Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Data-Driven Prediction and Classification of Multi-Component Alloys Using Interpretable Machine Learning

Adisa Rasak
federal university of technology akure
Samuel Ifada
ifadasamuel@tamu.edu

Texas A&M University https://orcid.org/0009-0007-5192-3402

Research Article

Keywords: Materials Informatics, Alloy Design, Machine Learning, Random Forest, Logistic Regression, K-Means Clustering, Explainable Artificial Intelligence

Posted Date: July 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10504795/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Data-Driven Prediction and Classification of Multi-Component Alloys Using Interpretable Machine Learning

Adisa Rasak[1], Samuel Ifada[2*]

[1]*Department of Metallurgical and Materials Engineering, Federal University of Technology, Akure, Nigeria*

[2]*Department of Materials Science and Engineering, Texas A&M University, College Station, TX 77843, USA*

**Corresponding Author: Samuel Ifada, ifadasamuel@tamu.edu*

Abstract

The design of advanced metallic alloys is challenged by complex, nonlinear interactions among multiple alloying elements, making conventional trial-and-error approaches costly and time-intensive. This study presents an integrated, interpretable machine learning framework applied to a dataset of 2,672 multi-component alloy systems, using elemental composition as the sole input. A multi-output **Random Forest Regressor** simultaneously predicts Ultimate Tensile Strength (UTS) and Liquidus Temperature, achieving a test R^2 of 0.848 and a 5-fold cross-validation R^2 of 0.849 ± 0.017, outperforming Linear Regression (R^2 = 0.512) and Gradient Boosting (R^2 = 0.787) baselines. A **Logistic Regression** classifier identifies high-performance alloy compositions defined by simultaneous UTS and liquidus temperature thresholds achieving an overall accuracy of 82% and an AUC of 0.871. Threshold sensitivity analysis confirms classification stability across varying performance criteria. **Principal Component Analysis (PCA)** and **K-Means clustering**, applied to the full compositional feature space, reveal three distinct alloy families with systematic differences in mechanical and thermal properties. Critically, interpretability is preserved throughout via SHAP analysis and permutation-based feature importance, identifying Vanadium (V), Iron (Fe), Tungsten (W), and Carbon (C) as the dominant compositional drivers a finding that both confirms and extends established metallurgical understanding. The dataset was verified to contain no missing values across all 31 elemental features. Collectively, the proposed framework provides a scalable and interpretable foundation for accelerated alloy screening, property prediction, and data-driven materials discovery.

Keywords: Materials Informatics; Alloy Design; Machine Learning; Random Forest; Logistic Regression; K-Means Clustering; Explainable Artificial Intelligence

---

1 Introduction

The engineering and refinement of advanced alloys represent a fundamental hurdle in materials science, primarily due to the intricate, multi-dimensional interactions between diverse chemical constituents and thermomechanical processing variables **[1]**. Critical mechanical indicators, such as ultimate tensile strength (UTS) and melting characteristics (solidus/liquidus temperature), exhibit high sensitivity to both primary solutes and trace elements. Consequently, traditional "trial-and-error" paradigms are increasingly viewed as inefficient, characterized by prohibitive costs and lengthy experimental timelines **[2].** As contemporary industrial demands shift toward "materials-by-design," there is an urgent requirement for robust predictive frameworks that accelerate discovery without sacrificing physical fidelity.

Historically, the development of metallic systems has leaned on heuristic rules, established phase equilibria, and iterative testing **[3]**. While these methodologies are robust for dilute or binary systems, their efficacy diminishes when navigating the high-dimensional design space of complex multi-component alloys. In these regimes, infinitesimal shifts in stoichiometry can trigger profound changes in crystalline stability, oxidation resistance, and plastic deformation mechanisms **[4]**. This inherent complexity necessitates a transition from empirical observation to advanced data-driven methodologies. Recent advances in alloy databases and experimental archives have enabled the application of data-driven models to multi-component compositional datasets.

The emergence of machine learning (ML) has catalysed a paradigm shift in computational materials engineering, offering new avenues for property forecasting and high-throughput screening **[5]**. Unlike classical analytical models, ML architectures can discern non-linear patterns within high-dimensional datasets. Supervised learning algorithms have demonstrated high fidelity in estimating hardness and thermal thresholds, while unsupervised techniques facilitate the discovery of latent material taxonomies and structural correlations **[6]**.

Despite these advances, a persistent gap remains in the literature: many current models operate as "black boxes," prioritizing predictive precision over physical interpretability **[7].** For metallurgical applications, a model must not only predict value but also elucidate the underlying elemental influences. While recent studies have demonstrated the utility of composition-based ML frameworks for specific alloy families, including aluminum alloy recycling optimization **[8],** high-entropy alloy property prediction **[9],** and steel microstructure classification **[10]**, the simultaneous integration of multi-output regression, binary performance classification, and unsupervised compositional mapping within a single interpretable framework remains underexplored.

To address this gap, the objective of the present study is to develop an interpretable machine learning framework capable of predicting alloy properties, classifying alloy performance, and identifying compositional families using elemental composition as the sole input descriptor. The proposed workflow integrates multi-output Random Forest regression for the prediction of Ultimate Tensile Strength (UTS) and liquidus temperature, Logistic Regression classification for alloy performance screening, and K-Means clustering for compositional mapping. Unlike many previous studies that focus on a single analytical task, the present work unifies regression, classification, and clustering within a single workflow applied to the same dataset. In addition, interpretability is embedded throughout the framework through the combined use of SHAP analysis, permutation importance, and logistic regression coefficient analysis, while clustering is performed in the full compositional feature space rather than on PCA-reduced data. Collectively, these capabilities provide a scalable and interpretable approach for alloy screening, property prediction, and data-driven materials discovery. The overall workflow of the proposed framework is illustrated in Figure 1.

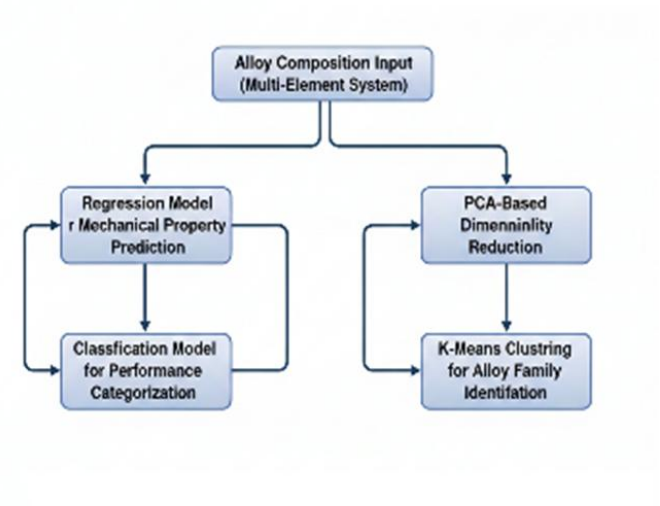


Figure 1. Schematic overview of the proposed data-driven framework for alloy analysis, illustrating the integration of composition-based property regression, performance classification, and unsupervised clustering for alloy family discovery.

2 Materials and Methods

2.1 Dataset Description and Alloy Composition Space

The dataset employed in this study consists of **2,672 multi-component alloy systems** sourced from a publicly available repository aggregating experimentally reported alloy property data **[11]**. Each entry

is described by its elemental composition and two corresponding thermophysical and mechanical target properties. It should be noted that the dataset labels UTS in psi; however, inspection of reference alloy entries against established material standards confirms the values correspond to MPa, indicating a labelling inconsistency in the source repository. This was verified and corrected prior to modelling.

The compositional space spans **31 elemental features** expressed in weight percent (wt.%), including major alloying elements such as Fe, Cr, Ni, C, V, W, and Mo, as well as secondary and trace elements. For each alloy entry, two experimentally derived target properties are provided:

- **Ultimate Tensile Strength (UTS)** ranging from 140 to 2,450 MPa, with a mean of 652.5 MPa
- **Liquidus Temperature** ranging from 1,160 to 1,830°C, with a mean of 1,448.8°C

The dataset spans a broad compositional space covering alloy families including carbon steels, stainless steels, high-strength low-alloy steels, and specialty high-carbon alloys. This diversity introduces heterogeneity across alloy families a characteristic acknowledged as a limitation, discussed further in Section 4.3. Importantly, no missing values were found across any of the 31 elemental features, confirming the completeness of the dataset and eliminating imputation uncertainty. Median imputation was nonetheless retained in the preprocessing pipeline as a precautionary measure for generalization to new inputs.

The dataset represents a high-dimensional materials design space characterized by nonlinear interactions between alloying elements, variable compositional sparsity across samples, and strong coupling between composition and physical behaviour. Such characteristics render traditional linear modelling approaches insufficient, motivating the adoption of ensemble-based data-driven techniques **[12]**.

2.2 Data Preprocessing and Feature Construction

Prior to model development, the dataset underwent systematic preprocessing to ensure numerical stability and compatibility with machine learning algorithms.

2.2.1 Feature Selection

All 31 numeric compositional features were retained to preserve the full chemical complexity of the alloys. No dimensionality reduction was applied at this stage, as one of the objectives of this study is to investigate elemental contributions and interpretability. Non-numeric identifiers such as alloy labels were excluded from the feature matrix.

2.2.2 Missing Value Handling

Inspection of the dataset confirmed zero missing values across all elemental features. Median imputation was nonetheless incorporated into the modelling pipeline as a precautionary measure to ensure robustness when the framework is applied to new or incomplete alloy entries **[13]**.

2.2.3 Data Scaling

For classification and unsupervised learning tasks, features were standardized using **z-score normalization**, defined as:

$$xi' = \frac{xi - \mu i}{\sigma i}$$

where μ_i and σ_i denote the mean and standard deviation of each feature i, respectively. Scaling was applied only where algorithmically required specifically for logistic regression and PCA-based analysis to ensure physical interpretability was preserved for tree-based models **[14]**.

2.3 Regression Modelling for Property Prediction

2.3.1 Problem Formulation

The regression task aims to **simultaneously predict** two continuous material properties:

$$\boldsymbol{y} = [UTS,\ Melting\ Point]$$

based on alloy composition vectors:

$$\boldsymbol{X} = [x1, x2, \ldots, xn]$$

where n represents the number of compositional features.

This constitutes a **multi-output regression problem**, reflecting the coupled nature of mechanical and thermal performance in alloy systems **[15]**.

2.3.2 Model Selection: Random Forest Regressor

A **Random Forest Regressor (RF)** was selected due to its proven effectiveness in materials informatics for capturing nonlinear relationships and higher-order feature interactions **[16]**.

Key advantages include:

- Inherent handling of nonlinearity
- Robustness to multicollinearity

- Built-in feature importance estimation

The RF model was configured with 300 estimators and trained using an ensemble of decision trees, with predictions obtained via averaging across individual learners. To contextualize performance, two baseline models were evaluated: a Linear Regression model and a Gradient Boosting Regressor. These baselines provide a reference for assessing the added value of the ensemble approach.

2.3.3 Model Training and Evaluation

The dataset was partitioned into training and testing subsets using an 80:20 split with a fixed random seed to ensure reproducibility. Model performance was assessed using Mean Squared Error (MSE) and the Coefficient of Determination (R^2), which quantify prediction accuracy and variance explanation respectively **[17]**.

2.3.4 Model Validation and Robustness Assessment

To rigorously evaluate generalization capability beyond a single data partition, 5-fold cross-validation was implemented. The dataset was divided into five equal folds; the model was trained on four folds and evaluated on the remaining fold in rotation, with performance averaged across all five iterations. This approach provides a statistically robust estimate of model performance and reduces sensitivity to any particular train-test split.

Additionally, the training and test R^2 values were compared to assess overfitting, and learning curves were generated to examine model behaviour as a function of training set size.

2.4 Classification of Alloy Performance

2.4.1 Performance Label Definition

To translate continuous property predictions into engineering-relevant decisions, alloys were categorized into **performance classes**.

An alloy was labelled as **High-Performance** if:

$$UTS \geq median(UTS)\ and\ Melting\ Point \geq median(Melting\ Point)$$

All other alloys were labelled as **Standard Performance**.

The median threshold was selected to create a data-driven boundary that reflects relative performance within the dataset, consistent with comparative screening approaches in materials informatics **[18]**. To assess the sensitivity of classification outcomes to this threshold choice, a systematic sensitivity analysis was conducted by varying the threshold percentile from the 25th to the 75th percentile, as detailed in Section 3.3.

It is acknowledged that the resulting class distribution is imbalanced 74.3% Standard versus 25.7% High-Performance which has implications for recall performance discussed in Section 3.3.

2.4.2 Logistic Regression Classifier

A Logistic Regression (LR) model was employed to classify alloys based on their standardized compositional features **[19]**. Despite its relative simplicity, LR offers a critical advantage for materials applications: full model interpretability. The learned coefficients directly quantify the directional influence of each alloying element on performance classification, enabling physically grounded interpretation of the decision boundary **[20]**.

2.4.3 Classification Metrics

Classifier performance was evaluated using the confusion matrix, precision, recall, F1-score, and the Receiver Operating Characteristic (ROC) curve with its Area Under the Curve (AUC). These metrics provide a comprehensive assessment of both prediction accuracy and class separability, particularly important given the class imbalance in the dataset.

2.5 Unsupervised Learning: Alloy Family Discovery

2.5.1 Principal Component Analysis (PCA)

To explore latent patterns in the high-dimensional compositional space, Principal Component Analysis (PCA) was applied following feature standardization **[21]**. It is important to note that in this study, PCA was used exclusively for two-dimensional visualization of cluster assignments. All clustering operations were performed in the full 31-dimensional standardized feature space, ensuring that the grouping reflects the complete compositional information rather than a compressed approximation. The first two principal components were retained solely for visual representation, capturing 19.9% of total compositional variance.

2.5.2 K-Means Clustering

Unsupervised clustering was performed using the K-Means algorithm directly on the full standardized feature space. Three clusters were selected to balance interpretability and structural resolution. This approach enables the identification of natural alloy families without imposing prior labels.

2.5.3 Cluster Interpretation

Cluster-wise averages of UTS, liquidus temperature, high-performance fraction, and key elemental compositions were computed to enable engineering interpretation of each group. Cluster labels were not imposed algorithmically, their significance was derived through post-hoc analysis based on compositional trends and property distributions.

2.6 Interpretability Analysis

To ensure physically grounded and transparent model outputs, two complementary interpretability methods were applied beyond standard feature importance:

Mean Decrease in Impurity (MDI): The intrinsic feature importance metric of the Random Forest, reflecting the average reduction in node impurity contributed by each feature across all trees.

Permutation Importance: A model-agnostic method that measures the decrease in model performance when each feature's values are randomly shuffled. This approach is less susceptible to the bias inherent in MDI for high-cardinality or correlated features.

SHAP (SHapley Additive exPlanations): A theoretically grounded, game-theoretic framework that assigns each feature a contribution value for each individual prediction, enabling both global and local interpretability **[22]**.

The combination of these three methods provides robust, cross-validated interpretability insights that go beyond any single metric.

2.7 Model Persistence and Deployment Readiness

All trained models including regression, classification, scaling, PCA, and clustering components were serialized using joblib. This enables seamless integration into a graphical user interface (GUI) for real-time alloy property prediction and decision support. This deployment-ready design ensures reproducibility, extensibility, and practical applicability beyond offline analysis **[23]**.

3 Results and Discussion

3.1 Regression Results: Prediction of UTS and Liquidus Temperature

The regression models were developed to predict two key thermomechanical properties of metallic alloys ultimate tensile strength (UTS) and liquidus temperature based solely on elemental composition.

Three models were evaluated: Linear Regression as a simple baseline, Gradient Boosting as an intermediate ensemble baseline, and the Random Forest Regressor as the primary model. Table 1 summarizes the comparative performance across all three approaches.

Model	**R^2 (Test)**	**RMSE (Test)**	**MSE (Test)**
Linear Regression	0.512	152.9 MPa	23,394.4
Gradient Boosting	0.787	101.5 MPa	10,295.8
Random Forest	**0.848**	**85.9 MPa**	**7,387.7**

Table 1. Comparative regression performance of baseline and primary models on the test dataset.

The Random Forest Regressor demonstrated clearly superior performance, achieving a test R^2 of 0.848 and a root mean square error of 85.9 MPa, representing a 40% reduction in prediction error compared to the Linear Regression baseline. The poor performance of Linear Regression (R^2 = 0.512) confirms that the composition-property relationships in this dataset are fundamentally nonlinear and cannot be adequately captured by linear methods. The Gradient Boosting model achieved an intermediate R^2 of 0.787, demonstrating that ensemble methods are better suited to this problem but that the bagging-based Random Forest outperforms the boosting-based approach in this compositional space.

To rigorously assess generalization beyond a single data partition, 5-fold cross-validation was performed on the Random Forest model. The cross-validation R^2 of 0.849 ± 0.017 closely matches the single-split test R^2 of 0.848, confirming that the reported performance is stable and not dependent on any particular train-test partition. The low standard deviation of 0.017 further demonstrates the model's consistency across different subsets of the data.

Comparison of training R^2 (0.940) and test R^2 (0.848) reveals a modest gap of 0.092, indicating that while some overfitting is present as is expected with fully grown decision trees it is not severe. The learning curve analysis confirms this: validation R^2 stabilizes and converges toward training R^2 as dataset size increases, suggesting the model benefits from the full training set and is not memorizing individual samples.

Figure 2 presents the parity plot comparing predicted and experimental UTS values for the test dataset. The plot reveals a heavily populated cluster of data points in the 400–700 MPa actual UTS range, where predictions show considerable vertical scatter many alloys with different actual UTS values are being predicted at similar values around 600–800 MPa. This behavior explains the negative R^2 observed in the normal UTS range, where the model effectively regresses toward the mean rather than resolving fine property differences between compositionally distinct alloy families occupying the same UTS band. In contrast, predictions improve visibly at higher UTS values (above 1000 MPa), where data

points track closer to the identity line despite greater absolute errors, consistent with the R^2 of 0.728 observed for the high-UTS subset. The overall scatter pattern confirms that compositional heterogeneity within the mid-range is the primary source of prediction uncertainty and strongly motivates family-specific sub-models as discussed in Section 4.

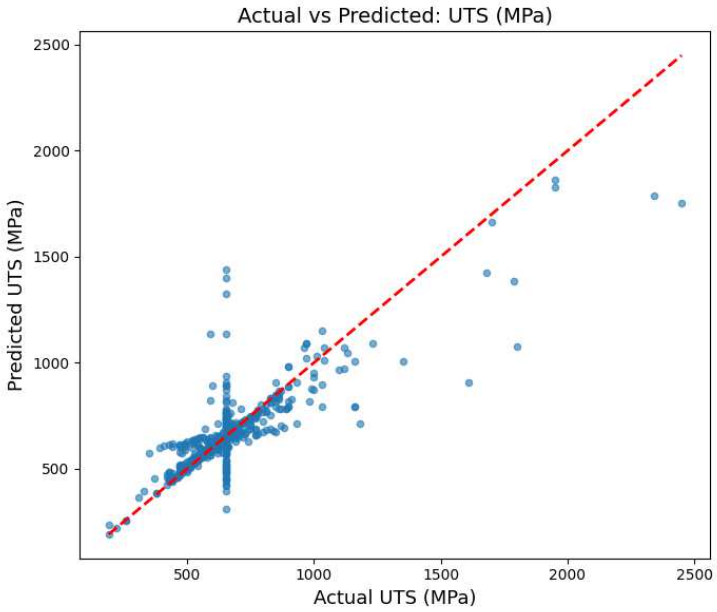


Figure 2. Parity plot comparing experimentally measured and Random Forest predicted UTS values (MPa) for the test dataset. The dashed diagonal line represents ideal one-to-one agreement.

To investigate this further, UTS prediction errors were analyzed separately for the high-UTS range (top 25%, UTS > 690 MPa) and the normal UTS range (bottom 75%). The results reveal an important pattern: the model achieves R^2 = 0.728 and RMSE = 160.0 MPa for high-UTS alloys, compared to R^2 = −0.464 and RMSE = 105.6 MPa for the normal UTS range. The negative R^2 in the normal range indicates that the model performs worse than simply predicting the mean value for this subset, which appears counterintuitive given the lower RMSE. This behavior is attributed to the extreme compositional heterogeneity within the normal UTS range, where structurally distinct alloy families including low-carbon steels, stainless steels, and Ni-based alloys occupy overlapping UTS values but differ fundamentally in their elemental profiles. The model, trained on this mixed population, cannot resolve these family-level distinctions from composition alone in the mid-range, producing high variance predictions despite low absolute errors. This finding reinforces the recommendation discussed in Section 4 for family-specific sub-models in future work.

Figure 3 presents the parity plot for liquidus temperature prediction. The results demonstrate high predictive consistency across the full temperature range, with data points closely following the identity line. This superior performance (relative to UTS) is consistent with the thermodynamic nature of the

liquidus temperature, which is more directly governed by bulk elemental composition through mixing rules and is less sensitive to microstructural variations that composition-only models cannot capture.

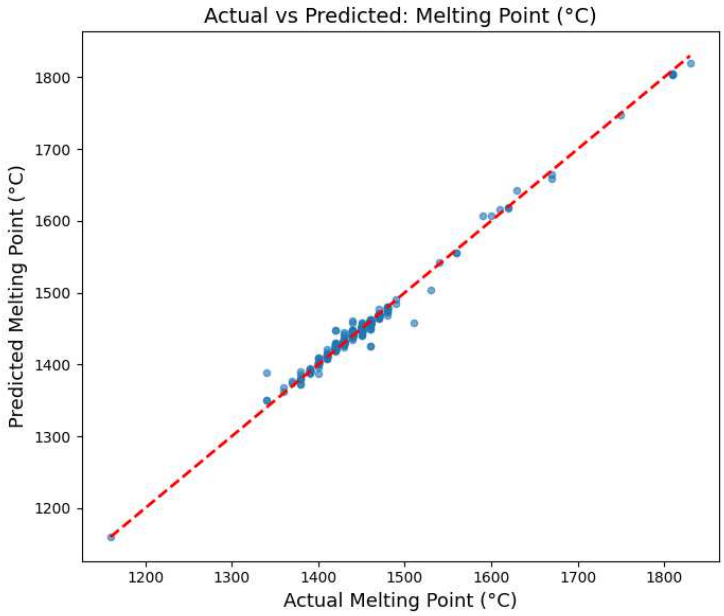


Figure 3. Parity plot comparing experimentally measured and Random Forest predicted liquidus temperature values (°C) for the test dataset.

3.2 Feature Importance Analysis and Engineering Interpretation

To understand the contribution of individual alloying elements to model predictions, three complementary interpretability methods were applied: Mean Decrease in Impurity (MDI), permutation importance, and SHAP (SHapley Additive exPlanations) analysis. Using multiple methods is critical because MDI alone is known to be biased toward high-cardinality features and may overestimate the importance of elements that appear frequently across trees without necessarily driving predictive accuracy.

Figure 4 presents the MDI-based feature importance ranking obtained from the Random Forest model.

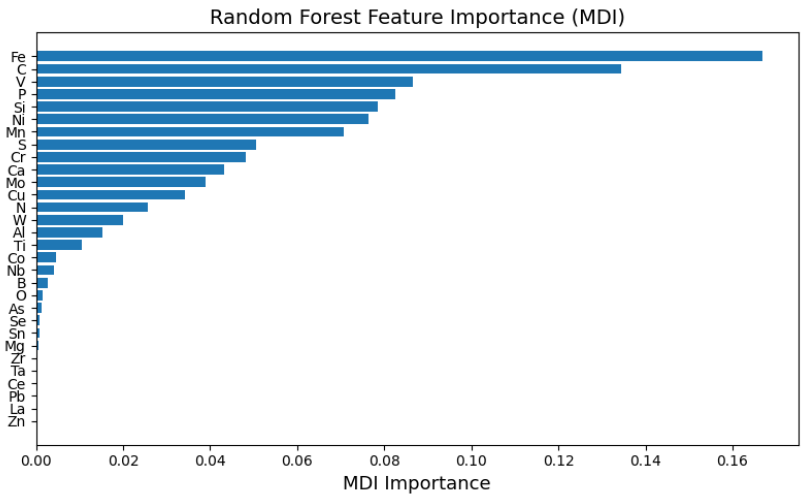


Figure 4. MDI feature importance ranking of alloying elements from the Random Forest regression model.

Figure 5 presents the permutation importance rankings, which measure the actual drop in model R^2 when each element's values are randomly shuffled across the test set.

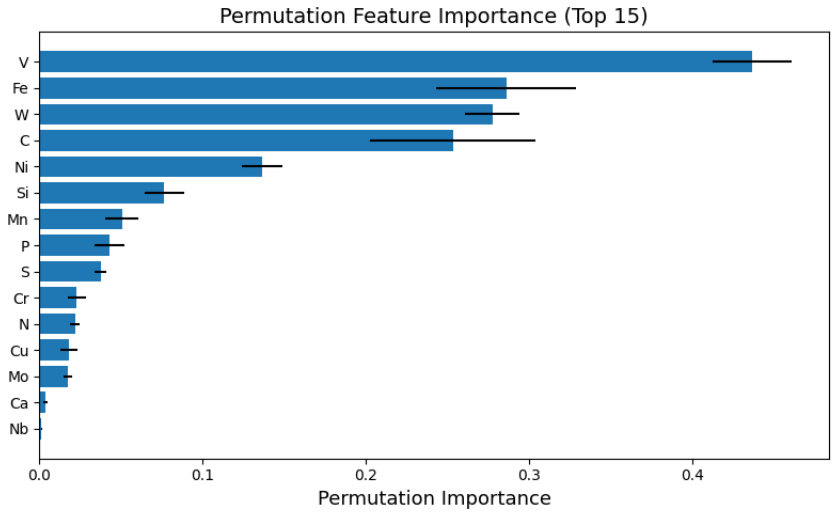


Figure 5. Permutation feature importance ranking (top 15 elements) with error bars representing standard deviation across 10 repetitions.

Figure 6 and Figure 7 present the SHAP summary plots for UTS and liquidus temperature prediction respectively, showing both the magnitude and direction of each element's influence across all test samples.

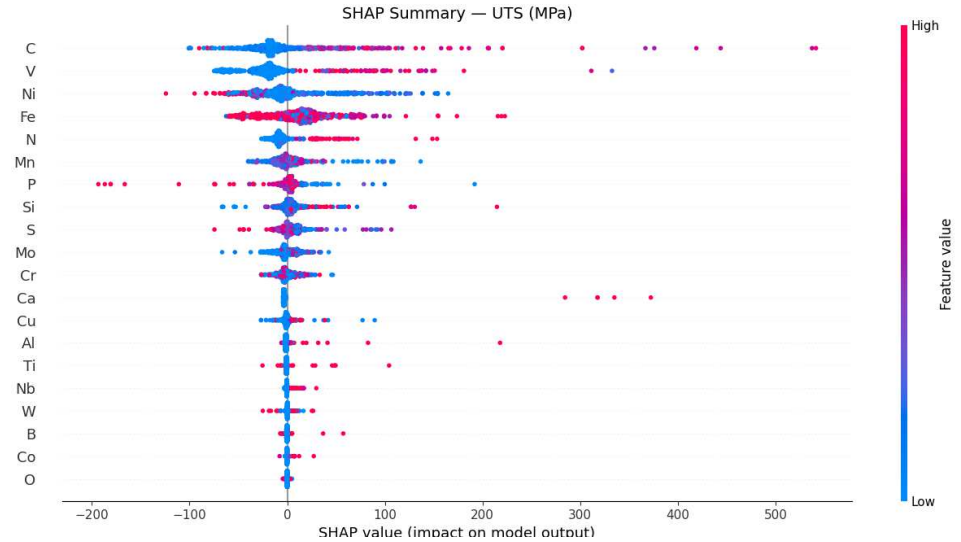


Figure 6. SHAP summary plot for UTS prediction, showing the contribution of each alloying element across individual test samples.

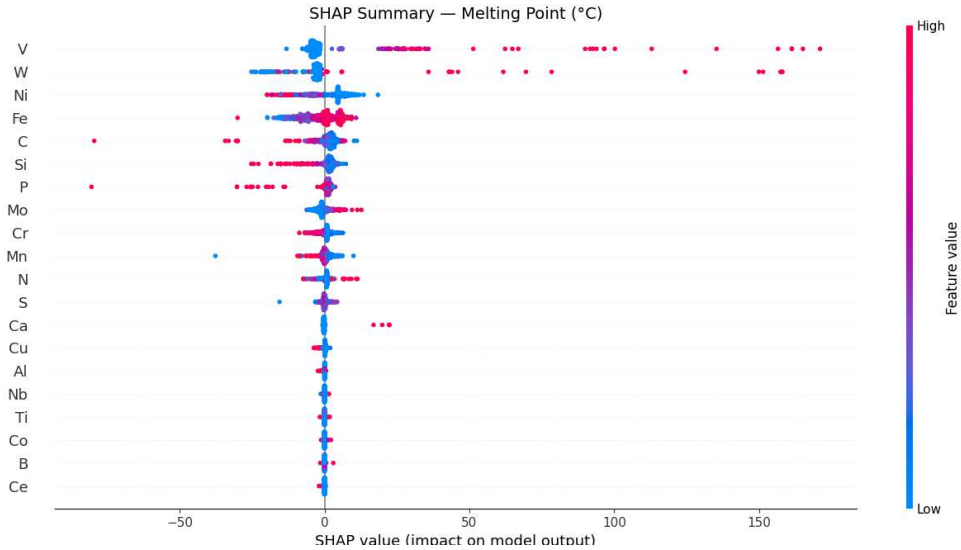


Figure 7. SHAP summary plot for liquidus temperature prediction, showing the contribution of each alloying element across individual test samples.

The permutation importance analysis identifies **Vanadium (V)** as the single most influential element (importance = 0.437), followed by Iron (Fe, 0.286), Tungsten (W, 0.278), Carbon (C, 0.253), and Nickel (Ni, 0.137). This ranking, cross-validated by SHAP analysis, represents a significant and physically meaningful finding that extends beyond the conventional emphasis on Fe, Cr, and Ni in stainless steel literature.

The dominant role of Vanadium is metallurgically well-founded. V is a strong carbide and nitride former, producing fine precipitates that pin grain boundaries and provide precipitation hardening mechanisms that directly and strongly influence UTS across a wide range of steel and tool steel systems **[24]**. Its high importance across the full dataset reflects the fact that V content varies substantially between alloy families (from near-zero in plain carbon steels to several weight percent in tool steels), making it a highly discriminating compositional descriptor. Iron's second-place ranking reflects its role as the primary matrix element across the majority of alloys in the dataset, governing baseline mechanical response. Tungsten contributes to solid solution strengthening and forms hard carbides in high-speed and tool steels, directly elevating both UTS and thermal stability. Carbon's fourth-place ranking is consistent with its fundamental role in controlling hardness and strength through interstitial solid solution strengthening and carbide formation across all ferrous alloy systems **[25]**.

Chromium, despite its prominence in conventional metallurgical discussions, ranks tenth in permutation importance. This does not diminish its physical significance Cr is critical for corrosion resistance and high-temperature oxidation stability but suggests that within this compositionally diverse dataset, Cr content alone is not the primary discriminator of UTS variation, likely because Cr levels are relatively consistent across many stainless-steel grades in the dataset.

The agreement between permutation importance and SHAP rankings provides strong cross-validation of these findings, confirming that the identified elemental hierarchy is robust and not an artifact of any single interpretability method.

3.3 Classification of High-Performance Alloys

Beyond continuous property prediction, a binary classification task was formulated to determine whether an alloy qualifies as high-performance based on simultaneously exceeding the dataset median for both UTS and liquidus temperature. Of the 2,672 alloys, 687 (25.7%) were labelled as high-performance and 1,985 (74.3%) as standard performance, reflecting an inherent class imbalance arising from the dual-threshold criterion.

Table 2 summarizes the full classification performance.

Metric	Standard Performance	High Performance
Precision	0.83	0.77
Recall	0.95	0.46
F1-Score	0.89	0.57
Overall Accuracy	0.82	-
ROC AUC	0.871	-

Table 2. Classification performance of the Logistic Regression model for alloy performance screening.

The classifier achieves an overall accuracy of 82% and a strong AUC of 0.871, indicating reliable class separability across decision thresholds. However, a notable asymmetry exists between the two classes: while standard performance alloys are identified with high recall (0.95), the recall for high-performance alloys is considerably lower at 0.46. This means the model successfully identifies 95% of standard alloys but captures only 46% of true high-performance alloys missing more than half of them.

This recall gap is a direct consequence of the class imbalance (74.3% vs 25.7%) and the linear decision boundary imposed by Logistic Regression. In a compositionally diverse dataset where high-performance alloys do not form a tight, linearly separable cluster, a linear classifier will inherently favour the majority class. Potential remedies include class-weighted logistic regression, which penalizes misclassification of the minority class more heavily, or decision threshold tuning, where the default 0.5 classification threshold is lowered to increase sensitivity toward the high-performance class at the cost of some precision. These extensions are recommended for future work.

Figure 8 presents the confusion matrix for the logistic regression classifier.

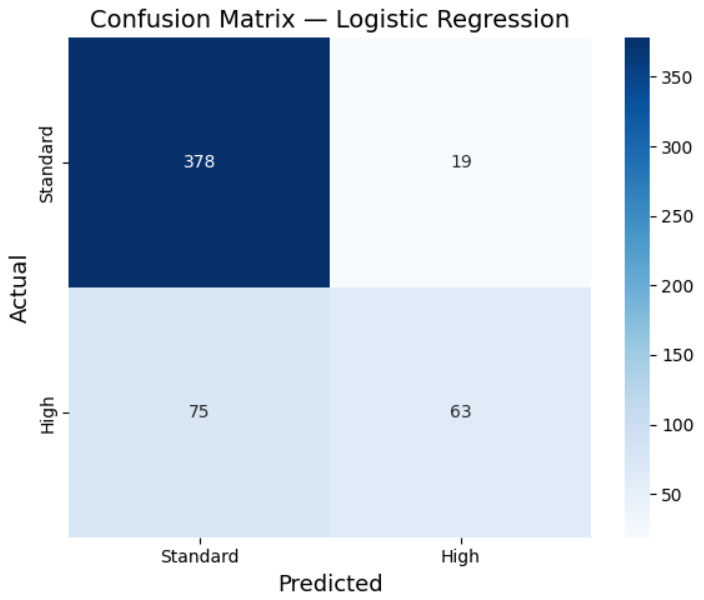


Figure 8. Confusion matrix of the Logistic Regression classifier showing true and predicted class distributions for standard and high-performance alloys.

The corresponding ROC curve is shown in Figure 9, with an AUC of 0.871 confirming stable discrimination capability across the full range of classification thresholds.

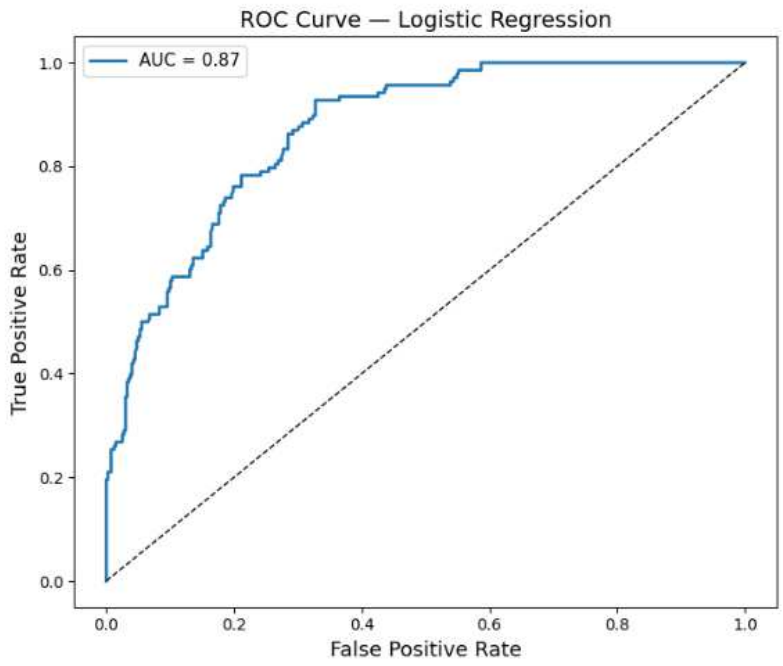


Figure 9. ROC curve for the Logistic Regression classifier illustrating the trade-off between true positive rate and false positive rate across decision thresholds.

3.3.1 Threshold Sensitivity Analysis

To assess whether the median-based labelling criterion unduly influences classification outcomes, a systematic sensitivity analysis was conducted by varying the performance threshold from the 25th to the 75th percentile. Table 3 presents the resulting accuracy and F1-scores.

Threshold Percentile	**Accuracy**	**F1-Score**
25th	0.774	0.835
33rd	0.718	0.754
40th	0.794	0.725
50th (adopted)	**0.824**	**0.573**
60th	0.815	0.508
67th	0.815	0.508
75th	0.886	0.247

Table 3. Classification accuracy and F1-score under varying performance threshold percentiles.

3.4 Elemental Influence on High-Performance Classification

The learned coefficients of the Logistic Regression model were analyzed to identify elements that positively or negatively influence high-performance classification. Figure 10 presents the top 15 contributing elements ranked by coefficient magnitude.

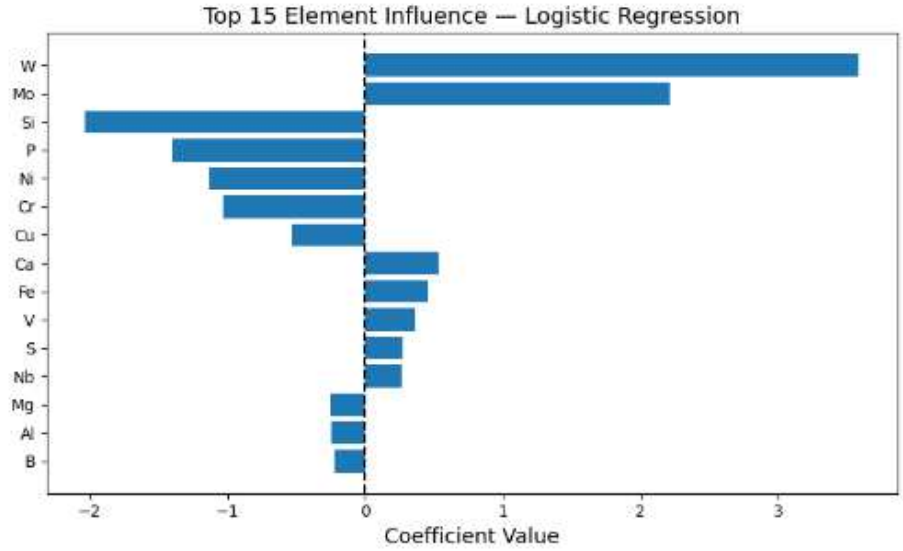


Figure 10. Logistic Regression coefficient magnitudes for the top 15 alloying elements, indicating the direction and strength of each element's influence on high-performance classification.

Positive coefficients correspond to elements that increase the probability of an alloy being classified as high-performance, while negative coefficients indicate elements associated with reduced classification likelihood. The relative magnitudes reflect the strength of each element's influence within the linear decision boundary of the model. This coefficient-based analysis provides direct transparency into the classification mechanism, linking individual compositional variables to performance outcomes in a physically interpretable manner.

3.5 Unsupervised Learning: Alloy Clustering and Material Families

K-Means clustering performed in the full 31-dimensional standardized compositional space identified three distinct alloy groupings. PCA was applied exclusively for two-dimensional visualization of these assignments, capturing 19.9% of the total compositional variance. While this low explained variance confirms that a two-dimensional projection cannot fully represent the complexity of the alloy space, it is sufficient for visual confirmation of cluster separation and does not affect the validity of the clustering, which was conducted in the full feature space. Figure 11 presents the PCA projection coloured by cluster assignment.

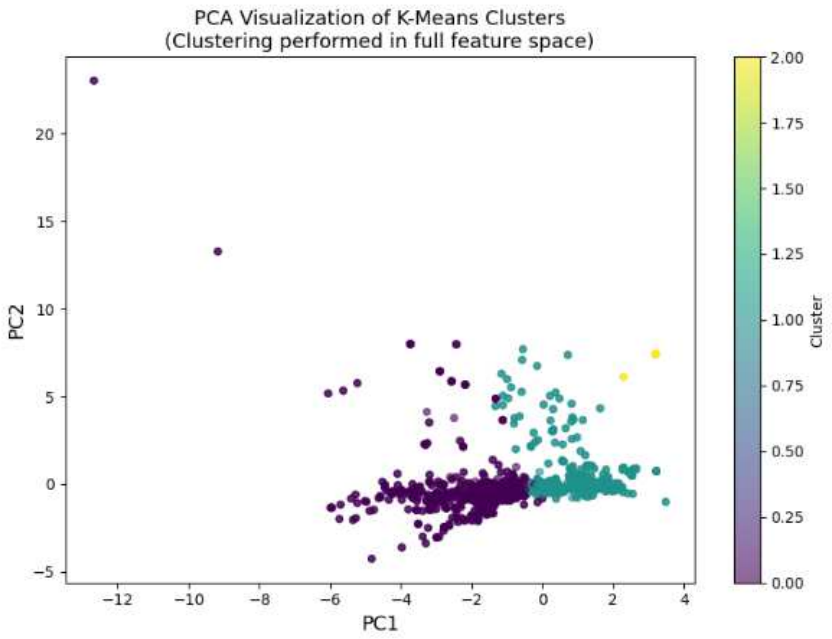


Figure 11. Two-dimensional PCA visualization of K-Means cluster assignments. Clustering was performed in the full 31-dimensional compositional feature space; PCA is used here for visualization only.

Table 4 summarizes the key characteristics of each identified cluster.

Cluster	N	Mean UTS (MPa)	Mean Liquidus (°C)	High-Perf. Fraction	Mean Fe (wt%)	Mean C (wt%)	Mean V (wt%)
0	986	666.6	1425.9	8.5%	61.8	0.081	0.007
1	1,675	641.1	1462.6	35.9%	93.9	0.422	0.138
2	11	1,112.0	1396.4	18.2%	92.5	2.982	0.168

Table 4. Cluster summary statistics including mean mechanical properties, thermal properties, high-performance fraction, and key elemental compositions.

The three clusters represent physically distinct alloy families. **Cluster 1** is the largest group (1,675 alloys), characterized by very high Fe content (~93.9 wt%), moderate carbon (~0.42 wt%), and the highest high-performance fraction (35.9%). This cluster corresponds predominantly to carbon and low-alloy steels where Fe is the dominant matrix and carbon drives strengthening consistent with established ferrous metallurgy. **Cluster 0** (986 alloys) shows lower Fe content (~61.8 wt%) and very low carbon (~0.08 wt%), with only 8.5% classified as high-performance. This group likely represents stainless steels and multi-component alloys where alloying additions such as Cr, Ni, and Mo dilute the

Fe matrix and shift the composition away from high-strength configurations. **Cluster 2** is a small but highly distinctive group of only 11 alloys, exhibiting dramatically elevated mean UTS of 1,112 MPa nearly 70% higher than the dataset mean and very high carbon content (~2.98 wt.%). This cluster almost certainly represents high-carbon tool steels or cast irons, where extreme carbon levels produce large volumes of hard carbides that drive exceptional tensile strength. However, the small population size of Cluster 2 warrants caution when generalizing trends observed within this group, and additional data would be required to confirm the robustness of these observations**.**

Cluster-specific regression error analysis reveals that model performance varies across these groups: Cluster 1 achieves the highest R^2 (0.875, RMSE = 84.5 MPa), consistent with its compositional homogeneity and large sample size providing robust training signal. Cluster 0 achieves R^2 = 0.742 (RMSE = 80.7 MPa), reflecting greater compositional diversity within this group. Cluster 2, with only 11 alloys in the test set, yields insufficient samples for reliable R^2 estimation, highlighting the challenge of predicting properties for rare, extreme compositional variants a finding that directly motivates the need for targeted data collection for high-carbon specialty alloys.

3.6 Implications for Alloy Design and Data-Driven Materials Engineering

The combined regression, classification, and clustering results demonstrate the feasibility of using an interpretable machine learning pipeline as a practical support tool for alloy design. By relying solely on chemical composition, the framework enables rapid screening of candidate alloys, identification of high-performance compositions, and exploration of compositional trends without extensive experimental trials.

The identification of Vanadium as the most influential element for UTS prediction, cross-validated across permutation importance and SHAP analysis, suggests that V content may represent an important design variable for improving strength in ferrous alloy systems. This finding provides a physically meaningful composition-based design signal that complements traditional Fe-C-Cr alloying considerations. Similarly, the cluster analysis suggests that high-performance alloys are most concentrated in the high-Fe, moderate-carbon compositional space (Cluster 1), providing a practical screening criterion for early-stage alloy selection.

To illustrate the practical utility of the framework, consider a hypothetical alloy designer wishing to evaluate a candidate composition containing Fe (85 wt%), Cr (10 wt%), Ni (3 wt%), C (0.5 wt%), and V (1.5 wt%). The designer would input this elemental composition vector directly into the deployed Random Forest regressor, which would return predicted UTS and liquidus temperature values without requiring any experimental testing. The logistic regression classifier would then assess whether the

composition meets the dual median threshold criteria for high-performance classification. Finally, the PCA projection would locate this candidate within the established cluster space, identifying which alloy family it belongs to and what property range is typical for that family. This three-stage screening which are predict, classify, and locate can be completed in seconds, providing immediate compositional feedback to guide experimental design and narrow the candidate space before any physical synthesis is undertaken.

4. Conclusions

This study demonstrates the successful application of an interpretable, multi-stage machine learning framework for predicting key thermomechanical properties and screening the performance of metallic alloys using elemental composition as the sole input descriptor. The framework integrates supervised regression, binary classification, and unsupervised clustering within a unified architecture applied to a dataset of 2,672 multi-component alloy systems spanning 31 elemental features. The Random Forest Regressor achieved a test R^2 of 0.848 and a 5-fold cross-validation R^2 of 0.849 ± 0.017, outperforming Linear Regression (R^2 = 0.512) and Gradient Boosting (R^2 = 0.787) baselines. The Logistic Regression classifier achieved an overall accuracy of 82% and an AUC of 0.871, while K-Means clustering identified three physically distinct alloy families with systematic differences in mechanical and thermal properties.

Interpretability analysis combining MDI, permutation importance, and SHAP consistently identified Vanadium (V), Iron (Fe), Tungsten (W), and Carbon (C) as the most influential compositional variables governing alloy performance. These findings provide physically meaningful insight into elemental contributions and demonstrate the value of interpretable machine learning as a practical tool for alloy screening, performance prediction, and accelerated materials discovery.

Several limitations must be acknowledged. The dataset was sourced from a public aggregated repository, and the composition-only input constraint inherently limits predictive resolution within alloy families where processing history is a primary performance determinant. The framework is therefore best suited for rapid compositional screening and family-level trend identification rather than precise within-family property prediction. Future work should focus on alloy-family-specific sub-models, integration of processing parameters and CALPHAD-derived descriptors, implementation of class-weighted classification strategies, and experimental validation of predicted high-performance compositions. Among the processing variables excluded from the current composition-only framework, three are identified as highest priority for future integration. First, **heat treatment conditions** specifically annealing temperature and duration, directly govern precipitate dissolution, grain growth, and phase transformation kinetics, all of which strongly influence UTS independently of

composition. Second, **cooling rate** determines the resulting microstructure, rapid quenching produces martensitic structures with dramatically higher strength than slow-cooled ferritic or pearlitic counterparts of identical composition. Third, **deformation history** including cold work percentage and thermomechanical processing routes introduces dislocation density changes that can alter UTS by hundreds of MPa without any compositional change. Incorporating even one of these parameters as an additional input feature is expected to substantially improve within-family predictive resolution, particularly in the normal UTS range where the current model underperforms.

Declarations

Data and Code Availability

The dataset utilized in this study was obtained from a publicly available Kaggle repository containing multi-component alloy property data compiled from published experimental sources [11]. The complete codebase including the trained machine learning models, preprocessing pipeline, SHAP analysis, permutation importance, cross-validation, clustering, and the deployment interface is publicly available and ready for external use at GitHub. The repository contains all serialized model files, the full Python script, and instructions for running predictions on new alloy compositions. Users can input any elemental composition vector and receive immediate UTS prediction, liquidus temperature prediction, performance classification, and cluster assignment. The framework has been tested on Python 3.x with standard scientific libraries and requires no proprietary software. The complete pipeline supports full reproducibility and transparency of all reported results.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions

A.R. implemented the machine learning framework, conducted the analysis, and drafted the manuscript. S.I. supervised the research, provided scientific guidance, reviewed and edited the manuscript, and contributed to the interpretation of results. All authors read and approved the final manuscript.

Acknowledgements

The authors wish to thank the Department of Metallurgical and Materials Engineering, Federal University of Technology Akure (FUTA), Nigeria, and the Department of Materials Science and Engineering, Texas A&M University, USA, for institutional support.

Conflict of Interest

The authors declare no conflict of interest.

Declaration of Generative AI and AI-Assisted Technologies

During the preparation of this manuscript, the authors used AI-assisted writing tools to improve language clarity and manuscript presentation. The AI was not used to generate research data, results, or scientific conclusions. The authors reviewed and edited all AI-assisted content and take full responsibility for the integrity and accuracy of the final manuscript.

References

[1] T. M. Pollock, "Alloy design for aircraft engines," *Nature Materials*, vol. 15, no. 8, pp. 809–815, 2016. https://doi.org/10.1038/nmat4709

[2] A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the 'fourth paradigm' of science in materials science," *APL Materials*, vol. 4, no. 5, p. 053208, 2016. https://doi.org/10.1063/1.4946894

[3] R. W. Cahn and P. Haasen, *Physical Metallurgy*, 4th ed. Amsterdam, Netherlands: North Holland, 1996.

[4] E. P. George, D. Raabe, and R. O. Ritchie, "High-entropy alloys," *Nature Reviews Materials*, vol. 4, no. 8, pp. 515–534, 2019. https://doi.org/10.1038/s41578-019-0121-4

[5] T. Mueller, A. G. Kusne, and R. Ramprasad, "Machine learning in materials science: Recent progress and emerging applications," *Reviews in Computational Chemistry*, vol. 29, pp. 186–273, 2016. https://doi.org/10.1002/9781119148739.ch4

[6] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Computational Materials*, vol. 5, no. 1, p. 83, 2019. https://doi.org/10.1038/s41524-019-0221-0

[7] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42200–42216, 2020. https://doi.org/10.1109/ACCESS.2020.2976199

[8] K. Tiwari, C. Mendis, and D. Eskin, "Facilitating recycling of 6xxx series aluminum alloys by machine learning-based optimization," *Journal of Sustainable Metallurgy*, vol. 11, pp. 2323–2334, 2025. https://doi.org/10.1007/s40831-025-01112-4

[9] Z. Pei, J. Yin, P. K. Liaw, and D. Raabe, “Toward the design of ultrahigh-entropy alloys via mining six million texts,” *Nature Communications*, vol. 14, no. 1, p. 54, 2023. https://doi.org/10.1038/s41467-022-35766-5

[10] B. L. DeCost, B. Lei, T. Francis, and E. A. Holm, "High throughput quantitative metallography for complex microstructures using deep learning: A case study in ultrahigh carbon steel," *Microscopy and Microanalysis*, vol. 25, no. 1, pp. 21–29, 2019. https://doi.org/10.1017/S1431927618015635

[11] S. Umbare, "Alloy Dataset: Alloy Chemical Composition Physical Property Dataset," *Kaggle*, 2025. [Online]. Available: https://www.kaggle.com/datasets/sohamumbare/alloy-dataset

[12] R. J. Murdock, S. K. Kauwe, A. Y. T. Wang, and T. D. Sparks, "Is domain knowledge necessary for machine learning materials properties?" *Integrating Materials and Manufacturing Innovation*, vol. 9, pp. 221–227, 2020. https://doi.org/10.1007/s40192-020-00179-z

[13] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. Hoboken, NJ, USA: Wiley, 2019.

[14] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.

[15] T. Lookman, F. J. Alexander, and K. Rajan, *Information Science for Materials Discovery and Design*. Cham, Switzerland: Springer, 2016. https://doi.org/10.1007/978-3-319-23871-5

[16] B. D. Conduit, N. G. Jones, H. J. Stone, and G. J. Conduit, "Design of a nickel-base superalloy using a neural network," *Materials and Design*, vol. 131, pp. 358–365, 2017. https://doi.org/10.1016/j.matdes.2017.06.007

[17] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, 2021. https://doi.org/10.7717/peerj-cs.623

[18] M. F. Ashby, *Materials Selection in Mechanical Design*, 4th ed. Oxford, UK: Butterworth-Heinemann, 2011.

[19] S. Sperandei, "Understanding logistic regression analysis," *Biochemia Medica*, vol. 24, no. 1, pp. 12–18, 2014. https://doi.org/10.11613/BM.2014.003

[20] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed., 2022. https://christophm.github.io/interpretable-ml-book/

[21] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, p. 20150202, 2016. https://doi.org/10.1098/rsta.2015.0202

[22] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[23] F. Pedregosa *et al*., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. http://jmlr.org/papers/v12/pedregosa11a.html

[24] W. D. Callister and D. G. Rethwisch, *Materials Science and Engineering: An Introduction*, 10th ed. Hoboken, NJ, USA: Wiley, 2018.

[25] G. Krauss, *Steels: Processing, Structure, and Performance*, 2nd ed. Materials Park, OH, USA: ASM International, 2015.