Prediction of Mechanical Properties of Wrought Aluminium Alloys Using Feature Engineering Assisted Machine Learning Approach

MINGWEI HU, QIYANG TAN, RUTH KNIBBE, SEN WANG, XUE LI, TIANQI WU, SAMS JARIN, and MING-XING ZHANG

Data-mining based machine learning (ML) method is emerging as a strategy to predict aluminium (Al) alloy properties with the promise of less intensive experimental work. However, ML models for wrought Al alloys are limited due to the difficulty in feature digitalization of the variety of manufacturing processes. Hence, most previous studies were constrained to specific alloy designations, which impeded the applicability of those ML models to broader wrought Al alloys. In the present work, we propose a novel feature engineering, called procedure-oriented decomposition (POD), assisting prediction framework to address the complexity introduced by manufacturing processes for wrought Al alloys. In this model, both chemical compositions and manufacturing processes are integrated as features. Correlation mapping of these features to the wrought Al alloys mechanical properties is established using the support vector regressor (SVR) model. The prediction framework demonstrates a high prediction accuracy and potential to design new alloys.

https://doi.org/10.1007/s11661-021-06279-5

I. INTRODUCTION

ALUMINIUM (Al) alloys are workhorse materials in aerospace and automotive industries owing to their low density, high specific strength, good corrosion resistance and formability, and low manufacturing and maintenance costs.[1–3] However, compared with steels, the strength of Al alloys is still relatively low, undermining their widespread application, particularly in the heavy-duty and rigorous environment.[4] Therefore, pursuing new Al alloys with high strength and ductility has been a long-term challenge for materials scientists and engineers to achieve weight reduction in aircraft and vehicles.[3,5] To facilitate the development process, great effort has been devoted to predicting the mechanical properties of Al alloys.

The mechanical properties of alloys are governed by their microstructures, which in turn are determined by chemical composition and manufacturing processes. Hence, typical metallurgical-property prediction uses a two-step process. In the first step, the microstructure is predicted or simulated from the composition/processes and in the second step, the material property is linked to the microstructure.[6] Many empirical, physical metallurgy models were established to simulate the strengthening mechanisms and contributions to strength. Typical models include the grain boundary effect—depicted by the Hall-Petch relation;[7,8] solid-solution strengthening—described by the Fleischer equation;[9] dislocation strengthening—expressed by the Bailey-Hirsch relation;[10] precipitation strengthening—governed by the Orowan equation or dislocation shearing mechanism[11] and *etc*. These constitutive models quantitively connect the microstructure of polycrystalline metallic alloys to a particular strengthening mechanism, which can be employed to predict the alloy strength by linear summation. Cao *et al.*[12] calculated the total strength of an ultrafine-grained (UFG) 5083 Al alloy, using linear summation based on the abovementioned mechanistic contributions. However, the computational result deviated from the experimental result due to uncertainties related to dislocation density and precipitates.[12] Therefore, Ma *et al.*[13] only used these mechanisms to identify the dominant strengthening mechanism in the UFG 7075 Al alloy. Although these classic models provided quantitative theories of strengthening mechanisms, the microstructure parameters used in the models are experimentally obtained,

MINGWEI HU, QIYANG TAN, RUTH KNIBBE, SAMS JARIN, and MING-XING ZHANG are with the School of Mechanical and Mining Engineering, The University of Queensland, St. Lucia, QLD, 4072, Australia. Contact e-mail: q.tan@uq.edu.au; mingxing.zhang@uq.edu.au. SEN WANG, XUE LI, and TIANQI WU are with the School of Information Technology and Electrical Engineering, The University of Queensland, St. Lucia, QLD, 4072, Australia.

Manuscript submitted September 16, 2020; accepted March 31, 2021.

Article published online April 26, 2021

considered as a high-cost and time-consuming method.[14] Moreover, there is no effective physical-metallurgy model to quantify the ductility of metals due to the complexity in the plastic stage.

In the last few decades, computational material science (CMS), such as integrated computational materials engineering (ICME),[15] has been the predominant strategy to simulate the microstructure of materials.[16–18] This strategy largely relies on the integration of physics-based multiscale models/theories and mathematic equations to tackle the process-structure-property correlation.[19–21] Du *et al*.[22] proposed a CALPHAD-coupled mathematical simulation for microstructural features of precipitates in Al alloys. This mathematical model accurately predicted the mean radius, but not the volume fraction, of the precipitates—probably due to the inadequate assumption of the thermodynamic data. Another result was reported later by Gu *et al*.[23] who used cellular automaton (CA) method in conjunction with process model and successfully simulated the grain size of Al alloy at different cooling rates. These two instances illustrate the application of CMS in predicting microstructural features, which can then feed into the abovementioned metallurgical models to determine strength. For ductility prediction, Hannard's group[24] integrated finite element (FE) and CA method into the multiscale, void nucleation models. This framework was then employed to simulate the deformation process of 6× series Al alloys with different heat treatments. The fracture strains were adequately predicted and the substantial influence of heat treatment on mechanical properties was also foreseen. CMS tools have successfully understanded the influences of individual factors, such as solute and thermal processing parameters, on the microstructure and properties of alloys. However, the remaining challenges need to see the integration of multiscale models to clarify multiple factors, such as properties, defects, and interactions of alloying elements.[19,25] Moreover, as CMS is strongly dependent on the comprehension of existing alloy systems, it can hardly predict for the new system.[26]

Machine learning (ML), based on big data analytics, is a burgeoning tool to predict the macroscopic properties of metallic materials[26–28] directly from their compositions and processes. This algorithm builds surrogate models that directly connect inputs and outputs with non-linear relationships, revealing the underlying patterns, trends, and connections.[29] By obviating the microstructure as a "bridge" between composition/process and material property, ML models could avoid the accumulation of errors generated through the complexity of microstructure quantification.[30] Thus, with ML algorithm, alloy properties can be directly correlated to the composition and the process without consideration of the microstructure. Applications of ML strategy on predicting the Al alloys properties have emerged in recent years.[4,31,32] In Belayadi's work,[33] the hardness of 7× series Al alloys was successfully predicted by an artificial neural network (ANN)—a typical ML model. However, the samples in the database fell into narrow input ranges, incapable of embodying the applicability of the model on new samples. Dey *et al*.[34] also used an ANN model (with 259 samples) to predict the mechanical properties of age-hardenable wrought Al alloys. Considered composition and tempering conditions as the inputs, a new age-hardenable Al alloy composition was designed. Unfortunately, the experimental mechanical properties of this new alloy significantly diverged from the prediction, implying inadequate generalizability (*i.e.* the prediction ability on new samples) of the trained ML models. This inaccuracy was likely attributed to the insufficient samples and imbalanced data points in the distribution range.[35] In addition, this ANN model only took the numerical processing parameters, such as heat-treatment temperature and time, into account without consideration of the manufacturing process (*e.g.* rolling, extrusion, etc.) and production shape (*e.g.* plate, bar, etc.). These non-numerical (*i.e.* categorical) inputs have significant impacts on the alloy properties but are difficult to get involved in ML models. Therefore, the abovementioned ML models can be exclusively used for specific types, or a small portion, of Al alloys where limited processing features are involved. Accordingly, the extracted correlation is not comprehensive, limiting the generalizability of prediction results.

To address these limitations, a new prediction framework driven by feature engineering assisted ML model is proposed in the present work. This algorithm aims to establish quantitative inference mapping from chemical composition and manufacturing process (processing parameters and method), combined as features, to mechanical properties including yield strength (YTS), ultimate tensile strength (UTS) and elongation (ELONG), called targeted properties, for wide-range wrought Al alloys. The feature engineering effectively modifies the troublesome feature set into robust factors of targeted properties with accelerated computational process and boosted accuracy.

II. MODELLING STRATEGY

Figure 1 schematically depicts the prediction framework applied to the mechanical properties of wrought Al alloys in the present study. The framework begins with the construction of the database, which contains both features (*i.e.* chemical composition and manufacturing process) and targeted properties (*i.e.* YTS, UTS and ELONG). The database is preprocessed to ensure efficient incorporation of features in the ML training and some irrelevant features are eliminated by feature engineering subsequently. ML regression models are trained based on the modified database to infer a quantitative relationship between features and targeted properties. The robustness of the obtained ML model is finally assessed by validating predicted properties of the testing set, in which all samples are reported in recent literatures and independent of model training.

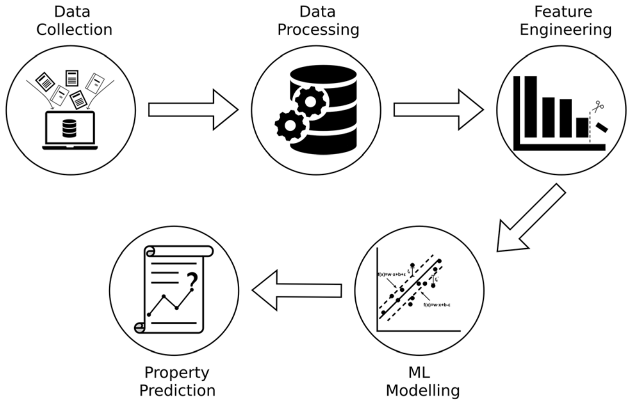


Fig. 1—Flow chart of the present prediction framework driven by feature engineering assisted ML model.

III. MACHINE LEARNING MODELLING

A. *Database*

In this work, commercially available engineering wrought Al alloys and new alloys with modified composition are collected to construct the database. Total of 930 samples were collected from both the Al alloy handbook[36] and peer-reviewed literatures (corresponding references are listed in Electronic Supplementary Table S1). Following criteria were used to select samples from literature: (i) alloy composition and processing method are not included in the ASM handbook; (ii) alloys were processed with common thermomechanical processes *e.g.* rolling, extrusion and forging (nano-crystallization, such as ECAP, is excluded), and heat-treatment methods, *e.g.* T3, T4 and T6; and (iii) tensile property data is available. It should be noted that, to constrain the feature space, only Al alloys processed with conventional forming methods (*e.g.* rolling, forging and extrusion) are considered, whereas the advanced severe plastic deformation (SPD) technologies, such as equal channel angular pressing (ECAP) and tube channel pressing (TCP), were not included to avoid excessive features brought by their complex processing steps. Furthermore, compared with the conventional processes, the SPD technologies are not widely used in industrial production due to the complex procedure and size limitation.[37] Considering that the variables in a feature (*e.g.* aging temperature and aging time) may interact with others and synergistically affect the targeted properties, different subsets of features are included in the database. In the database, each sample is described by its unique feature set, which consists of chemical composition and manufacturing process. Manufacturing process may either be in numerical (*e.g.* annealing temperature) or string format (*e.g.* extruded manufacturing method).

Unlike the ML databases in previous studies,[33,34] which only involved numerical features, the categorical features are included in the present database. This enables the integration of complex fabrication characteristics of wrought Al alloys into ML modelling. As these manufacturing processes significantly influence the properties of alloys, the prediction accuracy can be improved. Table I lists the minimum, maximum, average, and standard deviation of the numerical features and targeted properties of collected samples and the categorical features will be discussed in following sections.

B. *Data Preprocessing*

Normalization is a standard data preprocessing (DP) method for numerical features, especially when these features cover a wide range in magnitude, inducing a skewed correlation map. Proper normalization not only accelerates the training process but also reduces the prediction error of models.[38] In the present database, most of the compositional values are orders of magnitude less than the processing parameters (*e.g.* heat treatment temperature and time) as shown in Table I. Min-max and z-score normalizations[39] are two conventional approaches to scale the value magnitudes into the interval of [0, 1] and [− 1, 1], expressed by the Eqs. [1] and [2] respectively:

$$x^{'} = \frac{x - \min}{\max - \min} \qquad [1]$$

$$x^{'} = \frac{x - \text{mean}}{\text{std}} \qquad [2]$$

where x and $x^{'}$ are the original and normalized features, respectively;max and min denote the upper and lower limits of the original features; mean and std stand for the average and standard deviation of the features.

Categorical features can be encoded from string format into numerical format for ML modelling. Normal encoding methods include ordinal label encoding, binary encoding, one-hot encoding, and sum encoding. Ordinal label encoding assigns an integer to each unique category in alphabetical order. Binary encoding converts integers into binary form and then separates them into columns upon the length of digits. One-hot encoding is a widely-used encoding method for unordered features. It replaces the original feature column with k binary columns, where k is the number of distinct variables within the feature. Value '1' indicates the presence of a specific category, while '0' means absence. The sum encoding creates $k - 1$ columns, which follows the same principle as the one-hot encoding method for $k - 1$ categories, but the last category employs '− 1' for all columns. Here, we employ a DP scheme composed of z-score normalization and one-hot encoding method on the database and the effect of different DP schemes will be illustrated in Section IV–B.

C. *Feature Engineering*

In addition to the chemical composition, the manufacturing process is another critical factor that determines the microstructure and consequently influences the mechanical properties of wrought Al alloys. Hence, the relevant manufacturing process, generally represented by a temper designation, should be considered as inputs

Table I. Numerical Features and Targeted Properties in the Database with Minimum, Maximum, Average and Standard Deviation Values

	Variable (Unit)	Min.	Max.	Average	Std.
Features	Si (wt pct)	0	12.25	0.381	0.761
	Fe (wt pct)	0	1.6	0.343	0.305
	Cu (wt pct)	0	6.3	1.587	2.013
	Mn (wt pct)	0	2.06	0.324	0.380
	Mg (wt pct)	0	5.8	1.247	1.361
	Cr (wt pct)	0	0.5	0.068	0.092
	Ni (wt pct)	0	2	0.041	0.225
	Zn (wt pct)	0	8.69	0.682	1.707
	Ga (wt pct)	0	0.03	0.001	0.003
	V (wt pct)	0	0.158	0.006	0.022
	Ti (wt pct)	0	0.2	0.048	0.070
	Zr (wt pct)	0	0.4	0.029	0.059
	Bi (wt pct)	0	0.55	0.011	0.068
	Pb (wt pct)	0	0.55	0.011	0.068
	Li (wt pct)	0	4	0.171	0.625
	B (wt pct)	0	0.06	0.001	0.006
	Sc (wt pct)	0	1.44	0.050	0.172
	Be (wt pct)	0	0.1	0.000	0.007
	Solution/annealing T (°C) (SA Temp)	25	649	327.129	230.421
	Ageing T (°C) (A Temp)	25	290	89.121	84.413
	Ageing Time (h) (A Time)	0	1440	18.973	105.008
Targeted Properties	YTS (MPa)	9	684	246.562	139.443
	UTS(MPa)	34	732	311.470	151.844
	Elongation (pct EI)	0.5	50	12.583	8.078

for the ML model. However, the large number of temper designations gives rise to the challenge on the encoding process of categorical inputs. This, in turn, may undermine computational speed and robustness of the inference model.[40] In the present study, 14 strain hardening methods and 51 temper designations are involved in the original database, resulting in a feature space with around one hundred dimensions after one-hot encoding. The number of features should be reasonably controlled to avoid the curse of dimensionality.[41] Therefore, it is necessary to efficiently shrink down the high dimensionality of the feature space without abandoning any specified processing characteristics and relevant features.

In this work, we adopt a novel feature engineering method combining the random forest (RF) algorithm[42] and the procedure-oriented decomposition (POD) method. The categorical features 'strain-hardening method' (SM) and 'temper designation' (TD) in the initially constructed database are joined to define the tempering treatment for each sample. This description (i) generates a total of 65 features during data preprocessing; and (ii) implicitly contains excessive thermomechanical information in the TD. To overcome these drawbacks, a new POD method is proposed. It deconstructs the TD feature into several descriptive processing steps, including 'heat treatment medium' (HTM), 'ageing type' (AT), and 'treatment afterward' (TA). Figure 2 schematically shows the new POD method in the present work. Under these new categorical features, discrete values denoting specific thermomechanical treatment are listed in Table II. Through distinct permutations of these discrete values, the number of tempering treatment features is effectively reduced from 65 to 24 columns after one-hot encoding. Detailed examples of final feature vector translated from real-world processing treatment are shown in the electronic supplementary Section B.

After tackling the lengthy feature set, whether the characteristics in the temper process is neglected after the POD method should also be examined. Herein, RF algorithm is adopted to evaluate the variable importance (VI) of features. This algorithm is an ensemble model with a multitude of decision trees, where each tree contains a randomly sampled value vector within the same distribution.[42] The bootstrap with replacement method selects samples for the construction of an individual decision tree, while out-of-bag (OOB) samples (samples not included in the bootstrapped data) are used to assess the VI for each feature via Eqs. [3][43]:

$$VI(x^j) = \frac{1}{n}\sum\nolimits_{i=1}^{n}(err{\sim}OOB_i^j - errOOB_i) \qquad [3]$$

where $VI(x^j)$ is the variable importance of variable x^j; n denotes the number of decision tree in the RF; $errOOB_i$ is the error of tree i on OOB samples; and $err{\sim}OOB_i^j$ is the error of tree i on OOB samples with random permutation of variable x^j. The RF models are constructed based on two databases separately, one of which has the original temper designations and the other contains the thermomechanical processes decomposed from temper designations by the POD method. More

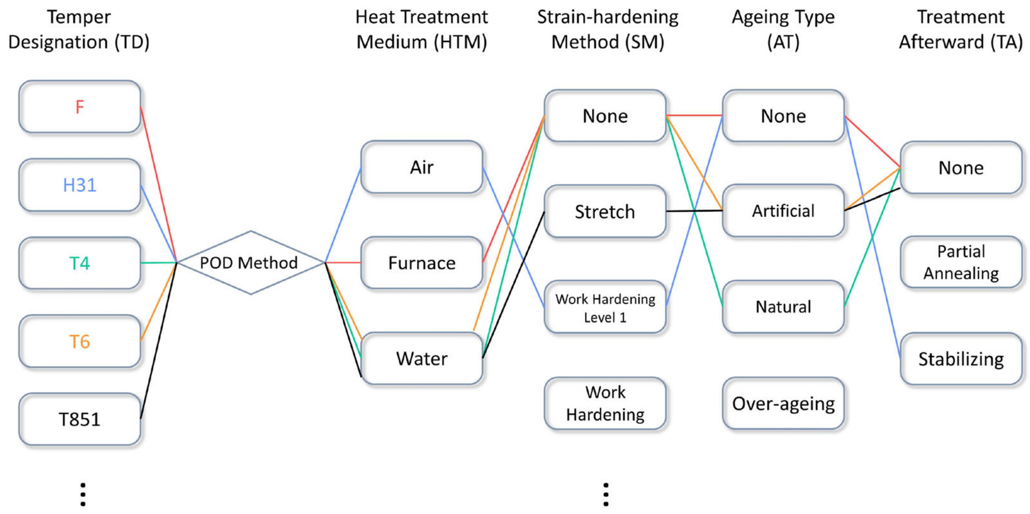


Fig. 2—Schematic illustration of the principle of the POD. This method decomposes the temper designation (TD) symbols into discrete thermomechanical processes under categorical features of ‘heat treatment medium’ (HTM), ‘strain-hardening method’ (SM), ‘ageing type’ (AT), and ‘treatment afterward’ (TA).

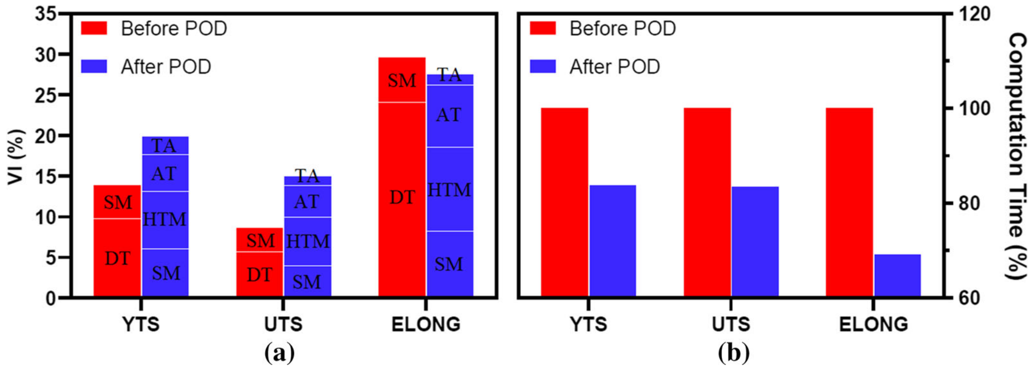
Table II. Categorical Features in the Database with All Distinct Values Enumerated Correspondingly

Categorical Features	Distinct Values
Manufacturing Method (MM)	drawn, extruded, forged, rolled, thixo-formed, unspecified
Production Shape (PS)	bar, die forging, fin stock, foil, forging, ingot, plate, ring, rod, sheet, tube, wire, unspecified
Heat Treatment Medium (HTM)	water, air, furnace
Strain-hardening Method (SM)	6 pct deformation, 7 pct deformation, compression, cry-rolling, ht work hardening, rolling, stretch, work hardening, work hardening level 1, work hardening level 2, work hardening level 4, work hardening level 5, work hardening level 6, work hardening level 8, work hardening level 9, none
Ageing Type (AT)	none, artificial, natural, over-ageing
Treatment Afterward (TA)	partial annealing, stabilizing, none

Note that in the SM, the hardening level refers to the degree of increase in tensile strength in H tempers as per handbook[36] whilst the x pct (“x” refers to numbers) deformation describes the thickness reduction during cold work. Work hardening levels 3 and 7 are not included due to the limited data to avoid data scattering.

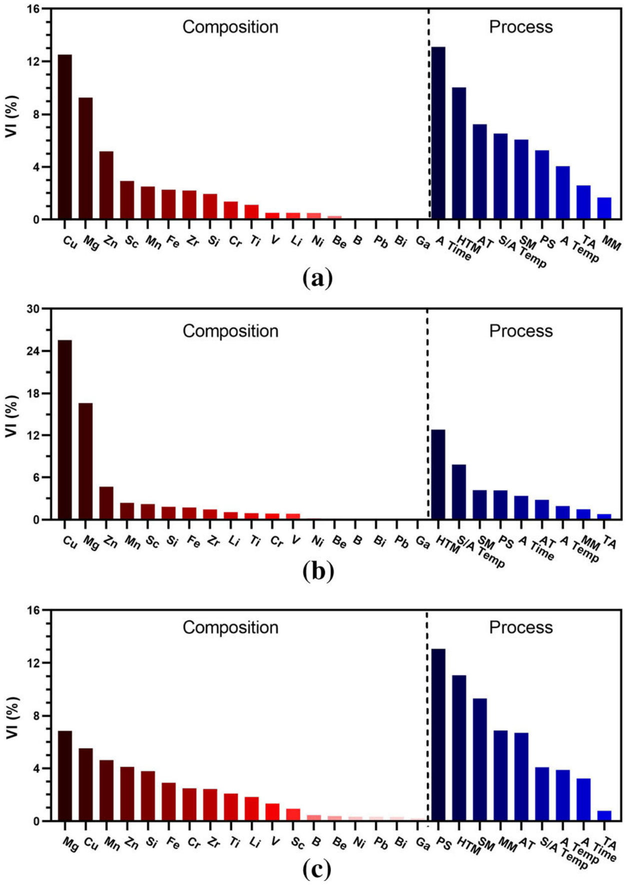
details of the RF models for feature engineering can be found in the electronic supplementary Section C. Figure 3(a) shows the comparison of VIs for the three targeted properties (YTS, UTS & ELONG), summed from only the temper-related categorical features. After data processing using the POD method, the overall VIs for YTS and UTS has been increased. It also shows a small reduction in VI for ELONG, indicating that the POD method has a negligible impact on creating an improved relevance between the temper features and the ELONG property. The computation time presented in Figure 3(b) indicates a significant improvement in the computation efficiency in terms of all the target properties after the POD processing. As such, the POD method successfully narrowed the feature columns down without sacrificing temper characteristics, but also accelerated the computational process.

Moreover, VI evaluation is also employed to understand the relationship between each feature and targeted properties quantitatively. In order to reduce the impact of different frequencies of each element, one hundred subsets of the database are randomly generated with more evenly distributed elements. The average VI values of features for the YTS, UTS, and ELONG of wrought Al alloys are shown in Figure 4, reflecting the effect of features on the prediction of properties. It is worth noting that the manufacturing process plays a critical role in determining the properties of wrought Al alloys, especially for YTS and ELONG. This highlights that ML method which overlooks the manufacturing process may lead to inaccuracy of predictions. In addition, some alloying elements are shown to have minimal impact on mechanical properties. However, it cannot be concluded that the effect of these alloying elements is negligible. For example, Li-containing Al alloys normally possess



(a) (b)

Fig. 3—The RF algorithm is adopted to confirm the effectiveness of the POD method. The red bars represent results obtained with the original database (*i.e.* includes categorical features of 'SM' and 'TD'), while the blue bars represent results obtained from the database processed by the POD method (*i.e.* consists of 'HTM', 'SM', 'AT', and 'TA'). (*a*) The VI comparison indicating the correlations between each temper treatment and the three targeted properties before and after POD processing. (*b*) The comparison of computation time for the three targeted properties before and after POD processing.



(a)

(b)

(c)

Fig. 4—VI ranking of compositional and processing features for (*a*) YTS; (*b*) UTS; (*c*) ELONG of wrought Al alloys. The black dotted lines separate the composition features from process features.

high strength. But, the majority of commercial high-strength, wrought Al alloys do not contain Li. Most of the alloying elements, with a high VI, are common alloying elements for improving wrought Al alloy mechanical properties—such as Cu, Mg, and Zn.

The VI results for YTS, shown in Figure 4(a), indicate that alloying elements, Cu, Mg and Zn, and process features, ageing time, heat treatment medium and ageing type, are predominant in strengthening the Al alloys. All the current high strength wrought Al alloys, including both 2××× and 7××× series, contain Cu, Mg, and Zn as core alloying elements. Heat treatment, ageing in particular, is the principal technique to increase the yield strength. Interestingly, these factors coincided with Chinh *et al*. study,[44] which demonstrated the synergistic effect of Cu, Zn, and Mg in improving the strength of a Cu-containing Al-Zn-Mg alloy by producing more ellipsoidal GP zones during the ageing process. One exception is the ageing temperature, which shows low VI in Figure 4. This does not mean the ageing temperature is not important. The low VI of ageing temperature is attributed to its small variation as only optimized ageing temperatures are listed in the current Al alloy handbooks, from which the majority of samples are collected. As shown in Figure 4(b), the critical features of UTS are similar to that of YTS but the composition shows stronger impacts—especially for Cu and Mg elements. In addition to the ageing time and heat treatment medium, the strain-hardening method is also critical in predicting UTS, as reported in other studies.[45,46] Interestingly, ELONG is more sensitive to the processing features, production shape, heat treatment medium, and strain-hardening method, than to the composition. This can be understood in terms of the strengthening mechanisms of wrought Al alloys. For a typical substitutional solid solution, the impact of solid solution strengthening on the Al alloy strength is relatively small compared with precipitation strengthening and work hardening. Precipitation strengthening is achieved through ageing, which generally reduces ductility due to the formation of precipitates and the introduction of thermal stress. The effect of work hardening on ductility depends on the process that produces various products with different shapes. For example, for the same alloy, extruded product may have higher elongation than the cry-rolled product. Hence, the VI values of the manufacturing processes for ENLOG are relatively higher than that of the alloying elements.

To further streamline the feature set, features with VI of less than 1 pct are considered as irrelevant factors in prediction and excluded from the original database. For YTS, features of V, Li, Ni, Be, B, Pb, Bi, and Ga elements are eliminated; features including Ti, Cr, V, Ni, Be, B, Bi, Pb, Ga, and treatment afterward are ignored for UTS; and for ELONG, features of B, Be, Ni, Pb, Bi, Ga, and treatment afterward are removed. Through the employment of feature engineering, the original database has been shrunk down from over one hundred features to 57, 53, and 56 features for YTS, UTS, and ELONG respectively. Since the number of features is different for these three properties, three databases were reconstructed separately for modelling by removing the corresponding irrelevant features, and the same method will also be used to predict the testing set later.

D. *Machine Learning Modelling Process*

The database with the processed and optimized features is used to train the ML model. In recent years, different ML models have been utilized to predict material properties in distinct scenarios.[47]. The well-known artificial neural network (ANN) is normally used for databases with a large sample size, while the support vector machine (SVM) provides improved prediction from much smaller databases.[30] Due to the high-dimensional feature set in the present database, the collected samples are not sufficient for the construction of an ANN model. SVM regressor, also called SVR, is considered more suitable to establish a correlation between the features (*i.e.* composition and manufacturing process) and the targeted properties (*i.e.* YTS, UTS and ELONG) of wrought Al alloys in this case.

The fundamental scheme behind the SVR model is to find the mapping function $f(x)$, expressed by Eq. [4], and this function is simultaneously subjected to the objective function, given by Eq. [5], and several constraints, given by Eq. [6] as follows:[48]

$$f(x) = w \cdot x + b \tag{4}$$

$$\min_{w \in X, C \in (0,\infty)} \frac{1}{2} w^2 + C \sum_{i=1}^{n} \left(\xi_i + \xi_i^* \right) \tag{5}$$

$$\text{subjectto} \begin{cases} y_i - w \cdot x_i - b \leq \varepsilon + \xi_i \\ w \cdot x_i + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{6}$$

where w denotes the parameter vector of the SVR model; x is the feature set of samples; b represents the constant number; C stands for the regularization parameter and it may be any real number between 0 and ∞; ξ_i and ξ_i^* are the slack variables, allowing for infeasible samples that exceed the precision of the model; x_i and y_i are the features and targeted properties of the sample i, respectively; and ε represents the precision of the model (*i.e.* the maximum toleration of predicted value can deviate from actual value). L2 norm of the vector w is used in the regularization term. This mapping function illustrates the basic idea for a linear model, while the kernel function can be applied for more complex nonlinear problems in the present work. The kernel function transfers the original feature space (x) to a higher dimensionality (x') used for mapping the nonlinear correlation.[49] Considering the vast number of the features in the present database, the kernel function is more suitable to automatically capture the synergies between different features in determining the targeted properties. Thus, it is a critical factor in the generalizability of the SVR model.

$$k\left(x - x'\right) = \exp(-\gamma \|x - x'\|^2) \tag{7}$$

In this work, a radial basis function (RBF), defined as Eq. [7], is selected as the kernel function in the SVR model, called the SVR-RBF model. In Eq. [7], x and x' represent pairs of samples in the training set. γ is the kernel coefficient, which governs the influence range of a single training sample in an inversely proportional pattern. The regularization parameter C, mentioned in Eq. [5], is another parameter incorporated in the cost function of the SVR model. A larger C value results in reduced strength of the model regularization. This signifies a higher accuracy in the training set, but a higher difficulty in generalizing the testing set—known as overfitting. Whereas, underestimation of C can expand the acceptable margin of the decision function, decreasing the training accuracy—known as the underfitting effect. Therefore, the generalization performance of the SVR-RBF model is also significantly affected by the hyperparameters C and γ.[50] Here, the grid search algorithm is performed to select the optimal combination of C and γ.

To prevent the trained SVR-RBF model from overfitting or underfitting, a Monte Carlo cross-validation method[50] is adopted with 4:1 split ratio (*i.e.* 80 pct samples are randomly selected as the training set to construct and train the model while the remaining 20 pct samples are retained as the validation set). The validation set provides an unbiased evaluation of a model fit to the training set when tuning hyperparameters. Since the sample size in the present database is relatively small, the training of the SVR-RBF model is sensitive to different partitions of training set and validation set. Hence, Monte Carlo cross-validation method is applied to 200 randomly partitioned databases to minimize the impact of imbalanced data distribution; correspondingly, 200 distinct SVR-RBF models are then established. The average performance of these 200 trained models in validation sets determines the optimal hyperparameters embedded. Herein, the root mean square error (RMSE) is considered as an evaluation function to provide prediction error with physical meaning. Meanwhile, the coefficient of determination (R^2), a scale-free metric, is also used to directly quantify the predictive accuracy of models. These two metrics are defined in Eqs. [8] and [9]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ia} - y_{ip})^2} \qquad [8]$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{ia} - y_{ip}\right)^2}{\sum_{i=1}^{n}(y_{ia} - \hat{y})^2} \qquad [9]$$

where n is the number of samples; y_{ia} and y_{ip} are the actual and the predicted output of sample i, respectively; and $\hat{y}$ is the mean value of the targeted properties in the validation set.

IV. RESULTS AND DISCUSSION

A. *Targeted Properties Prediction*

After the SVR-RBF models were trained, they are validated using the remaining 20 pct samples retained in the validation sets by predicting the targeted properties with the feature sets. As the cross-validation repeats 200 iterations, the same sample may be included in different validation sets for multiple times and also predicted by different SVR-RBF models. The average predicted properties of samples are plotted against the reported experimentally determined values in Figures 5(a) through (c) for YTS, UTS, and ELONG along with average RMSE and R^2 results, respectively. Different colours represent the number of samples near those corresponding points. The dashed straight lines in Figure 5 are the lines of identity (*i.e.* with a slope of 1) and are used to visualize the accuracy of models. The data points closer to the line of identity represent the higher consistency of the predictions with the experimental results, indicating more accurate inference. As shown in Figures 5(a) and (b), for the YTS and UTS, the data points are concentrated along the line of identity with the high regression coefficient R^2 of 91.17 (± 1.76) pct and 94.70 (± 0.99) pct, respectively. This indicates the high accuracy of the established correlations between the features and strength. In comparison, the data points of the ELONG, shown in Figure 5(c), are relatively dispersive, which is also reflected by the low corresponding R^2 value of 63.65 (± 5.83) pct. This signifies a relatively low accuracy of the established feature-to-ELONG correlation. Meanwhile, the frequency histogram of RMSE values are presented in the insets of Figures 5(a) through (c), from which clustered distributions can be shown. Scatter plots for widely-used wrought Al alloys (2024, 6061, 7075) in the present database are illustrated in Figures 5(d) through (f). These properties are evenly distributed within the shown range. The predictions on strength show high accuracy. The elongation results are more scattered. In particular, the 6061 alloy presents a larger deviation from the actual results than the other two alloys.

To further assess the model, additional data (not in the current database) collected from the most recently published literature[52–58] are assembled as the testing set to provide an unbiased evaluation of final models. These data have never been involved in model training; hence, they can well reflect the generalization ability of tuned models. The predicted mechanical properties of these alloys are plotted against the observed results in Figure 6. The strength values are primarily distributed around the high-end region, while the elongation values are intermediate. For the YTS and UTS prediction (Figures 6(a) and (b)), the scatter points are located near the lines of identity—indicating the generalizability of these models on wrought Al alloys with outstanding strength. This is also reflected by the RMSE values of 58.09 and 39.27 MPa for YTS and UTS respectively. Compared with the results of the validation set (Figures 5(a) through (c)), the RMSE values are slightly higher, possibly due to relatively high strengths of the testing set. A portion of ELONG predictions (Figure 6(c)) deviates from the line of identity—showing relatively poor prediction ability. The corresponding RMSE result is 3.31 pct EL, which is better than the result of the validation set.

B. *The Effect of Data Preprocessing*

As mentioned in section III–B, four encoding methods can be combined with two normalization approaches—-forming eight different DP schemes. Their corresponding errors in the validation sets are evaluated using the RMSE method as shown in Eq. [8] above. Table III lists the average RMSE results with uncertainties of these DP schemes for YTS, UTS, and ELONG respectively. It is obvious that the unordered encoding methods (*i.e.* one-hot encoding and sum encoding) outperformed the other two ordered methods as expected. Although the ordered encoding methods minimize the dimension of the feature set, they introduce a meaningful yet non-existent scales (*i.e.* unrealistic relation between categories) to the features[59]. Meanwhile, the Min-max normalization exhibits lower variance, but the Z-score method possesses considerably smaller errors. The DP scheme with the combination of Z-score normalization and one-hot encoding (used in previous modelling) yields the smallest RMSE value for all targeted properties, suggesting the best applicability for the present database.

C. *The Effect of ML Model*

As there is no clear indicator as to which machine learning model is the most suitable for our database, the performance of six renowned models were evaluated using Monte Carlo cross-validation, mentioned above. The models used are the SVR model with linear kernel (SVR-LIN), polynomial kernel (SVR-POLY), sigmoid kernel (SVR-SIG) and radial basis function (SVR-RBF), backpropagation neural network (BPNN) and random forest (RF). More details of the BPNN and RF are included in the electronic supplementary Section D. The average RMSE values for different mechanical properties as well as respective standard deviations are presented in Figure 7. It can be concluded that the SVR-RBF and RF outperform the other models across all three targeted properties in average error. However, the smaller standard deviation

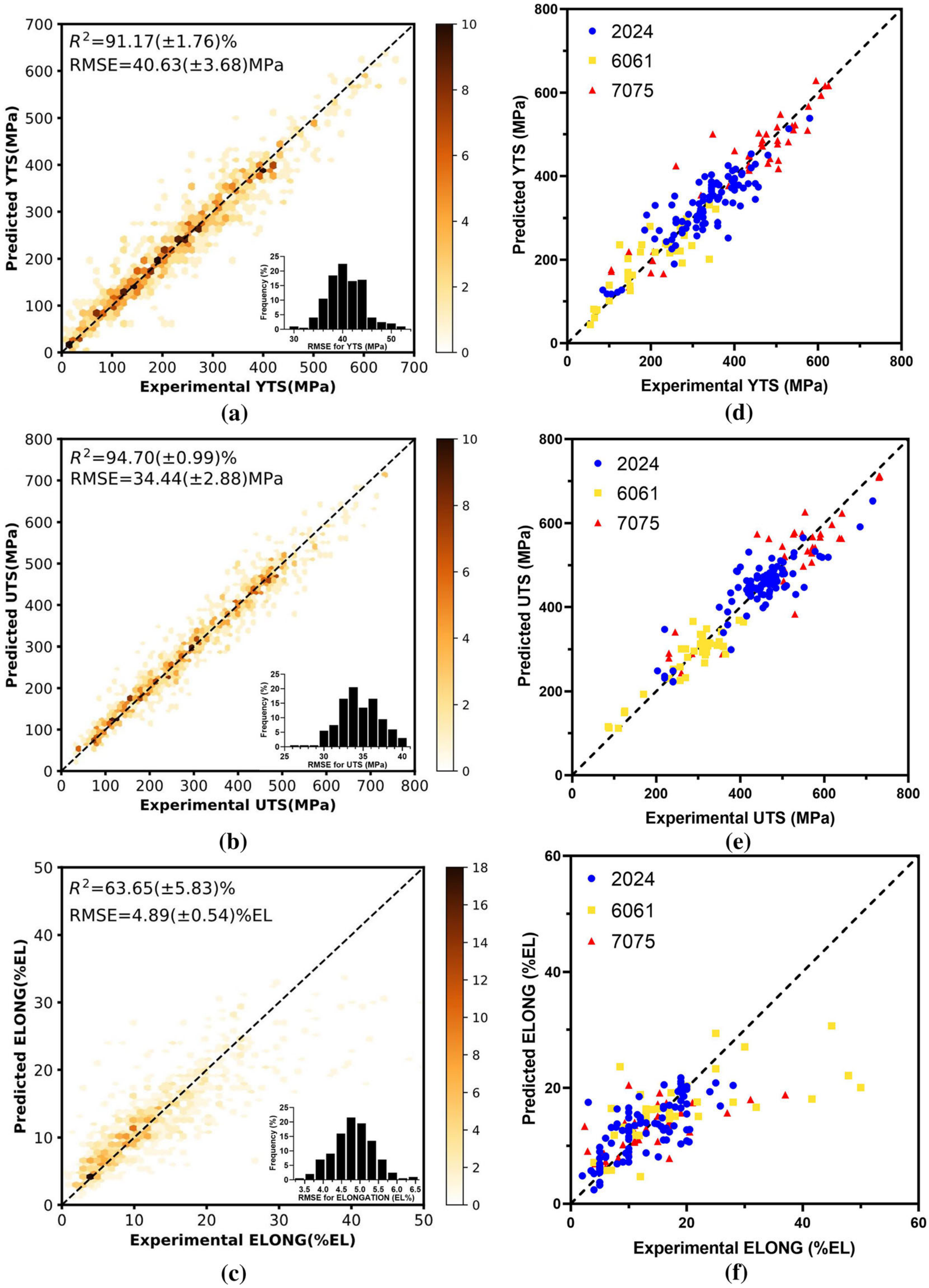


Fig. 5—Density plots of predicted results from tuned SVR-RBF models against the observed results in the validation sets: (*a*) YTS; (*b*) UTS; (*c*) ELONG. The insets in (a) (b) and (c) show the distribution of RMSE values in 200 validation sets. Scatter plots of predictions against the actual data for common Al alloys (*i.e*. 2024, 6061, 7075): (*d*) YTS; (*e*) UTS; (*f*) ELONG.

of the SVR-RBF shows its more stable prediction on validation sets (*i.e*. less affected by different partitions of database). Thus, the SVR-RBF model is chosen as the optimal surrogate model to establish the feature-property correlation. Moreover, the predictive accuracy of the SVR-RBF model with default kernel coefficient and regularization parameters are also quantified (as shown in Figure 6), indicating a critical influence of the C and γ parameters on the generalizability of the SVR-RBF model.

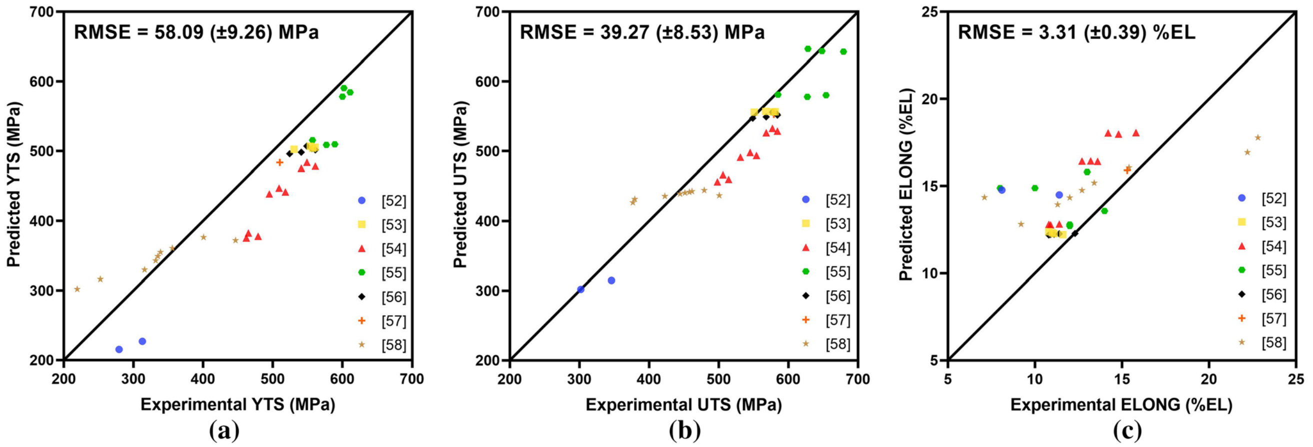


Fig. 6—Scatter plots of experimental and predicted properties for novel wrought Al alloys. Experimental data were collected from recent published papers[52–58]: (*a*) YTS; (*b*) UTS; (*c*) ELONG.

Table III. The RMSE Results for Different Properties With Database Preprocessed by Different DP Schemes

	YTS (MPa)		UTS (MPa)		ELONG (Percent)	
	Normalization					
Encoding	Z-Score	Min–Max	Z-Score	Min–Max	Z-Score	Min–Max
Ordinal Encoding	46.41 (± 4.18)	57.15 (± 3.34)	38.24 (± 3.58)	50.46 (± 3.32)	5.04 (± 0.53)	5.56 (± 0.47)
Binary Encoding	42.35 (± 3.60)	53.93 (± 3.25)	36.35 (± 2.84)	48.78 (± 2.66)	5.06 (± 0.54)	5.63 (± 0.43)
One-Hot Encoding	**40.63** (± 3.67)	53.21 (± 3.13)	**34.44** (± 2.88)	47.93 (± 2.56)	**4.89** (± 0.52)	5.53 (± 0.45)
Sum Encoding	41.76 (± 3.65)	53.69 (± 3.11)	35.64 (± 3.17)	48.29 (± 3.67)	4.98 (± 0.55)	5.58 (± 0.46)

Lowest achieved RMSE values are shown in bold.

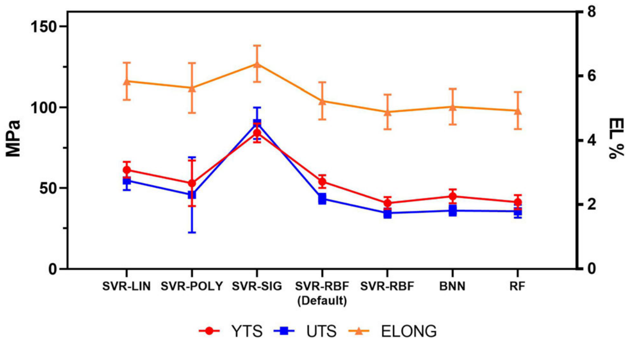


Fig. 7—The performance of six ML models on the targeted properties of wrought Al alloys in term of average RMSE value and the associated standard deviation.

D. *The Inadequate Accuracy in Predicting Elongation*

From Figure 5, it can be seen that the corresponding R^2 values of YTS and the UTS are higher than that of the ELONG. This phenomenon was also reported in Mohanty's work.[60] In the current database, the number of samples containing the ELONG, YTS and UTS property is 783, 860 and 896 respectively. The lower accuracy of the present SVR-RBF model in correlating the feature set with the elongation is probably attributed to the scarcity of samples in the training set. To confirm this assumption, the variation of R^2 values of ELONG with different training sample ratios is plotted in Figure 8. The accuracy of the SVR-RBF model on predicting the ELONG increases with the increasing size of the training set and has the tendency to be further improved with more samples. Recently, Guo's team used an industrial database with 63,137 samples to infer the mechanical properties of steels and Rˆ2 value of elongation reached 82.6 pct, which was still inferior to that of strength.[61] The insufficient accuracy may also be attributed to different sources of the database. Unlike the strength, the ELONG, *i.e.* the ductility of an alloy, is more sensitive to the testing conditions such as strain rate and geometry and size of the testing samples.[62] Although the tensile testing standard is available as per ASTM E8/E8M-13, the sample size and geometry, the required testing force and strain rate, may vary, causing data scatter. In addition, the measurement of elongation in Al alloy handbook,[36] from which the majority of samples are collected, was commonly done by measuring the gauge length of the tensile samples after putting the two broken pieces together. This method may cause higher experimental errors. Hence, the present ML model shows relatively low accuracy in predicting the ELONG of wrought Al alloys. Accordingly, an effective method to address this issue is to further optimize the database by using the samples with the same testing condition and sample geometry. However, this will

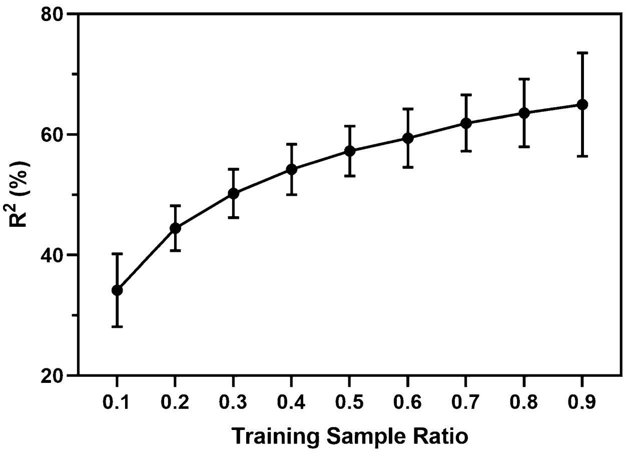


Fig. 8—Variation of R^2 value of SVR-RBF model in predicting ELONGATION property with different training sample sizes.

significantly shrink the number of samples due to the non-standardization in tensile testing throughout literature. Previous work by Guo *et al.* [61] and the present result shown in Figure 8 suggest that further expansion of the database appears to be a more practical and robust solution to improve the prediction accuracy.

V. CONCLUSIONS

In this study, the YTS, UTS, and ELONG of wrought Al alloy have been predicted by feature engineering assisted SVR-RBF model. By incorporating both numerical and categorical features as input, integral manufacturing processes are considered in the inference of a quantitative relationship. The key conclusions are listed as follows:

1. The manufacturing process has been effectively included in ML modelling. A procedure-oriented decomposition (POD) method was used, where a large number of categorical variables (temper designations) have been successfully described with reduced dimensionality. This has been achieved without sacrificing important information while improving computational speed.
2. The variable importance of each feature calculated by the random forest (RF) algorithm indicates the sensitivities of different mechanical properties (yield strength, tensile strength, and elongation) of wrought Al alloys to different alloying elements, heat treatment, and manufacturing process. The results agree very well with the physical metallurgy of wrought Al alloys and also eliminate irrelevant features.
3. SVR-RBF model is considered the most suitable machine learning model with the best generalization performance for this high-dimensional database. After tuning hyperparameters, the prediction accuracy on YTS, UTS, and ELONG are 91.17 (± 1.76) pct, 94.70 (± 0.99) pct, and 63.65 (± 5.83) pct in validation set, respectively. The model is capable of predicting the properties of recently reported wrought Al alloys (regarded as testing set) with potent consistency between the predictions and the experimental results.
4. This established inference map from composition and manufacturing process to the strength and ductility in the field of wrought Al alloys could accelerate the design of new materials with targeted tensile properties. Furthermore, this prediction framework can also be applied to other targeted properties or systems.

ACKNOWLEDGMENTS

The authors are very grateful to ARC Discovery Project for funding support (No. DP180102454).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

SUPPLEMENTARY INFORMATION

The online version contains supplementary material available at https://doi.org/10.1007/s11661-021-06279-5.

REFERENCES

1. Y. Li, S. Brusethaug, and A. Olsen: *Scripta Mater*., 2006, vol. 54, pp. 99–103.
2. W. Miller, L. Zhuang, J. Bottema, A. Wittebrood, P. De Smet, A. Haszler, and A. Vieregge: *Mater. Sci. Eng., A*, 2000, vol. 280, pp. 37–49.
3. T. Dursun and C. Soutis: *Mater. Des*., 2014, vol. 56, pp. 862–71.
4. J. Wang, A.Y. Nobakht, J.D. Blanks, D. Shin, S. Lee, A. Shyam, H. Rezayat, and S. Shin: *Adv. Theory Simul*., 2019, vol. 2, p. 1800196.
5. TA. Ivanoff, J.T. Carter, L.G. Hector, and E.M. Taleff: *Metall. and Mater. Trans. A*., 2019, vol. 50 (3), pp. 1545–61.
6. N. Reddy, J. Krishnaiah, H.B. Young, and J.S. Lee: *Comput. Mater. Sci*., 2015, vol. 101, pp. 120–26.
7. EO. Hall: Proceedings of the Physical Society*Section B*, 1951, vol. 64, pp. 747–53.
8. N. Petch: *Journal of the Iron and Steel Institute*, 1953, vol. 174, pp. 25–28.
9. R.L. Fleischer: *Acta Metall*., 1962, vol. 10, pp. 835–42.
10. J.E. Bailey and P.B. Hirsch: *Phil. Mag*., 1960, vol. 5, pp. 485–97.
11. E.A. Bloch: *Metallurgical Reviews*, 1961, vol. 6, pp. 193–240.
12. B. Cao, S.P. Joshi, and K. Ramesh: *Scripta Mater*., 2009, vol. 60, pp. 619–22.
13. K. Ma, H. Wen, T. Hu, T.D. Topping, D. Isheim, D.N. Seidman, E.J. Lavernia, and J.M. Schoenung: *Acta Mater*., 2014, vol. 62, pp. 141–55.
14. S. Curtarolo, G.L. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, and O. Levy: *Nat. Mater*., 2013, vol. 12 (3), pp. 191–201.
15. Y.W. Wang, J. Li, W. Liu, and Z.-K. Liu: *Comput. Mater. Sci*., 2019, vol. 158, pp. 42–48.
16. J. Smith, W. Xiong, J. Cao, and W.K. Liu: *Comput. Mech*., 2016, vol. 57, pp. 359–70.

17. Q. Du, W.J. Poole, M.A. Wells, and N. Parson: *JOM*, 2011, vol. 63, pp. 35–39.
18. T. Kitashima: *Phil. Mag.*, 2008, vol. 88, pp. 1615–37.
19. S.R. Kalidindi, A.J. Medford, and D.L. McDowell: *JOM*, 2016, vol. 68, pp. 2126–37.
20. J. Ling, E. Antono, S. Bajaj, S. Paradiso, M. Hutchinson, B. Meredig and B. M. Gibbons: *ASME Turbo Expo 2018: Turbomachinery Technical Conference and Exposition*, Oslo.
21. K. Rajan: *Mater. Today*, 2005, vol. 8, pp. 38–45.
22. Q. Du, W. Poole, and M. Wells: *Acta Mater.*, 2012, vol. 60, pp. 3830–39.
23. C. Gu, Y. Lu, E. Cinkilic, J. Miao, A. Klarner, X. Yan, and A.A. Luo: *Comput. Mater. Sci.*, 2019, vol. 161, pp. 64–75.
24. F. Hannard, T. Pardoen, E. Maire, C. Le Bourlot, R. Mokso, and A. Simar: *Acta Mater.*, 2016, vol. 103, pp. 558–72.
25. D. Xue, P.V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman: *Nat. Commun.*, 2016, vol. 7, p. 11241.
26. Y. Liu, T. Zhao, W. Ju, and S. Shi: *J. Mater.*, 2017, vol. 3, pp. 159–77.
27. C. Wen, Y. Zhang, C. Wang, D. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman, and Y. Su: *Acta Mater.*, 2019, vol. 170, pp. 109–17.
28. M.S. Ozerdem and S. Kolukisa: *Mater. Des.*, 2009, vol. 30, pp. 764–69.
29. KP. Murphy: *Machine Learning: A Probabilistic Perspective*, 1st ed., MIT Press, Cambridge, 2012.
30. C. Shen, C. Wang, X. Wei, Y. Li, S. van der Zwaag, and W. Xu: *Acta Mater.*, 2019, vol. 179, pp. 201–14.
31. A. Patra, S. Ganguly, M. Kaiser, P. Chattopadhyay, and S. Datta: *Int. J. Mechatron. Manufact. Syst.*, 2010, vol. 3, pp. 144–54.
32. T. Varol, A. Canakci, and S. Ozsahin: *J. Alloy. Compd.*, 2018, vol. 739, pp. 1005–14.
33. A. Belayadi and B. Bourahla: *Phys. B*, 2019, vol. 554, pp. 114–20.
34. S. Dey, N. Sultana, M.S. Kaiser, P. Dey, and S. Datta: *Mater. Des.*, 2016, vol. 92, pp. 522–34.
35. B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, and T.Y.J. Han: *NPJ Computat. Mater.*, 2019, vol. 5, p. 108.
36. J. Davis: *Aluminum and aluminum alloys*, ASM International, Materials Park, OH, 1993.
37. R.G. Guan and D. Tie: *Acta Metall. Sin.*, 2017, vol. 30 (5), pp. 409–32.
38. J. Sola and J. Sevilla: *IEEE Trans. Nucl. Sci.*, 1997, vol. 44, pp. 1464–68.
39. S.B. Kotsiantis, D. Kanellopoulos, and P.E. Pintelas: *Int. J. Comput. Sci.*, 2006, vol. 1, pp. 111–17.
40. S. Klement, A.M. Mamlouk, and T. Martinetz: *ICANN*, 2008, vol. 2008 (5163), pp. 41–50.
41. S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, and J. Wang: *Nat. Commun.*, 2018, vol. 9, p. 3405.
42. L. Breiman: *Mach. Learn.*, 2001, vol. 45, pp. 5–32.
43. R. Genuer, J.-M. Poggi, and C. Tuleau-Malot: *Pattern Recogn. Lett.*, 2010, vol. 31, pp. 2225–36.
44. N. Chinh, J. Lendvai, D. Ping, and K. Hono: *J. Alloy. Compd.*, 2004, vol. 378, pp. 52–60.
45. Z. Jin and P. Mallick: *J. Mater. Eng. Perform.*, 2006, vol. 15, pp. 540–48.
46. D. Ortiz, M. Abdelshehid, R. Dalton, J. Soltero, R. Clark, M. Hahn, E. Lee, W. Lightell, B. Pregger, J. Ogren, P. Stoyanov, and O. Es-Said: *J. Mater. Eng. Perform.*, 2007, vol. 16, pp. 515–20.
47. R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim: *NPJ Comput. Mater.*, 2017, vol. 3 (1), pp. 1–13.
48. AJ. Smola and B. Schölkopf: *Stat. Comput.*, 2005, vol. 14, pp. 199–222.
49. C.J. Burges: *Data Min. Knowl. Disc.*, 1998, vol. 2, pp. 121–67.
50. B. Üstün, W. Melssen, M. Oudenhuijzen, and L. Buydens: *Anal. Chim. Acta*, 2005, vol. 544, pp. 292–305.
51. RR. Picard and R.D. Cook: *J. Am. Stat. Assoc.*, 1984, vol. 79 (387), pp. 575–83.
52. W. Tu, J. Tang, Y. Zhang, L. Ye, S. Liu, J. Lu, X. Zhan, and C. Li: *Mater. Sci. Eng. A*, 2020, vol. 770, p. 138515.
53. B. Li, Q. Pan, C. Chen, H. Wu, and Z. Yin: *J. Alloy. Compd.*, 2016, vol. 664, pp. 553–64.
54. B. Li, Q. Pan, X. Huang, and Z. Yin: *Mater. Sci. Eng. A*, 2014, vol. 616, pp. 219–28.
55. G. Teng, C. Liu, Z. Ma, W. Zhou, L. Wei, Y. Chen, J. Li, and Y. Mo: *Mater. Sci. Eng. A*, 2018, vol. 713, pp. 61–66.
56. B. Li, Q. Pan, C. Chen, and Z. Yin: *Trans. Nonferrous Metals Soc. China*, 2016, vol. 26, pp. 2263–75.
57. X. Peng, Y. Li, G. Xu, J. Huang, and Z. Yin: *Meter. Mater. Int.*, 2018, vol. 24, pp. 1046–57.
58. Z. Tang, F. Jiang, M. Long, J. Jiang, H. Liu, and M. Tong: *Appl. Surf. Sci.*, 2020, vol. 514, p. 146081.
59. A. Von Eye and C.C. Clogg: *Categorical Variables in Developmental Research*, Elsevier, Burlington, 1996.
60. I. Mohanty, D. Bhattacharjee, and S. Datta: *Comput. Mater. Sci.*, 2011, vol. 50, pp. 2331–37.
61. S. Guo, J. Yu, X. Liu, C. Wang, and Q. Jiang: *Comput. Mater. Sci.*, 2019, vol. 160, pp. 95–104.
62. R. Smerd, S. Winkler, C. Salisbury, M. Worswick, D. Lloyd, and M. Finn: *Int. J. Impact Eng.*, 2005, vol. 32, pp. 541–60.