (%)	0	0	0	0	0	0	0	0	0
Max (%)	12.2	1.1	6.3	1.2	5.5	0.25	2	8.1	0.07
Mean (%)	0.30	0.02	0.98	0.32	1.57	0.05	0.03	1.00	0.00
Std. Dev (%)	1.16	0.10	1.55	0.39	1.45	0.08	0.21	2.32	0.01

by using the default hyperparameters, we selected random forest as the main approach for predicting the material's properties.

We developed RF models for both mechanical and technological properties since the RF algorithm can be used to model both regression and classification problems. Random procedures can be reproduced by choosing the same random seed, so the prediction performance was initially evaluated for one specific seed which had a similar distribution of test/train sets to the whole dataset. In order to take the randomness effect into account, the performance was further validated by testing for 100 different seeds and comparing the mean and standard deviation of evaluation metrics on the 100 unseen test sets. In addition, RF models were validated using the leave-one-out cross-validation (LOOCV) method. LOOCV is an unbiased technique to investigate the influence of different data split on the model performance, by leaving one sample for test and training the model with the rest of the samples, and then repeating for all samples within the dataset. The 100 seeds and LOOCV validation methods led to robust results and conclusions.

2.3. Evaluation metrics for ML models

Mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE), and R^2 are some of the common evaluation metrics for regression models. In the current study, MAPE and R^2 are used to evaluate and compare the performance of regression models. MAPE is a measure of accuracy in regression models which presents the percent of deviation between paired observations expressing the same phenomenon. R2 is the coefficient of determination and is the proportion of the variation in the dependent variable that is predictable from the independent variable(s).

Classification models were usually evaluated using measures including precision, recall, and F1-score. Precision, recall, and F1-score were calculated based on the number of true and false predictions for each class. The overall performance of a classifier model for predicting a property was presented by the average of the metric for different classes. For most of the technological properties in this study, the number of samples in different classes was not balanced, so the weighted average of metrics was calculated according to the number of samples in each class.

2.4. Feature importance analysis

The input features of an ML model can be ranked according to their importance and effect on the model's prediction. In this study, the mean decrease in impurity (MDI) is used for measuring feature importance [24,25]. MDI calculates the sum of the number of usages of each feature across all trees and the feature importance scores are computed as the mean and standard deviation of accumulation of the impurity decrease within each tree.

2.5. Statistical analysis

A student's *t*-test is a hypothesis testing procedure that demonstrates the significant differences in group means [26]. A *t*-test's p-value represents the likelihood that the findings occurred by coincidence. Low P-values suggest that the observation was not made by chance. In this study, a student's *t*-test is used to carry out a statistical analysis of the significance of prediction improvements brought by feature engineering. We applied two-tailed distribution and two-sample unequal variance.

Table 2
Statistical information of mechanical properties in the dataset.

	UTS	TYS	El	BHN	Shear	Endurance
Min	90	35	2	23	30	9
Max	730	695	42	200	435	238
Mean	325.0	259.9	13.5	87.5	191.6	108.7
Std. Dev	143.1	143.0	6.8	38.9	83.1	39.7

3. Results and discussion

3.1. A comprehensive dataset and its encoded statistical correlation among compositions, mechanical properties, and technological properties

We collected and cross-validated from a variety of sources [17–21] 236 wrought aluminum (Al) alloy experimental data points across five major alloy series with comprehensive mechanical and technological properties. The nine main alloying elements: Si, Fe, Cu, Mn, Mg, Cr, Ni, Zn, and Ti were used to create these Al alloys through a total of 66 designated heat treatment methods. Table 1 shows the statistical information of these nine main alloying elements across five major alloy series.

A total of six mechanical properties were included and described as continuous variables in the dataset: ultimate tensile strength (UTS), tensile yield strength (TYS), elongation (El), Brinell hardness number (BHN), ultimate shear strength (Shear) and endurance limit (Endurance or Endu). Their statistical information is presented in Table 2. A total of nine technological properties were included and described as categorical variables in the dataset: resistance to general corrosion (RGC), resistance to stress corrosion cracking (RSCC), Extrudability (Extr), cold workability (CW), machinability (Mach), resistance and spot weldability (RSW), brazeability (Braz), gas weldability (GW) and arc weldability (AW). These technological properties are categorized by comparative and linguistic phrases, including "Unacceptable", "Borderline", "Fair", "Very Good", and "Excellent". Table 3 shows the number of samples in each class for each property.

Using the Spearman correlation, we explored the statistical relationships among the different compositions, mechanical properties, and technological (Fig. 1). Several interesting rank correlations are found among those variables.

- El is not much correlated to composition or any properties except TYS. This is most likely due to the defect-related nature of the elongation. Even among the same batch of specimens, the elongation readings in tensile testing might vary significantly. The observed negative correlation with TYS is also consistent with strength-ductility trade-off.
- The mechanical properties (except El) have a strong positive correlation with each other as indicated by the blue square at the center of the heat map.
- Technological properties (except Mach) have a strong positive correlation with each other as indicated by the blue square at the bottom right corner of the heat map.
- Cu and Zn, in general, show a positive correlation with mechanical properties (except El) but a negative correlation with technological

Table 3
Statistical information of technological properties in the dataset.

	RGC	RSCC	Extr	CW	Mach	RSW	Braz	GW	AW
Unacceptable	NA	NA	NA	NA	16	NA	NA	NA	NA
Borderline	39	7	NA	65	75	3	127	85	35
Fair	47	36	61	79	55	2	15	39	29
Very Good	25	85	93	62	88	100	16	3	22
Excellent	125	108	82	30	2	131	78	109	150

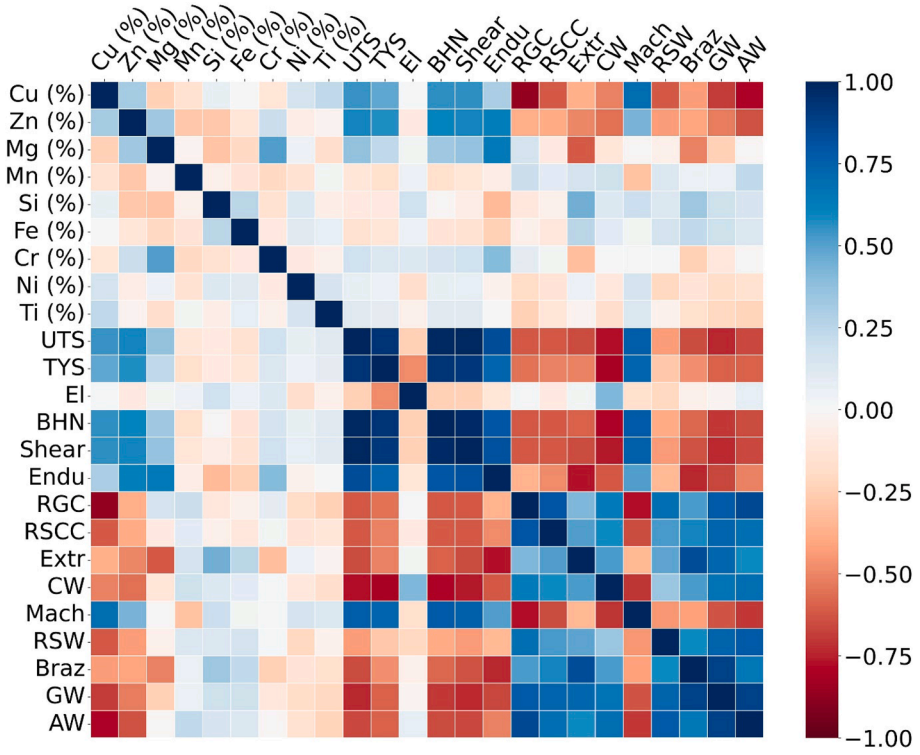


Fig. 1. The heatmap of statistical correlation among compositions, mechanical and technological properties. The bar on the right shows the color scheme representing the different values of Spearman's rank correlation coefficients.

properties (except Mach). This also sounds reasonable. These two elements are the top strengthening phase formers for Al alloys and are the major alloying elements for high strength wrought 2xxx and 7xxx alloys. However, these alloys also have a strong tendency to hot tearing due to the development of a large solidification range and high hot tearing susceptibility [27]. Therefore, their castability and weldability, such as GW and AW, are usually poor.

- Corrosion resistance including RGC and RSCC are both strongly and negatively related to Cu content. This is consistent with the fact that the introduction of Cu into Al alloys strongly increases the corrosion potential compared with pure Al.
- RSW shows a strong negative correlation with Cu and Zn. RSW of Al alloys generally depends on the electrical conductivity. With the appearance of numerous fine secondary phase particles, the electrical conductivity becomes much lower in age-hardened Al alloys. Therefore, it is making sense that those strong secondary phase formers, Cu and Zn, show a negative correlation with RSW.
- The technological properties (except Mach) are found to be negatively correlated with mechanical properties (except El) from the heat map. This statistical correlation is consistent with our observation that those alloys with better mechanical properties, such as high strength and/or high hardness, are in general more difficult to fabricate indicating poor manufacturability. This is also the reason why we believe those technological properties need to be considered for alloy design.

Overall, the results from the above correlation analysis are consistent with and supported by experimental observations or well-known materials theory/principles. This suggested that with a reliable and comprehensive dataset, one can readily extract rich and meaningful information with data mining techniques.

3.2. *Machine learning model selection and evaluation*

Considering those interesting rank correlations encoded in this dataset, we investigated whether we could build supervised machine learning models to predict this comprehensive list of properties from compositions and tempers with reasonable bias-variance trade-off. The bias-variance trade-off is a fundamental concept in supervised machine learning: increasing bias will decrease variance and vice versa. The bias error comes from erroneous assumptions of the learning algorithm; high-bias learning methods typically produce simpler models that miss the encoded relations between features and target variables (i.e., underfitting). The variance error results from the sensitivity of the model to small fluctuations in the training set; high variance learning algorithms may represent training data well but not generalize well to the unseen test data when they are modeling the random noise in the training data or unrepresentative training data (i.e., overfitting). Despite the tension between bias and variance, one would want to build or choose a supervised machine learning model with balanced bias and variance for the problem space.

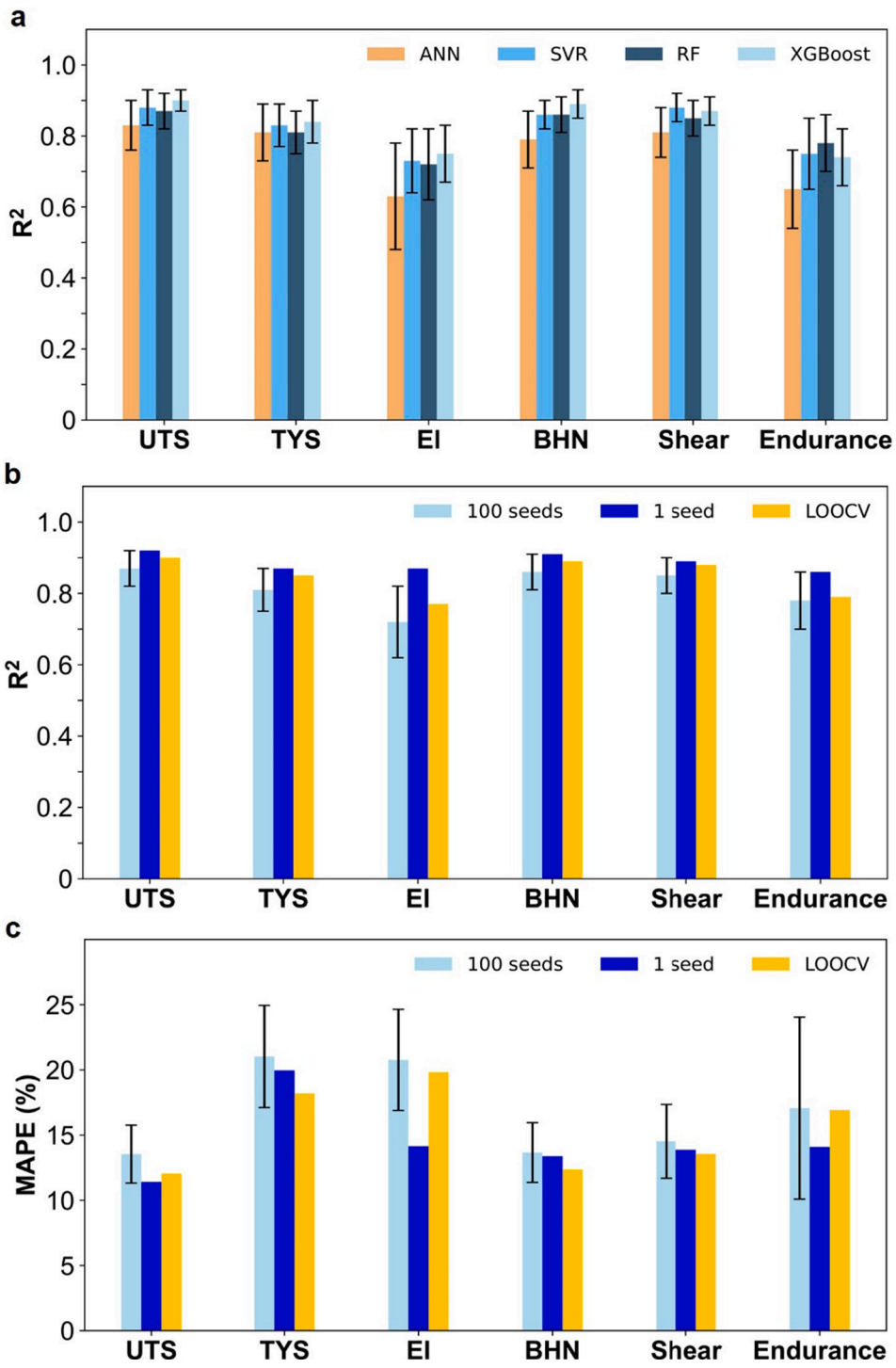


Fig. 2. Evaluation metrics of the developed ML models in predicting mechanical properties from compositions and tempers. a. The R^2 metric of various ML models was developed for each of the six properties. The color bars show the average R^2 from 100-seeds experiments and the small bars show the variance. **b.** The R^2 metric of RF models was developed for each of the six properties. **c.** The metric of Mean absolute percentage error (%) for RF models was developed for each of the six properties.

In this study, supervised ML algorithms were trained by 80 % of the data to build the models that were further tested by 20 % of the data samples which have never been seen by the learning algorithm during the training stage. The training and test samples were split randomly so that each set would be properly representative of the original/entire dataset and similar data distribution would be observed between the training and test set. This is normally true for a very large dataset. With a comparatively small dataset in our case, one random split may lead to biased train/test sets where the data distribution in the training and test set may be very different. In this case, the model trained and built from the training set would not perform well in the distinct test set.

To take the randomness effects into account and to vigorously evaluate the robustness of the developed ML models regarding both bias and variance, we trained and evaluated our models by three different methods: (i) using a single random train-test split returned by a single seed where the produced train/test sets look similar (1 seed), (ii) using 100 different random train-test splits returned by 100 random seeds (100 seeds), (iii) using leave one out cross-validation (LOOCV). We further used different metrics of regression models to evaluate our model performance in predicting mechanical properties including R^2 and MAPE. With the changed training/test sets, the 100-seeds experiment can tell both the bias and the variance of the developed ML models.

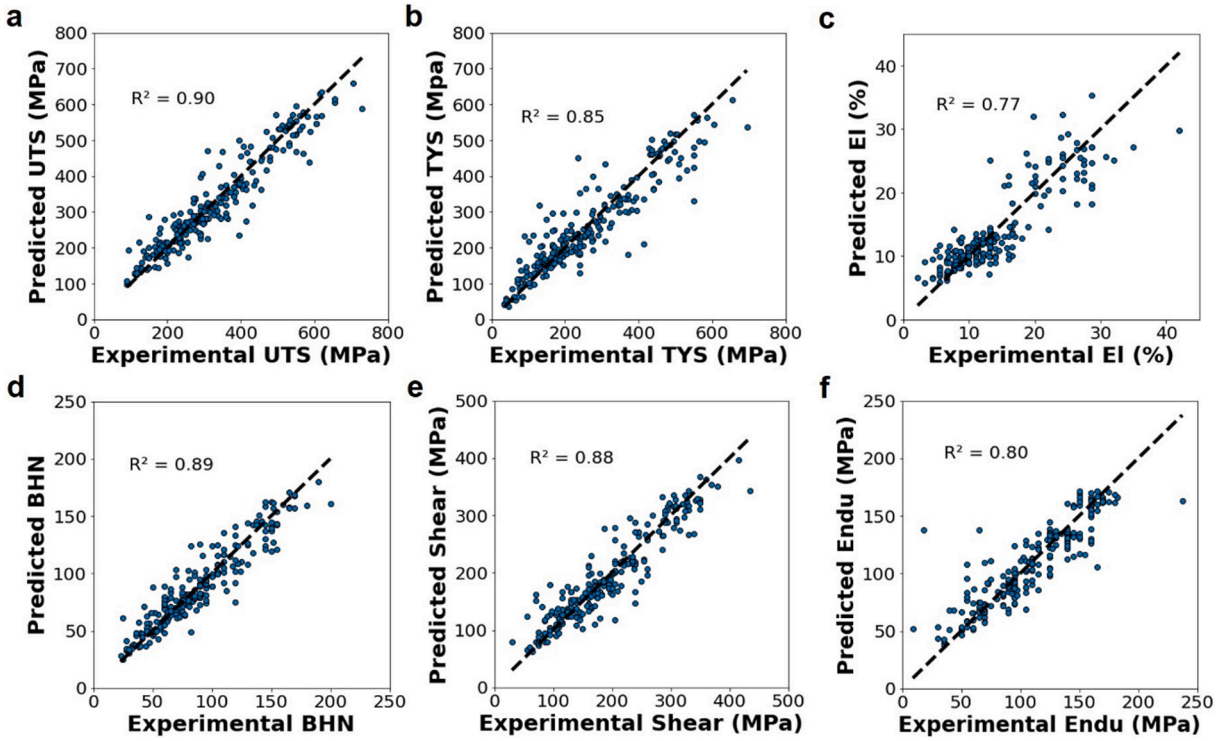


Fig. 3. Scatter plots of predicted and experimental true values for six mechanical properties from the LOOCV experiments. a. UTS **b.** TYS **c.** El **d.** BHN **e.** Shear **f.** Endurance. The black diagonal line shows where the predictions are exactly equal to the actual values.

For LOOCV experiments, each ML model was trained using all the samples except one test sample; a total of 236 ML models were trained in this way to make predictions on each unseen test sample of all the 236 data points; we then calculated the R^2 and MAPE across all the 236 samples using the true values and predicted values of target variables. As is commonly known, by using almost all the data available for training, LOOCV gives a reliable and unbiased estimate of model performance, especially when the sample size is relatively small.

In this study, we first tested the performance of several popular supervised ML algorithms on this small dataset that includes artificial neural networks (ANN), support vector machine for regression (SVR), random forest (RF), and extreme gradient boosting (XGBoost). ANN is a machine learning technique that is inspired by the human brain and neural system for data analysis [28,29]. ANNs are more computationally complex and usually need a larger number of training data compared to other ML models in order to deliver robust results. SVR fits an appropriate line to the data and works well for both linear and non-linear problems [30]. RF algorithm randomly chooses a subset of features to build an ensemble of decision trees [31]. RF needs minimum preprocessing on data and can deal with features of various scales. Moreover, RF reduces the variances in prediction by combining many decision trees into one ensemble. XGBoost is the optimized implementation of the gradient boosting algorithm that increases the speed and performance of the ensemble of sequential decision tree models [32]. Although XGBoost is faster than RF, it is relatively harder to tune and is more prone to overfitting.

During the training process, the hyperparameters of each model were tuned using 5-fold cross-validation within the training set for each property. Prediction performances in the unseen test set of the hyperparameter-tuned ANN, SVR, RF, and XGBoost models from 100-seeds experiments are compared (Fig. 2a). The performances of SVR, RF, and XGBoost, measured by the R^2 metric, are similar for all the properties in general (Fig. 2a). ANN has a lower R^2 and is also less reliable indicated by the larger variance (Fig. 2a). Other studies have reached similar conclusions about ANN, SVR, and RF [11,13]. However, RF was reported to be comparatively robust to hyperparameter tuning and the default hyperparameters typically perform well without the need for further hyperparameter tuning [33]. We verified this within this dataset. We decided to use RF with the default hyperparameters in this study considering the following factors: (1) the good performance of RF with default hyperparameters without the need for further hyperparameter tuning is convenient for building ML models for a variety of target variables, 6 mechanical and 9 technological properties in this study; (2) our study could be easily replicated or adopted by researchers who are not very experienced with ML models.

Six RF models with the default hyperparameters were finally developed to predict each of the six mechanical properties according to the features of compositional elements and tempers. Across all the six mechanical properties, the model performance on the unseen test dataset measured by R^2 and MAPE was very similar between 100-seeds and LOOCV experiments, and relatively small variance was observed for 100-seeds experiments indicated by small error bars (Fig. 2b,c). For some properties such as EI and Endurance, 1-seed experiments showed better model performance in prediction than 100-seeds (Fig. 2b,c). This could be because this 1-seed split yielded a desired situation for training ML models: similar data distribution between the training set and test set in terms of features (compositions and tempers) and outputs (properties). The histograms of the features and outputs in training and test sets split with seed 26 look similar (data not shown). On the other hand, for some random splits among the 100 seeds, potentially distinct training/test sets were generated and resulted in worse prediction performance in the test set, making the average performance from 100-seeds experiments slightly worse. Nevertheless, the good average performance with small variance from 100-seeds experiments together with the good performance from LOOCV experiments suggested that all the RF models developed for predicting these six mechanical properties were robust and reliable.

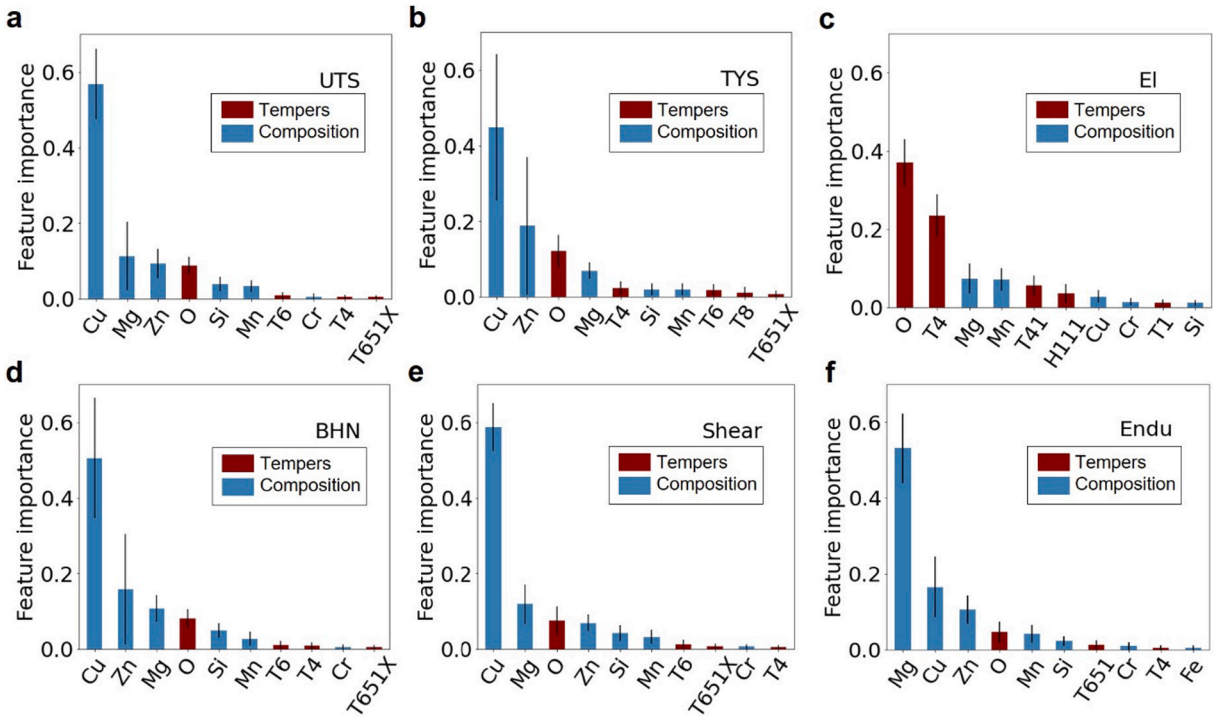


Fig. 4. The importance scores of the top-10 most important features identified by the RF model for each mechanical property. a. UTS **b.** TYS **c.** El **d.** BHN **e.** Shear **f.** Endurance. Red bars refer to temper types while blue bars refer to the compositions. The thin black lines are the error bars.

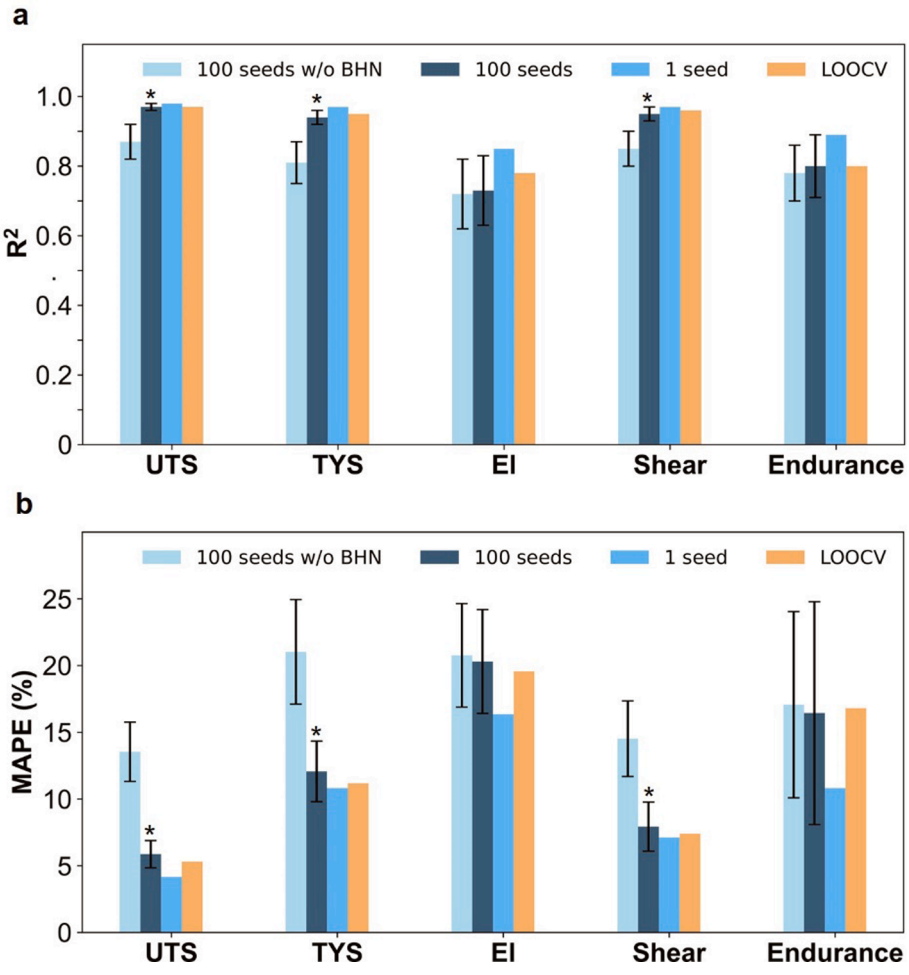


Fig. 5. Evaluation metrics of the developed RF models in predicting mechanical properties from both compositions/tempers and BHN. a. R-squared values. **b.** Mean absolute percentage error (%) values. For each property, the first bar shows the prediction accuracy without the BHN feature from 100-seeds experiments; the following three bars show the prediction accuracy with the addition of the BHN feature from 100-seeds, 1-seed, and LOOCV experiments. The stars suggest a significant difference in the R-squared values or the mean absolute percentage error values between the models developed without the BHN input and with the BHN input. P-values lower than 1e-10 are considered as a significant difference and marked with a star (*). UTS, TYS, and Shear have R^2 P-values of 2.36e-40, 1.65e-37, and 1.72e-40, respectively, whereas El and Endurance have P-values of 3.20e-1 and 1.72e-2. UTS, TYS, and Shear have MAPE P-values of 4.64e-65, 1.04e-44, 2.49e-45 while El and Endurance have P-values of 3.98e-1, and 5.65e-1.

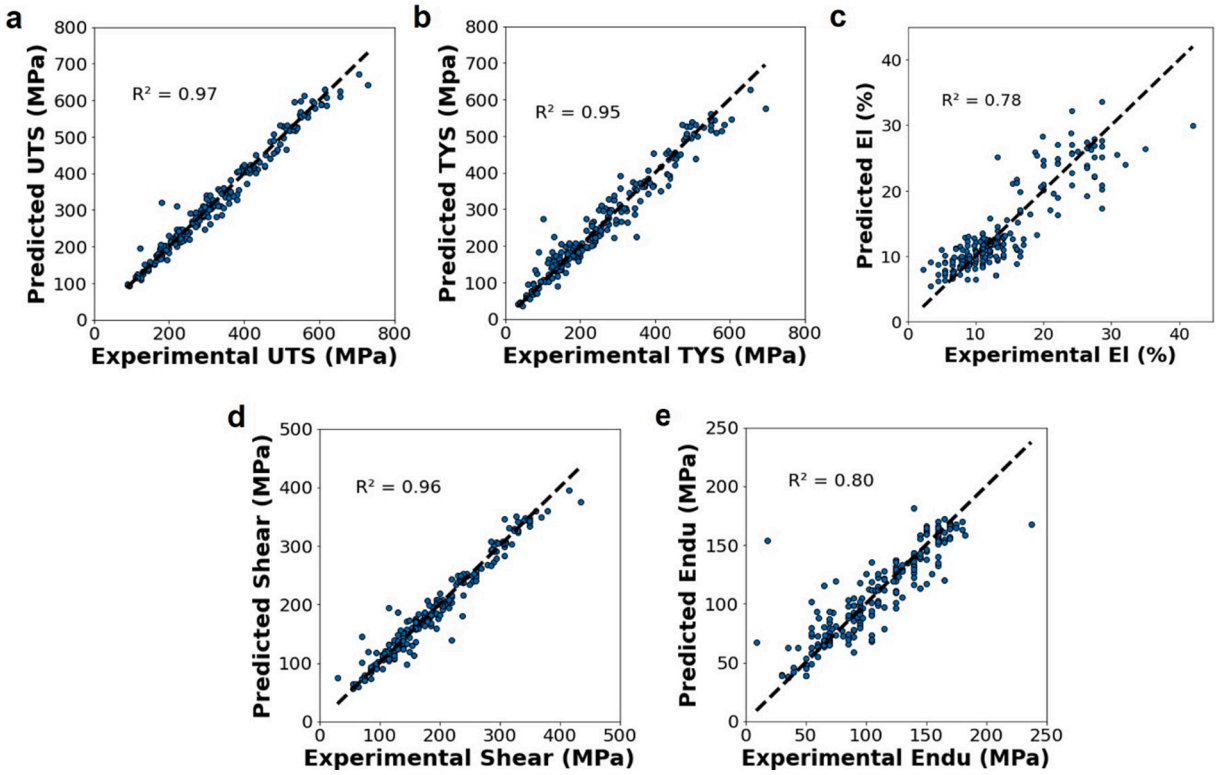


Fig. 6. Scatter plots of predicted and experimental true values for six mechanical properties with the additional BHN feature from the LOOCV experiments. a. UTS **b.** TYS **c.** El **d.** Shear **e.** Endurance. The black diagonal line shows where the predictions are exactly equal to the actual values.

3.3. Prediction of six mechanical properties from compositions and tempers

Among all the six mechanical properties, UTS, BHN, and Shear were predicted well from compositions and tempers with R^2 equal to or close to 0.90 for unseen test samples from LOOCV experiments (Fig. 2b). The prediction for TYS was a little bit worse with an R^2 of 0.85 (Fig. 2b). The predictions for EI and Endurance from compositions and tempers were slightly worse with R^2 close to 0.80 (Fig. 2b). Still, our prediction for El appeared to be better than others who also claimed EI to be more challenging to predict than UTS or TYS [13]. This is probably because it is more sensitive to internal defects, which requires more information on impurities, fabrication processes, and other defect formation-related factors for a better prediction. Similar trends were observed for the metric of MAPE (Fig. 2c).

Interestingly, although we simply used designated temper types as processing information, our predictions were comparable to others on Al alloys in literature which used more complicated parameters such as aging time and temperature [13]. Although they also had a larger dataset (~800 samples), we could achieve a better R^2 of 77.7 % (Fig. 2b) for predicting elongation in our LOOCV experiments. This suggests that it may be sufficient to use the designated temper as the only processing feature for predicting properties.

We further examined the scatter plots for predicted and actual experimental values of each property acquired from LOOCV experiments since LOOCV tends to give a reliable and unbiased estimate of model performance, especially for a small dataset (Fig. 3). The predictions for UTS, BHN, and Shear are scattered close to the diagonal line suggesting more accurate predictions, which is consistent with a high value of R^2 (Fig. 3a,d,e). The predictions of TYS and Endurance (Fig. 3b, f) for most samples are close to the diagonal line suggesting accurate predictions for most samples; the several samples that are away from the diagonal line (not predicted well) lead to lower R^2. For El, it is interesting to notice that the RF model predicted well (predictions close to the diagonal line) for samples with smaller El and generally worse (predictions away from the diagonal line) for samples with bigger El (Fig. 3c).

To understand better how each of the input features (compositions and tempers) relate to the outputs (mechanical properties), we further performed the feature importance study (Fig. 4). The feature importance score suggests how important a feature is in predicting the output. Since all the 236 models from the LOOCV experiments are similar with only one sample difference for training, we randomly choose one LOOCV model for each property to show the importance scores of the top-10 features. For all the mechanical properties except El, compositions are the most important predictors, among which Cu, Mg, and Zn are the most influential alloying elements (Fig. 4). In terms of tempers, O, T4, and T6 are the most important in the prediction of mechanical properties (Fig. 4). As the age hardening is the most effective strengthening method for Al alloys (~30 times increase in strength), the secondary strengthening phase-forming elements (Cu, Mg, Zn) and tempers (O, T4, T6) have the strongest influence on general mechanical properties including UTS, TYS, and El. Therefore, the feature importance ranking by our developed RF models is consistent with and supported by the material knowledge. These findings are also consistent with the previous correlation analyses (Fig. 1). It may also be deducted that heat treatment methods, especially temper O and T4 are the most defining and predictive features for elongation (Fig. 4c). This also makes sense because elongation is the highest in the annealed state (O temper) and the lowest in the peak aged state (T4/T6 temper).

3.4. The inclusion of BHN as an input feature significantly improves the prediction of UTS, TYS, and Shear, but not for El and Endurance

The performance of the model may be improved if provided with more useful information as inputs. BHN is known to be correlated with tensile properties and is relatively easy to be measured experimentally. Thus, including BHN as an additional input is a viable and practical idea

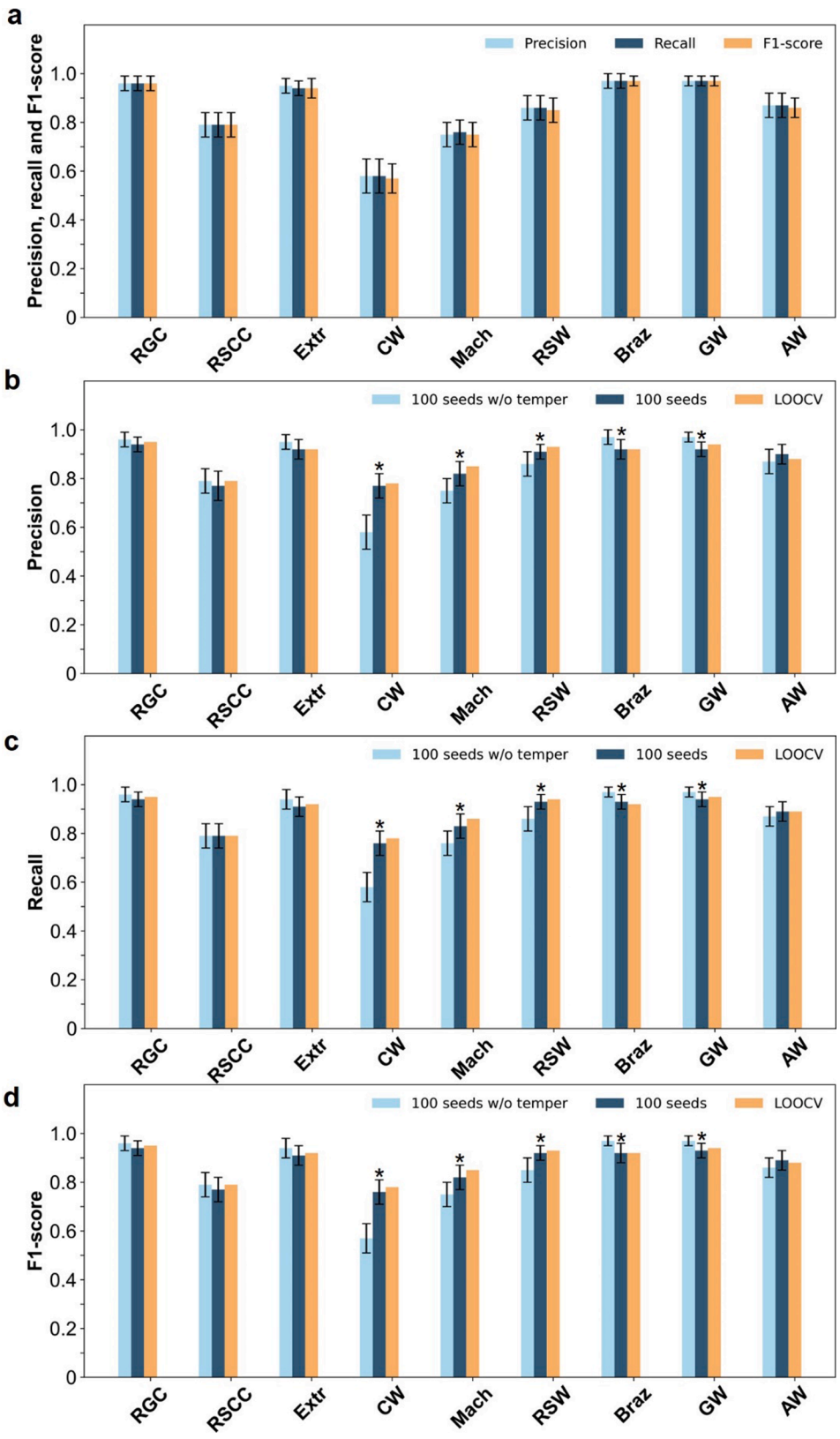


Fig. 7. Evaluation metrics of the developed RF models in predicting the technological properties. a. Precision, recall, and F1-score of the RF models that take compositions only as features or inputs. **b.** The precision of the RF models that take both compositions and tempers as feature inputs. The precision *P*-values of CW, Mach, RSW, Braz, GW are as low as 3.77e-54, 1.45e-21, 1.14e-14, 8.43e-20, 1.25e-30 while for RGC, RSCC, Extr and AW the *P*-values are 3.03e-5, 1.37e-2, 6.1e-8 and 5.56e-6. **c.** Recall of the RF models that take both compositions and tempers as feature inputs. The recall *P*-values of CW, Mach, RSW, Braz, GW are as low as 3.30e-56, 2.90e-20, 3.46e-26, 1.18e-23, 3.04e-15 while for RGC, RSCC, Extr and AW the *P*-values are 6.60e-6, 3.07e-1, 1.82e-7 and 4.36e-7. **d.** F1-score of the RF models that take both compositions and tempers as feature inputs. The F1-score *P*-values of CW, Mach, RSW, Braz, GW are as low as 9.79e-57, 7.76e-20, 1.58e-23, 1.28e-23, 1.64e-21 while for RGC, RSCC, Extr and AW the *P*-values are 1.75e-5, 5.60e-2, 2.10e-7 and 1.78e-6. The stars suggest a significant difference in the evaluation metrics between the models developed without the temper input and with the temper input. *P*-values lower than 1e-10 are considered as a significant difference and marked with a star (*).

for ML-assisted alloy development. Therefore, we performed all three experiments (1-seed, 100-seeds, and LOOCV experiments) to explore whether the inclusion of BHN as an additional feature would improve the predictions of other mechanical properties.

All three experiments suggested that by including BHN the predictions of UTS, TYS, and Shear were greatly improved while predictions of elongation and endurance were not influenced much, based on the evaluation metrics of R^2 and MAPE (Fig. 5a,b). For instance, not only the average R^2 of 100-seeds experiments was increased by 10 % or more for UTS, TYS, and Shear, but also the standard deviation of R^2 was

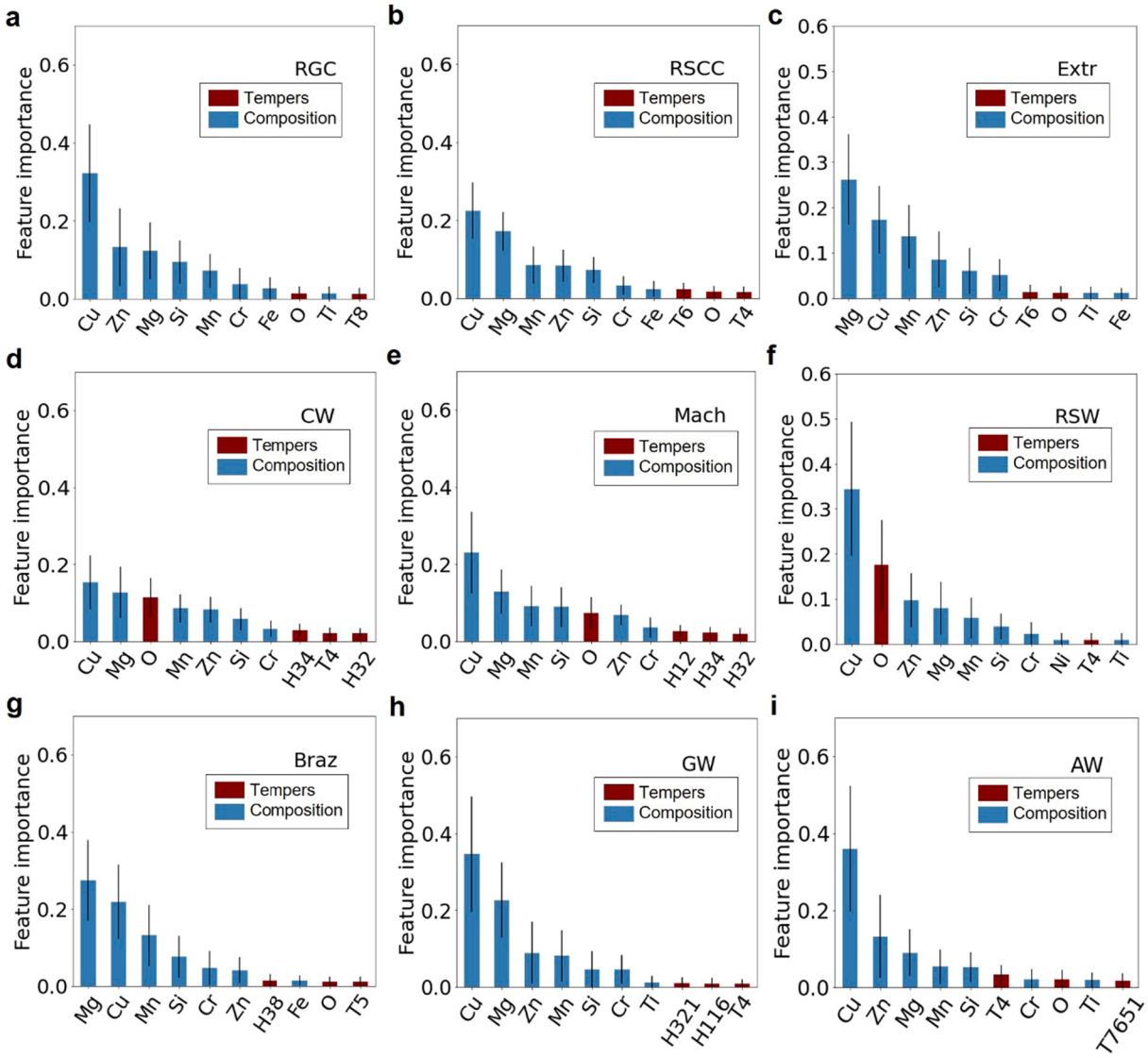


Fig. 8. The importance scores of the top-10 most important features identified by the RF model for each technological property. a. RGC. **b.** RSCC. **c.** Extr. **d.** CW. **e.** Mach. **f.** RSW. **g.** Braz. **h.** GW. **i.** AW. Red bars represent tempers while blue bars represent compositions. The thin black lines are the error bars.

much smaller after including BHN (Fig. 5a). It suggested that including BHN as an input, both improves the accuracy and reduces the uncertainty of prediction for UTS, TYS, and Shear. We performed the student's *t*-test to statistically examine the significance of prediction improvement brought by the inclusion of BHN. When testing whether there is a significant difference of the R^2 metric, very low *P*-values of 2.36e-40, 1.65e-37, and 1.72e-40 were respectively acquired for UTS, TYS, and Shear, whereas much larger *P*-values of 3.20e-1 and 1.72e-2 were obtained for El and Endurance (Fig. 5a). A similar pattern was observed for the MAPE metric (Fig. 5b). This supported our finding that including BHN into the feature space significantly improved the prediction for UTS, TYS, and Shear, not for EI and Endurance (Fig. 5a,b). We further examined the predicted property values of all the samples from the LOOCV experiments against their experimental values (Fig. 6). For UTS, TYS, and Shear, the predicted values are closely aligned to the diagonal line the location of which represents a perfect match between the predicted and experimental values (Fig. a,b,d). For El and Endurance, the scatter plots look very similar to the ones not including BHN as an input (Fig. 6c,e). Therefore, including BHN in ML alloy design for predicting UTS, TYS, and Shear is a viable option.

A potential problem in working with small datasets is that the results may not be generalized well since the small test set might be biased and unreliable [16]. However, our results are verified for 100 distinct random test splits with small but reasonable sample sizes (188 and 48 samples for training and test specifically). Therefore, testing the ML models by considering the randomness effect delivers a robust and reliable understanding of the model's performance, especially on a small dataset. Furthermore, large datasets are usually not available in real-life alloy discovery problems and may not necessarily result in significantly more accurate models. For example, in a study [14] a relatively large dataset of 713 samples was used to predict UTS using composition, temper, and the Brinell hardness and achieved a 10-fold cross-validation average error of 4.26 % which is very close to our LOOCV MAPE of 5.31 % (Fig. 5a).

3.5. *Prediction of technological properties directly from compositions or from compositions and tempers*

Technological properties are desirable for the manufacturing processes, such as casting, welding, forming, and machining. They define the suitability of the material for different fabrication processes and applications. Different from mechanical properties, some technological properties may not be affected by temper much. For instance, weldability and castability are only related to the composition, since they are

liquid-state processes. Therefore, we first examined how accurate compositions themselves could predict these nine technological properties using the RF classifier. Like the mechanical properties, the default hyperparameters of the RF classifier showed similar performance to the optimized hyperparameters for each of the technological properties (data not shown). Thus, we only showed the evaluation metrics of precision, recall, and F1-score from the developed RF models with the default hyperparameters. We did LOOCV and 100-seeds experiments to examine the model robustness to randomness caused by data splits. RF models yielded very similar values of precision, recall, and F1-score (Fig. 7a); thus, in the following sections, we only describe the F1-scores of the models for simplicity. Using only compositions, RGC, Extr, Braz, and GW were predicted with excellent F1-score of 0.96, 0.94, 0.97, and 0.97 specifically (Fig. 7a). RSCC, RSW, and AW were predicted well with F1-scores of 0.79, 0.85, and 0.86 specifically (Fig. 7a). Mach was predicted slightly worse with an F1-score of 0.75 (Fig. 7a). CW was predicted the worst with an F1-score of 0.57 (Fig. 7a).

We further investigated the influence of including both compositions and tempers as inputs on the prediction performance measured by precision, recall, and F1-score (Fig. 7b,c,d). Both LOOCV and 100-seeds experiments were repeated as above except in this case RF models were trained and developed with the additional inputs of tempers. We compared the 100-seed results of the RF models trained by composition only as well as by composition and temper. Including tempers as inputs greatly improved the F1-scores of the RF models in predicting CW, Mach, and RSW; a slight F1-score increase was observed for AW (Fig. 7d). By contrast, the prediction of RGC, RSCC, Extr, Braz, and GW got worse by including tempers as additional inputs (Fig. 7d). The student's *t*-test was performed to examine whether there is a significant difference of the model performance between including and excluding tempers. The F1-score *P*-values of CW, Mach, RSW, Braz, and GW were as low as 9.79e-57, 7.76e-20, 1.58e-23, 1.28e-23, 1.64e-21 respectively. While for RGC, RSCC, Extr, and AW, much larger F1-score *P*-values were acquired as 1.75e-5, 5.60e-2, 2.10e-07, and 1.78e-06, specifically (Fig. 7d). This similar trend was observed for both precision and recall *P*-values (Fig. 7b,c). With a stringent significance threshold of 1e-10, we may conclude that by including tempers, the model's predictive performance has significantly increased for CW, Mach and RSW, decreased for Braz and GW (Fig. 7b,c,d), and has no significant difference for RGC, RSCC, Extr, and AW. We hypothesized the properties would be predicted better by including tempers as inputs if these properties were affected much by tempers; on the contrary, for properties not affected much by tempers, the inclusion of tempers as inputs could worsen their predictions since it produced a much higher input dimension without providing much useful information.

The feature importance study supported our hypothesis (Fig. 8). As introduced in the mechanical property section, the feature importance score suggests how important a feature is in predicting the output. For RGC, RSCC, Extr, Braz, GW, and AW, tempers were ranked low with importance scores smaller than 0.05 (Fig. 8a, b, c, g, h, i); for CW, Mach, and RSW, among the top 10 features ranked by the importance score, the O temper was ranked very high, 3rd, 5th, and 2nd specifically (Fig. 8d,e, f). This explained why the inclusion of tempers predicted CW, Mach, and RSW significantly better but didn't help much in the prediction of all the others, compared to using compositions only. In particular, the predictions of Braz and GW were significantly worse; this is likely due to the dramatically increased dimension of input features that do not contain much useful information.

More interestingly, when looking at the top predictive features ranked by our RF models, some ranks are supported by the established processing (composition/temper)-property relationships while a few others provide indications on how one property could be affected by compositions or tempers that is not known. For example, Cu content is the most predictive feature for RGC (Fig. 8a) and RSCC (Fig. 8b) with the highest feature importance score. This is consistent with the strong influence of Cu content to Al corrosion potential. It is interesting to notice that Mg content is also a strong predictor, followed by Mn and Zn, for RSCC. Since the fundamental mechanism of stress corrosion cracking is still unclear, exploring how Mg, Mn and Zn affect RSCC may significantly improve our understanding of this phenomenon. Extr of an alloy is the maximum relative extrusion speed, which allows to obtain extruded profile without tearing or cracking. It is dependent on alloy compositions. In general, the existence of Mg (5xxx), Cu (2xxx) and Zn (7xxx) is responsible for poor Extr [34]. Consistent with these observations, compositions such as Mg, Cu, Mn, and Zn were ranked as the most-predictive features while tempers were showed to have little influence in predicting Extr (Fig. 8c). CW is defined as the relative ease with which a metal can be shaped through plastic deformation. It is related to TYS that is strongly affected by Cu and O temper. It is thus making sense for RF to rank Cu and O temper among the top 3 predictors for CW (Fig. 8d). In addition, work hardening, which is responsible for CW, is the major strengthening mechanism for Mn-containing and Mg-containing solid-solution Al alloys. Consistent with this, Mn and Mg stand out in the feature importance chart as well (Fig. 8d). For Mach, the compositions of Cu, Mg, Mn and Si, and O temper are among the top predictors (Fig. 8e). As the property of Mach is complicated, how Mach is quantitatively related to the above predictors needs further investigation. RSW is a property related to solid state welding process and is determined by the alloy's electrical conductivity that are mostly related to Cu content and O temper (highest conductivity state). Remarkably, our RF model for RSW also ranked Cu content and O temper as the top-2 predictive features (Fig. 8f). For predicting Braz, GW and AW (Fig. 8g, h, i), tempers in general have very low scores. GW and AW are mostly dependent on compositions since they are the properties related to the solidification. Therefore, the low predictive ability of tempers for GW (Fig. 8h) and AW (Fig. 8i) is in good agreement with this. Overall, this feature importance study suggested that our RF model was able to capture many of the true underlying processing-property relationships and further use these for reliable and accurate prediction. At the same time, it revealed a few novel relationships that warrant further investigation.

So far, there are limited research efforts in understanding the performance of ML algorithms in predicting technological properties [15,36–40]. One of the reasons could be the lack of data on technological properties[35]. Here, we have collected a small but decent set of data points (236) containing a comprehensive set of nine important technological properties for commercial wrought Al alloys. We have thoroughly studied the performance of ML algorithms in predicting the nine technological properties by examining the bias-variance trade-off, filling this knowledge gap on ML and technological properties of wrought Al alloy. We have also shown that the inclusion of tempers into the feature space can be beneficial for predicting some properties (CW, Mach and RSW), but not all of them. The subsequent RF feature importance studies are consistent with the established processing (and composition)-microstructure-property relationship, which further confirmed the effectiveness of our ML models.

4. Conclusions

In this work, we collected a small but comprehensive dataset of commercial wrought Al alloys and investigated the applicability of supervised ML models in alloy design using such a realistic small dataset. We systematically assessed the ML models with a focus on their robustness and effectiveness on predicting a full set of alloy properties. We performed 1-seed, 100-seeds, and LOOCV experiments to better understand the bias-variance trade-off of the developed ML models in predicting unseen test data points. We demonstrated that the developed RF models were accurate and reliable in predicting the full set of alloy properties, with generally low bias and variance. Our predictions are comparable with if not better than previous studies that used similar input features/outputs but larger datasets. We have effectively modeled a comprehensive collection of technological properties that are critical

for the manufacturing processes but have not been thoroughly studied before. In addition, we studied the impact of engineering the input (feature) space on the effectiveness of the ML models, by adding BHN to mechanical property predictions and adding temper to technological property predictions. We found the inclusion of BHN significantly improved the prediction of UTS, TYS and Shear but didn't affect the prediction of El and Endurance. The addition of temper greatly improved the prediction of CW, Mach, and RSW, but substantially worsened the prediction of Braz and GW. These findings may benefit other researchers in selecting or measuring the proper collection of features for developing ML models for different properties. The feature importance study not only confirmed the soundness of our ML models but also revealed the underlying processing-(micro)structure–property relationship. In summary, using a realistic small commercial dataset, supervised ML models such as RF can be applied to predict the whole set of properties including both continuous mechanical properties and discrete technological properties for wrought Al alloy. The methodology used in this study can be readily applied to study the processing-property relationship in other alloys.

CRediT authorship contribution statement

Yasaman J. Soofi: Data curation, Formal analysis, Methodology, Investigation, Validation, Writing – original draft, Visualization. **Md Asad Rahman:** Methodology, Visualization. **Yijia Gu:** Conceptualization, Data curation, Methodology, Supervision, Writing – review & editing, Funding acquisition. **Jinling Liu:** Conceptualization, Data curation, Methodology, Supervision, Writing – review & editing, Project administration, Funding acquisition.

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.

Data and code availability

The compiled dataset used in this study is included as supplemental materials. The codes are publicly available on GitHub at https://github.com/yasamanjs/ML-assisted-alloy-design-using-wrought-aluminum-alloys.git.

Acknowledgement

This work was supported in part by the seed grant from the Intelligent Systems Center (ISC) at Missouri University of Science and Technology.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.commatsci.2022.111783.

References

[1] J. Allison, D. Backman, L. Christodoulou, Integrated computational materials engineering: a new paradigm for the global materials profession, Jom. 58 (2006) 25–27.
[2] W. Xiong, G.B. Olson, Cybermaterials: materials by design and accelerated insertion of materials, npj Comput. Mater. 2 (2016) 15009, https://doi.org/10.1038/npjcompumats.2015.9.
[3] W. Xiong, G.B. Olson, Integrated computational materials design for high-performance alloys, MRS Bull. 40 (2015) 1035–1044, https://doi.org/10.1557/mrs.2015.273.
[4] C. Shen, C. Wang, X. Wei, Y. Li, S. van der Zwaag, W. Xu, Physical metallurgy-guided machine learning and artificial intelligent design of ultrahigh-strength stainless steel, Acta Mater. 179 (2019) 201–214.
[5] J.F. Durodola, Machine learning for design, phase transformation and mechanical properties of alloys, Prog. Mater Sci. 123 (2022), https://doi.org/10.1016/j.pmatsci.2021.100797.
[6] G.L.W. Hart, T. Mueller, C. Toher, S. Curtarolo, Machine learning for alloys, Nat. Rev. Mater. 6 (2021) 730–755, https://doi.org/10.1038/s41578-021-00340-w.
[7] C. Wen, Y. Zhang, C. Wang, D. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman, Y. Su, Machine learning assisted design of high entropy alloys with desired property, Acta Mater. 170 (2019) 109–117, https://doi.org/10.1016/j.actamat.2019.03.010.
[8] S. Malinov, W. Sha, Application of artificial neural networks for modelling correlations in titanium alloys, Mater. Sci. Eng., A 365 (2004) 202–211, https://doi.org/10.1016/j.msea.2003.09.029.
[9] C. Zou, J. Li, W.Y. Wang, Y. Zhang, D. Lin, R. Yuan, X. Wang, B. Tang, J. Wang, X. Gao, H. Kou, X. Hui, X. Zeng, M. Qian, H. Song, Z.K. Liu, D. Xu, Integrating data mining and machine learning to discover high-strength ductile titanium alloys, Acta Mater. 202 (2021) 211–221, https://doi.org/10.1016/j.actamat.2020.10.056.
[10] C. Wang, H. Fu, L. Jiang, D. Xue, J. Xie, A property-oriented design strategy for high performance copper alloys via machine learning, npj Comput. Mater. 5 (2019), https://doi.org/10.1038/s41524-019-0227-7.
[11] J. Peng, Y. Yamamoto, J.A. Hawk, E. Lara-Curzio, D. Shin, Coupling physics in machine learning to predict properties of high-temperatures alloys, npj Comput. Mater. 6 (2020), https://doi.org/10.1038/s41524-020-00407-2.
[12] J. Li, Y. Zhang, X. Cao, Q. Zeng, Y. Zhuang, X. Qian, H. Chen, Accelerated discovery of high-strength aluminum alloys by machine learning, Communications Materials. 1 (2020), https://doi.org/10.1038/s43246-020-00074-2.
[13] M. Hu, Q. Tan, R. Knibbe, S. Wang, X. Li, T. Wu, S. Jarin, M.X. Zhang, Prediction of Mechanical Properties of Wrought Aluminium Alloys Using Feature Engineering Assisted Machine Learning Approach, Metallurgical and Materials Transactions A: Physical Metallurgy and Materials, Science 52 (2021) 2873–2884, https://doi.org/10.1007/s11661-021-06279-5.
[14] D. Merayo, A. Rodríguez-Prieto, A.M. Camacho, Prediction of mechanical properties by artificial neural networks to characterize the plastic behavior of aluminum alloys, Materials. 13 (2020) 1–22, https://doi.org/10.3390/ma13225227.
[15] J. Ling, E. Antono, S. Bajaj, S. Paradiso, M. Hutchinson, B. Meredig, B.M. Gibbons, MACHINE LEARNING FOR ALLOY COMPOSITION AND PROCESS OPTIMIZATION, 2018. https://citrination.com.
[16] T. Thankachan, K.S. Prakash, C. David Pleass, D. Rammasamy, B. Prabakaran, S. Jothi, Artificial neural network to predict the degraded mechanical properties of metallic materials due to the presence of hydrogen, Int. J. Hydrogen Energy 42 (2017) 28612–28621, https://doi.org/10.1016/j.ijhydene.2017.09.149.
[17] J.G. Kaufman, Introduction to aluminum alloys and tempers, ASM international (2000).
[18] J.G. Kaufman, Aluminum Alloy Database, Knovel. (2015). https://app.knovel.com/hotlink/toc/id:kpAAD00001/aluminum-alloy-database/aluminum-alloy-database (accessed April 15, 2020).
[19] Matmatch, (n.d.). https://matmatch.com/ (accessed April 15, 2020).
[20] MakeItForm, (n.d.). https://matmatch.com/ (accessed April 15, 2020).
[21] ALUSELECT Material Database, (n.d.). https://matplus.eu/en-us/products/databases/aluselect/ (accessed April 15, 2020).
[22] I.O. for S. (ISO), Steel–conversion of elongation values–part 1: carbon and low alloy steels, (1984).
[23] L. Myers, M.J. Sirois, Spearman correlation coefficients, differences between, Encyclopedia of Statistical Sciences. 12 (2004).
[24] G. Louppe, L. Wehenkel, A. Sutera, P. Geurts, Understanding variable importances in forests of randomized trees, Adv Neural Inf Process Syst. 26 (2013).
[25] B.H. Menze, B.M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, F. A. Hamprecht, A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC Bioinf. 10 (2009) 1–16.
[26] M. Krzywinski, N. Altman, Significance, P values and t-tests, Nat Methods. 10 (2013) 1041–1042.
[27] S. Kou, A simple index for predicting the susceptibility to solidification cracking, Weld J. 94 (2015) 374s–388s.
[28] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol Rev. 65 (1958) 386.
[29] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks. 61 (2015) 85–117.
[30] N. Cristianini, J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge University Press, 2000.
[31] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans Pattern Anal Mach Intell. 20 (1998) 832–844.
[32] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[33] P. Probst, M.N. Wright, A. Boulesteix, Hyperparameters and tuning strategies for random forest, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 9 (2019), e1301.
[34] T. Sheppard, Extrusion of aluminium alloys, Springer Science & Business, Media (1999).
[35] L.B. Coelho, D. Zhang, Y. van Ingelgem, D. Steckelmacher, A. Nowé, H. Terryn, Reviewing machine learning of corrosion prediction in a data-oriented perspective, npj Mater. Degrad. 6 (2022) 1–16.