JMEPEG (2021) 30:434–443
https://doi.org/10.1007/s11665-020-05345-0


1059-9495/$19.00

# Prediction of Mechanical Properties of Steel Tubes Using a Machine Learning Approach

*Marcelo V. Carneiro, Turíbio T. Salis, Gustavo M. Almeida, and Antonio P. Braga*

Submitted: 4 June 2020 / Revised: 12 October 2020 / Accepted: 7 November 2020 / Published online: 5 January 2021

**Steel tubes produced in steelmaking plants are generally subjected to severe in-service conditions. Hence, quality control plays a key role in this process. The bottleneck is that this information is made available only after tube production from laboratory analysis. Given process complexity and current data availability, this work employs a series of machine learning techniques, namely neural networks, random forests and gradient boosting trees, to predict critical mechanical properties for steel tubes, namely yield strength, ultimate tensile strength and hardness. The model performance was kept high by combining different variable selection procedures. The prediction error was less than the inherent variability of each mechanical property, i.e., it is equal to 20 MPa for yield strength and ultimate tensile strength, and to 2 HRC, for hardness. This information in advance allows interventions before complete tube production contributing to more stable operations and, ultimately, to reduce rework and customer lead time. In sequence, an optimization problem for set point definition is illustrated. The neural predictive model previously identified for the yield strength was used in this application, exploring its predictive capabilities. The optimal solution yielded to lower amount of molybdenum and tube exit temperature from the tempering furnace, while keeping quality aspects, which means reduction in material and energy costs. Concluding, steelmaking processes, which are complex by nature, can strongly benefit from data-driven approaches, since data availability and computational processing are no longer a problem.**

**Keywords** mechanical property, steel tube, machine learning, variable selection, process optimization, steelmaking plant

## 1. Introduction

Steel tubes are usually subjected to severe in-service conditions. As a result, specific mechanical properties are required by industrial sectors in general, including oil and gas. Such properties depend mainly on tube geometry, chemical composition of the alloy steel and on heat treatment conditions during tube manufacturing (Ref 1, 2). Before being released to the final (external) customer, each batch of tubes must wait for laboratory tests. In addition to the increase in stock, out-of-specification tubes generate rework and even production losses. This creates a bottleneck in the process, since laboratory analysis may require a few days to become available. Hence, prediction of tube properties, prior to complete tube production, would be very beneficial from an operational point of view. It could also prevent infrequent or even improper process parameter values over time, contributing to more stable operations. Ultimately, it would improve tube quality control in several aspects such as design optimization, cost reduction and customer lead time, to mention a few.

The highly nonlinear and complex relationships involving steel tube production hinder the use of a pure mathematical description and, then, final properties' estimation of the tubes (Ref 3). Such restricted scenario on one side, and the current availability of massive amounts of data on the other side, favors the use of machine learning techniques. Data-driven approach has shown to be successful to describe complex industrial processes.

In the last decades, studies employing simulated or real data sets have been carried out with the aim to predict mechanical properties of steel products in general. Most of them applied linear regression and neural networks. For example, Pattanayak et al. (Ref 3) employed neural networks to a multi-objective optimization problem to improve mechanical properties of API (American Petroleum Institute) grade microalloyed pipeline steel tubes, namely strength, impact toughness and ductility, using chemical composition and processing parameters as design variables. Sampaio et al. (Ref 4) applied ensemble learning through a set of neural network models in a thermal treatment plant of steel tubes later used for online monitoring. Agrawal et al. (Ref 5) employed multivariate polynomial regression, decision trees and neural networks to predict fatigue properties of steels, and Jones et al. (Ref 6) used linear and nonlinear regression analysis and neural networks to predict mechanical properties of rolled steels.

After identification and validation, a model can be employed for several purposes, namely operating safety, clean production and economic efficiency. This is also the reality of steelmaking plants. For example, machine learning techniques may contribute to production planning and scheduling, where one challenge is the varied product portfolio (Ref 7–9). Other area

**Marcelo V. Carneiro** and **Turíbio T. Salis**, Vallourec Soluções Tubulares do Brasil, Av. Olinto Meireles, 65, Barreiro, Belo Horizonte, MG 30640-010, Brazil; **Gustavo M. Almeida**, Department of Chemical Engineering, Federal University of Minas Gerais, Av. Pres. Antonio Carlos, 6627, Pampulha, Belo Horizonte, MG 31270-901, Brazil; and **Antonio P. Braga**, Department of Electronic Engineering, Federal University of Minas Gerais, Av. Pres. Antonio Carlos, 6627, Pampulha, Belo Horizonte, MG 31270-901, Brazil. Contact e-mail: galmeida@deq.ufmg.br.

with great opportunities is process monitoring, with a particular interest in developing soft sensors (Ref 10). One example refers to severe operating conditions, as the determination of the steel temperature in the ladle furnace. Another concerns monitoring the output yield of steel, what is important for the control of the conversion of scrap and iron ore to steel lingots (Ref 11). There is also interest in the inference of quality parameters (as is the case of the present work), whose information is usually made available late from laboratory analysis (Ref 4). Another very profitable use of machine learning refers to process optimization mainly the multi-objective approach (Ref 3). This is strengthened, from one hand by increasingly tighter market, government and society regulations, and on the other hand by the complexity of the operations, given its multivariate and nonlinear nature. Naturally, adjustments are required prior to its full implementation on the shop floor. In addition to the operational issue, it is essential to address human, organizational and technological aspects to succeed (Ref 12, 13). The models in this work were developed with the participation of the process team of the steelmaking plant of the case study. Also, given the use of the R free statistical software (Ref 14), computational costs with respect to hosting and general maintenance tasks are relatively low. Practices like these facilitate implementations on the shop floor.

The present study develops predictive models for mechanical properties of steel tubes using machine learning techniques. The real case study refers to a steelmaking plant in Brazil and, more specifically, to its heat treatment process regarding quenching and tempering operations. The common mechanical characteristics of interest, namely yield strength, ultimate tensile strength and hardness, are predicted. They were defined to promote a fine adjustment in the quality control of this unit. The models are based on neural networks, random forests and gradient boosting trees. Next, a process optimization problem for defining set points is illustrated using the previously identified predictive model for yield strength.

Section 2 depicts the methodology, and section 3 describes the case study and its data set. The results of the predictive models are presented and discussed in section 4. Section 5 illustrates a machine learning application in steelmaking by presenting an optimization problem that employs one of the predictive models previously obtained. Final considerations are given in section 6.

## 2. Methodology for Construction of the Predictive ML Models

Figure 1 depicts the steps of the methodology adopted for constructing the predictive models. Each step is described next.

### 2.1 Data Preprocessing

Given the objective of this study, the raw data set is composed by operation and quality data. The former is obtained through the Quenching and Tempering Material Tracking System that has the operational history of every tube produced. This system is composed of process variables (flow rates, temperatures and pressures), design variables (diameter and wall thickness) and process parameters (the soaking time index and the Tsuchiyama parameter). The latter, obtained from the Laboratory Information Management System, stores the chemical composition analysis of the steel and the mechanical property tests, which are the response variables to be predicted by the machine learning models. Both concern the heat treatment process. The role of this step is to construct a reasonable working data set, which is crucial in any data-driven model description. Variables with considerable register errors, missing values, or relative low variance are then removed. Besides literature, it is essential to consider process team expertise for data cleaning.

### 2.2 Variable Selection

The goal is to identify the most compact and informative variable subset, that is, with minimum redundancy and maximum relevance among variables (Ref 15–17). Given the previous variable selection based on literature and process expertise, a second one using two statistical stepwise methods, namely forward and backward, was carried out (Ref 18). These procedures, with the objective of a fine adjustment on the number of predictors, respectively, add/remove variables, one at a time, to/from a linear regression model. They used the root-mean-squared error as decision criterion. The variables were initially normalized in the [0, 1] range to give them the same importance.

### 2.3 Model Identification and Evaluation

For the prediction of mechanical properties, two regression techniques were investigated, namely, artificial neural networks (ANN) and tree ensemble models. The first approach used the well-known multi-layer perceptron (MLP) architecture (Ref 19). The starting point of the second approach is the decision tree technique, which partitions the prediction space into subregions. Given a new sample, model output is generally given by the average response of the training samples in the closer partition. The tree ensemble model is composed of a set of decision trees (Ref 20). An ensemble approach often improves any individual performance. The present work employs random forests (RF) and gradient boosting trees (XGB) ensemble algorithms. To mitigate the problem of generalization capacity, the $k$-fold cross-validation procedure

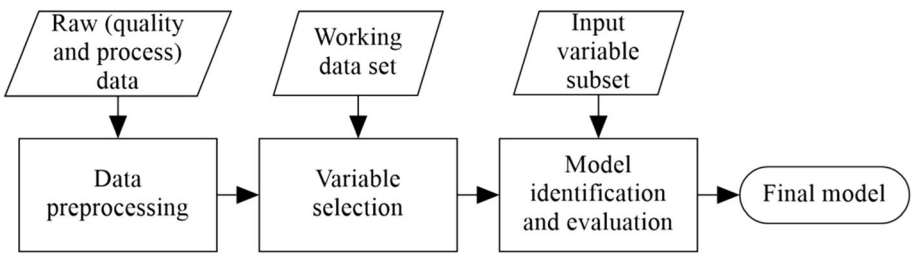


**Fig. 1** Methodology steps for obtaining the predictive machine learning models

was used for model identification and selection. This procedure randomly splits the data set into $k$ subsets, where $(k-1)$ are used for model estimation, and the remaining, for model evaluation. They are, respectively, called training and validation sets. This is repeated $k$ times, each one with a particular validation set. The root-mean-squared error (RMSE) and the mean absolute error (MAE) performance metrics are calculated in each run. They are, respectively, given by Eqs. 1 and 2, where $y$ is the target value, $\hat{y}$ is the corresponding model estimate, and $N$ is the size of the validation set. It aims at selecting the model that provides the lowest average (considering the $k$ runs) RMSE and MAE values among all candidate models. The generalization capacity of the selected models is also evaluated by the residues analysis, by verifying normality and proximity to zero mean.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad \text{(Eq 1)}$$

$$\mathrm{MAE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|} \quad \text{(Eq 2)}$$

A critical task in machine learning concerns model tuning. After some trials, the parameter set investigated for each technique during cross-validation is given as listed below.

- Multi-layer perceptron (MLP): Number of neurons in the first hidden layer [10:2:40], number of neurons in the second hidden layer [0:2:20], number of epochs [100:100:400], regularization coefficient [0.0001:0.0001:0.001] and learning rate parameter [0.01:0.01:0.1].
- Random forests (RF): Number of variables randomly chosen for each branch [4:2:12], number of trees [100:100:1000] and maximum number of nodes for each tree [50:50:300].
- Gradient boosting trees (XGB): Regularization coefficient [0:0.1:0.5], learning rate parameter [0.01:0.02:0.2], maximum depth of each tree [3:3:12] and number of rounds [200:100:800].

For instance, the number of neurons in the first hidden layer of the neural network models (MLP) was varied from 10 up to 40 in steps of 2. Two models are obtained for each parameter set, one for each variable selection method (that is, forward and backward). For comparison purposes, linear regression analysis (LM) was also applied in this work (Ref 18).

## 3. Case Study

The case study concerns the heat treatment plant of a Vallourec unit in Brazil that produces seamless steel tubes. Their main equipments are the quenching tank and the tempering furnace, as shown in Fig. 2. In the quenching operation, the tube is heated and then cooled abruptly for hardness increasing. Next, in the tempering operation, this tube is reheated during a certain time interval for adjusting internal stresses that arose in the previous operation.

Two data sets were used for modeling, namely the Quenching and Tempering Material Tracking System, and the Laboratory Information Management System, as previously mentioned. They consist of around two years of laboratory analysis. The variables sampled from them were based on literature and process expertise. Table 1 shows this initial selection containing twenty-seven variables. All are related to the heat treatment unit. Besides the inputs often associated with quenching and tempering processes, related to time, temperature and chemical composition, the Tsuchiyama parameter was also considered. According to Gomes et al. (Ref 21), it presents good correlations with mechanical properties of heat-treated materials. By providing more information than the soaking time index, in conjunction with the average tube temperature at the furnace exit, it is more suitable for more complex thermal cycles, as in this work. The Tsuchiyama parameter can be seen as an improvement of the Hollomon-Jae parameter. In short, it is obtained by dividing the entire thermal cycle into small intervals for which an equivalent time is calculated at the reference temperature (Ref 4). Also, the input for the heat treatment process is delivered by the factory itself in an integrated manner. That is, it is in line with a steel refining plant, a casting unit and a hot rolling mill line. Given the knowledge of their quality control policies, mechanical or chemical analyses of this input material were not considered in modeling.

The predicted mechanical properties are also shown in Table 1. They usually reflect the material behavior in response to physical forces, which are measured through a series of standardized mechanical tests. Yield strength (in MPa) and the ultimate tensile strength (in MPa) are obtained by a tensile test. Whereas the former is the stress at which a material starts to suffer plastic (permanent) deformation, the latter is given by the maximum stress it can withstand before failing. Hardness (in HRC; Rockwell C scale) is obtained from a hardness testing. The smaller the indentation given an indenter and a constant load on material surface, the harder the material (Ref 2, 22). Hardness is not an intrinsic property of a material, but a measure of resistance concerning plastic deformation. This property often obeys a relationship with yield strength, which depends on material structure. Song et al. (Ref 23), Hashemi (Ref 24) and Zhu and Xuan (Ref 25) reported a linear correlation between these properties for steels in general. Figure 3(a) shows the relationship between yield strength and hardness found in this work, from which an approximate linear behavior is also verified (Ref 26). The ranges for yield strength and hardness considered in this work are equal to [550, 640] MPa and [14.5, 21.0] HRC, respectively.

The steelmaking plant of the case study produces three grades of steel (that is, steel families 1, 2 and 3) that serve distinct specifications of the oil and gas industry. They differ from each other by combining the chemical composition of the steel and the external diameter and wall thickness of the tube. This segmentation, which may vary from plant to plant, is adopted in the unit of the case study. Table 2 shows the number of samples for each mechanical property, extracted for each steel grade. Due to a low number of samples, no models were obtained for hardness in family 2.

Since the steel grades have their proper modes of operation in the process, the operation data in the Quenching and Tempering Material Tracking System (section 2) were used to investigate them. Given problem dimensionality, with fifteen out of the twenty-seven variables, by disregarding chemical compositions and design variables (Table 1), this was accomplished using principal component analysis (PCA). PCA is a

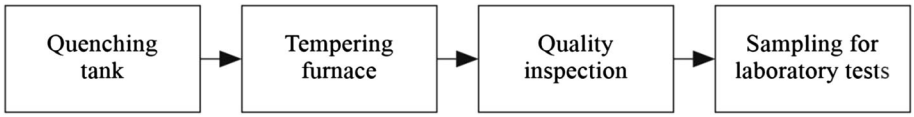


**Fig. 2** Main equipments of the heat treatment plant of the case study

**Table 1 Variables collected for each tube**

| Input | Unit |
|---|---|
| Pipe diameter | mm |
| Pipe wall thickness | mm |
| Chemical composition (a total of 14 elements) | % |
| Equivalent carbon content | % |
| Outlet temperature at hardening furnace | °C |
| Soaking time at hardening furnace | s |
| Tsuchiyama parameter for hardening furnace | – |
| Outlet temperature at tempering furnace | °C |
| Soaking time at tempering furnace | s |
| Tsuchiyama parameter for tempering furnace | – |
| Retreatment index | – |
| Immersion time in the quenching tank | s |
| Water flow rate in the quenching tank | l/s |
| Water pressure in the quenching tank | bar |
| **Output (mechanical properties of the steel pipes)** | **Unit** |
| Yield strength (YS) | MPa |
| Ultimate tensile strength (UTS) | MPa |
| Hardness (H) | HRC (Rockwell C scale) |

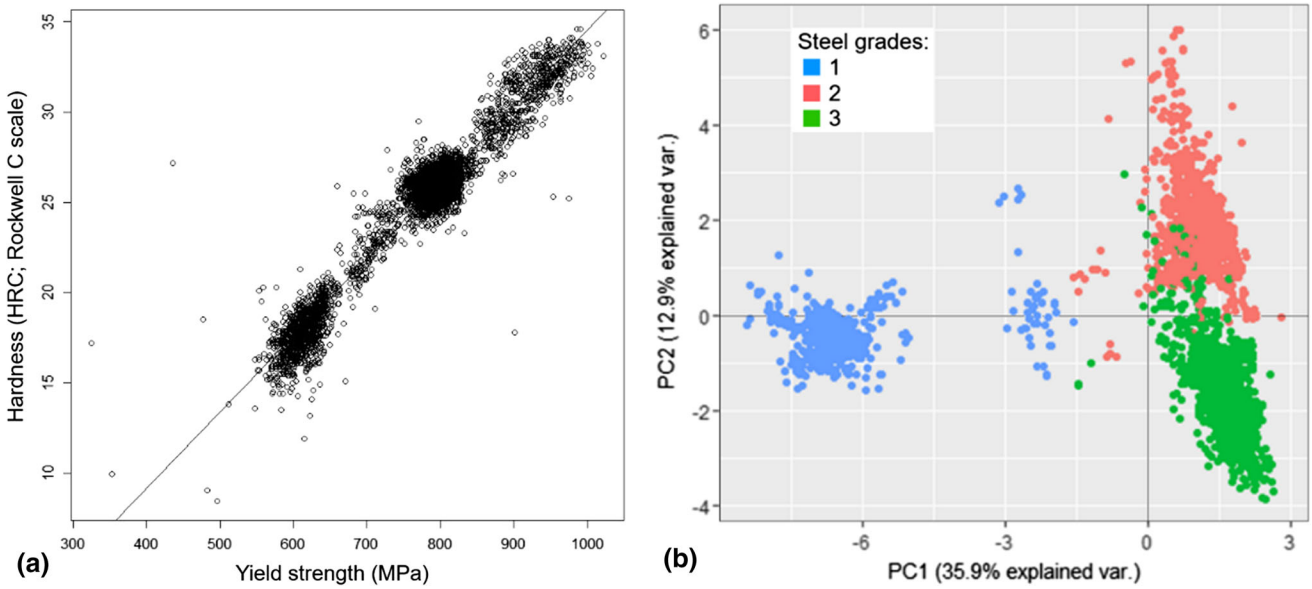


**Fig. 3** (a) Yield strength and hardness relationship in this work and (b) steel families grouping by principal component analysis

**Table 2 Number of samples for each mechanical property, extracted per steel family**

| Steel family | Yield strength | Ultimate tensile strength | Hardness |
|---|---|---|---|
| 1 | 833 | 833 | 792 |
| 2 | 407 | 407 | – |
| 3 | 162 | 162 | 134 |

multivariate statistical technique commonly used for dimensionality reduction. This is achieved by an orthogonal rotation of the original coordinate system of the original variables. The axes of this resulting system are given by the called principal components (PC), which are linear combinations of the original variables. The more correlated these variables are, the greater the problem reduction (Ref 27). Figure 3(b) depicts the score plot for the first two principal components that explain almost 50% (35.9% by PC1 + 12.9% by PC2) of the total variation of the operation data. Given the two reference lines ($x = 0$ and $y = 0$), it can be seen that each operating condition, that is, grade of steel, is located in a particular quadrant of the plot. Thus, it can be seen that these grades differ considerably with respect to operating conditions. Besides, data imbalance can be verified for all mechanical properties (Table 2). This condition generally favors the majority class, impairing the overall model performance in case of considering all of them together (Ref 28). Thus, to achieve greater performance, the predictive models were obtained separately, one for each steel family.

## 4. Results and Discussion

Predictive models for each mechanical property, namely yield strength (YS), ultimate tensile strength (UTS) and hardness (H), were obtained and evaluated. Firstly, the forward and backward variable selection methods were applied to reduce the initial subset of twenty-seven candidate predictors (Table 1). Dimensionality reduction is usually applied to reduce computational processing and to mitigate the numerical impact of redundant information. For instance, chemical compositions are not independent of each other since they are extracted from a common specimen. This aspect can even worsen model performance. By combining steel family, mechanical property and variable selection method, different subsets of predictors were obtained. As an example, Table 3 summarizes the selected groups of variables for steel family 1. In this case, there are 14 and 16 variables for yield strength, 17 and 12, for ultimate tensile strength, and 19 and 10, for hardness, given the forward

**Table 3 Variable selection (number of variables in parentheses) according to the stepwise methods (forward (For.) and backward (Back.)), for each mechanical property, given steel family 1**

| Number | Variable | Yield strength | | Ultimate tensile strength | | Hardness | |
|---|---|---|---|---|---|---|---|
| | | For. (14) | Back. (16) | For. (17) | Back. (12) | For. (19) | Back. (10) |
| 1 | Pipe diameter | | | | | | |
| 2 | Pipe wall thickness | | | | | | |
| 3 | C (Carbon) | | | | | | |
| 4 | B (Boron) | | | | | | |
| 5 | Co (Cobalt) | | | | | | |
| 6 | Cr (Chromium) | | | | | | |
| 7 | Cu (Copper) | | | | | | |
| 8 | Mn (Manganese) | | | | | | |
| 9 | Mo (Molybdenum) | | | | | | |
| 10 | N (Nitrogen) | | | | | | |
| 11 | Nb (Niobium) | | | | | | |
| 12 | Ni (Nickel) | | | | | | |
| 13 | Si (Silicon) | | | | | | |
| 14 | Ti (Titanium) | | | | | | |
| 15 | V (Vanadium) | | | | | | |
| 16 | W (Tungsten) | | | | | | |
| 17 | Equivalent carbon content | | | | | | |
| 18 | Outlet temperature at hardening furnace | | | | | | |
| 19 | Soaking time at hardening furnace | | | | | | |
| 20 | Tsuchiyama parameter for hardening furnace | | | | | | |
| 21 | Outlet temperature at tempering furnace | | | | | | |
| 22 | Soaking time at tempering furnace | | | | | | |
| 23 | Tsuchiyama parameter for tempering furnace | | | | | | |
| 24 | Retreatment index | | | | | | |
| 25 | Immersion time in the quenching tank | | | | | | |
| 26 | Water flow rate in the quenching tank | | | | | | |
| 27 | Water pressure in the quenching tank | | | | | | |

and backward procedures, respectively. These relatively smaller sizes corroborate with the high redundancy in the initial variable subset. Also, it can be noted that the subset sizes and the variables themselves vary for a property and steel family in particular. However, both subsets yielded to similar prediction performances (Tables 4, 5, 6). Thus, obtaining association between statistically based results and physico-chemical phenomena is usually not easy. However, a general map can be obtained.

Despite the differences between the variable subsets, there are variables that commonly appear in most of them. For instance, it is known that carbon, manganese, silicon and vanadium are positively correlated with resistance, while all of them, except vanadium, with hardness. This first group of chemical elements usually appeared for yield strength and ultimate tensile strength, whereas the second one, without vanadium, for hardness. Redundant information is critical for chemical elements considering the proper specifications of a steel family in particular. This fact may explain the selection of less expected variables. The design variables, namely pipe diameter and wall thickness, were selected for all subsets. This may be due to the fact that the definition of a steel family is a function of them, as previously mentioned (section 3). This is also the case for the tube outlet temperatures and process parameters, namely soaking time and Tsuchiyama index, which may vary between steel families. These variables are related to the hardening and tempering furnaces. The water flow rate in the quenching tank also appears in all subsets. This variable is responsible for the sudden cooling of the tubes, whose resulting difference in temperature affects tube mechanical characteristics (Ref 1, 2). Thus, in general lines, the selected subsets make sense in a practical point of view.

For simplification purposes, next step shows the results obtained for steel family 1. The analysis for families 2 and 3 is similar.

### 4.1 Yield Strength (YS)

A key point is the definition of the acceptable prediction error for model estimates in relation to laboratory results. This was carried out jointly by the process team and the laboratory experts. Given a steel family and a mechanical property in particular, differences in laboratory measurements between tubes presenting very similar operating conditions were calculated. After evaluating the distribution of a set of such differences, the 95th percentile of the maximum one was used as a measure of accuracy for the (steel family, mechanical property)-combination. This threshold can be seen as an inherent variability, as a result of all uncertainties involving laboratory tests and plant sensors. For yield strength, it is equal to 20 MPa.

Table 4 presents the best result obtained for each machine learning technique (RF: random forests; XGB: gradient boosting trees; MLP: neural network; and LM: linear regression), given the parameter set combinations (section 2). It shows the RMSE (Eq. 1) and the MAE (Eq. 2) values for the residues, given by the differences between laboratory analyses and corresponding model estimates. The best average (*avg*) and standard deviation (*std*) values, for each variable selection method (forward and backward), are highlighted in bold. From MAE, it can be observed that all models meet the acceptable prediction error (20 MPa). Indeed, they perform similarly, and anyone could be initially chosen. However, most accurate models, with lower MAE values, should be adopted in practice. Also, the backward procedure resulted in less complex models (fourteen inputs) in relation to the forward one (sixteen variables), without loss of generalization. This strategy was then more efficient in generating a compact and relevant input variable subset. Moreover, the MLP neural network models provided greater generalization capacity. Figure 4(a) and (b) show the parity relations between targets and model estimates for yield strength. The dashed lines, parallel to the line defining perfect correlation, comprise the acceptable prediction error. As desired, the distribution of the residues is approximately normal with zero mean.

**Table 4 Selected models for yield strength (MPa), given steel family 1 (*avg.*: average, *std.*: standard deviation)**

| Model | RMSE avg. | RMSE std. | MAE avg. | MAE std. |
|---|---|---|---|---|
| Forward alg. (16 variables) | | | | |
| RF | 12.22 | 1.34 | 9.44 | 1.00 |
| XGB | 11.05 | 1.30 | 8.51 | 0.90 |
| MLP | **10.86** | **0.74** | **8.50** | **0.52** |
| LM | 11.06 | 0.92 | 8.62 | 0.80 |
| Backward alg. (14 variables) | | | | |
| RF | 12.30 | 1.30 | 9.37 | 0.96 |
| XGB | 11.26 | 1.02 | 8.76 | **0.56** |
| MLP | 10.91 | **0.82** | **8.43** | 0.69 |
| LM | **10.90** | 0.87 | 8.50 | 0.58 |

### 4.2 Ultimate Tensile Strength (UTS)

The performance of all selected models for the ultimate tensile strength is also very similar (Table 5). All also meet the acceptable prediction error of 20 MPa, which was set in the same way as before for yield strength. The MLP models also resulted in the best generalization results (Fig. 4c and d). Regarding variable selection, the forward procedure, with twelve inputs, was more efficient in relation to the backward strategy, with seventeen variables, without loss of generalization capacity.

### 4.3 Hardness (H)

A prediction error less than 2 HRC was considered satisfactory for hardness. This threshold was set in the same way as before for yield strength. Table 6 shows that all models present similar results for MAE complying with this target. The results for RMSE are also very similar overall, mainly with respect to MLP, XGB and LM. Figure 4(e) and (f) shows the parity

**Table 5 Selected models for ultimate tensile strength (MPa), given steel family 1 (*avg.*: average, *std.*: standard deviation)**

| Model | RMSE avg. | RMSE std. | MAE avg. | MAE std. |
|---|---|---|---|---|
| Forward alg. (12 variables) | | | | |
| RF | 10.26 | 0.90 | 7.91 | 0.60 |
| XGB | 9.31 | 0.88 | 7.10 | 0.64 |
| MLP | **9.06** | 0.57 | **6.94** | **0.40** |
| LM | 9.49 | **0.47** | 7.31 | 0.42 |
| Backward alg. (17 variables) | | | | |
| RF | 10.17 | 0.89 | 7.75 | 0.60 |
| XGB | 8.91 | 0.94 | 6.74 | 0.67 |
| MLP | **8.64** | 0.92 | **6.62** | 0.75 |
| LM | 9.06 | **0.53** | 7.02 | **0.46** |

**Table 6 Selected models for hardness (H), given steel family 1 (*avg.:* average, *std.:* standard deviation)**

| Model | RMSE avg. | RMSE std. | MAE avg. | MAE std. |
|---|---|---|---|---|
| Forward alg. (10 variables) | | | | |
| RF | 0.66 | 0.09 | 0.49 | 0.06 |
| XGB | **0.64** | 0.07 | **0.48** | **0.05** |
| MLP | **0.64** | 0.07 | **0.48** | 0.06 |
| LM | 0.65 | **0.06** | 0.49 | **0.05** |
| Backward alg. (19 variables) | | | | |
| RF | 0.67 | 0.09 | 0.50 | 0.07 |
| XGB | 0.64 | **0.06** | 0.49 | **0.04** |
| MLP | 0.63 | 0.07 | **0.48** | 0.05 |
| LM | **0.62** | **0.06** | **0.48** | 0.05 |

relations between the targets and the MLP model estimates. Finally, it can be verified that the forward procedure, with ten variables, led to much less complex models in comparison with the backward strategy containing nineteen variables, without loss of performance.

Lastly, considering the results obtained for all mechanical properties, it can be verified the importance of investigating a set of machine learning techniques, in conjunction with variable selection methods, since none of the combinations is the most appropriate for all cases.

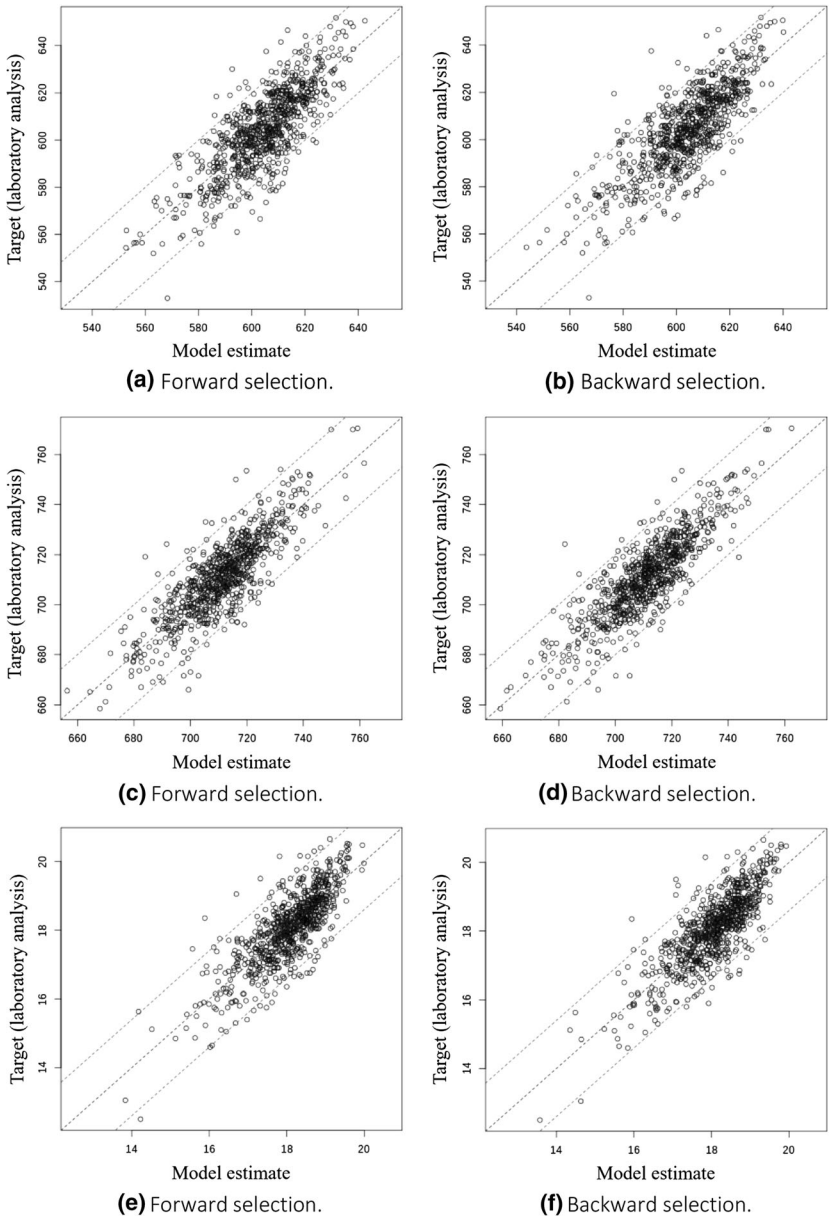


**Fig. 4** Parity relations between targets and MLP estimates for (a, b) yield strength, (c, d) ultimate tensile strength and (e, f) hardness, given steel family 1 (dashed lines comprehend the region of acceptable prediction error)

## 5. Process Optimization

Data-driven modeling has become more and more relevant in process industries worldwide for more rational decision making. After model identification and validation (section 4), an optimization problem is illustrated. Before that, a general framework is presented.

The objectives of the heat treatment process in steelmaking plants can be summarized in tube quality, defined by mechanical properties; production cost, associated with raw materials and process conditions; and in productivity, given by the processing time for tube production. In short, the challenge concerns the search for lower production cost and higher productivity, subject to mechanical property specifications for the tubes. The general framework for this optimization problem can be formulated according to the objective function ($F_{obj}$) in Eq. 3, and a set of restrictions, namely inequality equations (Eqs. 4-7), lower and upper bounds (Eq. 8) and equality equations (Eqs. 9-10), where $\vec{x}$ is the design variable vector.

$$\min F_{obj} = \beta \times C(\vec{x}) + \gamma \times P(\vec{x}) \quad \text{(Eq 3)}$$

subject to

$$\mathrm{YS}_{min} \leq \mathrm{YS}(\vec{x}) \leq \mathrm{YS}_{max} \quad \text{(Eq 4)}$$

$$\mathrm{UTS}_{min} \leq \mathrm{UTS}(\vec{x}) \leq \mathrm{UTS}_{max} \quad \text{(Eq 5)}$$

$$\mathrm{H}_{min} \leq \mathrm{H}(\vec{x}) \leq \mathrm{H}_{max} \quad \text{(Eq 6)}$$

$$d(\vec{x}) \leq d_{max} \quad \text{(Eq 7)}$$

$$\vec{x}_{min} \leq \vec{x} \leq \vec{x}_{max} \quad \text{(Eq 8)}$$

$$P_{Temp} = T_{Temp} \times [c + \log(t_{Soak})] \quad \text{(Eq 9)}$$

$$P_{Aust} = T_{Aust} \times [c + \log(t_{Soak})] \quad \text{(Eq 10)}$$

In Eq. 3, $C(\vec{x})$ describes the raw material costs, relative to the steel alloy elements, and the production costs, given by energy consumption. It can be obtained by $\vec{\omega} \times \vec{x}$, where $\vec{\omega}$ is a vector of weights. $P(\vec{x})$ represents line productivity, which refers to the time to heat each tube at the heat treatment furnaces. In other words, it concerns the cycle time between pipes or the quantity of pipes produced per hour. The shorter it is, the lower the energy consumption and the greater the productivity. This productivity term can be described by $\max(\vec{x}_p)$, where $\vec{x}_p$ represents the entire cycle time of the heat treatment furnaces, for which the shorter the better. These cycles are correlated with raw material and energy use. For instance, distinct combinations between steel alloying elements have different effects on furnace setup. Also, furnace cycle time can vary with the type of gas, whether LHV (lower heating value) higher or lower, air–fuel ratio and with the steel temperature at the furnace outlet. All these aspects are related to energy consumption. Thus, this optimization problem formulation seeks to minimize production costs in general while keeping quality control. Parameters $\beta$ and $\gamma$ are the weights for the cost and productivity terms, respectively. Equations 4 up to 6 are given by process models, one for each mechanical property, namely yield strength (YS), ultimate tensile strength (UTS) and hardness (H).

Equation 7 seeks to keep the design variables close to the usual operation. Besides avoiding infrequent, or even infeasible values, it helps to control the error in the predictive models (Eqs. 4–6). This becomes worse when only partial knowledge is available, which is the case of this work. The Manhattan metric, given by $d(\boldsymbol{a}, \boldsymbol{b}) = \| \boldsymbol{a} - \boldsymbol{b} \| = \sum_{i=1}^{p} |a_i - b_i|$, was adopted for this purpose. It computes the shortest distance ($d$) between a new sample ($\boldsymbol{a}$) and the closest observation in the training set ($\boldsymbol{b}$), where $p$ is the number of variables (Ref 29). The candidate design variable vector ($\vec{x}$) is then discarded if $d > d_{max}$. Equation 8 preserves the operating ranges of the design variables by setting lower and upper bounds for them, and Eqs. 9 and 10 are necessary to keep consistency concerning process variables' relationships. They refer to the pressure ($P$) and temperature ($T$) of the tempering ($Temp$) and austenization ($Aust$) processes, a parameter $c$ and the soaking time index ($t_{Soak}$).

The previous framework was employed to illustrate an optimization problem in the heat treatment furnaces of the steelmaking plant of the case study. For visualization purposes, two design variables were employed, namely the percentage of the molybdenum chemical element (Mo) ($x_1$) and the tube exit temperature after the tempering furnace ($x_2$) (Table 1). The greater this temperature, the greater the energy consumption. The objective was to define set points for both while keeping the remaining variables in their respective nominal values. The plant involved in this work is itself the raw material supplier for its heat treatment unit. That is, the factory is integrated by a steel refining plant, a casting unit, that produces steel cylindrical bars and a seamless tube rolling mill. The next operation is exactly the quenching–tempering line. This arrangement involving the entire steel production chain facilitates process changes, which explains the use of an alloying element as a design variable. Thus, the objective function (Eq. 3) was based on material and operation costs, with $\beta = 1$. Only the term relative to these costs, $C(\vec{x})$, was used in this example. In short, the objective regards to achieve lower material costs, by using lower contents of alloying elements (in this case, molybdenum), and lower production costs, in relation to energy consumption.

The following set of restrictions were considered. The inequality in Eq. 4 made use of a predictive model previously obtained in this work for yield strength (section 4). The MLP neural network with fourteen inputs, given by the backward variable selection method (Table 4), was adopted. This choice considered its lower MAE value (equal to 8.43) out of the eight candidate models, when considering both selection procedures. Due to process complexity, a purely mathematical description is infeasible. As previously mentioned (section 4.1), the inherent measurement uncertainty for this property can be of up to 20 MPa. To meet process conditions, a narrower range was then adopted, with $\mathrm{YS}_{min} = 640$ MPa and $\mathrm{YS}_{max} = 650$ MPa. More relaxed ranges, mainly for the lower bound, can be applied. For inequality in Eq. 7, the maximum distance ($d_{max}$) was set to 1. The lower and upper bounds for the design variables in Eq. 8 considered typical values. Namely, $\vec{x}_1 = [0; 0.45]$ for the molybdenum content (in %), and $\vec{x}_2 = [670; 740]$, for the tube exit temperature after the tempering furnace (in °C). These ranges may vary from customer to customer. For the equality restrictions in Eqs. 9 and 10, it is reasonable to consider simple cycles of heat treatments in the

furnaces. In this case, the Tsuchiyama equivalent time can be used as the soaking time index ($t_{Soak}$), and the reference temperature, as the tube exit temperature ($T_{Aust}$).

The use of neural models implies that no derivatives are available, which is required for gradient descent-based optimization techniques. Moreover, the current optimization problem involves continuous and discrete variables as well as a nonconvex and discontinuous search space. According to Rao (Ref 30), the use of conventional nonlinear optimization techniques in this context is inefficient, computationally expensive, and the final solution is generally close to the starting point. According to the same author, genetic algorithm (GA) is suitable in such cases. Thus, this work used a GA package (Ref 31) to perform the search in the solution space. Table 7 shows the parameter set adopted.

**Table 7 Parameter set used for the genetic algorithm**

| Parameter | Value |
|---|---|
| Population size | 100 individuals |
| Probability of mutation | 10% |
| Probability of cross-over | 80% |
| Number of generations | 50 |
| Elitism | 5 individuals |
| Encoding type | Real |

Figure 5 shows the ($x_1, x_2$)–search space after 1, 5, 10 and 50 generations. The level curves correspond to the yield strength property (in MPa) calculated using Eq. 4. The color map ([0.1 (green), 1.0 (gray)] scale) represents the distance of every feasible solution from the training set domain, according to Eq. 7. The red point in each plot highlights the best objective function evaluation. As the number of generations increases, the design variables converge to solutions that lead to lower production costs according to the objective function (Eq. 3), which is given by lower amounts of molybdenum ($x_1$) and lower tube exit temperatures ($x_2$), while satisfying the set of restrictions. The optimal solution found was equal to $x_1 = 0.047\%$ and $x_2 = 670$ °C. The chance of using less alloying elements and operating at lower temperatures is beneficial over time.

## 6. Conclusions

The high complexity of industrial operations from one side and the availability of massive amounts of data on the other side have been supporting machine learning applications in general. This work applied a set of data-driven techniques to

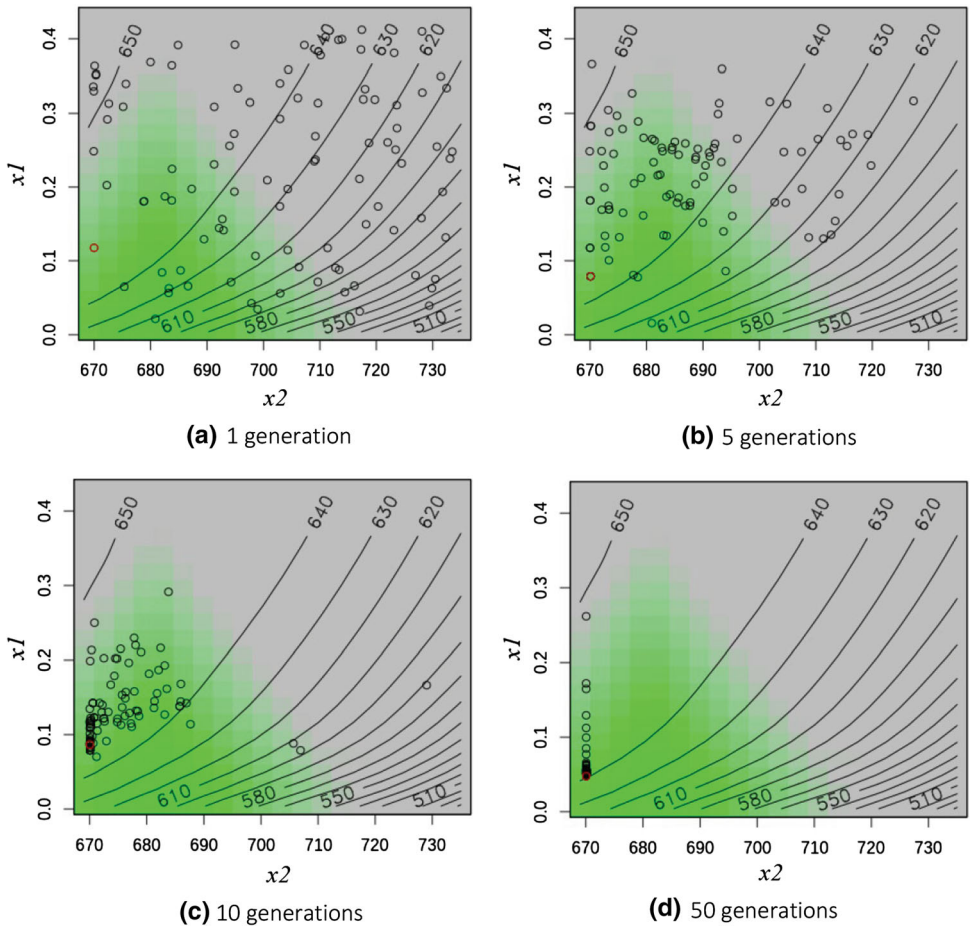


**Fig. 5** (a–d) Optimization results given the number of generations, where level curves are given by restriction in Eq. 4, and colormap ([0.1 (green), 1.0 (gray)] scale), by restriction in Eq. 7. ($x_1$ = Molybdenum chemical element (Mo; in %), and $x_2$ = Tube exit temperature after the tempering furnace (in °C) (Table 1)

predict commonly used mechanical properties in steelmaking plants, namely yield strength, ultimate tensile strength and hardness, of steel tubes produced. The availability of this information, commonly obtained from laboratory analysis, in advance, can contribute to more stable operations and, ultimately, to reduce rework and customer lead time. Two variable selection procedures were employed, which favors data collection, storage and processing, model interpretation and online implementations. Satisfactory results concerning acceptable prediction errors were achieved for all properties. Also, relatively smaller input subsets were able to keep model generalization capacity.

In this work, an optimization problem for minimizing costs in general was also illustrated. One of the predictive models previously obtained, for a mechanical property in particular, was used in this application. Due to process complexity, its description using a purely mathematical model does not look to be feasible. The optimal solution resulted in less use of a particular alloying element and in lower energy consumption, while keeping quality aspects. Medium and large process industries, and more specifically steelmaking plants, can greatly benefit from data-driven approaches, mainly in the present scenario, with more process data available and increased computational capacity.

## Acknowledgments

The authors thank the Vallourec unit in Brazil for both the data sets and the financial support. They also thank Rodolfo Dollinger, a data scientist in this steelmaking plant, for his contribution with respect to process data analysis. Antonio P. Braga thanks CNPq for the financial support.

## References

1. W. Smith and J. Hashemi, *Foundations of Materials Science and Engineering*, 5th ed., McGraw-Hill, New York, 2009
2. G. Totten and M. Howes, *Steel Heat Treatment Handbook*, 2nd ed., CRC Press, Boca Raton, 1997
3. S. Pattanayak, S. Dey, S. Chatterjee, S. Chowdhury, and S. Datta, Computational Intelligence Based Designing of Microalloyed Pipeline Steel, *Comput. Mater. Sci.*, 2015, **104**, p 60–68
4. P. Sampaio, R. Corrêa, A. Braga, Modelagem das propriedades mecânicas de tubos de aço utilizando redes neurais artificiais (Mechanical properties modeling of steel tubes using artificial neural networks). Automation and IT Seminar, ABM (Brazilian Association of Metallurgy, Materials and Mining), 2015
5. A. Agrawal, P. Deshpande, A. Cecen, G. Basavarsu, A. Choudhary, and S. Kalidindi, Exploration of Data Science Techniques to Predict Fatigue Strength of Steel from Composition and Processing Parameters, *Integr. Mater. Manuf. Innov.*, 2014, **3**, p 3–8
6. D. Jones, J. Watson, and K. Brown, Comparison of Hot Rolled Steel Mechanical Property Prediction Models Using Linear Multiple Regression, Non-linear Multiple Regression and Non-linear Artificial Neural Networks, *Comput. Mater. Sci.*, 2013, **32**(5), p 435–442
7. J. Mori and V. Mahalec, Planning and Scheduling of Steel Plates Production. Part II: Scheduling of Continuous Casting, *Comput. Chem. Eng.*, 2017, **101**(9), p 312–325
8. J. Mori and V. Mahalec, Planning and Scheduling of Steel Plates Production. Part I: Estimation of Production Times via Hybrid Bayesian Networks for Large Domain of Discrete Variables, *Comput. Chem. Eng.*, 2015, **79**(4), p 113–134
9. L. Tang, J. Liu, A. Rong, and Z. Yang, A Review of Planning and Scheduling Systems and Methods for Integrated Steel Production, *Eur. J. Oper. Res.*, 2001, **133**(1), p 1–20
10. X. Chen, X. Chen, J. She, and M. Wu, A Hybrid Just-In-Time Soft Sensor for Carbon Efficiency of Iron Ore Sintering Process Based on Feature Extraction of Cross-sectional Frames at Discharge End, *J. Process Control*, 2017, **54**, p 14–24
11. D. Laha, Y. Ren, and P. Suganthan, Modelling of Steelmaking Process with Effective Machine Learning Techniques, *Expert Syst. Appl.*, 2015, **42**(10), p 4687–4696
12. M. Bevilacqua, E. Bottani, F.E. Ciarapica, F. Costantino, L. di Donato, A. Ferraro, G. Mazzuto, A. Monteriù, G. Nardini, M. Ortenzi, M. Paroncini, M. Pirozzi, M. Prist, E. Quatrini, M. Tronci, and G. Vignali, Digital Twin Reference Model Development to Prevent Operators' Risk in Process Plants, *Sustainability*, 2020, **12**(1088), p 1–17
13. N.A. Carvalho, L.F. Scavarda, and L.J. Lustosa, Implementing Finite Capacity Production Scheduling: Lessons from a Practical Case, *Int. J. Prod. Res.*, 2014, **52**(4), p 1215–1230
14. R, The R Project for Statistical Computing, 2020. http://www.r-project.org/. Accessed 5 Oct 2019
15. F. Coelho, A.P. Braga, and M. Verleysen, A Mutual Information Estimator for Continuous and Discrete Variables Applied to Feature Selection and Classification Problems, *Int. J. Comput. Intell. Syst.*, 2016, **9**(4), p 726–733
16. F. Coelho, A.P. Braga, and M. Verleysen, Multi-objective semi-supervised feature selection and model selection based on Pearson's correlation coefficient, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Lecture Notes in Computer Science*, Vol 6419, I. Bloch and R.M. Cesar, Ed., Springer, Berlin, 2010
17. I. Guyon, S. Gunn, M. Nikravesh, and L. Zaded, *Feature Extraction: Foundations and Applications*, Springer, New York, 2006
18. D.C. Montgomery, E.A. Peck, and G.G. Vining, *Introduction to Linear Regression Analysis*, Wiley, New York, 2012
19. S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, Upper Saddle River, 1999
20. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, New York, 2013
21. C. Gomes, A.L. Kaiser, J.P. Bas, A. Aissaoui, and M. Piette, Predicting the Mechanical Properties of a Quenched and Tempered Steel Thanks to a Tempering Parameter, *Rev. Métall.*, 2010, **107**, p 293–302
22. W. Callister, *Materials Science and Engineering: An Introduction*, Wiley, New York, 2007
23. M. Song, C. Sun, Y. Chen, Z. Shang, J. Li, Z. Fan, K.T. Hartwig, and X. Zhang, Grain Refinement Mechanisms and Strength-Hardness Correlation of Ultra-fine Grained Grade 91 Steel Processed by Equal Channel Angular Extrusion, *Int. J. Press. Vessels Pip.*, 2019, **172**, p 212–219
24. S.H. Hashemi, Strength–Hardness Statistical Correlation in API, X65 Steel, *Mater. Sci. Eng. A*, 2011, **528**(3), p 1648–1655
25. M.-L. Zhu and F.-Z. Xuan, Correlation Between Microstructure, Hardness and Strength in HAZ of Dissimilar Welds of Rotor Steels, *Mater. Sci. Eng. A*, 2010, **527**(16–17), p 4035–4042
26. E.J. Pavlina and J. Van Tyne, Correlation of Yield Strength and Tensile Strength with Hardness for Steels, *J. Mater. Eng. Perform.*, 2008, **17**, p 888–893
27. I.T. Jolliffe, *Principal Component Analysis*, Springer, New York, 2002
28. Y.S. Aurelio, G.M. Almeida, C.L. Castro, and A.P. Braga, Learning from Imbalanced Data Sets with Weighted Cross-entropy Function, *Neural Process. Lett.*, 2019, **50**, p 1937–1949
29. E. Krause, *Taxicab Geometry: An Adventure in Non-Euclidean Geometry*, Dover Publications, New York, 1987
30. S. Rao, *Engineering Optimization: Theory and Practice*, Wiley, New York, 2009
31. L. Scrucca, GA: A Package for Genetic Algorithms in R, *J. Stat. Softw.*, 2013, **53**(4), p 1–37