Original Article

Mechanical properties prediction of dual phase steels using machine learning

Tairine Berbert Tavares [1*]
Felipe Pereira Finamor [1]
Julio Cezar de Sousa Zorzi [1]

Abstract

The use of artificial intelligence techniques, with the increase of data generation capacity and the advancement of computational resources, has enabled the industries to develop and improve products without compromising laboratory and industrial resources. In this paper, a supervised machine learning (ML) based technique was used to predict the yield strength (YS), ultimate tensile strength (UTS), and elongation (EL) of dual phase steels with minimum tensile strengths of 590 and 780 MPa. The computational analysis was done from industrial data information containing the chemical composition and the thermomechanical processing parameters of the referred materials. The proposed ML model reached values of coefficient of determination above 0.94, with an accuracy of ±30 MPa for YS and UTS, and ±3% for EL. These results demonstrated the rationality and reliability of the tested method, allowing its application in future research works and in decision making that aim to optimize the steels industrial processing parameters.

Keywords: Machine learning; Dual phase steels; Mechanical properties.

1 Introduction

Dual phase (DP) steels are well suited for automotive application due to their attractive mechanical properties, which combine high strength with good formability. These features basically result from the DP characteristic microstructure, mainly consisting of hard martensite islands distributed in a soft and ductile ferrite matrix. In the case of cold rolled sheets, the microstructure is strongly dependent on the chemical composition and on the steel sheet processing, especially at the continuous annealing stage. Since the applicability of these industrial products relies on the guarantee of their mechanical properties, here expressed by the yield strength (YS), ultimate tensile strength (UTS), and elongation (EL), a better understanding of the effects and interactions of the fundamental variables becomes of great importance for the development and improvement of the DP steels grades [1,2].

Although the metallurgical mechanisms involved in the steel manufacturing process are theoretically well understood, it is extremely difficult to mathematically express the relationships between the influencing factors, since the steel production is a complex and dynamic nonlinear system. Traditional regression methods, such as linear and multiple regression models, are not capable of meeting the required accuracy as the number of variables increases. In the case of steel, these variables are composed of chemical composition, which consists of iron, carbon, and several alloying elements, and many other parameters regarding the processing line, such as the heating rate, intercritical temperature and cooling rate. As a result, it is almost impossible to evaluate all different combinations of chemical elements and process parameters interactions without increasing the production costs and compromising laboratory and industrial resources [3,4].

To solve this problem, artificial intelligence techniques have been applied in the field of materials science to obtain improvements in the prediction of the overall behavior of steels. For instance, Lalam and coauthors [5] evaluated the prediction of the yield strength and ultimate tensile strength of industrial galvanized steel coils from its chemistry and key galvanizing process parameters using a feed-forward back-propagation artificial neural network (ANN). The model predicted the mechanical properties with an accuracy of ±10 MPa for 90% of the data, which was within the acceptance levels of the industrial operation team. Xu et al. [3] studied the prediction of mechanical properties of a hot rolled alloy steel based on its chemical composition and process parameters, but using a convolutional neural network (CNN) based method to convert the production data into two-dimensional data images. The results showed that the proposed CNN model successfully predicts the properties of the material used in the experiments, being consistent

[1]*Centro de Pesquisa e Desenvolvimento, Usiminas, Ipatinga, MG, Brasil.*

**Corresponding author: tairine.tavares@usiminas.com*

with what was metallurgically expected. Guo et al. [4] proposed a machine-learning-based method using nonlinear programming to process the restrictions of the materials properties from the mapped functions in the industrial data, achieving relatively good prediction performance.

With the improvement of computing technology and the advances in artificial intelligence methods, machine learning (ML) has become a promising route to predict materials properties for various applications, since it is capable of dealing with multiscale problems [4]. This method can be simply described as an automated analytical model building that enables the computer to iteratively learn from data and recognize patterns inside them without being explicitly programmed, producing reliable results.

In the case of materials engineering, where big data is generally not accessible for most empirical work, ML techniques can be applied to small and intermediate datasets with successful outcomes [6,7]. Hence, in the present work, a supervised ML model was proposed to predict the mechanical properties of dual phase steels, with minimum tensile strengths of 590 and 780 MPa, as a function of alloy composition and thermomechanical processing parameters extracted from industrial data information.

2 Materials and methods

In this section, it is presented the designing steps of the proposed model for predicting the mechanical properties of the DP steels based on ML algorithms, which are constituted by data collection, data processing, model training and testing, and model validation.

2.1 Dataset

For the model development, production data of dual phase steels processed via continuous annealing route were collected from the Usiminas database. The data refer to DP590 and DP780 steels fabricated during eight months of 2019. The original data contained around 3,500 samples, where samples are defined as the data information of 3,500 coils produced at Usiminas continuous annealing line during the mentioned period, from which 1,533 referred to DP590, and 1,947 to DP780. To make sure that errors in the raw data were not considered in the prediction analysis, missing and undefined information were removed, so at the end of cleaning step a total of 2,596 samples remained as entries of the ML model, from which 1,118 were DP590 samples, and 1,478 were DP780.

The next procedure was characterized by the features selection, where all the features that contribute to the prediction of the dependent variables were identified, and the irrelevant or uncorrelated ones were excluded from the dataset before training the model, in order to avoid unnecessary coefficients. The construction of a good database is the key for the success of the ML model, since the accuracy of predictions will depend on the independent variables and the complexity of correlations. For the study, 29 influence factors were considered, which included process parameters, chemical compositions, and the three mechanical properties (YS, UTS, EL) of each sample. The list of selected variables for the model development is given in Table 1.

The correlation matrix, Figure 1, illustrates the linear correlation between two considered features, with values lying between -1 and 1. A negative correlation coefficient means that when the value of one feature increases, the value of the other decreases. On the other hand, a positive value implies a positive correlation, while 0 means that there is no linear correlation between them.

2.2 ML approach

To determine the appropriate algorithms that predict the three mechanical properties of steel with a good approximation, in the present work, widely used supervised ML methods, belonging to the Caret [8] package (short for classification

Table 1. Features selected from the Usiminas steel production database

	Feature		Feature
1	Hot coil thickness (HCT)	16	Molybdenum content (Mo)
2	Rolling reduction (RR)	17	Boron content (B)
3	Final thickness (FT)	18	Finish rolling temperature (FRT)
4	Width (WT)	19	Coiling temperature (CT)
5	Carbon content (C)	20	Speed (SPD)
6	Silicon content (Si)	21	Heating furnace temperature (HF)
7	Manganese content (Mn)	22	Soaking furnace temperature (SF)
8	Phosphorus content (P)	23	Slow cooling furnace temperature (SCF)
9	Sulfur content (S)	24	Overaging furnace temperature (OA)
10	Aluminum content (Al)	25	Skin-pass mill (SPM)
11	Copper content (Cu)	26	Skin-pass mill load (SPML)
12	Chromium content (Cr)	27	Yield Strength (YS)
13	Niobium content (Nb)	28	Ultimate Tensile Strength (UTS)
14	Nitrogen content (N)	29	Elongation (EL)
15	Titanium content (Ti)		

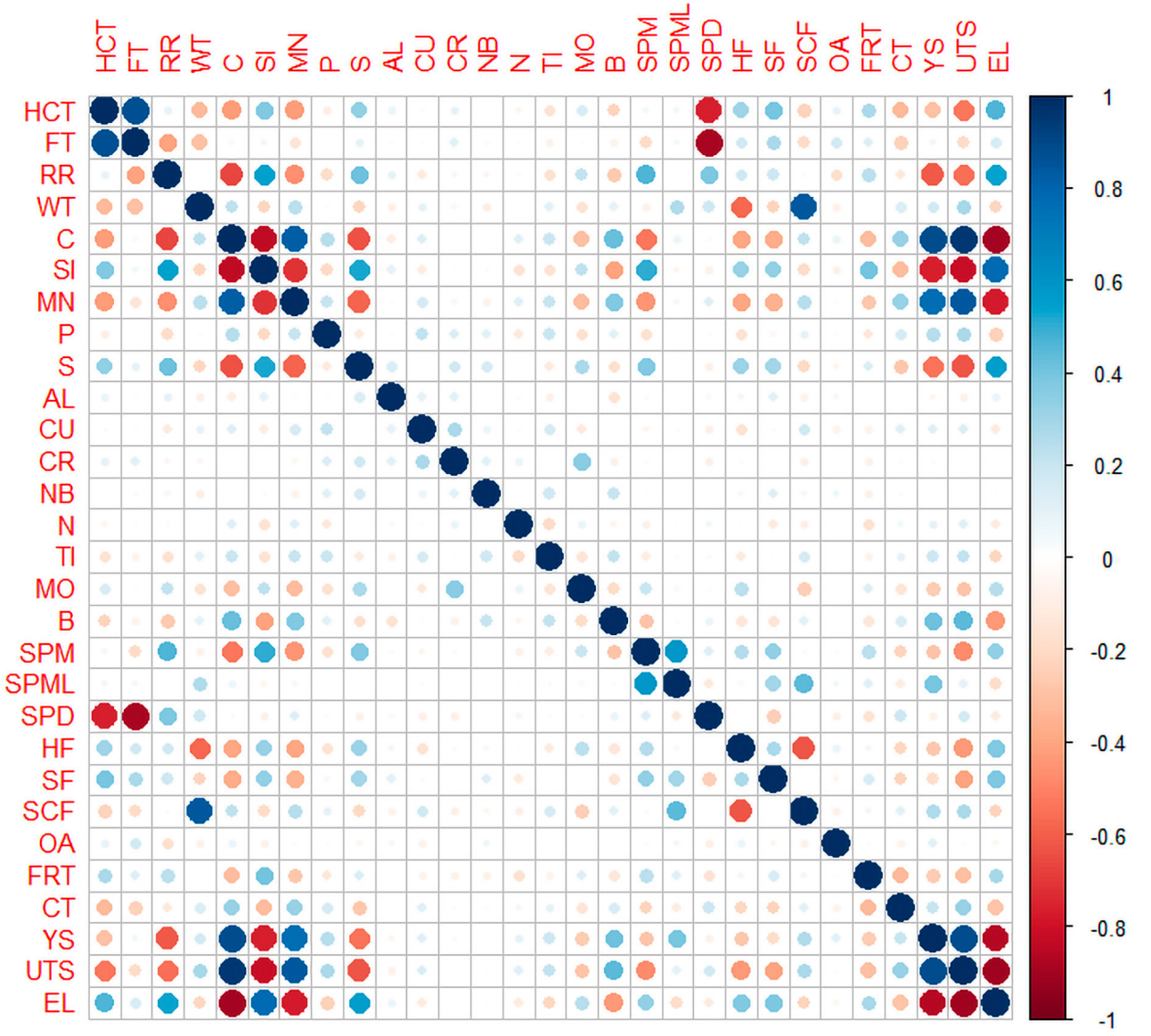

Figure 1. Correlation matrix plot for the variables considered in the present work

and regression training) available in the statistical software R, were applied for evaluation purpose as follows:

1. K-Nearest Neighbor (KNN);
2. Generalized Linear Model (GLM);
3. Support Vector Machine (SVM);
4. Cubist (CUB);
5. Random Forest (RF);
6. Lasso and Elastic-Net Regularized Generalized Linear Model (GLMNET)

The prediction performance of these methods was evaluated by two metrics, the coefficient of determination (R^2) and the root mean squared error (RMSE). R^2 characterizes the degree of fit by the change in data. It is always a value between 0 and 1, where the closer to 1, the stronger is the ability of the model's equation to predict the results found in the practical analysis. RMSE is the square root of the ratio of the square of the deviation between the observed value and the predicted value to the number of observations n. It assumes a value equal to or greater than 0, where 0 implies a statistically perfect fit to the analyzed data [9]. These metrics are individually defined by the Equation 1 and Equation 2:

$$R^2 = \frac{\sum_{i=1}^{n}\left(y_i^{obs} - \bar{y}^{obs}\right)^2 - \sum_{i=1}^{n}\left(y_i^{obs} - y_i^{pred}\right)^2}{\sum_{i=1}^{n}\left(y_i^{obs} - \bar{y}^{obs}\right)^2} \in [0,1] \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i^{obs} - y_i^{pred}\right)^2}{n}} \in [0,+\infty] \quad (2)$$

Where y_i^{obs} and y_i^{pred} represent the observed value and the corresponding predicted value, respectively, $\bar{y}^{obs}$ is the mean of the observed responses, and n is the total number of data.

The dataset was divided into two groups, 80% for training and 20% for validation. In order to compare different combinations of the different algorithms, a leave-one-out cross validation score was used to evaluate the models and obtain the averaged prediction performance. This method is commonly used for small datasets, such as in this work. Through this method, the original dataset is randomly divided into k groups (k-fold, here 10-fold) of approximately the same size, where k - 1 groups of the dataset are used as the training set, and the left-out group is used to validate the predictive accuracy of the model after training. Consequently, the problem of overfitting tends to be minimized, since the results in the experiments were independent of the training dataset, and a generalized model can be created [10,11]. Then, the models that responded with best performances when trained for 10 times were identified and considered in the properties prediction evaluation.

3 Results and discussion

3.1 Performance of the model

The performance of the ML model on the test data was considered for selecting the optimal model for predicting each mechanical property for DP590 and DP780 grades. This was done based on the results of the indicators adopted as the evaluation metrics to assess the prediction capability. In Figure 2, the interval charts represent the variation in R^2 and RMSE for the different algorithms tested for the prediction of the mechanical properties. It can be observed that, for the prediction of UTS, the algorithm that presented the greater value of R^2 and smallest value of RMSE was Cubist, and for the prediction of YS and EL, it was Random Forest, which imply that these models perform better than the other ones analyzed for this study purpose. It is also noticed, on the other hand, that the major adjustment failures are related to the elongation (EL), since it shows the worst value and greater dispersion of R^2. In this case, the model tends to provide the biggest errors during the prediction of the mechanical properties.

In the Cubist experiments, the Caret package in R is utilized to build a rule–based model through a combination of regression trees.

This technique randomly selects a specific set of features and, based on a best fit rule, returns, at the end, a linear regression model for prediction. In this way, the Cubist regression model can be described as a tree reduced

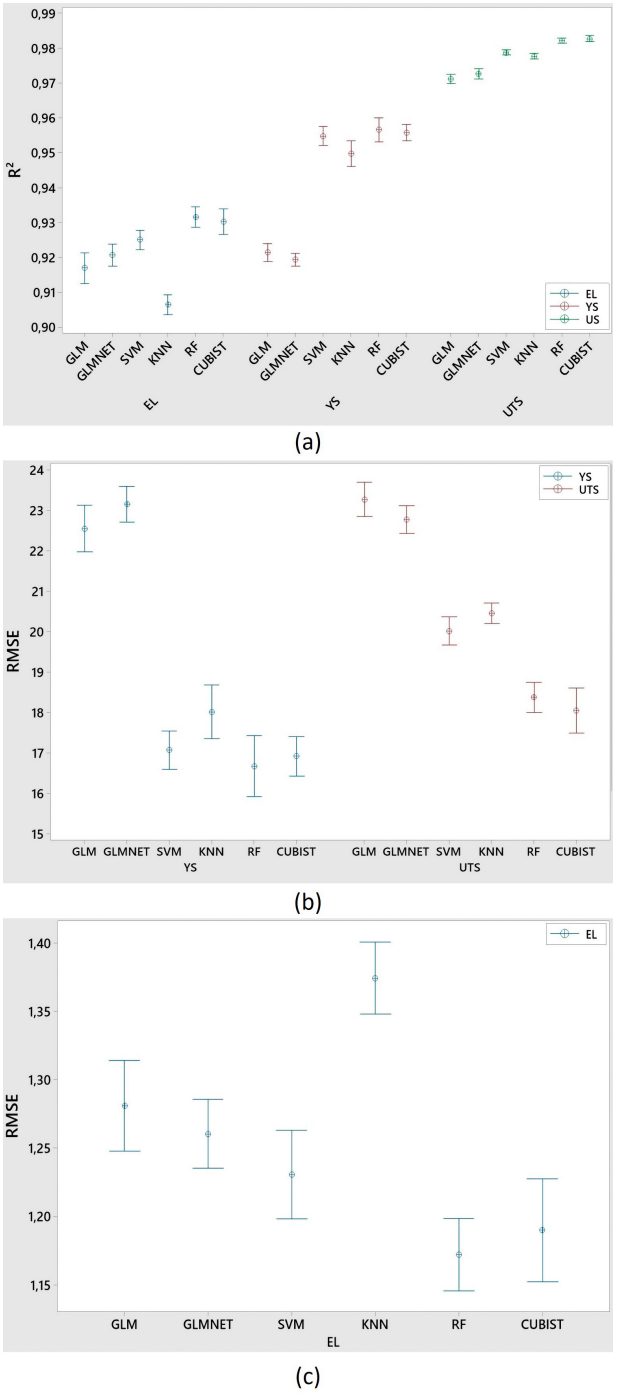

Figure 2. Interval charts of the test data evaluation metrics for different ML prediction models of both DP590 and DP780 steels grades. (a) coefficient of determination (R^2), (b) root mean squared error (RMSE) for YS and US, and (c) RMSE for EL.

to a set of rules that establish paths from the top of the tree to the bottom, where a prediction is made according to previous splits. The branches can be considered a series of "if-then" rules, while the terminal leaves can be regarded as an associated multivariate linear model, and as long as the set of covariates satisfies the conditions of the rule,

the corresponding model is used to calculate the predicted value [9,10].

Random Forests, in its turn, is also a tree-based ensemble technique, where each tree is built from a random subset of training data with a random subset of predictor variables. Different from traditional statistical methods, this technique contains many easy-to-interpret decision trees models instead of parametric models. In this case, the final predicted values are produced by the aggregation of the results of all the individual trees that make up the forest [9,10].

3.2 Model validation and analysis

Given the previous results, Cubist and Random Forest algorithms were used to predict each mechanical properties, as can be seen in Figure 3. The dotted center line in each graph represents the best fit on the predicted data and the vertical spread indicates the error in the predicted values.

For UTS prediction, R^2 of the optimal method was 0.99. For YS prediction, R^2 was 0.97, and for EL, 0.94. The deviation from the center line shows an error of ±30 MPa for YS and UTS, and ±3% for EL, which indicates, after the model validation, a good agreement between the industrial and predicted values. As previous mentioned, the EL presented the greater dispersions from the center line, but the results were within accepted levels. This expected error is mainly related to a greater variation of this property values, for a same steel grade, in the industrial data, as can be seen in Figure 4. Therefore, these results demonstrate the rationality and reliability of the tested ML method, pointing out that it has a great potential in modeling the mechanical behavior of the DP590 and DP780 steels. It is also noteworthy that the database used as the model input is based on the average values of the processing parameters, which inevitably influences the prediction of mechanical properties, reducing the generalization ability of the model. Despite that, the results of the method were promising and satisfactory, allowing its application in future research works and in decision making.

It can be seen from Figure 4, where individual value plot and median difference test are represented, that the difference between the mechanical properties values predicted by the proposed ML model and the real data are not statistically significant, suggesting its use feasibility. Also, the connection line between two medians presents an almost horizontal behavior, with the predicted YS, UTS and EL values variation within the range obtained in industrial data. Moreover, the Mann-Whitney Test [12] result showed, with a confidence index of 95%, that there are not sufficient evidences to affirm that the medians of the predicted and actual conditions have discrepant values, which proves that the ML model has a good prediction performance.

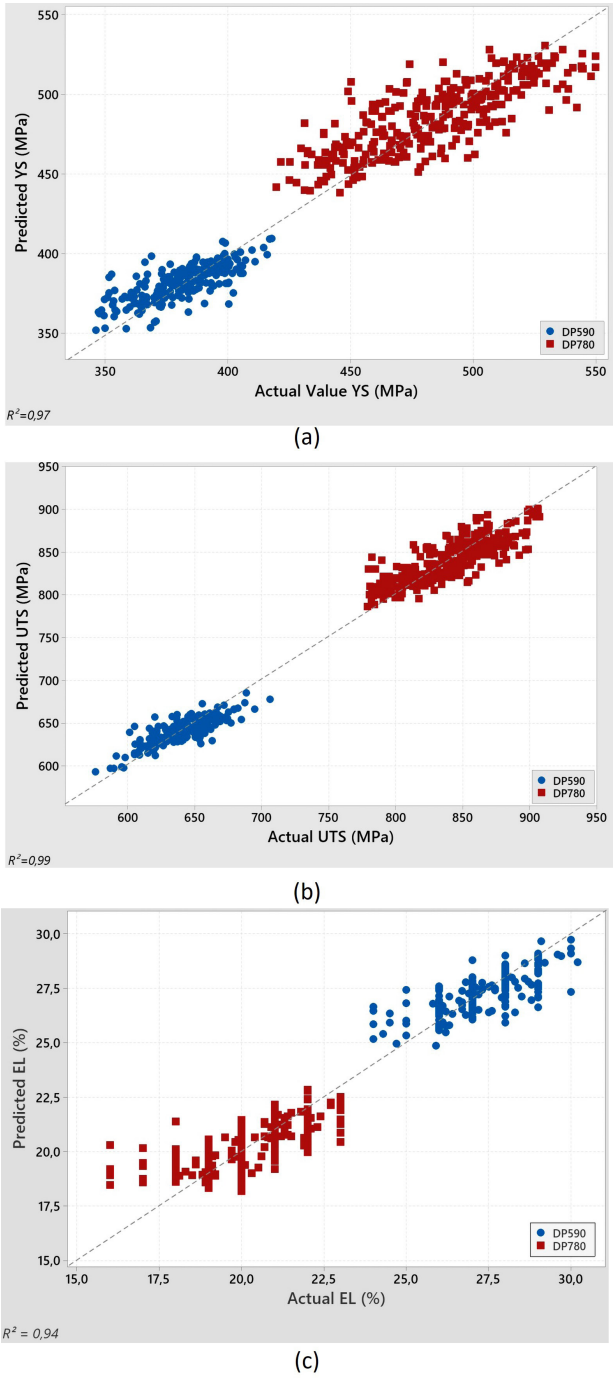

Figure 3. Comparisons of predicted and actual (a) yield strength, (b) tensile strength and (c) elongation determined by the best fit algorithm.

In summary, the ML method developed in this study successfully predicted the YS, UTS and EL of DP590 and DP780 steels based on their chemical composition and processing conditions. The differences between the predicted and industrial values are within the accepted error, which is reasonable considering that such a prediction was done based on average values of the processing parameters. To evaluate other steel types and DP grades, additional data with a wide range of

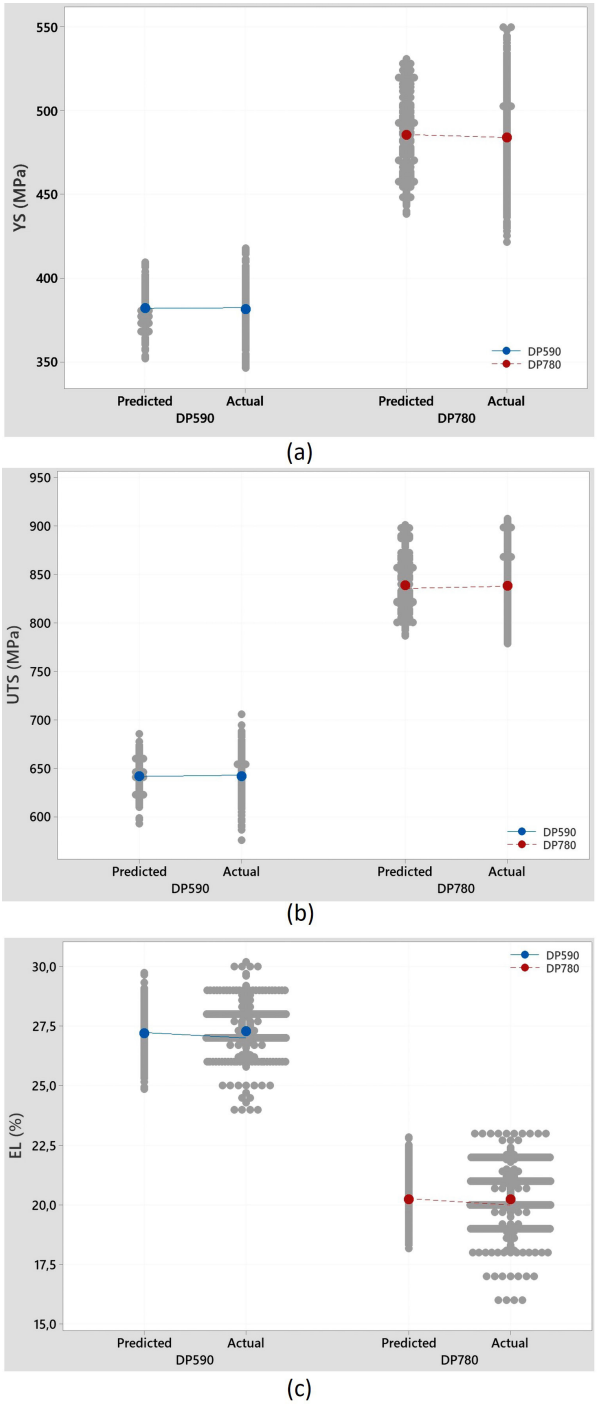

Figure 4. Individual value plots of predicted versus actual mechanical properties and the connected median lines of (a) yield strength, (b) tensile strength and (c) elongation determined by the best fit algorithm.

composition and process parameters needs be included in order to increase the capability of the developed model, since the current model was trained using only DP590 and DP780 steels data. The future research works can focus on microstructure and online mechanical properties prediction.

4 Conclusions

In this work, a supervised machine learning model was used for the prediction of the mechanical properties of cold rolled DP590 and DP780 steels, based on their chemical composition and thermomechanical processing parameters. The following conclusions were obtained:

- The Cubist and Random Forest regression algorithms were more suitable for predicting the mechanical properties of the dual phase steels when compared to other conventional supervised ML methods, in terms of R^2 and RMSE values. The Cubist model was used to predict the ultimate tensile strength, while the Random Forest was used for the yield strength and elongation.
- The accuracy of the proposed model was ±30 MPa for the YS and UTS, and ±3% for the EL, reaching R^2 values above 0.94. These results show that the proposed model has a good agreement between the industrial and predicted values, which can be considered as satisfactory even with the use of average values of the processing parameters.
- The major adjustment failures are related to the elongation values, which tend to provide the biggest errors in the application of the method. However, the difference between the predicted and the observed values are not statistically significant.
- The obtained results demonstrate the rationality and reliability of the tested ML method, which indicates a great potential in mechanical behavior modeling of the evaluated DP steels, allowing its application in future research works and in decision making that aim to optimize the industrial processing parameters of these steels.

References

1 Ramazani A, Mukherjee K, Prahl U, Bleck W. Modelling the effect of microstructural banding on the flow curve behavior of dual-phase (DP) steels. Computational Materials Science. 2012;52:46-54.

2 Arruda MVP, Melo TMF, Costa FS, Santos DB. Microstructural evolution during continuous annealing of a 980 MPa cold rolled steel grade. Journal of Physics: Conference Series. 2019;1270(012020):1-6.

3 Xu Z, Liu X, Zhang K. Mechanical properties prediction for hot rolled alloy steel using convolutional neural network. IEEE Access: Practical Innovations, Open Solutions. 2019;7:47068-47078.

4 Guo S, Yu J, Liu X, Wang C, Jiang Q. A predicting model for properties of steel using the industrial big data based on machine learning. Computational Materials Science. 2019;160:95-104.

5 Lalam S, Tiwari PK, Sahoo S, Dalal AK. Online prediction and monitoring of mechanical properties of industrial galvanized steel coils using neural networks. Ironmaking & Steelmaking. 2019;46(1):89-96.

6 Liu Y, Zhao T, Ju W, Shi S. Materials discovery and design using machine learning. Journal of Materiomics. 2017;3:159-177.

7 Rahaman M, Mu W, Odqvist J, Hedström P. Machine learning to predict the martensite start temperature in steels. Metallurgical and Materials Transactions. A, Physical Metallurgy and Materials Science. 2019;50A:2081-2091.

8 Kuhn M. Building predictive models in R using the Caret package. Journal of Statistical Software. 2008;28(5):1-26.

9 Zhou J, Li E, Wei H, Li C, Qiao Q, Armaghani DJ. Random forests and cubist algorithms for predicting shear strengths of rockfill materials. Applied Sciences (Basel, Switzerland). 2019;9(1621):1-16.

10 Noi PT, Degener J, Kappas M. Comparison of multiple linear regression, cubist regression, and random forest algorithms to estimate daily air surface temperature from dynamic combinations of MODIS LST data. Remote Sensing. 2017;9(398):1-23.

11 Chheda AM, Nazro L, Sen FG, Hegadekatte V. Prediction of forming limit diagrams using machine learning. IOP Conference Series. Materials Science and Engineering. 2019;651(012107):1-8.

12 Wijnand HP, Velde RV. Mann-Whitney/Wilcoxon's nonparametric cumulative probability distribution. Computer Methods and Programs in Biomedicine. 2000;63:21-28.

Received: 22 June 2021
Accepted: 24 Jan. 2022