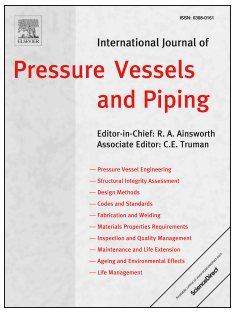
Journal Pre-proof



Machine Learning Assessment of the Importance of Unirradiated Yield Strength as a Variable in Embrittlement Trend Forecasting

Diego Ferreño, Marjorie Erickson, Mark Kirk, José A. Sainz-Aja

PII: S0308-0161(25)00014-6

DOI: https://doi.org/10.1016/j.ijpvp.2025.105444

Reference: IPVP 105444

To appear in: *International Journal of Pressure Vessels and Piping*

Received Date: 27 April 2024

Revised Date: 14 January 2025

Accepted Date: 20 January 2025

Please cite this article as: Ferreño D, Erickson M, Kirk M, Sainz-Aja JA, Machine Learning Assessment of the Importance of Unirradiated Yield Strength as a Variable in Embrittlement Trend Forecasting, *International Journal of Pressure Vessels and Piping*, https://doi.org/10.1016/j.ijpvp.2025.105444.

Machine Learning Assessment of the Importance of Unirradiated Yield Strength as a Variable in Embrittlement Trend Forecasting

Diego Ferreño[1], Marjorie Erickson[2], Mark Kirk[2], José A. Sainz-Aja[1].

[1] Laboratory of Science and Engineering of Materials Division (LADICIM), University of Cantabria, E.T.S. de Ingenieros de Caminos, Canales y Puertos, Av. Los Castros 44, 39005 Santander, Spain.

[2] Phoenix Engineering Associates, Inc., Unity, New Hampshire, USA.

Corresponding author: sainzajaj@unican.es (José A. Sainz-Aja)

Abstract

This paper presents an investigation into the possible influence of pre-irradiation hardening of RPV steel on the transition temperature shift, ΔT_{41J}. Using the ASTM PLOTTER-22 database supplemented with unirradiated yield stress, YS(u), data the study uses machine learning regression algorithms to construct a predictive model that accounts for YS(u) alongside more well-established predictor variables (e.g., copper, fluence, …). The Gradient Boosting algorithm emerged as the most efficient, with performance metrics $R^2 = 0.89 \pm 0.02$ and root-mean square error (RMSE) = 11.2 ± 0.7 ºC. Comparative analyses via bootstrapping underscore the beneficial effect of incorporating YS(u) as a regressor, resulting in a RMSE reduction by 7% and R^2 improvement of 15%. Feature interpretation techniques demonstrate that the significance of YS(u) is comparable to elements like nickel and irradiation temperature and above others such as manganese, phosphorus, or the product form of the steel. The revealed trend — higher YS(u) corresponding to lower ΔT_{41J} — and the lack of significant interactions between YS(u) and the chemical composition, supports the roughly independent role of YS(u). These results underscore the value of incorporating YS(u) as a predictor variable for irradiation embrittlement.

Keywords: Nuclear Reactor vessel; Unirradiated Yield Strength; Machine Learning; Transition Temperature Shift, ΔT_{41J}; Neutron Embrittlement.

1. Introduction

Nuclear power provides a clean source of energy, relative to coal, oil and gas-powered plants, and so is considered an important part of the clean energy strategy necessary to control global climate change [1]. However, few new nuclear power plants are being constructed due to the economic costs and political difficulties involved. Instead, power companies are seeking to extend the operable licensing timeframes for existing nuclear plants to better capitalize sunk costs and optimize clean energy production. Service life extension to 60, 80 and even 100 years has been considered, with research efforts focused on providing the technology to evaluate life extension decisions.

A key factor impacting service life extension decisions is the reliability with which degradation in mechanical properties due to exposure to irradiation can be predicted [2]. Plant surveillance programs were designed to include Charpy V-notch specimens to monitor degradation of impact toughness and tensile specimens to monitor degradation

in the ductility of the base and weld metals used to construct the beltline region of the plant's reactor pressure vessel (RPV) [3]. RPV material degradation due to embrittlement is most often quantified by the transition temperature shift, TTS, which is defined as the change in T_{41J} (the temperature at which the mean Charpy V-notch energy, CVE, is 41J), ΔT_{41J}, from the unirradiated condition to the irradiated condition. Embrittlement trend curve (ETC) models are needed to assess RPV material integrity throughout the planned licensing period.

Radiation embrittlement degradation of RPV steels has been studied since the advent of nuclear reactors, leading to advancements in the understanding of the mechanisms of radiation embrittlement as well as to identification of the key factors controlling degradation. Most current ETC models include terms to account for the primary effects of copper (Cu) content and fluence (Φ) exposure, as well as secondary factors such as nickel (Ni), phosphorous (P), manganese (Mn), exposure temperature (T), and product form (plate (P), forging (F) and weld (W)) [4,5]. Some models also include flux (ϕ), but none of the current models include a term to account for the initial hardening condition of RPV steels. All of the current models fit the data well over the ranges to which they were calibrated, but prediction uncertainty increases in regions in which data is limited. Data limitations impact ETC prediction reliability for current RPV materials to the high fluence conditions expected for some pressurized water reactors during license extension to 60, 80 and even 100 years, and effect embrittlement predictions for new reactor design conditions, such as those envisioned for small modular reactors. Research efforts to refine model prediction accuracy are continuing.

Machine learning (ML) methods are increasingly utilized for predicting radiation effects on steel materials, as evidenced by various recent studies [6–13]. These methods facilitate the identification of numerical correlations between independent and dependent variables and are capable of operating in high-dimensional spaces to reveal complex relationships that might be obscured by traditional analytical methods. The accuracy of ML predictions heavily depends on the quality and representativeness of the training data, necessitating datasets that are both comprehensive and reflective of the varied conditions under which the models will be applied. Recent studies have leveraged ML to analyze and predict neutron embrittlement in nuclear steels. For instance, Morgan et al. [6] provide an extensive review of this application. Cottrell et al. [7] utilized a Bayesian neural network to predict the shift in ductile-to-brittle transition temperature (ΔT_{41J}) for low-activation martensitic steels, highlighting the significant effects of irradiation temperature and dose. They also identified areas lacking data through the analysis of modeling uncertainties. Kemp et al. [8] developed a model using an artificial neural network to estimate irradiation hardening effects, using data covering a range of up to 90 displacements per atom (dpa) and temperatures between 273–973 K. Specific studies focused on nuclear vessel steels include work by Ferreño et al. [9], who applied ML regression models to the ASTM PLOTTER database, achieving notable prediction accuracy with a gradient boosting algorithm. Their approach surpassed existing models in predictive capability, notably reducing uncertainty in ΔT_{41J} predictions. They also employed feature importance analysis to identify key influencing factors on ΔT_{41J}. Xu et al. [10] utilized a XGBoost ML algorithm, achieving a low prediction error for ΔT_{41J} with a dataset of 390 instances, and identified critical dependencies on copper and nickel content, as well as temperature and flux influences. Mathew et al. [11] and Liu et al. [12] further expanded the application of ML in predicting yield stress and ΔT_{41J}, exploring the effects of material composition, flux, and temperature on irradiation embrittlement. Liu et al. evaluated the ratio of effective to actual fluence, providing insights that align with existing physical models. These studies collectively demonstrate the potential of ML in

enhancing our understanding and prediction of radiation effects on steel materials, emphasizing the importance of comprehensive and representative training data.

One area of continued study has been to understand, and more reliably account for, the effects of the RPV steel condition before irradiation on embrittlement behavior. Both hardness, as measured by yield or flow stress, and embrittlement, as measured by toughness, are controlled by the ability of a material to absorb applied strain via movement of dislocations. In a recent paper Erickson and Kirk [13] used a theoretical, dislocation-mechanics based understanding of steel deformation and fracture behavior to argue that the degree of hardening present in a steel prior to its exposure to neutron irradiation should influence the subsequent irradiation embrittlement capacity as quantified by ΔT_{41J}. Using the Zerilli-Armstrong (Z-A) constitutive model of body-centered cubic (bcc) steel flow behavior to model true-stress / true-strain [14] data, they demonstrated the equivalence of hardening due to mechanical strain and hardening due to exposure to radiation. The Z-A constitutive equation includes three terms accounting for the effects of barriers to dislocation motion on the flow behavior of steels upon loading:

$$\sigma_{ZA} = \sigma_0 + K\varepsilon^n + B_0 e^{-\beta T} \quad \textbf{(1)}$$

where *K, B_0,* and β are material constants, ε is the true strain, *n* is the strain hardening exponent (taken to be 0.5 following Taylor's theorem [15]) and *T* is the temperature. The first two terms on the right-hand side of Eq. (2) are athermal as they account for the effects of barriers to dislocation motion that well-exceed the interatomic spacing. σ_0 is the increment of true stress due to obstacles to dislocation motion present in the material prior to loading and can thus represent the initial yield strength of a material, as follows:

$$\sigma_0 = \sigma_G + K\left(\sqrt{\varepsilon_0}\right) \quad \textbf{(2)}$$

where ε_0 is a constant that quantifies the degree of prior hardening (precipitates, point defects and other dislocations) and σ_G is the stress to move dislocations in the fully annealed material. Since ε_0 accounts for all barriers to dislocation motion created by prior hardening, it provides a good indicator of the degree of hardening to which a material has been exposed, including due to the effects of irradiation. Accounting for the effects of prior strain, and irradiation, Erickson et al. showed that true stress / true strain curves for a steel exposed to various levels of fluence overlay each other in the same way that true stress true strain curves for a steel exposed to various levels of strain hardening overlay each other [16]. This demonstration motivated the use of the maximum load condition, $d\sigma/d\varepsilon = \sigma$, to argue that there is a limit to the hardening and embrittlement capacity of a material such that the higher the unirradiated hardness/embrittlement, the less additional hardening/embrittlement can be imparted to the material by irradiation damage. In other words, ΔT_{41J} should decrease with increasing the unirradiated yield stress, YS(u), as shown in [13,16]. Erickson and Kirk concluded that while the effect of YS(u) was not as strong as that of the primary embrittlement variables of copper and fluence, it may equal or exceed the effect of secondary variables such as manganese and phosphorus that are commonly accounted for by ETCs. Recalibrating the ASTM E900-15 ETC [17] to include a term to account for YS(u) showed as much as a 25°C difference between the ΔT_{41J} prediction for a 375 MPa and a 650 MPa yield strength steels at high fluences (this is the current range of YS(u) data for RPV steels). This finding confirms the effect of unirradiated yield strength to be secondary to copper and fluence but on par with those of other variables currently considered in ETCs. Considering the effect of unirradiated yield strength on ΔT_{41J} may refine our

understanding of the effects of other variables on ΔT_{41J} thereby improving ETC accuracy as well as confidence in using ETC's to predict behavior at higher fluences. This paper describes the results of an effort to use ML to assess the importance of YS(u) in predicting embrittlement of RPV steels.

2. Methods

2.1. The Yield Strength Augmented ASTM PLOTTER Database

The present study utilized the ASTM PLOTTER-22 Database, which is an extended version of the database used in a prior study [18]. The dataset used for training the machine learning models was the BASELINE subset of the PLOTTER database. This subset consists of commercial-grade steels with all known descriptive variables, including copper, nickel, manganese, phosphorus, fluence, flux, temperature, and product form defined. The steels were exposed to neutron irradiation in either a pressurized water reactor (PWR) or a boiling water reactor (BWR), and their embrittlement was measured by the ΔT_{41J}, using full-size Charpy V-notch specimens. The BASELINE subset consisted of 2,053 TTS surveillance data from 13 countries, namely Brazil, Belgium, France, Germany, Italy, Japan, Mexico, The Netherlands, South Korea, Sweden, Switzerland, Taiwan, and the United States. The predictor (or regressor) variables included numeric values to describe the chemical composition (Cu, Ni, P, Mn) and irradiation conditions (neutron fluence, flux, and temperature). Indicator/categorical variables were also included to describe the product form (welds, plates or forgings).

Since all surveillance capsules contain tensile specimens, abundant information on the tensile properties of RPV steels exists. To date this information has not been aggregated in a useful way. To provide the data needed for this study we began with the tensile information already aggregated into the NRC's REAP (Reactor Embrittlement Archive Project) database [19], cross-checking for accuracy by referring to the original surveillance reports in some cases. Only unirradiated data were captured. Data described as being tested at "room temperature" (which, when noted in the report, ranged from 18-27 °C) was used because tests at room temperature were conducted for the great majority of capsules. The (usually) 2-3 measurements performed at room temperature were averaged. These data were combined with information provided by Hieronymus Hein of Framatome to the on-going ASTM E10.02 effort to improve its PLOTTER-22 database [20]. This process produced unirradiated yield strength data at room temperature (YS(u)) for 874 ΔT_{41J} data records (686 from USA, 68 from Germany, 111 from South Korea, and 6 from Brazil). The majority are from PWRs (835) and a few from BWRs (36). This dataset is identical to that used in a previous study by the same authors [16]. Except for the variable temperature, which is limited in range because the data on YS(u)) comes primarily from PWRs, the 43% of PLOTTER-22 data for which YS(u) values are known provide reasonable coverage of the data ranges of the four primary embrittlement variables (see Figure 1). These data provide an adequate basis to support this exploratory study concerning the potential effects of un-irradiated yield strength. Efforts are underway within ASTM to increase the amount of data available by incorporating data from both BWRs and PWRs as well as data from other countries. These efforts at data entry and validation will, in the future, provide a more comprehensive data set against which the model developed in this paper, and other models, may be assessed.

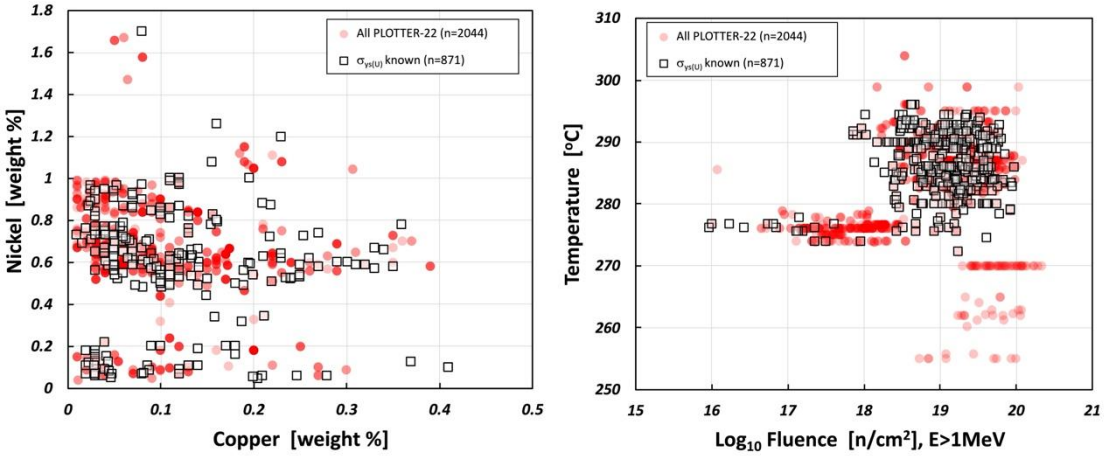


Figure 1. Comparison of conditions for which YS(u) is known with the overall PLOTTER-22 database for the four primary embrittlement variables: copper, nickel, fluence, and temperature.

2.2. Machine Learning

The ML models used in this study were developed and evaluated using the Python 3 programming language and several libraries, including Numpy, Pandas, Scikit-Learn, Matplotlib, Seaborn, and SHAP. The project workflow, described in previous papers [21,22], is summarized in the following sections.

2.2.1. Scope of the Analysis

The regression analysis is conducted with the objective of predicting a numeric value for new input data. The target variable in this case is the TTS (ΔT_{41J}) while the predictors consist of nine features, namely, copper, nickel, phosphorus, manganese, fluence, flux, temperature and product form, all of which are included in the PLOTTER-22 database. Moreover, the yield strength of the unirradiated material is considered here as an additional predictor variable. The dataset utilized for this analysis comprises a total of 874 instances.

2.2.2. Data Preprocessing

Preprocessing involves various stages, including data cleaning, which facilitates the optimization of the model [21,22]. In this study, data outliers, multicollinearity, standardization and nominal categorical variables, were addressed. The techniques described next are recommended for ensuring the efficient implementation of ML algorithms and reliable prediction of outcomes.

- Outliers in data can have a significant impact on the performance of ML models, leading to longer training times and less accurate results. To address this issue, a criterion based on the z-score ($|z|>3.0$) was implemented to identify outliers. However, no outliers were observed.

- Multicollinearity, characterized by high correlations between features, can negatively affect the performance of the ML model by reducing its statistical significance and complicating the determination of feature importance. To identify and address this issue, the Pearson's correlation matrix of the dataset was estimated (see Figure 2). Features exhibiting a correlation coefficient exceeding (in absolute value) 0.60 were removed from the dataset, with one feature eliminated from each correlated pair. However, since the maximum correlations observed were between Cu and P ($r = 0.43$), no features were eliminated.
- The StandardScaler, a widely used scaling technique, provided by Scikit-Learn [23] was employed in this study. The StandardScaler standardizes the features by removing the mean and scaling to unit variance. This method is mandatory for some ML algorithms and recommended for others, as it reduces the influence of large-scale features on the model and improves convergence.
- Additionally, the only categorical variable, product form, was subjected to the Scikit-Learn [23] OneHotEncoder. OneHotEncoder is a widely used method for encoding categorical variables into a binary matrix, which can be used as input to an ML algorithm. The encoded matrix ensures that categorical variables do not impart any inherent order to the model and allows the algorithm to consider each category as a separate feature. The variable 'Product_Form' within the study comprises three distinct categories: forgings (F), welds (W) and plates (P). Implementation of the OneHotEncoder results in the substitution of this categorical feature with three distinct binary variables: 'Product_Form_F', 'Product_Form_W', and 'Product_Form_P'. Each instance within the dataset is characterized by one '1' and two '0' values corresponding to these new features. For instance, a forging sample would be represented by the binary tuple {1,0,0}; a weld sample would be {0,1,0}; and a plate sample would be {0,0,1}. It should be noted that these binary variables demonstrate a perfect correlation, essentially offering redundant information. Consequently, the binary feature 'Product_Form_F' has been eliminated to address this multicollinearity, leading to a modified dataset post the application of the OneHotEncoder. This revised dataset includes only 'Product_Form_W' and 'Product_Form_P' as distinct binary variables, effectively representing the original 'Product_Form' categorical feature.

To perform these manipulations, the Scikit-Learn classes of Pipelines and Column Transformer were used. The purpose of a Pipeline is to assemble several steps that can be cross-validated together while setting different parameters. For example, it can be used to make sure that transformations/preprocessing steps are performed inside the cross-validation loop, thus preventing data leakage. Column Transformer is a Scikit-Learn class used to create and apply separate transformers for numerical and categorical data. This is useful for heterogeneous or columnar data, to be able to apply different preprocessing steps to different columns.

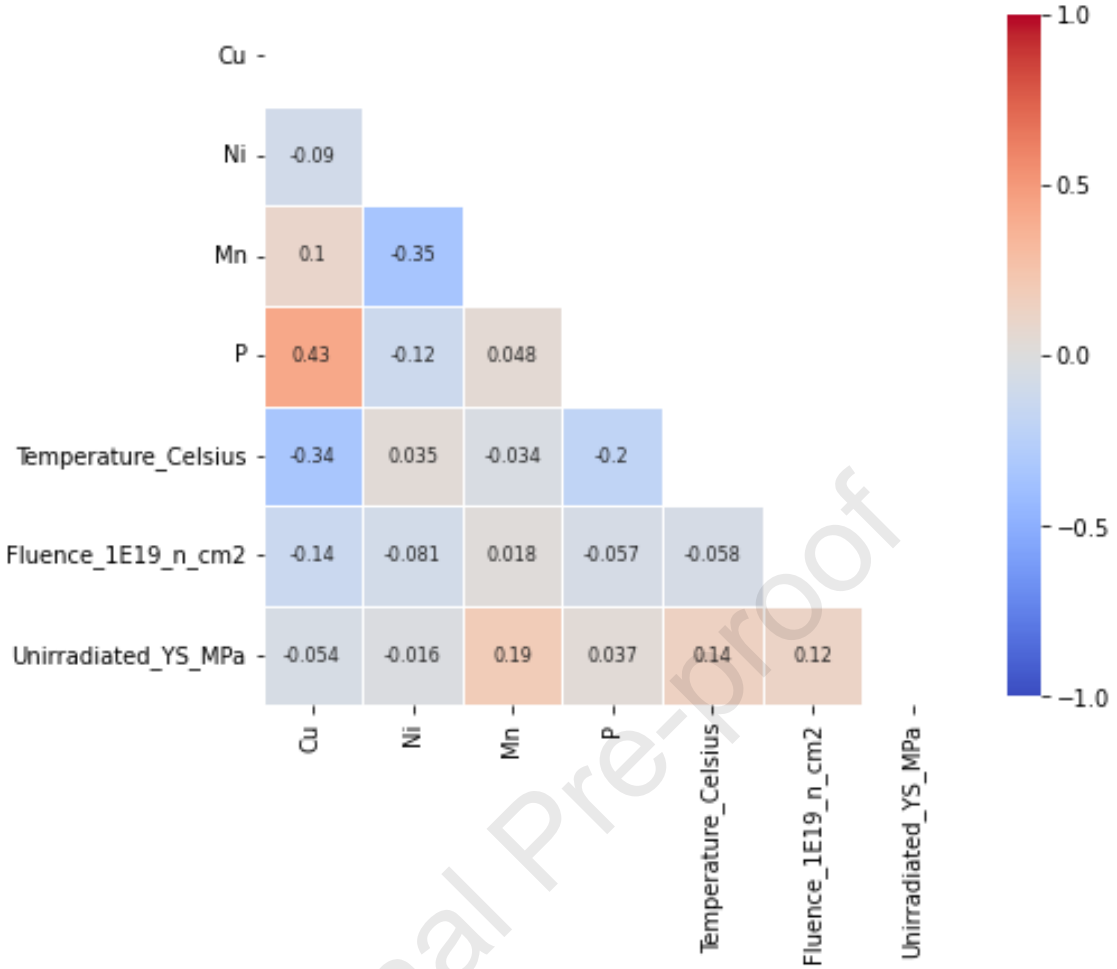


Figure 2. Pearson's correlation matrix of the dataset (only numerical features are considered).

2.2.3. ML Algorithms

The "No Free Lunch theorem" of ML, as stated by Wolpert and Macready [24], posits that for any two learning algorithms, there exist an equivalent number of situations, appropriately weighted, where algorithm one is superior to algorithm two as vice versa, according to any measure of superiority. Furthermore, it has been shown that if an algorithm performs exceptionally well on average for one class of problems, it must do worse on average over the remaining problems. Therefore, comparisons based on the performance of a particular algorithm with a particular parameter setting on a few sample problems are of limited utility.

To address this challenge, this study has implemented a wide range of ML regression algorithms, including Multiple Linear Regression (MLR), k-Nearest Neighbors (kNN), Classification and Regression Tree (CART), Support Vector Regression (SVR), four Ensemble Methods (Random Forest, RF; Gradient Boosting, GB; AdaBoost, AB; Extreme Gradient Boosting, XGB), and Artificial Neural Networks (ANNs). Specifically, the ANNs used in this study are Multi-Layer Perceptron (MLP). These algorithms are briefly described next (the reader will find more details in previous contributions of the same authors [18] or in technical texts [21,22]):

Multiple Linear Regression (MLR) is a commonly used regression algorithm that models the linear relationship between the response variable and one or more predictor variables. k-Nearest Neighbors (KNN) is a non-parametric algorithm that classifies a new data point based on the k nearest data points in the training set. Classification and Regression Tree (CART) is a decision tree-based algorithm that recursively partitions the data based on the most significant predictor variable. Support Vector Regression (SVR) is a regression algorithm that uses support vector machines to find the hyperplane that maximizes the margin between the predicted and actual values. The four Ensemble Methods used in this study, Random Forest (RF), Gradient Boosting (GB), AdaBoost (AB), and Extreme Gradient Boosting (XGB), are all algorithms that combine multiple weak predictors to create a strong predictor. Finally, Artificial Neural Networks (ANNs) are a set of algorithms that model the relationship between the input and output variables through layers of interconnected nodes.

2.2.4. Evaluation of Machine Learning Algorithms

To ensure an unbiased evaluation of the performance of the models, a test dataset was randomly extracted from the available data, prior to any preprocessing [21,22]. Specifically, 25% of the observations were used to form the test dataset, while the remaining 75%, referred to as the train dataset, were used for model training and hyperparameter refinement. To ensure representativeness of both the train and test datasets to the product forms represented in the PLOTTER-22 database, the train/test separation was conducted by stratifying the categorical variable PRODUCT_ID.

Hyperparameter tuning was conducted using GridSearchCV, in which threefold cross-validation was performed (see [18] for details). GridSearchCV is a popular method for hyperparameter tuning that exhaustively searches the hyperparameter space to identify the optimal hyperparameters for a given model. The threefold cross-validation technique partitions the available data into three non-overlapping subsets or folds. The model is trained on two of the three folds and evaluated on the remaining fold, and this process is repeated three times, with each fold serving as the validation set exactly once. This technique helps to prevent overfitting of the model to the training data and provides a more reliable estimate of the model's performance on unseen data.

To avoid the problem of information leakage, it is crucial to ensure that the dataset splitting during cross-validation is performed prior to any preprocessing steps. Specifically, any process that extracts information from the dataset should only be applied to the training subset of the data, and cross-validation should be the "outermost loop" in the overall modeling process. To achieve this objective, the Pipeline class can be utilized within the Scikit-Learn framework.

The Pipeline class is a powerful tool that enables the seamless integration of multiple processing steps into a single Scikit-Learn estimator. With its fit, predict, and score methods, the Pipeline class behaves similarly to any other model in the Scikit-learn library. The primary use case of the Pipeline class is for chaining together preprocessing steps, such as data scaling, with a supervised model (classifier or regressor). By combining these processing steps into a single estimator, the Pipeline class preserves of the integrity of the data throughout the modeling process, ensuring that preprocessing steps are applied only to the training data, and that cross-validation is performed correctly. Consequently, the use of the Pipeline class can help to enhance the accuracy

and reliability of machine learning models by preventing information leakage and other issues that can arise from improper preprocessing techniques.

The coefficient of determination, R^2, and the root mean square error, RMSE, were the scores selected to measure the quality of the ML regression models.

2.2.5. Model Explanation

Interpretability is a common issue in ML models. While complex ML algorithms can generate accurate predictions, their "black box" nature hinders explainability. Typically, there is a trade-off between accuracy and interpretability. Simple models, such as linear regression, are easier to interpret but less accurate than more complex options, such as deep learning and ensemble models. Various conventional approaches are available to compute feature importance. Tree ensemble ML algorithms, such as RF and GB, can provide feature importance measures computed from the impurity decrease within each tree. Impurity is determined by the splitting criterion of decision trees, such as Gini, Log Loss, or Mean Squared Error. However, this impurity-based feature importance can be misleading for high cardinality features, such as those in categorical attributes [25]. Permutation feature importance overcomes these limitations by not having a bias towards high-cardinality features and can be computed on a left-out test set. This technique involves randomly shuffling the values of a single feature and measuring the decrease in the model score. However, when two features are correlated, permuting one feature may not result in a corresponding decrease in the model score. As an alternative, the drop-column importance technique is a model-agnostic measure that determines feature importance by training the model without a particular feature and observing whether the model's performance degrades. The procedure includes four steps: i) Train the ML model with the full feature set and record its performance. ii) For each feature in the dataset, remove that feature and then re-train the model on the reduced data set. iii) Record the performance of each model trained on the reduced data set. iv) Compare the performance of each model trained without a specific feature to the performance of the original model. The impact of removing a feature provides an indication of its importance — a significant drop in performance suggests that the feature is important, while a small change implies that the feature may not be very consequential. While simple to implement, drop-column importance requires re-training the model as many times as the number of features, resulting in higher computational costs compared to other importance measures. However, it provides a robust and interpretable way to measure feature importance.

A recent alternative was proposed by Lundberg et al. [26] who developed the Shapley Additive Explanation (SHAP) method to address the interpretability issues commonly encountered with ML models. The SHAP method is based on SHAP values, a concept from cooperative game theory developed by Lloyd Stowell Shapley [27] that provides a fair profit allocation among stakeholders based on their contribution. In the context of ML, the players in game theory correspond to the features and the game corresponds to an observation, with the objective being to obtain a prediction. The SHAP method offers a unique solution for allocating the prize fairly by satisfying the properties of efficiency, symmetry, linearity and null player [28]. In the context of ML, a SHAP value is the average marginal contribution of a feature value across all the possible combinations of features. Although theoretically straightforward, computing SHAP values can be computationally expensive as it requires averaging over all possible feature orderings. To overcome this

challenge, Lundberg et al. proposed faster methods to compute SHAP values for tree-based models.

A notable characteristic of the SHAP approach is its capability to provide both global and local interpretability, whereas traditional variable importance algorithms (impurity-based, permutation-based or drop-column) only provide global interpretability. Global interpretability is achieved by computing the collective SHAP values, which demonstrate the contribution of each predictor, positively or negatively, to the target variable. In contrast, local interpretability provides a set of SHAP values for each observation, thereby increasing transparency. This permits identification of the factors contributing to a specific prediction and contrasts their impacts across different instances. The SHAP approach's global and local interpretability properties are represented mathematically by eq. (1), where $TTS_{pred,i}$ represents the model prediction for a specific observation 'i' and TTS_{mean} denotes the mean measured TTS of the entire dataset. The summation term calculates the sum of the contributions of each of the 'n' features to explain the difference ($TTS_{pred,i} - TTS_{mean}$). For example, if the model predicts $TTS_{pred,i}$ = 65ºC for the specific instance 'i' while TTS_{mean} = 45ºC, each of the n features (note that the summation extends from j=1 to j=n) will additively contribute to explain the 20ºC difference. The SHAP approach's local interpretability property is particularly useful because it permits an explanation of the model's reasoning for each individual instance, unlike traditional variable importance algorithms that only provide results for the entire population without considering individual cases.

$$TTS_{pred,i} = TTS_{mean} + \sum_{j=1}^{n} \phi_i^{(j)} \quad (1)$$

Furthermore, the SHAP approach is capable of generating a SHAP dependence plot that illustrates the connection between a feature and its influence on the outcome [28,29]. The SHAP dependence plot surpasses the partial dependence plot in terms of information, as the latter only exhibits the average effects and variation of a specific attribute on the target response, while the SHAP dependence plot displays both the average effects and the variation, which may reveal interactions with other features.

In this study, the explication of the optimized model will be conducted utilizing two distinct methodologies: the drop-column method and the SHAP approach. As mentioned above, impurity-based methods for feature importance can be biased towards high cardinality features and are not reliable when features are correlated. Moreover, permutation-based methods can be computationally expensive and may also be biased when features are correlated. The selection of drop-column and SHAP methods is based on their respective strengths. The drop-column method provides a simple and intuitive understanding of the model. Conversely, the SHAP approach provides both local and global metrics, allowing the effects of different features on TTS to be better understood. In combination, these two methodologies provide a well-rounded explanation of the optimized model. Therefore, it is common to see discrepancies in the feature importance reported by different methods. Ideally, multiple methods are used to investigate feature importance and take all the findings into account when making decisions. Different methods can shed light on different aspects of feature importance, and so using a single method might provide a limited view.

3. Results

3.1. Selection of the Optimum Algorithm

In this study, the performance of nine distinct ML algorithms was evaluated, with the aim of identifying the best-performing regression model. The initial stage of the analysis involved training the algorithms using their default hyperparameters, without any tuning strategies. The results of this stage are presented in Table 1, which shows that, in general, the scores for the training set were higher than those for the test set. This trend was particularly pronounced for algorithms that are prone to overfitting, such as the CART or RF. The various performance metrics used in the evaluation conveyed consistent trends across the algorithms. The GB, XGB, and RF algorithms demonstrated the highest R^2 and lowest RMSE in the test set (best means) in absolute terms. Based on these findings, the GB algorithm was selected for hyperparameter tuning.

Table 1. Scores (R^2 and RMSE) obtained in the train and test datasets for the nine algorithms implemented without tuning. MLR: multiple linear regression. KNN: k-nearest neighbors. CART: classification and regression tree. SVR: support vector regression. RF: random forest. AB: AdaBoost. GB: gradient boosting. XGB: extreme gradient boosting. MLP: multi-layer perceptron.

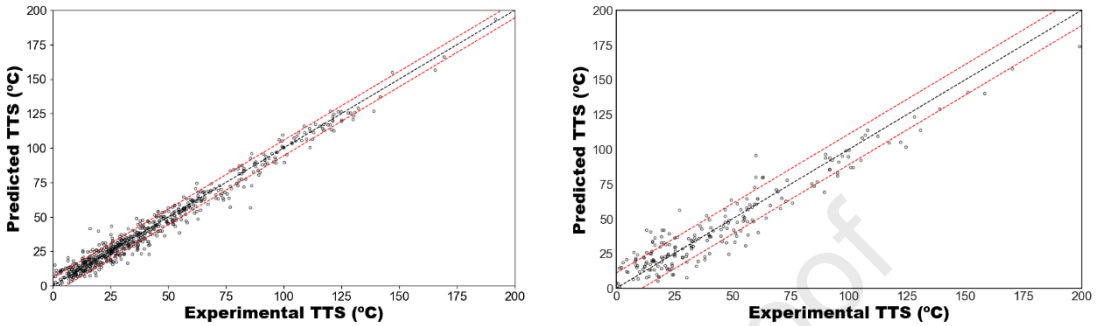
	R^2 - train	R^2 - test	RMSE - train (ºC)	RMSE - test (ºC)
MLR	0.71	0.68	18.17	18.94
KNN	0.85	0.75	12.89	16.67
CART	1.00	0.72	0.70	17.58
SVR	0.43	0.41	25.35	25.70
RF	0.97	0.81	6.16	14.48
AB	0.84	0.76	13.25	16.30
GB	0.95	0.85	7.71	12.95
XGB	1.00	0.83	0.82	13.60
MLP	0.87	0.81	12.06	14.64

3.2. Regression with Gradient Boosting

In view of the former results, the hyperparameters were tuned for the GB algorithm using a grid search and cross-validation scheme to obtain the optimal values of R^2 and RMSE for the train dataset. Hyperparameters are not directly learned within estimators, and they can significantly affect the performance of the model. For reproducibility, the optimal hyperparameters are shown next: learning_rate=0.01, max_features=2, min_samples_leaf=5, min_samples_split=25, n_estimators=5000, random_state=123. The performance of the model was then evaluated on the test dataset to assess the generalization ability of the model. The following scores were obtained: R^2 (train) = 0.972, RMSE (train) = 5.62 ºC, R^2 (test) = 0.906, RMSE (test) = 10.99 ºC. The model exhibits a

moderate amount of overfitting, as indicated by the slightly worse values of R^2 and RMSE for the test dataset compared to the train dataset. However, given the complexity of the model and the nature of the data, this level of overfitting is acceptable.

The scatterplots in Figure 3 represent the TTS predicted with the tuned GB model as a function of the experimental values for the train (a) and test (b) datasets.



(a) (b)

Figure 3. Scatterplots comparing the experimental TTS with the predicted values obtained from the optimized GB model in the train set (a) and test set (b). A 1:1 black dotted line as well as two red dotted lines vertically separated from the former by a distance equal to the RMSE have been included in the figure.

Performing an assessment of the distribution of residuals is a recommended practice in descriptive statistics. The GB algorithm generated a residuals plot for both the train and test datasets, as depicted in Figure 4, which demonstrates a random pattern, and no clear trends are observable. Of significant note are the large residuals visible for low TTS values, which may be associated with the errors inherent in determining TTS from a limited number of Charpy tests, typically ranging from 8 to 12.

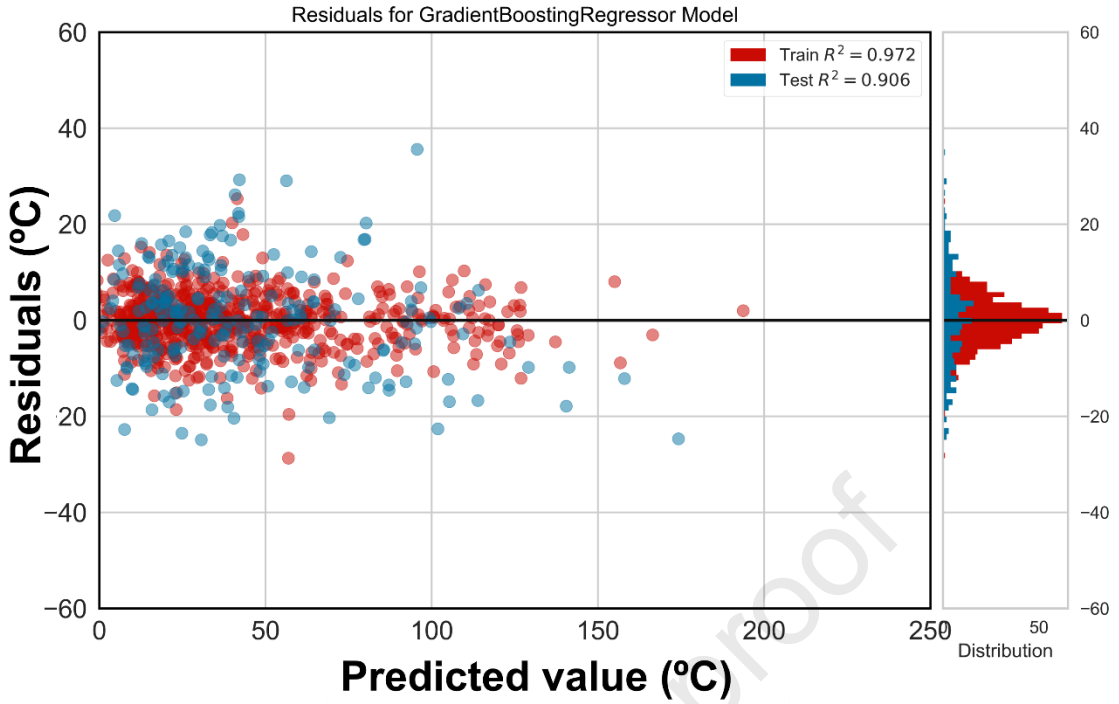
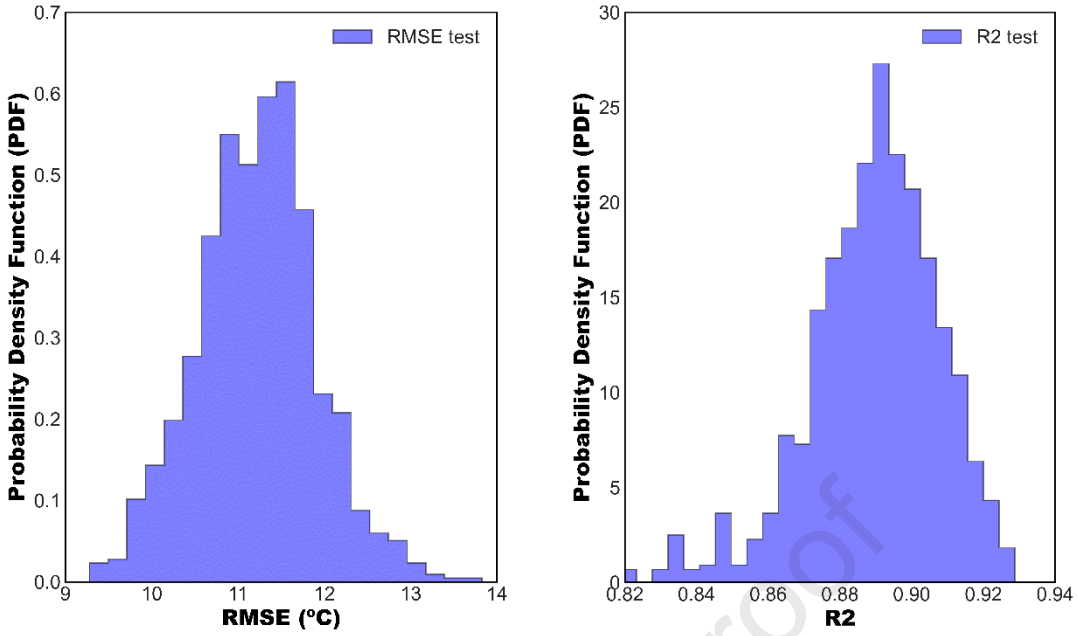


Figure 4. Residuals plot obtained from the GB algorithm in the train and test datasets.

The random train/test split approach is commonly used for estimating the generalization error, which is the prediction error on withheld data, in order to prevent overfitting. However, this approach disregards the spatial context of the data, which can result in suboptimal performance [30]. Specifically, the randomly selected training and test sets may contain test data locations that are spatially close to training data locations in the feature space, which can bias the estimation of the model's performance. Furthermore, spatial autocorrelation can lead to correlation between the test data and the available training data, such that the model may have access to information that overestimates its performance. As a consequence, reliance on a purely random train/test split may lead to an overly optimistic assessment of the model's predictive ability.

Several authors [30–35] have previously acknowledged the impact of autocorrelation on random train/test split, and proposed various correctional methods to address this issue. In the current study, to evaluate the reliability of the aforementioned results, an analysis to assess the sensitivity of the random train/test split on the performance of the ML model was conducted. Rather than employing any corrective technique, the random split was carried out 1000 times, and for each split, the algorithm was trained on the train set to obtain the distribution of R^2 and RMSE. The results for the test set are represented in the histograms in Figure 5; the mean ± standard deviation for the two scores are, respectively, R^2 (test) = 0.89 ± 0.02 and RMSE (test) = 11.2 ± 0.7 ºC. This result validates the result obtained before (R^2 (test) = 0.906, RMSE (test) = 10.99 ºC).



(a) (b)

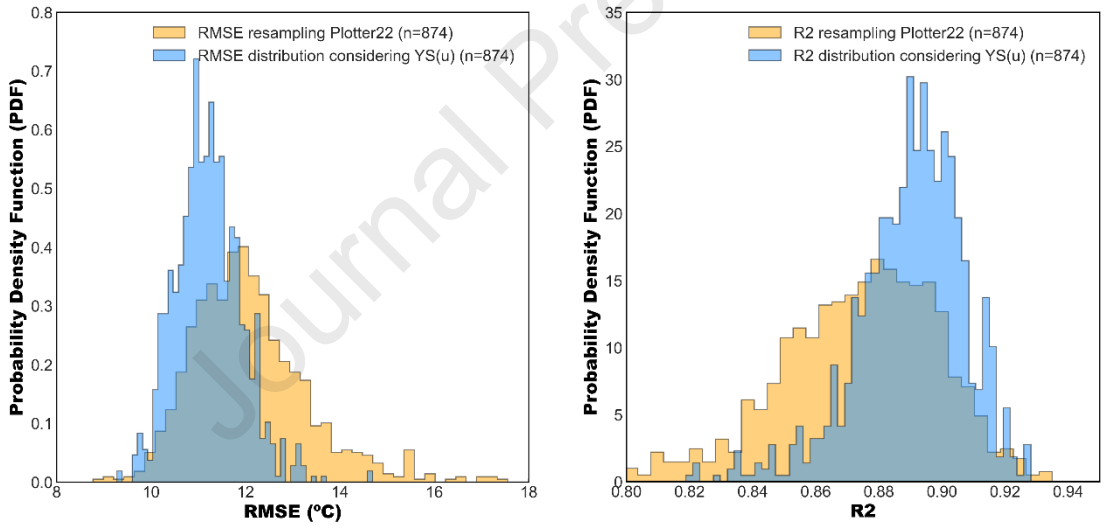
Figure 5. Histograms showing the distribution of the RMSE (a) and R2 (b) after applying 1000 random train/test splits and training the optimized GB algorithm in each of the cases.

3.3. Improvement after including the unirradiated yield strength as a predictor

In a prior work [18], the authors utilized the PLOTTER-15 database, which encompasses 1878 instances from power test reactors, to develop a ML regression model for predicting TTS. The best performing model was a GB algorithm, which yielded RMSE = 10.5 ºC and R^2 = 0.914 on the test set. In comparison, the current study includes the yield stress of unirradiated material along with the nine predictors from the PLOTTER-22 database [13] to predict TTS, yielding RMSE = 10.99 ºC and R^2 = 0.906 on the test set. However, given the discrepancy in sample sizes between the two studies, with 1878 observations in PLOTTER-15 and 874 in the present study, direct comparison of the performance metrics is inappropriate.

The performance of a supervised ML model largely depends on the size of the training dataset. Previous research [36–40] has demonstrated that the ability of ML algorithms to detect patterns is directly proportional to the size of the dataset. In general, smaller datasets tend to result in less powerful and less accurate ML models, particularly when dealing with high-dimensional input samples, which is not the case in this study. To overcome this limitation, data augmentation techniques have been developed, such as random transformations to generate more training data from existing samples in the field of Deep Learning [41]. Similarly, different random oversampling techniques have been developed to address imbalanced learning problems in conventional ML [42,43]. Therefore, comparing the results obtained from two datasets with different sample sizes is not meaningful. A reliable comparison requires the same sample size to properly assess the impact of including the yield stress of the unirradiated material as a predictor.

To achieve a fair comparison, 1000 random subsets each containing 874 samples were bootstrapped (random sampling with replacement) from the PLOTTER-22 database. Each of the subsets was then randomly split into a training set (75%) and a test set (25%), with the training set used to train the GB model with the hyperparameters described in Section 3.2. The trained GB model was then used on the corresponding test dataset to obtain the distribution scores, which are represented in Figure 6 and compared with the distributions obtained previously and represented in Figure 5. Despite the histograms overlapping, both figures suggested an observable improvement after including the yield stress of the unirradiated material, which resulted in a reduction in the RMSE and an increase in R^2. When the yield stress was considered, the RMSE was found to be 11.21 ± 0.7 ºC and R^2 was 0.89 ± 0.02, while using the bootstrapped data yielded an RMSE of 12.10 ± 1.3 ºC and R^2 of 0.87 ± 0.03. Thus, considering the unirradiated yield stress led to a reduction in the mean RMSE of approximately 7% and an increase in R^2 equivalent to 15% of the possible margin of improvement (from 0.87 to 0.89; note that R^2 cannot exceed 1.00). A review of Figure 6 reveals that the enhancement in both R^2 and RMSE scores is primarily attributed to the elimination of less accurate instances characterized by high RMSE and low R^2 values. Moreover, the figure also discloses that upon the introduction of YS(u) as a predictor, the distribution of both scores shifts towards increased symmetry, resembling a Gaussian distribution.



(a) (b)

Figure 6. Histograms comparing the distributions of the RMSE (a) and R2 scores (b) obtained, respectively, after resampling the PLOTTER-22 dataset using 874 instances and including YS(u) as a predictor.

3.4. Interpretation of the model

3.4.1. Feature Importance: Drop-column approach

The barplots in Figure 7 exhibit the comparative performance between the ML model after removing each of the features and the original model to estimate the relative importance of the omitted feature. The first row shows the importance measured in terms

of R^2 and the second uses the RMSE. It is worth noting that the sorting of features in the train set is equivalent regardless of the score employed and that the same occurs for the test set. Nevertheless, there are minor discrepancies between the results on the train and tests sets. In all cases fluence and Cu stand out as the most prominent predictors while the product form, Mn and P are less important. The importance of YS(u) is comparable to other features such as Ni or the irradiation temperature whose use can be justified both in physical terms as well as by data-driven approaches as important contributors to the mechanisms of neutron irradiation embrittlement. Thus, the importance of YS(u) as a feature for predicting TTS appears clear.

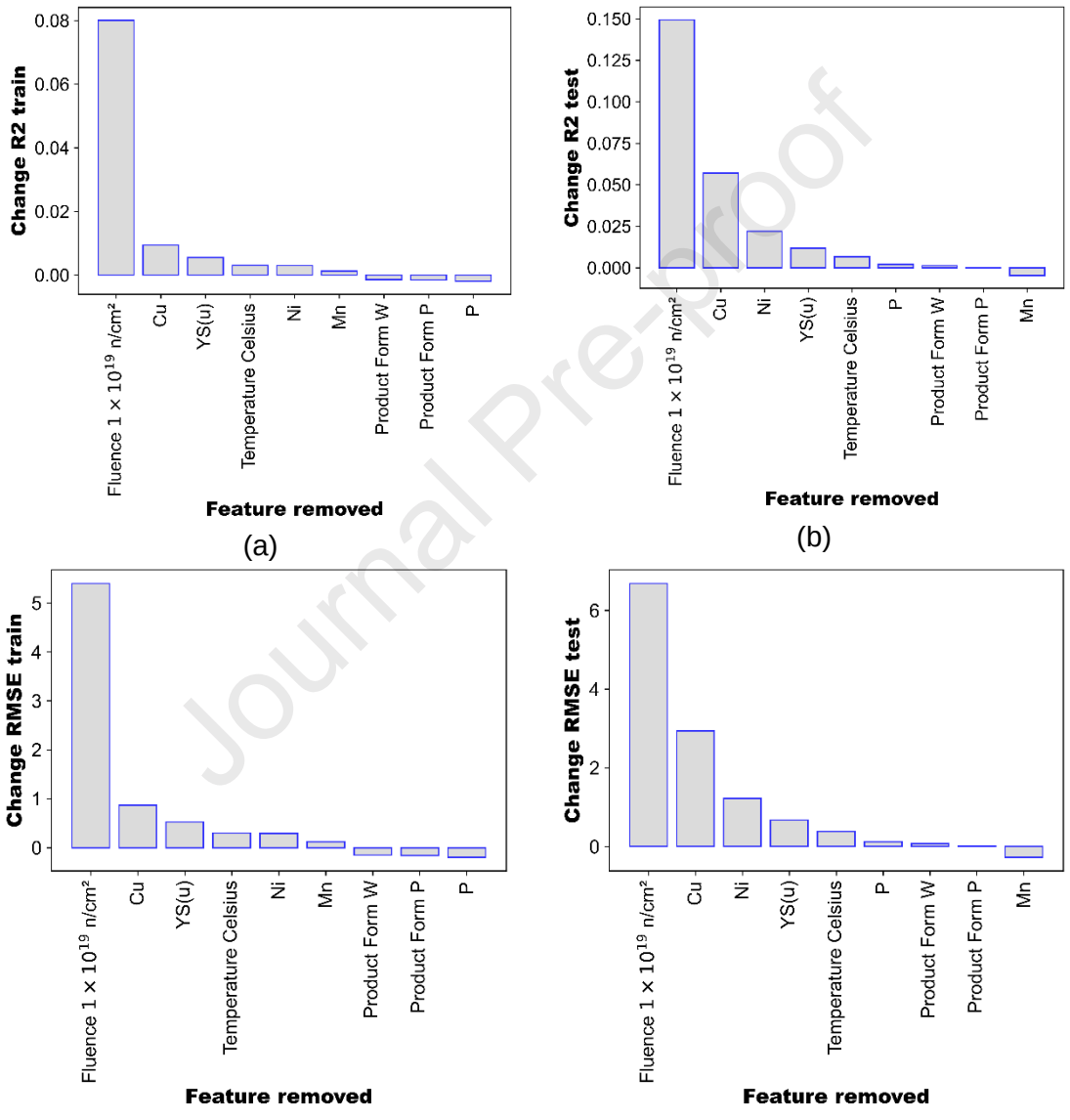


Figure 7. Barplots describing the change experienced by the two scores selected, R^2 and RMSE, after removing each of the features at a time.

3.4.2. Feature Importance: SHAP approach

A characteristic captured by eq. (1) is that, for a specific instance, the SHAP values of all the input features will always sum to the difference between the baseline model output (which corresponds to the expected TTS, E[f(x)] = 41.15ºC, i.e. the mean of the observed data) and the current model output for the observation being explained. The validity of this property can easily be understood through the waterfall plot in Figure 8, which has been generated for a randomly selected instance. As can be seen, the waterfall plot begins with the background prior expectation for the TTS and progressively includes features, one at a time, until reaching the current output of the model (TTS_{pred} was 25.917ºC for this instance). Features are arranged bottom-up in increasing order of influence on the TTS for the selected observation (in this example, the effect of Cu is negligible and the most influential feature is the fluence).

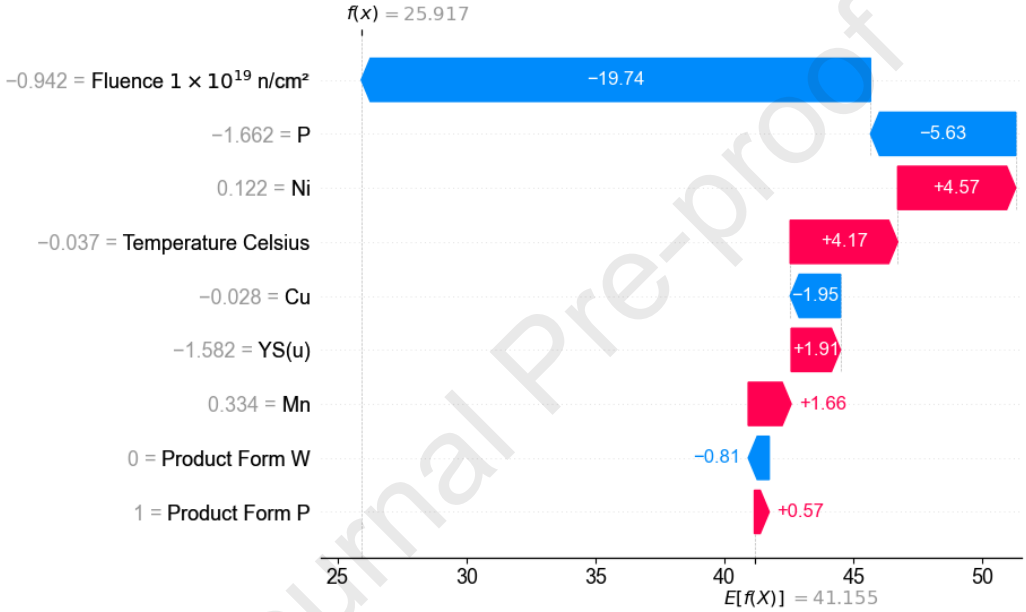


Figure 8. Waterfall plot for an instance randomly selected from the dataset with a TTS_{pred} of 25.917ºC.

To gain an overall understanding of the most influential features of a model, one effective approach is to visualize the SHAP values of each feature for every sample in a summary plot [44], see Figure 9. This plot combines feature importance with feature effects, where each point on the plot corresponds to a SHAP value for a feature and an instance. The feature and SHAP value for a point determine its position on the y-axis and x-axis, respectively, while the color of the point represents the feature value in ascending order. To mitigate overlapping points, they are dispersed along the y-axis to provide a depiction of the distribution of SHAP values for each feature. In the summary plot, the features are sorted in descending order according to their importance, rendering a clear visualization of the most influential features of the model.

The analysis of the summary plot depicted in Figure 9 provides understanding of the underlying patterns. Notably, it can be observed that both Cu and fluence exert a marked positive marginal effect on the target TTS. High Cu values are linked to a substantial increase in embrittlement (up to 60ºC with respect to the mean TTS), while low TTS values (approximately 40ºC below the mean) are associated with a substantially reduced

exposure to fluence. Furthermore, the plot demonstrates that higher irradiation temperatures result in smaller TTS values, whereas the opposite occurs for low temperatures. In addition, higher values of Ni, Mn, and P correspond to a larger TTS. Interestingly, the distribution of Ni displays several outliers for high values, which may artificially increase the significance of this feature. Moreover, the range of the distribution of P is smaller than the other features, typically ranging between -10ºC and 10ºC. When the product form corresponds to a weld, a slightly higher TTS can be observed, while the effect of the plate is negligible. The signal associated with the unirradiated yield stress manifests some blurriness, although a negative correlation can be observed such that higher values of this feature correspond to lower levels of irradiation embrittlement, which aligns well with the rationale presented in Section 1. The range of the TTS for this feature is approximately between -10ºC and 10ºC. A summary plot provides some first indications of the relationship between the value of a feature and the impact on the prediction; to provide more insight into this relationship SHAP dependence plots were produced (see section 3.4.3).

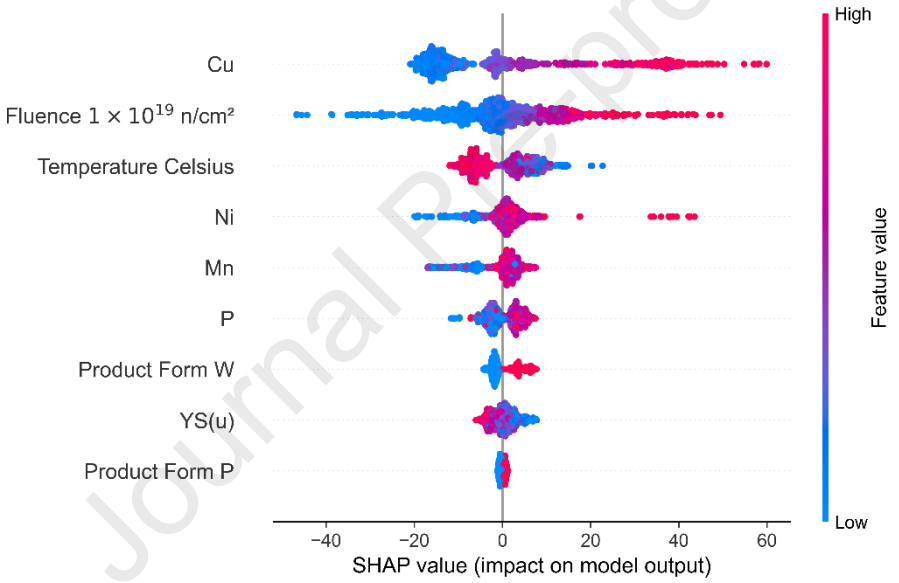


Figure 9. Summary plot indicating of the relationship between the value of a feature and the impact on the prediction

The idea behind SHAP feature importance is simple: features with large absolute SHAP values are important. This way, the global importance of the 'j' predictor, $I^{(j)}$, is defined as the average of the absolute SHAP values per feature across the data [44], see eq. (2) where 'j' represents the specific feature being selected, 'm' characterizes the number of observations in the dataset:

$$I^{(j)} = \frac{1}{m}\sum_{i=1}^{m}\left|\phi_i^{(j)}\right| \qquad (2)$$

Employing the aforementioned definition, the distribution of feature importance, which is presented in Figure 10, is derived. The features can be grouped into three categories

based on their SHAP Importance values. Firstly, Cu, fluence, and temperature emerge are the most prominent variables for determining embrittlement. Secondly, Ni, Mn, P, PF-W, and YS(u) occupy an intermediate tier of importance. Finally, the plate product form feature appears to be of negligible importance in comparison to the other variables.

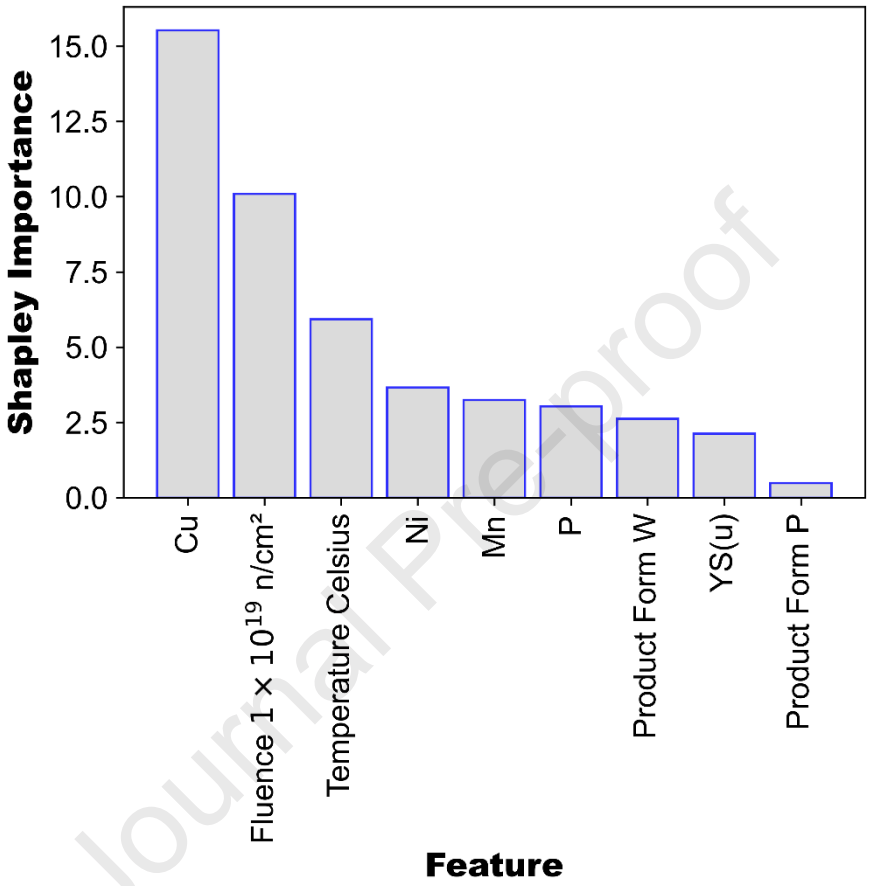


Figure 10. Barplot showing the global importance of each of the predictors, as defined in eq. (2).

While Figure 10 provides a useful high-level summary, mean values nevertheless provide an incomplete description of the importance of the various model features. Therefore, in Section 3.4.3, the SHAP dependence plots are introduced as a more comprehensive visualization tool to characterize the relationship between the predictors and the response variable.

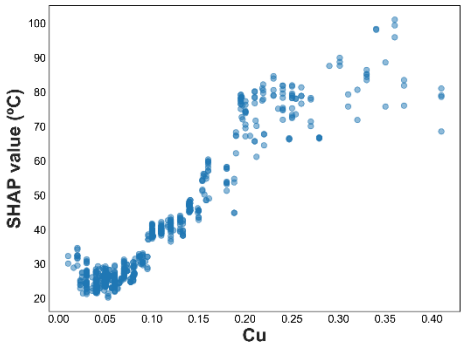
3.4.3. SHAP Dependence Plots

Figure 11 shows the SHAP Dependence Plots of the features involved in this analysis. SHAP dependence plots are an improved alternative to partial dependence plots which only show average effects while SHAP dependence also shows the variance on the y-

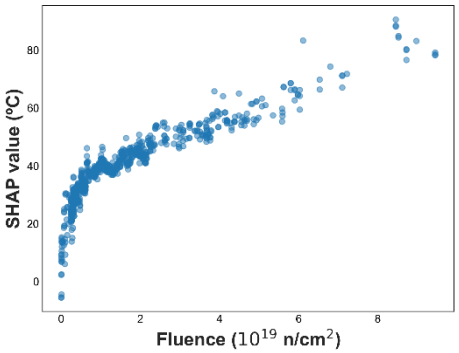
axis since, as explained in Section 2.2.5, the SHAP approach provides local explainability (the SHAP value corresponding to a particular variable represents its expected marginal contribution to the TTS). Dispersion in the y axis may be associated with interactions between features (see section 3.4.4).

The analysis using SHAP dependence plots reveals interesting patterns in the data that in many cases agree well with physical expectations, as follows:

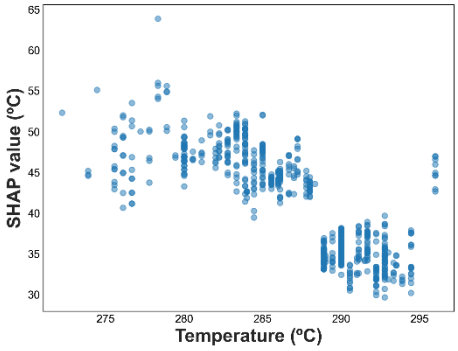
- **Copper**: There is relatively little effect of copper below 0.08 wt% or above 0.3 wt%. Between these values the SHAP value increases approximately linearly from 30 °C to 80 °C. Below 0.05 wt% Cu remains in solid solution and does not contribute significantly to embrittlement via cluster formation.
- **Fluence**: The effect of fluence increases non-linearly to a value of 1.0×10^{19} n/cm^2 and increases with approximate linearity thereafter. For low fluences, embrittlement is controlled by the formation of Cu-rich solute clusters. This mechanism tends to saturate at fluences of approximately 0.5×10^{19} n/cm^2 with matrix damage becoming the controlling embrittlement mechanism at fluences above this value.
- **Temperature**: Higher temperatures correspond to lower levels of embrittlement, with an approximate linear dependence over the range of available data. The total magnitude of the temperature "signal" in these data is approximately 20ºC, significantly lower than for Cu or fluence.
- **Nickel**: In analytic trend curves like E900-15 that have been calibrated to a broad range of data nickel plays a dominant role. While Figure 11 (d) shows a large overall effect of nickel (approximately 50-60 °C) this is produced by three distinct regions of nearly constant values. Thus, for Ni<0.25 wt%, SHAP ≈ 30 °C but with considerable scatter, for 0.25<Ni<1.0 wt%, SHAP ≈ 40 °C with less scatter and, for Ni>1.0 wt%, SHAP ≈ 80 °C for the few observations that are available. These apparent data clusters coupled with the broad range of embrittlement magnitude between them could help to explain the difficulties analytic ETCs have had in establishing a continuous function for nickel that well represents the entire data set.
- **Manganese and Phosphorus**: For these variables no clear patterns emerge other than a slight upward trend (increasing either variable increases SHAP). The significance of these variables seems to be influenced by the scattered clusters of data points at the distribution extremes.
- **Unirradiated Yield Strength**: In this case, although the trend appears noisy, it is more systematic and clearer than for any of Nickel, Manganese, or Phosphorus. Higher values of YS(u) are associated with smaller changes in the TTS. The marginal effect of YS(u) over the data range is ≈ 12ºC. Quantitatively, its influence on the TTS is comparable to that of the chemical features, yet the trend is less blurry.
- **Product Form**: Combining the information represented in Figure 11 (h) and (i) it is possible to identify the quantitative contribution to embrittlement associated with the product form. Thus, Figure 11 (h) indicates that in a nuclear vessel fabricated with a plate (Product_Form_P = 1.0) the expected TTS is ~1.0ºC higher than for a non-plate material. Equivalently, Figure 11 (i) shows that welds (Product_Form_W = 1.0) experience an embrittlement ~7ºC higher than the rest of materials.



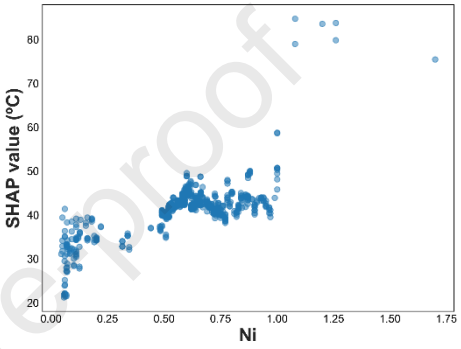
(a)



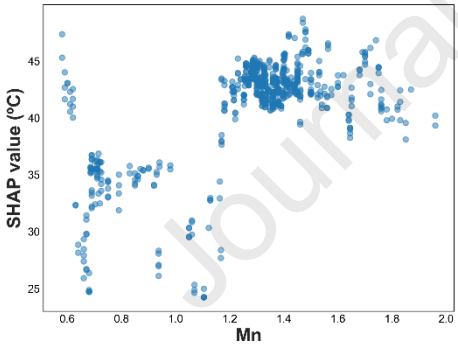
(b)



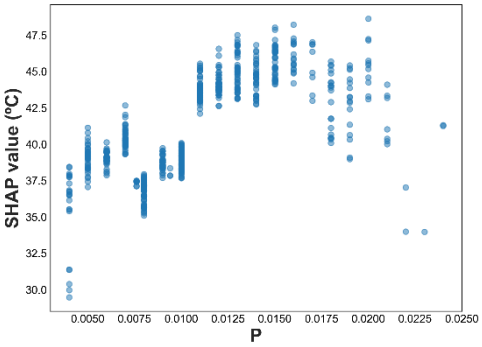
(c)



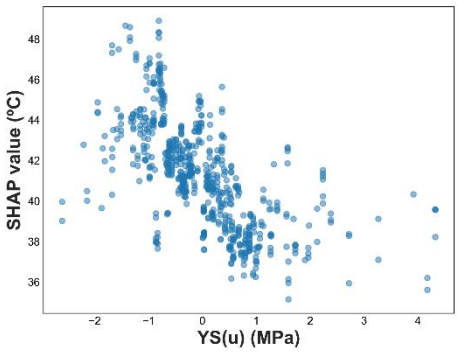
(d)



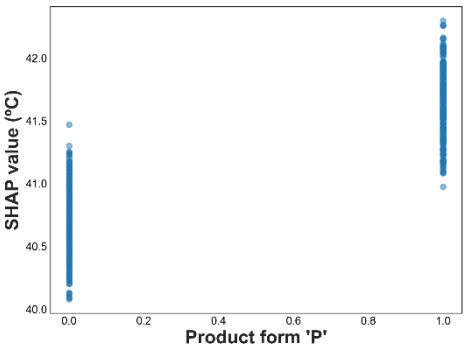
(e)



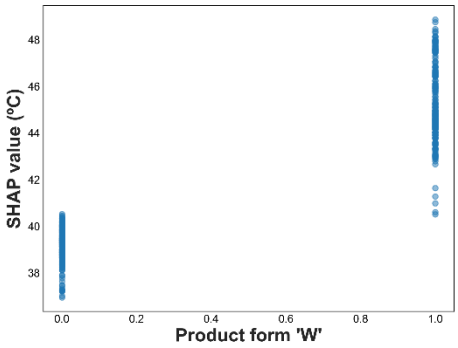
(f)



(g)



(h)



(i)

Figure 11. SHAP dependence plots of the features involved in the analysis. Note that the scale of the vertical axis is different on each plot. The specific scales were chosen to represent the ranges associated with each of the predictors.

3.4.4. Assessment of the interaction between the unirradiated yield stress and the rest of features

This study aims to evaluate the potential use of the unirradiated yield stress as an additional predictive parameter for estimating TTS. The inclusion of this new predictor may be affected by its correlation with the chemical composition of the steel, specifically the quantities of the intentional alloying elements nickel and manganese, which are also considered predictors. The incorporation of alloying elements in steel serves to modify its chemical composition and enhance its properties in comparison to carbon steel. Each alloying element possesses distinct effects on the properties of steel. Copper, manganese, and nickel promote the strengthening of steel by forming solid solutions within the ferrite structure. Consequently, it becomes necessary to assess the co-variation between the chemical features and the unirradiated yield stress in order to ascertain the true significance of the latter as a reliable predictive parameter.

Various methods are available to identify correlations between features in a dataset. The correlation matrix in Figure 2 displays the Pearson's correlation coefficients (r) obtained from pairwise linear regressions conducted on the predictors. It is evident that YS(u) demonstrates a negligible correlation with the chemical features, resulting in r(Cu) = -0.054, r(Ni) = -0.016, r(Mn) = 0.19, and r(P) = 0.037. Some researchers [45,46] propose using a cutoff value of 0.8 or 0.9 to indicate a high correlation between two predictors; clearly, the obtained correlations fall significantly below these thresholds. For the sake of completeness, the p-values of the pairwise regressions between YS(u) and the chemical elements were determined yielding p(Cu) = 0.252, p(Ni) = 0.220, p(Mn) = 7.2E-0.6 and p(P) = 0.268. Although YS(u) and Mn exhibit a significant p-value at the 0.05 significance level, caution must be taken when interpreting this result. In 2016, the American Statistical Association [47] cautioned against the misuse of statistical significance and p-values. Amrhein et al. [48] stated that categorizing results as "statistically significant" or "statistically non-significant" can lead to a mistaken belief that these categories represent distinct differences. Instead, it is recommended to describe the practical implications of values within the confidence interval. In this sense, the small value of the Pearson's coefficient suggests that the correlation between Mn and YS(u) plays a minor role. In addition, the results of the "drop column" approach for determining feature importance indicate that removing YS(u) leads to a reduction in RMSE by 0.53ºC and 0.67ºC in the training and test sets, respectively, while the removal of Mn results in

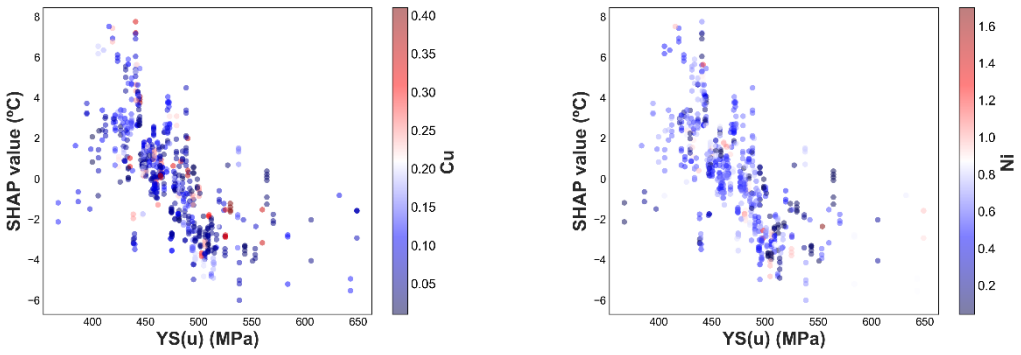
a reduction of 0.12ºC and -0.27ºC. Therefore, if concerns about correlation suggest the removal of one of these predictors, Mn should be removed and YS(u) retained.

A widely used indicator of multicollinearity is the Variation Inflation Factor (VIF) which is defined in equation (3):

$$VIF_j = \frac{1}{1-R_j^2} \qquad (3)$$

where R^2_j is the coefficient of determination for the regression of the 'jth' feature on the remaining variables. A VIF of 1 indicates the absence of correlation between the 'jth' predictor and the other predictors. Although no specific threshold exists to determine the presence of multicollinearity using VIF, a commonly accepted guideline suggests that VIF values exceeding 4 warrant further investigation, while values surpassing 10 indicate significant multicollinearity that necessitates corrective measures [46]. In our dataset, the maximum VIF between YS(u) and other variables was 1.36, corresponding to Cu. This value falls well below the threshold for serious multicollinearity, indicating that there is no substantial collinearity issue in the dataset.

Interactions between predictors are also an important aspect to consider. An interaction occurs when the relationship between one predictor and the target variable depends on another predictor. In the context of our analysis, it is relevant to explore the interaction between YS(u) and the remaining features. The SHAP dependence plot offers a visual method to examine such interactions, as it enables the assessment of vertical dispersion, which may indicate the presence of an interaction with another feature. Therefore, specific SHAP interaction plots for YS(u) were generated by color-coding the SHAP dependence plots based on the values of the other features. The resulting plots are shown in **Error! Reference source not found.**. If different slopes were observed between the SHAP dependence plots corresponding to low and high values of a particular feature (represented by blue and red dots, respectively, see the vertical scale to the right of each of the pictures), it would suggest the presence of an interaction between that feature and YS(u). However, none of the features appear to exhibit any notable interaction with YS(u). In summary, the analysis of SHAP dependence plots did not reveal any significant interactions between YS(u) and the other features in this dataset.



(a) (b)

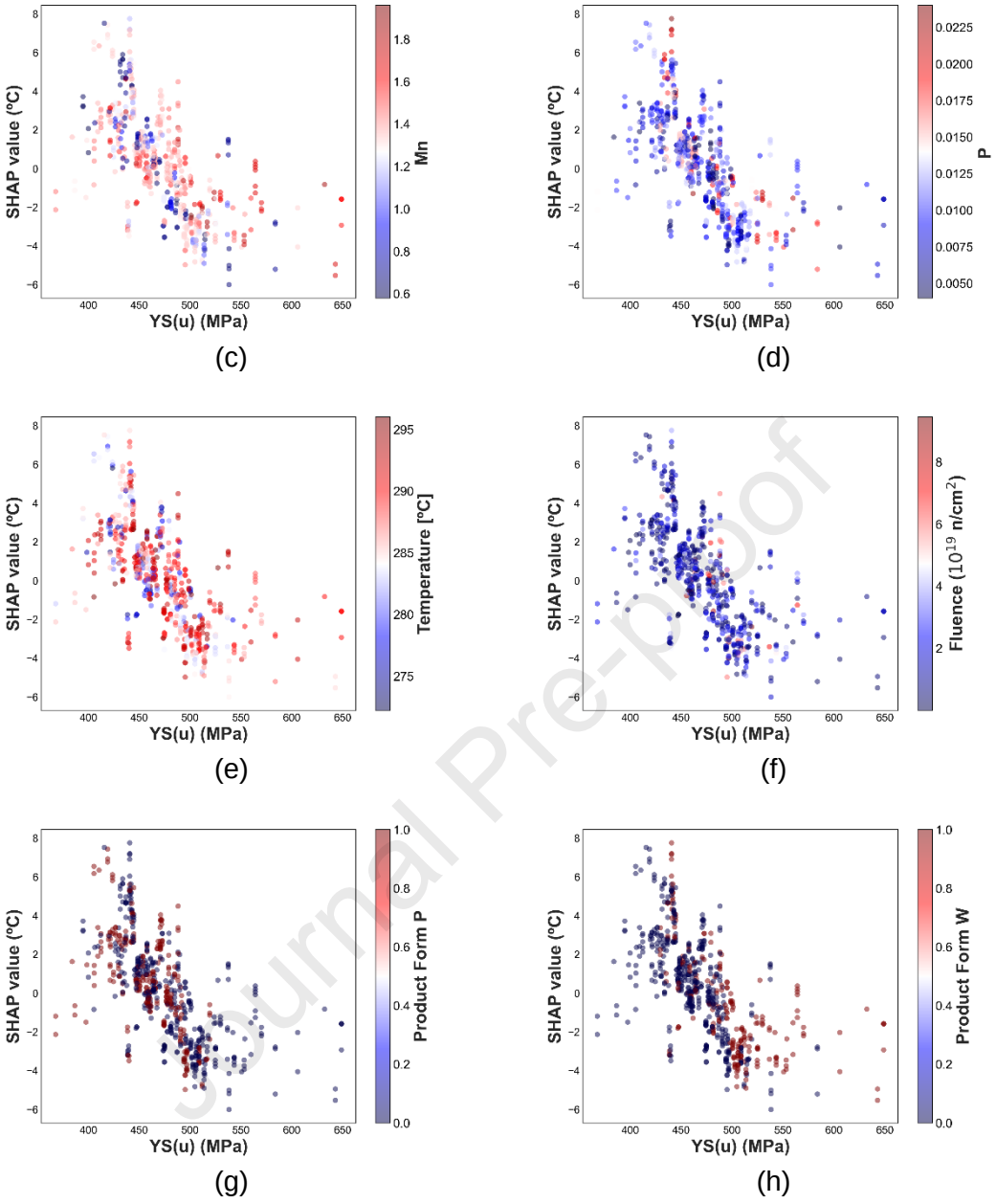


Figure 12. Interaction plots between YS(u) and the rest of predictors. Statistical interaction would imply that the slopes of large (red) and small (blue) values of the feature analyzed should be different.

4. Summary and discussion

We employed machine learning in this study to build on a previous statistical study using the same data [16]. While physics-based models, such as those relying on dislocation mechanics and constitutive equations, provide valuable insights, they often require simplifying assumptions that may fail to capture complex, non-linear interactions evident in data. Similarly, traditional statistical models as used in [16] must assume defined

relationships between input variables that may not represent the dependencies exhibited by embrittlement data, especially in heterogeneous datasets. Machine learning algorithms, particularly ensemble methods like Gradient Boosting, excel in identifying hidden patterns, non-linear relationships, and feature interdependencies without relying on such prior assumptions. Recent literature further corroborates the use of ML for complex material property forecasting, highlighting its ability to handle large, noisy datasets with greater robustness than traditional approaches [10,11].

As just stated, the present study further investigates the hypothesis proposed by Erickson and Kirk [13,16], who suggest that the extent of pre-neutron irradiation hardening in steel has an impact on its subsequent irradiation hardening capacity, as measured by the shift in transition temperature, ΔT_{41J} or TTS. This hypothesis is rooted in the theoretical framework of steel deformation and fracture, employing dislocation mechanics. Currently, there are no existing models or Embrittlement Trend Curves (ETC) that account for this prior hardening phenomenon. To investigate this hypothesis, the ASTM PLOTTER-22 database was utilized, consisting of 2053 data records of ΔT_{41J}, along with various predictors including numerical values for chemical composition (Cu, Ni, P, and Mn) and irradiation conditions (fluence and temperature) as well as a categorical value denoting product type (weld, plate, or forging). The database was supplemented with 874 data points of room temperature unirradiated yield strength, YS(u), obtained from surveillance reports. Several machine learning regression algorithms were trained and tested to establish a model capturing the relationship between the predictors and the target response, ΔT_{41J}. Among them, Gradient Boosting yielded the most favorable outcomes. To interpret the resulting model and identify the most influential features, two complementary approaches were employed, namely, the drop-column method and the SHAP Additive Explanation technique. The selection of these methods was guided by their respective strengths. The drop-column method offers conceptual clarity, providing an intuitive understanding of the model, while the SHAP approach is known for its mathematical rigor and its capacity to provide a range of interpretive tools.

The performance of the Gradient Boosting algorithm on the test set resulted in $R^2 = 0.89 \pm 0.02$ and RMSE = 11.2 ± 0.7 ºC. In a previous study [18], the authors employed the PLOTTER-15 database, an earlier version of PLOTTER-22, which consisted of 1,878 instances. Their regression model achieved an RMSE of 10.5 ºC and R^2 of 0.91 on the test set. However, due to the difference in sample sizes between the two studies, direct comparison of the performance metrics would be inappropriate, with 1878 observations in PLOTTER-15 and 874 in the current study (including YS(u) as an attribute). To enable a fair comparison, 1000 random subsets, each containing 874 samples, were bootstrapped from the PLOTTER-22 database. The trained Gradient Boosting model was then applied to the corresponding test datasets to obtain the distribution scores. The results clearly demonstrate an improvement when YS(u) is included as a predictor. Considering the yield stress, the RMSE was found to be 11.2 ± 0.7 ºC, and the R^2 was 0.89 ± 0.02. In contrast, when using the bootstrapped data, the RMSE increased to 12.1 ± 1.3 ºC, and the R^2 decreased to 0.87 ± 0.03. These findings indicate that the inclusion of YS(u) led to an approximate 7% reduction in the mean RMSE and a 15% increase in R^2.

While the Gradient Boosting model demonstrated strong predictive performance on the test set, a comparison between the train and test results reveals a moderate level of overfitting, as indicated by the slightly higher R^2 and lower RMSE for the train set. To minimize the overfitting risk, several measures were implemented, including careful

hyperparameter tuning via grid search with threefold cross-validation and bootstrapping with 1,000 random train-test splits to assess model stability. These steps ensured that the model's performance on unseen data remained consistent and reliable. The RMSE and R^2 scores across multiple splits demonstrated limited variance, reinforcing the robustness of the model.

The estimation of feature importance was initially conducted using the drop-column approach. The analysis revealed that fluence and Cu exhibited the highest significance as predictors, whereas the product form, Mn, and P were identified as the least relevant factors. In the context of this study, it is noteworthy that the importance of YS(u) is comparable to other features widely acknowledged as needed descriptors for predicting the magnitude of neutron embrittlement, such as Ni or irradiation temperature, which are currently incorporated into various ETCs. Additionally, the SHAP method was employed to estimate feature importance. The results demonstrated that the features could be categorized into three distinct groups based on their SHAP Importance values. Firstly, Cu, fluence, and temperature emerged as the most prominent variables in determining embrittlement. Secondly, Ni, Mn, P, PF-W, and YS(u) occupied an intermediate tier of importance. Finally, the plate product form feature appeared to have negligible importance compared to the other variables. Thus, both approaches provided a consistent assessment of the importance of YS(u), despite slight numerical differences resulting from the distinct definitions of importance used by each method. To gain a more detailed understanding of the influence of YS(u) on the transition temperature shift, SHAP dependence curves were determined. These curves revealed a robust trend, indicating that a higher unirradiated yield stress corresponded to lower ΔT_{41J}, which aligns with the mechanistic hypothesis underlying this study [13, 16]. The marginal effect of YS(u) on the transition temperature shift ranged approximately from -6ºC to +6ºC, resulting in a total impact of 12 ºC. Quantitatively, the influence of YS(u) on the transition temperature shift was comparable to that of chemical features such as Ni, Mn, and P, while exhibiting less noise.

In order to ensure rigor, various methods were employed to investigate the potential correlation between YS(u) and the chemical composition of the steel, specifically the quantities of copper, nickel, manganese, and phosphorus, which are also considered as predictors. Analysis of the Pearson's correlation matrix, pairwise linear regressions between each chemical attribute and yield stress, and the Variance Inflation Factor values yielded no indication of a significant correlation between the chemical features and YS(u) that could contribute to the observed importance of this variable.

All these findings provide strong evidence that the unirradiated yield strength exhibits effects on ΔT_{41J} comparable to those of other variables currently incorporated in Embrittlement Trend Curves. As a result, it is imperative to intensify efforts to gather as much information as possible on unirradiated yield stress as is currently being done by ASTM. A more comprehensive data set would contribute to further enhancing the validation and predictive capacity of this and other embrittlement trend models.

5. Acknowledgements

This work received partial financial support in the frame of the Euratom research and training programme 2019–2020 under grant agreement No 900018 (ENTENTE project).

6. Bibliography

[1] Nuclear Power in a Clean Energy System – Analysis - IEA. https://www.iea.org/reports/nuclear-power-in-a-clean-energy-system (accessed May 8, 2023).

[2] B.-S. Lee, M.-C. Kim, M.-W. Kim, J.-H. Yoon, J.-H. Hong, Master curve techniques to evaluate an irradiation embrittlement of nuclear reactor pressure vessels for a long-term operation, International Journal of Pressure Vessels and Piping 85 (2008) 593–599. https://doi.org/https://doi.org/10.1016/j.ijpvp.2007.08.005.

[3] E. Lucon, E. van Walle, M. Scibetta, R. Chaouadi, M. Willekens, M. Wéber, SCK·CEN contribution to the IAEA Round Robin exercise on WWER-440 RPV weld material: irradiation, annealing and re-embrittlement, International Journal of Pressure Vessels and Piping 79 (2002) 665–684. https://doi.org/https://doi.org/10.1016/S0308-0161(02)00070-4.

[4] L. Debarberis, F. Sevini, B. Acosta, A. Kryukov, D. Erak, Fluence rate effects on irradiation embrittlement of model alloys, International Journal of Pressure Vessels and Piping 82 (2005) 373–378. https://doi.org/https://doi.org/10.1016/j.ijpvp.2004.10.002.

[5] B.Z. Margolin, V.A. Nikolayev, E. V Yurchenko, Yu.A. Nikolayev, D.Yu. Erak, A. V Nikolayeva, Analysis of embrittlement of WWER-1000 RPV materials, International Journal of Pressure Vessels and Piping 89 (2012) 178–186. https://doi.org/https://doi.org/10.1016/j.ijpvp.2011.11.003.

[6] D. Morgan, G. Pilania, A. Couet, B.P. Uberuaga, C. Sun, J. Li, Machine learning in nuclear materials research, Curr Opin Solid State Mater Sci 26 (2022) 100975. https://doi.org/https://doi.org/10.1016/j.cossms.2021.100975.

[7] G.A. Cottrell, R. Kemp, H.K.D.H. Bhadeshia, G.R. Odette, T. Yamamoto, Neural network analysis of Charpy transition temperature of irradiated low-activation martensitic steels, Journal of Nuclear Materials 367–370 (2007) 603–609. https://doi.org/https://doi.org/10.1016/j.jnucmat.2007.03.103.

[8] R. Kemp, G.A. Cottrell, H.K.D.H. Bhadeshia, G.R. Odette, T. Yamamoto, H. Kishimoto, Neural-network analysis of irradiation hardening in low-activation steels, Journal of Nuclear Materials 348 (2006) 311–328. https://doi.org/https://doi.org/10.1016/j.jnucmat.2005.09.022.

[9] D. Ferreño, M. Serrano, M. Kirk, J.A. Sainz-aja, Prediction of the Transition-Temperature Shift Using Machine Learning Algorithms and the Plotter Database, Metals (Basel) 12 (2022). https://doi.org/10.3390/met12020186.

[10] C. Xu, X. Liu, H. Wang, Y. Li, W. Jia, W. Qian, Q. Quan, H. Zhang, F. Xue, A study of predicting irradiation-induced transition temperature shift for RPV steels with XGBoost modeling, Nuclear Engineering and Technology 53 (2021) 2610–2615. https://doi.org/https://doi.org/10.1016/j.net.2021.02.015.

[11] J. Mathew, D. Parfitt, K. Wilford, N. Riddle, M. Alamaniotis, A. Chroneos, M.E. Fitzpatrick, Reactor pressure vessel embrittlement: Insights from neural network modelling, Journal of Nuclear Materials 502 (2018) 311–322. https://doi.org/https://doi.org/10.1016/j.jnucmat.2018.02.027.

[12] Y. Liu, D. Morgan, T. Yamamoto, G.R. Odette, Characterizing the flux effect on the irradiation embrittlement of reactor pressure vessel steels using machine learning, Acta Mater 256 (2023) 119144. https://doi.org/https://doi.org/10.1016/j.actamat.2023.119144.

[13] M. Erickson, M. Kirk, The Effects of Prior Hardening on Irradiation Embrittlement of RPV Steels, in: ASTM Symposium on Embrittlement Trend Curves, Prague, 2022.

[14] F.J. Zerilli, R.W. Armstrong, Dislocation-mechanics-based constitutive relations for material dynamics calculations, J Appl Phys 61 (1987) 1816–1825. https://doi.org/10.1063/1.338024.

[15] G.I. Taylor, The mechanism of plastic deformation of crystals. Part II.—Comparison with observations, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 145 (1934) 388–404. https://doi.org/10.1098/rspa.1934.0107.

[16] M. Erickson, M. Kirk, Use of Unirradiated Yield Strength as a Variable in Embrittlement Trend Forecasting to Better Inform DT41J Predictions, in: International Symposium Contribution of Materials Investigations and Operating Experience to LWRs' Safety, Performance and Reliability, FONTEVRAUD 10, Avignon, France, 2022.

[17] Adjunct for ASTM E900-15: Technical Basis for the Equation used to Predict Radiation-Induced Transition Temperature Shift in Reactor Vessel Materials, 2015.

[18] D. Ferreño, M. Serrano, M. Kirk, J.A. Sainz-aja, Prediction of the Transition-Temperature Shift Using Machine Learning Algorithms and the Plotter Database, Metals (Basel) 12 (2022) 1–24. https://doi.org/10.3390/met12020186.

[19] REAP. https://reapdatabase.nrc-gateway.gov/ (accessed May 8, 2023).

[20] M. Kirk, D.F. Blanco, J.A. Sainz-Aja Guerra, Evaluation of the ASTM E900 ΔT41J Prediction Equation in Light of New Data, in: ASTM Special Technical Publication, ASTM, Prague, Czech Republic, 2023: pp. 233–258. https://doi.org/10.1520/STP164720220063.

[21] S. Guido, A. Müller, Introduction to Machine Learning with Python. A Guide for Data Scientists, O'Reilly Media, 2016.

[22] A. Geron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, O'Reilly Media, Inc., 2017.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825−2830.

[24] D.H. Wolpert, W.G. Macready, No free lunch theorems, IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION 1 (1997) 67–82. https://doi.org/10.1007/978-3-662-62007-6_12.

[25] ScikitLearn, 4.2. Permutation feature importance — scikit-learn 1.0.1 documentation, ScikitLearn Documentation (2021). https://scikit-learn.org/stable/modules/permutation_importance.html (accessed May 9, 2023).

[26] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Adv Neural Inf Process Syst, Curran Associates, Inc., 2017: pp. 4766–4775.

[27] A.E. Roth, Lloyd Shapley (1923-2016), Nature 532 (2016) 178. https://doi.org/10.1038/532178a.

[28] Y. Nohara, K. Matsumoto, H. Soejima, N. Nakashima, Explanation of machine learning models using shapley additive explanation and application for real data in hospital, Comput Methods Programs Biomed 214 (2022). https://doi.org/10.1016/j.cmpb.2021.106584.

[29] Interpretable Machine Learning using SHAP — theory and applications | by Khalil Zlaoui | Towards Data Science. https://towardsdatascience.com/interpretable-machine-learning-using-shap-theory-and-applications-26c12f7a7f1a (accessed May 12, 2023).

[30] J.J. Salazar, L. Garland, J. Ochoa, M.J. Pyrcz, Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy, J Pet Sci Eng 209 (2022) 109885. https://doi.org/10.1016/j.petrol.2021.109885.

[31] D.R. Roberts, V. Bahn, S. Ciuti, M.S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J.J. Lahoz-Monfort, B. Schröder, W. Thuiller, D.I. Warton, B.A. Wintle, F. Hartig, C.F. Dormann, Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography 40 (2017) 913–929. https://doi.org/10.1111/ecog.02881.

[32] G. Ruß, R. Kruse, Regression models for spatial data: An example from precision agriculture, in: P. Perner (Ed.), Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Berlin Heidelberg, Berlin,

Heidelberg, 2010: pp. 450–463. https://doi.org/10.1007/978-3-642-14400-4_35.

[33] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, A. Brenning, Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data, Ecol Modell 406 (2019) 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002.

[34] J. Pohjankukka, T. Pahikkala, P. Nevalainen, J. Heikkonen, Estimating the prediction performance of spatial models via spatial k-fold cross validation, International Journal of Geographical Information Science 31 (2017) 2001–2019. https://doi.org/10.1080/13658816.2017.1346255.

[35] A. Brenning, Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest, in: International Geoscience and Remote Sensing Symposium (IGARSS), 2012: pp. 5372–5375. https://doi.org/10.1109/IGARSS.2012.6352393.

[36] M. Sordo, Q. Zeng, On sample size and classification accuracy: A performance comparison, in: Oliveira José Luís, V. and Maojo, and M.-S. Fernando, and P.A. Sousa (Eds.), Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Berlin Heidelberg, Berlin, Heidelberg, 2005: pp. 193–201. https://doi.org/10.1007/11573067_20.

[37] D.D. Moghaddam, O. Rahmati, M. Panahi, J. Tiefenbacher, H. Darabi, A. Haghizadeh, A.T. Haghighi, O.A. Nalivan, D. Tien Bui, The effect of sample size on different machine learning models for groundwater potential mapping in mountain bedrock aquifers, Catena (Amst) 187 (2020) 104421. https://doi.org/10.1016/j.catena.2019.104421.

[38] P. Kokol, M. Kokol, S. Zagoranski, Machine learning on small size samples: A synthetic knowledge synthesis, Sci Prog 105 (2022) 00368504211029777. https://doi.org/10.1177/00368504211029777.

[39] D. Rajput, W.J. Wang, C.C. Chen, Evaluation of a decided sample size in machine learning applications, BMC Bioinformatics 24 (2023) 48. https://doi.org/10.1186/s12859-023-05156-9.

[40] C.A. Ramezan, T.A. Warner, A.E. Maxwell, B.S. Price, Effects of training set size on supervised machine-learning land-cover classification of large-area high-resolution remotely sensed data, Remote Sens (Basel) 13 (2021) 1–27. https://doi.org/10.3390/rs13030368.

[41] N. Ketkar, J. Moolayil, Deep Learning with Python, Manning Publications, New York, 2021. https://doi.org/10.1007/978-1-4842-5364-9.

[42] M. Cuartas, E. Ruiz, D. Ferreño, J. Setién, V. Arroyo, F. Gutiérrez-Solana, Machine learning algorithms for the prediction of non-metallic inclusions in steel wires for tire reinforcement, J Intell Manuf 32 (2021) 1739–1751. https://doi.org/10.1007/s10845-020-01623-9.

[43] M. Cuartas, E. Ruiz, D. Ferreño, J. Setién, V. Arroyo, F. Gutiérrez-Solana, Prediction of non-metallic inclusions in steel wires for tire reinforcement by means of machine learning algorithms, in: AIP Conf Proc, 2019. https://doi.org/10.1063/1.5138082.

[44] C. Molnar, 9.6 SHAP (SHapley Additive exPlanations) | Interpretable Machine Learning, (2019). https://christophm.github.io/interpretable-ml-book/shap.html#shap-summary-plot (accessed May 13, 2023).

[45] J.Y. Le Chan, S.M.H. Leow, K.T. Bea, W.K. Cheng, S.W. Phoong, Z.W. Hong, Y.L. Chen, Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review, Mathematics 10 (2022). https://doi.org/10.3390/math10081283.

[46] C.H. Mason, W.D. Perreault, Collinearity, Power, and Interpretation of Multiple Regression Analysis, Journal of Marketing Research 28 (1991) 268. https://doi.org/10.2307/3172863.

[47] R.L. Wasserstein, N.A. Lazar, The ASA's Statement on p-Values: Context, Process, and Purpose, American Statistician 70 (2016) 129–133. https://doi.org/10.1080/00031305.2016.1154108.

[48] V. Amrhein, S. Greenland, B. McShane, Retire statistical significance, Nature 567 (2019) 305–307.

Highlights:

- ML algorithms were developed to predict ΔT41J using YS(u) as a predictor.
- This extended model led to a 7% reduction in RMSE and a 15% boost in R^2.
- SHAP importance showed that YS(u) is on par with features such as Ni or temperature.
- Steels with a higher YS(u) exhibit a lower ΔT_{41J}, the rest of things being equal.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: