crystals

MDPI

Article

The Prediction of the Undercooling Degree of As-Cast Irons and Aluminum Alloys via Machine Learning

Yong Chen [1], Litao Wen [1,2,3], Shuncheng Wang [4], Zhibo Zhang [2,3,*], Cuicui Yin [2,3], Nan Zhou [2,3] and Kaihong Zheng [2,3,*]

1 College of Mechanical Engineering, University of South China, Hengyang 421000, China; 201820620164@stu.edu.cn (Y.C.); 201820620150@stu.edu.cn (L.W.)
2 Institute of Materials and Processing, Guangdong Academy of Sciences, Guangzhou 510651, China; yincuicui@gimp.gd.cn (C.Y.); zhounan@gimp.gd.cn (N.Z.)
3 Guangdong Provincial Key Laboratory of Metal Toughening Technology and Application, Guangzhou 510000, China
4 Guangdong Xingfa Aluminum Co., Ltd., Foshan 528137, China; dennispang@xingfa.com
* Correspondence: zhangzhibo@gimp.gd.cn (Z.Z.); zhengkaihong@gimp.gd.cn (K.Z.)

Abstract: As-cast irons and aluminum alloys are used in various industrial fields and their phase and microstructure properties are strongly affected by the undercooling degree. However, existing studies regarding the undercooling degree are mostly limited to qualitative analyses. In this paper, a quantitative analysis of the undercooling degree is performed by collecting experimental data and employing machine learning. Nine machining learning models including Random Forest (RF), eXtreme Gradient Boosting (XGBOOST), Ridge Regression (RIDGE) and Gradient Boosting Regressor (GBDT) methods are used to predict the undercooling degree via six features, which include the cooling rate (CR), mean atomic covalence radius (MAR) and mismatch (MM). Four additional effective models of machine learning algorithms are then selected for a further analysis and cross-validation. Finally, the optimal machine learning model is selected for the dataset and the best combination of features is found by comparing the prediction accuracy of all possible feature combinations. It is found that RF model with CR and MAR features has the optimal performance results for predicting the undercooling degree.

Keywords: undercooling degree; machine learning; as-cast irons; aluminum alloys; cooling rate; mean covalent atomic radius

Citation: Chen, Y.; Wen, L.; Wang, S.; Zhang, Z.; Yin, C.; Zhou, N.; Zheng, K. The Prediction of the Undercooling Degree of As-Cast Irons and Aluminum Alloys via Machine Learning. *Crystals* **2021**, *11*, 432. https://doi.org/10.3390/cryst11040432

Academic Editors: David Holec and Cyril Cayron

Received: 15 March 2021
Accepted: 14 April 2021
Published: 16 April 2021

1. Introduction

In terms of production volumes and application scales, iron and aluminum are two of the mostly utilized metals in the world. They have found applications in various industries such as mechanical engineering and shipping [1,2]. Due to its advantages of relatively low cost and wide processing adaptability, casting is one of the main methods of iron and aluminum material preparation. The casting process is always accompanied by the nucleation process, which plays an important role in metal solidification. The undercooling degree strongly affects the nucleation and additionally controls the phase composition, microstructure, properties and quality of as-cast materials [3–5].

There are many factors affecting the undercooling degree such as the metal nature, the cooling rate, the mismatch magnitude, the interfacial energy of the molten metals and the nucleated solid phase [6–9]. Generally, the undercooling degree increases with the cooling rate, which consequently increases both the nucleation and growth rates. Due to the limitation of the heat transfer process, rapid solidification technology, which is widely used in the industry, can only prepare alloys with extremely small dimensions. With the outstanding advances in deep undercooling technology, many metals and alloys have achieved relatively large undercooling degrees, which have greatly exceeded the

critical undercooling degree for the homogeneous nucleation of liquid metals [10]. Deep undercooling techniques primarily include melt immersion floatation [11], suspension without vessel treatment [12] and free fall methods [13]. These methods can ensure a liquid metal state hundreds of degrees Celsius below the liquid phase line and then suddenly achieve a fast-solidification microstructure via nucleation. The development of deep undercooling technology via rapid solidification can contribute to grain refinement, the elimination of segregation, the expansion of the solid solution limit and the formation of sub-stable phases. Thus, material properties are improved. Battersby et al. [14] used a melt encasement (fluxing) technique to achieve high undercooling and systematically studied the velocity-undercooling relationship in samples of pure Ge and Ge doped with 0.01 at % Fe at undercooling up to 300 K. Jian et al. [15] studied the effect of undercooling on crystal-liquid interface energy in the growth mode of undercooled semiconductors. Li et al. [16] utilized containerless electromagnetic levitation processing to obtain the undercooling of 420 K using a two-step heating method in elemental semiconductor silicon. Li et al. [17] investigated Fe alloy melts containing 7.5, 15, 22.5 and 30 at% Ni and found that the undercooled degree had a strong influence on the structure evolution especially for grain refinement and recrystallization. Previously conducted investigations have mainly focused on the mechanism and qualitative analysis of deep undercooling. A model that could accurately predict the undercooling degree and thus increase the experimental and industrial cost-effectiveness as well as improve the processing accuracy is required but is not yet established.

With the rapid development of material informatics, machine learning (ML) has emerged as a new method to quantitatively predict material parameters based on a specific dataset. Various useful predictions have been performed by ML to obtain a quantitative analysis. Agrawal et al. [18] established ML models to predict the fatigue strength of steel, quantitatively analyze its relationship with the composition and processing parameters and eventually develop steels with a high fatigue strength. By employing support vector regression, sequenced minimum optimization regression and a multi-layer perceptron algorithm, Jiang et al. [19] proposed a model that accounted for the chemical composition, dendrite crystal parameters and measured temperature to predict the interface mismatch. Its accuracy was verified by the empirical formula and validated by the experimental results. Based on a database of density functional theory calculations, an ML model was developed by Meredig et al. [20] to predict the thermodynamic stability of arbitrary components without any additional inputs. Javed et al. [21] proposed a lattice constant prediction model based on the support vector machine. This model could optimize the lattice constants of perovskites with predetermined structures. The model was proven to be more efficient, faster and robust than the models based on the artificial neural network. In addition, ML has proven effective in material data mining collected from experiments or simulations as well as in the accurate prediction of material behavior. However, as different ML algorithms result in a variation of the prediction accuracy, the ML algorithm that should be employed for a specific material still requires further discussion.

In this paper, nine popular ML algorithms are considered to build a model for the undercooling degree prediction by mining the data from previously conducted experimental results. In Figure 1, the workflow diagram of this paper is presented. First, data samples are collected and filtered based on the literature survey. Following standardization, the data are then divided into training and testing sets according to a certain ratio. Subsequently, after nine ML models are used to mine data samples, four ML models are chosen based on their superior performance. Next, to achieve the accurate prediction of the material undercooling degree, the optimal model is obtained by comparing its evaluation indexes with the ones from the remaining models. Finally, the influence of different feature combinations on the prediction accuracy is investigated using the selected optimal model. Compared with previous qualitative understandings, we establish a model that can accurately predict the undercooling degree and thus increase the experimental and industrial cost-effectiveness

as well as improve the processing accuracy. A quantitative analysis of the undercooling degree for the sake of an accurate industrial and experimental control is of great interest.

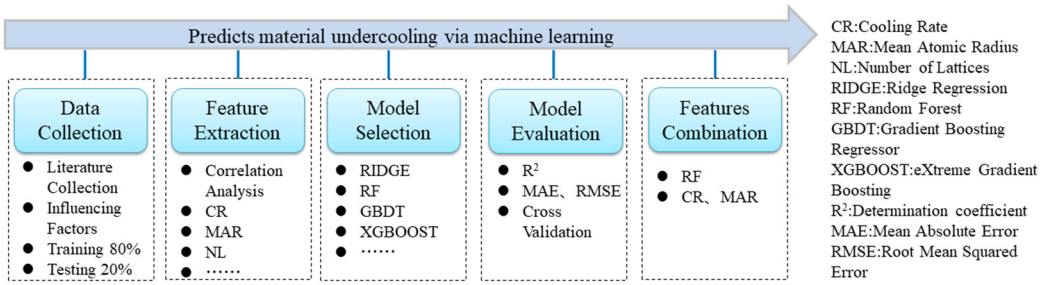


Figure 1. Strategy and machine learning (ML) model workflow for a given material ML prediction model.

2. Data Collection and Computation Method

2.1. Data Collection and Features Selection

In this paper, based on the conducted literature survey on undercooling [6–9,22], 63 datasets of undercooling are collected and screened with different substrate phases under nucleation phases such as β-Sn, BCC-Fe and FCC-Al (Table A1). The data are then divided into 50 training datasets and 13 testing datasets (the former for model construction and the latter for model validation). As features have a different effect on the target properties, the beneficial performance of the ML model heavily relies on the feature selection. Therefore, the reasonable selection of features is very important for establishing the model.

When selecting the features, substrate phases and nucleation phases are initially selected as two feature parameters involved in the establishment of the model. As substrate phases and nucleation phases are non-numerical data types that cannot be involved in the calculation, datasets have to be One-Hot encoded prior to establishing the model. In other words, the data are converted into zeroes and ones, i.e., the existing value is represented as a one while the non-existent value is represented as a zero. Following One-Hot encoding, the dimensionality of the features is changed. By considering the current data volume as a small sample of data, a change in the feature dimensionality is disadvantageous to the establishment of the ML model. During the model validation, the ML model did not perform well in choosing the former two features. Hence, other features are considered.

According to the literature survey, six feature variables that affect the undercooling are selected from a total of nine features (such as the substrate phase, nucleation phase, mismatch and lattice number) including the cooling rate (CR), the mean covalent atomic radius (MAR) [23], the number of lattices (NL), the mismatch (MM) [6], the mean Mendeleev number (MMN) and the nucleation and substrate plane (NSP). Here, the predicted target value is the undercooling. The MAR mean value is the average value of the mean atomic covalent nucleation radius and a substrate plane, which is used to express the properties of nucleation and the substrate phase. The NL mean value is the product of a substrate and the nucleation phase lattice constants. For a dense hexagonal structure, the NL is the value of a/c. The MMN mean value is the Mendeleev number mean value of nucleation and the substrate phase, which indicates the chemical properties of the nucleus phase and the base phase. The NSP mean value is the crystallographic representation of the mismatch between the nucleation and the substrate phase, which reveals the effect of different orientations of the crystallographic plane on the undercooling. For example, if the selected crystallographic surface of the nucleation phase is 111 and the selected crystallographic surface of the base phase is 100, then the NSP value is equal to 11,100. Furthermore, if the selected crystallographic surface of the nucleation phase is $\bar{1}00$ and the crystallographic surface of the base phase is 110, then the NSP value is equal to −1,100,110. Here, the negative sign indicates that $\bar{1}$ is present and the first digit indicates that several different types of numbers are present while the following digit indicates the position of those numbers. An information summary of the dataset for a simple statistical analysis is presented in Table 1.

Table 1. Statistical analysis of the dataset.

Quantity	Mean	Minimum	Maximum	Standard Deviation
MM	10.32	6.94	33.20	1.11
CR	119.75	313.84	1000.00	0.25
NL	13.22	12.59	54.44	1.07
MAR	110.92	6.19	138.70	103.50
MMN	21.18	15.25	65.25	11.20
NSP	-	-	-	-
UR	33.77	38.96	144.00	1.70

2.2. Computational Methods

2.2.1. Normalization Processing

When predicting material undercooling, if a large difference in magnitude between features is encountered, the ML model is affected. Therefore, the feature data have to be normalized, i.e., feature scaling has to be conducted. In this paper, z-score normalization is employed where datasets are modified by a mean of zero and a variance of one, as shown in Equation (1). This not only eliminates the impact of inconsistent data magnitude on ML but also ensures that data maintain the original distribution compared with the magnitudes of different features. It is worth noting that this treatment causes the loss of the meaning of the original data. However, it is beneficial for the establishment of ML models.

$$y_i = \frac{x_i - \overline{x}}{\sigma} \tag{1}$$

where x_i is the original data, $\overline{x}$ is the mean of the original data and σ is the standard deviation of the original data.

2.2.2. Correlation Analysis and Machine Learning

In the ML algorithm, a low correlation between features should be ensured. In this paper, the Pearson correlation coefficient r is used to observe the correlation between features [24], as shown in Equation (2). Here we use the original data for the correlation analysis; the correlation coefficient ranges between the values of –1 and 1. The closer the absolute value of the coefficient is to 1, the stronger the correlation between the two variables is. When the coefficient is equal to 0, the two variables are not correlated. In this paper, when the absolute value of the coefficient between features is greater than 0.95 [25], its correlation is considered to be high. Consequently, the feature should be removed.

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{2}$$

where X_i and Y_i are the values of the two undercooling degree factors and $\overline{X}$ and $\overline{Y}$ are the average values of these two factors.

For these datasets, nine ML models are used for the undercooling prediction. These models are the RF model [26], the gradient boosting regressor (GBDT) [27], TREE [28], XGBOOST [29], RIDGE [30], the Bayesian Ridge (BR) [31], k-nearest-neighbor (KNN) [32], the least absolute shrinkage and selection operator (LASSO) [33] and the support vector machine (SVM (kernel = linear)) [34] model. The variations between different ML models in the predicted results are compared to select the most suitable models for datasets and further analysis. Top-ranked ML models are selected and validated using the k-fold (with k = 5) cross-validation [35]. This method randomly divides the input datasets into k groups of equal size. These datasets are used for the ML model training while the remaining groups are denoted as the testing data. When evaluating different ML models, certain parameters are employed to indicate the strengths and weaknesses of the ML models. Model adjustment based on the feedback from the evaluation metrics is a key

parameter in model evaluation. In this paper, the mean absolute error (MAE) is employed, which demonstrates improved reflections of the actual prediction error. The root mean squared error (RMSE) is also employed, which is used to measure the deviation between the predicted and the actual value. Furthermore, the RMSE can eliminate the effect of different magnitudes between features. The average square of the Pearson product-moment correlation coefficient (R^2) is used as a generalized performance evaluation parameter controlling the goodness of fit of the ML model, as shown in Equations (3)–(5). The optimal combination of features is considered after selecting the best model by comparing the evaluation metrics. In this paper, the scikit-learn package [36] is used to process the datasets and establish the ML models.

$$E_{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_i\right| \tag{3}$$

$$E_{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\hat{y}_i - y_j\right)^2} \tag{4}$$

$$R^2 = 1 - \frac{\sum_{j=0}^{n-1}\left(y_j - \hat{y}_i\right)^2}{\sum_{j=0}^{n-1}\left(y_j - \overline{y_j}\right)^2} \tag{5}$$

where y_i are the actual values and $\hat{y}_i$ are the predicted values.

3. Results and Discussion

3.1. Correlation Analysis and Algorithm Selection

All of the features as well as the predicted values are briefly described in Table 1. In Figure 2, the correlation values between the features are shown. The color between the CR and the target feature undercooling (UR) is yellow to indicate that the CR had a high influence on the target feature. The remaining features demonstrated a low correlation. Therefore, they were retained.

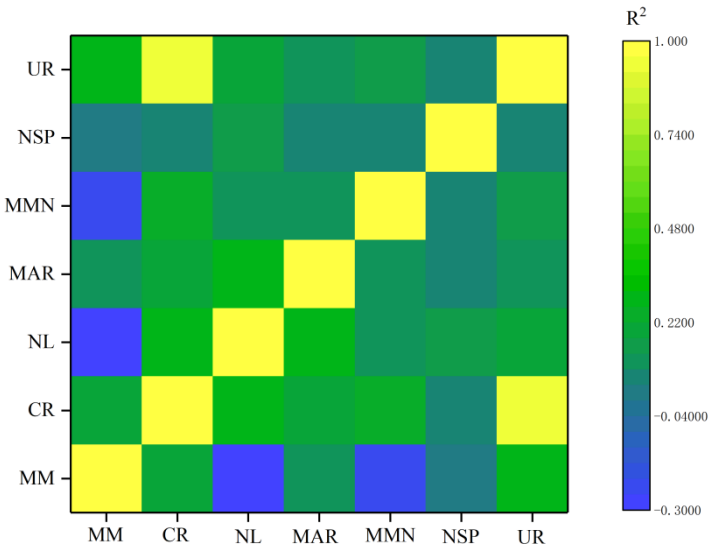


Figure 2. Correlation coefficient between the features.

Different ML models have different predictive capabilities. Due to the complexity of the datasets and material properties, researchers usually do not specify which ML algorithm is the most suitable. In addition, while the predicted values for specific attributes heavily depend on the ML algorithm selection, it is necessary to evaluate the performance and output of the chosen algorithm to assess the degree of uncertainty arising from its

choice. An R^2 comparison of nine ML models is presented in Figure 3. It can be seen that three of the top four ML models were integrated models. The RF model showed the best performance with a value of 0.831. There was a minor difference between the other three ML models. RIDGE was the next model performing relatively well while the worst model was the SVM (kernel = linear) with a value of 0.34. In summary, the prediction results varied significantly with respect to different ML models. Thus, it was necessary to carefully select ML models. In order to further analyze the models, the top four ML models were selected. In addition, the RIDGE model was substituted in place of the TREE model due to its overfitting problem. To summarize, the RF, XGBOOST, GBDT and RIDGE models were used to analyze the material undercooling. The results showed that when ML models were employed to predict the material properties, various ML models performed inconsistently for the same datasets, which was in accordance with [37–39].

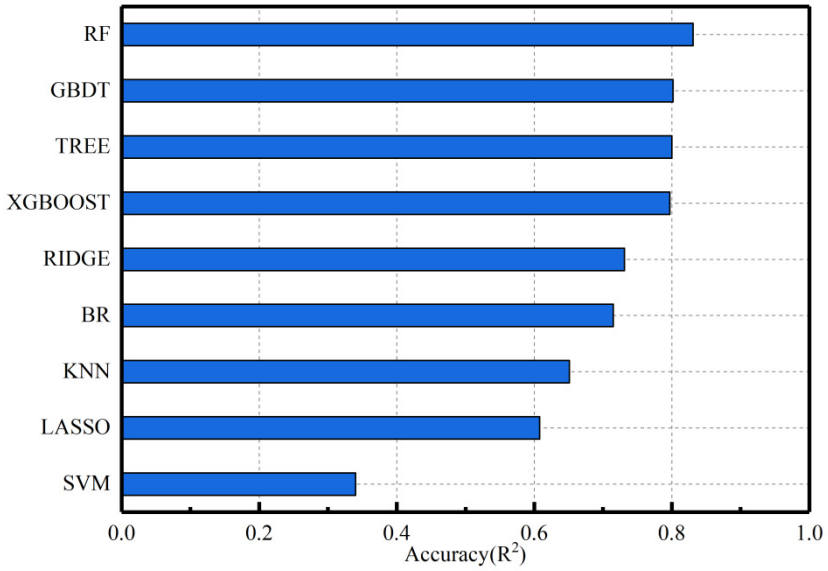


Figure 3. The R^2 prediction results of nine different machine learning algorithms.

In Figure 3, the results of a single training are displayed. In Figure 4, a comparison of R^2 values under four different ML algorithms trained for 10 times is shown. It displays that each algorithm fluctuated with different training times. In the second model training, the four algorithms showed relatively low R^2 values compared with other training times. More specifically, the R^2 value of the RIDGE algorithm was equal to 0.672. This was due to the random selection of training and testing datasets. When the selected training dataset was good, its R^2 value increased and vice-versa. This is further explained below. In addition, the prediction results of four ML models were almost equal and the R^2 of the remaining training results was close to 0.8. This was a relatively good result especially considering the R^2 value of the GBDT algorithm, which had the maximum value of 0.971 at the 9th training. This indicated that the selection of this training dataset was very representative. In conclusion, when studying the predictive ability of the ML model, the dataset selection should be considered because different training and testing sets lead to model differences. By considering the dataset selection, model differences were reduced and the generalization ability of the datasets was improved.

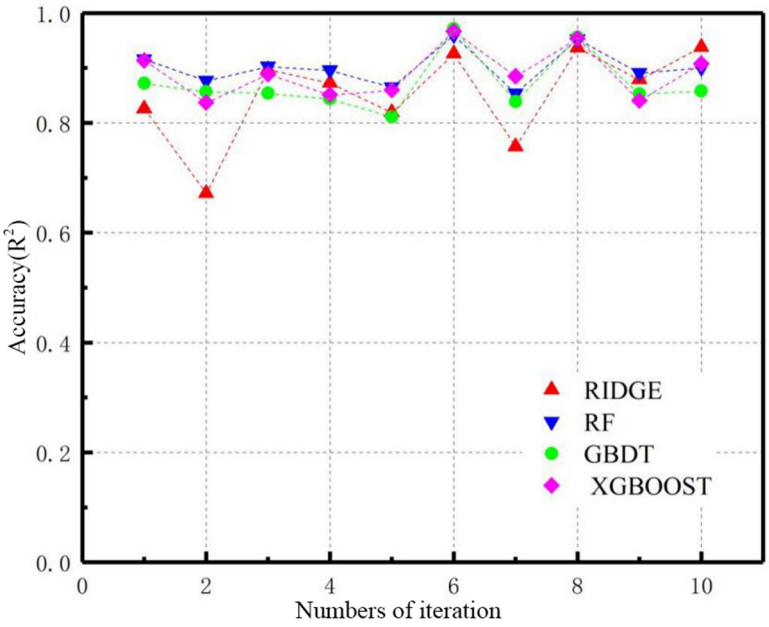


Figure 4. The R^2 prediction results of four different machine learning models in ten runs.

To further illustrate the performance capability of different algorithms for the same datasets, the strengths and weaknesses of each model were evaluated from the MAE and the RMSE. In Table 2, the average training and testing set evaluation results of different ML models according to Figure 4 are listed; each model was tuned up to 10 iterations. The R^2 values of four ML models were similar with all of them being above 0.85. The RF model had the highest R^2 value of 0.902 and its RMSE was also the lowest among the four ML models. However, the MAE was not the lowest of the four algorithms. This indicated that the gap between the predicted and the true values of the remaining models was greater, thus leading to a greater value of the RMSE. This meant that the gap between the predicted and the true values of 10 training RF results was smaller. In the training data, the performance of the GBDT and XGBOOST were better than the RF and R^2 was close to 1 so there might have been overfitting, making the evaluation standard of the testing data lower than the RF. In summary, the R^2, the MAE and the RMSE values of the RIDGE model were relatively poor performers among the four ML models, which also indicated that the integrated algorithm had an improved prediction ability when dealing with the datasets.

Table 2. Comparison of train and test sample results in ten runs.

Model	MAE		RMSE		R^2	
	Train	Test	Train	Test	Train	Test
RIDGE	9.824	12.635	12.735	15.434	0.882	0.852
RF	3.457	10.137	4.962	12.926	0.982	0.902
GBDT	1.388	10.723	2.286	14.632	0.996	0.871
XGBOOST	0.328	9.23	1.577	13.551	0.997	0.891

3.2. Cross-Validation

Different algorithms resulted in different prediction results due to statistical dataset features being evaluated from the sub-datasets. This sometimes might not be representative of the entire datasets and it might lead to sampling uncertainly. In this paper, a five-fold cross-validation was employed, which randomly selected both the training and the testing set. However, unlike the previous selection methods, the selection of the training and testing set ensured that all samples could serve as either the training set or the testing set. In order to reduce the sampling uncertainty, various training models were iterated several

times to broaden the distribution of the validation subsets for a given set of parameters. In Table 3, the R^2 results for the cross-validation of four ML models are listed. It divides the data into five pieces, i.e., fold 1–fold 5. Each fold from fold 1 to fold 5 was then successively used as validation data while others were used as training data. As the selected data are different each time, the R^2 of the different ML models fluctuated with training. The cross-validation R^2 values of the four algorithms fluctuated widely from 0.28 to 0.98. Among them, the R^2 prediction of the RF in the five-fold cross-validation fluctuated from 0.61 to 0.98. Its fluctuation range was relatively small, which indicated that the RF algorithm was more stable in coping with the datasets.

Table 3. Cross-validation of test sample results.

		Five-Fold Cross-Validation				
Model	**Run**	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**
RIDGE	1	0.545	0.255	0.831	0.885	0.868
	2	0.864	0.536	0.402	0.891	0.435
RF	1	0.917	0.976	0.611	0.935	0.765
	2	0.729	0.828	0.922	0.913	0.933
GBDT	1	0.894	0.967	0.786	0.419	0.924
	2	0.830	0.849	0.856	0.308	0.861
XGBOOST	1	0.893	0.914	0.942	0.329	0.865
	2	0.800	0.831	0.634	0.885	0.880

The mean R^2 value for two five-fold cross-validations of four ML algorithms is shown in Figure 5. Through cross-validation, it could be concluded that the RF algorithm had the best prediction result for the datasets, which also verified the previous conclusion and further showed that the generalization ability of the RF was relatively strong. The RF had the highest R^2 value in four cross-validated models (0.865). This was a better result compared with the previously obtained R^2 mean value of 0.848. This was because the dataset selection was improved and could better represent the entire dataset features. To summarize, the cross-validation results indicated that different algorithms responded to different attributes of datasets with various stability effects. Furthermore, the RF had a relatively high stability for the dataset selection.

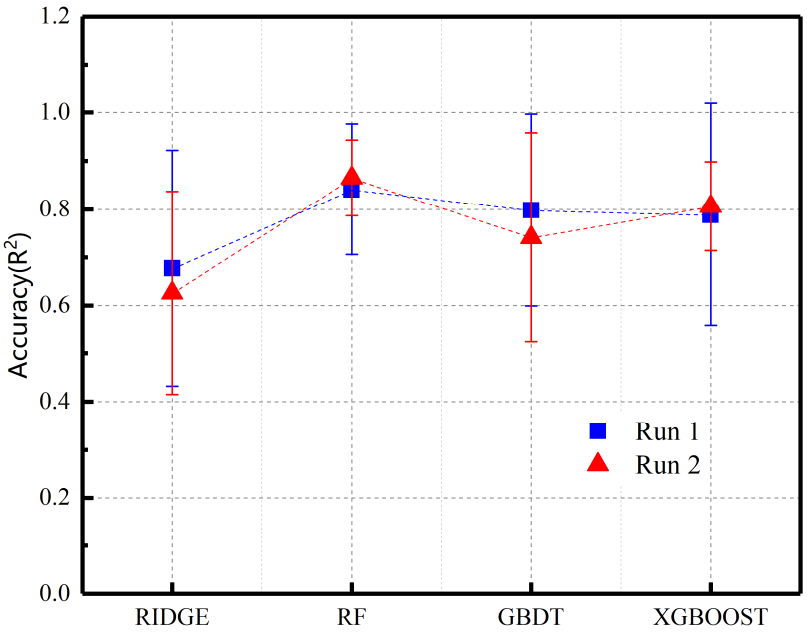


Figure 5. Cross-validation results under different machine learning in two runs.

In order to more intuitively illustrate the advantages and disadvantages of ML models, RF and XGBOOST were selected from the four ML models for further analysis. The prediction results of both models were compared. In Figure 6, the prediction results of R^2 and the MAE under the RF and XGBOOST are shown. Interestingly, many data points were perfectly organized in the diagonal for the training data (Figure 6a) in the XGBOOST model. This meant that, for these data, the XGBOOST was much better (almost 100% fit) than the RF. However, an ML model with 100% fit for a large fraction of the data may be caused purely by overlearning, which was confirmed by testing data (Figure 6b). In addition, we noted that most data via XGBOOST fitted well in the training set but others did not. The reason was that XGBOOST might have misjudged them because several pieces of data were so similar with several similar features after checking the raw data corresponding with the deviation points. In conclusion, this indicated that their prediction results were not significantly different when compared with the actual values while the results of XGBOOST were somewhat worse. This indicated that the RF had a stronger performance capability.

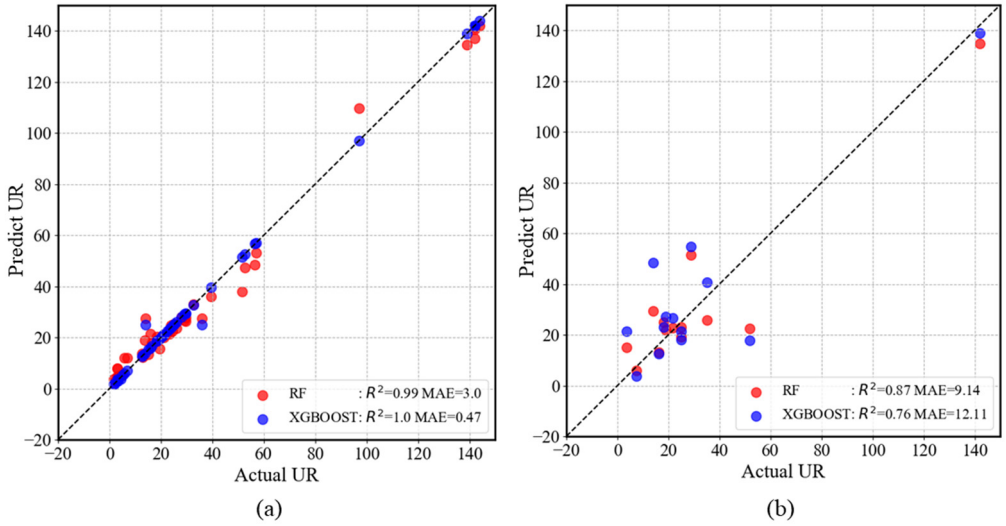


Figure 6. Prediction results under RF and XGBOOST ML models, (**a**) training data (**b**) testing data.

3.3. Combination of Features

The selection of different features had a significant effect on the ML model. In this paper, the performance capability of different ML models for the same datasets is discussed. In order to further discuss the influence of different features on establishing the ML model, the RF model with the optimal performance was used to analyze feature combinations with the purpose of obtaining the feature with the greatest influence on the model establishment. In Figure 7, the importance of six different features is shown. The ranking was done via the RF based on the Gini coefficient approach. It was known that the CR had the greatest influence on undercooling, which was consistent with materials science knowledge. This was followed by the MAR and the MM while the least influential one was the NSP.

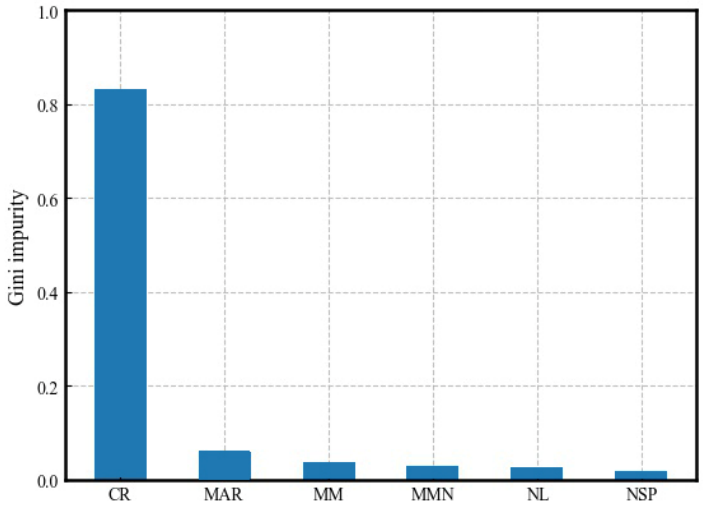


Figure 7. Ranking the importance of features.

In Figure 8, the degree to which the selection of features affected the RF model is shown. It demonstrated that 2–6 combinations of features from high to low (according to Figure 4) were required to predict the undercooling by taking the average value after 100 training sessions. With an increase in the number of features, R^2 demonstrated a slight decline while the MAE increased. In other words, the selection of the top two features could improve the representation of the dataset features. Furthermore, the increased feature values did not significantly affect the accuracy of the model. To summarize, the selection of features had a significant impact on the establishment of the model. However, this was based on the selection of beneficial features while inferior features only increased the workload. The improvement of the model was smaller, which decreased the predictive power of the model. This, in turn, showed that the number of features should not be selected randomly but rather appropriately.

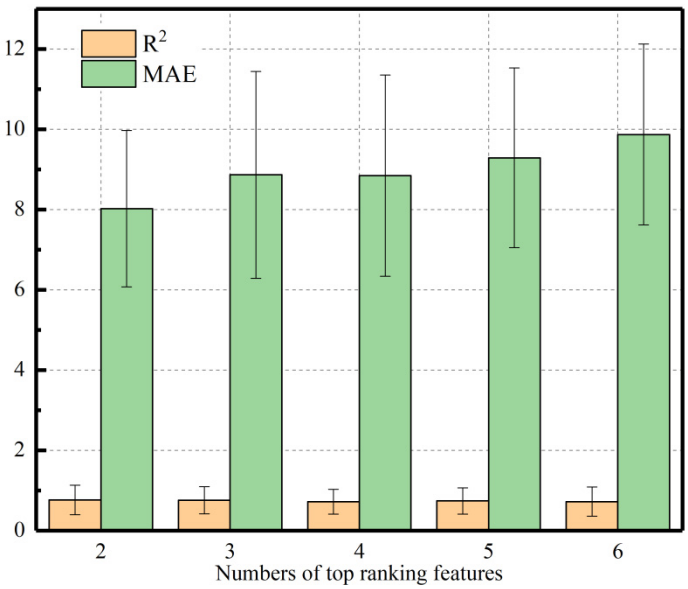
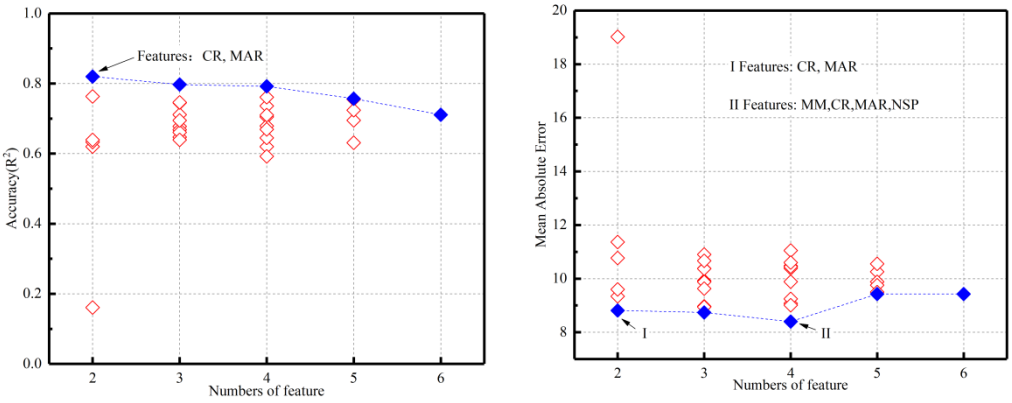


Figure 8. The R^2 and MAE results of ranking feature combinations in 100 training sessions.

The above presented feature combinations were only sequential superimposed combinations of the feature importance ranking. It was not possible to discern how the remaining feature combinations affected the prediction results. Thus, a further analysis was required. In Figure 9a, $R^2 > 0$ results for 100 runs under 2–6 feature combinations in the RF algorithm are shown. Under multiple combinations of 2–6 major factors, the optimal R^2 value for

each of its feature combinations fluctuated between 0.7 and 0.85 with the highest R^2 being 0.82 under two feature combinations represented by the CR and the MAR. This was similar to the importance ranking results provided in Figure 7. It could be seen that the R^2 of the testing set decreased with an increase in the number of features, which might produce overfitting and a loss of the generalization ability of the model. The results indicated that the number of features should favor quality over quantity. In Figure 9b, the results of the MAE run for 100 times under 2–6 combinations of features in the RF algorithm are shown. Point I indicates that under the combination of the CR and the MAR features, the MAE was equal to 8.813. Point II indicates that the value of the MAE under the combination of the MM, CR, MAR, NSP and MAE features was 8.397. Its value was lower than the value of point I. However, R^2 was equal to 0.792. The features of the CR and the MAR had favorable performances regarding R^2 values. Thus, they had a significant influence on the undercooling and could be considered as the key factors.



(**a**) (**b**)

Figure 9. (**a**) $R^2 > 0$ results of different RF algorithm feature combinations with 100 trainings and (**b**) the MAE results of different RF algorithm feature combinations with 100 trainings.

Based on the RF model, the CR and the MAR were the most important factors for the undercooling degree related to processing and the properties of both nucleation and the substrate phase, respectively. Obviously, the higher the CR, the larger the undercooling degree because the melt metal cannot nucleate on time and thus keeps liquid far below the solidification temperature when following the rapid change of temperature. The MAR was used to describe the atomic information of both the nucleation and the substrate phase. It expressed lattice distortion in a solid solution and consequently electron means free path [23], which is supposed to affect nucleation and thus the undercooling degree.

In summary, after comparing the evaluation indexes under different ML models, it could be concluded that the RF had the best performance ability in the material undercooling prediction. The CR had the greatest influence on the prediction results when the correlation analysis was performed on the features. The importance of the CR and the MAR was also demonstrated in the features ranking using the RF models, which further increased the reliability of its feature selection. The results obtained in this study can serve as a beneficial reference for obtaining key undercooling factors.

4. Conclusions

This study developed ML models for the prediction of the undercooling degree of as-cast irons and aluminum alloys. Here, 63 datasets with six features were collected and standardized from the experimental results. Furthermore, nine ML algorithms were used to mine the datasets. Four models were selected for a detailed analysis of nine ML models. It was found that differences in algorithm, features and data have a significant influence on the performance of the ML models. After comparing the evaluation indexes,

the RF model was considered to be the optimal model for the accurate prediction of the undercooling degree with the corresponding R^2 value of 0.85 and an MAE of 8.43. Various factors affected the undercooling degree differently according to their importance in the following sequence: cooling rate (CR), mean covalent atomic radius (MAR), mismatch (MM), mean Mendeleev number (MMN), number of lattices (NLs) and the nucleation and substrate plane (NSP). Two key features, the cooling rate (CR) and the mean covalent atomic radius (MAR), were selected as an optimal combination after comparing all possible combinations among six features and were enough to build the ML model for the prediction of the undercooling degree. In this study, the ML model based on the RF algorithm could accurately predict the undercooling degree for as-cast iron materials and aluminum alloys, which has a potential application in both industrial and experimental areas.

Author Contributions: Y.C. conceived the total investigation; L.W. analyzed the data and wrote the manuscript; S.W., C.Y. and N.Z. helped data collections and discussed the results; Z.Z. designed the whole work and revised the manuscript; K.Z. supervised the whole work. All authors have read and agreed to the published version of the manuscript.

Funding: Simulations were performed at the Guangdong Province Engineering Research Center of Metal Matrix Composite Database. We acknowledge financial support by the Guangdong Academy of Sciences through project of [2020GDASYL-20200302017] and [2020GDASLY-20200103141]. This research was also funded by Guangdong Province Key Area R and D Program (2019B010940001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data, models or code generated or used during the study are available from the corresponding author by request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The original data used in this study.

Table A1. The original data used in this study.

MM	CR	NL	MAR	MMN	NSP	UR
3.57	1.32	12.62	109.75	20.25	100,100	1.7
5.32	1.32	12.83	115.75	20	100,100	1.8
1.2	0.25	33.38	108.95	11.5	100,100	2.97
4.49	0.5	1.88	109.5	27	111,001	3
5	0.33	1.88	109.5	27	111,001	3
3.9	1.32	12.63	109.75	20.25	100,100	3.1
5.9	1.32	12.87	115.75	20	100,100	3.3
7.77	20	1.50	107.1	11.2	100,223	3.5
1.2	0.25	33.38	108.95	11.5	110,110	4.4
1.21	0.25	33.38	108.95	11.5	111,111	5.03
9.45	20	1.50	107.1	11.2	100,100	5.8
11.42	1.32	13.58	114.25	49.5	100,100	7
6	1.32	12.89	105.75	22.35	100,100	7.5
11.2	1.32	13.51	114.25	49.5	100,100	12.6
3.9	0.33	12.63	109.75	20.25	100,100	13
14.25	1.32	13.92	115.25	24.5	100,100	13.6
8.04	0.33	1.07	107.1	11.2	111,001	13.9
16.1	0.33	1.07	107.1	11.2	111,001	13.9
3.57	0.33	12.63	109.75	20.25	100,100	15.2
12.49	20	1.50	107.1	11.2	−1,511,112	16

Table A1. *Cont.*

MM	CR	NL	MAR	MMN	NSP	UR
3.46	0.25	1.50	107.1	11.2	100,110	16.2
8.4	0.25	1.50	107.1	11.2	−15,110,112	16.6
16.01	20	33.38	108.95	11.5	311,110	18
20.67	20	1.50	107.1	11.2	111,100	18.5
8.02	0.25	1.50	113.5	11.5	110,001	18.5
25.03	20	1.50	107.1	11.2	311,110	19
4.49	0.33	1.88	109.5	27	111,001	19.4
1.11	20	33.38	108.95	11.5	100,100	20.5
1.11	20	33.38	108.95	11.5	110,110	20.5
1.11	20	33.38	108.95	11.5	111,111	20.5
10	20	17.09	108.95	11.5	111,110	21
6.2	20	33.38	113.5	11.5	311,111	21
23.88	20	1.50	113.5	11.5	111,110	21.7
21.42	20	1.50	107.1	11.2	100,312	22.4
3.14	20	1.50	113.5	11.5	111,111	23.2
14.49	20	17.09	113.5	11.5	311,110	24
14.4	1.32	13.91	115.25	24.5	100,100	24.5
10.83	20	17.09	108.95	11.5	111,100	25
12.49	20	33.38	113.5	11.5	111,100	25
17.1	20	1.50	107.1	11.2	111,113	25.1
16.36	20	1.50	107.1	11.2	100,001	25.2
3.12	20	17.09	113.5	11.5	111,111	26
13.87	20	17.09	113.5	11.5	110,111	28
6.61	15	15.41	104.1	19.8	111,003	28.7
12.7	1.32	8.59	111	33	110,001	29
14.02	0.33	20.96	138.7	15	111,111	29.5
22	0.33	20.97	138.7	15	111,111	29.5
12.77	20	1.50	107.1	11.2	100,101	32.7
8.04	0.33	1.07	109.75	17.7	111,001	35
16.1	0.33	1.07	107.1	11.2	111,001	36
12.9	20	1.50	107.1	11.2	111,104	39.6
8.04	15	1.07	107.1	11.2	111,004	51.6
3.58	15	12.35	108.5	18	100,100	51.8
9.49	20	13.34	103.5	21	111,111	52.7
6.61	0.33	15.41	104.1	19.8	111,002	56.6
26	0.33	1.11	104.1	19.8	111,001	57

References

1. Olakanmi, E.O.; Cochrane, R.F.; Dalgarno, K.W. A review on selective laser sintering/melting (SLS/SLM) of aluminium alloy powders: Processing, microstructure, and properties. *Prog. Mater. Sci.* **2015**, *74*, 401–477. [CrossRef]
2. Flower, H.M. Light alloys: Metallurgy of the light metals. *Int. Mater. Rev.* **1992**, *37*, 196. [CrossRef]
3. Xu, C.L.; Jiang, Q.C. Morphologies of primary silicon in hypereutectic Al–Si alloys with melt overheating temperature and cooling rate. *Mater. Sci. Eng. A* **2006**, *437*, 451–455. [CrossRef]
4. Vijeesh, V.; Prabhu, K.N. Review of Microstructure Evolution in Hypereutectic Al–Si Alloys and its Effect on Wear Properties. *Trans. Indian Inst. Met.* **2014**, *67*, 1–18.
5. Xu, Y.; Deng, Y.; Casari, D.; Mathiesen, R.; Liu, X.; Li, Y. Growth kinetics of primary Si particles in hypereutectic Al-Si alloys under the influence of P inoculation: Experiments and modelling. *J. Alloys Compd.* **2021**, *854*, 155323. [CrossRef]
6. Bramfitt, B.L. The effect of carbide and nitride additions on the heterogeneous nucleation behavior of liquid iron. *Metall. Trans.* **1970**, *1*, 1987–1995. [CrossRef]
7. Wang, L.; Yang, L.; Zhang, D.; Xia, M.; Wang, Y.; Li, J.G. The Role of Lattice Misfit on Heterogeneous Nucleation of Pure Aluminum. *Metall. Mater. Trans. A* **2016**, *47*, 5012–5022. [CrossRef]
8. Perepezko, J.; Uttormark, M. Undercooling and Nucleation during Solidification. *ISIJ Int.* **1995**, *35*, 580–588. [CrossRef]
9. Ohashi, T.; Hiromoto, T.; Fujii, H.; Nuri, Y.; Asano, K. Effect of Oxides on Nucleation Behaviour in Supercooled Iron. *Tetsu Hagane* **1976**, *62*, 614–623. [CrossRef]
10. Mueller, B.A.; Perepezko, J.H. The undercooling of aluminum. *Metall. Mater. Trans. A* **1987**, *18*, 1143–1150. [CrossRef]

11. Kalb, J.A.; Spaepen, F.; Wuttig, M. Kinetics of crystal nucleation in undercooled droplets of Sb- and Te-based alloys used for phase change recording. *J. Appl. Phys.* **2005**, *98*, 054910. [CrossRef]
12. Kelton, K.F.; Lee, G.W.; Gangopadhyay, A.K. First X-Ray Scattering Studies on Electrostatically Levitated Metallic Liquids: Demonstrated Influence of Local Icosahedral Order on the Nucleation Barrier. *Phys. Rev. Lett.* **2003**, *90*, 195504. [CrossRef] [PubMed]
13. Sang, U.; Yang, M. Nucleation modes of the drop tube processed $Nd_{70}Fe_{20}Al_{10}$ droplets. *Mater. Lett.* **2004**, *58*, 975–979.
14. Battersby, S.E.; Cochrane, R.F.; Mullis, A.M. Growth velocity-undercooling relationships and microstructural evolution in undercooled Ge and dilute Ge-Fe alloys. *J. Mater. Sci.* **1999**, *34*, 2049–2056. [CrossRef]
15. Jian, Z.; Kuribayashi, K.; Jie, W. Critical undercoolings for the transition from the lateral to continuous growth in undercooled silicon and germanium. *Acta Mater.* **2004**, *52*, 3323–3333. [CrossRef]
16. Li, D.; Herlach, D.M. High undercooling of bulk molten silicon by containerless processing. *EPL* **2007**, *34*, 423. [CrossRef]
17. Li, J.F.; Jie, W.Q.; Yang, G.C. Solidification structure formation in undercooled Fe–Ni alloy. *Acta Mater.* **2002**, *50*, 1797–1807. [CrossRef]
18. Ankit Agrawal, P.; Ahmet Cecen. Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. *Integr. Mater. Manuf. Innov.* **2014**, *3*, 1–19. [CrossRef]
19. Jiang, X.; Yin, H.-Q.; Zhang, C. An materials informatics approach to Ni-based single crystal superalloys lattice misfit prediction. *Comput. Mater. Sci.* **2018**, *143*, 295–300. [CrossRef]
20. Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J.E.; Doak, J.W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **2014**, *89*, 094104. [CrossRef]
21. Javed, S.G.; Khan, A.; Majid, A. Lattice constant prediction of orthorhombic ABO3 perovskites using support vector machines. *Comput. Mater. Sci.* **2007**, *39*, 627–634. [CrossRef]
22. Nakajima, K.; Hasegawa, H.; Khumkoa, S. Effect of a catalyst on heterogeneous nucleation in pure and Fe-Ni alloys. *Metall. Mater. Trans. B* **2003**, *34*, 539–547. [CrossRef]
23. Hong, Z.; Hua, F.; Xing, H.; Chang, W.; Lei, J.; Long, C.; Jian, X. Dramatically Enhanced Combination of Ultimate Tensile Strength and Electric Conductivity of Alloys via Machine Learning Screening. *Acta Mater.* **2020**, *200*, 803–810.
24. Pearson, K. Note on Regression and Inheritance in the Case of Two Parents. *Proc. R. Soc. Lond.* **1895**, *58*, 240–242.
25. Yuan, R.; Liu, Z.; Balachandran, P.V. Accelerated Discovery of Large Electrostrains in $BaTiO_3$-Based Piezoelectrics Using Active Learning. *Adv. Mater.* **2018**, *30*, 1702884. [CrossRef]
26. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
27. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
28. Quinlan, J.R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
29. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
30. Hoerl, A.E.; Kannard, R.W.; Baldwin, K.F. Ridge regression: Some simulations. *Commun. Stat.* **1975**, *4*, 105–123. [CrossRef]
31. Tipping, M.E. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244.
32. Altman, N.S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175–185.
33. Tibshirani, R.J. Regression Shrinkage and Selection via the LASSO. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *73*, 273–282. [CrossRef]
34. Cortes, C.; Vapnik, V.N. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
35. An, S.; Liu, W.; Venkatesh, S. Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression. *Pattern Recognit.* **2007**, *40*, 2154–2162. [CrossRef]
36. Pedregosa, F.; Varoquaux, G.l.; Gramfort, A. Scikit-learn: Machine Learning in Python. *Comput. Sci.* **2012**, *12*, 2825–2830.
37. Kauwe, S.K.; Graser, J.; Vazquez, A. Machine Learning Prediction of Heat Capacity for Solid Inorganics. *Integr. Mater. Manuf. Innov.* **2018**, *7*, 43–51. [CrossRef]
38. Peng, J.; Yamamoto, Y.; Brady, M.P. Uncertainty Quantification of Machine Learning Predicted Creep Property of Alumina-Forming Austenitic Alloys. *JOM* **2021**, *73*, 164–173. [CrossRef]
39. Sun, W.; Zheng, Y.; Yang, K. Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Sci. Adv.* **2019**, *5*, 4275. [CrossRef] [PubMed]