Machine Learning-Based Prediction and Composition Optimization of Tensile Strength in Cast Titanium Alloys

Feng Li [1,2], Wangneng Liao[1], Xuejun LIU[1], Hai Nan[2]

1. Nanjing University of Aeronautics and Astronautics，Nanjing，211106, China ;
2. AECC Beijing Institute of Aeronautical Materials，Beijing，100095, China.

*Corresponding address: AECC Beijing Institute of Aeronautical Materials，Beijing，100095, China. e-mail: lifengmaterial@nuaa.edu.cn

Abstract: This study presents a machine learning framework for predicting tensile strength and optimizing compositions of cast titanium alloys using a dataset of 13,210 industrial production records. Sixteen machine learning models were systematically evaluated, with ensemble methods demonstrating superior performance. The ExtraTrees Regression model achieved optimal prediction accuracy (R^2=0.895, RMSE=20.1 MPa), effectively capturing nonlinear relationships between alloy compositions and mechanical properties. Feature importance analysis identified aluminum (Al) as the dominant strengthening element (69.9% contribution), exhibiting a nonlinear enhancement effect where tensile strength increased by 10 MPa per 1 wt.% Al. A genetic algorithm-based optimization strategy enabled the design of novel high-strength alloys exceeding 1,000 MPa tensile strength, characterized by balanced compositions of 6.46 wt.% Al, 4.14 wt.% Zr, and 2.8 wt.% Mo. The work innovatively integrates industrial big data with ensemble learning, addressing limitations of traditional trial-and-error approaches in alloy development. This data-driven paradigm provides a scalable solution for accelerating the discovery of advanced cast titanium alloys, particularly for aerospace applications requiring high-strength cast titanium alloy.

Keywords: Casting; Titanium alloys; Tensile strength prediction; Machine learning; Ensemble learning; Composition optimization; Feature importance analysis.

1 Introduction

Titanium alloys are extensively employed in aerospace, biomedical, and energy sectors owing to their exceptional strength-to-weight ratio, low density,and superior corrosion resistance[1-3]. Titanium alloys are classified into four primary categories based on their crystalline structures: α-titanium alloys, near-α titanium alloys, α+β titanium alloys, and β-titanium alloys[4]. α-Titanium alloys, equivalent to commercially pure titanium (CP-Ti), exhibit moderate tensile strength ranging from 170 MPa to 480 MPa[5]. Near-α titanium alloys, classified as high-temperature titanium alloys, are composed primarily of aluminum (Al), tin (Sn), and zirconium (Zr), with β-stabilizing elements (e.g., Mo, Nb, Si) limited to ≤2 wt.%. These alloys exhibit service temperatures of approximately 500°C–600°C[6]. A representative example, Ti-6Al-2Sn-4Zr-2Mo-0.1Si, demonstrates a tensile strength of 990 MPa[7].

α+β titanium alloys, the most extensively utilized titanium alloys, contain 4–6 wt.% β-stabilizing elements. The widely adopted alloy Ti-6Al-4V exemplifies this category, demonstrating a tensile strength of approximately 900 MPa [8]. β-titanium alloys, recognized as the highest-strength titanium alloys, retain 100% metastable β-phase when quenched from above the β-transus temperature. A representative alloy, Ti-5Al-5Mo-3Cr (designated as Ti-5553), achieves a tensile strength of 1,159 MPa [9].

Casting processes enable direct fabrication of complex thin-walled structures (e.g., large-scale thin-walled intermediate casings for aero-engines, vertical tail components) that are challenging to achieve via traditional forging or machining. However, as aerospace technology advances toward higher thrust-to-weight ratios, extended service lifetimes, and enhanced extreme-environment adaptability, cast titanium alloys face unprecedented performance challenges. Breaking the specific strength limits of conventional cast titanium alloys has become a critical developmental focus.

Traditional approaches to establishing composition-process-performance relationships rely on empirical methods such as the molybdenum equivalent method[10] or computational models including phase diagram calculations, phase-field simulations, first-principles analyses, and multi-physics modeling[11]. These methods exhibit significant limitations in novel alloy development due to their heavy dependence on trial-and-error experimentation, prolonged cycles, and escalating costs — particularly when alloying elements increase, complicating the mapping between composition and mechanical properties

2 Experimental section

2.1 Data Acquisition and Preprocessing

The data were derived from a real-world experimental dataset collected through the Huazhu ERP system, encompassing historical records of casting titanium alloy production. The dataset comprises 13,210 sets of titanium alloy compositions and tensile strength values, with measurement samples obtained from attached test specimens during casting production. The alloy compositions include 11 elements: C (Carbon), O (Oxygen), N (Nitrogen), H (Hydrogen), Al (Aluminum), V (Vanadium), Fe (Iron), Si (Silicon), Y (Yttrium), Mo (Molybdenum), and Zr (Zirconium), as summarized in Table 1.

Tab. 1 Composition and Strength of Titanium Alloys

Composition	Mean	Min	Max	Percentage
C	0.011	0.001	0.096	100.00%
O	0.152	0.015	18.000	100.00%
N	0.021	0.001	22.000	100.00%
H	0.003	0	0.050	99.86%
Al	6.424	0	6.800	98.57%
V	3.578	0	5.100	98.13%
Fe	0.103	0.001	16.000	100.00%
Si	0.019	0	0.250	68.42%
Y	0.001	0	0.010	32.78%
Mo	0.527	0	5.220	28.16%
Zr	0.634	0	4.200	28.15%
Strength	918.7	391.5	1033.5	100.00%

The preprocessing workflow encompassed data cleaning, outlier detection, normalization, and dataset partitioning. The dataset was divided into a training set and a testing set at a ratio of 80% and 20%, respectively. Based on impurity content specifications (C, O, N, H) for cast titanium alloys, data exceeding threshold limits and statistically anomalous entries were removed. Specifically, one sample with oxygen content (18%), one with nitrogen content (22%), and three samples with iron content (14%-16%) were identified as outliers and subsequently eliminated.

The tensile strength distribution is presented in Fig. 1, showing a minimum value of 391.5 MPa, a maximum value of 1033.5 MPa, and an average value of 918.4 MPa. The distribution exhibited a distinct bimodal pattern characterized by two prominent peaks. The left peak in the low-strength range (453-541 MPa) accounted for 1.25% of the data, corresponding to pure titanium regions. The right peak dominated the medium-to-high strength range (823-1008 MPa), representing 98.1% of the total data. Minimal data points were observed in other intervals.

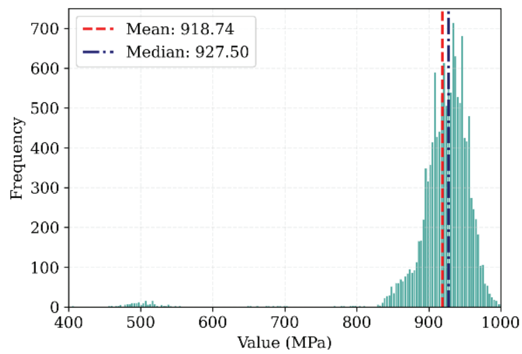

Fig. 1: Distribution of tensile strength

2.2 Correlation analysis

Correlation analysis is employed to identify the degree of association between a dependent variable and other variables [15]. As illustrated in the matrix, the lower triangular portion contains scatterplots within each cell, while the upper triangular region displays correlation coefficients (Spearman's rank coefficients) and color intensity. The correlation coefficients range from -1 to +1, with blue indicating negative correlations, red representing positive correlations, and intermediate grey denoting no significant correlation. This dual visualization method (numerical values + color gradient) enhances interpretability of both linear and non-linear relationships across variables.

The scatterplots illustrate the relationships between variables, while the correlation coefficients quantify the strength and direction of these associations. As shown in the Fig. 2, the alloying elements Al, Si, Mo, and Zr exhibit positive correlations with tensile strength, whereas V and Y display negative correlations with tensile strength. Notably, Mo and Zr demonstrate a strong positive correlation (r=0.96). Additionally, Si correlates positively with Zr (r=0.60) and Mo (r=0.61), while V shows strong negative correlations with Mo (r=-0.82) and Zr (r=-0.82).

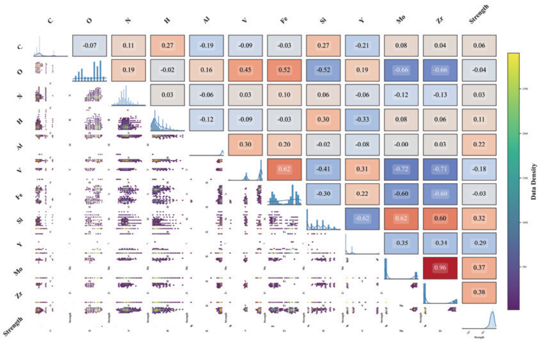

Fig. 2: Intervariable Correlation Analysis

This study formulates a typical regression problem where alloy composition serves as the input variable and tensile strength as the output. Leveraging the Python programming language and the scikit-learn machine learning library, we implemented sixteen distinct machine learning regression algorithms. These algorithms were systematically categorized into four principal classes based on their underlying methodologies: linear regression models(Linear, Ridge, Lasso, ElasticNet, BayesianRidge), tree-based models(DecisionTree, RandomForest , ExtraTrees),ensemble learning models(GradientBoost, XGBoost , AdaBoost , LightGBM , CatBoost)and other specialized models (SVR, KNN, MLP)[16].

2.3 Model evaluation metrics

Model evaluation metrics are used to assess the discrepancy between predicted and actual values. Through a unified evaluation algorithm, the performance of predictions from different models can be compared. Commonly used performance evaluation metrics for regression algorithms include the coefficient of determination (R^2), mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) [17]. The value of the coefficient of determination ranges from 0 to 1, with values closer to 1 indicating a better model fit. This evaluation metric is also well - interpretable. Both mean squared error and root mean squared error can be used to analyze the accuracy of predictions, but mean squared error is more susceptible to the influence of outliers. Mean absolute error cannot distinguish between positive and negative errors. Therefore, in this project, the coefficient of determination (R^2) and root mean squared error (RMSE) are chosen as the evaluation metrics for the model.

The calculation formula for the coefficient of determination (R^2) [18]:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (2.1)$$

where:

SSR denotes the regression sum of squares;

SST represents the total sum of squares;

SSE indicates the residual sum of squares.

The formula for Root Mean Square Error (RMSE) [19]:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (2.2)$$

where:

y_i denotes the actual observed value;

$\hat{y}_i$ represents the predicted value;

n indicates the total number of data points.

2.4 Feature importance analysis

In machine learning, feature importance refers to the metric that measures the degree of influence each feature has on the model's prediction results. It primarily quantifies the contribution of features to model performance, enabling the selection of key features to optimize model performance and enhance interpretability [20]. In this project, two calculation methods are employed: tree-based methods and Partial dependence plots methods. Partial dependence plots are a global interpretability tool used to explain the predictions of machine learning models. Their core objective is to illustrate the marginal impact of one or several features on model predictions.

3 Results and discussion

3.1 Model result analysis

The performance of the 16 models is ranked from highest to lowest as shown in Table 1.2. Ensemble models (especially tree-based models) dominate the top five, demonstrating significant advantages. Linear models and simple regularization methods (such as Ridge and Lasso) show moderate performance, while SVR and ElasticNet perform the worst. Extra_Trees achieves the best result with an R^2 of 0.895, RMSE 为 20.1. This indicates that the model can identify 89.4% of the regular variations in the data.

Tab. 2 Model Performance Comparison

Number	Models	R^2	RMSE
1	ExtraTrees	0.895	20.1
2	LightGBM	0.893	20.3
3	CatBoost	0.893	20.3
4	RandomForest	0.890	20.6
5	XGBoost	0.886	20.9
6	GradientBoosting	0.884	21.1
7	DecisionTree	0.868	22.5
8	MLP	0.859	23.3
9	KNN	0.854	23.6
10	LinearRegression	0.815	26.61
11	Ridge	0.815	26.61
12	BayesianRidge	0.815	26.61
13	Lasso	0.814	26.68
14	AdaBoost	0.760	30.30
15	ElasticNet	0.727	32.37
16	SVR	0.361	49.48

To visually assess model performance, we constructed a comparative plot where the x-axis represents true values and the y-axis represents predicted values. A diagonal reference line indicates the ideal scenario where predictions perfectly match actual values. Data points are color-coded according to their absolute error values, with a gradient color scheme ranging from purple (indicating smaller absolute errors) to yellow (representing larger absolute errors).

3.1.1 Linear Regression Algorithm

Linear regression is a statistical method for making predictions by identifying the linear relationship between independent variables (features) and the dependent variable (target) [21].

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where:

y is the dependent variable;

xi is the independent variable;

$\beta 1$ to βn are the regression coefficients.

The calculation results (Fig.3 -Fig. 7) show that the coefficient of determination and root mean squared error of the four regression algorithms are basically consistent, and the performance of the Elastic Net regression model is the worst. This indicates that there is a nonlinear relationship between the composition of titanium alloy and tensile strength.

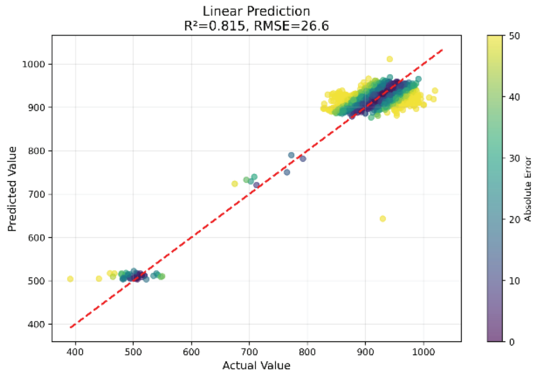

Fig. 3: Simple Linear Regression

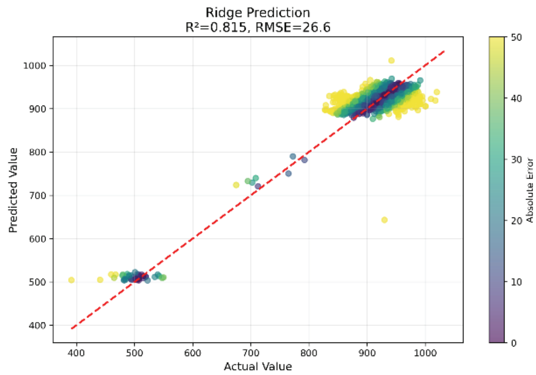

Fig. 4: Ridge Regression

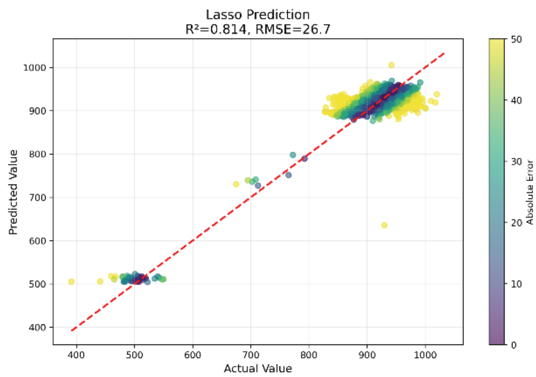

Fig. 5: Lasso Regression

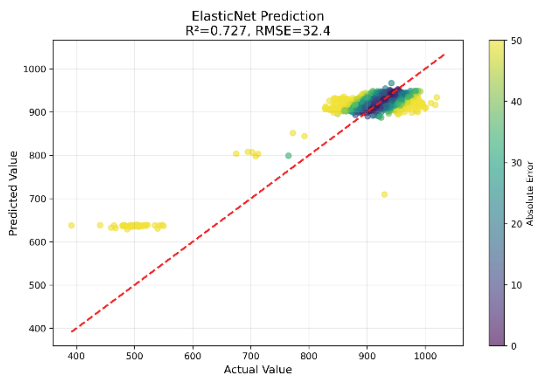

Fig. 6: ElasticNet Regression

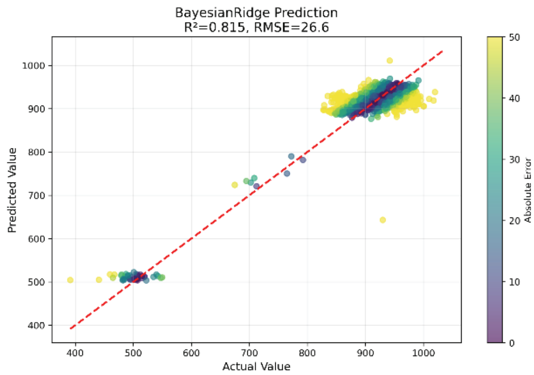

Fig. 7 :BayesianRidge Regression

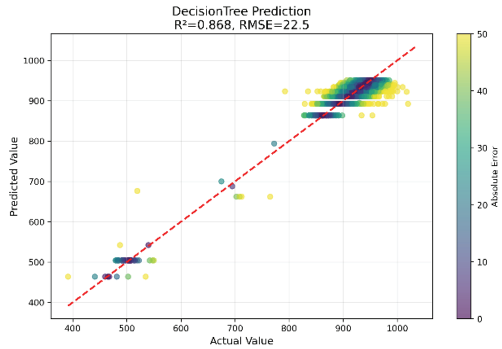

Fig. 8: DecisionTree Regression

3.1.2 Tree-based Regression

Decision Tree Regression is a tree-based regression method. The fundamental concept involves recursively partitioning the dataset to construct a tree-like structure, where each internal node represents a split condition based on a feature, and each leaf node corresponds to a predicted value. Random Forest Regression, an ensemble learning approach, enhances overall performance and stability by constructing multiple decision trees and aggregating their predictions. Extremely Randomized Tree Regression (Extra-Trees Regression), a variant of Random Forest, further mitigates overfitting risks by employing entirely random selections of features and split points during tree construction [21]. The splits in these methods are determined by minimizing the mean squared error (MSE). As shown in Fig. 8 to Fig.10, the ExtraTrees Regression model demonstrated the best performance.

As shown in Fig.10, In the low-strength region, The data points are densely distributed along the ideal line (dashed red line) and are predominantly purple in color (error < 10), indicating that the model has excellent prediction stability in this range. This may reflect that there is a significant relationship between the features and the target variable.

In the high-strength region , it can be observed that the errors increase. This is because when approaching the strength limit of cast titanium alloy, the factors influencing tensile strength become more complex and the effects of production processes need to be considered. These production processes include casting parameters during the pouring process (such as pouring temperature and cooling rate), heat treatment (such as soaking time and soaking temperature), and hot isostatic pressing (HIP) process (such as temperature, pressure, and time).

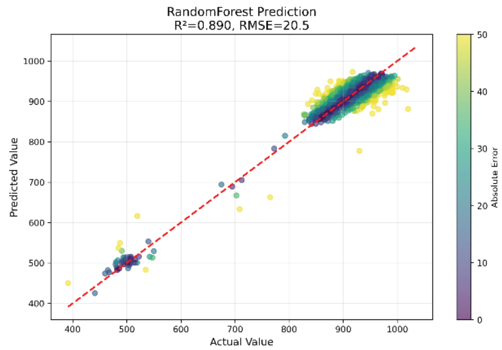

Fig. 9: RandomForest Regression

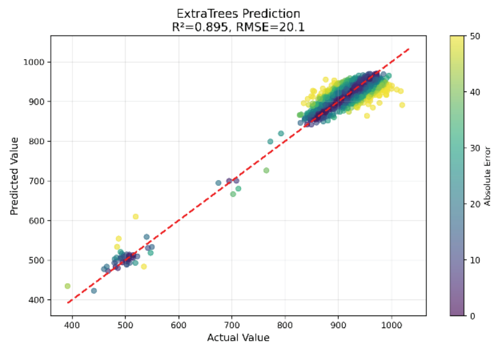

Fig. 10: ExtraTrees Regression

3.1.3 Ensemble Learning Algorithms

Ensemble Learning is a machine learning methodology that constructs a more robust model by combining multiple base learners (weak learners) [22]. The ensemble learning results are illustrated in Fig.11-Fig.15.The performance of the ensemble learning model is superior.

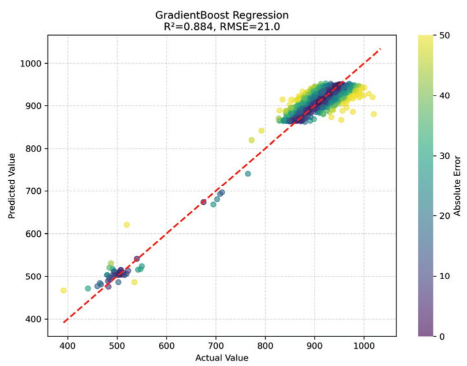

Fig. 11: GradientBoosting Regression

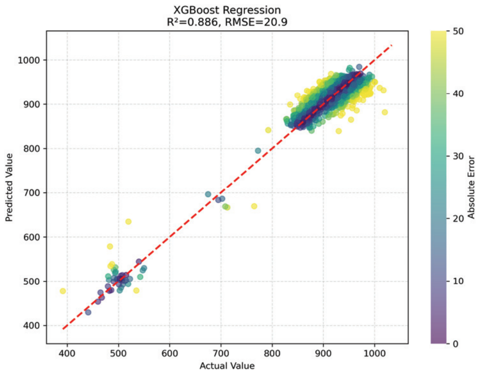

Fig. 12: XGBoost Regression

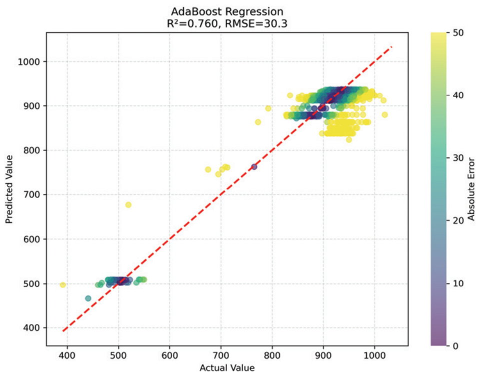

Fig. 13: AdaBoost Regression

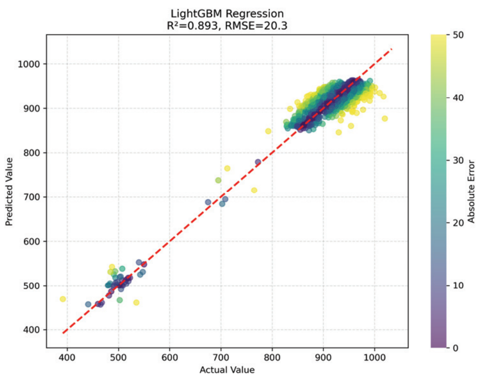

Fig. 14: LightGBM Regression

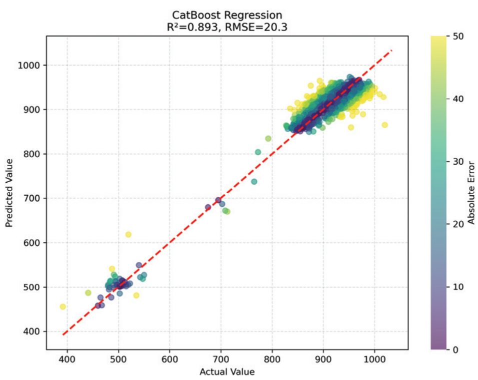

Fig. 15: CatBoost Regression

3.1.4 Other Learning Algorithms

In addition, comparisons were made with the Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP) algorithms.

As shown in Fig.16, In the low-strength region, the prediction error of the SVM model is very large. SVR transforms the input space into a high dimensional feature space through a nonlinear mapping and constructs a linear regression model in this space[23]. The core objective is to find the optimal hyperplane that allows the majority of samples to fall within the insensitive band.It is sensitive to imbalanced data types and tends to bias towards the majority class.

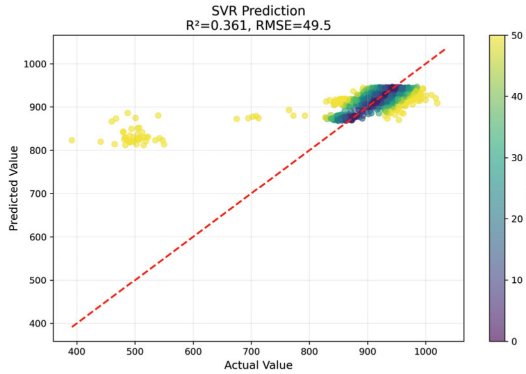

Fig. 16: SVR Regression

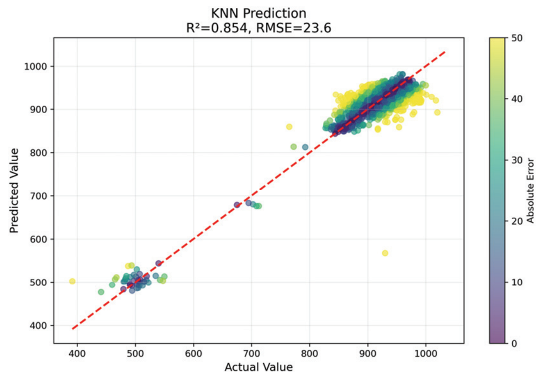

Fig. 17: KNN Regression

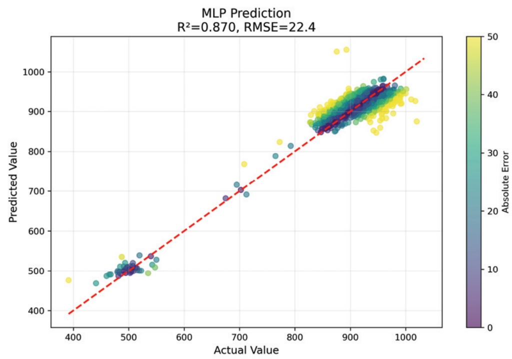

Fig. 18: MLP Regression

3.2 Feature Importance Analysis

The impact of different alloy compositions on tensile strength was obtained through the ExtraTrees tree-based method, as shown in the Fig. 19, Among them, the elements Al and V have a relatively large influence, with values of 0.699 and 0.102 respectively. The element Al is an α-stabilizing element, which contributes to solid solution strengthening in the α-phase.

The impact of aluminum content on tensile strength is shown in the Fig.20. The horizontal axis in the figure represents the aluminum content (ranging from 6.1% to 6.7%), and the vertical axis represents the tensile strength (ranging from 921 MPa to 926 MPa). The blue line in the figure shows the trend of tensile strength as the aluminum content changes. When the aluminum content is low, the tensile strength changes relatively gently; when the aluminum content exceeds 6.5%, the tensile strength increases significantly, reaching a peak before a slight decrease. Specifically, for every 1% increase in aluminum content, the tensile strength increases by approximately 10 MPa. The partial dependence plots for all the components are shown in the Fig.21.

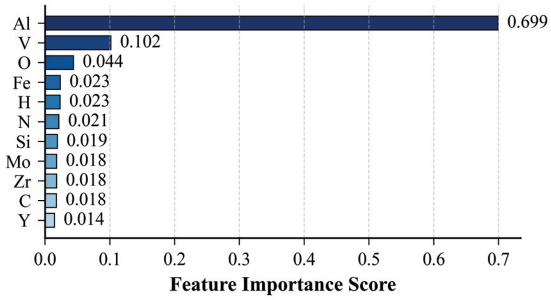

Fig. 19: Feature Importance Score

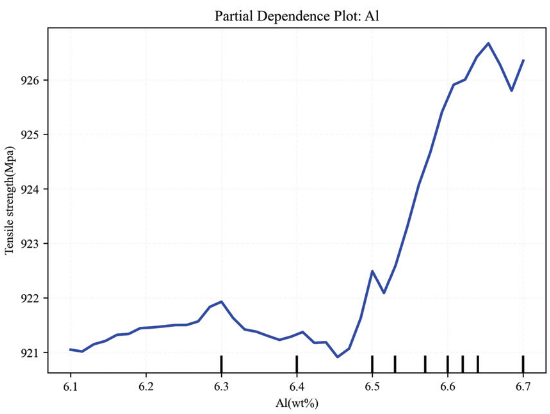

Fig. 20：PDP for the feature of aluminum

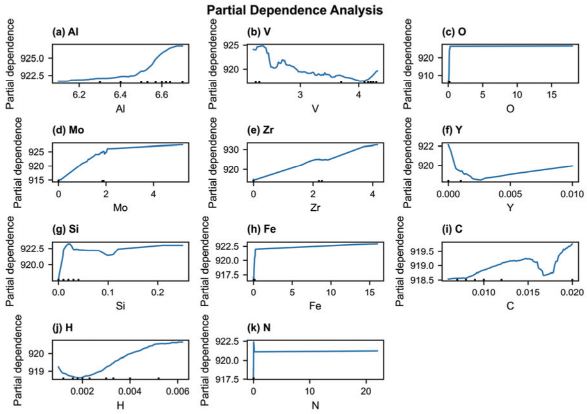

Fig. 21：PDP for the feature of all the elements

3.3 Model Optimization Analysis

The Bayesian optimization algorithm was introduced and combined with the characteristics of the data distribution to train the two peak regions separately based on Extra_Trees. The optimized coefficient of determination (R^2) is 0.896,RMSE is 19.9.

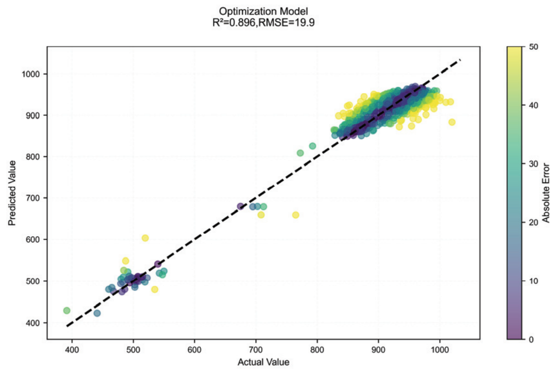

Fig. 22：Optimization Model Regression

As shown in the Fig.23, Alloy compositions with tensile strength greater than 1000 MPa have been designed using genetic algorithms. The specific compositions are shown in Table 3.

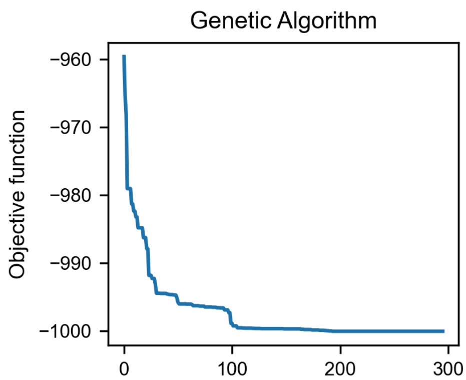

Fig. 23：Genetic Algorithm Optimization

Tab. 3 Design of High-Strength Titanium Alloy Composition

Number	Element	Content (wt.%)
1	Al	6.463
2	Zr	4,141
3	Mo	2.8
4	Y	0.672
5	Si	0.167
6	Fe	0.028
7	V	0.008
8	C	0.005
9	N	0.134
10	O	0.012
11	H	0.002

4 Conclusions

This study developed a machine learning framework for predicting tensile strength and optimizing compositions of cast titanium alloys. Key findings demonstrate that ensemble learning models, particularly ExtraTrees Regression, achieved superior predictive performance (R^2=0.896, RMSE=20.1 MPa) by effectively capturing nonlinear composition-strength relationships. Feature importance analysis revealed aluminum content as the dominant factor (69.9% contribution), exhibiting a nonlinear strengthening effect with optimal performance at 6.5 wt.% Al. The proposed genetic algorithm-based optimization successfully designed novel high-strength alloy compositions exceeding 1,000 MPa tensile strength through multi-element synergistic effects.

The principal innovations include: (1) Establishment of a data-driven prediction system integrating 16 machine learning algorithms and 13,210 industrial production datasets; (2) Development of a dual-strategy optimization framework combining Bayesian hyperparameter tuning with genetic algorithm-based composition search; (3) Identification of critical element interaction mechanisms, particularly the Mo-Zr-Si synergistic strengthening cluster and V-Y antagonistic effects.

Future research directions should focus on three aspects: (1) Expansion of the prediction system to incorporate process parameters (e.g., heat treatment conditions, HIP parameters) through multi-modal data fusion; (2) Development of cross-property optimization models considering fatigue resistance and creep performance; (3) Experimental validation of designed alloys and implementation of active learning strategies for closed-loop materials development. This machine learning-driven paradigm shows significant potential for accelerating the discovery of next-generation high-performance cast titanium alloys in aerospace applications.

Conflicts of 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.

References

[1] Schutz R W, Watkins H B. Recent developments in titanium alloy application in the energy industry[J]. Materials Science and Engineering: A, 1998, 243(1-2): 305-315.

[2] Peters M, Kumpfert J, Ward C H, et al. Titanium alloys for aerospace applications[J]. Advanced engineering materials, 2003, 5(6): 419-427.

[3] Rack H J, Qazi J I. Titanium alloys for biomedical applications[J]. Materials Science and Engineering: C, 2006, 26(8): 1269-1277.

[4] Williams J C, Boyer R R. Opportunities and issues in the application of titanium alloys for aerospace components[J]. Metals, 2020, 10(6): 705.

[5] Lütjering G, Williams J C. Commercially pure (CP) titanium and alpha alloys[J]. Titanium, 2007: 175-201.

[6] Tabie V M, Li C, Saifu W, et al. Mechanical properties of near alpha titanium alloys for high-temperature applications-a review[J]. Aircraft Engineering and Aerospace Technology, 2020, 92(4): 521-540.

[7] Veiga C, Davim J P, Loureiro A J R. Properties and applications of titanium alloys: a brief review[J]. Rev. Adv. Mater. Sci, 2012, 32(2): 133-148.

[8] L. Nastac, K. Klug, M. Gungor and t. Tack: ‘Microstructure evaluation and prediction during the investment casting of Ti-6Al-4V plates', TMS Annual Meeting, San Francisco, CA, 13-17 February 2005.

[9] Boyer R R, Briggs R D. The use of β titanium alloys in the aerospace industry[J]. Journal of Materials Engineering and Performance, 2005, 14: 681-685.

[10] Yang Y L, Wang W Q, Li F L, et al. The effect of aluminum equivalent and molybdenum equivalent on the mechanical properties of high strength and high toughness titanium alloys[C]//Materials Science Forum. Trans Tech Publications Ltd, 2009, 618: 169-172.

[11] George E P, Raabe D, Ritchie R O. High-entropy alloys[J]. Nature reviews materials, 2019, 4(8): 515-534.

[12] Wang W, Ma J. A review: Applications of machine

learning and deep learning in aerospace engineering and aero-engine engineering[J]. Advances in Engineering Innovation, 2024, 6: 54-72.

[13] Tan S, Hao X, Wang D, et al. Numerical simulation and intelligent optimization of wax injection in investment casting[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1653(1): 012016.

[14] Wang D, Sun J, Dong A, et al. An optimization method of gating system for impeller by RSM and simulation in investment casting[J]. The International Journal of Advanced Manufacturing Technology, 2018, 98: 3105-3114.

[15] Xiao C, Ye J, Esteves R M, et al. Using Spearman's correlation coefficients for exploratory data analysis on big dataset[J]. Concurrency and Computation: Practice and Experience, 2016, 28(14): 3866-3878.

[16] Bonaccorso G. Machine Learning Algorithms: Popular algorithms for data science and machine learning[M]. Packt Publishing Ltd, 2018.

[17] Plevris V, Solorzano G, Bakas N P, et al. Investigation of performance metrics in regression analysis and machine learning-based prediction models[J]. 2022.

[18] Nagelkerke N J D. A note on a general definition of the coefficient of determination[J]. biometrika, 1991, 78(3): 691-692.

[19] Balachandran P V, Xue D, Theiler J, et al. Importance of feature selection in machine learning and adaptive design for materials[M]//Materials discovery and design: By means of data science and optimal learning. Cham: Springer International Publishing, 2018: 59-79.

[20] Tranmer M, Elliot M. Multiple linear regression[J]. The Cathie Marsh Centre for Census and Survey Research (CCSR), 2008, 5(5): 1-5.

[21] Breiman L, Friedman J, Olshen R A, et al. Classification and regression trees[M]. Routledge, 2017.

[22] Sagi O, Rokach L. Ensemble learning: A survey[J]. Wiley interdisciplinary reviews: data mining and knowledge discovery, 2018, 8(4): e1249.

[23] Joachims T. Making large-scale SVM learning practical[R]. Technical report, 1998.