Supplementary Materials

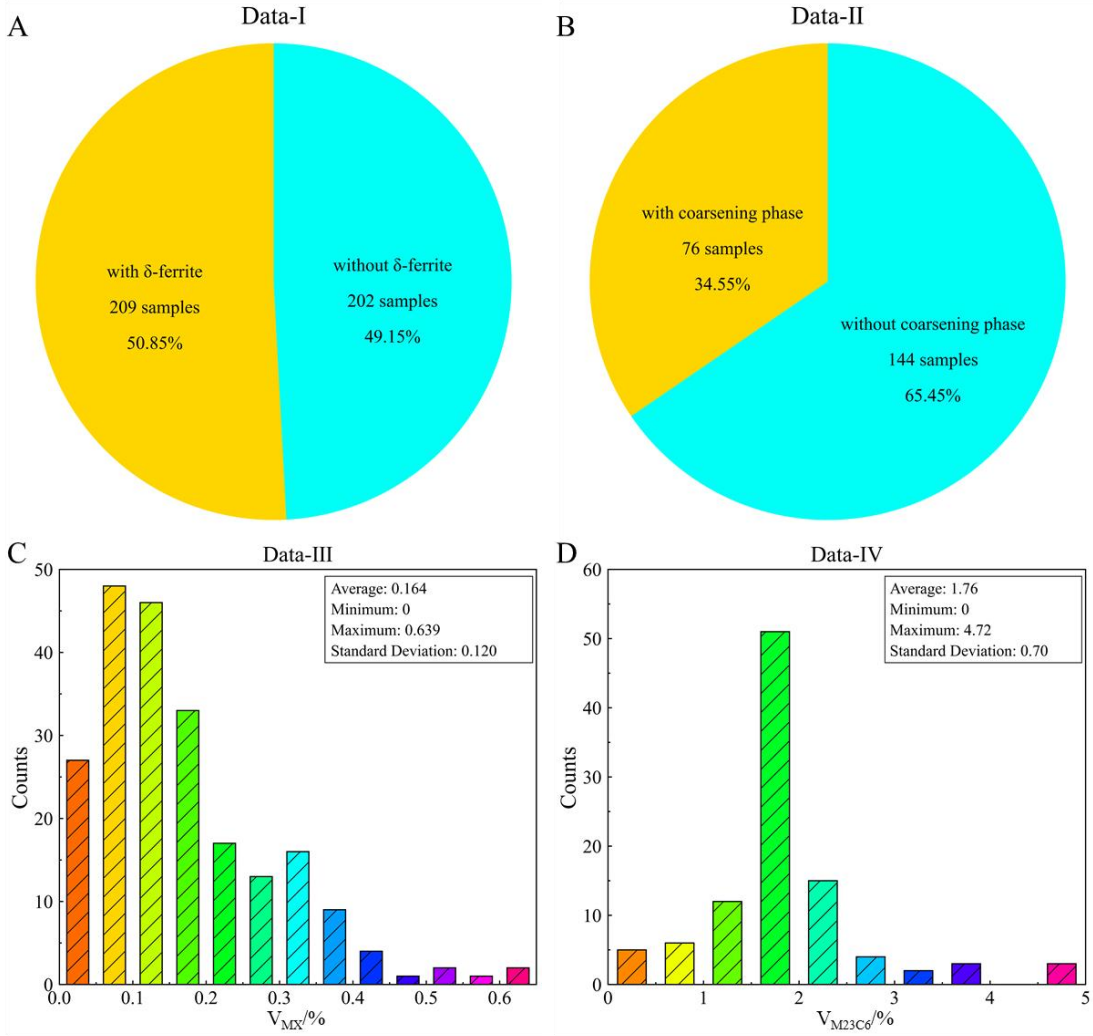
An integrated design of novel RAFM steels with targeted microstructures and tensile properties using machine learning and CALPHAD

Xiaochen Li[1,2], Mingjie Zheng[1,3,*], Hao Pan[1,3,4], Chunliang Mao[5], Wenyi Ding[1]

[1]Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, Anhui, China.
[2]School of Physics and Electronic Engineering, Jining University, Qufu 273155, Shandong, China.
[3]University of Science and Technology of China, Hefei 230026, Anhui, China.
[4]Department of Mechanical Engineering, City University of Hong Kong, Hong Kong 999077, China.
[5]College of Mechanical Engineering, Yanshan University, Qinhuangdao 066004, Hebei, China.

[*]**Correspondence to:** Prof. Mingjie Zheng, Hefei Institutes of Physical Science, Chinese Academy of Sciences, 350 Shushanhu Road, Hefei 230031, Anhui, China.
E-mail: mingjie.zheng@inest.cas.cn

1. Supplementary information on the data distribution of the microstructural dataset



Supplementary Figure 1. The data distribution of the microstructural dataset.

2. Supplementary information on the overview of commonly used machine learning algorithms

In this study, various machine learning algorithms which were commonly used in materials research, such as decision tree (DT), random forest (RF), support vector machine (SVM), gradient boosting (GB), k-nearest neighbor (KNN), and artificial neural network (ANN) were used to develop classification and regression models. The following section provides an overview of these algorithms, highlighting their unique strengths and limitations.

DT is a robust and prevalent algorithm that utilize a tree-like flowchart to effectively partition data into groups for solving classification and regression problems. It does not require complex domain knowledge [1], making them accessible for various applications. However, DT can readily result in significant prediction deviations from actual results [2]. Additionally, it is more suitable for predicting categorical features than for estimating numerical variables [3].

RF is highly regarded for its strength, flexibility, and capability in processing high-dimensional data [4]. It helps reduce overfitting compared to individual decision trees, leading to improved predictive performance. RF works well with large datasets, can accommodate a wide range of input features, and excels at determining feature importance [5]. Furthermore, it effectively handles missing values and addresses class imbalances [5]. However, using RF requires careful attention to hyperparameter tuning, such as the number of trees and the features chosen for splits. Due to its ensemble nature, the model is less interpretable than a single decision tree, making it

challenging to understand the contribution of each individual tree.

SVM is a supervised binary classification method introduced by VaPnik [6]. It is primarily designed to identify a separating hyperplane that maximizes the margin in the feature space, leading to higher classification accuracy. This segmentation maximizes the margin, transforming the problem into a convex optimization challenge. Initially designed for linear classification, SVM has evolved to tackle non-linear and high-dimensional data while effectively addressing overfitting [7]. It is known for its robustness and accuracy in managing complex, high-dimensional, and small-sample challenges [8]. However, SVM requires high computational resources and relies heavily on selecting an optimal hyperplane [9]. Despite this, its framework is effortlessly generalized for various issues, making it highly versatile [10].

GB is a machine learning algorithm for regression and classification that builds models in stages but extends this approach by optimizing a chosen differentiable loss function [11]. This algorithm assembles multiple weak models, usually decision trees, to form a more powerful predictive model [12]. The effectiveness of GB stems from the proven superiority of ensemble methods over other machine learning algorithms in various situations, making it particularly powerful for complex predictive tasks [13–15]. However, GB builds models in a stage-by-stage way [12], which may result in higher computational costs and difficulty in achieving parallelization.

KNN is a non-parametric supervised machine learning algorithm utilized for both classification and regression tasks [16]. It classifies a new data point by finding its ‘k’ nearest neighbors in the training set based on similarity. The prediction for the new

data point is then calculated as the average or weighted average of the outcomes from its ‘k’ nearest neighbors [17]. KNN is a straightforward algorithm suitable for applications across various fields and supports multiple distance measures (e.g., Euclidean, Manhattan, Minkowski), making it adaptable to different data types and problem requirements. Despite its usefulness, the KNN algorithm can be sensitive to outliers in the data, which may disproportionately affect its prediction performance.

ANN is non-linear computational model inspired by biological neural networks in the brain, utilizing interconnected neurons and weighted connections to recognize the pattern and tackle complex problems [18]. ANN is particularly effective for non-linear problems and perform well with large datasets. A basic ANN algorithm consists of input, hidden, and output layers: the input layer receives primary data, the hidden layer processes it, and the output layer generates results [19]. Common types of ANN include Feed Forward Neural Network (FFNN), Back Propagation Neural Network (BPNN), etc., among which FFNN is the most widely used [20]. ANN is well-suited for tasks like pattern recognition and matching, grouping, and classification, but they require extensive computational resources and large datasets to perform effectively and often lack interpretability.

References

[1] Pouriyeh S, Vahid S, Sannino G, et al. A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. *2017 IEEE Symposium on Computers and Communications (ISCC)*. 2017;204–207. https://doi.org/10.1109/ISCC.2017.8024530.

[2] Olu-Ajayi R, Alaka H, Sulaimon I, et al. Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. *Journal of Building Engineering* 2022; 45: 103406. https://doi.org/10.1016/j.jobe.2021.103406.

[3] Yu Z, Haghighat F, Fung BCM, et al. A decision tree method for building energy demand modeling. *Energy and Buildings* 2010; 42(10): 1637–46. https://doi.org/10.1016/j.enbuild.2010.04.006.

[4] Yang L, Shami A. IoT data analytics in dynamic environments: From an automated machine learning perspective. *Engineering Applications of Artificial Intelligence* 2022; 116: 105366. https://doi.org/10.1016/j.engappai.2022.105366.

[5] Khalid H, Khan A, Zahid Khan M, et al. Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease. *Comput Intell Neurosci* 2023; 2023: 9266889. https://doi.org/10.1155/2023/9266889.

[6] Chapelle O, Haffner P, Vapnik VN. Support vector machines for histogram-based image classification. *IEEE Trans Neural Netw* 1999; 10(5): 1055–1064. https://doi.org/10.1109/72.788646.

[7] Hussain SF. A novel robust kernel for classifying high-dimensional data using Support Vector Machines. *Expert Syst Appl* 2019; 131: 116–131. https://doi.org/10.1016/j.eswa.2019.04.037.

[8] Xu X, Liang T, Zhu J, et al. Review of classical dimensionality reduction and sample selection methods for large-scale data processing. *Neurocomputing* 2019; 328: 5–15. https://doi.org/10.1016/j.neucom.2018.02.100.

[9] Nalepa J, Kawulok M. Selecting training sets for support vector machines: a review. *Artif Intell Rev* 2019; 52(2): 857–900.

https://doi.org/10.1007/s10462-017-9611-1.

[10] Wei Y, Zhang X, Shi Y, et al. A review of data-driven approaches for prediction and classification of building energy consumption. *Renew Sust Energ Rev* 2018; 82: 1027–1047. https://doi.org/10.1016/j.rser.2017.09.108.

[11] Flores V, Keith B. Gradient Boosted Trees Predictive Models for Surface Roughness in High-Speed Milling in the Steel and Aluminum Metalworking Industry. *Complexity* 2019; 1536716. https://doi.org/10.1155/2019/1536716.

[12] Ahmed N, Ahammed R, Islam MdM, et al. Machine learning based diabetes prediction and development of smart web application. *International Journal of Cognitive Computing in Engineering* 2021; 2: 229–241. https://doi.org/10.1016/j.ijcce.2021.12.001

[13] Dietterich TG. Ensemble Methods in Machine Learning. Multiple Classifier Systems. Berlin, Heidelberg: Springer; 2000;1–15. https://doi.org/10.1007/3-540-45014-9_1

[14] Berk RA. An introduction to ensemble methods for data analysis. *Sociol Methods Res* 2006; 34(3): 263–295. https://doi.org/10.1177/0049124105283119.

[15] Zhang C-X, Zhang J-S. A novel method for constructing ensemble classifiers. *Stat Comput* 2009; 19(3): 317–327. https://doi.org/10.1007/s11222-008-9094-7.

[16] Ortiz-Bejar J, Graff M, Tellez ES, et al. k-Nearest Neighbor Regressors Optimized by using Random Search. 2018 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC). 2018;1–5. https://doi.org/10.1109/ROPEC.2018.8661399.

[17] Chen Y, Hao Y. A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. *Expert Systems with Applications* 2017; 80: 340–355. https://doi.org/10.1016/j.eswa.2017.02.044.

[18] Amasyali K, El-Gohary NM. A review of data-driven building energy consumption prediction studies. *Renew Sust Energ Rev* 2018; 81: 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095.

[19] Bourdeau M, Zhai XQ, Nefzaoui E, et al. Modeling and forecasting building energy consumption: A review of data-driven techniques. *Sust Cities Soc* 2019;

48: 101533. https://doi.org/10.1016/j.scs.2019.101533.

[20] Ahmad MW, Mourshed M, Rezgui Y. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy Build* 2017; 147: 77–89. https://doi.org/10.1016/j.enbuild.2017.04.038