Received: 15 May 2018 | Revised: 19 December 2018 | Accepted: 10 January 2019

DOI: 10.1002/sam.11406

RESEARCH ARTICLE

WILEY

Materials data analytics for 9% Cr family steel

Vyacheslav N. Romanov[1] | **Narayanan Krishnamurthy**[1] | **Amit K. Verma**[2] | **Laura S. Bruckman**[2] | **Roger H. French**[2] | **Jennifer L.W. Carter**[2] | **Jeffrey A. Hawk**[3]

[1] U.S. Department of Energy, National Energy Technology Laboratory, Pittsburgh, Pennsylvania
[2] Department of Materials Science and Engineering, Case Western Reserve University, Cleveland, Ohio
[3] U.S. Department of Energy, National Energy Technology Laboratory, Albany, Oregon

Correspondence
Vyacheslav N. Romanov, U.S. Department of Energy, National Energy Technology Laboratory, 626 Cochrans Mill Road, Pittsburgh, PA 15236.
Email: romanov@netl.doe.gov

Funding Information
U.S. Department of Energy, DE-FE0028685.

A materials data analytics (MDA) methodology was developed in this study to evaluate publicly available information on 9% Cr family steel and to handle nonlinear relationships and the sparsity in materials data for this alloy class. The overarching goal is to accelerate the design process as well as to reduce the time and expense associated with qualification testing of new alloys for fossil energy applications. Data entries in the analyzed data set for 82 iron-base alloy compositions, several processing parameters, and results of tensile mechanical tests selected for this study were arranged in 34 columns by 915 rows. While detailed microstructural information was not available, it is assumed that the compositional space for the 9 to 12% Cr steels is limited such that all data entries have a tempered martensitic microstructure during service. Establishing a hierarchy of first-order trends in the publicly available data requires the MDA to filter out the biases. Complexity of the phase transformations and microstructure evolution in the multicomponent alloys (using 21 chemical elements) with major influence on mechanical properties, leads to inefficiency in direct application of unbiased linear regression across the entire data space. To address the nonlinearity, analyses of tensile data were performed in composition-based clusters. Clusters corresponding to moderately frequent patterns and maximized information gain were further refined by using *p*-norm distance measures, matching the alloy classification groups adopted by industry. The evolutionary method of propagating an ensemble of competing cluster-based models proved to be a viable option in dealing with scarce, multidimensional data.

KEYWORDS

alloy, clustering analysis, pattern discovery, strength

1 | INTRODUCTION

Motivation for this research comes from the desire to shorten the rigorous and time-consuming alloy qualification (standardization) process, for new fossil energy materials applications. The preliminary focus of the data science effort is on 9 to 12% Cr martensitic-ferritic steels used as structural materials in steam boiler and turbine applications in power generation. One main consideration for using this alloy class is its relatively high microstructural stability at the operating temperature over time, since power plants have a design lifetime expectation of over 30 years.

Abbreviations: A_{c3}, upper critical point (to austenitize alloy); AGS, prior austenite average grain size; AGS#, prior austenite grain size number; AIC, Akaike information criterion; ASME, American Society of Mechanical Engineers; ASTM, American Society for Testing and Materials; BIC, Bayesian information criterion; c-IG, combinatorial pattern search to maximize information gain; COST, European Cooperation in Science and Technology's specifications; CPJ, National Energy Technology Laboratory current program's specifications; Elong, sample elongation to failure; IG, information gain; kNN, k-nearest neighbors algorithm; MDA, materials data analytics; mod, modification; NETL, National Energy Technology Laboratory; NIMS, National Institute for Materials Science; P91/P92, American Society for Testing and Materials' specifications; PAM, partitioning around medoids; PCA, principal component analysis; PLS, projection to latent structures; RA, reduction in area; UTS, ultimate tensile strength; YS, yield strength; α-Fe, α-phase of iron (ferrite); γ-Fe, γ-phase of iron (austenite); %wt., % by weight.

Further improvements in efficiency of a power plant can only be gained by using materials that allow for higher temperature or pressure (or both) for the cumulative hundreds of thousands of hours of operation [1,2,13,33,52,53,54]. The objectives of this work were to develop a data-centric framework for analysis and characterization of materials that could be used in fossil energy power plants and, by doing so, be in a position to better predict their mechanical properties.

Similar studies have tested a variety of unbiased machine learning approaches [29]. For example, a neural network with 51 predictor variables was used to model crack growth rate (under a fatigue stress regime) in nickel-based superalloys. This neural network was able to virtually explore new phenomena in such instances where certain vital information cannot be directly obtained via experiments [16]. Subsequently, a neural network model was developed to predict yield and tensile strength of steel as a function of 108 variables, including chemical composition and metalworking parameters [49]. Hancheng et al. [20] proposed an adaptive fuzzy neural network model to predict strength based on composition and microstructure. Alternatively, support vector regression combined with particle swarm optimization was utilized to set up a model for prediction of the corrosion rate of carbon steel exposed to seawater environment [55].

Some of the materials research and development activities (and a good portion of them too) are proprietary, which makes it particularly difficult to access and compile high-quality information. Consequently, many researchers have relied on public data. One source of data for materials for energy applications is made available through the National Institute for Materials Science (NIMS). A study on fatigue strength prediction using information available in the NIMS database [17] initially employed principal component analysis (PCA) and then performed partial least squares regression on the clusters identified by PCA. Large R^2 values, ranging between 0.88 and 0.94, were obtained for individual clusters. More recently, Agrawal et al. [3] using the same data set demonstrated that several advanced data analytics techniques such as neural networks, decision trees, and multivariate polynomial regression can significantly improve goodness of fit over the previous efforts, with R^2 values approaching and exceeding 0.97.

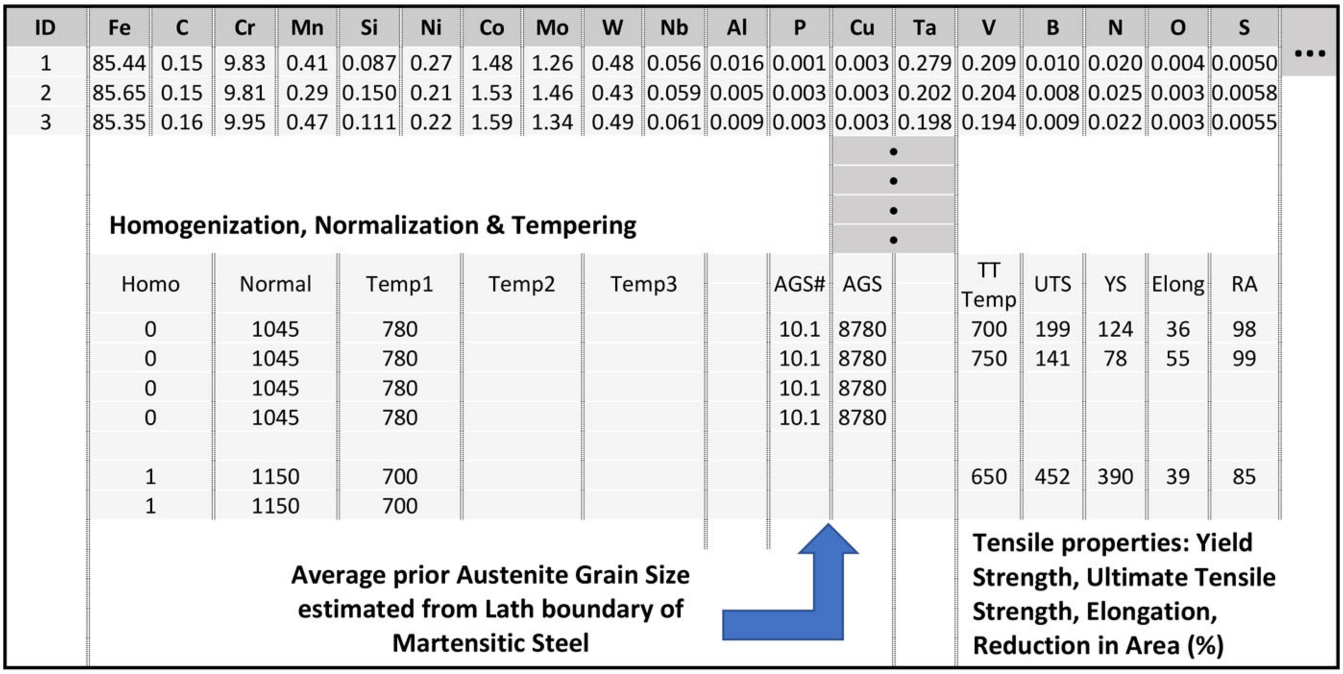
2 | DATA MANAGEMENT AND ANALYSIS SETUP

Data (see Table 1) used in this paper have come from a variety of sources: (a) NETL in-house research; (b) NIMS database [35–43]; (c) open literature; and (d) proprietary research [27]. A small subset of carbon steels with similar ferritic or martensitic lath microstructure (and average prior austenite grain size)—typically identified as 9 to 12% Cr (or 9% Cr family for simplicity) ferritic-martensitic steels (iron-chromium alloys with body-centered cubic crystal morphology)—were chosen for these data. Overall, the data spanned a slightly broader chemistry range for chromium (Cr), that is, 8 to 13% by weight. Data and codes were shared among the collaborators via the National Energy Technology Laboratory's (NETL) GitLab server. The data ID information was further screened and hidden from the data scientists participating in this effort, in terms of sources and pedigree. For this exercise, only elemental composition, processing parameters (homogenization, normalization, and tempering), prior austenite grain size, and mechanical properties information was extracted and arranged in 34 columns as shown in Table 1 (rhenium and hafnium columns are hidden because of the very limited data availability for those elements) for 82 alloy compositions with unique ID numbers. Here Homo stands for homogenization (1 = yes, 0 = none), Normal—the initial (normalization or austenitization) heat treatment temperature (°C), Temp1 (as well as 2 and 3)—the subsequent, tempering heat treatment cycle's (in that order) temperature, AGS—prior austenite (a solid solution of carbon in iron, with face-centered cubic crystals stable at high temperatures) average grain size (per Japanese Industrial Standards, number of grains/mm^2), AGS#—number (n) determined by the American Society for Testing and Materials (ASTM) standard test methods (N = $2^{[n-1]}$; where N = number of grains/inch^2), TT Temp—test temperature (°C), UTS—ultimate tensile strength (MPa), YS—yield strength (MPa), Elong—sample elongation to failure (%), RA—reduction in area (%). AGS/AGS# is determined prior to tempering cycles.

The dimensions of the entire data space can be grouped into compositional variables (relative concentrations of the elements, by weight), processing parameters (the homogenization descriptor and heat treatment temperatures), microstructural descriptors, test parameters, and test outcome values. The combined data ranges of the composition and processing groups collectively represent the composition-processing subspace of this data set. The variables in this subspace, along with the test parameters, are treated as primary contributors to data-driven models predicting the test outcome which is the primary response variable. On the other hand, the microstructure descriptors are secondary response variables controlled by the primary contributors. Microstructure is defined here as the structure of a prepared surface of material as revealed by a microscope with ×100 magnification. The microstructural information can be used either for indirect validation of the models or as a secondary (ie, dependent) contributor.

The initial data analysis setup was based on the Case Western Reserve University's informatics infrastructure, CRADLE, which is a Hadoop-based NoSQL technology for the ingestion and rapid processing of data [25]. Within CRADLE, data analytics using open-source programs, such as Python and R libraries, can be performed on the shared data sets. Figure 1 illustrates the key elements of the baseline methodology starting with unbiased exploratory analysis. The

TABLE 1 Tidy tensile strength data set: 34 columns by 915 rows; alloying elements, heat treatment (normalization and up to three tempering cycles) temperatures, prior austenite grain size (and AGS number), test temperature, UTS, YS, elongation to failure, and reduction in area



ID	Fe	C	Cr	Mn	Si	Ni	Co	Mo	W	Nb	Al	P	Cu	Ta	V	B	N	O	S	•••
1	85.44	0.15	9.83	0.41	0.087	0.27	1.48	1.26	0.48	0.056	0.016	0.001	0.003	0.279	0.209	0.010	0.020	0.004	0.0050	
2	85.65	0.15	9.81	0.29	0.150	0.21	1.53	1.46	0.43	0.059	0.005	0.003	0.003	0.202	0.204	0.008	0.025	0.003	0.0058	
3	85.35	0.16	9.95	0.47	0.111	0.22	1.59	1.34	0.49	0.061	0.009	0.003	0.003	0.198	0.194	0.009	0.022	0.003	0.0055	
													•							
													•							
													•							
													•							

Homogenization, Normalization & Tempering

Homo	Normal	Temp1	Temp2	Temp3		AGS#	AGS		TT Temp	UTS	YS	Elong	RA
0	1045	780				10.1	8780		700	199	124	36	98
0	1045	780				10.1	8780		750	141	78	55	99
0	1045	780				10.1	8780						
0	1045	780				10.1	8780						
1	1150	700							650	452	390	39	85
1	1150	700											

Average prior Austenite Grain Size estimated from Lath boundary of Martensitic Steel

Tensile properties: Yield Strength, Ultimate Tensile Strength, Elongation, Reduction in Area (%)

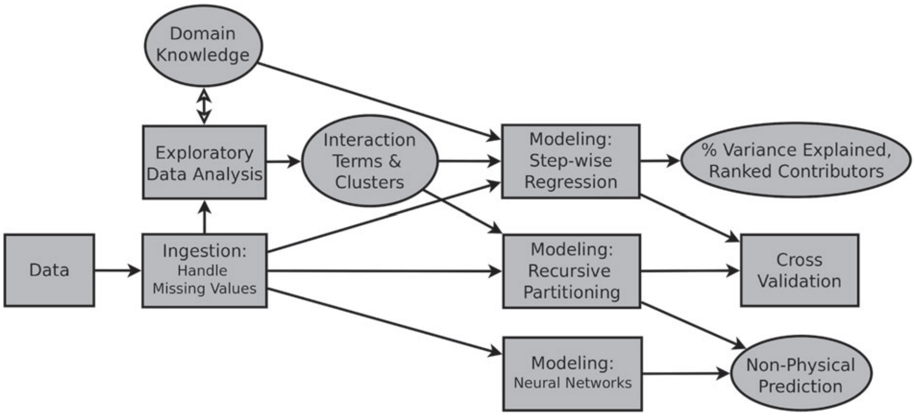

Domain Knowledge
Exploratory Data Analysis
Interaction Terms & Clusters
Modeling: Step-wise Regression
% Variance Explained, Ranked Contributors
Data
Ingestion: Handle Missing Values
Modeling: Recursive Partitioning
Cross Validation
Modeling: Neural Networks
Non-Physical Prediction

FIGURE 1 Data analytics baseline approach: After data ingestion, exploratory analysis sets the basis for identifying relations (eg, correlations and clustering), along with domain knowledge, to guide data-driven modeling, with further examination using cross-validation techniques to test for underfitting and overfitting

next step is to incorporate various data-driven analytical techniques and generate models. Validation of the selected model features and assumptions will ultimately be guided by domain knowledge.

This paper reflects a portion of the study aimed at distilling unbiased data-driven information. However, the authors were cognizant of the inherently biased nature of a database of the heritage data, given the biased nature of scientific experimental design and data reporting. For example, only the successful materials with best properties were intensely examined; other less promising lines of inquiry were abandoned and under-reported. This was identified during the exploratory data analysis and taken into consideration at the advanced stages of modeling, particularly, dealing with variable interdependencies and biases by design. Hence, some generalized or empirical form of domain knowledge was essential in developing even "unbiased" (i.e., without specific metadata or physics models) data-driven modeling. It turns out that such an approach is necessary for several other reasons such as data sparsity, data gaps, and the interpretability of the findings so that the domain scientists can naturally utilize the produced results.

3 | STATISTICAL ANALYSIS

An important part of the preliminary analysis was to characterize the single variable distributions and identify the variable interaction terms. Pairs of variables can be characterized by covariance and correlation measures to describe a degree

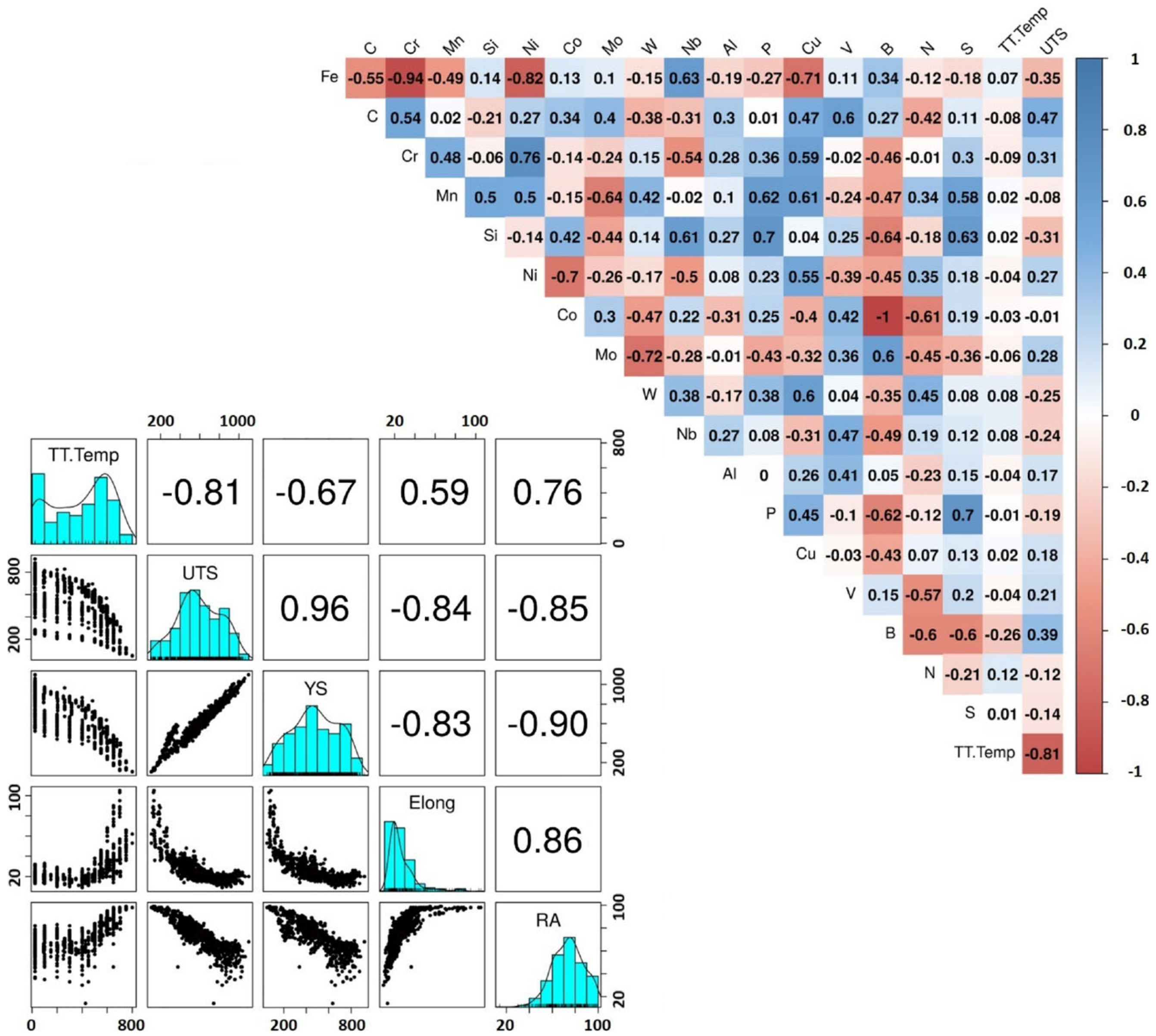


FIGURE 2 Bottom left: Pairwise scatter plots and Pearson correlation coefficients of tensile properties and test temperature. Top right: Heatmap highlighting the correlations. TT.Temp, test temperature, UTS, ultimate tensile strength, YS, yield strength, Elong, % elongation, RA, % reduction in area; the rest are chemical elements present in the alloys

of association among them [46]. The results were visualized (as in Figure 2) to facilitate better understanding of the data availability and detection of anomalies. Pair plots show strong correlation between the temperature and tensile test outcomes, while the associated heatmap shows correlations within the compositional space as well. Regression modeling can then be performed by searching for a combination of linear [31] and basic nonlinear parametric functions [9,51] that would minimize the number of parameters per number of available data points, provided that the models deliver acceptable levels of prediction accuracy. The primary justification for using various selection criteria to preferentially search for sparse models is that enabling fewer (i.e., most meaningful) features means reducing computational costs of model training on a limited (e.g., expensive) set of data, reducing the chance of overfitting, and making it easier for the domain scientists to interpret and test the underlying model assumptions. Linear regression can often be used as a practical alternative to more advanced statistical methods, if the underlying assumptions (essentially, the residuals being identically distributed and independent) are valid. It is necessary to check for the residuals' linearity, homoscedasticity, no correlation with predictors, and small Cook's distance [14].

Even within a linear approximation, selecting primary predictors can frequently be challenging if the input variables are strongly correlated. For example, an apparently strong relationship between a predictor and the outcome could be due to its strong spurious correlation with another predictor. By decomposing a secondary predictor y into its projection onto the primary predictor x (with hypothetical causal relation to the outcome z) and an orthogonal residual $\overline{y}$, it is elementary to show (Equation 1) that its correlation to z (e.g., expressed as the sample Pearson correlation coefficient r_{yz}) is roughly proportional to the natural correlation between x and z (i.e., ρ_{xz}) if the residuals do not correlate with the outcome:

$$r_{yz} = r_{xy} \times (\rho_{xz} + r_{\bar{y}z}). \tag{1}$$

For example, the heatmap in Figure 2 shows that tungsten is frequently used to replace molybdenum ($r_{xy} = -0.72$). If the observed correlation between UTS and molybdenum ($\rho_{xz} = 0.28$) is assumed to be natural (in the absence of strong correlation between molybdenum and other major predictors) then the expected correlation for tungsten ($-0.72 \cdot 0.28 \approx -0.2$) accounts for over 80% of the correlation between UTS and tungsten ($r_{yz} = -0.25$) which can tentatively be treated as spurious. Hence, tungsten can be ignored in lieu of a principal predictor variable (named either "Mo" or Mo-W pair) to account for a combined effect of the molybdenum and tungsten compositional changes, in a preliminary analysis.

Generic statistical techniques like PCA [56] and projection to latent structures (PLS) [19] are less transparent while their ease of use is outweighed by the difficulty of their output's interpretation in the domain science. They become increasingly less useful when expanded into new data spaces as their findings' applicability is limited to the space within the implicit prior design assumptions inherent to the original data set. Additionally, it is helpful and instructive to learn more about the actual, unintended consequences (such as $r_{\bar{y}z} \neq 0$) of having a secondary component (either composition or processing) added for some expected side benefits but with no intended direct effect on the primary target (as, for example, tensile strength in this study).

From a data science perspective, univariate pairwise correlations help to identify the designer biases, and likely, the associated empirical material design rules. Incidentally, it is important to distinguish between the shared (e.g., publicly available) data and the entire trail of historical trial and error outcomes of both successful and failed experimental and theoretical studies. From a historical perspective, these were used to forge the current design practices (e.g., for molybdenum/tungsten atomic substitution ratios, observed in the 8 to 13%wt. chromium steel data set).

In assessing data and models it became clear that if the relationships between the variables are highly nonlinear, partitioning of the overall parameter space into similarity-based clusters can decrease the extent of variance in predicting the response variable, while minimizing the prorated number of parameters per number of local data points in the composition-processing subspace. Another reason for using clusters is that microstructure and its evolution may vary between groups of polycrystalline alloys [10], leading to conflicting roles played by the chemical elements in the alloy compositions from different clusters. However, even if the element's impact is cluster-independent within a limited subset of compositions, it may be a major predictor within one cluster while being obscured by random errors within another. Its ultimate role depends on the product of predictor's variance within a cluster and the test outcome's sensitivity to the predictor's variation.

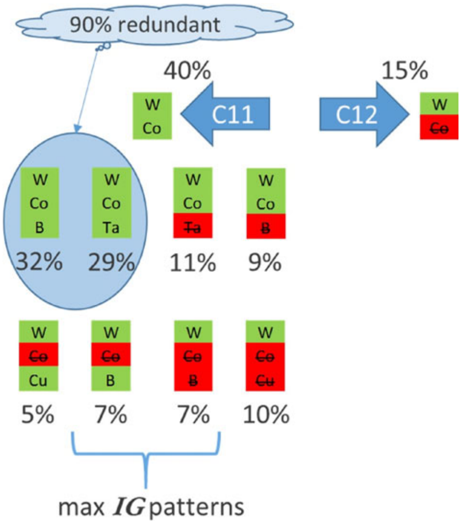


FIGURE 3 Composition-based clustering (shown for the tungsten-based, C1 cluster) to maximize information gain, c-IG [47]; C11 subcluster has two competing but 90%-redundant patterns (encircled) of splitting either by Ta (preferred) or by B; C12 has the patterns of splitting by B (preferred) and by Cu. IG is maximized by the more even partitioning as shown

A redundancy aware, moderate-frequency patterns-based approach was used to seed clusters in a meaningful way to maximize information gain (IG) [32]. Measures of a single random variable (e.g., modality of distribution, its mean and spread) were used to transform the data space prior to the combinatorial pattern search (Figure 3) hereafter referred to as c-IG. For the sake of a simplified classification exercise as an example, the nonzero contents were only discriminated by whether they were meaningful (i.e., sufficiently large) or not relative to their median value, for each element. Entries with values of less than 5% of the median were labeled not meaningful. There are multiple search algorithms available for frequent pattern mining [44] but discriminative pattern [12] analysis presented a useful strategy for effective classification of alloy compositions from the steel database.

The primary division of the composition space into major clusters C1 and C2 was based on the alloy's tungsten content (1 = yes and 2 = no in all splits). Next, C1 was split into C11 and C12—by the cobalt content. C2 was split into C21 and C22—by the vanadium content. Further partitioning was complicated by availability of redundant patterns and outliers as explained by subsequent refinement. C11 was partitioned by tantalum (competing with boron) into C111 and C112, C12—by boron into C121 and C122 (an alternative split, by copper as shown in Figure 3, polarizes C121 and creates an outlier group as shown in Figure 4). C22 was partitioned by molybdenum into C221 and C222. No moderately-frequent patterns were observed for C21, so it was not split by the simplified (yes/no) c-IG algorithm. To refine the cluster partitioning, the nearest neighbor algorithm [18] beginning with simultaneous origins at k preseeded cluster centers was used in combination with the $p \rightarrow \infty$ (known as

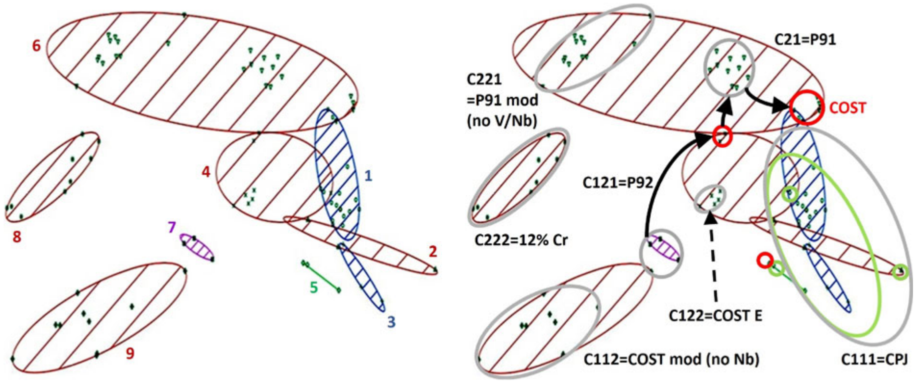


FIGURE 4 Left: 2D cluster visualization in the compositional space reduced to principal components, generated by using PAM. Right: c-IG/kNN [47] clusters (circled gray and labeled) plotted on top of the PAM clusters; homogenized compositions (including outliers) are circled green; nonhomogenized outliers are circled (and labeled where appropriate) red

infinity, supremum, max, uniform, or Chebyshev) limit of the *p*-norm (Equation 2) which was incorporated as a measure of distance between a given data point corresponding to a multidimensional composition vector and any of the clusters seeded in the previous step. This refinement is hereafter referred to as (modified) kNN.

Due to the equidistance problem arising in multidimensional spaces [7], by convention, cluster analysis is preceded by dimensionality reduction. Most common dimensionality reduction approaches make subsequent clustering solutions less flexible and more difficult to interpret within the framework of domain science. However, this can be circumvented by using increasingly larger *p* values in the *p*-norm defined below:

$$\|\boldsymbol{z}\|_p = \left(\sum_{i=1}^{n} |z_i|^p\right)^{1/p}, \text{for p} \in \mathbf{N} \text{ (natural number, } 1 \le \text{p} < \infty) \tag{2}$$

where $z_i = (y_i - x_i)$ is ith component (element concentration) of the change vector z: $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{S}$ (set of alloy composition vectors), $z \in \mathbf{R}^n$ (real coordinate space of n dimensions). In this case, the prior dimensionality reduction may not be necessary as the equidistance problem is alleviated.

The data points were sequentially added to the nearest cluster in the order of increasing nearest-neighbor distance. The resulting clusters were mapped over the clusters identified and visualized by using a common partitioning around medoids (PAM) [6,45] algorithm (Figure 4).

The PAM-generated clusters (nine elliptical shapes filled with diagonal lines, including cluster number 5 shown as a line segment) were visualized as a projection on two principal components (Figure 4). As the refined c-IG cluster patterns (encircled by solid gray lines) were mapped onto the PAM projection, it became clear that—except for several unassigned or questionable outliers (circled green if homogenized and red if not)—the proximity of the individual data points in the composition vector space is quite similar, using either approach. C222 perfectly matched one of the PAM clusters. C221 and C21 matched a couple of compact groups (subclusters within one PAM cluster). Similarly, C122 corresponds to a compact core group within a PAM cluster. C121 and C112 generally match two PAM clusters, aside from a transfer of the borderline-area data points from one of the two clusters into another. However, C111 roughly encompasses four PAM cluster objects, except for a couple of adjacent outliers (from PAM cluster number 5) with nearly identical compositions. Notably, this is the only c-IG cluster entirely made of homogenized alloys (green ellipse and two green-circled outliers from PAM clusters 1 and 2). The other clusters are nonhomogenized alloys. This was not a partitioning criterion; hence, it is a confirmation of the c-IG algorithm effectiveness.

The c-IG/kNN clusters can be tentatively classified as CPJ- (NETL current program's specifications), COST- (European Cooperation in Science and Technology's specifications), and P91/92-like (ASTMs' specifications) groups and their modifications to closely match the industry classification (as shown in Figure 4). More importantly, there is now some transparency on what separates the groups/clusters. The kNN refinement identified several outlier groups that were far apart from other data points as well as from the original cluster's median, including a compact group (red-circled object labeled COST, in the top-right corner, Figure 4) which moved out of C21 and merged with C111 outliers. The latter occurrence is an instance of c-IG/kNN producing a better match to a standard classification grouping (P91 and COST) than a simplified c-IG. Note, that the polarized (by multiple elements, in 3:1 ratio by data points) C21 was not easily split by c-IG. The outlier object at the intersection of PAM clusters 4 and 6 is conventionally classified as belonging to P92 and is better represented by c-IG grouping shown in Figure 3 (ie, C12 split by boron and copper, where B = yes and redundant Cu = no). The kNN refinement moved it out of C121 into C21 core. This is an instance of c-IG/kNN producing an improved match to PAM clustering but resulting in inferior performance relative to c-IG alone.

Transparency with respect to what separates the clusters is a crucial difference between the methods based on dimensionality reduction and the c-IG clustering approach. For one, it is now possible to clearly observe what composition elements are more prevalent in a certain cluster compared to its nearest

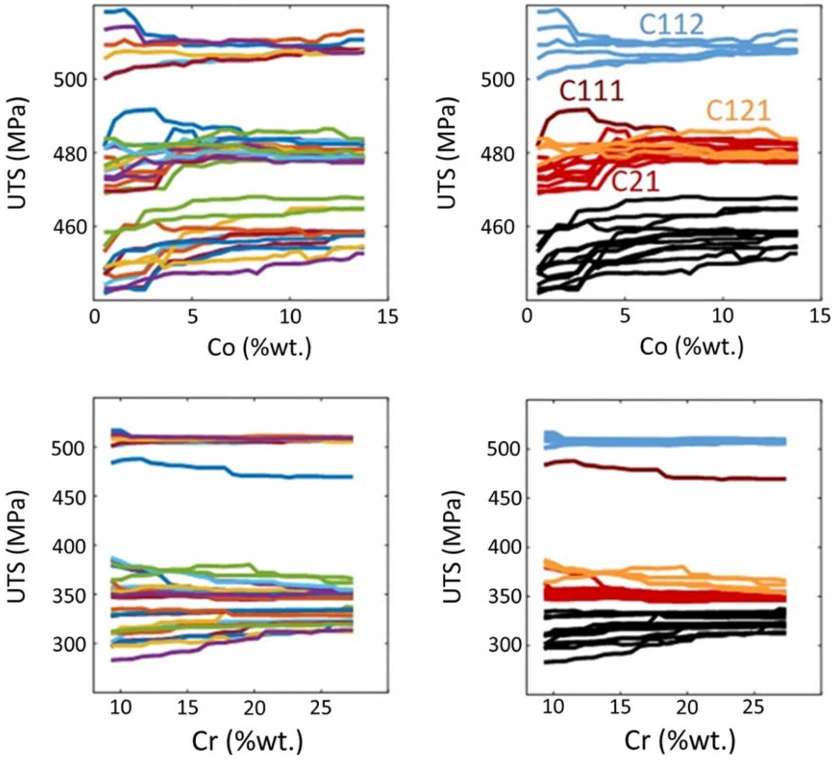


FIGURE 5 The examples of random forest [8,30] implementation: nonlinear training and forecasting of ultimate tensile strength by cobalt and chromium; left – by standard algorithm, right – color map used to match the composition clusters. In all the simulations, normalization temperature = 1045°C, tempering temperature = 650°C, test temperature = 600°C

neighbors. For example, the kNN refinement uncovered a pattern of increase in molybdenum (typically at the expense of tungsten) diagonally from the bottom left to top right corner, while vanadium and niobium tend to increase from left to right (in the PAM representation, Figure 4). The cluster-based data analysis also revealed that the alloys were processed at the same or similar normalization and tempering conditions (per reported data set) within each c-IG/kNN cluster. Analysis of the results also noted some similarity in prior austenite grain size within the clusters. C22 appears to be the only cluster with a moderate degree of within-cluster variability with respect to thermal processing and prior austenite grain size. Once again, those were not partitioning criteria but an additional confirmation of the c-IG algorithm effectiveness.

The following analysis of the data on ultimate tensile strength is used to illustrate advantages of the cluster-based approach to data-driven nonlinear model development. The initial modeling was implemented in Python using random forest [8] algorithms (Figure 5). Interestingly, as illustrated in the figure, the predicted patterns of UTS performance of alloys are distinctly similar within each c-IG/kNN cluster. However, the jagged individual plots demonstrate the problems associated with sparsity of the available data, particularly for individual clusters. Precision of the ensemble model predictions, with randomForest [30] algorithm trained on all clusters, was not adequate either. Typically, such approaches require very large quantities of data to achieve reasonable accuracy, which reduces their applicability to small cluster-based model development.

The alternative strategy for minimizing the number of parameters per number of data points is summarized as follows: (a) identify the global contributors with significant effect on the response variable and (b) add local variables, as long as the marginal benefit of adding a local variable is greater than that of adding an equivalent number of global variables on a “per corresponding number of data points” basis. Since the thermal processing parameters had been identified as the main global predictors, C22 will not be considered in the subsequent analysis, because both C221 and C222 subclusters had much lower corresponding process temperatures than the rest of the composition clusters. Additionally, this discussion is limited to the tensile test temperature at 600°C. The temperature was selected just above the high-temperature break-point on the UTS vs test temperature plots (not shown) which generally had three distinct piecewise-linear segments for the majority of analyzed alloys. The effects of minor-to-moderate variations in heat treatment (particularly, tempering) temperatures were assumed to be linear.

Despite the apparent simplicity of the general strategy above, its implementation is not straightforward and is sensitive to estimated (here from data reproducibility) uncertainty in the data. To develop generalizable solutions and to avoid overfitting the data, the search was biased toward the global contributors that were consistently strong across multiple clusters and had the most reproducible relationship with the response variable.

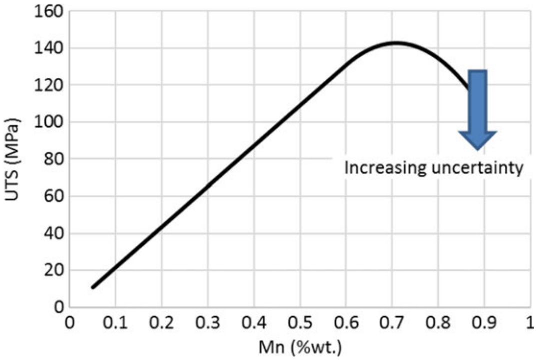


FIGURE 6 Differential contribution to ultimate tensile strength from a global variable: Mn

Goodness of fit measures like adjusted R^2 capture the extent of variance in predicting the response variable [21], but only within the input data subset used for the response-surface fitting. This information cannot be used for predicting just how well the locally trained model will perform outside of initial confinement. There is no reason why a single solution should be selected based on the best fit to data. Instead, an evolutionary approach was employed in this work to generate a limited number of alternative models originating on each cluster and then propagate them to neighboring clusters, with respect to either supremum distance measure or a specific predictor variation. The best-performing models can then be evaluated regarding potential physics insights and physics-based model refinement.

Regarding UTS, manganese and carbon (convoluted by correlation with nitrogen) were most consistently found to belong with the global predictor-variables. Manganese contribution appears to be almost universally linear up to 0.5% concentration (by weight), with consistently narrow range of the best-fit slope of UTS vs manganese concentration (Figure 6). The role of carbon (nitrogen) is more complex (Figure 7). Within 0.13 to 0.18% (by weight) range for carbon, a Gaussian feature appears to reside on top of the linear increase in UTS with increasing relative concentration (C/Fe). No clear trends were identified for carbon content outside of this compositional range. The feature's magnitude depends primarily on molybdenum content. Tungsten substitution with the equivalent impact on eutectoid carbon content [4] does not match the molybdenum effect. However, it was convoluted by nitrogen (or N/Fe) and some unidentified latent parameters, perhaps, variations in process or uncontrolled impurities. As was seen with carbon, a step-wise UTS increase (by 35 to 55 MPa) was observed with respect to nitrogen (at 0.018 to 0.022%, with its onset and magnitude being sensitive to alloy composition). Several alternative models may "survive" the selection-under-uncertainty process or further mutate as new data become available, for which they can serve as the (Bayesian) prior hypotheses. This process is ongoing and is complementary to the domain science knowledge discovery.

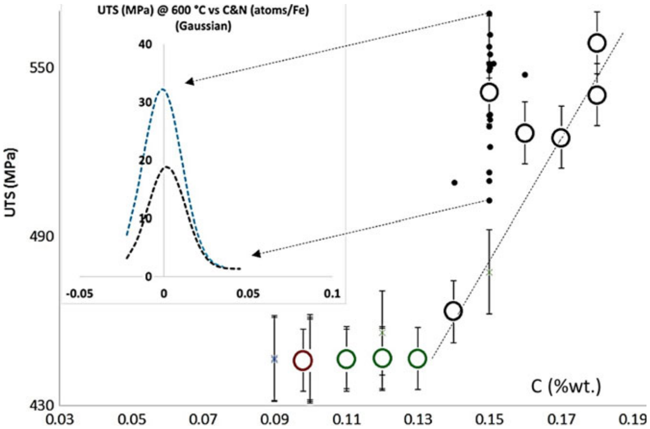


FIGURE 7 Global variable: C; while carbon and manganese are distinctly global contributors, the trends within clusters are convoluted; for example, in homogenized C111 (black dots), by molybdenum and copper (inset: black dashes = lower molybdenum concentration; blue dashes = higher molybdenum concentration) as well as nitrogen

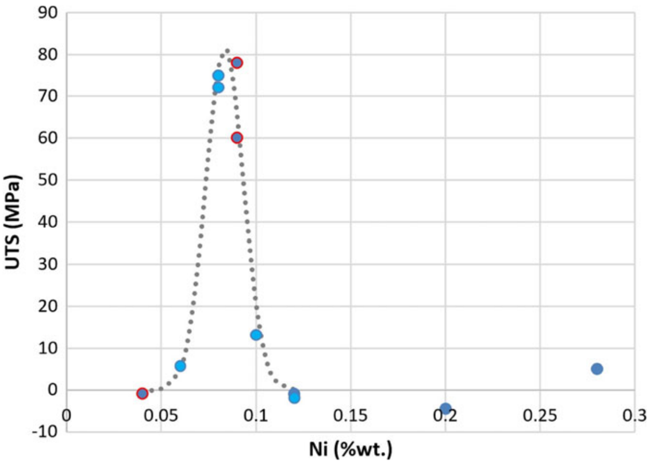


FIGURE 8 Differential contribution to ultimate tensile strength from a local (C21) variable: Ni, after a linear correction for chromium; the process temperature-related corrections (represented by color) were applied globally, uniformly across the entire data set

During the cluster-based generation of the predictor candidates pool exercise, some elements (most notably, chromium, copper, and nickel) exhibited strong local correlations with UTS. For example, C121 was fully represented by copper, with linear, negative trend. Interestingly, nickel data produced a sharp Gaussian shaped feature (after a linear, local correction for Cr in C21) near the lowest nickel concentration (Figure 8). However, such models were not consistent across the composition clusters, which may imply that either their effects are composition dependent or that the observed correlations were spurious. Clearly, more data are needed to further test and refine these models.

4 | DISCUSSION

The purpose of this study was to demonstrate the benefits of the transparent, evolutionary, modeling ensemble-generation for heterogeneous, multidimensional engineering data of a generally clustered nature, aggregated from multiple research groups/communities, within the broader scientific field. In multivariate analysis, with independent variables used to predict a response variable, the quality measures such as Akaike information criterion (AIC) or Bayesian information criterion (BIC) are helpful in selecting the best-performing models, while penalizing the number of predictor variables in the models, with the intention of preventing overfitting. Additionally, to ensure that the models are generalizable, cross-validation is done by randomly partitioning the data into training and test sets to verify if the generated model predicts well across the two sets [21]. However, the ultimate goal of materials data analytics is not to achieve a single, reasonable statistical description of data, but to provide domain science with new insights and multiple viable hypotheses allowed by the data, which suggest the least expensive ways of refining, or refuting, them.

Examination of the 8 to 13%wt. chromium steel data set revealed that the reported experimental data (specifically, the controlled parameters that are expected to contribute to variation in mechanical properties) are far from presenting independent variables randomly distributed in the input data space. In part, it represents the larger research community's historical proclivity of reporting mostly "good" data. It is also a reflection of the experiment design biases, either due to predetermined validated physics models or based on empirical "rules of thumb" commonly used by the practitioners.

Clustering analysis proved to be an effective MDA tool for mapping out the data for highly convoluted multidimensional alloying systems. The use of random forests, in this application, resulted in distinctly cluster-defined patterns, with nearly discontinuous predictive curves indicative of data insufficiency. The limited number of data points, in combination with nonlinearity associated with multiple and varying phases that are present in complex alloys, leads to poor performance of popular "one-size-fits-all" algorithms. It makes incorporation of domain knowledge almost imperative for data-driven predictive modeling of nonlinear relationships. However, the comprehensive knowledge acquisition process via experiments and physics-based simulations is usually expensive and time consuming.

To address these challenges, an evolutionary approach was developed, focused on generating ensembles of progressively better performing predictive models among the least complex ones. The response reproducibility errors extracted from the data were used to set the lower limit of primary differentiation between the models regarding the data fit. The secondary differentiation criterion was based on the effective number of available data points per one model parameter (preferably, a single element's concentration) with the aim of reducing model complexity, in conformity with the common statistical principle of parsimony: The generalizable model should be sufficiently complex to fit the data well, but it should not be more complex than the underlying relations it is designed to capture.

It was observed that UTS in C111 cluster is primarily governed by optimization of C/Fe ratio (and of the correlated N). The addition of molybdenum significantly increases the prominence of the feature, and copper appears to decrease it. Homogenization, particularly of the high-tungsten alloys within this compositional cluster, did not seem to affect the alloy tensile strength. Globally, as indicated by the optimized cluster models, UTS is not sensitive to carbon concentration below 0.13% (by weight) but increases with carbon in the range of 0.13 to 0.18% and, likely, beyond that as well (at least, in low-carbon steel, albeit with altered carbide morphology). Mechanical properties of dual-phase steel in the carbon range of 0.10 to 0.15% (i.e., near the solubility limit and with potential coarsening of carbonitrides) are controlled by martensite and ferrite fractions, martensite carbon content, grain sizes and strength of both phases. These microstructure features are particularly sensitive to variation in thermal history, which cause variation in ferritic-martensitic microstructure [28]. It is also likely that the addition of the mobile molybdenum atom may contribute to a coupled solute drag effect due to interaction of molybdenum and carbon near the grain boundaries [48].

C112 features a saturation of the manganese-induced strength increase, at about 0.70%, which matches the maximum limit specified in the American Society of Mechanical Engineers (ASME) Boiler and Pressure Vessel Code for austenite stabilization. (Higher concentration of manganese may promote cracking.) Globally, manganese concentration below this point positively correlated with UTS. Manganese is considered the most important, after carbon, as an addition to steel. It prevents the formation of embrittling grain-boundary cementite (Fe_3C) and plays a key role in controlling the overall precipitation process. The addition of manganese (Mn) facilitates the formation of carbides, particularly Mn_3C carbide which forms solid solutions with Fe_3C. Mn has similar atomic size to Fe. As such it can reduce the solubility of carbon in ferrite (α-Fe phase). For the alloys selected in this study, which are a low carbon (<0.30%) and low manganese (<1%) subset of the broader family of steels, the temperature at which austenite begins to decompose decreases with an increase in concentration of either carbon or manganese [24,34]. This extends the metastable austenitic (γ-Fe, face-centered cubic morphology) region, causing substantial grain refinement and increased dispersion hardening. Carbon may also slow down the temper reactions in metastable martensite, or increase temper embrittlement, unless carbon content is very low and trace-element impurities are minimal. Carbon may contribute to decreasing the difference in hardness between ferrite (body-centered cubic) and martensite (hexagonal close-packed, transition → distorted body-centered cubic) with increasing tensile strength. Sufficiently large amounts of manganese (as well as nickel) can make steel austenitic even at room temperature [26,50]. Additional contributing factors are: Manganese forms manganese sulfide morphologies dependent upon the state of oxidation of the steel, improving surface quality. However, manganese additions also reduce the number of cycles to failure under high strain conditions and increase the propensity for weld cracking because of hardenability issues [23]. The effect of manganese on hardenability is greater than that of any other alloying element [5] except molybdenum, especially at lower quenching temperatures [22].

Other observations are primarily related to cluster-localized patterns. C121 core (highest copper) strength is entirely controlled by copper (negatively correlated with ultimate tensile strength). Ultimate tensile strength in C21 core is positively correlated with chromium (within 8.3–8.8% cluster range) and has an intriguing, extra-low interstitial nickel (used for grain size control) [11] feature. Interestingly, chromium concentration and ultimate tensile strength tend to correlate increasingly negatively in clusters with much higher (10.5 to 12.9%) chromium concentration at which chromite-like passivation can occur [15].

The C122 set did not have enough data for meaningful comparisons. It does seem to fit the global trends detected elsewhere. Yet, overall, more data is required to meaningfully link the two of the C12 subclusters with the other clusters.

Cluster-based models can take advantage of compositional and structural similarities between the analyzed alloys. However, strong intracluster correlations between the key variables occasionally present a challenge. The importance of physics-based interpretation of the data-driven inferencing can be illustrated with a simple example. For M sets of data, each containing values for K independent input variables $\{x_k\}$ and a response variable R, there is a response function that provides an approximation of the response variable values R_m with some residual error e_m on each set:

$$\widetilde{R}_m(\{x_k\}) = R_m - e_m; m = 1 \ldots M. \tag{3}$$

The response function can be a combination of normalized linear, L and nonlinear, N variable transformations:

$$\widetilde{R}(\{x_k\}) = \sum_{k=1}^{q<K-1} s_k \sigma_k L(x_k) + s \times N(p_1, p_2; x_K), \tag{4}$$

where s is sensitivity, σ is variable's spread, and p_1 and p_2 are optimal parameter values. There are several ways to reduce the residual error: one way is to add another linear function,

$$\widetilde{R}'^{(\{x_k\})} = \sum_{k=1}^{q+1<K} s_k \sigma_k L(x_k) + s \times N'(p_1, p_2; x_K); \tag{5}$$

while another way is to add complexity to the nonlinear function,

$$\widetilde{R}''^{(\{x_k\})} = \sum_{k=1}^{q<K-1} s_k \sigma_k L(x_k) + s \times N''(p_1, p_2, p_3; x_K). \tag{6}$$

If a norm (over M data sets) of the residual error is comparable for the two approximations,

$$\begin{aligned} &\|s \times \{N''(p_1, p_2, p_3; x_K) - N'(p_1, p_2; x_K)\} - e_m\| \\ &\sim \|s_{q+1} \sigma_{q+1} L(x_{q+1}) - e_m\|, \end{aligned} \tag{7}$$

there are no conclusive statistical criteria to select one over the other, unless there is a physics-based justification. Incremental addition of new experimental data for the same set of variables may not overturn this reasoning. However, it is likely that some of these models may not survive the bootstrapping [29] used in ensemble approaches, or they may mutate [32] if evolutionary approaches are employed. Regardless of statistical techniques employed in model selection, such selections are not final and can be reversed with subsequent tests. There are no statistical criteria to confirm that even an apparently stable solution is global either. Only the critical test data and physics models that establish causality links can provide the valid criteria, evolving with the domain knowledge as discussed earlier.

5 | CONCLUSIONS

In multivariate analysis, with independent variables used to predict a response variable, quality measures such as AIC or BIC are helpful in selecting the best performing models, while penalizing the number of predictor variables in the models, with the intention of preventing overfitting. Alternatively, concurrent development of ensemble of competing models is a viable option in dealing with scarce, multidimensional data.

Examination of the 8 to 13% Cr steel data set revealed that the reported experimental data are far from presenting independent predictors randomly distributed in the input data space. The biases resulted in partially collapsed dimensions and in similarity-based subspace groupings within the alloy composition-and-process data space. Pattern search helped to identify the biases. The data clusters corresponding to the moderately frequent patterns and maximized IG, and further refined by using p-norm distance measures, match the alloy classification groups used by industry.

The limited number of data points per individual data cluster, in combination with nonlinearity associated with multiple and varying phases that are present in these steels, leads to poor performance of popular "one-size-fits-all" algorithms. The use of random forests in this application resulted in distinct cluster-defined patterns, with nearly discontinuous predictive curves indicative of data insufficiency.

Specific findings of interest to help domain scientists in designing new 9% Cr steel alloys are:

- Heat treatment and test temperature parameters were found to be the primary contributors to the steels' mechanical strength.
- If publicly available data are for the steels normalized at 20 to 50°C above the upper critical point (i.e., A_{c3}, as is common practice), there is no apparent correlation between the actual normalization temperature and the tensile strength.
- In the publicly available data, the reported tempering temperatures are bound within a narrow range for any standard 9% Cr steel subset. The effect of their moderate variation (between the groups of similar alloys) on tensile strength can be accounted for by using linear approximation.

- At test temperature of 600°C, manganese additions up to 0.70% (by weight) increase tensile strength regardless of the moderate composition variations (within the subset of similar alloy groups considered in this study), where it reaches apparent saturation.
- The chromium impact on tensile strength is highly nonlinear, with the correlation ranging from very strong positive at the levels below 9% (and involving cooperative effects with nickel) to relatively neutral at moderately higher levels to increasingly negative at the levels above 10.5%. However, more data are needed to define and quantify this effect.
- The contributions of carbon (specifically, at concentrations above 0.13%) and nitrogen (specifically, at concentrations above 0.02%) appear to be highly nonlinear and involve cooperative effects (mutual and with molybdenum). The stepwise increases in tensile strength induced by nitrogen are particularly steep, with the exact onset being dependent on the steel's composition.
- High copper concentration (~1%) within the range of compositions similar to P92 specifications (8.5 to 9.5% chromium, 1.5 to 2.0% tungsten, 0.3 to 0.6% molybdenum micro-alloyed with vanadium and niobium, and with controlled boron and nitrogen contents) has a very strong negative correlation with tensile strength.

ACKNOWLEDGMENTS

This project was supported in part by an appointment to the Internship/Research Participation Program at the National Energy Technology Laboratory, U.S. Department of Energy, administered by the Oak Ridge Institute for Science and Education. The work conducted by the co-authors at Case Western Reserve University was funded under the U.S. Department of Energy contract DE-FE0028685.

DISCLAIMER

This project was funded by the Department of Energy, National Energy Technology Laboratory, an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ORCID

Vyacheslav N. Romanov https://orcid.org/0000-0002-8850-3539

Laura S. Bruckman https://orcid.org/0000-0003-1271-1072

REFERENCES

1. F. Abe, *Strengthening mechanisms in creep of advanced ferritic power plant steels based on creep deformation analysis*, in *Advanced steels: The recent scenario in steel science and technology*, Y. Weng, H. Dong, and Y. Gan, Eds., Springer, Berlin, 2011, 409–422.
2. F. Abe, T.-U. Kern, and R. Viswanathan, Eds., *Creep-resistant steels*, Woodhead Publishing, Boca Raton, 2008.
3. A. Agrawal, P. D. Deshpande, A. Cecen, G. P. Basavarsu, A. N. Choudhary, and S. R. Kalidindi, Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters, Integr. Mater. Manuf. Innov. 3 (2014), 8.1–8.19.
4. ASM International Handbook Committee, *Heat treating, in ASTM handbook*, Vol 4, Book News, Portland, 1991.
5. AZoM, *The properties and effects of manganese as an alloying element*, AZO Materials, Sydney, 2016, available at https://www.azom.com/article.aspx?ArticleID=13027.
6. P. Berkhin, *A survey of clustering data mining techniques*, in *Grouping multidimensional data: Recent advances in clustering*, J. Kogan, C. Nicholas, and M. Teboulle, Eds., Springer, Berlin, 2006, 25–71.
7. K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, *When Is "Nearest Neighbor" meaningful?* in *Database theory - ICDT'99: Proceedings of the 7th international conference on database theory*, Catriel Beeri and Peter Buneman, Eds., Springer, Jerusalem, 1999, 217–235.
8. L. Breiman, Random forests, Mach. Learn. 45 (2001), 5–32.
9. L. S. Bruckman, N. R. Wheeler, J. Ma, E. Wang, C. K. Wang, I. Chou, J. Sun, R. H. French, Statistical and domain analytics applied to PV module lifetime and degradation science, IEEE Access 1 (2013), 384–403.
10. W. D. Callister and D. G. Rethwisch, *Materials science and engineering: An introduction*, 9th ed., John Wiley & Sons, New York, 2011.
11. J. Charles, J.-D. Mithieux, P.-O. Santacreu, and L. Peguet, The ferritic stainless family: the appropriate answer to nickel volatility? La Rev. Métall. 106 (2009), 124–139.
12. H. Cheng, X. Yan, J. Han, and C.-W. Hsu, *Discriminative frequent pattern analysis for effective classification*. IEEE 23rd International Conference on Data Engineering, April 15-20, 2007. Istanbul, IEEE, 2007.
13. K. K. Coleman and W. F. Newell Jr., P91 and beyond: Welding the new-generation Cr-Mo alloys for high-temperature service, Weld J. 86 (2007), 29–33.
14. R. D. Cook, Detection of influential observation in linear regression, Dent. Tech. 19 (1977), 15–18.
15. Pierre-Jean Cunat, *Alloying elements in stainless steel and other chromium-containing alloys*, Euro Inox. Paris, International Chromium Development Association, 2004.
16. H. Fujii, D. MacKay, and H. Bhadeshia, Bayesian neural network analysis of fatigue crack growth rate in nickel base superalloys, ISIJ Int. 36 (1996), 1373–1382.
17. B. Gautham, R. Kumar, S. Bothra, G. Mohapatra, N. Kulkarni, and K. Padmanabhan, *More Efficient ICME through materials informatics and process modeling*, in *Proceedings of the 1st world congress on integrated computational materials engineering*, J. Allison, P. Collins, and G. Spanos, Eds., John Wiley & Sons, Inc., Hoboken, 2011, 35–42.
18. G. Gutin, A. Yeo, and A. Zverovich, Traveling salesman should not be greedy: domination analysis of greedy-type heuristics for the TSP, Discrete Appl. Math. 117 (2002), 81–86.
19. M. Haenlein and A. M. Kaplan, A Beginner's Guide to Partial Least Squares Analysis, Understand. Stat. 3 (2004), 283–297.
20. Q. Hancheng, X. Bocai, L. Shangzheng, and W. Fagen, Fuzzy neural network modeling of material properties, J. Mater. Process. Technol. 122 (2002), 196–200.
21. T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed., Springer, New York, 2009.

22. D. H. Herring, Gear heat treatment: The influence of materials and geometry, Gear Technol. 21 (2004), 35–40.
23. D. H. Herring, *The influence of manganese in steel*. Industrial Heating, 2010, available at https://www.industrialheating.com/publications/3/editions/1201.
24. D. H. Herring, *Influence of alloying elements on austenite*, in *Industrial Heating,* 2015, available at https://www.industrialheating.com/publications/3/editions/1257.
25. Y. Hu, V. Y. Gunapati, P. Zhao, D. Gordon, N. R. Wheeler, M. A. Hossain, T. J. Peshek, L. S. Bruckman, G.-Q. Zhang, and R. H. French, A nonrelational data warehouse for the analysis of field and laboratory data from multiple heterogeneous photovoltaic test sites, IEEE J. Photovolt. 7 (2017), 230–236.
26. R. L. Klueh, P. J. Maziasz, and E. H. Lee, Manganese as an Austenite Stabilizer in Fe-Cr-Mn-C Steels, Mater. Sci. Eng. A 102 (1988), 115–124.
27. N. Krishnamurthy, et al. *Data analytics for alloy qualification,* NETL Tech. Rep. Ser. NETL-PUB-21550, Pittsburgh, U.S. Department of Energy, National Energy Technology Laboratory, 2017.
28. V. L. de la Concepcióna, H. N. Lorussoa, and H. G. Svoboda, Effect of carbon content on microstructure and mechanical properties of dual phase steels, Procedia Mater. Sci. 8 (2015), 1047–1056.
29. B. Lantz, *Machine learning with R*, 2nd ed., Packt Publishing, Birmingham, 2013.
30. A. Liaw and M. Wiener, Classification and regression by randomForest, R News 2 (2002), 18–22.
31. T. Lumley, *Leaps: Regression subset selection*, R package (based on Fortran code by Alan Miller) Version 3.0, 2017, available at https://CRAN.R-project.org/package=leaps.
32. D. J. C. MacKay, *Information theory, inference, and learning algorithms*, 4th ed. (version 7.2), Cambridge University Press, Cambridge, 2005.
33. F. Masuyama, History of power plants and progress in heat resistant steels, ISIJ Int. 41 (2001), 612–625.
34. H. Matsuda, R. Mizuno, Y. Funakawa, K. Seto, S. Matsuoka, and Y. Tanaka, Effects of auto-tempering behaviour of martensite on mechanical properties of ultra high strength steel sheets, J. Alloys Compd. 577S (2013), S661–S667.
35. NIMS, Creep data sheet, no. 13B. *Online*, National Institute for Materials Science, Tsukuba-shi, 1994.
36. NIMS, Creep data sheet, no. 19B. *Online*, National Institute for Materials Science, Tsukuba-shi, 1997a.
37. NIMS, Creep data sheet, no. 44. *Online*, National Institute for Materials Science, Tsukuba-shi, 1997b.
38. NIMS, Creep data sheet, no. 10B. *Online*, National Institute for Materials Science, Tsukuba-shi, 1998.
39. NIMS, Creep data sheet, no. 46A. *Online*, National Institute for Materials Science, Tsukuba-shi, 2005.
40. NIMS, Creep data sheet, no. 48A. *Onine*, National Institute for Materials Science, Tsukuba-shi, 2012.
41. NIMS, Creep data sheet, no. 51A. *Online*, National Institute for Materials Science, Tsukuba-shi, 2013a.
42. NIMS, Creep data sheet, no. 52A. *Online*, National Institute for Materials Science, Tsukuba-shi, 2013b.
43. NIMS, Creep data sheet, no. 43A. *Online*, National Institute for Materials Science, Tsukuba-shi, 2014.
44. Laxmi Parida, *Pattern discovery in bioinformatics: Theory and algorithms*, Chapman & Hall, New York, 2008.
45. G. Pison, A. Struyf, and P. Rousseeuw, Displaying a clustering with CLUSPLOT, Comput. Stat. Data Anal. 30 (1999), 381–392.
46. W. Revelle, *psych: Procedures for personality and psychological research*, R package, Version 1.8.6, 2017, available at https://personality-project.org/r/psych/.
47. Vyacheslav Romanov, *Combinatorial pattern search for information gain (c-IG)*, U.S. Department of Energy, National Energy Technology Laboratory, Pittsburgh 2016.
48. T. Schambron, A. Dehghan-Manshadi, L. Chen, T. Gooch, C. Killmore, and E. Pereloma, Effect of Mo on dynamic recrystallization and microstructure development of microalloyed steels, Metals Mater. Int. 23 (2017), 778–787.
49. S. Singh, H. Bhadeshia, D. MacKay, H. Carey, and I. Martin, Neural network analysis of steel plate processing, Ironmak. Steelmak. 25 (1998), 355–365.
50. Total Materia, *Influence of alloying elements on steel microstructure*, 2001, available at https://www.totalmateria.com/page.aspx?ID=CheckArticle&LN=NL&site=kts&NM=50.
51. A. K. Verma, R. H. French, and J. L. W. Carter, Physics-informed network models: A data science approach to metal design, Integr. Mater. Manuf. Innov. 6 (2017), 279–287.
52. R. Viswanathan, *Damage mechanisms and life assessment of high-temperature components*, ASM International, Metals Park, 1989.
53. R. Viswanathan and W. Bakker, Materials for ultrasupercritical coal power plants—Boiler materials: Part 1, J. Mater. Eng. Perform. 10 (2001a), 81–95.
54. R. Viswanathan and W. Bakker, Materials for ultrasupercritical coal power plants – Turbine materials: Part II, J. Mater. Eng. Perform. 10 (2001b), 96–101.
55. Y. F. Wen, C. Z. Cai, X. H. Liu, J. F. Pei, X. J. Zhu, and T. T. Xiao, Corrosion rate prediction of 3C steel under different seawater environment by using support vector regression, Corros. Sci. 51 (2009), 349–355.
56. S. Wold, K. Esbensen, and P. Geladi, Principal component analysis, Chemom. Intel. Lab. Syst. 2 (1987), 37–52.

How to cite this article: Romanov VN, Krishnamurthy N, Verma AK, et al. Materials data analytics for 9% Cr family steel. *Stat Anal Data Min: The ASA Data Sci Journal* 2019;1–12. https://doi.org/10.1002/sam.11406