

Article

Interpretable Machine Learning for Mechanical Property Prediction of 5Cr-0.5Mo Steel: SHAP Explainability, Multi-Model Comparison, and Uncertainty Quantification

Saurabh Tiwari ^{1,*} , Hyoju Ahn ¹, Jongwon Lee ¹ and Nokeun Park ^{1,2,*} ¹ School of Materials Science and Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea; ahnalysis@yu.ac.kr (H.A.); jongwon1229@gmail.com (J.L.)² Institute of Materials Technology, Yeungnam University, Gyeongsan 38541, Republic of Korea

* Correspondence: saurabh@yu.ac.kr (S.T.); nokeun_park@yu.ac.kr (N.P.)

Abstract

5Cr-0.5Mo ferritic steels are widely used in high-temperature power-plant components. Although artificial neural network (ANN) models have shown good performance in predicting tensile properties, they provide limited insight into predictions and generally do not quantify the uncertainty. In this study, three tree-based machine learning models—Random Forest (RF), XGBoost (XGB), and Gradient Boosting (GB)—were developed using 36 unique alloy grade–temperature observations from a validated NIMS 5Cr-0.5Mo tensile dataset. The model performance was evaluated using leave-one-grade-out (LOGO) cross-validation, with pooled out-of-fold (OOF) predictions used to assess the overall performance. SHapley Additive exPlanations (SHAP) were used to examine feature contributions, whereas Gaussian Process Regression (GPR) was evaluated as a proof-of-concept for uncertainty quantification of yield strength (YS). GB showed the strongest performance for ultimate tensile strength (UTS) and reduction in area (RA), achieving pooled OOF R^2 values of 0.9698 and 0.9570, respectively. RF achieved corresponding R^2 values of 0.9406 and 0.9488, respectively. SHAP identified the test temperature as the most influential feature across all four properties, whereas the Cr content and austenite grain size contributed significantly to the strength predictions. For YS, the GPR achieved complete empirical coverage of the 95% predictive intervals, although the relatively large mean interval width indicated conservative uncertainty estimates. Given the limited dataset and feature correlations, the SHAP results should be regarded as exploratory, rather than mechanistic. Overall, this study demonstrates the potential of interpretable, uncertainty-aware ML for small alloy datasets, while emphasizing the need for larger, compositionally diverse datasets and independent validation.



Academic Editors: Yunfei Jia and Haijie Wang

Received: 15 July 2026

Revised: 21 August 2026

Accepted: 22 August 2026

Published: 7 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.**Keywords:** 5Cr-0.5Mo steel; machine learning; SHAP; Gradient Boosting; Gaussian Process Regression; uncertainty quantification

1. Introduction

Ferritic chromium-molybdenum steels, particularly 5Cr-0.5Mo steel (4.6–5.0 wt. % Cr and 0.49–0.52 wt. % Mo), are widely used in power plants for heater tubes, pressure vessels, and high-temperature piping. These components typically operate between 500 and 700 °C under long-term mechanical loading conditions [1]. The alloy combines the good oxidation resistance of chromium with solid-solution and precipitation strengthening from molybdenum. Its ferritic body-centered cubic (BCC) microstructure also provides

good toughness over a wide temperature range [2]. The tensile properties of 5Cr-0.5Mo steel depend on several interacting microstructural factors. Carbon influences the amount and morphology of carbide precipitates, which evolve during thermal exposure through the sequence of ϵ -carbide, cementite, M_2C , M_7C_3 , $M_{23}C_6$, and M_6C [3]. Chromium- and molybdenum-rich carbides, especially $M_{23}C_6$, mainly form along the grain boundaries above approximately 400 °C. These precipitates initially increase the strength but gradually dissolve at higher temperatures, leading to softening [3]. Austenite grain size (AGS) also affects the available grain boundary area for carbide precipitation and therefore influences both strength and ductility.

Machine learning (ML) has become an effective tool for predicting the mechanical properties of engineering alloys [4–7]. Tree-based ensemble methods, including Random Forest (RF) [8], Gradient Boosting (GB) [9], and XGBoost [10], are particularly attractive because they capture nonlinear relationships while also supporting model interpretation. Current studies further demonstrate the growing application of data-driven approaches to Cr-Mo steels and their related steel systems. Machine-learning models have been applied to predict the creep rupture life of 5Cr-0.5Mo steel using composition, temperature, stress, and microstructural descriptors, while Gradient Boosting has been used to model microstructural evolution and mechanical properties across heat-treated steel grades. Recently, ANN-based analysis has been applied to quantify temperature-induced variations in the mechanical properties of 5Cr-0.5Mo steel. These studies demonstrate the potential of data-driven methods for capturing complex composition–processing–property relationships, while also emphasizing the importance of appropriate model development and validation [11–13]. A recent study by Ishtiaq et al. [14] used an artificial neural network (ANN) to predict the tensile properties of 5Cr-0.5Mo steel with good accuracy. However, the model does not explain how individual input variables influence the predictions and does not provide prediction uncertainty. In addition, the random train–test split used in that study did not fully assess model generalization because measurements from the same alloy grade could appear in both the training and testing sets. SHapley Additive exPlanations (SHAP) [15,16] provides a reliable way to quantify the contribution of each input feature to model predictions. Gaussian Process Regression (GPR) [17,18] complements this analysis by providing both predicted values and associated uncertainty through predictive intervals, making it useful for engineering applications where prediction reliability is important [4].

Recent studies have demonstrated the potential of machine-learning-based surrogate and inverse-design frameworks for engineering applications. Sarfarazi et al. developed explainable inverse-learning approaches for structural design, demonstrating how surrogate models, SHAP-based interpretation, and systematic validation can support performance-driven engineering decisions [19,20]. These developments further highlight the importance of model interpretability and transparent validation when data-driven predictions are used to support engineering designs. Accordingly, this study evaluates Random Forest (RF), XGBoost, and Gradient Boosting (GB) using leave-one-grade-out (LOGO) cross-validation and discusses their performance in a methodological context with a previously reported ANN study. SHAP analysis was used to explore the contribution of each input variable to the predicted mechanical properties, with the observed associations cautiously discussed in the context of established metallurgy. In addition, Gaussian Process Regression (GPR) was employed under the same LOGO cross-validation framework to quantify the prediction uncertainty through 95% predictive intervals.

2. Materials and Methods

2.1. Dataset Description and Preparation

The dataset used in this study was obtained from the supplementary material of Ishtiaq et al. [14], who compiled it from the tensile property database of the National Institute for Materials Science (NIMS, MatNavi) in Japan. Tensile tests were performed on specimens with a diameter of 10 mm and gauge length of 50 mm according to ISO 9513 [21] at a strain rate of 10^{-3} s^{-1} [14]. Each alloy grade was described by 14 input features, including 11 elemental composition variables (C, Si, Mn, P, S, Ni, Cr, Mo, Cu, Al, and N; wt. %), austenite grain size (AGS, μm), non-metallic inclusion content (NMI, wt. %), and test temperature ($^{\circ}\text{C}$). The target properties were yield strength (YS, MPa), ultimate tensile strength (UTS, MPa), elongation (EL, %), and reduction in area (RA, %). The reported supplementary dataset contains 238 records corresponding to 36 unique alloy grade–temperature combinations (nine alloy grades $\times$ four test temperatures: 500, 550, 600, and 650°C). Multiple identical records were identified for each unique alloy grade–temperature combination. To avoid potential bias from repeated identical observations, duplicate records were removed before model development, and all machine learning analyses were performed using the 36 unique observations. The 36 observations represent a complete set of unique alloy grade–temperature combinations available in the published dataset. The processed dataset used in this study is provided in Supplementary Data S1. Descriptive statistics for all input and output variables are summarized in Table 1.

Table 1. Descriptive statistics of the input and output variables for the 5Cr-0.5Mo steel dataset (36 unique observations from 9 alloy grades evaluated at four test temperatures: 500, 550, 600, and 650°C).

Variable	Minimum	Maximum	Mean	Std Dev
C (wt. %)	0.090	0.120	0.108	0.012
Si (wt. %)	0.270	0.370	0.331	0.035
Mn (wt. %)	0.440	0.560	0.499	0.048
P (wt. %)	0.007	0.022	0.015	0.005
S (wt. %)	0.005	0.012	0.007	0.003
Ni (wt. %)	0.043	0.083	0.060	0.016
Cr (wt. %)	4.610	5.020	4.835	0.151
Mo (wt. %)	0.490	0.520	0.503	0.011
Cu (wt. %)	0.050	0.130	0.072	0.031
Al (wt. %)	0.004	0.008	0.006	0.001
N (wt. %)	0.0104	0.0178	0.014	0.003
AGS (μm)	4.800	6.900	5.967	0.583
NMI (wt. %)	0.030	0.100	0.086	0.045
Temp ($^{\circ}\text{C}$)	500	650	—	—
YS (MPa)	106	338	194	64
UTS (MPa)	175	421	291	65
EL (%)	28	80	42	10
RA (%)	78	96	89	5

2.2. Data Preprocessing and Model Development

2.2.1. Data Preprocessing

Before model training, all 14 input features were standardized using the Standard-Scaler implementation of Scikit-learn [22]. To prevent information leakage, the scaler was fitted only on the training data within each Leave-One-Grade-Out (LOGO) fold and then applied to transform the corresponding test data. The dataset contained no missing values; therefore, data imputation was not required.

2.2.2. Machine Learning Models

Three tree-based ensemble learning algorithms were evaluated: Random Forest (RF) [8], XGBoost [10], and Gradient Boosting (GB) [9]. Random Forest constructs an ensemble of decision trees using bootstrap sampling and random feature selection at each node split. The RF model was configured with 300 estimators and no restriction on the maximum tree depth, allowing the model to capture the nonlinear relationships among the input variables. XGBoost builds decision trees sequentially, with each tree learning from the residual errors of the previous tree. The model was configured with 300 estimators, a maximum tree depth of five, and a learning rate of 0.05. Gradient Boosting follows a similar sequential boosting strategy and is implemented using scikit-learn. The model employed 300 estimators, a maximum tree depth of four, and a learning rate of 0.05. StandardScaler was applied to the input features before the model fitting. Standardization did not affect the predictions of RF, XGBoost, or GB because these tree-based models are insensitive to feature scaling. Scaling was nevertheless applied uniformly to maintain a consistent preprocessing procedure across the tree-based models and GPR analysis, for which feature scaling is relevant. The same preprocessing procedure and predefined model configurations were applied independently to each target property (YS, UTS, EL, and RA), and separate models were trained for every target. The model configurations were defined manually a priori before the LOGO analysis and were kept unchanged across all nine LOGO folds and target properties. Specifically, RF was configured with 300 estimators and no maximum depth constraint; XGB with 300 estimators, `max_depth = 5`, and `learning_rate = 0.05`; and GB with 300 estimators, `max_depth = 4`, and `learning_rate = 0.05`. These configurations were selected a priori as reasonable, near-default settings and were not adjusted according to the observed LOGO performance. No grid search, randomized search, Bayesian optimization, or other automated hyperparameter search procedures were performed at any stage. Nested grouped cross-validation for hyperparameter optimization was not performed because the dataset contained only 36 unique observations distributed across nine alloy grades. An additional inner grouped-validation procedure would further reduce the number of independent alloy grade groups available for model fitting and hyperparameter selection, and could produce unstable results. Accordingly, the reported results represent the performance of these predefined model configurations rather than optimized configurations, and alternative hyperparameter choices could yield somewhat different performances.

2.3. Model Evaluation and Interpretation

2.3.1. Leave-One-Grade-Out Cross-Validation

The dataset comprised 36 unique observations from nine alloy grades, with four test temperatures per grade. A conventional random train–test split may place measurements from the same alloy grade in both training and test sets, leading to optimistic estimates of predictive performance. To provide a more rigorous evaluation, leave-one-grade-out (LOGO) cross-validation was performed. In each of the nine folds, one alloy grade was used as the test set, whereas the remaining eight alloy grades were used for training. This strategy evaluates the model’s ability to predict the mechanical properties of an unseen alloy grade and minimizes potential information leakage arising from repeated identical observations. Given the limited number of unique observations relative to the number of input features, particular attention was paid to reducing the risk of optimistic performance estimation. Duplicate observations were removed before model development, and LOGO cross-validation was used to ensure that each test fold consisted of an entirely unobserved alloy grade. Predefined model configurations were consistently applied across all folds without hyperparameter selection based on the held-out fold performance. In addition, the tree depth was constrained for GB and XGB, whereas RF employed bootstrap aggregation

and random feature selection. These measures reduce the risk of overfitting and information leakage; nevertheless, the limited dataset size requires the resulting performance estimates to be interpreted as exploratory results.

2.3.2. SHAP Explainability Analysis

Model interpretability was investigated using SHapley Additive exPlanations (SHAP) [15,16]. SHAP values were calculated using the TreeExplainer algorithm [16], which computes the exact Shapley values for tree-based models. For each prediction, the SHAP value of a feature represents its contribution to the deviation of the prediction from the model's average prediction while satisfying the efficiency, symmetry, dummy, and linearity properties of the cooperative game theory. The global feature importance was quantified using the mean absolute SHAP value ($\text{mean}(|\phi_i|)$) across all observations. Beeswarm plots were used to visualize the magnitude and direction of the feature effects. SHAP analysis was performed using the Random Forest model retrained on the complete dataset used in this study after model evaluation.

2.3.3. Gaussian Process Regression for Uncertainty Quantification

Prediction uncertainty was evaluated using Gaussian Process Regression (GPR) [17,18] for the yield strength (YS) as a proof-of-concept uncertainty quantification analysis. GPR models the response as a Gaussian process and provides both the predicted mean and associated predictive uncertainty. A Matérn 5/2 kernel combined with a WhiteKernel noise term was adopted following Rasmussen and Williams [17]. The kernel hyperparameters were optimized by maximizing the marginal log-likelihood within each training fold. To ensure methodological consistency, the GPR was evaluated using the same LOGO cross-validation protocol as the ensemble models. For each test observation, the GPR model returned a predictive mean and standard deviation using the predict (return_std = True) implementation of Scikit-learn's GaussianProcessRegressor. The 95% predictive interval was calculated as $\hat{y} \pm 1.96\sigma$, where σ is the standard deviation returned by the fitted models. Because the fitted covariance kernel is the sum of a Matérn 5/2 component and a WhiteKernel component, the estimated WhiteKernel noise variance contributes to the predictive variance returned at the test points. Therefore, the reported intervals incorporate both the uncertainty in the latent function and the estimated observation noise contribution represented by the WhiteKernel. The additional $\alpha = 10^{-2}$ term is a fixed regularization value added separately to the diagonal of the training covariance matrix during fitting and is distinct from the WhiteKernel noise parameter.

3. Results and Discussion

3.1. Dataset Correlations

Figure 1 presents the Pearson correlation matrix for the 14 input features and four mechanical properties. Among all variables, the test temperature exhibited the strongest correlation with the mechanical properties, showing a negative correlation with YS ($r = -0.61$) and UTS ($r = -0.90$), and a positive correlation with EL ($r = +0.72$) and RA ($r = +0.92$). These trends reflect the expected reduction in strength and improvement in ductility with increasing temperatures. Among the compositional variables, the Cr content exhibited a moderate positive correlation with YS ($r = +0.64$). This association is metallurgically plausible, given the established roles of Cr in solid-solution strengthening and carbide formation; however, these mechanisms cannot be directly confirmed from the present dataset. Phosphorus and sulfur also show strong positive correlations with NMI, reflecting their well-established roles in the formation of nonmetallic inclusions [14]. Overall, the

test temperature showed the strongest correlation with the tensile properties, whereas Cr showed a comparatively strong compositional association within the investigated dataset.

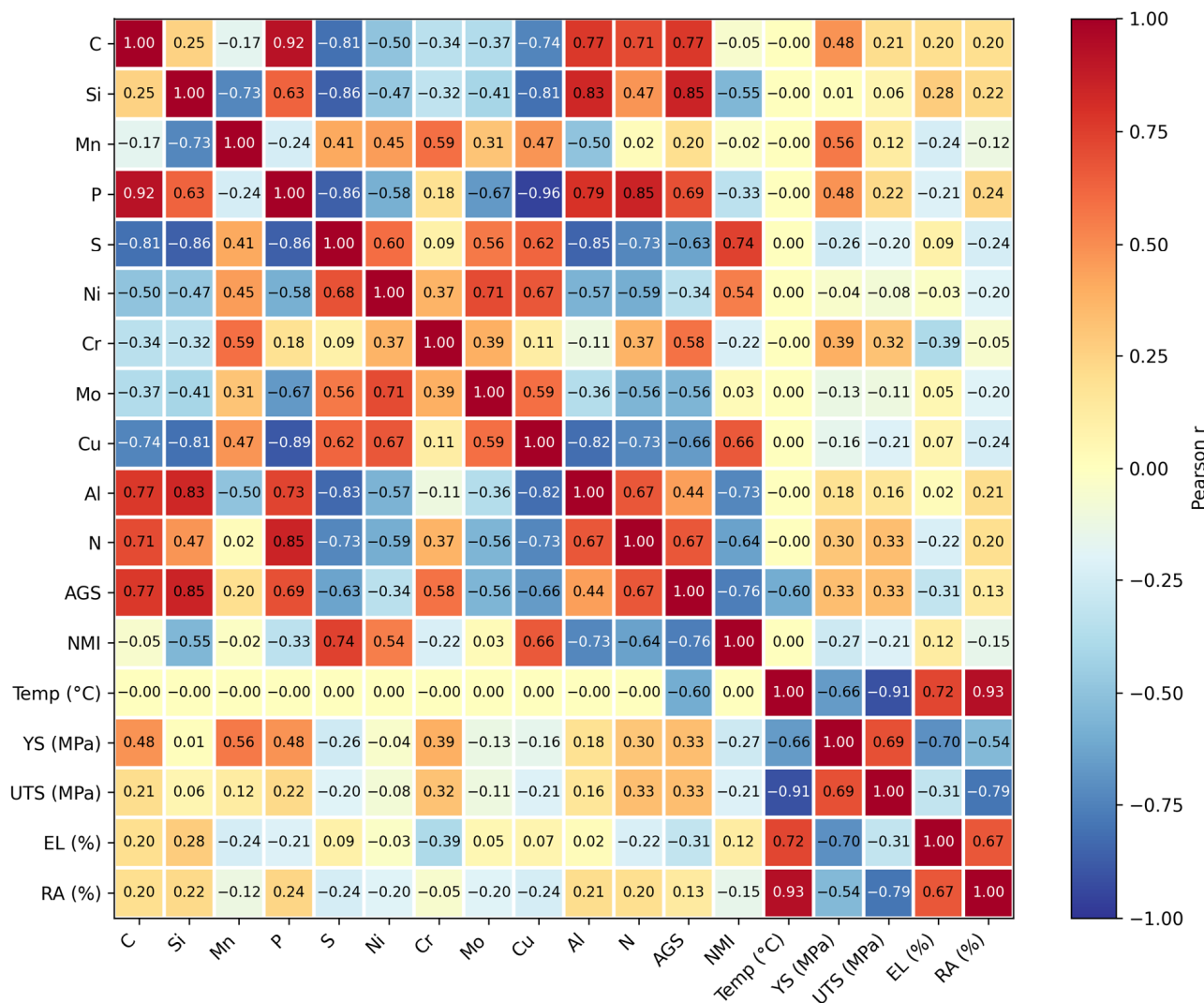


Figure 1. Pearson correlation matrix of the 14 input features and 4 mechanical properties for the 5Cr-0.5Mo steel dataset (36 unique observations). AGS = austenite grain size; NMI = non-metallic inclusion content; Temp = test temperature.

3.2. Leave-One-Grade-Out Cross-Validation Performance

Table 2 and Figure 2 summarize the performance of the three ensemble models under leave-one-grade-out (LOGO) cross-validation. The model accuracy strongly depends on the target property. UTS and RA were predicted with consistently high accuracy, whereas YS and EL were more difficult to predict. For UTS, Gradient Boosting (GB) achieved the highest performance, with a mean R^2 of 0.964 ± 0.020 , MAE of 10.27 ± 3.57 MPa, and RMSE of 11.40 ± 3.49 MPa. Random Forest (RF) also performed well ($R^2 = 0.929 \pm 0.060$), whereas XGBoost (XGB) showed lower accuracy ($R^2 = 0.844 \pm 0.243$) and substantially higher fold-to-fold variability. The substantially greater fold-to-fold variability indicates that the predefined XGBoost configuration was less stable than RF and GB across small LOGO training sets. This behavior may reflect the limited number of training observations in each fold ($n = 32$), sensitivity to the selected model complexity and boosting parameters, or a combination thereof.

Table 2. Leave-One-Grade-Out (LOGO) cross-validation performance. Fold-wise values are reported as mean $\pm$ standard deviation across the nine LOGO folds. Because each test fold contained only four observations, pooled out-of-fold (OOF) metrics calculated from all 36 predictions were used as the primary summary of the overall predictive performance.

Model	Target	R ² (Mean $\pm$ SD) Fold-Wise	MAE (Mean $\pm$ SD) Fold-Wise	RMSE (Mean $\pm$ SD) Fold-Wise	Pooled R ²	Pooled MAE	Pooled RMSE
RF	YS	0.759 $\pm$ 0.231	15.45 $\pm$ 9.78 MPa	17.18 $\pm$ 10.01 MPa	0.9015	15.45 MPa	19.89 MPa
XGB	YS	−0.343 $\pm$ 2.861	25.68 $\pm$ 31.62 MPa	28.11 $\pm$ 33.21 MPa	0.5288	25.68 MPa	43.51 MPa
GB	YS	0.803 $\pm$ 0.182	13.73 $\pm$ 7.71 MPa	15.32 $\pm$ 7.93 MPa	0.9259	13.73 MPa	17.25 MPa
RF	UTS	0.929 $\pm$ 0.060	13.94 $\pm$ 6.94 MPa	15.21 $\pm$ 6.98 MPa	0.9406	13.94 MPa	16.74 MPa
XGB	UTS	0.844 $\pm$ 0.243	17.47 $\pm$ 16.49 MPa	18.71 $\pm$ 16.22 MPa	0.8699	17.47 MPa	24.76 MPa
GB	UTS	0.964 $\pm$ 0.020	10.27 $\pm$ 3.57 MPa	11.40 $\pm$ 3.49 MPa	0.9698	10.27 MPa	11.92 MPa
RF	EL (%)	−0.273 $\pm$ 2.556	4.85 $\pm$ 2.61%	5.83 $\pm$ 3.62%	0.6054	4.85%	6.86%
XGB	EL (%)	−0.954 $\pm$ 3.050	7.12 $\pm$ 2.36%	8.40 $\pm$ 2.89%	0.3380	7.12%	8.88%
GB	EL (%)	−0.183 $\pm$ 2.013	5.53 $\pm$ 2.63%	6.39 $\pm$ 3.33%	0.5648	5.53%	7.20%
RF	RA (%)	0.934 $\pm$ 0.081	0.91 $\pm$ 0.40%	1.08 $\pm$ 0.51%	0.9488	0.91%	1.20%
XGB	RA (%)	0.824 $\pm$ 0.285	1.17 $\pm$ 0.71%	1.52 $\pm$ 0.96%	0.8842	1.17%	1.80%
GB	RA (%)	0.937 $\pm$ 0.078	0.80 $\pm$ 0.32%	1.01 $\pm$ 0.43%	0.9570	0.80%	1.10%

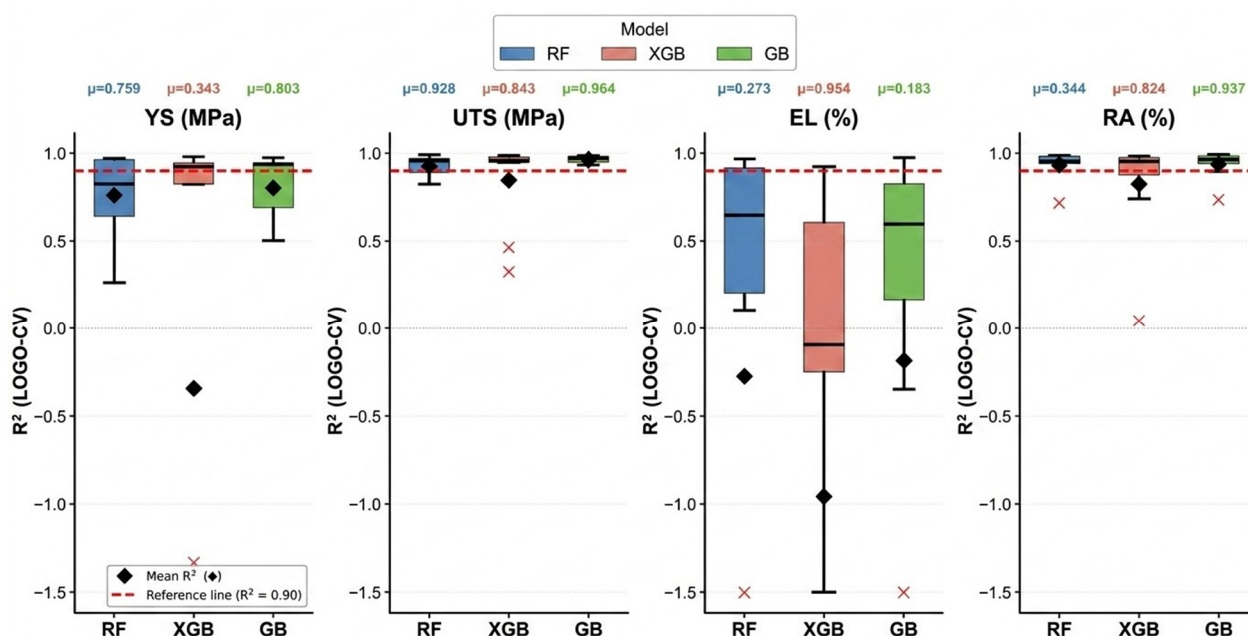


Figure 2. Leave-One-Grade-Out cross-validation R² scores for all three ensemble models and all four target mechanical properties (36 unique observations, 9 alloy grades). Each boxplot represents the distribution of R² across the nine LOGO folds. Diamond markers (◆) indicate fold means. The dashed horizontal line at R² = 0.90 denotes the performance reference threshold. RF = Random Forest; XGB = XGBoost; GB = Gradient Boosting.

Because no systematic hyperparameter or sensitivity analysis was performed, the specific cause of the poorer XGBoost performance could not be established from the present results, and alternative XGBoost configurations may yield different performances.

For RA, GB achieved the highest mean R² (0.937 $\pm$ 0.078) with the lowest prediction error (MAE = 0.80 $\pm$ 0.32%), closely followed by RF (R² = 0.934 $\pm$ 0.081, MAE = 0.91 $\pm$ 0.40%). Both models exhibited comparatively strong and consistent predictive performances across the held-out alloy grades. The prediction of YS was more challenging, with mean R² values of 0.803 $\pm$ 0.182 for GB and 0.759 $\pm$ 0.231 for RF. The lower predictive performance for YS may reflect microstructural factors that are not explicitly represented by the available descriptors, such as the carbide state, grain boundary characteristics, and other aspects of

microstructural evolution. These factors were not directly measured in the present dataset and, therefore, cannot be evaluated independently.

EL was the most difficult property to predict. All three models produced R^2 values close to zero or negative, indicating that composition, grain size, inclusion content, and test temperature alone are insufficient to explain the variation in the elongation across the unseen alloy grades. This observation is consistent with the large scatter reported in a previous ANN study [14]. Fold-wise R^2 values should be interpreted with caution because each LOGO test fold contained only four observations (one alloy grade $\times$ four temperatures). Consequently, the fold-wise R^2 can be highly sensitive to individual prediction errors and may become strongly negative even when the absolute prediction errors are moderate. Therefore, pooled out-of-fold (OOF) metrics calculated from all 36 held-out predictions were used as the primary summary of the overall predictive performance. Pooled R^2 was calculated as $R^2 = 1 - SS_{\text{res}}/SS_{\text{tot}}$, with SS_{res} and SS_{tot} calculated across all 36 OOF predictions simultaneously. The fold-wise R^2 was retained to characterize the variation in performance among the nine held-out alloy grades rather than as an independent estimate of the overall predictive accuracy. Figure 3 presents the predicted versus experimental values for the RF model using all 36 out-of-fold predictions generated under LOGO cross-validation, providing a complementary visualization of the model performance across the four target properties.

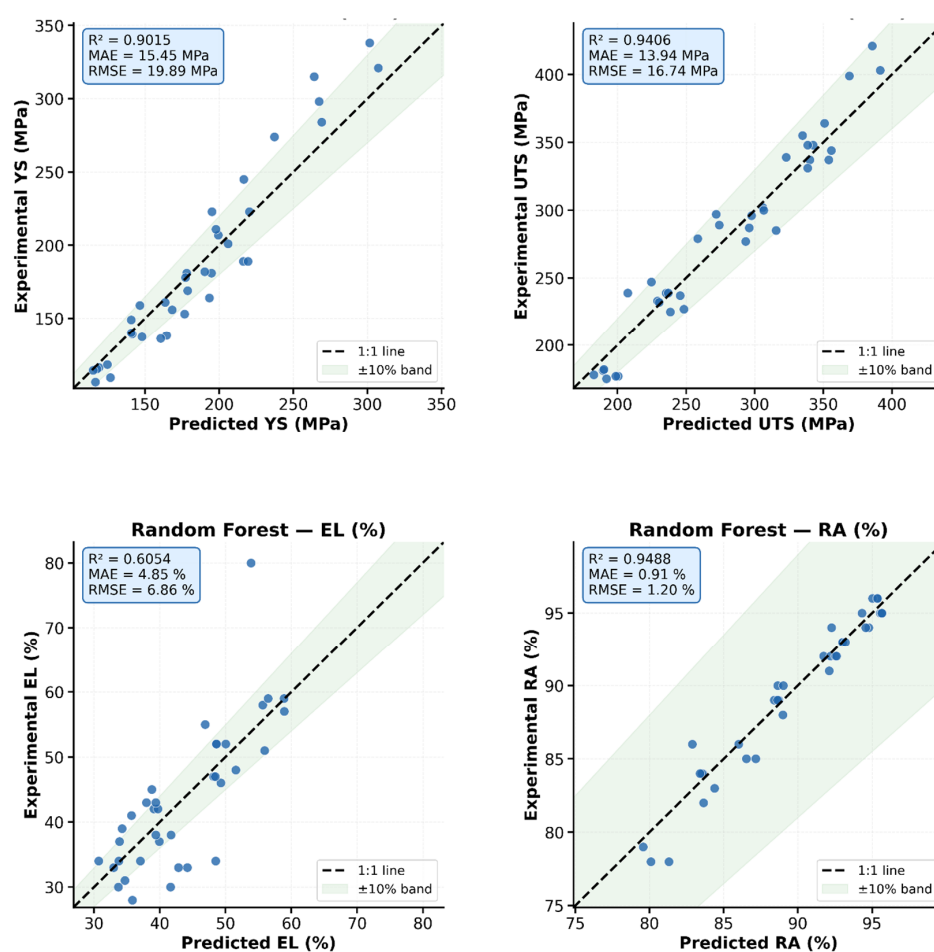


Figure 3. Predicted versus experimental mechanical properties for Random Forest, evaluated under leave-one-grade-out cross-validation (36 unique observations, within-fold StandardScaler). Each point represents an individual measurement of the withheld test alloy grade. The dashed line represents the ideal 1:1 correspondence, and the shaded band indicates a $\pm 10\%$ deviation.

3.3. SHAP Feature Importance Analysis

Figure 4 presents the global SHAP feature importance (mean absolute SHAP value) for the four target properties obtained from the Random Forest model trained on the complete 36-observation dataset. The SHAP results provide an exploratory ranking of the feature contributions. Because several compositional and microstructural variables are strongly correlated (Figure 1), their individual SHAP attributions may be shared or redistributed among the correlated features. Therefore, the relative rankings of variables such as Cr, AGS, P, S, and NMI should be interpreted cautiously and as hypothesis-generating rather than as independently validated features. The test temperature is more readily interpretable because it varies systematically within each alloy grade, whereas the composition and AGS remain constant across the four temperature measurements of a given grade. The test temperature was the most influential feature for all four properties, with mean $|\text{SHAP}|$ values of 35.80 (YS), 56.07 (UTS), 6.75 (EL), and 4.42 (RA). These values were substantially higher than those of the compositional variables, indicating that temperature had the strongest SHAP association with the predicted tensile response in the model. This behavior is expected because increasing temperature promotes dislocation mobility and dynamic recovery while accelerating the dissolution and coarsening of strengthening M_{23}C_6 carbides, resulting in reduced strength and improved ductility [3]. Among the compositional variables, the Cr content was the second most important feature for both YS (mean $|\text{SHAP}| = 14.27$) and UTS (mean $|\text{SHAP}| = 5.82$). The positive contribution of Cr is consistent with its known role in solid solution strengthening and carbide formation in Cr–Mo steels. However, this metallurgical interpretation remains a physically plausible hypothesis because the present dataset does not contain direct information on the carbide fraction, size, morphology, phase state, or dislocation density. Similarly, the austenite grain size (AGS) ranked third for both strength properties. This association is consistent with the established relationship between grain size and strength; however, the present model cannot independently establish the Hall–Petch mechanism because other microstructural variables are not available in the dataset. For ductility-related properties (EL and RA), the test temperature remained the dominant feature. Si and S also contributed to the prediction of RA; however, the underlying metallurgical mechanisms could not be established from the present dataset. Overall, the SHAP analysis showed the strongest association for the test temperature across the predicted tensile properties, whereas the Cr content and AGS showed comparatively strong contributions to the predicted strength response. These associations should be interpreted as exploratory rather than confirmed mechanistic relationships.

3.4. SHAP Beeswarm Analysis: Direction of Feature Effects

Figure 5 shows the SHAP beeswarm plots for YS and UTS, illustrating the magnitude, direction, and distribution of the feature contributions across the 36 unique observations. For both YS and UTS, the test temperature showed the strongest influence on the model predictions. High temperatures (red points) were associated with negative SHAP values, indicating a lower predicted strength, whereas low temperatures (blue points) produced positive SHAP values and increased the predicted strength. This trend agrees with the well-established reduction in the strength of ferritic steels at elevated temperatures. The Cr content exhibited a clear positive association with the predicted YS. Higher Cr content (red points) generally produced positive SHAP values, whereas lower Cr content (blue points) tended to produce lower or negative SHAP contributions. This pattern is consistent with the known metallurgical roles of Cr in Cr–Mo steels, including solid solution strengthening and carbide formation. However, these mechanisms cannot be confirmed directly from the present dataset because the relevant microstructural state variables are not available.

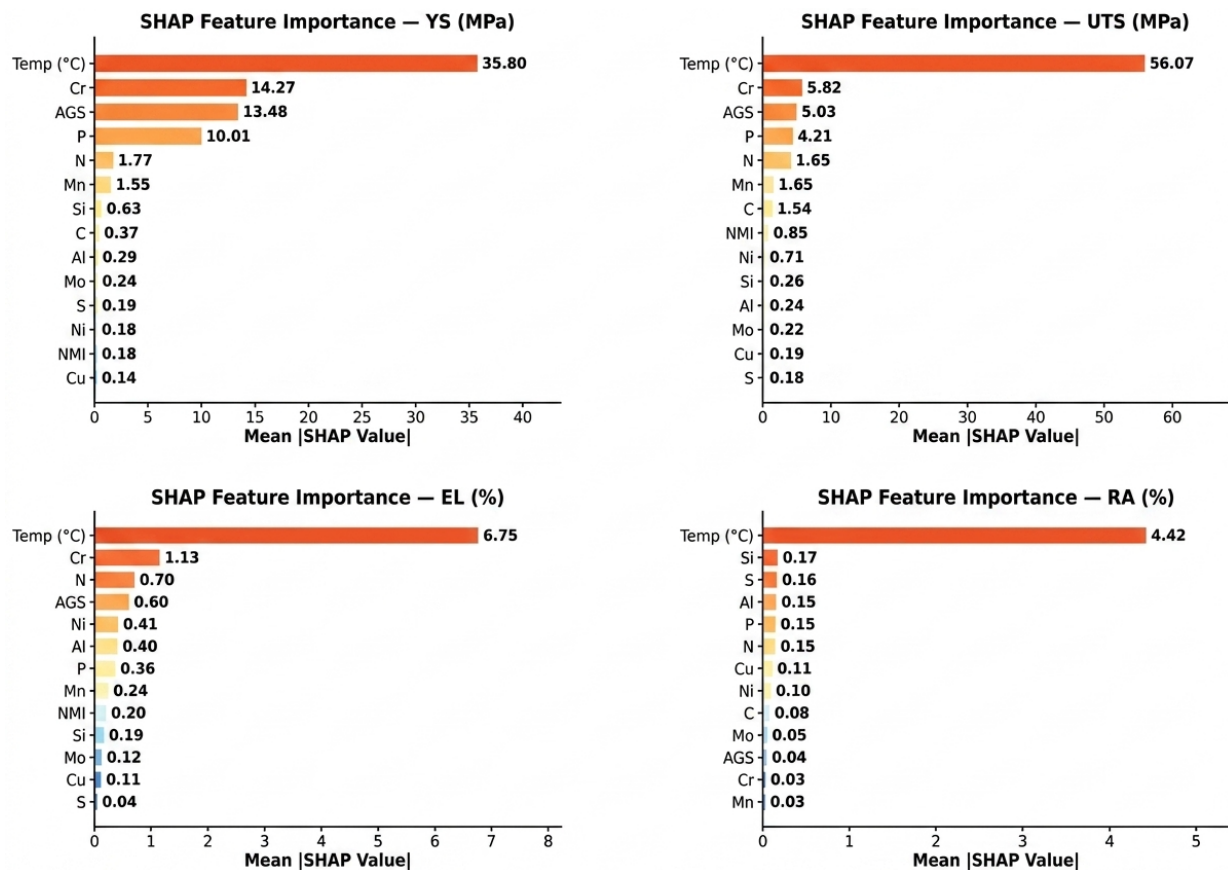


Figure 4. Global SHAP feature importance (mean absolute SHAP value) for the four target mechanical properties obtained from the Random Forest model trained on the complete dataset used in this study. The features were ranked by the mean absolute SHAP value. AGS = austenite grain size; NMI = non-metallic inclusion content; Temp = test temperature.

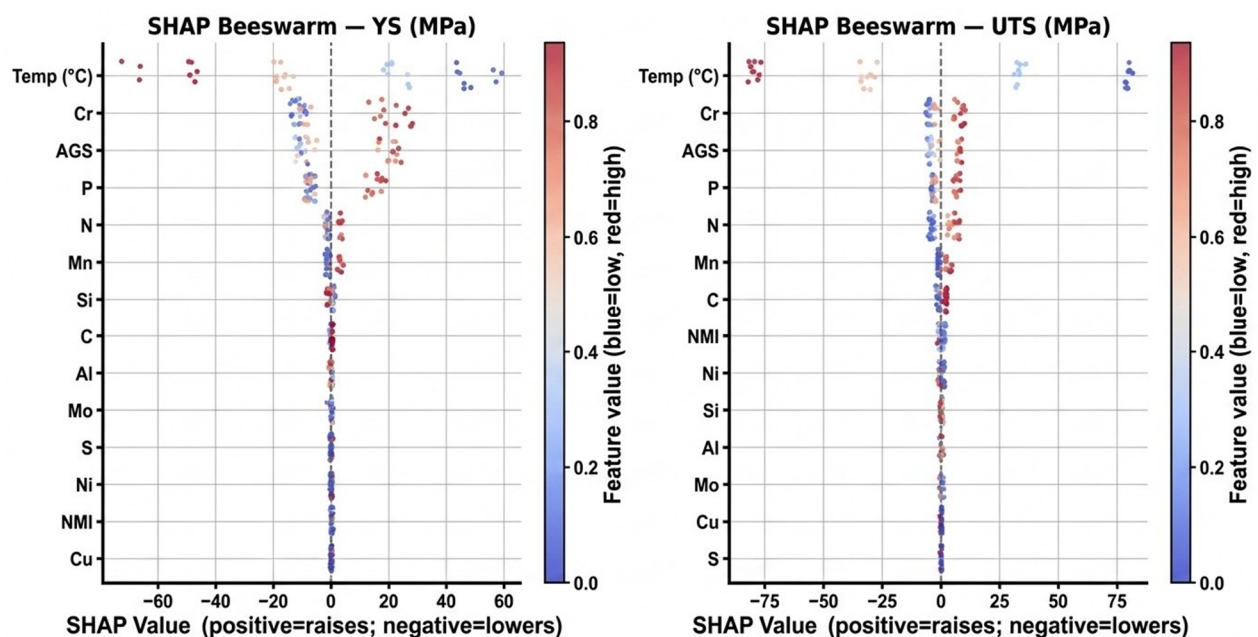


Figure 5. SHAP beeswarm plots for yield strength (YS) (left) and ultimate tensile strength (UTS) (right) for all 36 unique observations. The x-axis position indicates the SHAP value (impact on model output), and the point color encodes the normalized feature value (red = high, blue = low). The features are sorted by the mean |SHAP value| from top to bottom.

In contrast, non-metallic inclusion (NMI) content had SHAP values close to zero across most observations, indicating a limited contribution to the predicted tensile properties within the range of 0.03–0.10 wt. % considered in this dataset. Overall, these SHAP patterns suggest that the test temperature has the strongest association with the predicted tensile strength in the fitted model, whereas Cr shows a comparatively strong compositional contribution. These associations should be interpreted cautiously, considering the correlated features and limited dataset size.

3.5. Uncertainty Quantification via Gaussian Process Regression

Prediction uncertainty was evaluated using Gaussian Process Regression (GPR) for yield strength (YS) only, as a proof-of-concept demonstration under the same LOGO cross-validation protocol. YS was selected because it exhibited greater variation in the fold-wise predictive performance than UTS and RA, making the uncertainty analysis particularly informative for this target. GPR was not evaluated for UTS, EL, or RA; therefore, similar predictive performance or uncertainty characteristics should not be assumed for these properties. The GPR achieved an overall LOGO R^2 of 0.803 with a mean absolute error (MAE) of 20.56 MPa across the 36 predictions. The 95% predictive intervals ($\pm 1.96\sigma$) covered 100% of all experimental observations (Figure 6a). The mean predictive interval width was 147.9 MPa, corresponding to a normalized width of 0.64 relative to the observed YS range, indicating that complete empirical coverage was accompanied by relatively wide intervals. The mean predictive standard deviation was 37.8 MPa, and the 95th percentile was 61.4 MPa (Figure 6b). The variation in the predictive standard deviation across the LOGO predictions indicates that the estimated uncertainty was not uniform among the held-out observations. These predictive intervals provide additional information beyond point predictions; however, their relatively large widths should be considered when interpreting prediction uncertainty. The extension of the GPR uncertainty analysis to UTS, EL, and RA using larger and more diverse datasets would be valuable in future studies.

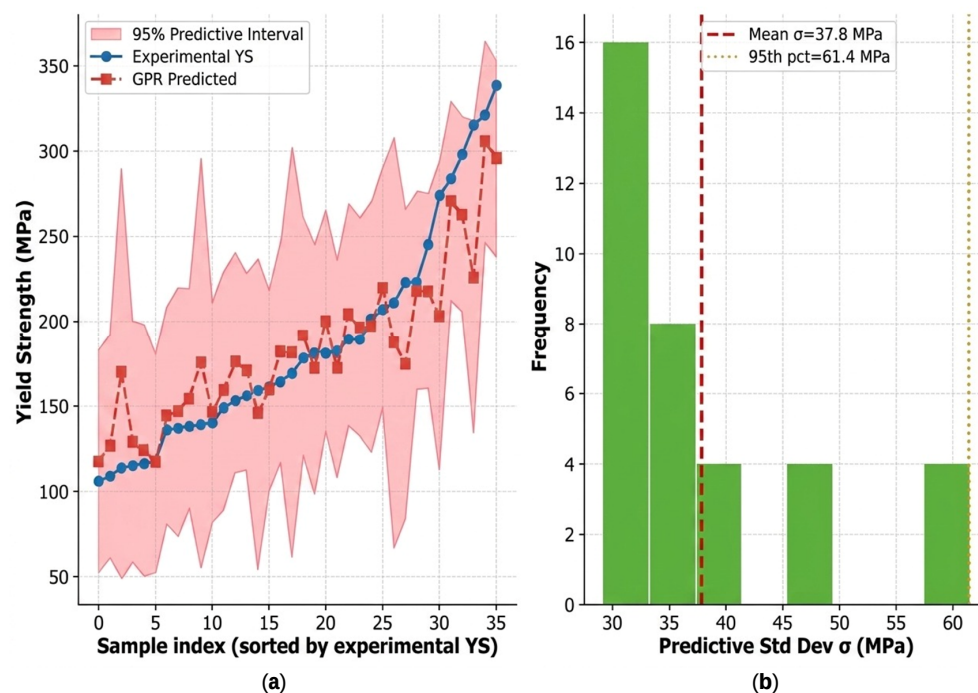


Figure 6. GPR uncertainty analysis for yield strength under LOGO cross-validation. (a) Experimental and predicted YS with 95% predictive intervals ($\hat{y} \pm 1.96\sigma$) for 36 out-of-fold predictions. (b) Distribution of predictive standard deviation (σ); dashed and dotted lines indicate the mean (37.8 MPa) and 95th percentile (61.4 MPa), respectively.

3.6. Per-Grade Generalization Analysis

The per-grade leave-one-grade-out (LOGO) results provide further insight into the model generalization across individual alloy grades (Figure 7). For UTS, both Random Forest (RF) and Gradient Boosting (GB) achieved R^2 values above 0.90 for most alloy grades, indicating consistent predictive performance. Lower accuracy was observed for the MDE and MDF grades, which exhibited the largest differences in Cr content and austenite grain size (AGS) compared with the other alloys. This variation in prediction accuracy suggests that the model performance is influenced by the coverage of the available training data. When a withheld alloy grade is well represented by the compositional range of the training grades, prediction primarily involves interpolation, whereas grades near the boundaries of the available design space may require greater extrapolation and consequently show reduced accuracy. This behavior reflects limited training data coverage near the boundaries of the investigated compositional space rather than evidence of model overfitting. These results indicate that expanding the dataset with additional alloy grades, particularly those spanning intermediate compositions and microstructures, could further improve the model robustness and generalization across the 5Cr-0.5Mo alloy system.

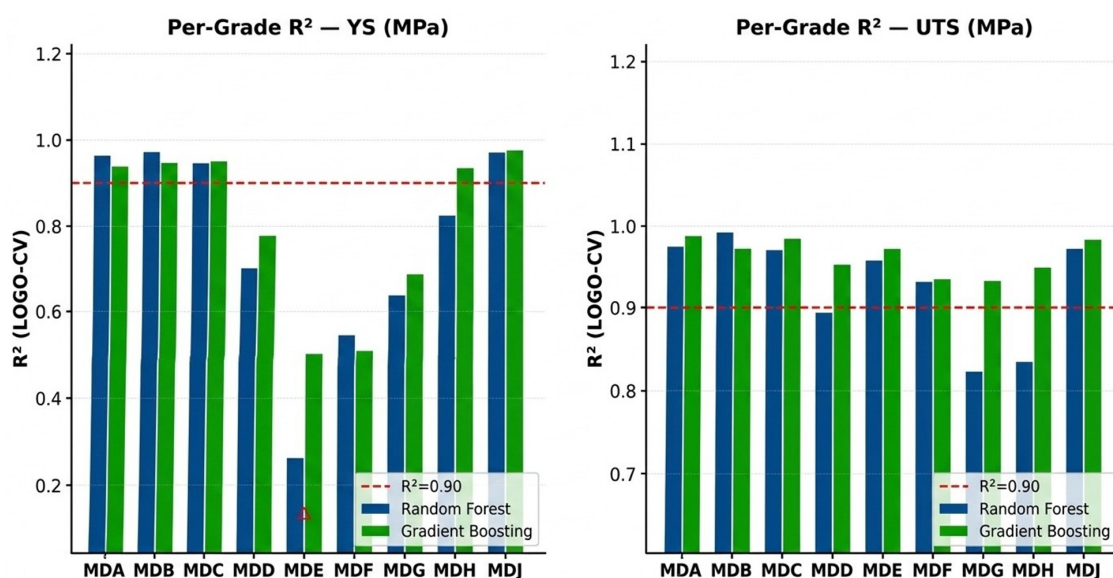


Figure 7. Per-alloy-grade Leave-One-Grade-Out (LOGO) cross-validation R^2 values for yield strength (YS, (left)) and ultimate tensile strength (UTS, (right)) obtained using Random Forest (RF) and Gradient Boosting (GB). Each pair of bars represents the performance when one alloy grade was used as a test set. The dashed red line indicates the reference performance level ($R^2 = 0.90$), and the warning symbol identifies grades with R^2 values < 0.50 .

3.7. Methodological Comparison with the Previous ANN Study

Figure 8 summarizes the LOGO R^2 and MAE performance across all three models and four target properties. For methodological context, the present results are discussed alongside those of the ANN model reported by Ishtiaq et al. [14], which was developed from the original version of the same dataset using a different data processing and validation strategy. The ANN achieved adjusted R^2 values of 0.997 for the training set and 0.993 for the testing set for YS, based on a random 81:18 train-test split applied to the 238 records. As described in Section 2.1, these records correspond to 36 unique alloy grade-temperature combinations with multiple identical entries for each combination. Under a random train-test split, identical observations may appear in both the training and test sets, potentially leading to optimistic estimates of predictive performance. In contrast, this study removed duplicate records and evaluated the model performance using Leave-One-Grade-Out

(LOGO) cross-validation, which provides a more rigorous assessment of the prediction for unseen alloy grades. Under the LOGO protocol, Gradient Boosting achieved R^2 values of 0.964 and 0.937 for UTS and RA, respectively, whereas random forest achieved 0.929 and 0.934, respectively. Because the two studies used fundamentally different data processing and validation protocols, their reported R^2 values are not directly comparable.

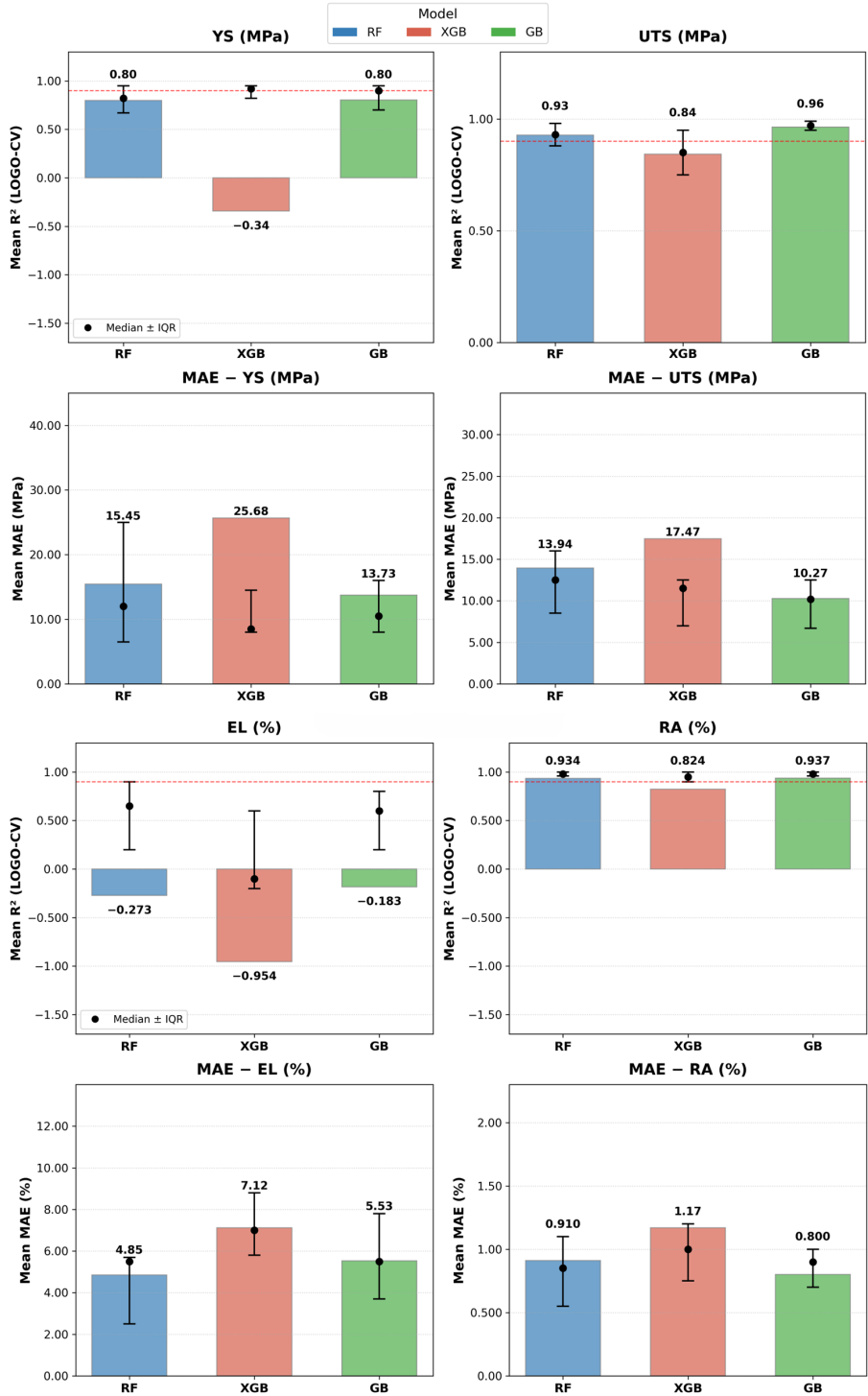


Figure 8. Comparison of model performance under Leave-One-Grade-Out (LOGO) cross-validation. The top row shows the mean R^2 values, and the bottom row shows the mean MAE values for the four target mechanical properties. The red dashed horizontal line indicates the $R^2 = 0.90$ reference threshold. The error bars represent the interquartile range (IQR). RF = Random Forest; XGB = XGBoost; GB = Gradient Boosting.

The ANN results were obtained using a random train–test split on the original 238 records, whereas the present results were obtained using LOGO cross-validation on 36 unique deduplicated alloy grade–temperature observations. Therefore, differences in the reported performance may reflect both the validation strategy and the effective dataset used for model development and should not be interpreted as evidence of intrinsic algorithmic superiority or inferiority of the model. A direct comparison of ANN under the same LOGO protocol was not performed because the original ANN implementation and complete training procedure were not publicly available. Consequently, the LOGO performance of previously reported ANN remains unknown. A direct comparison of ANN and ensemble models using the same deduplicated dataset and LOGO protocol will be valuable in future studies. In addition to prediction accuracy, the current analysis provides several additional contributions. SHAP analysis provides exploratory information on the relative contributions of input variables, with the observed associations interpreted cautiously in the context of established metallurgies. Gaussian Process Regression (GPR) provides complementary uncertainty information through 95% predictive intervals for YS, although the relatively wide intervals indicate that this proof-of-concept uncertainty analysis should be interpreted cautiously. In addition, the curated dataset and explicitly described LOGO evaluation protocol improve the transparency and reproducibility of future machine learning studies based on this dataset. Physically motivated feature engineering was also considered but not pursued in the present analysis. Interaction terms, such as $\text{Temp} \times \text{Cr}$, could potentially represent temperature-dependent compositional effects, whereas ratio descriptors, such as Cr/Mo , may provide additional information on the combined variation of these alloying elements. However, the present dataset contains only 36 unique observations with 14 existing input features, and introducing additional engineered descriptors would further increase the feature-to-sample ratio. Given the narrow compositional ranges and small sample sizes, the independent contributions and generalizability of these engineered features could not be evaluated reliably. Therefore, the present exploratory analysis retained the original experimentally reported descriptors. Future studies using larger and more compositionally diverse datasets should systematically evaluate the physically motivated interactions and ratio features. The applicability of the trained models is confined to the compositional and temperature domains represented by the present 5Cr-0.5Mo dataset. In particular, the narrow ranges of major alloying elements, including Cr (4.61–5.02 wt. %) and Mo (0.49–0.52 wt. %), do not support extrapolation to substantially different Cr–Mo steel compositions. Broader applications would require larger and more compositionally diverse datasets encompassing additional Cr–Mo grades, wider alloying-element ranges, processing and service conditions, and relevant microstructural descriptors, where available. Thus, the present models should be regarded as exploratory models for the investigated material domain rather than general alloy-design tools.

4. Conclusions

This study presents an exploratory interpretable machine learning case study for predicting the tensile properties of 5Cr-0.5Mo steel using 36 unique alloy grade–temperature combinations available in a published dataset. The findings provide insights into the relationships between the composition, temperature, and tensile properties within the investigated 5Cr-0.5Mo steel dataset and provide a basis for future validation using larger and more compositionally diverse datasets. The main conclusions are as follows.

- Gradient Boosting and Random Forest showed the strongest predictive performance for UTS and RA. Based on the primary pooled OOF evaluation, Gradient Boosting achieved R^2 values of 0.9698 and 0.9570 for UTS and RA, respectively, whereas Random

Forest achieved 0.9406 and 0.9488. These results indicate promising performance across the nine closely related alloy grades represented in the present dataset.

- SHAP analysis showed the strongest association with the test temperature across all four predicted mechanical properties, with mean $|SHAP|$ values of 35.80 for YS and 56.07 for UTS. Among the compositional and microstructural descriptors, the Cr content and austenite grain size (AGS) showed comparatively strong contributions to the predicted strength. Owing to feature correlations and the limited dataset size, these SHAP patterns should be regarded as exploratory associations rather than confirmed metallurgical mechanisms.
- Gaussian Process Regression (GPR), evaluated as a proof-of-concept for YS, produced 95% predictive intervals with 100% empirical coverage under LOGO cross-validation. However, the relatively large mean interval width (147.9 MPa; normalized width = 0.64) indicates conservative uncertainty estimates, and complete empirical coverage alone should not be interpreted as evidence of a tight calibration.
- Elongation (EL) remained difficult to predict under the LOGO framework, indicating that the available compositional and processing variables do not fully describe the microstructural factors that control ductility. Future models may benefit from incorporating descriptors such as dislocation density, precipitate characteristics, and grain boundary features into the model.
- Although the 36 observations represent the complete set of unique alloy grade–temperature combinations available in the published dataset, the narrow compositional ranges restrict the applicability of the trained models to the investigated 5Cr-0.5Mo domain. The extension to broader Cr–Mo steel systems will require larger and more compositionally diverse datasets covering additional alloy grades, processing conditions, and relevant microstructural descriptors.

Overall, this exploratory study demonstrates how prediction, model interpretation, and uncertainty analysis can be integrated into a transparent ML workflow for a small 5Cr-0.5Mo steel dataset. Further validation on larger and more compositionally diverse datasets is required before extending this approach to broader Cr–Mo steel systems or alloy-design applications.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/met16090995/s1>, Supplementary Data S1: Dataset Preparation Summary; Alloy Grade–Temperature Matrix; Dataset Used in This Study.

Author Contributions: Conceptualization, S.T.; methodology, H.A. and S.T.; software, J.L. and S.T.; validation, J.L. and H.A.; formal analysis and investigation, H.A., S.T. and N.P.; resources, N.P. and S.T.; data curation, J.L. and S.T.; writing—original draft preparation, H.A. and S.T.; writing—review and editing, S.T. and N.P.; visualization, J.L. and N.P.; supervision, S.T. and N.P.; project administration, N.P.; funding acquisition, N.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Gyeongbuk ANCHOR system through the Gyeongbuk ANCHOR CENTER, funded by the Ministry of Education (MOE) and the Gyeongsangbuk-do, Republic of Korea (2026-ANCHOR-15-115).

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Materials. Further inquiries can be directed to the corresponding authors.

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

References

1. Viswanathan, R. *Damage Mechanisms and Life Assessment of High-Temperature Components*; ASM International: Materials Park, OH, USA, 1989.
2. Lyu, Z.; Kannan, R.; Saini, N.; Li, J.; Li, L. Degradation of a Cr-Mo Steel by Carbide Precipitation over Long-Term Service. *Mater. Sci. Eng. A* **2024**, *898*, 146354. [\[CrossRef\]](#)
3. Saucedo-Muñoz, M.L.; Lopez-Hirata, V.M.; Dorantes-Rosales, H.J.; Villegas-Cardenas, J.D.; Rivas-Lopez, D.I.; Beltran-Zuñiga, M.; Ferreira-Palma, C.; Moreno-Palmerin, J. Phase Transformations of 5Cr-0.5Mo-0.1C Steel after Heat Treatment and Isothermal Exposure. *Metals* **2022**, *12*, 1378. [\[CrossRef\]](#)
4. Narayana, P.L.; Lee, S.W.; Park, C.H.; Yeom, J.-T.; Hong, J.-K.; Maurya, A.K.; Reddy, N.S. Modeling High-Temperature Mechanical Properties of Austenitic Stainless Steels by Neural Networks. *Comput. Mater. Sci.* **2020**, *179*, 109617. [\[CrossRef\]](#)
5. Wang, Y.; Wu, X.; Li, X.; Xie, Z.; Liu, R.; Liu, W.; Zhang, Y.; Xu, Y.; Liu, C. Prediction and Analysis of Tensile Properties of Austenitic Stainless Steel Using Artificial Neural Network. *Metals* **2020**, *10*, 234. [\[CrossRef\]](#)
6. Chai, M.; He, Y.; Wang, J.; Wu, Z.; Lei, B. Predicting Creep Life of CrMo Pressure Vessel Steel Using Machine Learning Models with Optimal Feature Subset Selection. *Int. J. Press. Vessels Pip.* **2024**, *212*, 105349. [\[CrossRef\]](#)
7. Mamun, O.; Wenzlick, M.; Hawk, J.; Devanathan, R. A Machine Learning Aided Interpretable Model for Rupture Strength Prediction in Fe-Based Martensitic and Austenitic Alloys. *Sci. Rep.* **2021**, *11*, 5466. [\[CrossRef\]](#)
8. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
9. Friedman, J. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2000**, *29*, 1189–1232. [\[CrossRef\]](#)
10. 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: New York, NY, USA, 2016; pp. 785–794.
11. Ishtiaq, M.; Hong, H.J.; Reddy, N.G.S. Data-Driven Quantification of Temperature-Induced Mechanical Property Variations in 5Cr-0.5Mo Steel Using Artificial Neural Networks. *Processes* **2026**, *14*, 2208. [\[CrossRef\]](#)
12. Ishtiaq, M.; Tariq, H.M.R.; Reddy, D.Y.C.; Kang, S.-G.; Reddy, N.G.S. Prediction of Creep Rupture Life of 5Cr-0.5Mo Steel Using Machine Learning Models. *Metals* **2025**, *15*, 288. [\[CrossRef\]](#)
13. Tiwari, S.; Dash, K.; Heo, S.; Park, N.; Reddy, N.G.S. Machine Learning-Driven Prediction of Microstructural Evolution and Mechanical Properties in Heat-Treated Steels Using Gradient Boosting. *Crystals* **2026**, *16*, 61. [\[CrossRef\]](#)
14. Ishtiaq, M.; Tiwari, S.; Nagamani, M.; Kang, S.-G.; Reddy, N.G.S. Data-Driven ANN-Based Predictive Modeling of Mechanical Properties of 5Cr-0.5Mo Steel: Impact of Composition and Service Temperature. *Crystals* **2025**, *15*, 213. [\[CrossRef\]](#)
15. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
16. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [\[CrossRef\]](#)
17. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; The MIT Press: Cambridge, MA, USA, 2005.
18. Schulz, E.; Speekenbrink, M.; Krause, A. A Tutorial on Gaussian Process Regression: Modelling, Exploring, and Exploiting Functions. *J. Math. Psychol.* **2018**, *85*, 1–16. [\[CrossRef\]](#)
19. Sarfarazi, S.; Shamass, R.; Rabi, M.; Abarkan, I.; Ferreira, F.P.V.; Tsavdaridis, K.D. Inverse Machine Learning for the Design of Perforated Beams: Parent Section and Material Prediction. *Eng. Appl. Artif. Intell.* **2026**, *164*, 113275. [\[CrossRef\]](#)
20. Sarfarazi, S.; Ferreira, F.P.V.; Shamass, R.; Tsavdaridis, K.D.; Abarkan, I.; Rabi, M. An Intelligent Predictive-Inverse Design Framework for Cellular Steel Beams with Integrated Robustness, Sustainability Metrics, and Automated CAD Generation. *Adv. Eng. Softw.* **2026**, *221*, 104262. [\[CrossRef\]](#)
21. ISO 9513:2012; Metallic Materials—Calibration of Extensometer Systems Used in Uniaxial Testing. International Organization for Standardization: Geneva, Switzerland, 2012.
22. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.