International Journal of Minerals, Metallurgy and Materials
Volume 29, Number 4, April 2022, Page 836
https://doi.org/10.1007/s12613-022-2437-0

Evaluating data-driven algorithms for predicting mechanical properties with small datasets: A case study on gear steel hardenability

Bogdan Nenchev[1)], *Qing Tao*[1)], *Zihui Dong*[1)], *Chinnapat Panwisawas*[1)], *Haiyang Li*[2)], *Biao Tao*[2)], *and Hongbiao Dong*[1),✉]

1) NISCO UK Research Centre, School of Engineering, University of Leicester, Leicester LE1 7RH, UK
2) Nanjing Iron & Steel United Co., Ltd., Nanjing 210044, China

(Received: 12 December 2021; revised: 27 January 2022; accepted: 11 February 2022)

Abstract: Data-driven algorithms for predicting mechanical properties with small datasets are evaluated in a case study on gear steel hardenability. The limitations of current data-driven algorithms and empirical models are identified. Challenges in analysing small datasets are discussed, and solution is proposed to handle small datasets with multiple variables. Gaussian methods in combination with novel predictive algorithms are utilized to overcome the challenges in analysing gear steel hardenability data and to gain insight into alloying elements interaction and structure homogeneity. The gained fundamental knowledge integrated with machine learning is shown to be superior to the empirical equations in predicting hardenability. Metallurgical-property relationships between chemistry, sample size, and hardness are predicted via two optimized machine learning algorithms: neural networks (NNs) and extreme gradient boosting (XGboost). A comparison is drawn between all algorithms, evaluating their performance based on small data sets. The results reveal that XGboost has the highest potential for predicting hardenability using small datasets with class imbalance and large inhomogeneity issues.

Keywords: machine learning; small dataset; XGboost; hardenability; gear steel

1. Introduction

In steel metallurgy, the mechanical response of materials directly depends on the microstructure and its homogeneity across multiple length scales including chemical microalloying, phases, and grain size distribution. Linking the microstructure to the mechanical properties requires solving complex coupled multi-physics approaches consisting of several underlying differential equations. The task becomes extremely challenging to resolve with finite element analysis (FEA) when non-linear microstructure response, inhomogeneous element distribution, and rapid phase transformation are involved [1–2]. As an alternative, numerically solving non-linear solid mechanics problems such as the quenching of high-alloyed steel, is commonly achieved via empirical based modelling, relying on thousands of costly laboratory tests.

The mechanical performance of gear steels is evaluated through hardenability i.e., the ability of a material to change its hardness as a result of a given heat treatment. Thus, hardness is the most important indicator for gear steel fatigue life, resistance to deformation, and performance, directly affecting the application of the manufactured steel component. In steelmaking, quenching is an abrupt cooling process inducing martensitic transformation through diffusionless shear crystallographic deformation [3]. The homogeneity and fraction of the formed meta-stable structures, martensite, and bainite depends on the kinetics of transformation as a function of the micro-alloying element additions and cooling rate. However, building a computational model which relies upon thermally activated process simulation of the different phases and differential cooling rates is complex, time-consuming and requires high calibre skills. Consequently, in industry, determining the processing–metallurgical–microstructure–property causal relationship in gear steel manufacturing typically relies upon the empirical knowledge of composition and hardness.

A number of empirical hardenability mixture models exist in literature, most established of which is Maynier equation [4]. His equation takes into account martensite, bainite, and ferrite fraction, each expressed through simplified physical approximation determined via parametric calibration. However, the fraction of phases varies along the steel production bars due to the afore mentioned differential cooling rate. Performing microscopy identification of phases on multiple locations per test component is not only be highly time consuming but also costly and ineffective. Hence, alterations of Maynier equation purely based on alloy chemistry were developed through the years [5]. Nowadays, steelmaking industry still puts considerable resources into measuring harde-

✉ Corresponding author: Hongbiao Dong E-mail: h.dong@le.ac.uk

nability through Jominy end quench test to expand the ever-increasing standard steel families and their parametric hardenability expressions. However, the ever-increasing need for rapid alloy modification and improvement in gear performance gives rise to considerable research efforts towards developing better predictive models, forecasting hardenability of high-performance steels.

In recent years, a number of attempts to achieve more robust non-linear material modelling were made using artificial intelligence, specifically focusing on neural networks (NNs) and decision trees [6–8]. The industrial digitalization is increasing in attention and importance attributed to the data-driven materials science for new high-added value materials design. Since the end of the last century, Badeshia [9] highlighted the role of NNs as a highly supportive technique in material science. However, it is only in the last decade that the evolution of the machine learning (ML) techniques and the availability of large and low-cost computational resources, allow material scientists to investigate and utilize measured data in novel ways. ML techniques, and in particular NNs, are now increasingly applied in steel alloy optimization and in striving for superior target properties, as they appear capable to overcome the lack of efficiency in traditional experimental and industrial alloy characterization [10–11]. Through NN regression, even small data sets can produce simple quantitative expression capturing the complex relation between chemical compositions and resulting properties [6]. However, researchers face a number of challenges when dealing with NN modelling such as overfitting, high oscillations and lack of transparency.

Gradient tree boosting [12], also known as gradient boosting machine (GBM) or gradient boosted regression tree (GBRT), belongs to the family of ensemble models. It is a state-of-the-art technique for solving regression problems with a wide implementation in real-world applications. Effective implementation of statistical models that capture the variables dependencies and scalable learning systems is key to construct machine-learning-based analytics for complex non-linear processes. Extreme gradient boost (XGBoost) is a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning proposed by Chen and Guestrin [13]. It has well-acknowledge impact among ML community due to exceptional performance in ML and data mining challenges [14]. Due to its outstanding performance, >10 times faster than other machine learning algorithms, XGBoost's scalability in multi-disciplinary fields across all scientific fields in both regression and classification [13].

XGBoost provides several merits in terms of data processing. First, it has higher accuracy in comparison with GBRT, which only used first derivative of Taylor expansion, while, XGBoost offers degree of freedom in defining loss function in expansion function [15]. Introduction of regularized objective controls the complexity of models by using parameters, including number of leaf nodes, optimal weights of leaves. Shrinkage, sharing similar principles of learning rate, reduces the influence from individual tree enabling further improvement of model via future trees. Subsampling prevents overfitting and accelerates computations of the parallel algorithm. Thus, XGBoost holds a promising potential towards industrial applications where accuracy, fluctuations and transparency are critical.

In this work we investigate the performance of three different data-driven models with the specific application of small industrial hardenability dataset on gear steel alloy: 20CrMnTi. An empirical model is generated through a guided parametric calibration with non-linear multivariate minimization. A state-of-the art NN and the novel XGboost models are also utilized to generate target hardness prediction. In this paper, empirical and data-driven approaches are compared based on their ability to predict and understand hardenability through investigating alloying additions influences and microstructure homogeneity in gear steel samples.

2. Hardenability prediction for industrial applications

The Jominy profile, Fig. 1, is used for characterizing each steel grade hardenability. Specific requirements are imposed by the customers to the steel producers, in the form of upper and lower bound for the hardness value corresponding to specific Jominy distance values (J). Thus, controlling the range of hardenability for each steel grade is crucial for meeting the customers' needs and reducing scrappage. Each steel alloy possesses unique hardenability target range. The main alloying elements for gear steel, which affect hardenability include carbon (C), chromium (Cr), manganese (Mn), molybdenum (Mo), silicon (Si), nickel (Ni), and titanium (Ti), as shown in Fig. 1(d). The influence of these microalloying elements can be separated into direct and indirect effect. Carbon strongly affects the hardness of the martensite, as C delays the onset of pearlite formation hence stimulates the formation of martensite at slower cooling rates. However, the effect is not significant enough to be purely used for phase formation control, hence other elements are commonly used to control the hardenability. Cr, Mo, Mn, Si, Ni, and V (especially Cr, Mo, and Mn) retard the martensitic and bainitic transformation. The distribution of these microalloying additions have a direct effect on the microstructure transformation from austenite to ferrite and pearlite. On the other hand, elements such as Al, Ti, N show complex interactions among each other, indirectly affecting the temperature during the transformation phase. For instance, for the investigated gear alloy 20CrMnTi, the Ti content interacts with N_2. A TiN precipitate is formed reducing the interstitial solid solution of N which in turn causes lattice distortion influencing the phase transformation. Empirical methods are designed for direct influences hence there is an apparent gap in the investigation of the indirect influences.

Other than controlling the hardenability range and increasing the consistency of the hardness, the industry is focused on reducing the time required for the steel grade design

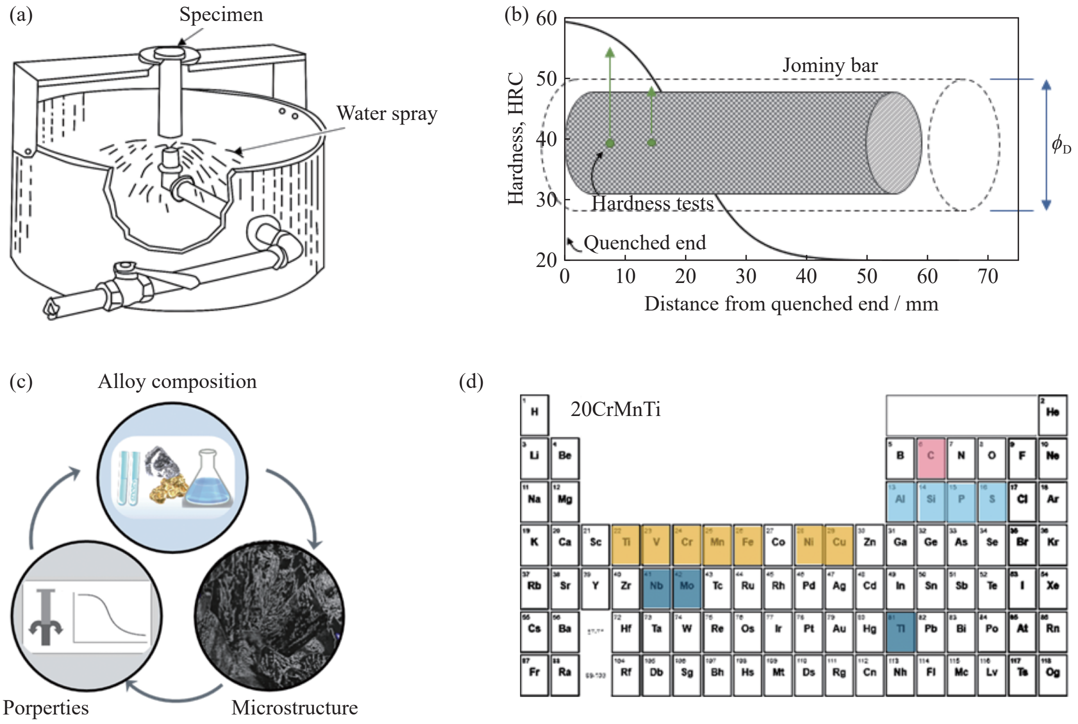


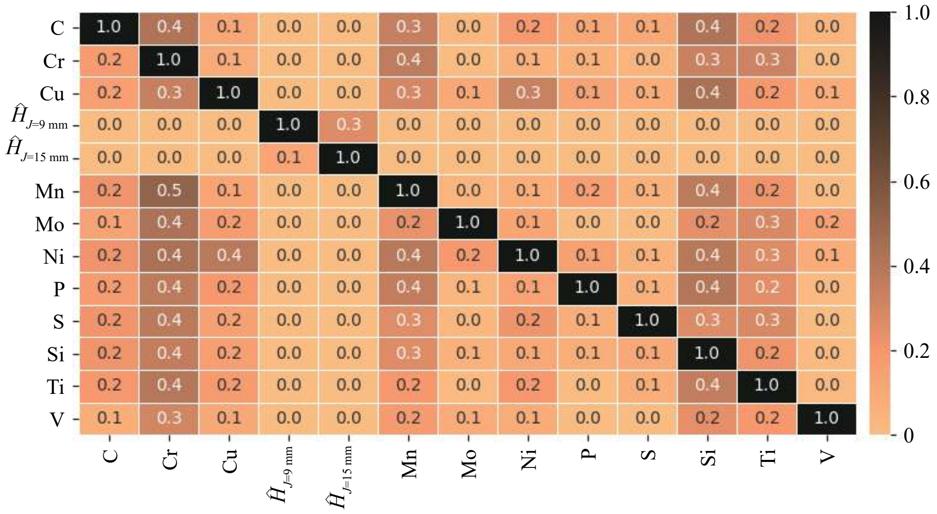
Fig. 1. Hardenability study of gear steel 20CrMnTi: (a) diagram of Jominy test set [16]; (b) Jominy profile is obtained by the measured hardness values as a function of the distance from the quenched end (*J*). The hardness along the bar depends on the cooling rate where both hardenability and cooling rate are at their maximum at the quenched (leading) end; (c) as illustrated in (c) the hardness depends on the alloy composition which in turns affects the formed microstructure; (d) illustrates all the alloying elements (highlight with colors) involved in 20CrMnTi alloy.

and testing. Microstructure phase fraction and grain testing is costly to perform on a large number of product specimens of various alloys. Any model predicting hardness without taking these variables into account is beneficial in mitigating the laboratory work burden and reducing costs. Further saving can be achieved by reducing the number of tests per Jominy location (*J*). In many cases an asymmetric hardness distribution is observed, i.e., the sensitivity of indentation at the same distance from quenched end is related to the inhomogeneity of microalloying elements distribution. Such inhomogeneity is also the primary reason for the reduced model accuracy and fluctuations and target predictions. Further indirect effect on the hardenability is caused by the component size. It is challenging to achieve consistency of the alloying and microstructure on larger components. Hence, the dataset examined in this study takes into account hardness measured on samples with a range of diameters. Thus, a model is required that increases the reliability of the hardenability models, designed for small data sets with limited variables and with a consideration of component size.

2.1. Industrial data analysis

The investigated in this study alloy 20CrMnTi is a relatively new gear steel alloy. Up to the knowledge of the authors, there isn't any readily available empirical model or chart for estimating its hardenability. Thus, steelmaking manufactures must perform Jominy quench tests for each batch taking into consideration both microalloying elements and component diameter (ϕ_D). The data in this study comes from such industrial quality control tests. Hardness measurements are taken at Jominy distances of 9 ($\widehat{H}_{J=9\,\mathrm{mm}}$) and 15 mm ($\widehat{H}_{J=15\,\mathrm{mm}}$) for component diameter in the range 30–130 mm, as shown in Fig. 1(b). The data contains measurements taken at total of 30 different sample diameters (ϕ_D), where between 2 to 16 hardness measurements are taken at each diameter per specific batch. Each sample batch has slightly different alloying chemistry and microstructure due to the prior processing. All hardness measurements are repeated twice and average is taken to remove any influences from the test equipment. Mass spectrometer is used to measure the steel chemistry for each steel bar. All samples are made from 20CrMnTi steel alloy with a manufacturing variation in the chemistry across the following elements: C, Mn, Ti, Cr, Cu, Mo, Ni, P, S, Si, and V recorded in weight percent. The obtained data is 370 heats total, hence it belongs to the group of small industrial data sets. Consequently, performing removal of data though data cleaning is not appropriate.

Both the hardness measurements and alloying constituents are investigated prior to the modelling. A novel Ppscore [17] algorithm is used to investigate the correlation between the parameters. Ppscore is an asymmetric, data-type-agnostic score that can detect linear or non-linear trends between variables. It assigns a predictive power value between 0 and 1. As seen in Fig. 2, Si, Cr, and Mn possess the highest interactivity out of all microalloying additives. Such mid-level correlation (0.02 >Ppscore value < 0.6) indicates that any information that these three elements add to the model will be partial (semi-dependant) in comparisons to the other fully independent elements (score <0.2). Interestingly, neither of the hardness measurements are directly related to the alloying chem-



	C	Cr	Cu	$\hat{H}_{J=9\text{ mm}}$	$\hat{H}_{J=15\text{ mm}}$	Mn	Mo	Ni	P	S	Si	Ti	V
C	1.0	0.4	0.1	0.0	0.0	0.3	0.0	0.2	0.1	0.1	0.4	0.2	0.0
Cr	0.2	1.0	0.1	0.0	0.0	0.4	0.0	0.1	0.1	0.0	0.3	0.3	0.0
Cu	0.2	0.3	1.0	0.0	0.0	0.3	0.1	0.3	0.1	0.1	0.4	0.2	0.1
$\hat{H}_{J=9\text{ mm}}$	0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{H}_{J=15\text{ mm}}$	0.0	0.0	0.0	0.1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mn	0.2	0.5	0.1	0.0	0.0	1.0	0.0	0.1	0.2	0.1	0.4	0.2	0.0
Mo	0.1	0.4	0.2	0.0	0.0	0.2	1.0	0.1	0.0	0.0	0.2	0.3	0.2
Ni	0.2	0.4	0.4	0.0	0.0	0.4	0.2	1.0	0.1	0.1	0.4	0.3	0.1
P	0.2	0.4	0.2	0.0	0.0	0.4	0.1	0.1	1.0	0.1	0.4	0.2	0.0
S	0.2	0.4	0.2	0.0	0.0	0.3	0.0	0.2	0.1	1.0	0.3	0.3	0.0
Si	0.2	0.4	0.2	0.0	0.0	0.3	0.1	0.1	0.1	0.1	1.0	0.2	0.0
Ti	0.2	0.4	0.2	0.0	0.0	0.2	0.0	0.2	0.0	0.1	0.4	1.0	0.0
V	0.1	0.3	0.1	0.0	0.0	0.2	0.1	0.1	0.0	0.0	0.2	0.2	1.0

Fig. 2. Predictive power score (Ppscore) correlation between chemistry and hardness of 20CrMnTi gear steel alloy. The score indicates the level of cross interaction and/or dependency of the variables within the data set.

istry. The lack of such direct dependence between hardness and the chemistry is suggested to be one of the main reasons why parametric studies struggle to predict consistently alloy hardenability.

2.2. Unsupervised learning for data classification

Unsupervised learning uses ML algorithms to analyze and cluster unlabelled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. From observations performed in the industrial hardenability dataset, Fig. 3, it is apparent that there are two distinct distributions within both the hardness at 9 and 15 mm Jominy distances. To better understand this hardenability separation, multivariate gaussian analysis was applied to both the hardness measurements at 9 and 15 mm, as shown in Fig. 3. In engineering, the Gaussian mixture model (GMM) is the one of the most commonly used probabilistic clustering methods, where data points are clustered based on the likelihood that they belong to a particular normal distribution.

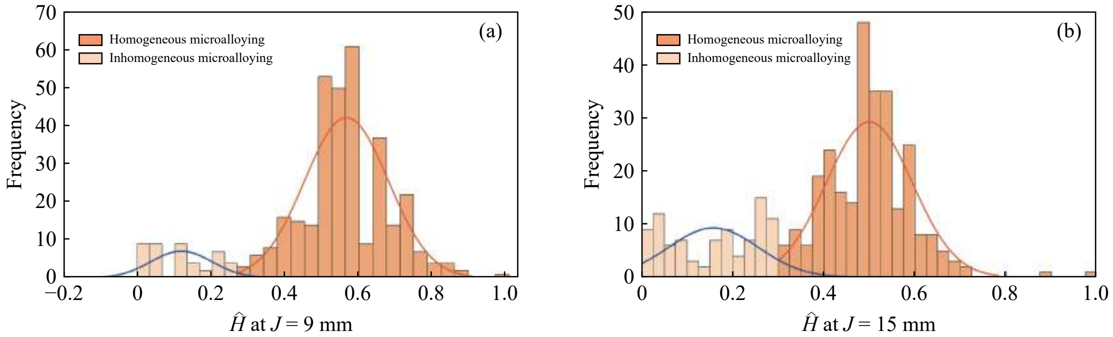


Fig. 3. Data classification according to multivariate gaussian distribution (GMM) analysis: (a) labelled frequency data for hardness measurements at *J* = 9 mm; (b) labelled frequency data for hardness measurements at *J* = 15 mm.

The GMM analysis on the hardness measurements reveals that there are two distinct peaks within the data. The primary peak is distributed around the target hardenability value $\widehat{H}_{J=9\,\mathrm{mm}} = 0.57$. However, the secondary peak is located at much lower values (0.1) and is undesired by the steelmaking manufacturers. It is crucial to understand its origin and mitigate its occurrence. As an unsupervised ML algorithm, GMM are capable of quantitatively separating and labelling the clusters of measured data points, however, a knowledge-based approach is required to metallurgically describe their occurrence. As described in the introduction section, two effects are not accounted for in the data variables: the fraction of martensite and bainite and the grain size, both of which are a function of the alloying elements, assuming the samples underwent identical austenitization and quenching procedure in laboratory-controlled setting. Consequently, the samples with lower hardenability have higher percentage of retained austenite, i.e., lower volume fraction of martensite and bainite. The multivariate gaussian analysis was implemented as an alternative method to account for this distinct change in microstructural conditions. By labelling this change into a separate variable which is then subsequently encoded into the machine learning algorithms to increase the data-driven modelling accuracy and provide a reference which guides the model optimization.

All data is consistently normalized between 0 and 1 prior to any of the processing and modelling. Standard approach in ML is to split the data into training, testing, and validation, however, by reducing the training data, we risk losing important patterns/trends in data set, which in turn increases error induced by bias. Thus, a method is required which provides sample data for training the model and still leaves sample data for validation. This is achieved via *k*-fold cross validation. In *k*-fold cross validation, the data is divided into *k* number of subsets. The holdout method is repeated *k* times, such that each time, one of the *k* subsets is used as the test set and the other *k*−1 subsets are put together to form a training set. The model effectiveness is evaluated by averaging the error (MSE) over all $k = 5$ trials. In such way, every data point gets evaluated at least once alleviating the large categorical imbalance.

2.3. Empirical hardenability model

In parametric mathematical hardenability models, each of the parameters is linked to the steel chemical composition by non-linear equations. In this study, a constrained nonlinear multivariable function [18] is used for parameter optimization to generate an empirical hardenability model for steel grade 20CrMnTi. The algorithm finds a constrained minimum of a scalar function of several variables starting at initial estimates. It uses sequential quadratic programming (SQP) method where the objective function is minimized by iterating through 14 parameters. The minimized equation is of the type:

$$H = a_1\sqrt{w_{\rm C}} + a_2 J^2\sqrt{w_{\rm C}} + a_3 w_{\rm Cr} + a_4 w_{\rm Ni} + a_5 w_{\rm Mn} + a_6 w_{\rm Ti} + a_7 w_{\rm Si} + a_8 w_{\rm Mo} + a_9 w_{\rm V} + a_{10} w_{\rm Cu} + a_{11} w_{\rm P} + a_{12} w_{\rm S} + a_{13} w_{\rm Al} + a_{14}\sqrt{\phi_{\rm D}} \tag{1}$$

where H is hardness at a distance J from quenched end, and $\phi_{\rm D}$ is component diameter; $a_{1\rightarrow 12}$ are the tuneable parameters; w_M means the mass fraction of element M. Equation coefficients from similar studies but for a different alloy (20CrMn Mo) are used to obtain the initial estimates in Eq. (1) solver. A MSE function is used to evaluate the difference between the predicted with Eq. (1) and measured hardness for alloy 20CrMnTi. The resulting empirical equation is as follows:

$$H = 63.31\sqrt{w_{\rm C}} - 0.136J^2\sqrt{w_{\rm C}} + 10.45w_{\rm Cr} + 5.00w_{\rm Ni} + 2.92w_{\rm Mn} + 3.93w_{\rm Ti} - 1.29w_{\rm Si} + 32.34w_{\rm Mo} + 9.92w_{\rm V} + 3.04w_{\rm Cu} + 11.99w_{\rm P} - 8.58w_{\rm S} + 0.00w_{\rm Al} - 0.0807\sqrt{\phi_{\rm D}} \tag{2}$$

The interaction of the alloying elements and their effect on hardness is reflected through the empirically tuned interaction parameters. According to Eq. (1), in this study, hardness decreases with distance from quenched end and with increasing the component diameter, both of which are reasonable and expected. Interestingly, hardness also decreases with the increase in Si content. Si generally influences hardenability positively, however, Si is one of the elements with the highest recorded fluctuation both with varying microalloying and distance from quenched end [19]. By standard, S content is reduced as much as possible, but even traces are detrimental to the alloy.

2.4. Neural networks for small industrial dataset

As shown in Fig. 1, 14 chemical components, $N = 14$, are considered in this application (C, Mn, Si, P, S, Cu, Cr, Ni, Mo, Ti, Nb, Al, V). Component diameter and Jominy distance are also considered. Both the mean hardness at 9 and 15 mm is used as a target for the study. The implemented NN model is a three-layer feed-forward perceptron type network with a variable number of neurons in the hidden layers. The NN consists of two hyperbolic tangent (tanh) transfer functions and a linear activation (purelin). The weights and biases of which are trained by backpropagation algorithm that employs Bayesian regularization to improve the network generalization capabilities and robustness [20].

$$I = \tanh\left(\sum_{j=1}^{N} W_{0j}^{(1)} \boldsymbol{X}_j + b_j^{(1)}\right) \tag{3}$$

$$H_k = \tanh\left(\sum_{j=1}^{N_H^{k-1}} W_{ij}^{(k+1)} \boldsymbol{X}_j g_j^{(k)} + b_i^{(k+1)}\right), k = 1,2,3, i = 1,2\ldots N_H \tag{4}$$

$$O = f\left(\sum_{j=1}^{N_H^4} W_{ij}^{(5)} g_j^{(4)} + b_i^{(5)}\right), \left\{\begin{array}{l} f(x), x > 0 \\ \alpha f(x), x \le 0 \end{array}\right. \tag{5}$$

where I, H_k, and O stands for input layer, transfer layer, and output layer, respectively, with k being the number of layer. In this work we use 3 NN tanh layers, therefore k varies from 1 to 3 in H_k, plus an input ($k = 0$) and an output layer ($k = 4$). $W_{ij}^{(k+1)}$ are the trainable weights for node j in layer k+1 for incoming node i. For example, in the input layer I, $W_{ij}^{(k+1)}$ becomes $W_{0j}^{(1)}$, i.e., $i = 0$, $k = 0$, and j is in the range from 1 to N (number of input); in the output layer O, $W_{ij}^{(k+1)}$ becomes $W_{ij}^{(5)}$ and i varies from 1 to N_H^4 (number of neurons), $k = 4$. b_i^k are the biases for node i in layer k; g_j represents the first order gradient of the loss function for node i; $\boldsymbol{X}_j$ is the input matrix; $f(x)$ represents a linear transfer function which has a gradient of one for positive x, and gradient α for negative x. The upper bound of the internal summation in the input layer (I) depends on the number of inputs (N), whereas in the transfer layer (H_k), it depends on the number of nodes in the preceding layer (N_H^{k-1}). The maximum number of layers (k) for a regression type NN is recommended to be three [6], where $k > 3$ has been shown to cause overfitting for minor increases in accuracy.

To improve generalization and reduce the overfitting, regularized performance function MSE_{reg} have been used in training the NN.

$$\text{MSE}_{\text{reg}} = \gamma\text{MSE} + (1-\gamma)\text{MSW} \tag{6}$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(t_i - i_i)^2 \tag{7}$$

$$\text{MSW} = \frac{1}{N}\sum_{j=1}^{N} W_j^2 \tag{8}$$

where γ is the regularization; t_i is the predicted by NN target; W_j represents the trained weights.

The NN hyperparameters were fine-tuned to achieve the optimal network configuration. Five main hyper parameters were optimized via the Bayesian search algorithm namely: learning rate, regularization (γ), number of neurons in each layer (N_H^k). The Bayesian search was performed >100 times using different random seeds and selected hyperparameters, then the optimal NN were chosen for this study. The best NN with lowest MSE out of the all NNs within the training grid was 14-(4-8-17)-2, where the number represents the following: input–(N_H^1–N_H^2–N_H^3)–output. Learning rate was determined to be 0.01 and best regularization parameter γ = 0.91. All calculations were carried out in Matlab 2021a.

2.5. Decision trees

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In regression problems with small-to-medium structured data, decision tree-based algorithms are considered state-of the-art.

The objective function, $\mathcal{L}(\phi)$, for training is expressed in Eq. (9) consisting of sum of loss function and regularized objective:

$$\mathcal{L}(\phi) = \sum_i l(\widehat{y_i}, y_i) + \sum_k \Omega(f_k) \tag{9}$$

where the loss function $l(f)$ measures the difference between the predicted data $\widehat{y_i}$ and target data y_i (i.e., variance), while $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda|\omega|^2$ is called regression tree functions which penalize the complexity of the model. T,λ, and ω represents the number of leaves in the tree, weight decay (also commonly known as $L2$ regularization), and leaf weights, respectively.

Since the tree ensemble model in Eq. (4) includes functions as parameters and cannot be optimized using traditional optimization methods in Euclidean space, second-order approximation of Taylor expansion is introduced to optimize objective without interfering other settings.

With definition of as the instance set in leaf j. The objective function can be derived as Eq. (10) by expanding Ω:

$$\mathcal{L}(\phi) \approx= \sum_i \left[\sum_{i\in \boldsymbol{I}_j} g_i W_j + \frac{1}{2}\left(\sum_{i\in \boldsymbol{I}_j} h_i + \lambda\right) W_j^2\right] + \gamma T \tag{10}$$

where $\boldsymbol{I}_j$ is defined as the instant set for leaf j while g and h represents first and second order gradient statistics on the loss function using Taylor expansion.

By calculating the optimal weight W_j^* for leaf j that has a fixed tree structure $q(x)$:

$$w_j^* = -\frac{\sum_{i\in \boldsymbol{I}_j} g_i}{\sum_{i\in \boldsymbol{I}_j} h_i + \lambda} \tag{11}$$

Eventually, the optimal value in the objective function is calculated by Eq. (12):

$$\mathcal{L}(\phi) = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i\in \boldsymbol{I}_j} g_i\right)^2}{\sum_{i\in \boldsymbol{I}_j} h_i + \lambda} + \gamma T \tag{12}$$

Analogous to the NN, hyperparameter optimization was performed across selected optimizable parameters within Boost. A grid search was applied over the following parameters learning rate (η), depth of tree, maximum bins, and $L1$ and $L2$ regularization. The optimum XGboost configuration was found to be: η = 0.03 with 1000 iterations, depth = 6, bins = 50, $L1$ = 0, and $L2$ = 0.3. The algorithm and all calculations were carried out in python 3.8.

3. Results

3.1. Parametric model

The parametric model, Eq. (1), is evaluated on the set of input data outlined in section 2.1. As a result, a hardness prediction is obtained as a function of the microalloying chemistry, diameter (ϕ_D) and Jominy distance (J). To standardise the results, a mean is taken for all hardness values measured at the same ϕ_D. The predicted mean hardness ($\widehat{H}$) is plotted against the measured mean hardness in Fig. 4, for both J = 9 mm and J = 15 mm. The empirical model shows a good correlation with the measurements for the small (30–40 mm) diameters and is close comparison for the mid-range diameters (55–90 mm). However, as observed, the measured hardness varies significantly with ϕ_D. In the small ϕ_D range these oscillations are small with consistently decreasing hardness value, when error bars are considered. The predicted with parametric model (PM) hardness captures well some of the oscillations, especially where sufficient data was provided. The disparity between the empirical model and measurements is more pronounced at the larger diameters (>90 mm). Parametric models are designed to capture the mean of the target variable; however, owing to their highly conservative nature, they often underestimate the standard deviation in hardness at each diameter. This trend is even more apparent at the J = 15 mm distance where only one fluctuation in hardness at ϕ_D = 55 mm was captured. The PM provides a good average prediction. In overall, the PM has a MSE of 0.051 and 0.075 for J = 9 mm and J = 15 mm, respectively. For consistency, the error is taken for the normalized hardness

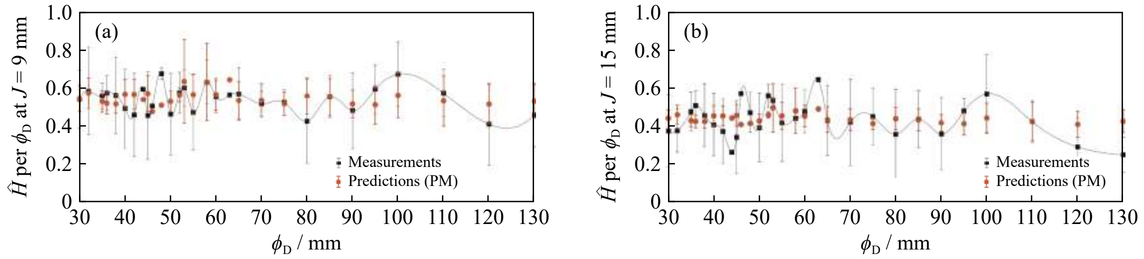


Fig. 4. Parametric model (PM) prediction against measured hardness values at (a) J = 9 mm and (b) J = 15 mm.

range between 0 and 1 in this and any further studies.

3.2. Neural networks

Once the network is trained using the training dataset and pre-processing outlined in section 2.1, the hardenability data is evaluated with the resulting NN model. The accuracy of the NN model is showed in Fig. 5, where the coefficient of determination for both predictions at $J = 9$ mm and $J = 15$ mm is 0.92. From the plot it can be seen that the lower range hardenability values are well predicted. There is a scatter of points around the mid-to-top range values ($\widehat{H} = 0.6$ to 0.9). To investigate this scatter in predictions, a plot of sorted measurements against predicted values is plotted in Fig. 6. The predicted with NN hardness (orange points) follow well the measured hardness across the entire range of values. All values were sorted in increasing order to provide clear visual contrast between the predicted and actual results. Again, the low and mid-range hardness are well predicted. The highest mismatch between prediction and measurements is seen at the location of largest change in hardness, at $\widehat{H} = 0.45$ to 0.60. This effect occurs at both Jominy distances. Identical procedure to the empirical model is followed where mean hardness measurements and NN predictions are plotted grouped according to diameter, as seen in Fig. 7. The NN model manages to capture the hardness variation much better than the empirical. The highest error is at $\phi_D = 45$ mm. This is contributed to the lack of data at this point. In overall, the MSE result for the neural network is $MSE_{NN} = 0.0048$, which is ten orders smaller than the parametric.

3.3. Extreme gradient boosting

The hardenability results using XGBoost are reproduced as shown in Fig. 8(a) and (b), which contains moderate improvements on predictions at $J = 9$ mm and $J = 15$ mm with R

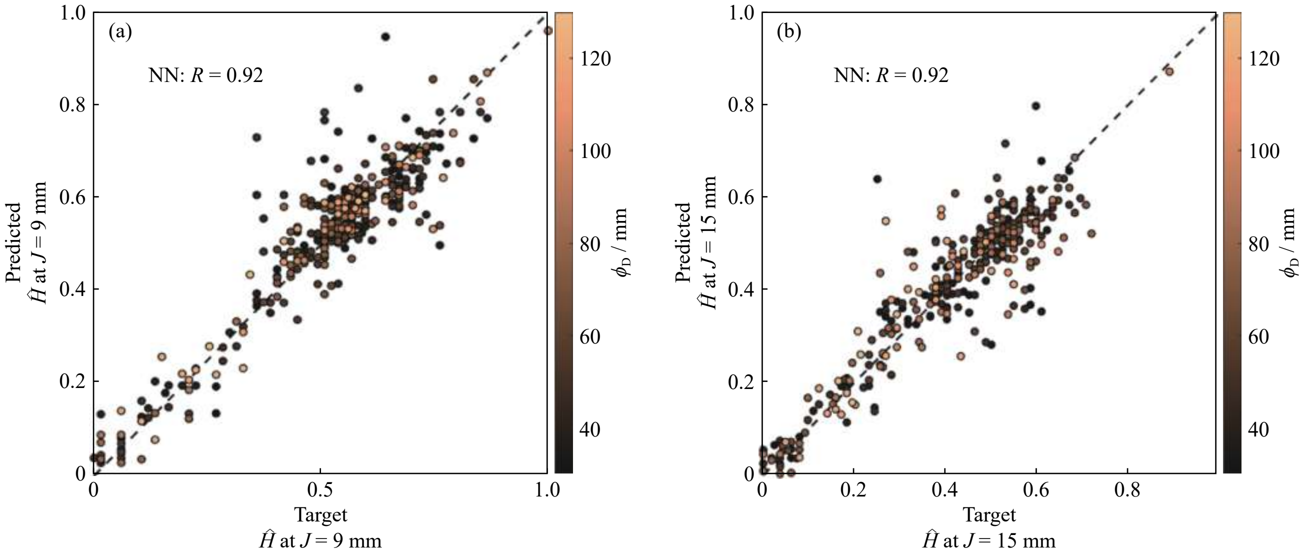


Fig. 5. Coefficient of determination obtained from NNs applied for hardness prediction at (a) 9 mm along the industrial bars and at (b) 15 mm distance from quenched end. The data is fitted across a range of diameters shown in by the colormap.

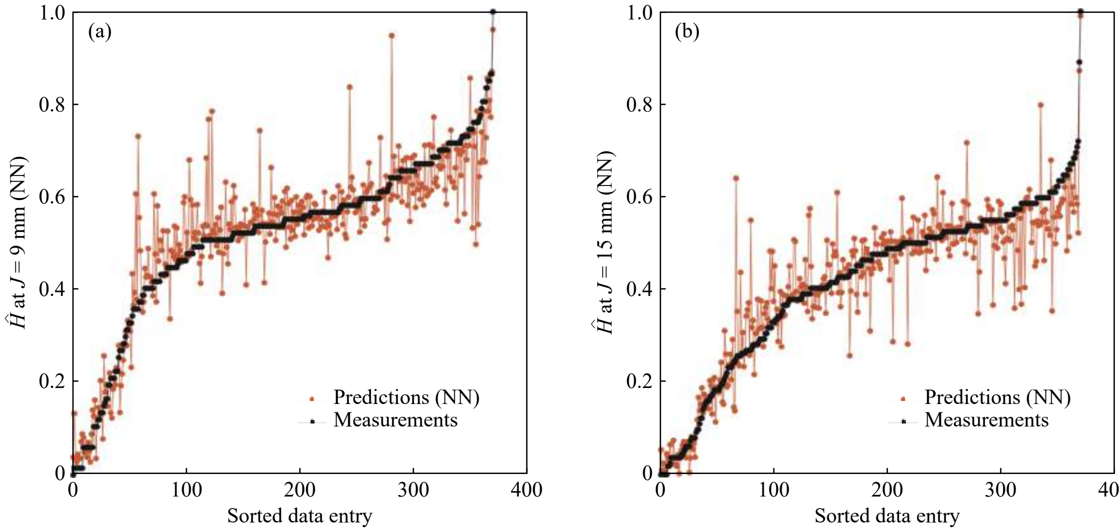


Fig. 6. Comparison between measured and predicted data on small industrial data set for steel grade 20CrMnTi at (a) 9 mm along the industrial bars and at (b) 15 mm distance from quenched end.

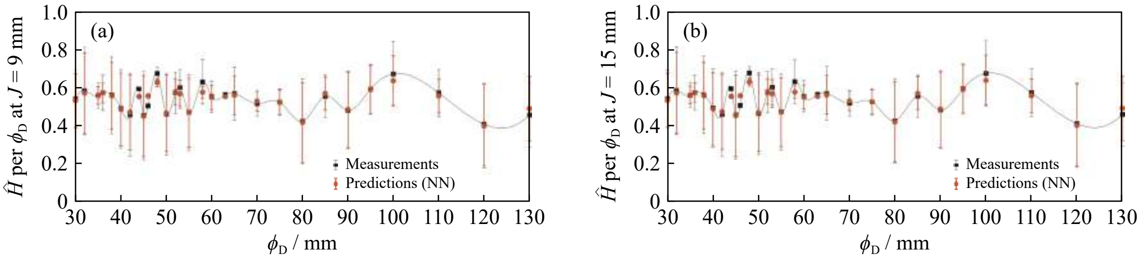


Fig. 7. Comparison between measured and predicted data on small industrial data set for steel grade 20CrMnTi for (a) J = 9 mm and (b) J = 15 mm. Values are grouped according to the diameter. The standard deviation shows the consistency in the achieved hardenability per ϕ_D both for the measurements and predictions of those measurements.

values of 0.94 and 0.96, respectively. In addition, XGBoost also has better performance in prediction of fluctuated data as shown in Fig. 8(c) and (d). Despite of experimental uncertainties during hardness measurement, the XGBoost prediction is showing less deviated scatters in comparison with experimental measurements. Mean hardness measurements grouped according to diameter and XGBoost predictions are plotted in Fig. 9. The model manages to capture the hardness variation best out of the three investigated models, where XGboost provides reasonable estimate even for diameters with only one measurement.

3.4. Importance of the data pre-processing

The added GMM pre-processing does increase the accur-

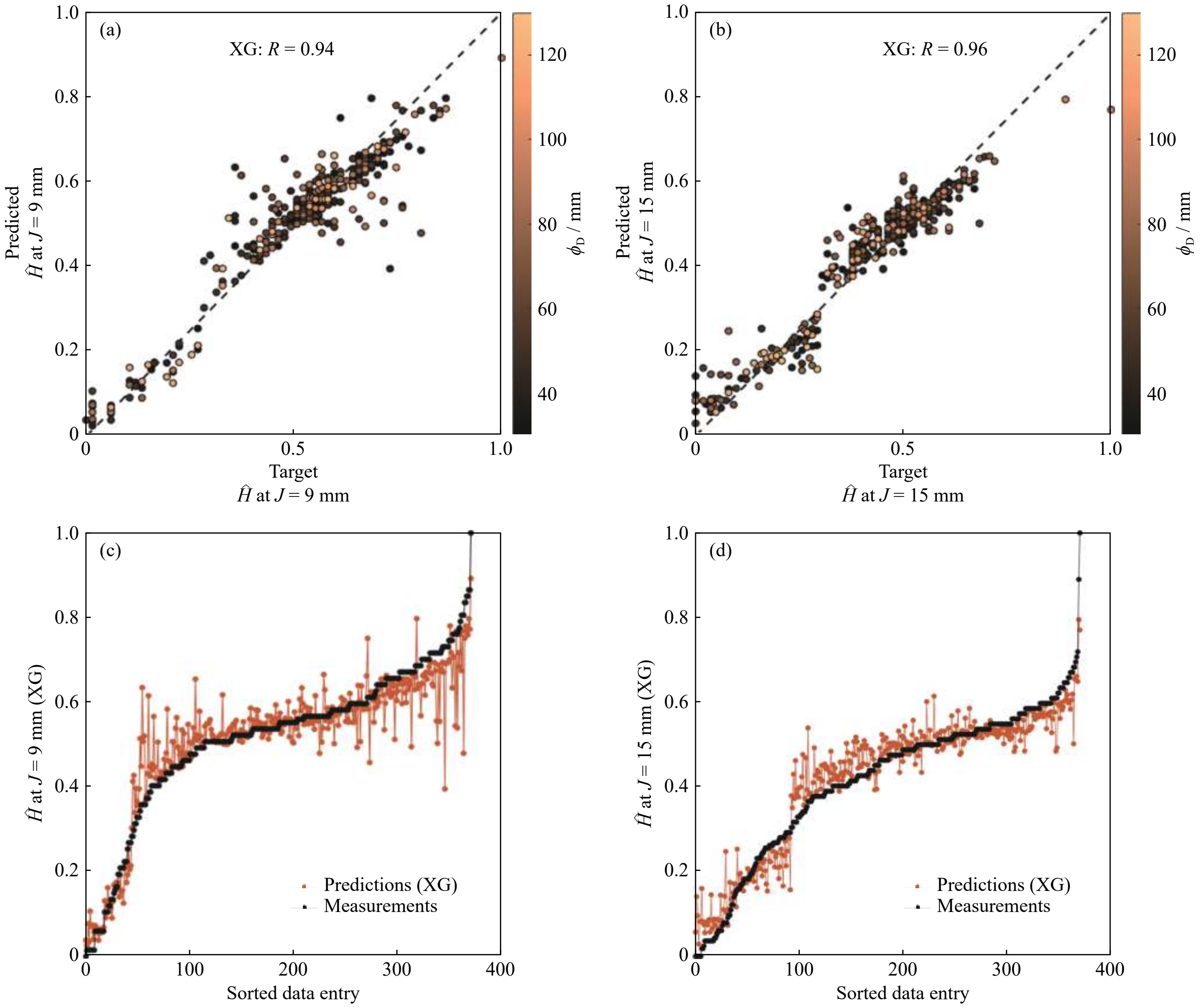


Fig. 8. Coefficient of determination obtained from XGboost applied for hardness prediction at (a) 9 mm along the industrial bars and at (b) 15 mm distance from quenched end. The data is fitted across a range of diameters shown in by the colormap. Comparison between measured and predicted data on small industrial data set for steel grade 20CrMnTi for sorted measurement entries in ascending order for (c) J = 9 mm and (d) J = 15 mm.

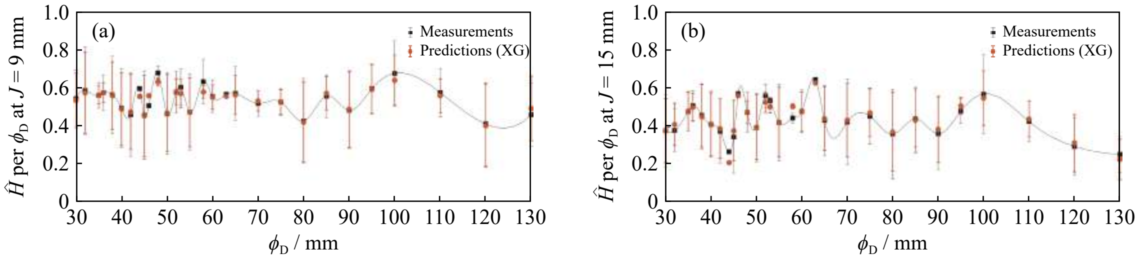


Fig. 9. Comparison between measured and predicted data on small industrial data set for steel grade 20CrMnTi obtained from XGboost for (a) J = 9 mm and (b) J = 15 mm. Values are grouped according to the diameter. The standard deviation shows the consistency in the achieved hardenability per ϕ_D both for the measurements and predictions of those measurements.

acy of the ML algorithms significantly. The introduced GMM pre-processing is added through encoding the labelled/categorised points as input to indicate the unaccounted microstructural effects. This engineered feature is in effect guiding the machine learning models indicating that there is a change in the conditions. As shown in Fig. 10, the NN coefficient of determination without taking account the GMM is R = 0.75 and R = 0.78, respectively for J = 9 mm and J = 15 mm. Similarly, the coefficient of determination for XGboost is R = 0.82 and R = 0.85, respectively for J = 9 mm and J = 15 mm. The spread of points is much larger, especially in the lower hardness range where the secondary peak in hardness

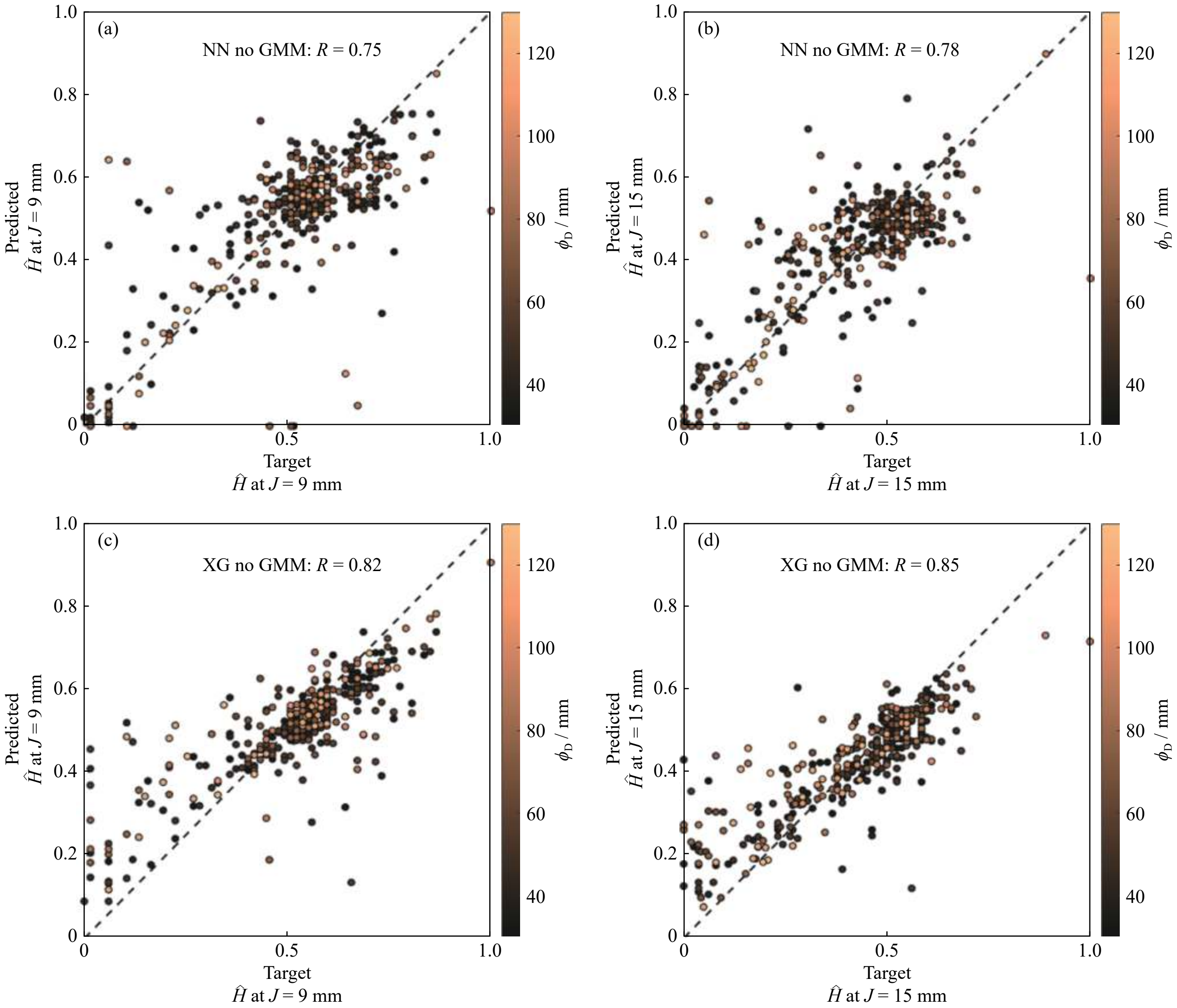


Fig. 10. Machine learning models without the implementation of the GMM classification. Neural network prediction on raw data without GMM classification is shown at Jominy distance of (a) 9 mm and (b) 15 mm. Similarly, prediction versus measurements (target) comparison for XGboost is shown in (c) for J = 9 mm and (d) for J = 15 mm. The coefficient of determination for both NN and XGboost is much lower. The spread of points is higher especially for the low range hardness values.

was observed.

3.5. Accuracy comparison—MSE and MAE

The algorithms are compared against two well established errors: mean absolute error (MAE) and MSE, as shown in Fig. 11. The ML algorithms perform much better than the parametric method. As seen above, XGboost has a minor advantage over the NN with a percent increase in accuracy and reduction in the overall deviation of the predicted values, seen from the error bars.

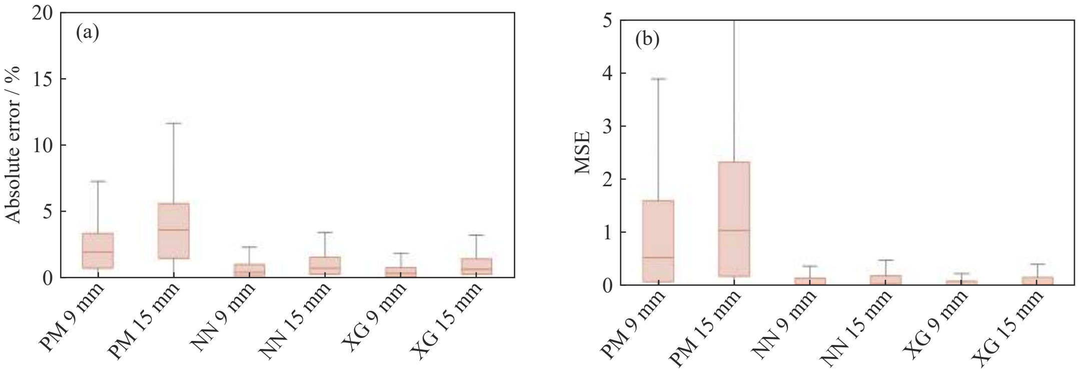


Fig. 11. Accuracy comparison between empirical/parametric modelling (PM), NNs, and XGboost. (a) Absolute error between predicted and measured values is shown. ML performs few orders of magnitude better whilst the difference between NN and XG is marginal. (b) MSE is also shown for all models. Similarly, XGboost has the lowest error, hence best performance.

4. Discussion

4.1. Challenges in industrial data for machine learning

Industrial data is subject to multiple challenges, which must be accounted for prior to any data-driven modelling, namely:

(1) Lack in quantity—small data set;

(2) Lack in diversity—limited variables;

(3) Class imbalance—few measurements per category.

Unlike controlled laboratory experiments, industrial test facilities do not have the time and resources to cover and test across all aspects of a process. Despite following established standards, the resulting data is limited, sufficient to cover customers' needs but lacking in quantity and diversity. For instance, the hardness data examined in this study is only limited to chemistry, Jominy distance, and diameter size. Metallurgical variables from the smelting and rolling process are considered in affect the hardenability although indirectly, as shown in Fig. 1. These additional unaccounted for process components such as the element, phase fraction (martensite and bainite) distribution, and grain size have their footprint on the hardness leading to the observed variations and inhomogeneity in hardenability across the test samples, Fig. 3. To account for these variable influences, the data in our study is categorized into homogeneous and inhomogeneous, as shown in Fig. 3. The inhomogeneous effect in hardenability is highly undesired by industry, hence its occurrence must be investigated and better understood. This effect is integrated in our data-driven algorithms through the performed GMM classification, Fig. 3. Each hardenability component is defined by its mean and covariance. The mixture of both homogeneous and inhomogeneous microstructure is defined by a vector of mixing proportions, where each mixing proportion represents the fraction of the population described by a corresponding component. This gained additional knowledge about the data set is directly integrated in all the machine learning algorithms.

Apart from the diversity of the data set, its quantity is also challenging. For the investigated 20CrMnTi gear steel grade, only 370 data points have been recorder spread out across 30 diameter sets. One can quickly notice that there is a potential class imbalance issue when modelling such data. To resolve this problem, a *k*-fold validation was implemented for both the NN and XGboost models. This allows all the data to be utilized covering a *k*-number of variations of training and testing data sets. This method significantly reduces bias before implementing data for fitting, and significantly reduces variance as most of the data is also being used in validation set. Interchanging, *k*-fold validation increases the effectiveness and robustness of the data-driven modelling.

4.2. Limitations of empirical models

In empirical equations, specific complex dynamical processes are replaced by simplified physical approximations whose associated parameter values are estimated from data i.e., PM. PM models are still relied upon in industry; however, they have multiple fundamental limitations:

(1) Limited to specific alloy grades;

(2) Time consuming and costly: require extensive testing and tabulation;

(3) Assume no interaction between microalloying elements;

(4) Assume homogeneous microalloying.

Most numerical models forecasting the Jominy profile of steels provide good subject to specific of steel grades, on which their internal parameters were tuned, and do not show good generalization properties when applied outside those ranges, as the relationship linking such parameters with the steel chemical composition are mostly empirical and difficult to extend. Moreover, often the accuracy is acceptable only on a few points of the Jominy curve. This is due to each alloy element being individually analyzed, while interactions are neglected. As shown in Fig. 2 and highlighted in the introduction, a lot of the microalloying elements cause indirect effects on the hardenability of the alloy. These interactions between elements are problematic to capture through empirical equations hence more advanced modelling methods are implemented. The investigated in this work empirical models clearly shows the drawback of such methods. As shown in Fig. 4, empirical methods can predict a mean value but fail to capture the range or deviation in hardness.

4.3. Comparison between NN and XGboost

Machine learning used collaboratively with fundamental knowledge of metallurgy has been shown to be superior to empirical models [10,21]. Hence, there is great interest in using ML models to gain new insight directly from observations and high-resolution model simulation. NNs as one of the main representatives of machine learning come into consideration when faced with small challenging data sets. NN are capable of compensating for the shortcomings of empirical formulas. NN can capture the non-linear interaction between the elements as well as the indirect effect of process parameters, once integrated with knowledge-based classification. As seen in Fig. 5, NN successfully predicts the measurements. The overall accuracy is satisfactory high with $R = 0.92$; $R = \sum_i \frac{(\widehat{y_i} - y_i)^2}{(y - y_i)^2}$, where (y_i) is the measured value and $(\widehat{y_i})$ is the predicted value, employed to evaluate the performance of the ML models. Naturally, the more the hardness measurement per ϕ_D, the better the NN prediction. This is evident from Fig. 7, where at diameters (ϕ_D) with only one or two measurements, the hardness predictions were severely over- or underestimated by the NN model. Thus, despite being able to account for multiple direct and indirect influences on the microstructure, NN falls short in predicting: (1) out-of-range and (2) limited data relationships, accounting for class imbalance.

Interestingly, the deviation in hardness predicted by the NN model corresponds well to the standard deviation range of the measurements, taken per diameter.

Another examined in this work machine learning algorithms is XGboost. Gradient boosting is an approach where multiple residual calculation models are created and added together to make the final prediction. These models utilize a gradient descent algorithm to minimize the loss across all combined models. XGBoost is one of the leading algorithms that effectively utilizes this approach through: (1) boosting; (2) regularization.

XGBoost offers a systematic methodology for combination of the predictive power of multiple learners. The resultant is a single decision tree model resulting from the aggregated output from several trees, where each subsequent tree reduces the errors of the previous one. Another advantage of XGboost is in its regularization. The algorithm has an option to penalize complex models through both $L1$ and $L2$ regularization. Regularization helps in preventing overfitting and as seen from our results Fig. 9, the predictions with XGboost are more conservative than the NN. The fluctuations are smaller especially for the predictions at J = 15 mm. The predicted hardness standard deviation is also smaller than that of the NN. Interestingly, XGboost manages to provide good estimation even for points where data is insufficient such as hardness at $\phi_\mathrm{D} = 45$ mm and $\phi_\mathrm{D} = 55$ mm. All this combined leads to the overall better performance $R = 0.94$ and $R = 0.96$ coefficient of determination, respectively for J at 9 and 15 mm and best performance in terms of MSE in overall as shown in Fig. 11. Thus, XGboost shows the highest potential for predicting small industrial data sets with class imbalance and large inhomogeneity issues in a conservative manner.

5. Conclusion

In this work, hardenability prediction of gear steels is investigated in both a data-driven and theoretical context. The limitations of current empirical methods are identified and clearly illustrated. A number of challenges are discussed, and solution proposed to handle small data sets with a large number of categories. Effective strategies for inferring microstructure homogeneity and microalloying interaction are utilized. Two modelling routes are investigated combining XGboost and NNs with Gaussian process algorithm to significantly improve the hardenability prediction. The data-driven, optimized and intelligent machine learning methods show significant advantages over the traditional costly and time-consuming experimental parametric studies. With absolute errors lower that 5%, XGboost and NN prove capable of dramatically accelerating process optimization and quality assessment. The high accuracy of prediction ensures a reliable forecast of gear steel performance contribution to lowering costs and improving efficiency of industrial production.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M. Vctor Li, D.V. Niebuhr, L.L. Meekisho, and D.G. Atteridge,

A computational model for the prediction of steel hardenability, *Metall. Mater. Trans. B*, 29(1998), No. 3, p. 661.

[2] V. Javaheri, A. Pohjonen, J.I. Asperheim, D. Ivanov, and D. Porter, Physically based modeling, characterization and design of an induction hardening process for a new slurry pipeline steel, *Mater. Des.*, 182(2019), art. No. 108047.

[3] E.C.H.C. O' Brien and H.K. Yeddu, Multi-length scale modeling of carburization, martensitic microstructure evolution and fatigue properties of steel gears, *J. Mater. Sci. Technol.*, 49(2020), p. 157.

[4] P.H. Maynier, J. Dollet, and P. Bastien. Prediction of microstructure via empirical formulas based on CCT diagrams, [in] *The 107th AIME Annual Meeting*, Denver, Colorado, 1978, p. 163.

[5] D. Khan and B. Gautham, Integrated modeling of carburizing-quenching-tempering of steel gears for an ICME framework, *Integr. Mater. Manuf. Innovation*, 7(2018), No. 1, p. 28.

[6] S. Feng, H.Y. Zhou, and H.B. Dong, Using deep neural network with small dataset to predict material defects, *Mater. Des.*, 162(2019), p. 300.

[7] C.G. Shen, C.C. Wang, X.L. Wei, Y. Li, S. van der Zwaag, and W. Xu, Physical metallurgy-guided machine learning and artificial intelligent design of ultrahigh-strength stainless steel, *Acta Mater.*, 179(2019), p. 201.

[8] F.E. Bock, R.C. Aydin, C.J. Cyron, N. Huber, S.R. Kalidindi, and B. Klusemann, A review of the application of machine learning and data mining approaches in continuum materials mechanics, *Front. Mater.*, 6(2019), art. No. 00110.

[9] H.K.D.H. Bhadeshia, Neural networks in materials science, *ISIJ Int.*, 39(1999), No. 10, p. 966.

[10] S.W. Wu, J. Yang, and G.M. Cao, Prediction of the Charpy V-notch impact energy of low carbon steel using a shallow neural network and deep learning, *Int. J. Miner. Metall. Mater.*, 28(2021), No. 8, p. 1309.

[11] Z.H. Deng, H.Q. Yin, X. Jiang, C. Zhang, G.F. Zhang, B. Xu, G.Q. Yang, T. Zhang, M. Wu, and X.H. Qu, Machine-learning-assisted prediction of the mechanical properties of Cu–Al alloy, *Int. J. Miner. Metall. Mater.*, 27(2020), No. 3, p. 362.

[12] J. Friedman, T. Hastie, and R. Tibshirani, Additive logistic regression: A statistical view of boosting (With discussion and a rejoinder by the authors), *Ann. Stat.*, 28(2000), No. 2, p. 337.

[13] T.Q. Chen and C. Guestrin, XGBoost: A scalable tree boosting system [in] *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, 2016, p.785

[14] J. Bennett and S. Lanning, The Netflix prize, [in] *Proceedings of KDD Cup and Workshop 2007*, San Jose, 2007.

[15] K. Song, F. Yan, T. Ding, L. Gao, and S.B. Lu, A steel property optimization model based on the XGBoost algorithm and improved PSO, *Comput. Mater. Sci.*, 174(2020), art. No. 109472.

[16] E.T. Akinlabi1, O.M. Ikumapayi, O.P. Bodunde, B.A. Adaramola, I.D. Uchegbu, and S.O. Fatoba, Impact of quenching on the hardenability of steels EN-3 (~1015), EN-8 (~1040) and EN-24 (~4340) during Jominy end quench technique*. Int. J. Emerging Technol*. 11(2020), No. 5, p. 290.

[17] F. Wetschoreck, T. Krabel, and S. Krishnamurthy, *8080labs/Ppscore: Zenodo Release* [2020-10-15]. DOI: 10.5281/zenodo.4091345

[18] R.A. Waltz, J.L. Morales, J. Nocedal, and D. Orban, An interior algorithm for nonlinear optimization that combines line search and trust region steps, *Math. Program.*, 107(2006), No. 3, p. 391.

[19] P. Schüler, Calculation of hardenability in the Jominy end quench test on the basis of the Chemical composition of steel, *Revue de Métallurgie*, 89(1992), No. 1, p. 93.

[20] F. Burden and D. Winkler, Bayesian regularization of neural networks, [in] D.J. Livingstone ed, *Artificial Neural Networks*, *Methods in Molecular Biology*™, Humana Press, 458(2008), p. 23.

[21] S. Feng, H.Y. Zhou, and H.B. Dong, Application of deep transfer learning to predicting crystal structures of inorganic substances, *Comput. Mater. Sci.*, 195(2021), art. No. 110476.