Machine Learning of Mechanical Properties of Steels

Jie XIONG[b,c], Tong-Yi ZHANG[a,*], San-Qiang SHI[b,c,*]

[a]Materials Genome Institute, Shanghai University, Shanghai, China

[b]Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong, China

[c]Shenzhen Research Institute, The Hong Kong Polytechnic University, Shenzhen, China

* corresponding author: T.-Y. ZHANG (zhangty@shu.edu.cn); S.-Q. SHI(san.qiang.shi@polyu.edu.hk).

Abstract: Knowledge of the mechanical properties of structural materials is essential for these practical applications. Three-hundred and sixty data samples on four mechanical properties of steels – fatigue strength, tensile strength, fracture strength and hardness – were selected for analysis from the Japan National Institute of Material Science database, comprising data on carbon steels and low-alloy steels. Five machine learning algorithms were used to predict the mechanical properties of the materials represented by the three-hundred and sixty data samples, and random forest regression showed the best predictive performance. Feature selection conducted by random forest and symbolic regressions revealed the four most important features that most influence the mechanical properties of steels: the tempering temperature of steel, and the alloying of steel with carbon, chromium or molybdenum. Mathematical expressions were generated via symbolic regression, and the expressions explicitly predicted how each of the four mechanical properties varied quantitatively with the four most important features. This study demonstrates the great potential of symbolic regression in the discovery of novel advanced materials.

Keywords: Materials Informatics; Steel; Fatigue Strength; Symbolic Regression

1. Introduction

The identification of structure-property relationships is fundamental to the discovery of new materials. However, the ability to comprehensively understand and manipulate structure-property relationships of materials is very challenging, due to the diversity and complexity of materials. As a result, data-driven discovery of novel advanced materials requires the use of advanced techniques such as big data and artificial intelligence, data mining and machine learning (ML) to accelerate research and development [1–8]. Materials data and ML provide the foundation of this data-driven materials discovery paradigm, which integrates materials domain knowledge and artificial intelligence technology to form the new research field of materials informatics.

In this new field, the Materials Genome Initiative aims to halve the cost and time from discovery to development to deployment of advanced materials [9]. This integrated approach applies materials data to explore structure-property relationships and to develop models and guidance for synthesis of new materials. For example, Homer et al. [10] and Zhu et al. [11] used ML tools to investigate grain boundaries in polycrystalline materials, and Raccuglia et al. [12] demonstrated a ML strategy to elucidate how to classify successful and failed synthesis conditions with the use of historically accumulated experimental data. Agrawal et al. [13,14] used ML algorithms to predict the fatigue strength of steels, which substantially improved the understanding of fatigue behavior. However, their ML predictions did not result in explicit mathematic expressions linking features and output properties, which are desirable for materials research, design, development and deployment.

The purpose of this study was to predict the four mechanical properties of steels using five ML algorithms, especially using random forest (RF) regression and symbolic regression (SR). The performances of the five algorithms were assessed, revealing that RF performing the best, and explicit mathematical expressions were obtained from SR.

2. Data Resource

The publicly available dataset for steels in the Japan National Institute of Material Science (NIMS) [15] was used in this study, as it is among the world’s largest experimental datasets of its type. The NIMS dataset contains materials chemical compositions, processing conditions and property information, including the mechanical properties of steels at room temperature, such as fatigue strength, tensile strength, fracture strength and hardness. Fatigue strength is defined as the critical value of an applied stress range, at or below which no fatigue failure will occur during a given material’s lifetime. In this study, the rotating bending fatigue strength of materials (hereafter ‘fatigue strength’) was measured at a fatigue life of 1×10^7 cycles.

Fatigue testing conditions, such as the loading frequency and profile, the testing temperature and environment, and the specimen dimensions, have significant effects on fatigue behavior. The 393 original data samples collected from the NIMS database had all been fatigue-tested under the same conditions, and thus the testing conditions are not considered in this study. The 393 original fatigue samples comprised 113 carbon steels, 258 low-alloy steels and 22 stainless steels, characterised by chemical composition, processing parameters, inclusion parameters and mechanical properties. In terms of chemical composition, the materials were composed of various proportions of nine alloying elements: carbon (C), silicon (Si), manganese (Mn), phosphorus (P), sulphur (S), nickel (Ni), chromium (Cr), copper (Cu), and molybdenum (Mo). The included parameters were the area fraction of non-metallic inclusions, namely dA (inclusions formed during plastic work), dB (inclusions that occur in discontinuous arrays) and dC (isolated inclusions). The processing parameters were the reduction ratio from the ingot to the bar, and the heat treatment parameters, as described in detail below.

(1) The heating rate and cooling rate are not considered, because no such data were available.

(2) Three types of heat treatments – normalizing, quenching and tempering – were conducted on the steels. The temperatures for normalizing, quenching and tempering were included, whilst the holding times at heat treatment temperatures were not, as data for only two holding times were available.

(3) After the heat treatment, the samples were cooled to room temperature to conduct the fatigue tests.

(4) Eleven carbon steels without normalizing treatment (SC25 steels) and 22 stainless steels without quenching and tempering treatment were excluded from the study, which reduced the original 393 data samples to 360.

The 360 data samples comprised 16 variables of nine alloying elements, one reduction ratio, three heat-treatment temperatures, three inclusions and four target properties (fatigue strength, tensile strength, fracture strength and hardness). These 16 variables were the named features in ML, and the minimum and maximum values of each feature are shown in Table 1.

Table 1. 16 Features of the 360 NIMS fatigue data

Features	Description	Min	Max	Mean	StdDev
NT	Normalizing Temperature (°C)	825	900	865.6	17.37
QT	Quenching Temperature (°C)	825	865	848.2	9.86
TT	Tempering Temperature (°C)	550	680	605	42.4
C (x_1)	wt% of Carbon	0.28	0.57	0.407	0.061
Si (x_2)	wt% of Silicon	0.16	0.35	0.258	0.034
Mn (x_3)	wt% of Manganese	0.37	1.3	0.849	0.294
P (x_4)	wt% of Phosphorus	0.007	0.031	0.016	0.005
S (x_5)	wt% of Sulphur	0.003	0.03	0.014	0.006
Ni (x_6)	wt% of Nickel	0.01	2.78	0.548	0.899
Cr (x_7)	wt% of Chromium	0.01	1.12	0.556	0.419
Cu (x_8)	wt% of Copper	0.01	0.22	0.064	0.045
Mo (x_9)	wt% of Molybdenum	0	0.24	0.066	0.089
RR	Reduction ratio	420	5530	971.2	601.4
dA	Plastic work-inclusions	0	0.13	0.047	0.032
dB	discontinuous array-inclusions	0	0.05	0.003	0.009
dC	isolated inclusions	0	0.04	0.008	0.01

*Weight percentage of iron is $x_{10} = 100 - \sum_{i=1}^{9} x_i$

3. Results and Discussion

3.1 ML models with all features

Four ML algorithms – RF, linear least-square (LLS), *k*-nearest neighbors (KNN) and architecture-neural network (ANN) – were conducted on the dataset comprising all 16 features (termed ‘All’). The performances of these algorithms were evaluated by ten-fold cross-validation, in which the data were divided into ten parts (nine parts for training data and one part for testing data) and the training and testing were cycled ten times to allow the use of all data in testing. The predictive power of an ML algorithm on the testing data was measured by the correlation coefficient (*R*) and the relative root-mean-square errors (RRMSE), which are defined by

$$R = \frac{\left|\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(\hat{y}_i - \bar{\hat{y}}\right)\right|}{\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2}}, \tag{1}$$

$$RRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}}{\bar{y}} \tag{2}$$

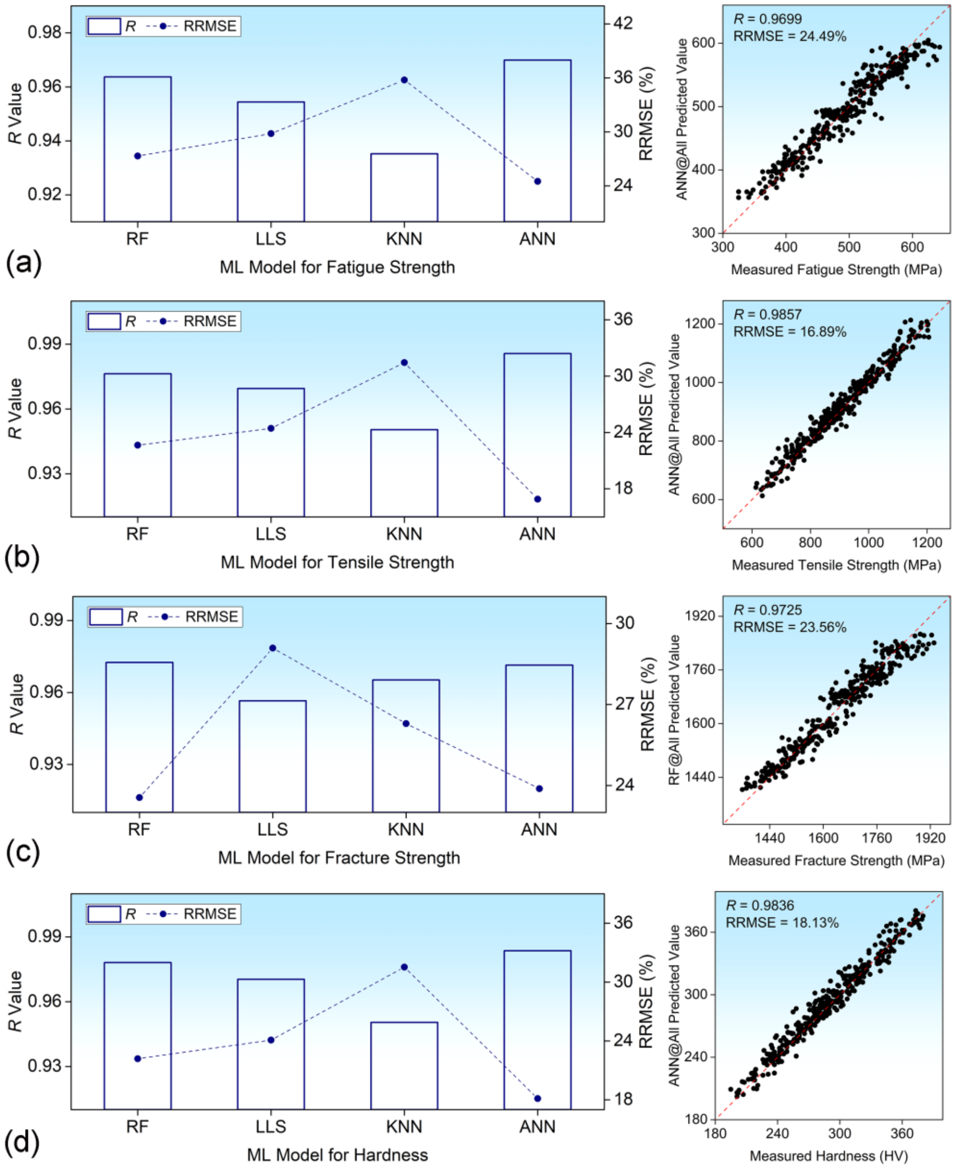

Figure 1. The *R* and RRMSE values of the RF (random forest), LLS (linear least-square), KNN (*k*-nearest neighbors) and ANN (architecture-neural network) models using all 16 features: (a) fatigue strength of all models, and the performance of the best model, ANN@All; (b) tensile strength and the performance of the best model, ANN@All; (c) fracture strength and the performance of the best model, RF@All; and (d) hardness and the performance of the best model ANN@All.

where n is the number of testing data, and y, $\hat{y}$ and $\bar{y}$ denote the actual value, the predicted value and the average value, respectively. R lies between 0 and 1, and a value of 1 indicates a perfect prediction. An RRMSE value of zero indicates a perfect fit. In general, a higher value for R and a lower value for RRMSE indicate a better ML algorithm [16].

Figure 1 shows the R and RRMSE values of the four ML algorithms and compares the best predicted values from one of the ML algorithms with the measured values for each of the four mechanical properties. As can be seen, the RF has the greatest predictive power for fracture strength (R = 0.9725, RRMSE = 23.56%), whilst the ANN algorithm gives the best results for fatigue strength (R = 0.9699, RRMSE = 24.49%), tensile strength (R = 0.9857, RRMSE = 16.89%) and hardness (R = 0.9836, RRMSE = 18.13%).

3.2 Feature selection

Feature selection is crucial in ML; given the fact that ML algorithms such as RF and SR have feature selection functions, these algorithms are emphasized here. The importance of the features computed by RF is denoted RFI, and that of the features computed by SR is called SRI. Figures 2 (a-b) show the RFI and SRI values, respectively, for each original feature. The RFI values indicate that the four most important features are the presence of Mo and Cr, the normalizing temperature and the tempering temperature, whilst the SRI values indicate that the four most important features are the tempering temperature, and the presence of C, Cr and Mo, which correspondingly yield two feature subsets of RFI (NT, TT, Cr, Mo) and SRI (TT, C, Cr, Mo).

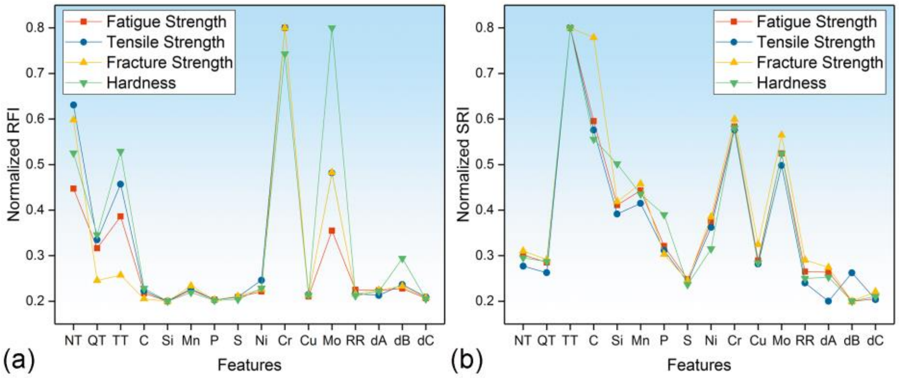

Figure 2. Normalized (a) random forest importance (RFI) and (b) symbolic regression importance (SRI) of the 16 features for fatigue strength, tensile strength, fracture strength, and hardness.

The four ML algorithms were conducted using the RFI features (NT, TT, Cr, Mo) and the SRI features (TT, C, Cr, Mo). Figure 3 shows the cross-validation *R* values and the predicted values of the best model against the measured value for each of the four features. The results illustrate that the RF algorithm with the feature subset SRI (TT, C, Cr, Mo) outperforms the other algorithms. The RF models with the feature subset SRI (TT, C, Cr, Mo) predict the four target properties with high predictive accuracy ($R > 0.9550$, RRMSE < 30.00%).

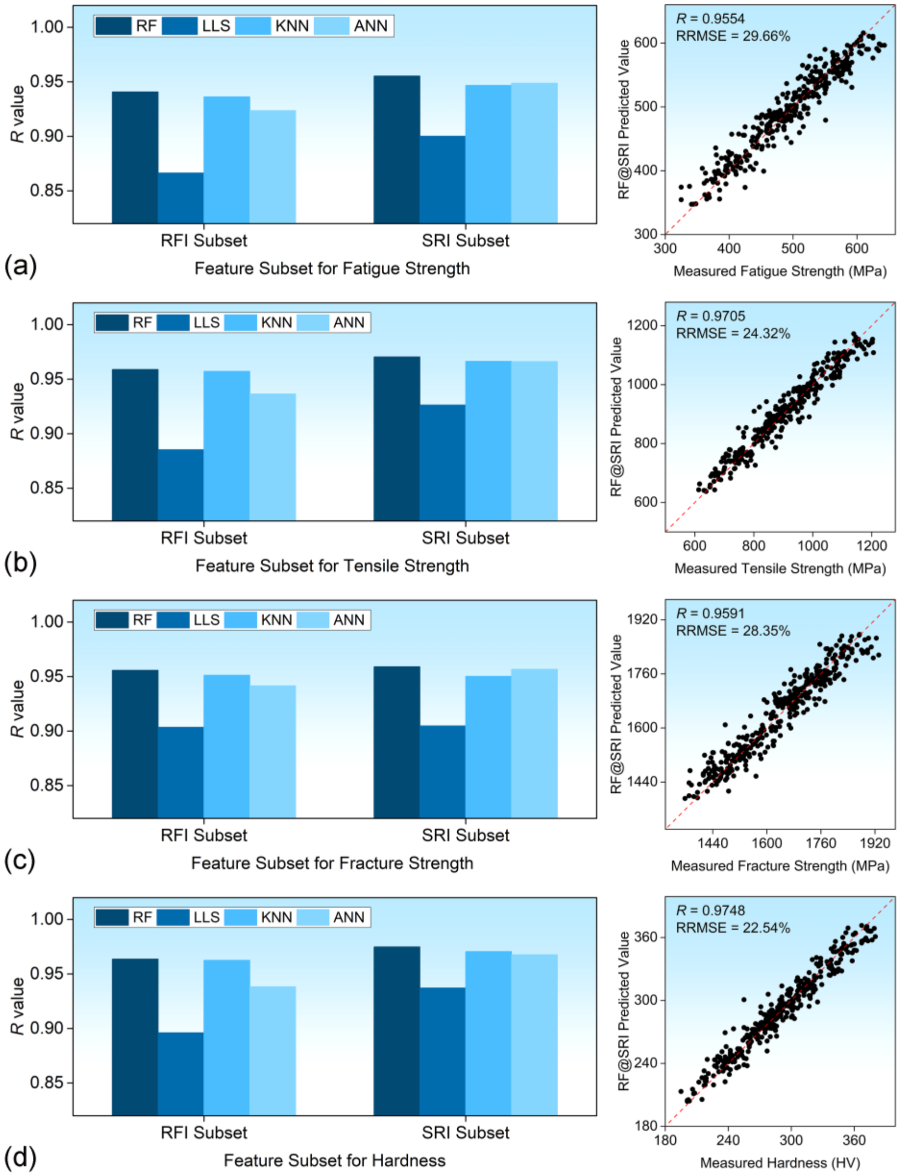

Figure 3. *R* values of the RF (random forest), LLS (linear least-square), KNN (*k*-nearest neighbors) and ANN (architecture-neural network) models with the selected RFI and SRI feature subsets: (a)

for fatigue strength, and the performance of the best model, RF@SRI; (b) for tensile strength, and the performance of the best model, RF@SRI; (c) for fracture strength and the performance of the best model, RF@SRI; and (d) for hardness, and the performance of the best model RF@SRI.

3.3 Mathematical expressions

With SRI features (TT, C, Cr, Mo), SR gave the following mathematical expressions for fatigue strength (FaS) (MPa), tensile strength (TS) (MPa), fracture strength (FrS) (MPa), and hardness (H) (HV).

$$\mathrm{FaS} = -0.8685\mathrm{TT} + 316.7\mathrm{C} + 367.6\mathrm{Cr} - 227.5\mathrm{Cr}^2 + 708.6\mathrm{Mo}^2 + 785.0 \tag{3}$$

$$\mathrm{TS} = -1.827\mathrm{TT} - 119.7/\mathrm{C} + 643.2\mathrm{Cr} - 379.9\mathrm{Cr}^2 + 1514\mathrm{Mo}^2 + 2122 \tag{4}$$

$$\mathrm{FrS} = -1.176\mathrm{TT} - 46.12/\mathrm{C} + 695.4\mathrm{Cr} - 415.3\mathrm{Cr}^2 + 1461\mathrm{Mo}^2 + 2267 \tag{5}$$

$$\mathrm{H} = -0.5839\mathrm{TT} - 38.41/\mathrm{C} + 191.2\mathrm{Cr} - 113.3\mathrm{Cr}^2 + 104.0\mathrm{Mo} + 681.9 \tag{6}$$

where all elements are expressed in wt.% and TT is expressed in (°C). The equations had strong predictive power ($R > 0.9425$, RRMSE < 33.30%), as shown in Figure 4. Equations (3-6) each include a minus sign with the tempering temperature, which indicates that lower tempering temperatures are should improve the strength and hardness of steels. The alloying elements of C, Cr and Mo are also strengthening elements.

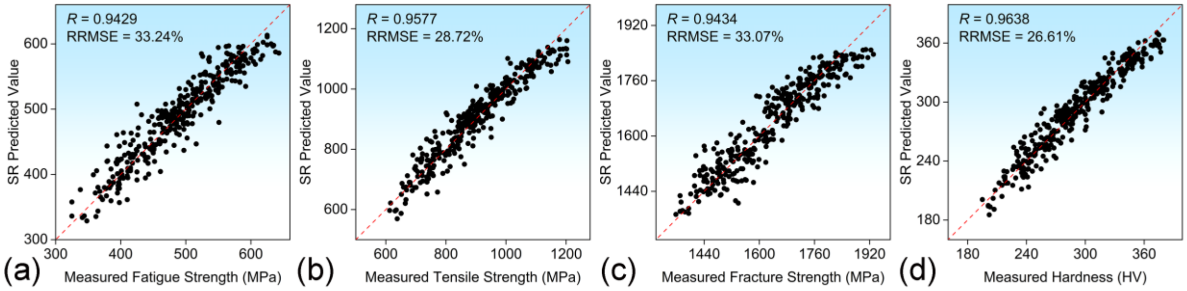

Figure 4. Performance illustrations of (a) Eq. (3) for fatigue strength, (b) Eq. (4) for tensile strength, (c) Eq. (5) for fracture strength and (d) Eq. (6) for hardness.

3.4 ML model based on atomic features

Atomic features were used in this study to generalize the predictive power of ML in new alloy discovery. Specifically, iron (Fe) is the matrix of steels, and alloying elements in steels may behave as solutes within the iron matrix 'solvent', forming metal carbides with carbon or intermetallic compounds with Fe and/or among the alloying elements themselves, which then may precipitate as clusters or/and tiny phases. Table 2 lists the atomic features, which, together with tempering temperatures, are denoted by the term 'All-AF' and are used in the following ML.

In Table 2, r_i and r_{Fe} denote the atomic radii of element i and Fe, respectively; VEC_i, $\mathrm{VEC}_{\mathrm{Fe}}$ and $\mathrm{VEC}_{\mathrm{C}}$ are the valance electrons of element i, Fe and C, respectively; and χ_i, χ_{Fe} and χ_{C} are the Pauling electronegativities of element i, Fe and C, respectively. Table S1 in the Supplementary Material lists the values of these atomic properties. In addition, a_i is the atomic percentage of element i, which is linked to the weight percentage x_i by the expression $a_i = \frac{x_i / M_i}{\sum_i (x_i / M_i)}$, where M_i is the atomic weight of element i.

Table 2. Atomic features used in this work

Features	Description	Formula
a_{Fe}	Atomic percentage of Fe	a_{10}
t_r	Total atomic radius	$\sum_{i=1}^{10} a_i r_i$
$d_{r\text{-Fe}}$	Atomic radius difference (Fe-based)	$\sqrt{\sum_{i=1}^{10} a_i \left(1 - {r_i}/{r_{\mathrm{Fe}}}\right)^2}$
t_{VEC}	Total valance electron	$\sum_{i=1}^{10} a_i \mathrm{VEC}_i$
$d_{\mathrm{VEC\text{-}Fe}}$	Valance electron difference (Fe-based)	$\sqrt{\sum_{i=1}^{10} a_i \left(1 - {\mathrm{VEC}_i}/{\mathrm{VEC}_{\mathrm{Fe}}}\right)^2}$
$d_{\mathrm{VEC\text{-}C}}$	Valance electron difference (C-based)	$\sqrt{\sum_{i=1}^{10} a_i \left(1 - {\mathrm{VEC}_i}/{\mathrm{VEC}_{\mathrm{C}}}\right)^2}$
t_χ	Total Pauling electronegativity	$\sum_{i=1}^{10} a_i \chi_i$
$d_{\chi\text{-Fe}}$	Electronegativity difference (Fe-based)	$\sqrt{\sum_{i=1}^{10} a_i \left(1 - {\chi_i}/{\chi_{\mathrm{Fe}}}\right)^2}$

$d_{\chi\text{-C}}$	Electronegativity difference (C-based)	$\sqrt{\sum_{i=1}^{10} a_i\left(1-{\chi_i}/{\chi_{\mathrm{C}}}\right)^2}$

The selection of atomic features was conducted by RF and SR. Thus, RFI selected four important features (t_{VEC}, $d_{\mathrm{VEC\text{-}Fe}}$, $d_{\mathrm{VEC\text{-}C}}$ and TT for fatigue strength and hardness and t_{VEC}, a_{Fe}, $d_{\mathrm{VEC\text{-}C}}$ and TT for tensile strength and fracture strength), and SRI selected four important features ($d_{\mathrm{VEC\text{-}C}}$, $d_{r\text{-}\mathrm{Fe}}$, a_{Fe} and TT for all four mechanical properties). The features selected by RF and SR are referred to as RFI-AF (TT, t_{VEC}, $d_{\mathrm{VEC\text{-}Fe}}$, $d_{\mathrm{VEC\text{-}C}}$), RFI-AF (TT, t_{VEC}, a_{Fe}, $d_{\mathrm{VEC\text{-}C}}$) and SRI-AF (TT, a_{Fe}, $d_{r\text{-}\mathrm{Fe}}$, $d_{\mathrm{VEC\text{-}C}}$), respectively.

The RF algorithm was conducted again with the All-AF, RFI-AF and SRI-AF features. Figure 5(a) shows the R values for each of the four properties. The results indicate that the RF model with SRI-AF (TT, a_{Fe}, $d_{r\text{-}\mathrm{Fe}}$, $d_{\mathrm{VEC\text{-}C}}$) has similar performance to that of the RF model with All-AF, and the RF model with SRI-AF performs better than the RF model with the two RFI-AF feature sets. Figures 5(b-e) show the values predicted by the RF model with SRI-AF against the measured values for the four mechanical properties, respectively (all $R > 0.9510$; all RRMSE < 31.00%).

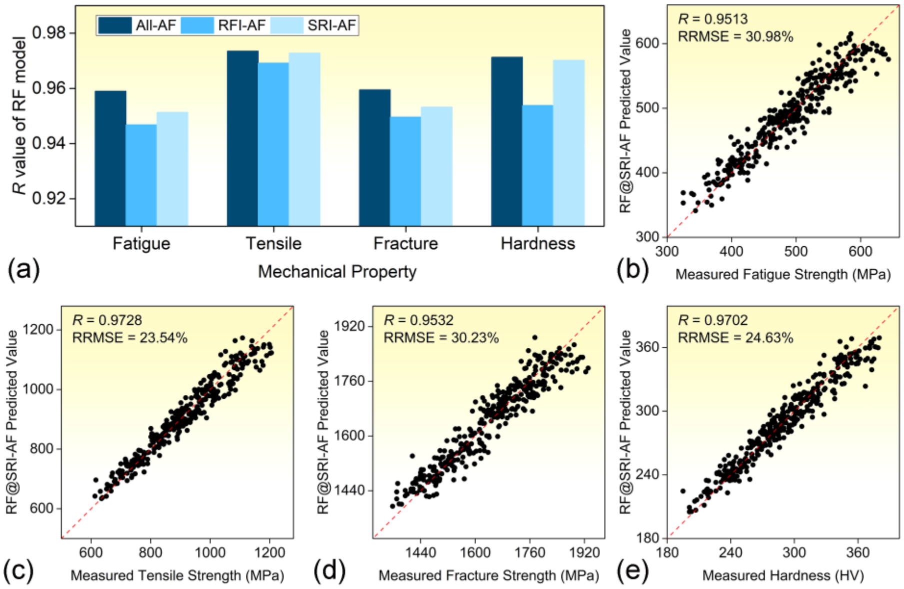

Figure 5. (a) R values of RF (random forest) models with All-AF, RFI-AF and SRI-AF features for the four mechanical properties. Predicted values of RF model with SRI-AF features against the measured values for (b) fatigue strength, (c) tensile strength, (d) fracture strength and (e) hardness.

Similarity, Equations (7-10) from SR gave the explicit correlations of FaS in MPa, TS in MPa, FrS in MPa, and hardness (H) in HV, with the SRI-AF, respectively.

$$\mathrm{FaS} = -0.8631\mathrm{TT} - 2771 a_{\mathrm{Fe}} + 6679 d_{r\text{-}\mathrm{Fe}} + 27690 d_{\mathrm{VEC\text{-}C}} - 10610 \tag{7}$$

$$\mathrm{TS} = -1.801\mathrm{TT} - 4438 a_{\mathrm{Fe}} + 14852 d_{r\text{-}\mathrm{Fe}} + 58552 d_{\mathrm{VEC\text{-}C}} - 24019 \tag{8}$$

$$\mathrm{FrS} = -1.148\mathrm{TT} - 4718 a_{\mathrm{Fe}} + 9863 d_{r\text{-}\mathrm{Fe}} + 60564 d_{\mathrm{VEC\text{-}C}} - 24003 \tag{9}$$

$$\mathrm{H} = -0.5724\mathrm{TT} - 1122 a_{\mathrm{Fe}} + 4810 d_{r\text{-}\mathrm{Fe}} + 18906 d_{\mathrm{VEC\text{-}C}} - 8062 \tag{10}$$

where a_{Fe} is expressed in at.% (atomic percentage) and TT is expressed in degrees Celsius (°C). Those equations indicate that the alloying elements enhance the strength of steels. Figures 6 (a-d) show the predictive performances of Equations (7-10), respectively, and the associated *R* and RRMSE values.

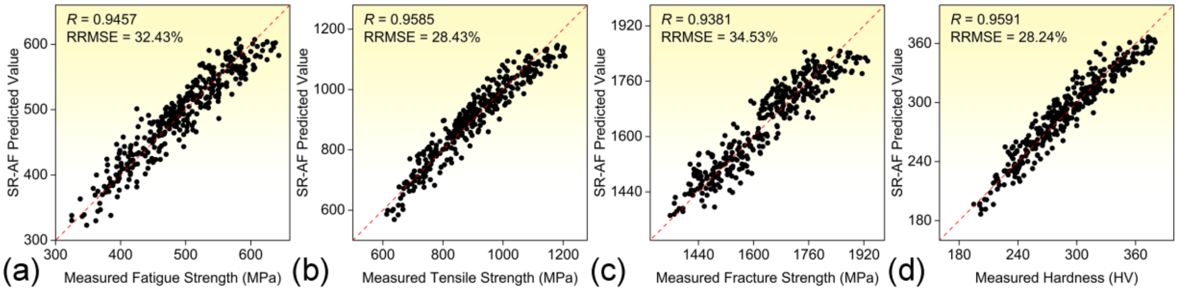

Figure 6. Performance illustrations of (a) Eq. (7) for fatigue strength, (b) Eq. (8) for tensile strength, (c) Eq. (9) for fracture strength and (d) Eq. (10) for hardness.

3.5 Development of anti-fatigue high strength steel

In the 360 data samples used, the lowest tempering temperature was 550°C for forming tempering sorbate, and the maximum proportions of C, Cr and Mo were 0.57 wt.%, 1.12 wt.% and 0.24 wt.%, respectively. Thus, the conditions required for the formation of a novel anti-fatigue high-strength steel were possibly discovered, i.e., the heat-treatment condition and compositions shown in Table 3, and the ML-predicted mechanical properties shown in Table 4. As can be seen, although the ML predictions from Equations (3-6) deviate slightly from the corresponding values from Equations (7-10), the average predicted fatigue strength (682.5 ± 27.5 MPa at a fatigue life of 10^7), tensile strength (1286 ± 48 MPa) and hardness (406 ± 16 HV) each exceed the corresponding maximum values, and the average predicted fracture strength (1922 ± 41 MPa) is comparable to the maximum reported fracture strength (1931 MPa).

Table 3. The tempering temperature and composition of the data-driven discovered anti-fatigue high strength steel

TT	C	Cr	Mo	Other Features
550 °C	0.57 wt%	1.12 wt%	0.24 wt%	Maximum value (minimize a_{Fe})

Table 4. The four mechanical properties of the data-driven discovered anti-fatigue high strength steel

Properties	Maximum value in the dataset	Predictions of Eq. (3)-(6)	Predictions of Eq. (7)-(10)	Average Predictions
FaS (MPa)	643	655	710	682.5 ± 27.5
TS (MPa)	1206	1238	1334	1286 ± 48
FrS (MPa)	1931	1881	1963	1922 ± 41
H (HV)	380	390	422	406 ± 16

4. Concluding Remarks

ML and feature selection were conducted on 360 data samples of steels to predict the fatigue strength at a fatigue life of 1×10^7 cycles and the tensile strength, fracture strength and hardness, and to find the features that were most important for the optimisation of these four mechanical properties. The ML results demonstrated that the tempering temperature and the presence of C, Cr and Mo were key to the mechanical properties of steels, with respect to which the RF model exhibited high validation accuracy (R > 0.9550, RRMSE < 30.00%). In particular, the SR gave explicit mathematic expressions of the four mechanical properties as functions of the four important features, and revealed the required features of novel an anti-fatigue high-strength steel.

Method and Software

Four ML algorithms (RF, LLS, KNN, ANN) in the WEKA software library [17] and the SR algorithm in HeuristicLab [18] were used in this study. All parameters of ML algorithms were set as the default in the open-source software, unless otherwise requested

RF: The number of features randomly chosen at each node is denoted by *numFeatures* and is determined via grid search to achieve the greatest predicting accuracy. The search results are shown in Table 5 for each feature subset. The RFI value was computed on the basis of the mean decrease impurity [19] in WEKA.

Table 5. The number of features randomly chosen for each subset

Training set	All	RFI	SRI	All-AF	RFI-AF	SRI-AF
numFeatures	7	1	2	5	2	2

KNN: The number of neighbors is denoted by *KNN* and is determined via grid search, *KNN* is recommend to be 4, 2, and 3 for All, RFI, and SRI feature subsets, respectively.

ANN: The number of hidden layers in the neural network and the learning rate of weight update are denoted by *hiddenLayers* and *learningRate*, respectively. The two hyper-parameters were determined via grid search to be *learningRate* = 0.1 and the *hiddenLayers* of 8, 7 and 7 for the All, RFI and SRI feature subsets, respectively.

SR: Genetic programming (GP) in HeuristicLab was used to search for an optimal expression. The parameters of GP that were used in this study are listed in Table 6. One hundred independent GP runs were conducted, and, based on these, the SRI value was computed as the fitness-weighted variable importance as defined in [20].

Table 6 The used parameters in GP

Parameter	*Population Size*	*Number of Generation*	*Mutation Probability*	*Crossover Probability*	*Maximum Tree Depth*	*Maximum Tree Length*
Value	1000	10000	20%	80%	10	15

Acknowledgments

This study was supported by the National Key Research and Development Program of China (No. 2018YFB0704404), the Hong Kong Polytechnic University (internal grant nos. 1-ZE8R and G-YBDH) and the 111 Project of the State Administration of Foreign Experts Affairs and the Ministry of Education, China (grant no. D16002).

References

1 Ramprasad R, Batra R, Pilania G, et al. Machine learning in materials informatics: Recent applications and prospects. npj Comput Mater, 2017, 3: 54.

2 Xue D, Xue D, Yuan R, et al. An informatics approach to transformation temperatures of NiTi-based shape memory alloys. Acta Mater, 2017, 125: 532–541.

3 Ward L, Agrawal A, Choudhary A, et al. A general-purpose machine learning framework for predicting properties of inorganic materials. npj Comput Mater, 2016, 2: 1–7.

4 Senderowitz H, Barad H-N, Yosipof A, et al. Data Mining and Machine Learning Tools for Combinatorial Material Science of All-Oxide Photovoltaic Cells. Mol Inform, 2015, 34: 367–379.

5 Agrawal A, Choudhary AN. Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. APL Mater, 2016, 4: 53208.

6 Xiong J, Shi SQ, Zhang TY. A machine-learning approach to predicting and understanding the properties of amorphous metallic alloys. Mater Des, 2020, 187: 108378

7 Takahashi K, Tanaka Y. Material synthesis and design from first principle calculations and machine learning. Comput Mater Sci, 2016, 112: 364–367.

8 Seshadri R, Wolverton C, Hill J, et al. Materials science with large-scale data and informatics: Unlocking new opportunities. MRS Bull, 2016, 41: 399–409.

9 Green ML, Choi CL, Hattrick-Simpers JR, et al. Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. Appl Phys Rev, 2017, 4: 011105.

10 Huber L, Hadian R, Grabowski B, et al. A machine learning approach to model solute grain boundary segregation. npj Comput Mater, 2018, 4: 64.

11 Zhu Q, Samanta A, Li B, et al. Predicting phase behavior of grain boundaries with evolutionary search and machine learning. Nat Commun, 2018, 9: 467.

12 Falk C, Wenny MB, Norquist AJ, et al. Machine-learning-assisted materials discovery using failed experiments. Nature, 2016, 533: 73–76.

13 Agrawal A, Deshpande PD, Cecen A, et al. Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. Integr Mater Manuf Innov, 2014, 3: 90-108.

14 Agrawal A, Choudhary A. An online tool for predicting fatigue strength of steel alloys based on ensemble data mining. Int J Fatigue, 2018, 113: 389-400

15 Yamazaki M, Xu Y, Murata M, et al. NIMS structural materials databases and cross search engine - MatNavi. VTT Symp, 2007.

16 Lison P. An introduction to machine learning. Language Technology Group: Edinburgh, UK, 2015.

17 Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. ACM SIGKDD

Explor Newsl, 2009, 11(1): 10-18.

18 Wagner S, Affenzeller M. HeuristicLab: a generic and extensible optimization environment. Adapt Nat Comput Algorithms. Vienna: Springer, 2005. 538-541.

19 Louppe G, Wehenkel L, Sutera A, et al. Understanding variable importances in forests of randomized trees. Adv Neural Inf Process Syst, 2013. 431-439.

20 Vladislavleva K, Veeramachaneni K, Burland M, et al. Knowledge mining with genetic programming methods for variable selection in flavor design. In: Proc 12th Annu Genet Evol Comput Conf, GECCO 2010. 941-948.