Scripta Materialia 209 (2022) 114366

Contents lists available at ScienceDirect

Scripta Materialia

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

Balancing data for generalizable machine learning to predict glass-forming ability of ternary alloys

Yi Yao[a], Timothy Sullivan IV[a], Feng Yan[a], Jiaqi Gong[b], Lin Li[a,*]

[a] Department of Metallurgical and Materials Engineering, The University of Alabama, Tuscaloosa, AL 35487, USA
[b] Department of Computer Science, The University of Alabama, Tuscaloosa, AL 35487, USA

ARTICLE INFO

Article history:
Received 19 May 2021
Revised 29 August 2021
Accepted 18 October 2021
Available online 30 October 2021

Keywords:
Machine learning
Data imbalance
Metallic glass
Glass-forming ability
Artificial neural network

ABSTRACT

Machine Learning has thrived on the emergence of data-driven materials science. However, the materials datasets acquired at existing research efforts have significant imbalance issues. This paper investigated the data imbalance for the glass-forming ability of ternary alloy systems, which consists of abundant, low-fidelity high-throughput data, and sparse, high-fidelity traditional experimental data. We demonstrated a new method to handle the data imbalance and trained artificial neural network (ANN) models on the original vs. balanced datasets. The ANN model trained on the balanced dataset solved the overfitting issue suffered by the model trained on the original dataset. More importantly, the generalizability in predicting the new alloy system was improved in the data-balanced model, evidenced by the leave-one-alloy-system-out validation. Our work highlights the importance of handling data imbalance in material datasets to solve the overfitting issues of machine learning models and further enhance generalizability in predicting the characteristics of the new material systems.

Metallic glasses (MGs), owing to the amorphous structure, exhibit various remarkable properties that are difficult to achieve in crystal materials [1,2]. The development and application of MGs are largely hindered by the limited glass-forming ability (GFA) of metallic systems. Tremendous efforts have been made to explore alloy systems and compositions with enhanced GFA [1,3-7]. Inoue summarized an empirical rule for high GFA systems, which are composed of more than three elements with negative heats of mixing (ΔH_{mix}) of the liquid phases and an atomic size difference (δ) above 12% [1]. Yang *et al.* have derived a thermodynamic model with a parameter Ω ($= T_m \cdot \Delta S_{mix} / |\Delta H_{mix}|$, T_m is melting point, ΔS_{mix} is entropy) based on the experimental observations. They demonstrated that MGs were formed when the parameter $\Omega \geq 1$ and $\delta \geq 6\%$ [7]. However, the key thermodynamic and physical factors to determine the GFA are still mysteries, as a result of the non-equilibrium nature of glass forming processes.

Machine learning (ML), capable of learning the underlying statistics of materials datasets, emerges as a powerful tool to tackle the long-standing issues in the materials field [8–19]. It has been accelerating the discovery of new materials and uncovering the hidden mechanisms that control material structure, processing, and properties [8-15,17–19]. Specifically, in the field of MG development, researchers have gained a deeper understanding of GFA with the assistance of ML models [20–25]. Sun *et al.* utilized a Support Vector Machine algorithm to study the GFA of as-cast alloys and found that the differences in liquidus temperature and fictive liquid temperature were the critical features to determine GFA [21]. Fang and Logan et al. combined ML and high-throughput experiments to explore the GFA of ternary alloys, discovering two new ternary alloy systems with high GFA that could be omitted by conventional alloy development approaches [25]. Till now, most of the work on GFA has focused on material featurization and ML algorithms, limited study considers the quality of the material dataset and its influence on the ML model performance.

The successful incorporation of ML models into materials research requires overcoming the challenge of data scarcity and imbalance inherent to the material datasets [26]. Traditional material data collections rely heavily on dedicated and costly experiments, which results in data scarcity in materials science when compared to other big data fields [27]. The published data contains high-fidelity and successful results, yet excludes the failures, leading to skewed data distributions (data imbalance, i.e. the successful samples dominate the datasets) [26]. Such imbalance mars the central assumption of ML models: both training and testing data should be independent and identically distributed [28]. Recently, high-throughput experiments and simulations have been widely adopted to address data scarcity and imbalance. The high-throughput efforts have yielded abundant data, including both improved and deteriorated results on material structure and property,

* Corresponding author.
E-mail address: lin.li@eng.ua.edu (L. Li).

https://doi.org/10.1016/j.scriptamat.2021.114366
1359-6462/

greatly enhancing the dataset size [25,29]. And yet the data yielded by the high-throughput efforts are in general low-fidelity. For instance, to explore the GFA in large compositional space, Fang *et al* employed an in-house sputtering model to estimate the composition of the data points, instead of experimentally measuring every point. Such estimation could lead to a maximum error of 5% [25]. The incorporation of diversified data sources (e.g. traditional literature data vs. high-throughput data, simulation vs. experimental data) poses a new data imbalance issue, i.e. the low-fidelity high-throughput data dominates the ML model performance. The high-fidelity traditional literature data are scarce but essential, which deliver more accurate material information to train the ML models. The high-throughput data overcome the issures of data scarcity and imbalance, but they are lower-fidelity and less accurate in materials information for the ML models. How to combine the two kinds of datasets to provide underlying knowledge to the ML model and further avoid the side effect of the domination of the high-throughput data become an emergent question to be answered.

In this work, we present a method to handle the dataset imbalance, i.e. the high-throughput data dominate the dataset in GFA of ternary alloy systems. We balanced the dataset by systematically reducing abundant, low-fidelity high-throughput data and augmenting sparse, high-fidelity traditional experimental data. Two artificial neural network (ANN) models were built based on the original (imbalanced) and balanced datasets. The ANN model trained on the balanced dataset solves the overfitting issue suffered by the model trained on the original datasets, exhibiting 42% improvement in the model performance on the alloy systems with sparse data, and maintaining nearly the same performance on the alloy systems with abundant data. More importantly, the ANN model trained on the balanced dataset has a 31% improvement in predicting unseen alloy systems in the leave-one-alloy-system-out validation when compared to the model trained on the original dataset. Our work highlights the importance of data balancing in applying ML model to the material dataset and provides a practical approach to solving the overfitting issues of ML models and improve the model's generalizability to explore new alloy systems.

Our training dataset consists of 5725 alloys fabricated by magnetic sputtering [25,29,30], and the data are categorized based on the alloy structures, i.e. crystal vs. amorphous. The collected dataset has 20 alloy elements, as shown in Fig. S1 in the supplementary materials (SM), and 1997 (34.88%) alloys are amorphous structure and 3728 (65.12%) alloys are crystal. The data distribution in the different alloy systems is illustrated in Fig. 1**a**. Specifically, the data were obtained from two types of experiments: high-throughput vs. traditional. The high-throughput dataset includes 5 ternary alloy systems (i.e. CoFeZr, CoTiZr, CoVZr, FeTiNb, and AlNiTi), consisting of 5568 data points (~97.3 % of the dataset). The traditional experiments provide the data on the 12 alloy systems but only have 157 data points (~2.7% of the dataset). A significant data imbalance issue emerges, the high-throughput experiments provide fewer but data-abundant alloy systems vs. traditional experiments have more but the data-sparse alloy systems. It is noteworthy that such a data imbalance issue is quite prevalent when compiling various data sources to train ML model for materials investigation [25,31].

To balance the dataset [28,32], we applied data reduction and data augmentation to the high-throughput data and the traditional data, respectively. For data reduction, the number of data points for each high-throughput ternary alloy system was reduced *uniformly* in the compositional space to 200 from ~ 1000. For data augmentation, we increased data from the existing compositions with a step of 0.1 at.%, which is within the error of compositional measurements. For instance, starting from the Al50Cu30Fe20 (at.%) alloy with the crystal structure, we varied the element compositions with 0.1 at.%, resulting in 6 new ternary alloys all labeled with the same crystal structure as the original one. To ensure fidelity, if two data points had different structures and their composition varies by less than 2 at.%, no augmentation is conducted. It is noteworthy that there is an inevitable error upon data augmentation. Our study on error tolerance of labeling the augmented data find that the models trained on the balanced dataset can tolerate 1% label change/error (details in Fig. S7, S8, and S9 in SM). After augmentation, a duplication check will be performed to remove any data with the same composition. And if the number of data points is larger than 200, data reduction would be applied to control the number to 200. Consequently, after data processing, as shown in Fig. 1**b**, the balanced dataset has 1983 alloys, in which 765 (38.58%) are crystals and 1218 (61.42%) are amorphous. The data distribution of the balanced dataset is illustrated in Fig. 1**b**, and Table S1 in SM.

According to the previous studies on the GFA of MGs [18,23,31], 131 features are used to describe the alloys as the input of the ML model (details in Table S2 in SM). The 131 features contain atomic properties (*e.g.* atomic radii, atomic radius mismatch), atomic packing properties, and thermodynamic properties (*e.g.* heat of mixing and configurational entropy of mixing), covering a comprehensive list used for general-purpose as well as for the MG systems. Specifically, the probability density distribution of data for the most relevant 13 features [1,8,14] before and after data processing is illustrated in Fig. 2. The dashed and solid lines denote the crystal (CR) and amorphous (AM) structures, respectively. After data balancing, the number of data points in each alloy system is roughly equal, leading to a more uniform and broader data distribution of the balanced dataset in the feature space. Especially for the AM data, the sharp peaks in the features of σ_{VEC}, ΔS_{mix}, *PE*, *R*, δ, T_m, σ_{Tm}, $\Delta Hmix$, $\sigma_{\Delta Hmix}$, *B*, σ_B have been smoothed after balancing the data numbers between the high-throughput and traditionally obtained alloys systems. The probability density distributions of all the features upon data processing are available at GitHub. Next, two ML models are trained on the original and balanced datasets, respectively. We demonstrate that the balanced data features can solve the overfitting of the ML models to the dominant data clusters, enhancing the generalizability of the ML models.

An ANN algorithm is used to build the ML model. The ANN has a feed-forward structure with one input layer (131 normalized features), two hidden layers (250 neurons in the first hidden layer, 25 neurons in the second hidden layer), and one output layer (one neuron). The architecture of the ANN model is shown in Fig. S3 in SM, and the convergence study of hyperparameters is provided in Table S5 in SM. Two ANN models are trained on the original and balanced datasets, respectively. To evaluate the ANN model performance, 10-fold cross-validation (CV) is performed. The receiver operating characteristic (ROC) curve is constructed according to the results of the 10-fold CV, and then the area under the ROC curve (AUC) is evaluated. Additionally, the root mean square error (*RMSE*) for different datasets and different alloy systems is also calculated as $RMSE = \sqrt{\sum_{i}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$, where n is the total number of data, $\hat{y}_i$ is the predicted GFA (between 0-1), and y_i is 0 for crystal and 1 for amorphous.

The performance of the ANN models trained on the original dataset (ANN-original) and balanced dataset (ANN-balanced) in the 10-fold CV validation are summarized in Fig. 3 and Table 1. Notably, both models show high accuracy in the overall performance with AUC = 0.9963, and *RMSE* = 0.1478 for the ANN-original, and AUC = 0.9958, and *RMSE* = 0.1384 for the ANN-balanced. When delving into the data subsets, the overfitting of the ANN-original is identified: the model shows a lower *RMSE* = 0.1399 on the high-throughput subset but much higher *RMSE* = 0.2008 on the tradi-

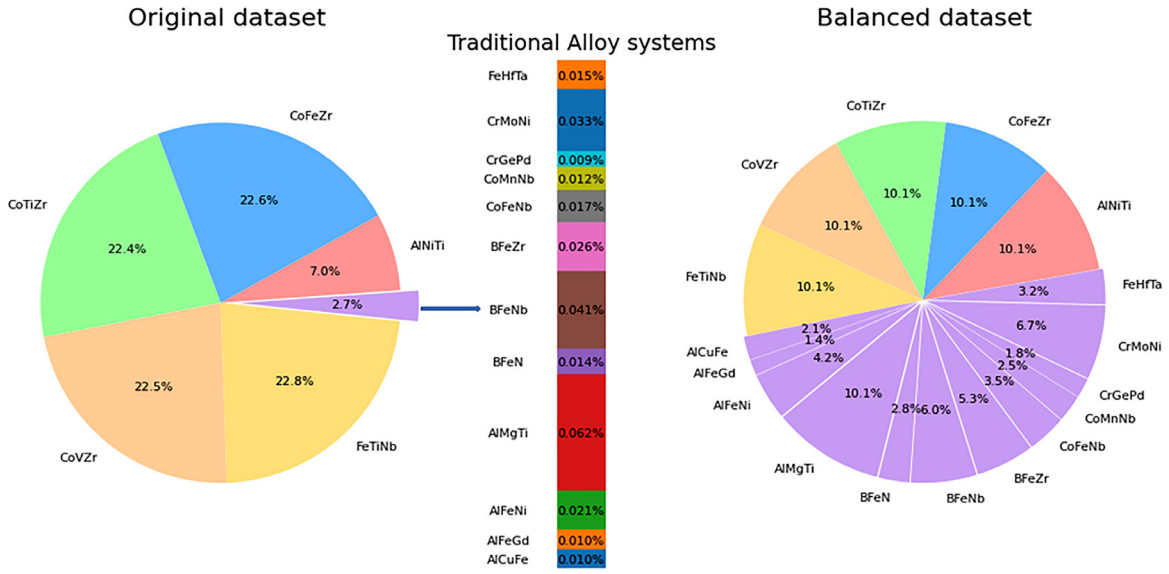

Fig. 1. The data distribution of the original vs. balanced datasets. After data processing, the balanced dataset shows a more uniform data distribution among different alloy systems.

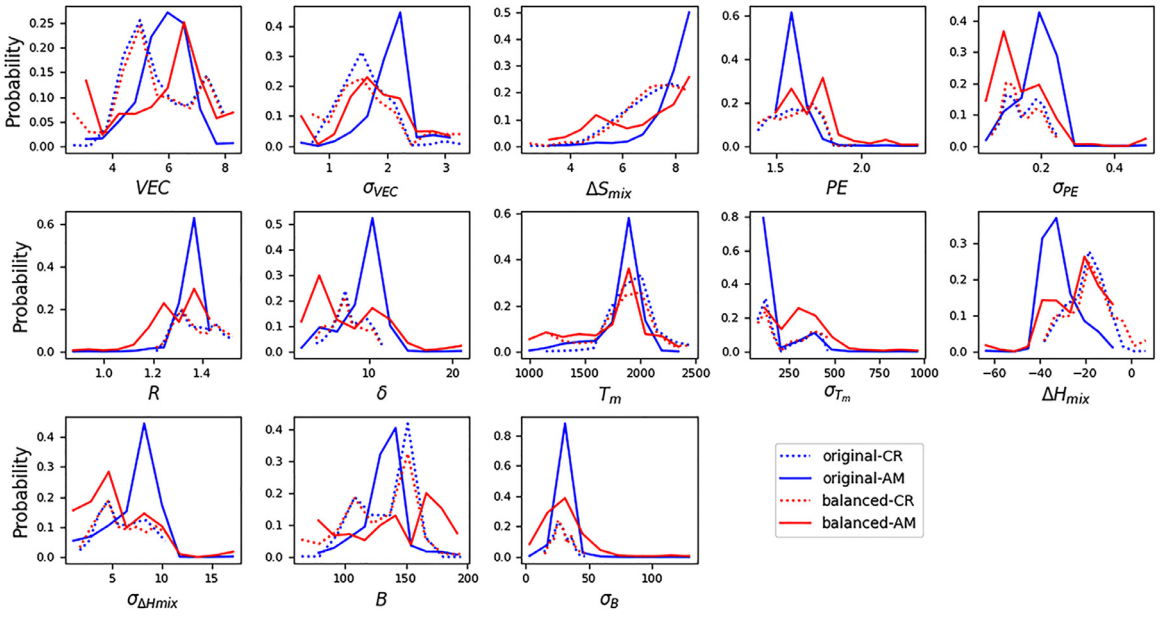

Fig. 2. The probability density distribution of data for 13 different features. Blue and red colors represent the original and balanced datasets, respectively. Dashed and solid lines denote the alloys with crystal (CR) and amorphous (AM) structures. After data processing, the distribution of the CR data changes slightly; whereas the distribution of AM data changes significantly. *VEC, PE, R,* T_m, ΔH_{mix}, *B* represent the mean value of the valence electron concentration, Pauling electronegativity, covalent radius, melting temperature, the heat of mixing, and bulk modulus; and σ calculates the standard deviation of those quantities. ΔS_{mix} and δ stand for entropy and covalent difference, respectively (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Table 1
The performance of the original and balanced ML models.

Models		AUC	*RMSE*
ANN-original	overall	0.9963	0.1478
	high-throughput	0.9974	0.1399
	traditional	0.9744	0.2008
ANN-balanced	overall	0.9958	0.1384
	high-throughput	0.9955	0.1579
	traditional	0.9920	0.1160

tional dataset. In contrast, the ANN-balanced significantly improves the performance on the traditional subset with *RMSE* = 0.1160 and maintains the performance on the high-throughput subset with *RMSE* = 0.1579 (ref. to Table 1). Moreover, the *RMSEs* of each ternary alloy system for the two models are calculated and shown in Fig. 3**(b)**. The ANN-original has smaller *RMSEs* for the 4 high-throughput alloy systems (CoFeZr, CoTiZr, CoVZr, and FeTiNb) than those of the ANN-balanced. The exception is the AlNiTi alloy system, in which the *RMSE* remains nearly unchanged. On the other hand, the ANN-balanced shows significant improvement on the alloy systems started with sparse data. For the AlCuFe and AlFeGd the *RMSEs* reduced by ~83% from ~0.46 to ~0.08. The detailed results of the *RMSEs* of each ternary alloy system in 10-fold CV are shown in Table S3 in SM. Even though the ANN-balanced does not have significantly better performance than the ANN-original on the overall dataset, it solves the over-fitting suffered by ANN-original. The ANN-original is predominantly controlled by the high-throughput alloy systems (*RMSE* = 0.1399), resulting in large errors

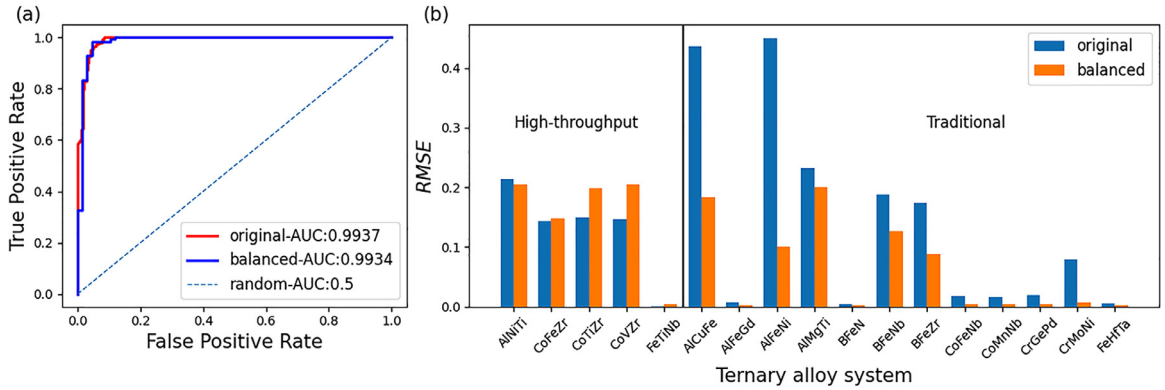

Fig. 3. The performances of the original and balanced ML models in the 10-fold CV. (a) The ROC curves and calculated AUCs of the two ML models. The AUCs show that both two models have overall good performances under 10-fold CV. (b) The *RMSE* of two models for each alloy system. The original model has slightly smaller *RMSE*s in high-throughput data than those of the balanced model. The balanced model has much better *RSME*s for the other ternary alloy systems in the traditional data. These results indicate that the ANN-original is suffering from overfitting.

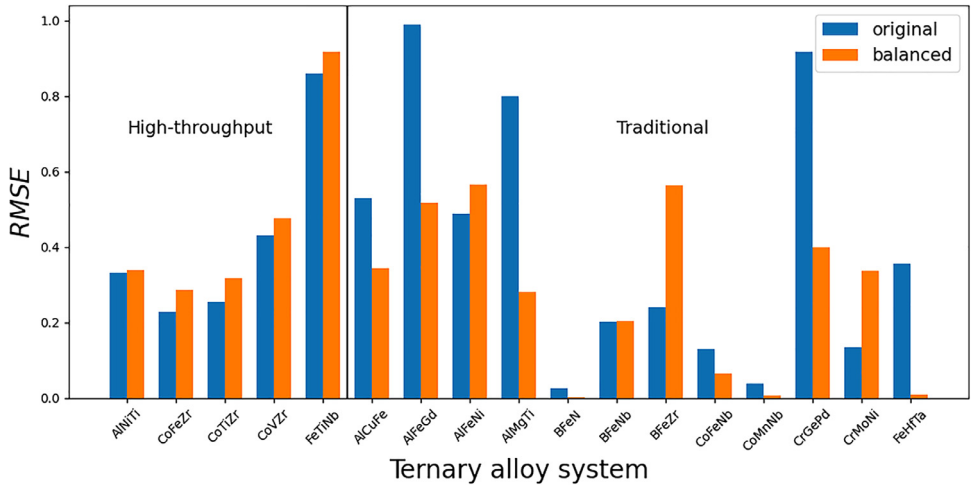

Fig. 4. The *RMSE*s of the original and balanced models for each ternary alloy system in the LEAVE-ONE-ALLOY-SYSTEM-OUT validation. The balanced model shows a much better performance than the original model, demonstrating enhanced generalizability in predicting the new alloy system data.

in the traditional data subset (e.g. *RMSE* for AlFeGd = 0.456186). Furthermore, we will demonstrate that the overfitted model is not able to explore new alloy systems with high accuracy, while the balanced model that avoids overfitting enhances the generalizability for new alloy systems.

The ultimate goal of the ML model is to predict the GFA of new alloy systems that have not been seen by the initial ML models, and we refer to this capability as the generalizability of the ML models. To evaluate the generalizability of the ML model, we perform leave-one-alloy-system-out validation. Specifically, upon developing the ML models, we leave the data in one alloy system completely out of the training dataset, and such alloy system only serves as the validation dataset. Since the ML model has not been trained on the alloy system, the performance of the model in the leave-one-alloy-system-out validation can be treated as the generalizability of the model to the new alloy system. It is noteworthy that this is distinct from the 10-fold CV performance, in which the training dataset includes approximately 90% data from all the alloy systems, and thus the performance cannot represent the generalizability of the ML models.

Fig. 4 displays the *RMSE*s of the ANN-original and ANN-balanced models for each alloy system in the leave-one-alloy-system-out validation. When comparing the two models, the ANN-balanced model shows nearly identical performance in the high-throughput alloy systems, but a much better performance in the traditional ones. The average *RMSE* of all the predicted alloy systems is reduced by 31%, from 0.4587 to 0.3170. For example, the *RMSE*s of the AlCuFe, AlFeGd, AlMgTi, CoFeNb, CrGePd, and FeHfTa systems significantly decrease after data balancing. The *RMSE*s of each ternary alloy system in the leave-one-alloy-system-out validation are shown in Table S4 in SM. The overall performance of the ANN-balanced model in the leave-one-alloy-system-out validation demonstrates its better generalizability to a new alloy system.

Fig. 5 illustrates the predicted GFA of the AlMgTi alloy system in the leave-one-alloy-system-out validation for (a) ANN-original vs. (b) ANN-balanced, along with the experimental data shown in blue (crystal) and red (amorphous). The ANN-original predicts low GFA throughout the entire compositional region of AlMgTi, and cannot identify the boundary between two the crystal and amorphous phases in this alloy system. Such prediction results from the overfitting in the ANN-original, i.e. the high-throughput data dominates the model performance. Notably, after data processing, the ANN-balanced can identify the two-phase boundary even without seeing this AlMgTi alloy system before. The balanced dataset of AlNiTi, AlCuFe, and AlFeGd provides more information for the ANN-balanced to do accurate prediction in AlMgTi. The comparison between the two models under leave-one-alloy-system-out validation for other alloy systems is shown in Fig. S4 vs. Fig. S5. It is noteworthy that our data balancing method and the ANN-balanced model so far do not handle extreme cases well when the initial data distribution is clustered around only one corner of the composition space, e.g. BFeZr alloy system.

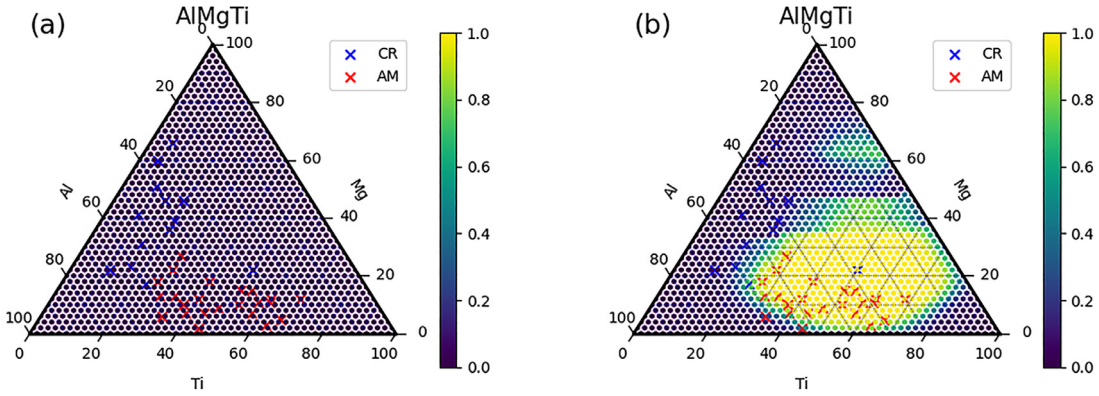

Fig. 5. The predicted GFA for AlMgTi alloy system under leave-one-alloy-system-out validation. (a)The ANN-original model; (b) the ANN-balanced model. The symbol x represents the experimental data points.

In this study, we investigated the imbalance issues of the datasets for GFA of ternary alloy systems, which consist of abundant, low-fidelity high-throughput data and sparse, high-fidelity traditional data. We proposed a method to handle the data imbalance issues and trained ANN models on the original and balanced datasets. The 10-fold cross-validation and leave-one-alloy-system-out validation were employed to evaluate the model performance. The 10-fold cross-validation results reveal that the model trained on the original dataset suffers from overfitting to the abundant high-throughput data. The model trained on the balanced dataset exhibits 42% improvement in the alloy systems with sparse data, and maintains the performance in the alloy systems with high-throughput data. More importantly, in the leave-one-alloy-system-out validation, the balanced model has a 31% improvement in predictive ability for unseen alloy systems when compared to the original model. In addition, the data-balancing approach also leads to better and more stable performance for the decision tree (DT) and support vector machine (SVM) algorithms, not only the ANN algorithms. Our work highlights the importance of data balancing in applying ML models to the material field and provides a practical approach to balance the dataset of ML models and improve the model's generalizability to explore new alloy systems. The data, codes, and trained ANN models that support the findings of the present work are available at GitHub, (https://github.com/linear85/TernaryAlloyGFA_ANN_Model).

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.

Acknowledgments

This work is supported by the U.S. National Science Foundation (CMMI-1727875). Additional support from the Alabama Cyber Institute at the University of Alabama is acknowledged.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.scriptamat.2021.114366.

References

[1] A. Inoue, Acta Mater. 48 (1) (2000) 279–306.
[2] W.H. Wang, C. Dong, C.H. Shek, Mater. Sci. Eng. 44 (2-3) (2004) 45–89.
[3] K.J. Laws, D.B. Miracle, M. Ferry, Nat. Commun. 6 (2015) 8123.
[4] K.J. Laws, K.F. Shamlaye, K. Wong, B. Gun, M. Ferry, Metall. Mater. Trans. A 41 (7) (2010) 1699–1705.
[5] D.B. Miracle, Nat. Mater. 3 (10) (2004) 697–702.
[6] H.W. Sheng, W.K. Luo, F.M. Alamgir, J.M. Bai, E. Ma, Nature 439 (7075) (2006) 419–425.
[7] X. Yang, Y. Zhang, Mater. Chem. Phys. 132 (2-3) (2012) 233–238.
[8] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, Nat. Commun. 9 (1) (2018) 3405.
[9] J.C. Mauro, A. Tandia, K.D. Vargheese, Y.Z. Mauro, M.M. Smedskjaer, Chem. Mater. 28 (12) (2016) 4267–4277.
[10] T.D. Sparks, M.W. Gaultois, A. Oliynyk, J. Brgoch, B. Meredig, Scr. Mater. 111 (2016) 10–15.
[11] D. Xue, P.V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, Nat. Commun. 7 (2016) 11241.
[12] Z. Fan, J. Ding, E. Ma, Mater. Today 40 (2020) 48–62.
[13] Y.-J. Hu, G. Zhao, M. Zhang, B. Bin, T. Del Rose, Q. Zhao, Q. Zu, Y. Chen, X. Sun, M. de Jong, L. Qi, npj Computat. Mater. 6 (1) (2020).
[14] X. Liu, X. Li, Q. He, D. Liang, Z. Zhou, J. Ma, Y. Yang, J. Shen, Acta Mater. 201 (2020) 182–190.
[15] C.W. Rosenbrock, E.R. Homer, G. Csányi, G.L.W. Hart, npj Computat. Mater. 3 (1) (2017).
[16] L. Tian, Y. Fan, L. Li, N. Mousseau, Scr. Mater. 186 (2020) 185–189.
[17] M. Wagih, P.M. Larsen, C.A. Schuh, Nat. Commun. 11 (1) (2020) 6376.
[18] P. Liu, H. Huang, S. Antonov, C. Wen, D. Xue, H. Chen, L. Li, Q. Feng, T. Omori, Y. Su, npj Comput. Mater. 6 (1) (2020).
[19] Z. Zhou, Y. Zhou, Q. He, Z. Ding, F. Li, Y. Yang, npj Comput. Mater. 5 (1) (2019).
[20] B. Deng, Y. Zhang, Chem. Phys. 538 (2020).
[21] Y.T. Sun, H.Y. Bai, M.Z. Li, W.H. Wang, J. Phys. Chem. Lett. 8 (14) (2017) 3434–3439.
[22] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, npj Comput. Mater. 2 (1) (2016).
[23] L. Ward, S.C. O'Keeffe, J. Stevick, G.R. Jelbert, M. Aykol, C. Wolverton, Acta Mater. 159 (2018) 102–111.
[24] J. Xiong, T.-Y. Zhang, S.-Q. Shi, MRS Commun. 9 (02) (2019) 576–585.
[25] F. Ren, L. Ward, T. Williams, K.J. Laws, C. Wolverton, J. Hattrick-Simpers, A. Mehta, Sci. Adv. 4 (4) (2018) eaaq1566.
[26] C. Suh, C. Fare, J.A. Warren, E.O. Pyzer-Knapp, Annu. Rev. Mater. Res. 50 (1) (2020) 1–25.
[27] Y. Zhang, C. Ling, npj Comput. Mater. 4 (1) (2018).
[28] B. Krawczyk, Prog. Artificial Intell. 5 (4) (2016) 221–232.
[29] H. Joress, B.L. DeCost, S. Sarker, T.M. Braun, S. Jilani, R. Smith, L. Ward, K.J. Laws, A. Mehta, J.R. Hattrick-Simpers, ACS Comb. Sci. 22 (7) (2020) 330–338.
[30] Y. Kawazoe, J.-Z. Yu, A.-P. Tsai, T. Masumoto, Springer (1997).
[31] M. Samavatian, R. Gholamipour, V. Samavatian, Comput. Mater.Comput. Mater. Sci. 186 (2021).
[32] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, J. Artificial Intelligence Res. 16 (2002) 321–357.