International Journal of Fatigue 159 (2022) 106785

Contents lists available at ScienceDirect

International Journal of Fatigue

journal homepage: www.elsevier.com/locate/ijfatigue

Application of machine learning algorithms for the optimization of the fabrication process of steel springs to improve their fatigue performance

Estela Ruiz [a], Diego Ferreño [b,*], Miguel Cuartas [c], Borja Arroyo [b], Isidro A. Carrascal [b], Isaac Rivas [b], Federico Gutiérrez-Solana [b]

[a] Global Steel Wire, GSW. Nueva Montaña s/n, 39011 Santander, Spain
[b] LADICIM (Laboratory of Science and Engineering of Materials Division), University of Cantabria, E.T.S. de Ingenieros de Caminos, Canales y Puertos, Av. Los Castros 44, 39005 Santander, Spain
[c] GTI (Group of Information Technologies), University of Cantabria, E.T.S. de Ingenieros de Caminos, Canales y Puertos, Av. Los Castros 44, 39005 Santander, Spain

ARTICLE INFO

Keywords:
Fatigue
Spring
Machine learning
Tempering

ABSTRACT

Machine Learning algorithms are aimed at building generalizable models to provide accurate predictions or to find patterns from noisy data. These characteristics are potentially beneficial for the fabrication of steel products. In this research, 529 rotating bending fatigue tests (R = -1 and σ_a = 400 MPa) were carried out on steel suspension spring bars fabricated using different combinations of manufacturing parameters. A reliable regression model (R^2 = 0.877 on the test dataset) based on the Gradient Boosting algorithm was obtained. The interpretation of the model was carried out through the Permutation Importance algorithm, revealing the relevance of the temperature in the tempering treatment applied after quenching on the fatigue lifespan. This pattern was quantitatively described by means of the Partial Dependence Plot of this feature. Besides, a specific study was carried out to obtain a reliable interpretation of the results derived from the Machine Learning analysis. In this sense, it has been observed that specimens subjected to high temperature tempering display a lower surface hardness that provokes a higher surface roughness after shot peening; this, in turn, facilitates the initiation of surface cracks during the fatigue tests reducing the fatigue lifespan. This study provides a reliable framework to optimize the suspension spring manufacturing conditions to increase their fatigue lifespan as well as an example, generalizable to other manufacturing processes, of the potential benefits of Machine Learning.

1. Introduction

The Paris Agreement establishes a long-term temperature rise of less than 2 °C above pre-industrial levels and its conclusions are relevant for those economic sectors with intensive greenhouse gas (GHG) emissions. Transport accounts for around one-fifth of global CO_2 emissions [1]. Conventional vehicles, based on the combustion engine, emit GHG (mainly CO_2, in addition to other gases such as carbon monoxide, nitrogen oxides, unburned hydrocarbons, lead compounds, sulfur dioxide and solid particles). One of the most effective and realistic measures to minimize GHG emissions in this context is to reduce the weight of vehicles. In this sense, many of the components present in cars, mostly made of steel until very recently, are being progressively replaced by lightweight materials, including metal alloys (aluminum, magnesium), plastics or composites. However, this strategy is not applicable to the elements of greater structural responsibility where steel is irreplaceable due to its high strength and low cost (compared to other structural metal alloys, such as nickel or titanium-based ones). The viable alternative is, therefore, to reduce the size of the structural components made of steel which implicitly requires the improvement of their intrinsic mechanical properties to avoid compromising the safety and functionality of the vehicle. The pressure to manufacture lighter and less polluting vehicles has produced a noticeable weight reduction in the last decades. In this scenario, structural materials must be capable of providing the same strength with smaller sections. This is relevant for fatigue, which represents the predominant failure mechanism of suspension springs, due to the repetitive loads to which they are subjected under in-service conditions.

Two industrial producers, a steelmaker and a manufacturer of suspension springs, have collaborated in this research. The steel wire rod, fabricated by the steelmaker in an electric arc furnace, was supplied to the spring manufacturer to apply the quality, mechanical and thermal

* Corresponding author.
E-mail address: ferrenod@unican.es (D. Ferreño).

https://doi.org/10.1016/j.ijfatigue.2022.106785
Received 30 November 2021; Received in revised form 2 February 2022; Accepted 4 February 2022
Available online 7 February 2022
0142-1123/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Table 1
Specified chemical composition (% weight) of the spring steel C54SiCr6.

Element	Min	Max
C	0.54	0.56
Mn	0.67	0.72
Si	1.4	1.5
P		0.015
S		0.015
Cr	0.65	0.7
Ni		0.08
Cu		0.08
Mo		0.02
Sn		0.015
Al		0.002
V		0.01
N		0.007
B		0.0005
Pb		0.03

treatments necessary to obtain a suspension spring. The study is aimed at optimizing the fatigue behavior of automotive suspension springs using Machine Learning (ML) algorithms to simultaneously model the steelmaking and the suspension spring wire manufacturing. This approach has been used successfully in previous recent studies [2–4]. The scope was designed to produce a number of realistic combinations of fabrication parameters during steelmaking and spring fabrication (fulfilling in all cases their respective quality specifications). These scenarios were defined according to the background of the manufacturers in their respective sectors. Subsequently, an experimental study was carried out in which the fatigue strength of the final steel was characterized by means of rotating bending fatigue tests (mean stress, $\sigma_m = 0$, stress ratio R = -1).

Fatigue crack initiation implies cyclic plastic deformation. For this reason, the fatigue performance of spring steels is strongly influenced by the presence of defects [5,6]. In this sense, the surface finish and the inclusion content represent two competing mechanisms for fatigue initiation [7]. Fatigue occurs at stress amplitudes below the yield stress where plastic deformation is limited to a small number of grains of the material. This microplasticity preferentially occurs in grains at the material surface -because they are less constrained by neighboring grains- than in subsurface grains [8]. Moreover, fatigue crack nucleation can also occur at non-metallic inclusions (NMI) of microscopic size (10 to 100 μm) because of the stress concentration they induce locally; for the sake of completeness, other fatigue crack origins may be present in non-conventional materials, such as those manufactured by selective laser melting [9]. High-strength steels are particularly sensitive to defects (either NMIs or local surface stress concentrators such as roughness) because they exhibit high notch sensitivity, even for micronotches, due to their high yield stress. In fact, it has been observed that the fatigue limit of steel increases approximately proportionally to its tensile strength, but at very high values of the tensile strength, this trend does not continue and lower fatigue limits are obtained [10]. In addition, lifetime can be increased through the application of shot peening [11,12].

In total, 527 fatigue tests were carried out in this study, corresponding to 27 manufacturing conditions (that is, combinations of process parameters, introduced by either the wire rod manufacturer or the spring manufacturer). The stress amplitude $\sigma_a = 400$ MPa was selected (based on prior knowledge about the fatigue behavior of the material) to promote surface crack initiation in all cases. For this reason, after each fatigue test, a fractographic examination was carried out using Scanning Electron Microscopy (SEM) to verify the initiation mechanism. After training and validating a reliable ML model, it was possible to identify the most relevant parameters of the manufacturing process in the fatigue lifespan. Subsequently, a specific experimental study was carried out based on mechanical and metallographic characterization methods to interpret the results derived from the ML study.

The remainder of the paper is organized as follows: the properties of the material, the description of the fabrication process of the steel rod and the springs and the ML methods are described in Section 2. The experimental results as well as the outcome of the ML models are presented in Section 3. Finally, the discussion of results and conclusions are shown in Section 4.

2. Materials and methods

2.1. Steel fabrication and properties

The experimental scope of this study included the characterization of the fatigue strength of C54SiCr6 steel bars, whose specified chemical composition can be seen in Table 1, by means of rotating bending tests.

The fabrication of steel rods for springs comprises the following four major stages: electric arc furnace (EAF), ladle furnace (LF), continuous casting (CC) and hot rolling (HR). They are briefly described hereafter [13–18]:

- In the EAF, high-current electric arcs melt a mixture of steel scrap, direct reduced iron and hot briquetted iron to obtain liquid steel with an adequate chemistry and temperature. The formation of slag is promoted through the addition of lime and dolomite: this favors the refining of steel and reduces heat losses. Molten steel is poured into the transportation ladle where ferroalloys and additives are added to form a new slag layer.
- The secondary metallurgy takes place in the LF where the final chemical composition and the temperature of the steel are adjusted. Deoxidizers, slag formers, and other alloying agents are added for refining. Molten steel is stirred by means of a stream of argon to homogenize the temperature and composition and to promote the flotation of NMIs within the slag.
- During CC, the solidification of steel in the form of billets occurs after pouring the molten material from the ladle into the tundish (a small distributer that controls the flow rates and feeds the mold).
- Rods are obtained from billets through HR. During HR, the steel is passed through several pairs of rolls to reduce the cross-section. To facilitate the process, the temperature of steel during forming is above the recrystallization temperature. Rods are coiled after HR.

The total number of attributes collected throughout the fabrication process of the rods was 277. These attributes will be the features coming from the steel fabrication process for the ML analysis. Coil rods, supplied to the spring manufacturer by the steelmaker, received the same manufacturing process and treatments as the actual springs except for the forming stage. This difference should not influence the final behavior of the material since, after forming, springs undergo heat treatment to relieve residual stresses. For the fabrication of springs, wire rods were subjected to the following processes:

- After straightening the coils, the material was subjected to defect inspection by Eddy currents and any defect detected was removed by grinding. Then, steel coils were drawn to final spring wire diameter.
- Then, the material underwent in-line induction quenching and tempering heat treatment. This was followed by a second inspection for longitudinal defects: in case of detection, the corresponding wire rod was rejected. The tensile strength of steel after these heat treatments is about 2000 MPa (approximately twice its initial strength).
- Finally, a shot-peening treatment was applied.

During the fabrication of springs, 20 parameters were recorded. These were included as features for the ML modelling. Therefore, the total number of attributes was 297. Fatigue tests were carried out on straight bars with a diameter between 10.0 and 12.5 mm supplied by the spring manufacturer.

Most of these 297 features are not freely modifiable by the

manufacturers, but rather are parameters recorded during the fabrication process (temperatures, speeds, flows, etc.). To generate the 27 manufacturing conditions considered in this study, it was decided to act on those variables that are, in theoretical terms, potentially influential on the fatigue behavior of the steel. Specifically, the steelmaker changed the chemical composition of the material -within the specified ranges (Table 1)- and also selected billets from different strands and sequences at continuous casting (during the fabrication of steel, the molten material is distributed by the tundish into six molds giving rise to six strands. Steel emerges horizontally in the form of a solid steel strand and at this point, it is cut to length to produce billets. The index that defines the position of the slab in each strand/line corresponds to the feature billet sequence). On the other hand, the spring manufacturer modified the parameters of the quenching and tempering heat treatments applied to the wire rod. Shot-peening is a standard procedure that is uniform among batches and does not introduce variability. In any case, the dataset considered for the study included all the 297 attributes recorded during fabrication since the manufacturing process itself necessarily induces a certain variability in the parameters recorded that could be influential on the fatigue performance of the final material.

2.2. Rotating bending fatigue tests

The characterization of the fatigue strength was carried out through rotating bending tests following the procedure of the standard ISO 1143: 2010 [19]. This method consists of subjecting a cylindrical bar to uniform bending while the specimen is rotating around its axis. In this way, each point of the specimen is subjected to alternating axial stresses over time (i.e., the mean stress is $\sigma_m = 0$ or, equivalently, the stress ratio is $R = -1$). As this is a bending state, maximum stresses occur at the contour of the specimen.

In order to establish an unbiased comparison of the fatigue strength of each casting and spring fabrication condition, a constant value for the stress amplitude was set for the entire test campaign. After reviewing the history of tests compiled in the context of quality control of the spring manufacturer, an amplitude $\sigma_a = 400$ MPa was established. Under these conditions, failure should presumably be initiated at the surface of the specimens. The frequency of the tests was 25 Hz in all cases. The experimental study enabled the characterization of 27 fabrication conditions through 529 fatigue tests (approximately 20 tests per condition; this number was selected to properly consider the intrinsic dispersion of fatigue results [20–22]).

2.3. Machine learning

The ML models were developed and evaluated in the Python 3 programming language using libraries such as Numpy, Pandas, Scikit-Learn [23], Matplotlib and Seaborn, among others. The workflow of this ML project is summarized [24,25] in the following sections.

2.3.1. Scope of the analysis

The objective of regression analysis is to predict a numeric value for new input data. Here, the target variable is the fatigue lifespan in rotating bending tests, the stress amplitude being $\sigma_a = 400$ MPa. The predictors correspond to the 297 features collected by the steel rod and spring manufacturers. The dataset included the 527 fatigue tests carried out.

2.3.2. Data preprocessing

The ability to learn from ML models and the useful information that can be derived may be extremely influenced by data preprocessing. This consists in cleaning the raw data to enable the optimization of the model. Preprocessing includes the following stages [24,25]:

- Data outliers can mislead the training process resulting in longer training times and less accurate models. In this research, outliers were defined as data points beyond a z-score $|z|>3.0$.
- Multicollinearity is potentially harmful for the performance of the model because it may reduce statistical significance and complicate the assessment of importance of a feature to the target variable. Pearson's correlation matrix of the dataset was obtained and one of the features of every couple with a correlation coefficient exceeding (in absolute value) 0.60 was removed (this selection was supported with engineering judgement).
- Standardization / feature scaling of datasets is mandatory for some ML algorithms and advisable for others. For this reason, the range of all features was normalized so that each one contributes approximately proportionately to the final distance. In this study, features were scaled through the StandardScaler provided by Scikit-Learn [23] which standardizes the features by removing the mean and scaling to unit variance.
- Imputation is the process of replacing the missing values of the dataset by an educated guess. To avoid sacrificing any instances or features, imputation was carried out by means of the KNNImputer provided by Scikit-Learn.
- Ordinal categorical variables were transformed using the Scikit-Learn LabelEncoder and nominal categorical variables were subjected to the Scikit-Learn OneHotEncoder.
- Due to the large number of features (297) as compared to the number of instances (529), the RFECV (recursive feature elimination with cross-validation) method of Scikit-Learn [23] was implemented for feature selection. This technique performs the feature elimination in a recursive fashion to determine the optimal number of features.

2.3.3. ML algorithms

According to the "No Free Lunch theorem" of ML established by Wolpert, "[…] for any two learning algorithms, there are just as many situations (appropriately weighted) in which algorithm one is superior to algorithm two as vice versa, according to any of the measures of superiority" [26]. In other words, "if an algorithm does particularly well on average for one class of problems then it must do worse on average over the remaining problems. […] Thus comparisons reporting the performance of a particular algorithm with a particular parameter setting on a few sample problems are of limited utility" [27]. For this reason, a number of regression algorithms was implemented in this study: Multiple Linear Regression (MLR), K-Nearest Neighbors (KNN), Classification and Regression Tree (CART), three Ensemble Methods (Random Forest, RF; Gradient Boosting, GB; Adaboost, AB) and Artificial Neural Networks (ANNs, in this case, Multi-Layer Perceptron, MLP). A brief description of these algorithms is presented hereafter:

- MLR models the relationship between two or more predictors and the response variable by fitting a linear equation to the observed data. MLR is considered as a baseline algorithm for regression, i.e., a simple model which has a reasonable chance of providing decent results. Baseline models are easy to deploy and provide a benchmark to evaluate the performance of more complex models.
- In KNN, regression is carried out for a new observation by averaging the output variable of the 'K' closest observations (the neighbors) with weights either uniform or proportional to the inverse of the distance from the query point. KNN is an example of an instance-based algorithm which depends on the memorization of the dataset; predictions are obtained by looking into these memorized examples. The distance between instances is expressed through the Minkowski metric which depends on the power parameter, 'p'. When $p = 1$, this is equivalent to using the Manhattan distance and for $p = 2$ the Euclidean distance.
- CART: Classification and Regression Trees were introduced in 1984 by Breiman et al. [28]. In a classification tree the target variable is categorical while in a regression tree it is continuous. A CART splits

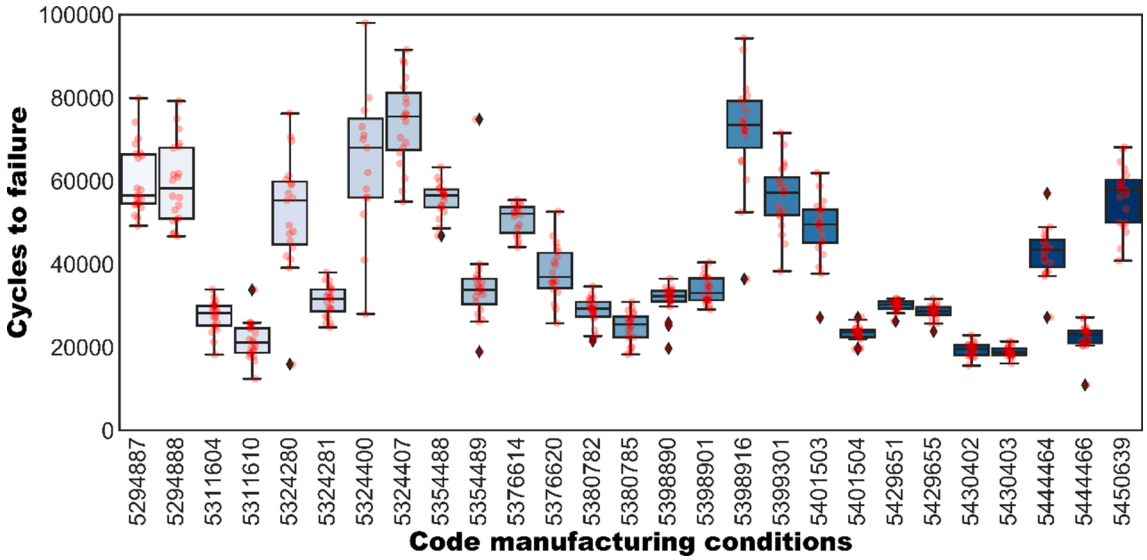

Fig. 1. Boxplot diagram representing the distribution of the fatigue lifespan for each of the 27 combinations of processing variables. A strip plot has been superimposed on each boxplot enabling the observations of each group to be fully appreciated.

the dataset in the form of a tree structure based on the homogeneity of data. The final result is a tree with decision nodes and leaf nodes. The Gini index and the entropy are the most common scores to measure the homogeneity of a sample and to decide which feature should be selected for the next split. Building a decision tree consists in finding the attribute that returns the highest information gain (which is defined as the entropy of the parent node minus the entropy of the child nodes after the dataset has been split on that attribute) or the highest reduction in the Gini index. The main advantages of decision trees are that the interpretation of results is straightforward and that they implicitly perform a feature selection since the top nodes of the tree are the most important variables within the dataset. The main limitation of CARTs is that when a decision tree grows and becomes very complex, it usually displays a high variance and a low bias which results in overfitting. This makes the model unable to generalize and to incorporate new data.

- Ensemble learning is a learning paradigm that focuses on training a large number of low-accuracy models (weak-learners), combining their predictions to obtain a high accuracy *meta*-model [23]. Shallow decision trees are the most widely used weak-learners; the idea behind ensemble learning is that, if the trees are not identical and they predict slightly better than random guessing, a combination of a large number of such trees will give rise to an accurate model. To obtain a prediction for an input, the predictions of each weak model are combined using some sort of weighted voting. Ensemble methods are classified into bagging-based and boosting-based, which are designed to reduce variance and bias, respectively. Bagging (which stands for Bootstrap Aggregation) is the application of the bootstrap procedure (i.e. random sampling with replacement.) to a high-variance ML algorithm. Many models are created and every model is trained in parallel. Each of the models is trained on a subset of the whole dataset composed of a number of observations randomly selected with replacement and a subset of features. Prediction is obtained as the average of the predictions from all the models. The most widely used bagging-based ML algorithm is RF which uses classification trees as weak learners. The most important hyperparameters to tune in a RF are the number of trees and the size of the random subset of the features to consider at each split. By using multiple samples of the original dataset, the variance of the final model is reduced, as is the overfitting. Boosting consists of using the original training data and iteratively creating multiple models by using a weak learner. Each new model tries to fix the errors made by previous models. Unlike bagging, which aims at reducing variance, boosting is mainly focused on reducing bias. In adaptative boosting (often called "Adaboost", AB) and in Gradient Boosting (GB), the ensemble model is defined as a weighted sum of weak learners; the best ensemble model is determined using an iterative optimization process.
- ANNs are used for data classification, regression and pattern recognition. A basic ANN contains a large number of neurons / nodes arranged in layers. A MLP contains one or more hidden layers (as well as one input and one output layer). The nodes of consecutive layers are connected and these connections have weights associated with them. In a feedforward network, the information moves in one direction from the input nodes, through the hidden nodes (if any) to the output nodes. The output of every neuron is obtained by applying an activation function to the linear combination of inputs (weights) to the neuron; sigmoid, tanh and ReLu (Rectified Linear Unit) are the most widely used activation functions. MLPs are trained through the backpropagation algorithm. Gradient descent, Newton, conjugate gradient and Levenberg-Marquardt are different algorithms used to train an ANN.

2.3.4. *Train, test, validation*

25% of the instances (132 observations) were randomly extracted to form a test dataset later used to provide an unbiased evaluation of the models. This way, the model's performance is evaluated on a new set of data that were not seen during the training phase. This approach helps in avoiding overfitting (in this case, the algorithm learns the noise of the training set but fails to predict on the unseen test data). The inconvenience of a train/test split is that the results can depend on the particular random choice of the test set. To avoid this, Scikit-Learn [23] was used to implement 3-fold cross-validation on the 75% of remaining instances (397 observations) in order to select the best models and to optimize their hyperparameters through training and validation, avoiding overfitting. Model selection and hyperparameter optimization were conducted with GridSearchCV.

2.4. *Extended experimental study*

A specific experimental study was carried out to interpret the results derived from the ML modelling, including Vickers microhardness tests, measures of surface roughness [29] and SEM fractographic examination of fatigue-broken specimens. HV0.05 microhardness tests were

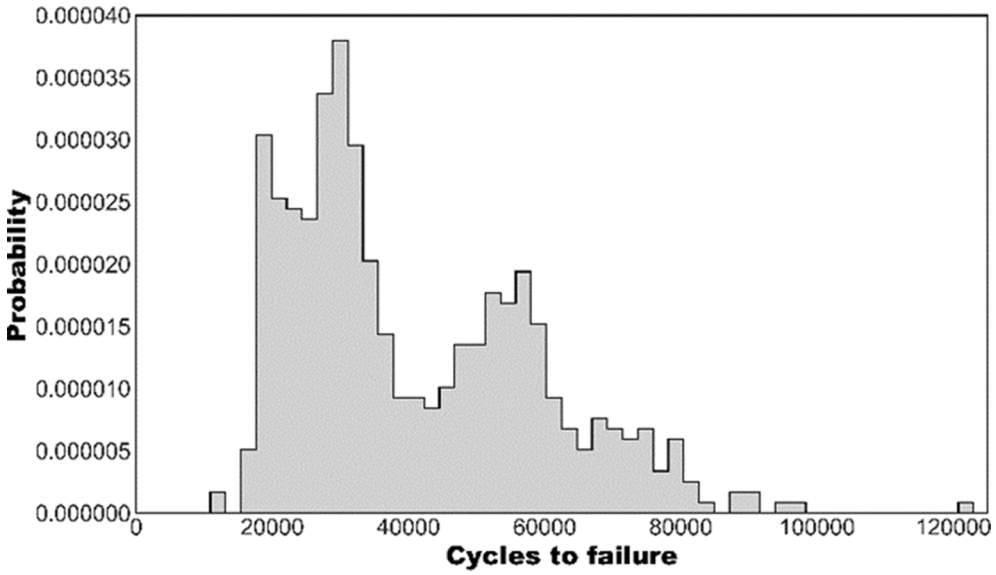

Fig. 2. Histogram of the distribution of fatigue lifespan obtained from the 529 tests carried out. Two modal values can be observed that approximately correspond to the two families of boxplots that can be discerned in Fig. 1.

conducted on the surface of the samples using a Qness Q10-Q30 device. The surface roughness R_a was determined on selected samples by means of a roughness tester PCE-RT 11 and the fractographic study was carried out with a Zeiss EVO MA 15 SEM device.

3. Results

3.1. Fatigue characterization

After carrying out each of the fatigue tests, the fracture surface of the specimen was examined by SEM microscopy: in all cases the failure started on the surface of the sample and no NMI was detected as the fatigue initiator. It follows that, for the experimental conditions imposed ($R = -1$ and $\sigma_a = 400$ MPa), the local plasticization state on the surface of the bar represents the predominant initiation micromechanism in this material.

The collection of boxplots in Fig. 1 shows, for each one of the 27 combinations of manufacturing conditions, the fatigue lifespan obtained from the rotating bending tests. The most striking aspect that emerges from this figure is the substantial variability among different groups in comparison with the intrinsic dispersion within each group. Thus, while in some cases the average fatigue lifespan is greater than ≈50000 cycles, in others it barely reaches ≈25000 cycles. The histogram shown in Fig. 2 reveals a significantly bimodal distribution for the fatigue lifespan in the 529 tests carried out, with a first modal value around 30,000 cycles and a second one around 50,000 cycles. It is worth noting that a bimodal distribution is an absolutely anomalous outcome. Not only there are many experimental evidences but also various analytical models in the literature [21,22,30] for the distribution the fatigue lifespan under ample conditions and in all cases distributions are unimodal. This difference of approximately 20,000 cycles observed between these two modal values will be a reason for consideration in this research.

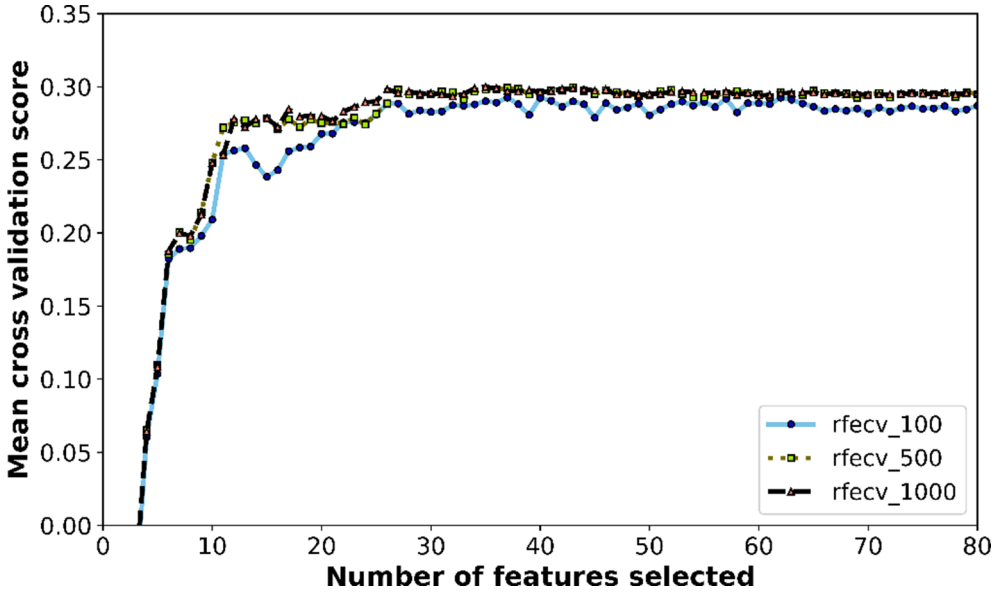

Fig. 3. The negative mean squared error regression loss is plotted against the number of features considered during the RFE process.

Table 2
Results obtained in the test set from the regression algorithms without hyper-parameter optimization.

Regressor	R^2	RMSE	MAE
Linear Regression	0.703	9893	6987
K-Nearest Neighbors (KNN)	0.855	6630	4930
Decision Tree (DT)	0.778	8339	5698
Support Vector Regressor (SVR)	−0.057	18,663	14,925
Random Forest (RF)	0.854	6644	4907
AdaBoost (AB)	0.718	9633	6560
Gradient Boosting (GB)	0.859	6607	4966
Multi Layer Perceptron (MLP)	−4.199	41,386	37,426

Table 3
Results obtained in the test set for the KNN, RF and GB regression algorithms optimized by Grid Search and Cross Validation.

Regressor	R^2	RMSE	MAE
K-Nearest Neighbors (KNN)	0.868	6598	4840
Random Forest (RF)	0.850	7022	5029
Gradient Boosting (GB)	0.877	6354	4584

3.2. Machine learning

3.2.1. Recursive feature elimination

RFE [23] is a technique to reduce the dimensionality of a problem by feature elimination. In this method the most relevant attributes are selected by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features (297 in this case) and the importance of each feature is estimated. Then, the least important features are pruned from the current set and the procedure is recursively repeated. The procedure has been repeated three times using RF estimators with different complexities (100, 500 and 1000 trees). Results are represented in Fig. 3 where the negative mean squared error regression loss is plotted against the number of features. As can be seen, the quality of the model does not improve beyond approximately 25 features, and this outcome is consistent regardless of the estimator. Therefore, these 25 features have been identified and selected, and the rest of analyses have been carried out on this subset. This reduction of dimensionality represents a great advantage in terms of computational effort and of interpretability of the model. In this sense, it has been observed that these variables are mostly associated with the manufacturing process of the spring manufacturer. This result is easily interpretable given that, as indicated in Section 2.1, it is at this stage that steel is subjected to a series of procedures -such as defect inspection, heat treatments and, finally, shot-peening- that may in practice

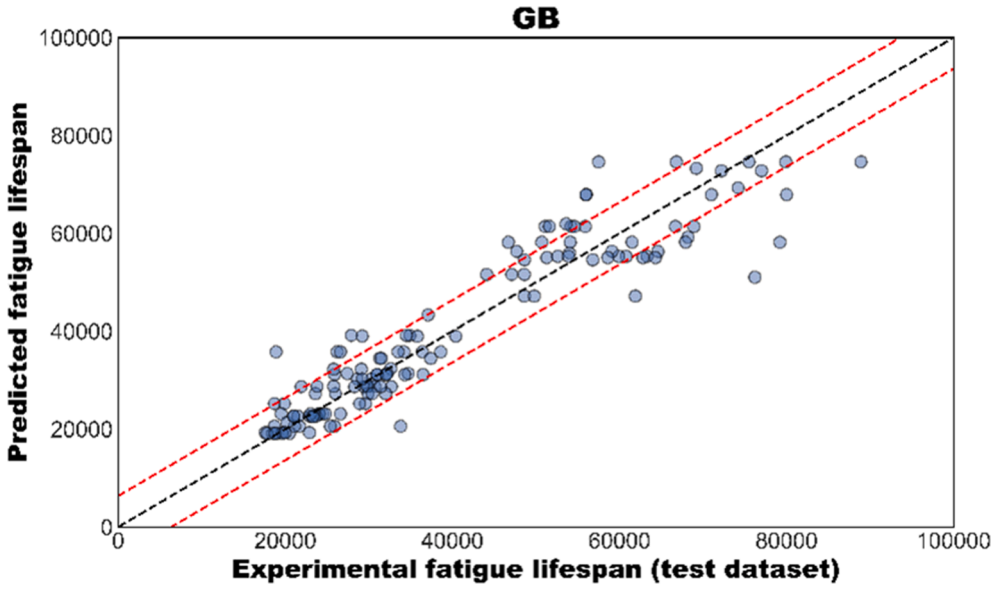

Fig. 4. Scatterplot showing the correlation between the experimental results of fatigue lifespan against the predictions obtained from the GB algorithm for the test set. The figure includes a 1:1 slope line and two confidence bands separated from the previous one by a distance equal to the RMSE.

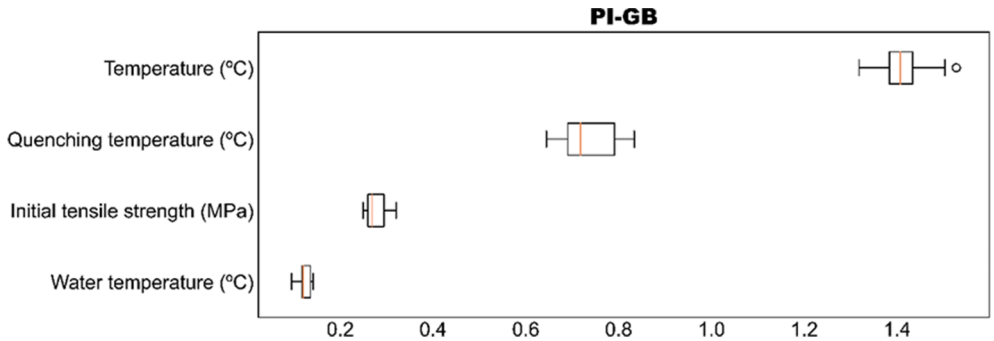

Fig. 5. Classification obtained through the Permutation Importance algorithm of the fabrication variables according to their relevance on fatigue lifespan. The importance, represented in the X axis represents the average reduction in R^2 after randomly shuffling the column in the dataset corresponding to the feature being assessed to generate a corrupted version of the data.

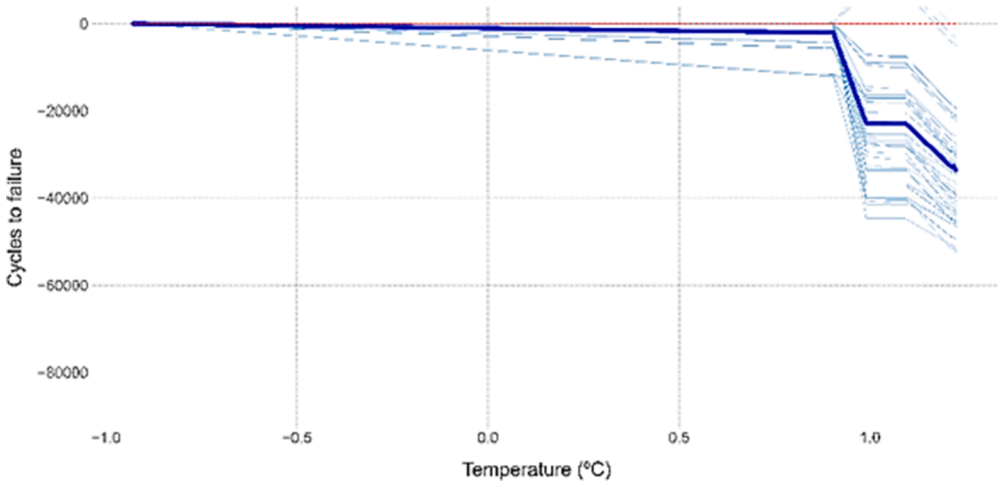

Fig. 6. The figure displays the Partial Dependence Plot of the feature 'Temperature (°C)', which exerts a marked influence on the total fatigue lifespan. The values of the temperature are standardized.

influence the final lifespan of the material.

3.2.2. Results of the regression model

In a first approximation, an evaluation of the results obtained from the eight algorithms described in Section 2.3.3 was carried out, without optimizing their hyperparameters, that is, using their default values provided by Scikit Learn [23]. For each of them, the coefficient of determination R^2, the root mean square error (RMSE) and the mean absolute error (MAE) have been obtained in the test set, as shown in Table 2. It can be seen that, even without optimizing, the KNN, RF and GB algorithms display very promising results.

Consequently, the hyperparameters of these three algorithms were optimized using Grid Search and Cross Validation (K-fold, $K = 3$). The results, reproduced in Table 3, highlight that GB is the algorithm with best prediction performance.

Fig. 4 presents a scatterplot where, for the observations belonging to the test set, the experimental results obtained for the total fatigue lifespan are compared with the predictions derived from the GB model. As can be seen, a very satisfactory agreement has been achieved. The figure also suggests the existence of two clusters of observations that approximately correspond to the two modal values shown in Fig. 2. These two clusters are approximately separated in Fig. 4 by a horizontal line corresponding to a lifespan of ~ 42000 cycles; this is also evident in Fig. 2. Roughly, 45% of the instances in this study exhibited a lifespan above that separation.

3.2.3. Assessment of feature importance

The identification of the features that most influence the fatigue lifespan was carried out through the impurity-based and the permutation-based algorithms, which were implemented in the Scikit-Learn library [23]. Both procedures provided equivalent results, which are represented in Fig. 5: features have been sorted by decreasing order of importance and, for the sake of simplicity, only the four most relevant are included in the figure since the importance attributed to the rest of them is negligible. As can be seen, the feature 'Temperature (°C)' clearly stands out from the rest: It represents the material temperature during the tempering heat treatment applied after quenching which was deliberately modified, among other variables, to study its influence on fatigue lifespan.

The forcefulness of the information shown in Fig. 5 is an invitation to focus on the attribute 'Temperature (°C)'. Its Partial Dependence Plot, represented in Fig. 6, shows that for values beyond ≈0.8 standard deviations above the mean (X-axis), the fatigue lifespan of the material is reduced on average by more than ≈30000 cycles (Y-axis) (for the experimental conditions imposed on the rotating bending tests). This result is consistent with the lifespan distributions represented in Fig. 1 and Fig. 2, as well as with the scatterplot in Fig. 3. The values of tempering temperature are standardized in Fig. 6 (this is one of the data preprocessing procedures previously explained in Section 2.3.2); tempering temperatures approximately range between 435 °C and 500 °C and the limiting temperature identified in Fig. 6 correspond approximately to 475 °C.

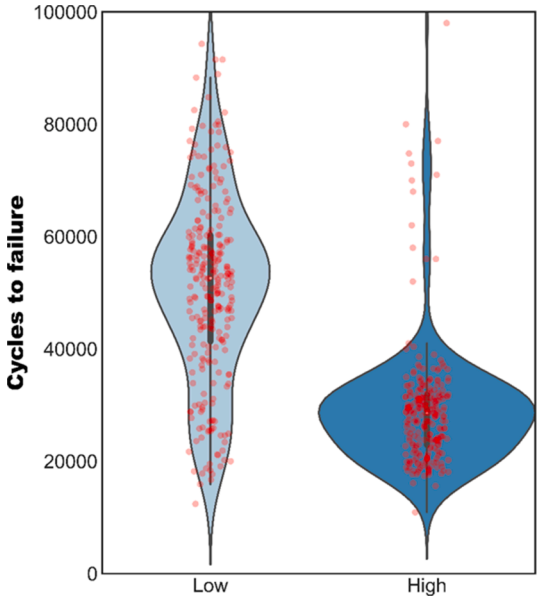

Fig. 7. These violin diagrams show the distribution of the lifespan in the rotating bending fatigue tests for those specimens that were not subjected to high tempering temperatures (left) and those subjected to it (right).

3.3. Mechanical interpretation

As previously seen, the tempering temperature stands out from the rest of fabrication attributes regarding the total fatigue lifespan of the specimens. To analyze the influence of this treatment properly, it is first necessary to understand its nature and the grounds that justify its application.

In this sense, in Fig. 7, the distributions of the fatigue lifespan obtained from the rotating bending tests were represented in the form of violinplots, distinguishing between those specimens with low or high

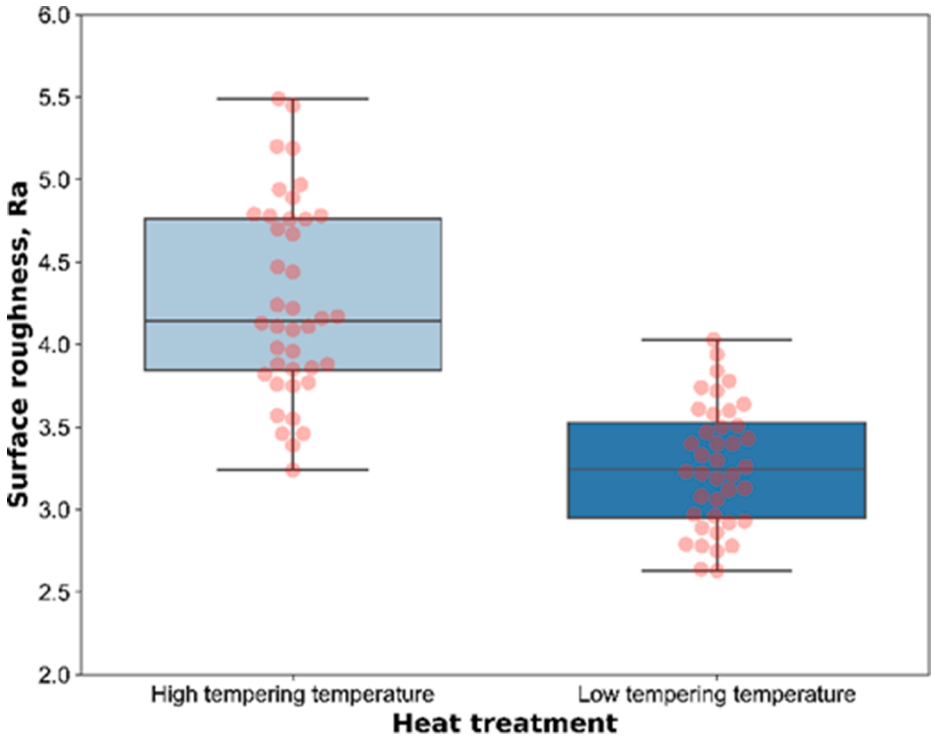

Fig. 8. Boxplots showing the surface roughness, R_a, in samples either subjected (left) or not (right) to high temperature tempering.

tempering temperature (below or above 475 °C, as explained in Section 3.2.3). When comparing Fig. 7 with the histogram in Fig. 2, it can be seen that the observations that constitute the first modal value approximately correspond to those of the diagram on the right in Fig. 7, while those of the second modal value can be identified with the violinplot on the left. Therefore, the bimodal distribution presented above was nothing else that the superposition of two unimodal distributions.

The very obvious differences between the distributions observed in Fig. 7 cannot, however, be considered as evidence of causality. For this reason, a specific empirical study was carried out to evaluate the influence of the additional tempering on the properties of the material and on its fatigue lifespan.

The Vickers surface microhardness was obtained on a series of specimens belonging to the same heat (which, consequently, have been subjected to the same fabrication conditions during steelmaking), that were split into two groups depending on whether they had or had not received high temperature tempering. The microhardness of specimens with high temperature tempering were 550 ± 10 HV while in low-temperature specimens it was 620 ± 11 HV.

Fatigue initiation is strongly influenced by the surface finish of the material [31]. For this reason, the surface roughness R_a of a series of bars from the same heat and previously tested under rotating bending fatigue conditions, was determined. The results obtained are shown in Fig. 8 in the form of boxplots. There are several aspects worthy of consideration. First, the surface roughness of the high-tempering temperature group (left) is appreciably higher than the rest (right). A *t*-test [32] was carried out to compare the mean values of the distributions of the pair of boxplots obtaining a p-value of $\approx 10^{-14}$; therefore, there is strong evidence to reject the null hypothesis (equal mean values). Before carrying out the *t*-test, the hypotheses of normality of the distribution of the means and homogeneity of the variance required by the test were verified. In the first case, the Shapiro test of normality was used and in the second, the Levene homoscedasticity test.

To complete the study, a series of specimens of the same heat, subjected respectively to high and low tempering temperature, were selected and subjected to fractographic examination to identify differential patterns in the failure micro-mechanisms that could justify the variations observed in the fatigue lifespan. The most relevant aspect that

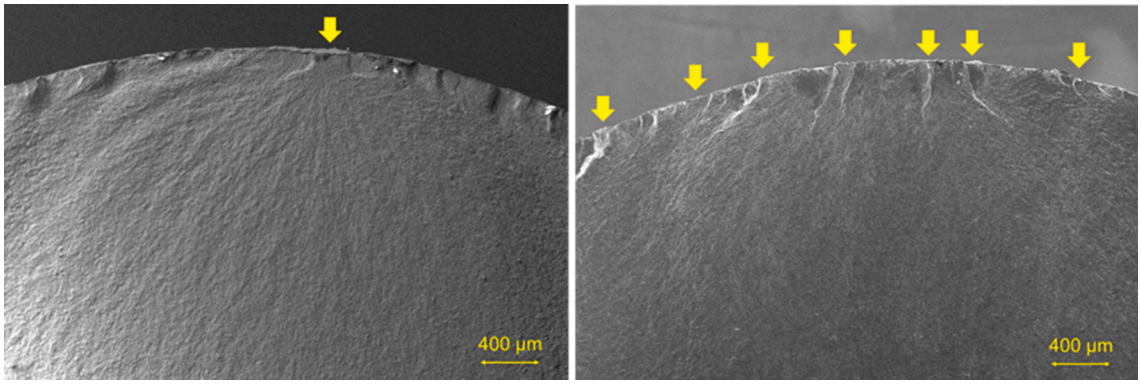

Fig. 9. The figure shows the fractographies of two specimens from the same heat range/group subjected to low (left) or high temperature tempering. In the first case, a single initiation site is seen, while the high-temperature tempered bar displays a large population of initiators.

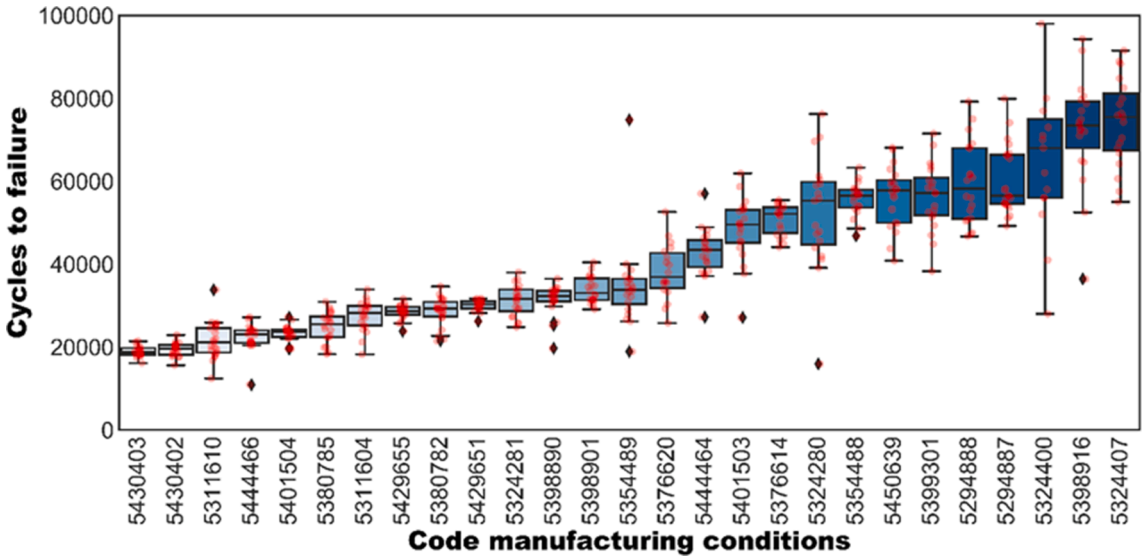

Fig. 10. Boxplot diagram showing the results obtained for the 27 combinations of manufacturing parameters analyzed. The data sets have been ordered in increasing order of the mean fatigue lifespan.

was observed, shown in Fig. 9, is associated with the fatigue initiation process. Thus, while a single initiator is observed in the left picture, the right one shows multiple initiation sites in the contour, giving rise to the overlap of crack growth surfaces that converge to one single crack front, leading to the final fracture of the bar [8]. These patterns have been systematically observed in the rest of the specimens included in the fractographic study.

4. Discussion and conclusions

The evidence derived from the ML modelling and the experimental studies allows the following evidences to be established:

- Substantial differences were identified in the distribution of the fatigue lifespan depending on whether the specimen had or had not undergone high-temperature tempering after quenching (see Fig. 7).
- Microhardness revealed that the main mechanical effect derived from the high-temperature tempering consists of the softening of the material (550 ± 10 HV vs. 620 ± 11 HV).
- Specimens subjected to high-temperature tempering exhibit greater surface roughness (Fig. 8).
- According to the fractographic study, see Fig. 9, the specimens treated with high-temperature tempering display a larger number of surface initiation fatigue sites.

All this evidence provides a consistent interpretation based on the theoretical and experimental foundations of the phenomenon of fatigue. In particular, it is important to distinguish between the initiation of fatigue (the generation of a crack from a surface) and its subsequent propagation; moreover, it is worth taking into account that, in general, initiation consumes most of the fatigue lifespan of a component [8]. Initiation from a non-cracked surface requires the confluence of a series of conditions to promote the plastification of the material. In particular, the local mechanical properties of the steel and the roughness of the initiation surface play a relevant role. Thus, a lower yield stress facilitates the initiation of fatigue and so does a higher surface roughness, due to the induced stress concentration [8] (the pioneering work of De Forest [33] revealed this influence of the surface finish on the initiation of fatigue). As has been experimentally proved, increasing the tempering temperature softens the material and thereby, in accordance with the above interpretation, facilitates initiation. The softening induced by the tempering heat treatment on steels has been previously reported by other authors. Tempering is applied on hardened martensitic steels precisely to improve toughness and ductility and lower hardness. As explained by Canale et al. [34], "During tempering, solid-state reactions occur and the as-quenched martensite is transformed into tempered martensite, which, at higher tempering temperatures, is composed of highly dispersed spheroids of cementite (carbides) dispersed in a soft matrix of ferrite, resulting in reduced hardness and increased toughness". These authors provide some reference curves that illustrate the effect of carbon content and tempering temperature on hardness of carbon steels (the higher the carbon content, the higher the initial hardness and the more pronounced the softening with tempering temperature). Krauss [35] distinguishes two extreme domains for tempering: Low-temperature tempering is typically applied between 150 and 200 °C and produces very high strength steels while high-temperature tempering is applied roughly between 500 and 650 °C, depending on desired properties and alloying. In this sense, the tempering treatment applied in this study belongs to the intermediate domain. In principle, a trade-off between low and high tempering temperature cannot be ruled out since the fracture toughness is expected to reduce after increasing the tempering temperature. In our opinion, this is not a source of concern for this specific application since fracture toughness is relevant for cracked components and steel springs are subjected to several quality controls to guarantee the absence of surface cracks. For this reason, fatigue lifespan in this case is controlled by the initiation process. Another consequence derived from the softening experienced by the steel is that the roughness of the specimens increases significantly. This is consistent with the multiple initiation sites observed in the fractographic analysis on high-temperature treated specimens. It has been widely described [31,33] that the number of initiators increases with the stress amplitude and that, for low stresses, fatigue usually displays a single initiator. In this experimental study, an amplitude $\sigma_a = 400$ MPa was applied in all cases, which is relatively moderate for a material with a yield stress exceeding 2000 MPa. This suggests that the change in local mechanical properties as a consequence of the tempering temperature is significantly severe in terms of the initiation of fatigue.

Fig. 10 reveals another relevant aspect. It contains the same information as Fig. 1 but, in this case, the datasets (each one corresponding to a fabrication condition) were sorted in increasing order of the mean value of the number of cycles to failure. A clear positive correlation can be seen between the dispersion within each group and its lifespan (scatter clearly increases from left to right in the figure). The available

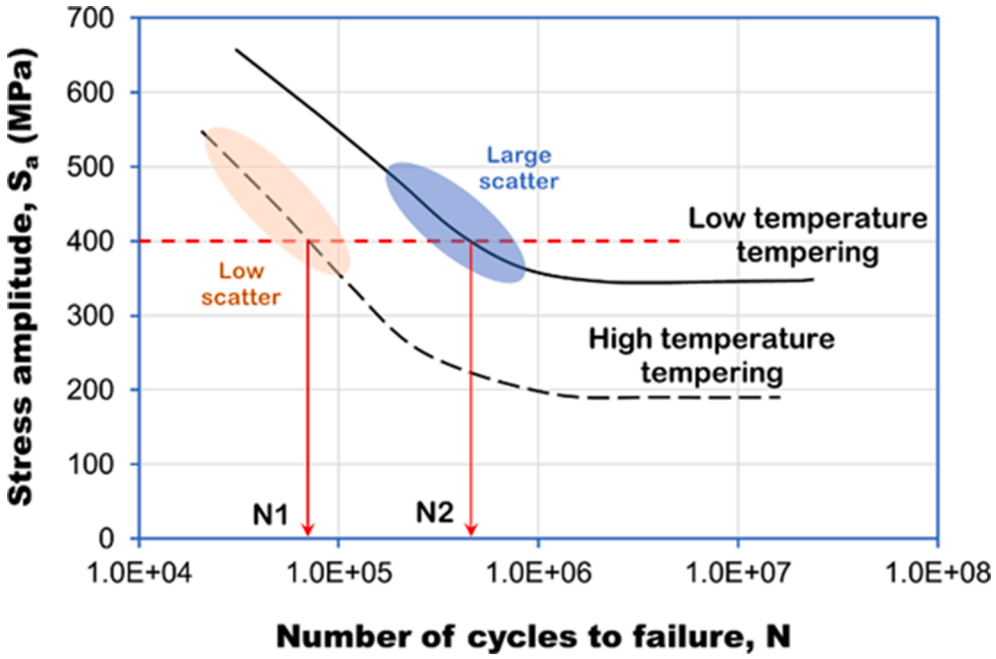

Fig. 11. Schematic description showing the influence of the tempering temperature on the S-N curves. The upper curve would correspond to the material treated at low temperature and the lower one to the steel treated at high temperature. For the stress amplitude applied in the fatigue tests, $\sigma_a = 400$ MPa, the former material would work in the low dispersion region while the latter would belong to the large scatter domain.

empirical evidence [8,31] proves that the typical dispersion along an S-N curve is not constant but tends to increase in the region of low amplitude, close to the fatigue limit. This empirical regularity is explained because large stresses promote the formation of a large number of initiators, reducing the variability of the process and, as a consequence, the dispersion in the number of cycles to failure. The pattern represented in Fig. 10 can be adequately interpreted by considering simplistically two materials, namely, one with low and another with high temperature hardening, whose hypothetical S-N curves are sketched in Fig. 11. As can be seen, for the amplitude imposed in the rotating bending fatigue tests, $\sigma_a = 400$ MPa, the material with high-temperature tempering would work in the upper part of its S-N curve, with reduced dispersion (due to the large number of initiators), while the material with low-temperature tempering would be located in the lower region of its S-N curve, with appreciable dispersion (due to the reduced number of initiators). In short, this mechanistic interpretation justifies the relationship observed between the number of cycles to failure and the dispersion within each group. In this case, manufacturing conditions before code 5,376,620 were subjected to high temperature tempering while the rest of combinations received a low temperature tempering.

5. Conclusions

This paper exemplifies the possibilities that the modelling methods based on ML algorithms can provide in the fields of steelmaking and design of steel components. The role played by the temperature of the tempering treatment applied as a part of the fabrication of springs -that has been elucidated through ML algorithms-, has provided a substantial improvement of the fatigue strength of the material. Specifically, the fatigue lifespan for the experimental conditions imposed (rotating bending fatigue tests, $R = -1$ and $\sigma_a = 400$ MPa) have improved from $\sim$ 25000 cycles to $\sim$ 50000 cycles. The contributions derived from this research that could be beneficial not only for other researchers concerned with the fatigue behavior of steel springs but, in general, with the final mechanical properties for materials after a complex manufacturing process are threefold: simplification, prediction and improvement. First, the implementation of ML algorithms carried out in this study has enabled to identify the attributes with the greatest relevance on the fatigue lifespan of the suspension springs among the 297 variables involved in the whole manufacturing process. To do this, the method of recursive feature elimination with cross-validation has been employed to aggressively reduce the dimensionality of the problem. Second, a reliable predictive ML regression model, based on the GB algorithm has been trained and validated, providing a R^2 in the test set of 0.877. Third, this model has been subsequently exploited to identify the variables of the manufacturing process with the greatest influence on the fatigue life of the material. In this case, it has been observed that one single attribute, the temperature during the tempering heat treatment of the steel, accounts for most of the observed variability, even screening the influence of the rest of variables. In fact, it has been observed that when this temperature exceeds a certain threshold, the fatigue life plummets. In this sense, it can be concluded that the manufacturing process would benefit notably if the tempering temperature is reduced below the indicated threshold. In addition, in this study the data-driven approach was combined with an exhaustive experimental study and the interpretation of the results has been based on concepts that belong to the subject of mechanics of materials and, especially, to the field of the fatigue behavior of metallic materials.

Declaration of Competing Interest

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

Acknowledgements

This Project was carried out with a financial grant from the program INNOVA 22018 (Grant 2018/INN/44). The financial contributions from GSW, the Government of Cantabria and the European Union through the program FEDER Cantabria are gratefully acknowledged. The authors would also like to express their gratitude to the technical staff of GSW and, especially, to Mr. Rafael Piedra, Mr. Santiago Pascual and Mr.

Enrique Gutiérrez, without whom it would not have been possible to do this research.

References

[1] Ritchie H. Cars, planes, trains: where do CO2 emissions from transport come from? Our World Data 2020. https://ourworldindata.org/co2-emissions-from-transport (accessed August 2, 2021).

[2] He L, Wang Z, Ogawa Y, Akebono H, Sugeta A, Hayashi Y. Machine-learning-based investigation into the effect of defect/inclusion on fatigue behavior in steels. Int J Fatigue 2022;155:106597. https://doi.org/10.1016/j.ijfatigue.2021.106597.

[3] Gan L, Wu H, Zhong Z. On the use of data-driven machine learning for remaining life estimation of metallic materials based on Ye-Wang damage theory. Int J Fatigue 2022;156:106666. https://doi.org/10.1016/j.ijfatigue.2021.106666.

[4] Barbosa JF, Correia JAFO, Júnior RCSF, Jesus AMPD. Fatigue life prediction of metallic materials considering mean stress effects by means of an artificial neural network. Int J Fatigue 2020;135:105527. https://doi.org/10.1016/j.ijfatigue.2020.105527.

[5] Murakami Y, Takada M, Toriyama T. Super-long life tension-compression fatigue properties of quenched and tempered 0.46% carbon steel. Int J Fatigue 1998;20:661–7. https://doi.org/10.1016/S0142-1123(98)00028-0.

[6] Lonkwic P, Różyło P. Theoretical and Experimental Analysis of Loading Impact From the Progressive Gear on the Lift Braking Distance With the Use of the Free Fall Method. Adv Sci Technol Res J 2016;10:103–9. https://doi.org/10.12913/22998624/62628.

[7] Cummings HN, Stulen FB, Schulte W. Tentative fatigue strength reduction factors for silicate-type inclusions in high-strength steels. ASTM Proceeding - 1958;58:505–14.

[8] Schijve J, editor. Fatigue of Structures and Materials. Dordrecht: Springer Netherlands; 2009.

[9] Qian G, Li Y, Paolino DS, Tridello A, Berto F, Hong Y. Very-high-cycle fatigue behavior of Ti-6Al-4V manufactured by selective laser melting: Effect of build orientation. Int J Fatigue 2020;136:105628. https://doi.org/10.1016/j.ijfatigue.2020.105628.

[10] Ransom JT. The effect of inclusions on the fatigue strength of SAE 4340 steels. Trans Am Soc Met 1954;46:1254–69.

[11] Ferreño D, González R, Carrascal IA, Cuartas M, García D, Eraña R, et al. Investigation through artificial neural networks on the influence of shot peening on the hardness of ASTM TX304HB stainless steel. J Test Eval 2021;49(1):20180819. https://doi.org/10.1520/JTE20180819.

[12] Tsumura K. Hierarchically Aggregated Optimization Algorithm for Heterogeneously Dispersed Utility Functions. IFAC-PapersOnLine 2017;50(1):14442–6. https://doi.org/10.1016/j.ifacol.2017.08.2287.

[13] Gasik M, editor. Handbook of Ferroalloys: Theory and Technology. 1st editio. Butterworth-Heinemann; 2013.

[14] Ruiz E, Cuartas M, Ferreno D, Romero L, Arroyo V, Gutierrez-Solana F. Optimization of the fabrication of cold drawn steel wire through classification and clustering machine learning algorithms. IEEE Access 2019;7:141689–700. https://doi.org/10.1109/ACCESS.2019.2942957.

[15] Wente EF, Nutting J, Wondris EF. Steel. Encycl Br 2019. https://www.britannica.com/technology/steel (accessed May 20, 2021).

[16] Cuartas M, Ruiz E, Ferreño D, Setién J, Arroyo V, Gutiérrez-Solana F. Machine learning algorithms for the prediction of non-metallic inclusions in steel wires for tire reinforcement. J Intell Manuf 2021;32(6):1739–51. https://doi.org/10.1007/s10845-020-01623-9.

[17] Ruiz E, Ferreño D, Cuartas M, Lloret L, Ruiz Del Árbol PM, López A, et al. Machine learning methods for the prediction of the inclusion content of clean steel fabricated by electric arc furnace and rolling. Metals (Basel) 2021;11. https://doi.org/10.3390/met11060914.

[18] Ruiz E, Ferreño D, Cuartas M, López A, Arroyo V, Gutiérrez-Solana F. Machine learning algorithms for the prediction of the strength of steel rods: an example of data-driven manufacturing in steelmaking. Int J Comput Integr Manuf 2020;33(9):880–94. https://doi.org/10.1080/0951192X.2020.1803505.

[19] ISO 1143:2010. Metallic materials-Rotating bar bending fatigue testing. ISO Stand., 2010, p. 26.

[20] Castillo E, Fernandez-Canteli A. A Unified Statistical Methodology for Modeling Fatigue Damage. Springer Netherlands; 2009. https://doi.org/10.1007/978-1-4020-9182-7.

[21] Castillo E, Fernández-Canteli A, Ruiz-Ripoll ML. A general model for fatigue damage due to any stress history. Int J Fatigue 2008;30:150–64. https://doi.org/10.1016/j.ijfatigue.2007.02.011.

[22] Qian G, Lei W-S. A statistical model of fatigue failure incorporating effects of specimen size and load amplitude on fatigue life. Philos Mag 2019;99(17):2089–125. https://doi.org/10.1080/14786435.2019.1609707.

[23] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn. J Mach Learn Res 2011;12:2825–30.

[24] Geron A. Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly M. 2017.

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

[26] Wolpert DH. The Supervised Learning No-Free-Lunch Theorems. 6th Online World Conf. Soft Comput Ind Appl, Springer 2001:1–20. https://doi.org/10.1007/978-1-4471-0123-9_3.

[27] Wolpert DH, Macready WG. No free lunch theorems. IEEE Trans Evol Comput 1997;1:67–82. https://doi.org/10.1007/978-3-662-62007-6_12.

[28] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. London: Chapman and Hall/CRC; 1984.

[29] ISO 4287:1997. Geometrical Product Specifications (GPS) — Surface texture: Profile method — Terms, definitions and surface texture parameters. ISO Stand., ISO International Organization for Standardization; 1997, p. 1–5.

[30] Castillo E, Ramos A, Koller R, López-Aenlle M, Fernández-Canteli A. A critical comparison of two models for assessment of fatigue data. Int J Fatigue 2008;30:45–57. https://doi.org/10.1016/j.ijfatigue.2007.02.014.

[31] Schijve J. Fatigue predictions and scatter. Fatigue Fract Eng Mater Struct 1994;17:381–96. https://doi.org/10.1111/j.1460-2695.1994.tb00239.x.

[32] Kottegoda N. Applied Statistics for Civil and Environmental. 2008.

[33] de Forest A V. The rate of growth of fatigue cracks. J Appl Mech 1936;3:A-23-A-25.

[34] Canale LCF, Vatavuk J, Totten GE. 12.02 - Introduction to Steel Heat Treatment. In: Hashmi S, Batalha GF, Van Tyne CJ, Yilbas BBT-CMP, editors., Oxford: Elsevier; 2014, p. 3–37. https://doi.org/10.1016/B978-0-08-096532-1.01202-4.

[35] Krauss G. 12.11 - Quench and Tempered Martensitic Steels: Microstructures and Performance. In: Hashmi S, Batalha GF, Van Tyne CJ, Yilbas BBT-CMP, editors., Oxford: Elsevier; 2014, p. 363–78. https://doi.org/10.1016/B978-0-08-096532-1.01212-7.