Integrating Materials and Manufacturing Innovation
a SpringerOpen Journal

RESEARCH Open Access

Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters

Ankit Agrawal[1]*, Parijat D Deshpande[2], Ahmet Cecen[3], Gautham P Basavarsu[2], Alok N Choudhary[1] and Surya R Kalidindi[3,4]

*Correspondence: ankitag@eecs.northwestern.edu
[1] Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA
Full list of author information is available at the end of the article

Abstract

This paper describes the use of data analytics tools for predicting the fatigue strength of steels. Several physics-based as well as data-driven approaches have been used to arrive at correlations between various properties of alloys and their compositions and manufacturing process parameters. Data-driven approaches are of significant interest to materials engineers especially in arriving at extreme value properties such as cyclic fatigue, where the current state-of-the-art physics based models have severe limitations. Unfortunately, there is limited amount of documented success in these efforts. In this paper, we explore the application of different data science techniques, including feature selection and predictive modeling, to the fatigue properties of steels, utilizing the data from the National Institute for Material Science (NIMS) public domain database, and present a systematic end-to-end framework for exploring materials informatics. Results demonstrate that several advanced data analytics techniques such as neural networks, decision trees, and multivariate polynomial regression can achieve significant improvement in the prediction accuracy over previous efforts, with R^2 values over 0.97. The results have successfully demonstrated the utility of such data mining tools for ranking the composition and process parameters in the order of their potential for predicting fatigue strength of steels, and actually develop predictive models for the same.

Keywords: Materials informatics; Data mining; Regression analysis; Processing-property linkages

Background

Causal relations are foundational to all advances in sciences and technology. In advancing materials science and engineering, the practitioners of the field have traditionally relied largely on observations made from cleverly designed controlled experiments and sophisticated physics-based models to establish the desired causal relations, e.g., process-structure-property (PSP) linkages. In recent vision-setting documents [1,2], experts in materials science and engineering have identified data science and analytics as offering a third set of distinct tools (i.e., experiments, models, and data analytics making up the three foundational components of an integrated approach) for establishing the desired causal relations. Data science and analytics is expected to positively impact the ongoing materials development efforts by maximizing the accuracy and reliability of the core knowledge mined from large ensembles of (often heterogeneous and/or incomplete) datasets, and providing clear guidance for investment of future effort in as yet unexplored "white" spaces with the highest potential for success/benefit. In fact, data analytics techniques have already been successfully making inroads in the quest for new material design and discovery. The considerable interest and progress in the recent years has resulted in developing this new field and terming it as "Materials Informatics" [3,4]. The Materials Genome Initiative [2] places a large emphasis on data-driven approaches. Progress in this direction is supplemented by the availability of large amounts of experimental and simulation data, enhanced computing tools and advances in data analytics, which is expected to augment rather than compete with existing analytics methods.

State-of-the-art data analytics

Over the last few decades, our ability to generate data has far exceeded our ability to make sense of it in practically all scientific domains, and materials science is no exception. This has led to the emergence of the fourth paradigm of science [5], which is data-driven science and discovery, and is based on developing predictive and discovery-based data mining approaches on big data in a comprehensive manner. Fourth paradigm compliments the three traditional scientific advancement models of mathematical modeling, experiments, and computer simulations. Indeed, the most advanced techniques in this field come from computer science, high-performance computing, machine learning and data mining algorithms, and via applications in business domain, climate science, bioinformatics, astronomy/cosmology, intrusion detection, network analysis and many others, where predictive data mining has been effectively used for decision making with significant gains in the outcomes relevant to that domain. For example, companies like Amazon [6,7], Netflix [8], Google [9], Walmart [10] and Target [11] use predictive modeling for recommendations, personalized news, cost reductions, predicting demand and supply chain management at a massive scale providing lifts in sales and satisfaction. Scientists use predictive mining on big data to discover new stars/galaxies, predict hurricane paths, or predict structure of new materials. The accurate prediction of the path of hurricane Sandy illustrates an example of how the use and analysis of much larger data sets and algorithms can significantly improve accuracy.

The impressive advances made in the last two decades in both materials characterization equipment and in the physics-based multiscale materials modeling tools have ushered the BIG DATA age of materials science and engineering. With the advent of big data came the recognition that advanced statistics and modern data analytics would have to play an important role in the future workflows for the development of new or improved materials.

Several case studies illustrating the potential benefits of this emerging new field of Materials Informatics (MI) have already been published in literature. Rajan et al. [12] applied principal component analysis (PCA) on a database consisting of 600 compounds of high temperature superconductors to identify patterns and factors which govern this important property. They observed that the dataset clusters according to the average valency, a criterion which has been reported in literature to be of utmost importance for superconducting property. They concluded that informatics techniques allow one to investigate complex multivariate information in an accelerated and yet physically meaningful manner. Suh and Rajan [13] applied informatics tools on a dataset consisting of AB2N4 spinel nitrides to find the statistical interdependency of factors that may influence chemistry-structure-property relationships. Using partial least squares (PLS), they developed a quantitative structure-activity relationship (QSAR) relating bulk modulus of AB2N4 spinels with a variety of control parameters. They observed a strong agreement between the properties predicted based on ab-initio calculations and the ones based strictly on a data-driven approach. Nowers et al. [14] investigated property-structure-processing relations during interpenetrating polymer network (IPN) formation in epoxy/acrylate systems using an informatics approach. They concluded that material informatics is a very efficient tool which can be utilized for additional materials development. Gadzuric et al. [15] applied informatics tools on molten salt database to predict enthalpy (δH_{form}) and Gibbs free energy of formation (δG_{form}) of lanthanide halides. The results of the analysis indicated a high level of confidence for the predictions. George et al. [16] applied similar approach on a dataset consisting of binary and ternary metal hydrides to investigate the interrelationships among material properties of hydrides. They developed a relationship between entropy of a hydride and its molar volume which was in close agreement with the theoretical predictions. Singh et al. [17] developed a neural network model in which the yield and tensile strength of the steel was estimated as a function of some 108 variables, including the chemical composition and an array of rolling parameters. Fujii et al. [18] applied neural network approach for prediction of fatigue crack growth rate of nickel base superalloys. They modeled the rate as a function of 51 variables and demonstrated the ability of such methods for investigation of new phenomena in cases where the information cannot be accessed experimentally. Hancheng et al. [19] developed an adaptive fuzzy neural network model to predict strength based on compositions and microstructure. Runway stiffness prediction and evaluation models have also been developed using techinques such as genetic programming [20] and artificial neural networks [21]. Wen et al. [22] applied support vector regression (SVR) approach for prediction of corrosion rate of steels under different seawater environments. They concluded that SVR is a promising and practical method for real-time corrosion tracking of steels. Rao et al. [23] applied SVR for prediction of grindability index of coal and concluded that SVR is a promising technique and needs smaller data set for training the model than artificial neural network (ANN) techniques. To the best of our knowledge, there is only one prior study [24] dealing with fatigue strength prediction using the NIMS database (same data that we use in this work; details of the data provided later). It applied PCA on the data and subsequently performed partial least square regression (PLSR) on the different clusters identified by PCA for making predictions. Large R^2 values ranging between 0.88 and 0.94 were obtained for the resulting clusters.

Motivation

The prior MI case studies cited above have established the unequivocal potential of this emerging discipline in accelerating discovery and design of new/improved materials. However, there still does not exist a standardized set of protocols for exploring this approach in a systematic manner on many potential applications, and thus, establishing the composition-processing-structure-property relationships still remains an arduous task. A report published by NRC [1] stated that materials design has not been able to keep pace with the product design and development cycle and that insertion of new materials has become more infrequent. This poses a threat to industries such as automotive and aerospace, in which the synergy between product design, materials, and manufacturing is a competitive advantage.

In this paper, we embark on establishing a systematic framework for exploring MI, and illustrate it by establishing highly reliable causal linkages between process variables in a class of steels, their chemical compositions, and their fatigue strengths. The approach described in this work comprises of four main steps: (i) Preprocessing for consistent description of data, which can include things like filling in missing data wherever possible, with the help of appropriate domain knowledge; ii) Feature selection for attribute ranking and/or identifying the best subset of attributes for establishing a given linkage; (iii) Predictive modeling using multiple statistical and advanced data-driven strategies for the establishment of the desired linkages, (iv) Critical evaluation of the different informatics approaches using appropriate metrics and evaluation setting to avoid model over-fitting.

Accurate prediction of fatigue strength of steels is of particular significance in materials science to several advanced technology applications because of the extremely high cost (and time) of fatigue testing and often debilitating consequences of fatigue failures. Fatigue strength is the most important and basic data required for design and failure analysis of mechanical components. It is reported that fatigue accounts for over 90% of all mechanical failures of structural components [25]. Hence, fatigue life prediction is of utmost importance to both the materials science and mechanical engineering communities. The unavailability of recorded research in using a large number of heat treatment process parameters, composition to predict extreme value properties such as fatigue strength has led us to work on this problem. The complex interaction between the various input variables have baffled the conventional attempts in pursuing this work and advanced data analytics techniques may lead the path. The aim of this study is thus to fill some of the gaps encountered in this work and serve as a preliminary guide for prospective researchers in this field. The scope of this paper includes application of a range of machine learning and data analytics methods applied for the problem of fatigue strength prediction of steels using composition and processing parameters. A conference version of this paper with preliminary results appeared in Proceedings of the 2nd World Congress on Integrated Computational Materials Engineering (ICME 2013) [26].

Data

Fatigue Dataset for Steel from National Institute of Material Science (NIMS) MatNavi [27] was used in this work, which is one of the largest databases in the world with details on composition, mill product (upstream) features and subsequent processing (heat treatment) parameters. The database comprises carbon and low-alloy steels, carburizing steels and spring steels. Fatigue life data, which pertain to rotating bending fatigue tests at room temperature conditions, was the target property for which we aimed to construct predictive models in the current study. The features in the dataset can be categorized into the following:

- Chemical composition - %C, %Si, %Mn, %P, %S, %Ni, %Cr, %Cu, %Mo (all in wt. %)
- Upstream processing details - ingot size, reduction ratio, non-metallic inclusions

- Heat treatment conditions - temperature, time and other process conditions for normalizing, through-hardening, carburizing-quenching and tempering processes
- Mechanical properties - YS, UTS, %EL, %RA, hardness, Charpy impact value (J/cm2), fatigue strength.

The data used in this work has 437 instances/rows, 25 features/columns (composition and processing parameters), and 1 target property (fatigue strength). The 437 data instances include 371 carbon and low alloy steels, 48 carburizing steels, and 18 spring steels. This data pertains to various heats of each grade of steel and different processing conditions. The details of the 25 features and given in Table 1.

Methods

The overall proposed approach is illustrated in Figure 1. The raw data is preprocessed for consistency using domain knowledge. Ranking-based feature selection methods are also used to get an idea of the relative predictive potential of the attributes. Different regression-based predictive modeling methods are then used on the preprocessed and/or transformed data to construct models to predict the fatigue strength, given the composition and processing parameters. All constructed models are evaluated using Leave-One-Out Cross Validation with respect to various metrics for prediction accuracy. Below we present the details of each of the 4 stages.

Table 1 NIMS data features

Abbreviation	Details
C	% Carbon
Si	% Silicon
Mn	% Manganese
P	% Phosphorus
S	% Sulphur
Ni	% Nickel
Cr	% Chromium
Cu	% Copper
Mo	% Molybdenum
NT	Normalizing Temperature
THT	Through Hardening Temperature
THt	Through Hardening Time
THQCr	Cooling Rate for Through Hardening
CT	Carburization Temperature
Ct	Carburization Time
DT	Diffusion Temperature
Dt	Diffusion time
QmT	Quenching Media Temperature (for Carburization)
TT	Tempering Temperature
Tt	Tempering Time
TCr	Cooling Rate for Tempering
RedRatio	Reduction Ratio (Ingot to Bar)
dA	Area Proportion of Inclusions Deformed by Plastic Work
dB	Area Proportion of Inclusions Occurring in Discontinuous Array
dC	Area Proportion of Isolated Inclusions
Fatigue	**Rotating Bending Fatigue Strength (10^7 Cycles)**

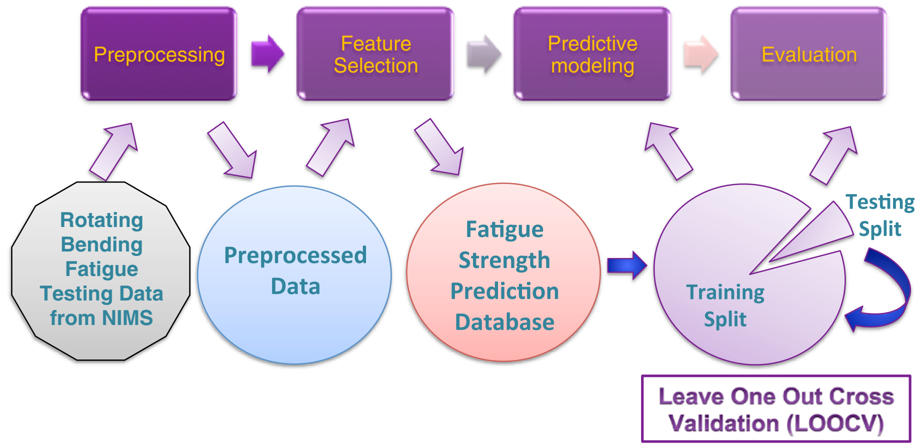
Preprocessing
Feature Selection
Predictive modeling
Evaluation
Rotating Bending Fatigue Testing Data from NIMS
Preprocessed Data
Fatigue Strength Prediction Database
Testing Split
Training Split
Leave One Out Cross Validation (LOOCV)

Figure 1 Block diagram. Schematic of overall proposed approach to fatigue strength prediction of steels with knowledge discovery and data mining.

Preprocessing

Understanding and cleaning the data for proper normalization is one of the most important steps for effective data mining. Appropriate preprocessing, therefore, becomes extremely crucial in any kind of predictive modeling, including that of fatigue strength. The dataset used in this study consists of multiple grades of steel and in some records, some of the heat treatment processing steps did not exist. In particular, different specimens are subjected to different heat treatment conditions. For example, some are normalized and tempered, some are through hardened and tempered, and others are carburized and tempered. There could be cases where normalization is done prior to carburization and tempering. In order to bring in a structure to the database, we have included all the key processes in the data-normalization, through hardening, carburization, quenching and tempering. For the cases where the actual process does not take place, we set the appropriate duration/time variable to zero with corresponding temperature as the austenization temperature or the average of rest of the data where the process exists. Setting the time to zero would essentially mean that no material transformation occurs. An artifact of our resulting data is that we are treating temperature and time as independent variables whereas they actually make sense only when seen together.

Feature selection

Information gain

This is an entropy-based metric that evaluates each attribute independently in terms of its worth by measuring the information gain with respect to the target variable:

$$IG(Class, Attrib) = H(Class) - H(Class|Attrib) \tag{1}$$

where $H(.)$ denotes the information entropy. The ranking generated by this method can be useful to get insights about the relative predictive potential of the input features.

SVD-PCA

Singular value decomposition is a matrix factorization defined as:

$$D = U \times S \times V \tag{2}$$

where, D is the data matrix such that every observation is represented by a row and each column is an explanatory variable, U is the matrix of left singular vectors, V is the matrix of right singular vectors and S is the diagonal matrix of singular values. In this case, $A = U \times S$ is a transformation of D where the data is represented by a new set of explanatory variables such that each variable is a known linear combination of the original explanatory parameters. The dimensions of A are also referred to as the Principal Components (PC) of the data.

Predictive modeling

We experimented with 12 predictive modeling techniques in this research study, which include the following:

Linear regression

Linear regression probably the oldest and most widely used predictive model, which commonly represents a regression that is linear in the unknown parameters used in the fit. The most common form of linear regression is least squares fitting [28]. Least squares fitting of lines and polynomials are both forms of linear regression.

Pace regression

It evaluates the effect of each feature and uses a clustering analysis to improve the statistical basis for estimating their contribution to overall regression. It can be shown that pace regression is optimal when the number of coefficients tends to infinity. We use a version of Pace Regression described in [29,30].

Regression post non-linear transformation of select input variables

A non-linear transformation of certain input variables can be done and the resulting data-set used for linear regression. In this study, the temperature variation effects on the diffusion equation are modelled according to the Arrhenius' empirical equation as $exp(-1/T)$ where T is measured in Kelvin.

Robust fit regression

The robust regression method [31] attempts to mitigate the shortcomings which are likely to affect ordinary linear regression due to the presence of outliers in the data or non-normal measurement errors.

Multivariate polynomial regression

Ordinary least squares (OLS) regression is governed by the equation:

$$\beta = (X'X)^{-1}X'Y \tag{3}$$

where β is the vector of regression coefficients, X is the design matrix and Y is the vector of responses at each data point. Multivariate Polynomial Regression (MPR) is a specialized instance of multivariate OLS regression that assumes that the relationship between regressors and the response variable can be explained with a standard polynomial. Standard polynomial here refers to a polynomial function that contains every polynomial term implied by a multinomial expansion of the regressors with a given degree (sometimes also referred to as a polynomial basis function). Polynomials of various degrees and number of variables are interrogated systematically to find the most suitable fit. There is a finite number of possible standard polynomials that can be interrogated due to the degree of freedom imposed by a particular dataset; the number of terms in the polynomial (consequently the number of coefficients) cannot exceed the number of data points.

Instance-based

This is a lazy predictive modeling technique which implements the K-nearest-neighbour (kNN) modeling. It uses normalized Euclidean distance to find the training instance closest to the given test instance, and predicts the same class as this training instance [32]. If multiple instances have the same (smallest) distance to the test instance, the first one found is used. It eliminates the need for building models and supports adding new instances to the training database dynamically. However, the zero training time comes at the expense of a large amount of time for testing since each test instance needs to be compared with all the data instances in the training data.

KStar

KStar [33] is another lazy instance-based modeling technique, i.e., the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that it uses an entropy-based distance function. The underlying technique used of summing probabilities over all possible paths is believed to contribute to its good overall performance over certain rule-based and instance-based methods. It also allows an integration of both symbolic and real valued attributes.

Decision table

Decision table is a rule-based modeling technique that typically constructs rules involving different combinations of attributes, which are selected using an attribute selection search method. It thus represents one of the simplest and most rudimentary ways of representing the output from a machine learning algorithm, showing a decision based on the values of a number of attributes of an instance. The number and specific types of attributes can vary to suit the needs of the task. Simple decision table majority classifier [34] has been shown to sometimes outperform state-of-the-art classifiers. Decision tables are easy for humans to understand, especially if the number of rules are not very large.

Support vector machines

SVMs are based on the Structural Risk Minimization (SRM) principle from statistical learning theory. A detailed description of SVMs and SRM is available in [35]. In their basic form, SVMs attempt to perform classification by constructing hyperplanes in a multidimensional space that separate the cases of different class labels. It supports both classification and regression tasks and can handle multiple continuous and nominal variables. Different types of kernels can be used in SVM models, including linear, polynomial, radial basis function (RBF), and sigmoid. Of these, the RBF kernel is the most recommended and popularly used, since it has finite response across the entire range of the real x-axis.

Artificial neural networks

ANNs are networks of interconnected artificial neurons, and are commonly used for non-linear statistical data modeling to model complex relationships between inputs and outputs. The network includes a hidden layer of multiple artificial neurons connected to the inputs and outputs with different edge weights. The internal edge weights are 'learnt' during the training process using techniques like back propagation. Several good descriptions of neural networks are available [36,37].

Reduced error pruning trees

A Reduced Error Pruning Tree (REPTree) [38] is an implementation of a fast decision tree learner. A decision tree consists of internal nodes denoting the different attributes and the branches denoting the possible values of the attributes, while the leaf nodes indicate the final predicted value of the target variable. REPTree builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning. In general, a decision tree construction begins at the top of the tree (root node) with all of the data. At each node, splits are made according to the information gain criterion, which splits the data into corresponding branches. Computation on remaining nodes continues in the same manner until one of the stopping criterions is met, which include maximum tree depth, minimum number of instances in a leaf node, minimum variance in a node.

M5 model trees

M5 Model Trees [39] are a reconstruction of Quinlan's M5 algorithm [40] for inducing trees of regression models, which combines a conventional decision tree with the option of linear regression functions at the nodes. It tries to partition the training data using a decision tree induction algorithm by trying to minimize the intra-subset variation in the class values down each branch, followed by back pruning and smoothing, which substantially increases prediction performance. It also uses the techniques used in CART [41] to effectively deal with enumerated attributes and missing values.

Evaluation

Traditional regression-based methods such as linear regression are typically evaluated by building the model (a linear equation in the case of linear regression) on the entire available data, and computing prediction errors on the same data. Although this approach works well in general for simple regression methods, it is nonetheless susceptible to over-fitting, and thus can give over-optimistic accuracy numbers. In particular, a data-driven model can, in principle learn every single instance of the dataset and thus result in 100% accuracy on the same data, but will most likely not be able to work well on unseen data. For this reason, advanced data-driven techniques that usually result in black-box models need to be evaluated on data that the model has not seen while training. A simple way to do this is to build the model only on random half of the data, and use the remaining half for evaluation. This is called the train-test split setting for model evaluation. Further, the training and testing halves can then also be swapped for another round of evaluation and the results combined to get predictions for all the instances in the dataset. This setting is called 2-fold cross validation, as the dataset is split into 2 parts. It can further be generalized to k-fold cross validation, where the dataset is randomly split into k parts. $k - 1$ parts are used to build the model and the remaining 1 part is used for testing. This process is repeated k times with different test splits, and the results combined to get preductions for the all the instances in the dataset using a model that did not see them while training. Cross validation is a standard evaluation setting to eliminate any chances of over-fitting.

Of course, k-fold cross validation necessitates builing k models, which may take a long time on large datasets.

Leave-one-out cross validation

We use leave-one-out cross validation (LOOCV) to evaluate and compare the prediction accuracy of the models. LOOCV is commonly used for this purpose particularly when the dataset is not very large. It is a special case of the more generic k-fold cross validation, with $k = N$, the number of instances in the dataset. The basic idea here is to estimate the accuracy of the predictive model on unseen input data it may encounter in the future, by withholding part of the data for training the model, and then testing the resulting model on the withheld data. In LOOCV, to predict the target attribute for each data instance, a separate predictive model is built using the remaining $N - 1$ data instances. The resulting N predictions can then be compared with the N actual values to calculate various quantitative metrics for accuracy. In this way, each of the N instances is tested using a model that did not see it while training, thereby maximally utilizing the available data for model building, and at the same time eliminating the chances of over-fitting of the models.

Evaluation metrics

Quantitative assessments of the degree to how close the models could predict the actual outputs are used to provide an evaluation of the models' predictive performances. A multi-criteria assessment with various goodness-of-fit statistics was performed using all the data vectors to test the accuracy of the trained models. The criteria that are employed for evaluation of models' predictive performances are the coefficient of correlation (R), explained variance (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error ($RMSE$), Standard Deviation of Error (SDE) between the actual and predicted values. The last three metrics were further normalized by the actual fatigue strength values to express them as error fractions. The definitions of these evaluation criteria are as follows:

$$R = \frac{\sum_{i=1}^{N}(y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2 \sum_{i=1}^{N}(\hat{y}_i - \overline{\hat{y}})^2}} \tag{4}$$

$$MAE = \overline{e} = \frac{1}{N}\sum_{N}|y - \hat{y}| \tag{5}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{N}(y - \hat{y})^2} \tag{6}$$

$$SDE = \sqrt{\frac{1}{N}\sum_{N}(|y - \hat{y}| - \overline{e})^2} \tag{7}$$

$$MAE_f = \overline{e_f} = \frac{1}{N}\sum_{N}\left|\frac{y - \hat{y}}{y}\right| \tag{8}$$

$$RMSE_f = \sqrt{\frac{1}{N}\sum_{N}\left(\frac{y - \hat{y}}{y}\right)^2} \tag{9}$$

$$SDE_f = \sqrt{\frac{1}{N}\sum_{N}\left(\left|\frac{y - \hat{y}}{y}\right| - \overline{e_f}\right)^2} \tag{10}$$

where y denotes the actual fatigue strength values (MPa), $\hat{y}$ denotes the predicted fatigue strength values (MPa), and N is the number of instances in the dataset.

The square of the coefficient of correlation, R^2, represents the variance explained by the model (higher the better), and is considered one of the most important metrics for evaluating the accuracy of regressive prediction models. Another useful metric is the fractional mean absolute error, MAE_f, which represents the error rate (lower the better).

Results and Discussion

We used the following statistical and data mining software for conducting the analysis reported in this paper: R [42], MATLAB [43], WEKA [44]. Default parameters were used unless stated otherwise.

The entire available data-set was assessed for visible clustering by employing K-means clustering technique. The cluster plot demonstrates inherent clustering in the available data, which agrees with the a priori knowledge of the dataset. The distinct clustering in the available data represents 3 clusters according to the grade of steels as depicted in Figure 2. These clusters however do not offer sufficient data-points to create individual meta-models for each cluster and hence, for all methods used, the entire data-set is used to develop predictive models.

The information gain metric was calculated for each of the 25 input attributes. For this purpose, the numerical fatigue strength was discretized into 5 equal-width bins as this only works for categorical target attributes. The relative predictive power of the 25 input attributes is shown in Figure 3. All the attributes were retained for building various predictive models as all of them were found to have significant predictive potential. We also looked at the correlation values of the 25 input features with the fatigue strength, as shown in Figure 4. Interestingly, TT (tempering temperature) shows up as the most important attribute for predicting fatigue strength in Figure 3. This is because the dataset

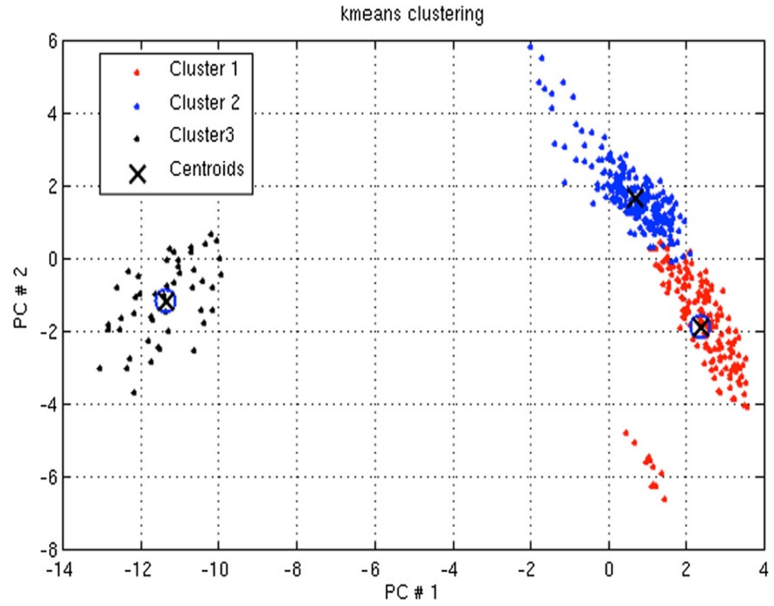

Figure 2 Cluster visualization. K-means clustering of the NIMS data using the top 2 principal components with K=3, clearly depiciting the 3 clusters corresponding to the 3 steel grades.

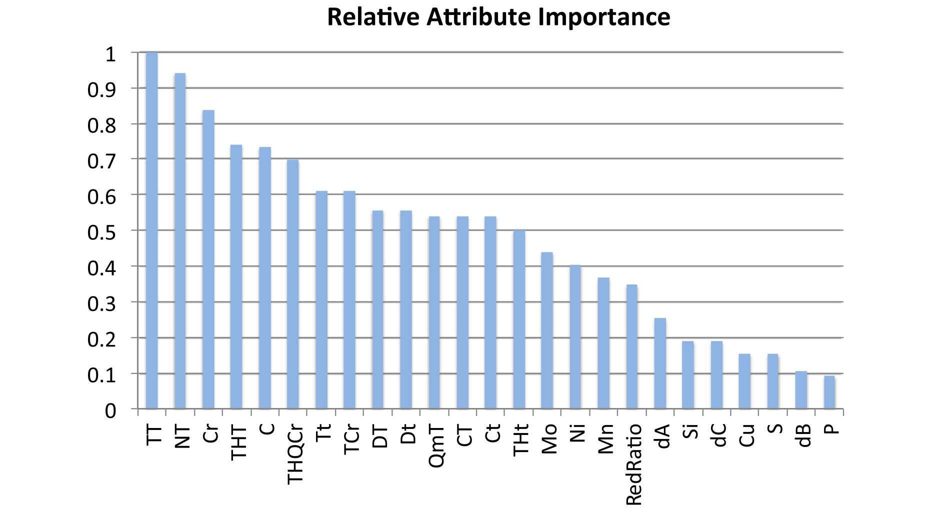

Figure 3 Information gain based feature ranking. The 25 input attributes along with their relative predictive power in terms of information gain.

consists of multiple grades of steel, each with a narrow yet significantly different range of TT. For example, TT for Through-hardened-tempered (without carburization) is around 400°C and that with carburization is around 200°C. These two situations will lead to a large difference in the fatigue strength. Thus, there is no surprise that the influence of TT seems high. However, the truth is that having a carburization step is what makes the key difference in the fatigue strength. Nevertheless, tempering will have a significant effect and this is reflected by the influence of tempering time in Figure 4. Figure 4 also identifies other variables such as carburizing temperature or through hardening temperature as important influencing factors. These are in line with expected results.

As mentioned before, we use Leave-One-Out Cross Validation (LOOCV) for model evaluation. Figure 5 and Table 2 present the LOOCV prediction accuracy of the 12 modeling techniques used in this work, in terms of the metrics discussed earlier. Clearly, many

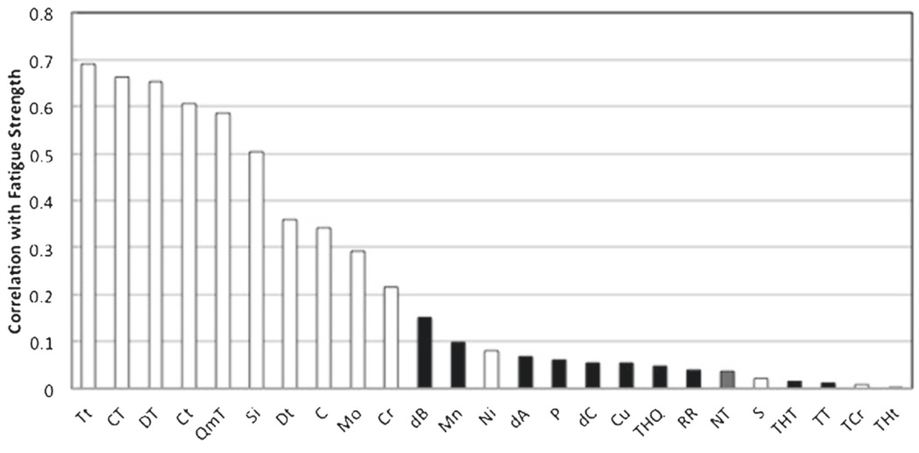

Figure 4 Correlation based feature ranking. The 25 input attributes along with their correlation values with fatigue strength. White (black) bars indicate positive (negative) influence.

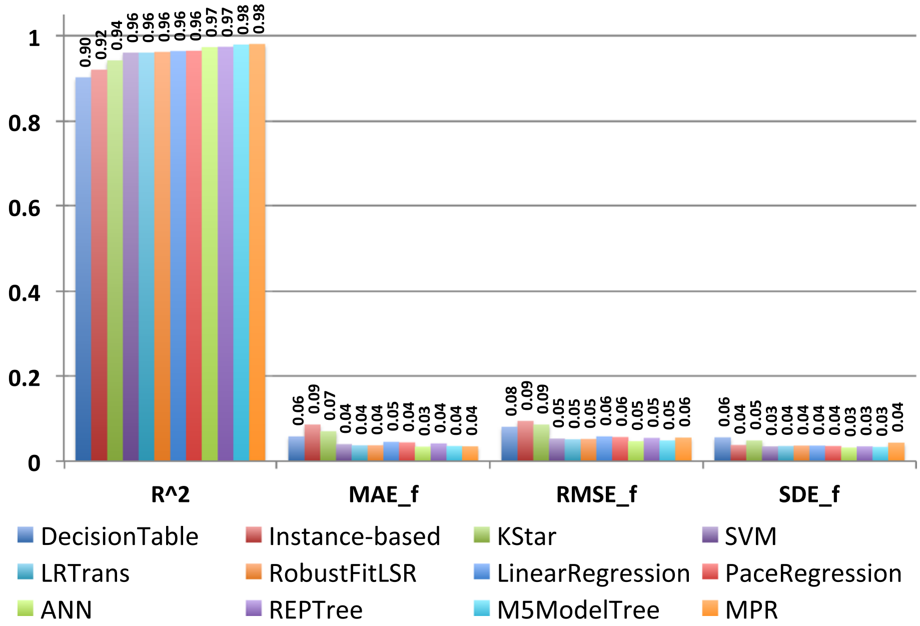

Figure 5 Results comparison. R^2, MAE_f, $RMSE_f$, and SDE_f for 12 predictive modeling techniques. Better fitting accuracy is characterized by higher values of R^2 and lower values of MAE_f, $RMSE_f$, and SDE_f.

of the employed data analytics techniques are able to achieve a high predictive accuracy, with R^2 values ~0.98, and error rate <4%. This is extremely encouraging since it significantly outperforms the only prior study on fatigue strength prediction [24], which reported R^2 values of <0.94. It is well known in the field of predictive data analytics that it becomes progressively more and more challenging to increase the accuracy of prediction beyond a certain point. To put it in context of this study, an increase in R^2 from 0.94 to 0.98 should not be viewed as simply an improvement of 0.04 or 4%. Rather, it should be seen with respect to the available scope for improvement of 0.06 (= 1.00 - 0.94). Thus, a more reasonable evaluation of the improvement accomplished by the current study over prior work would be about 66% (0.04/0.06), which is very significant.

Figure 6 presents the scatter plots for the 12 techniques. As can be seen from these plots, the three grades of steels are well separated in most of the techniques, and different

Table 2 Results comparison

Method	*R*	R^2	*MAE*	*RMSE*	*SDE*	MAE_f	$RMSE_f$	SDE_f
DecisionTable	0.9494	0.9014	34.8762	58.5932	47.1371	0.0584	0.0806	0.0557
IBk	0.9589	0.9195	46.0320	53.2749	26.8499	0.0859	0.0940	0.0382
KStar	0.9702	0.9413	36.9986	45.3779	26.3029	0.0706	0.0857	0.0487
SVM	0.9795	0.9594	24.2820	37.6250	28.7736	0.0400	0.0530	0.0349
LRTrans	0.9796	0.9596	22.3336	37.4748	30.1272	0.0370	0.0514	0.0357
RobustFitLSR	0.9804	0.9612	22.2152	37.2188	29.8960	0.0369	0.0520	0.0366
LinearRegression	0.9815	0.9633	25.6006	35.7168	24.9345	0.0456	0.0581	0.0360
PaceRegression	0.9816	0.9635	25.0302	35.5733	25.3065	0.0439	0.0565	0.0356
ANN	0.9861	0.9724	19.7778	31.0545	23.9695	0.0343	0.0470	0.0322
REPTree	0.9862	0.9726	22.5671	30.9401	21.1907	0.0414	0.0542	0.0349
M5ModelTree	0.9890	0.9781	19.3760	27.6065	19.6870	0.0353	0.0484	0.0332
MPR	0.9900	0.9801	18.5529	26.4378	18.8563	0.0350	0.0556	0.0432

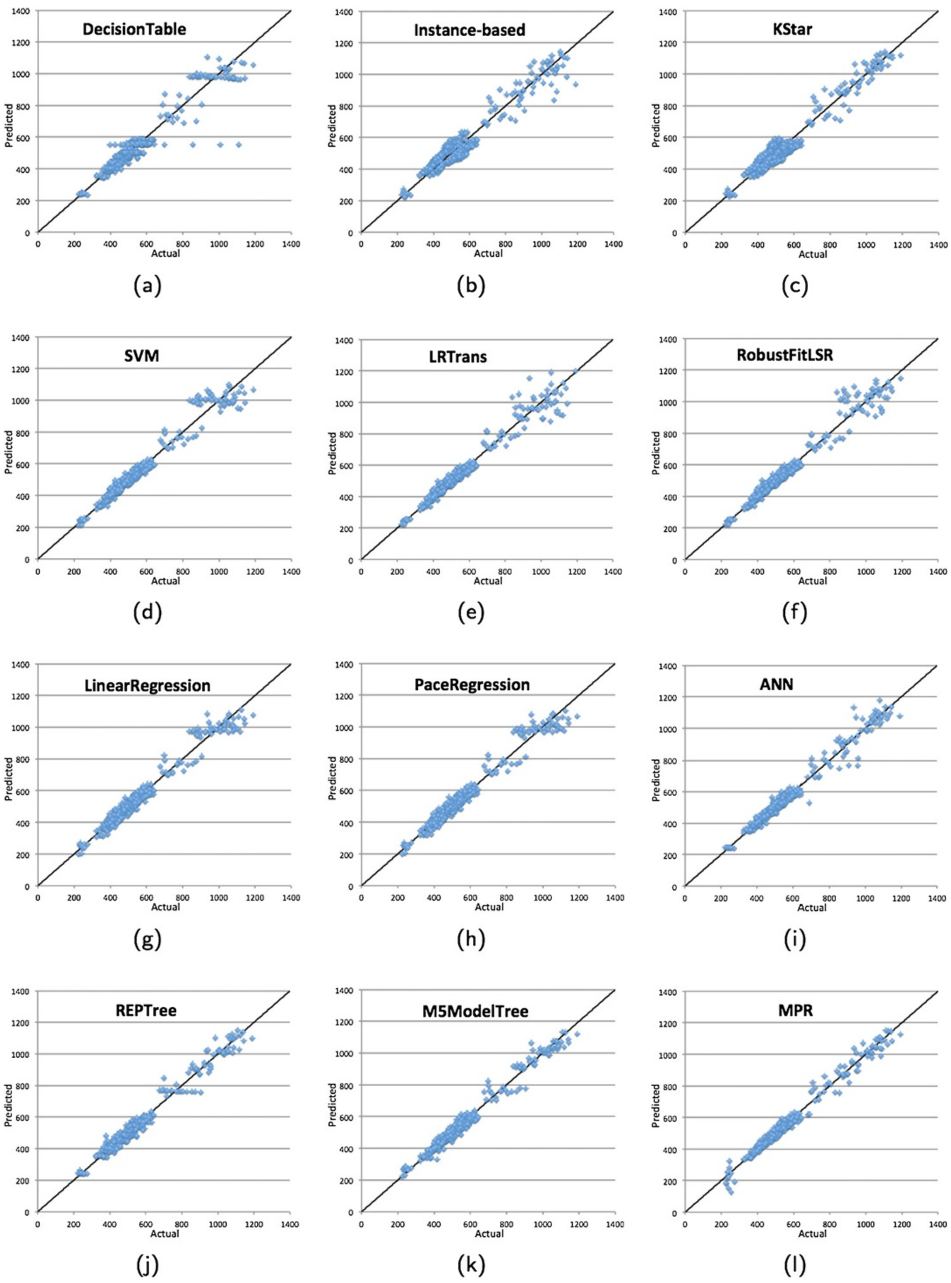

Figure 6 Scatter plots. Scatter plots for the 12 modeling techniques. X-axis and Y-axis denote the actual and predicted fatigue strength (in MPa) respectively. **a)** Decision Table; **b)** Instance-based; **c)** KStar; **d)** Support Vector Machines; **e)** Regression with Transformed Terms; **f)** RobustFit Regression; **g)** Linear Regression; **h)** Pace Regression; **i)** Artificial Neural Networks; **j)** Reduced Error Pruning Trees; **k)** M5 Model Trees; **l)** Multivariate Polynomial Regression.

techniques tend to perform better for different grades. Figure 7 shows the histograms of the error fractions for each of the techniques, to visualize the spread in the prediction errors. As expected, the spread in the error reduces as R^2 values improve. However, a point to be noted is that even though R^2 value is high, there are regions of data clusters where data fit is not sufficiently high and this is reflected in the nature of distribution of

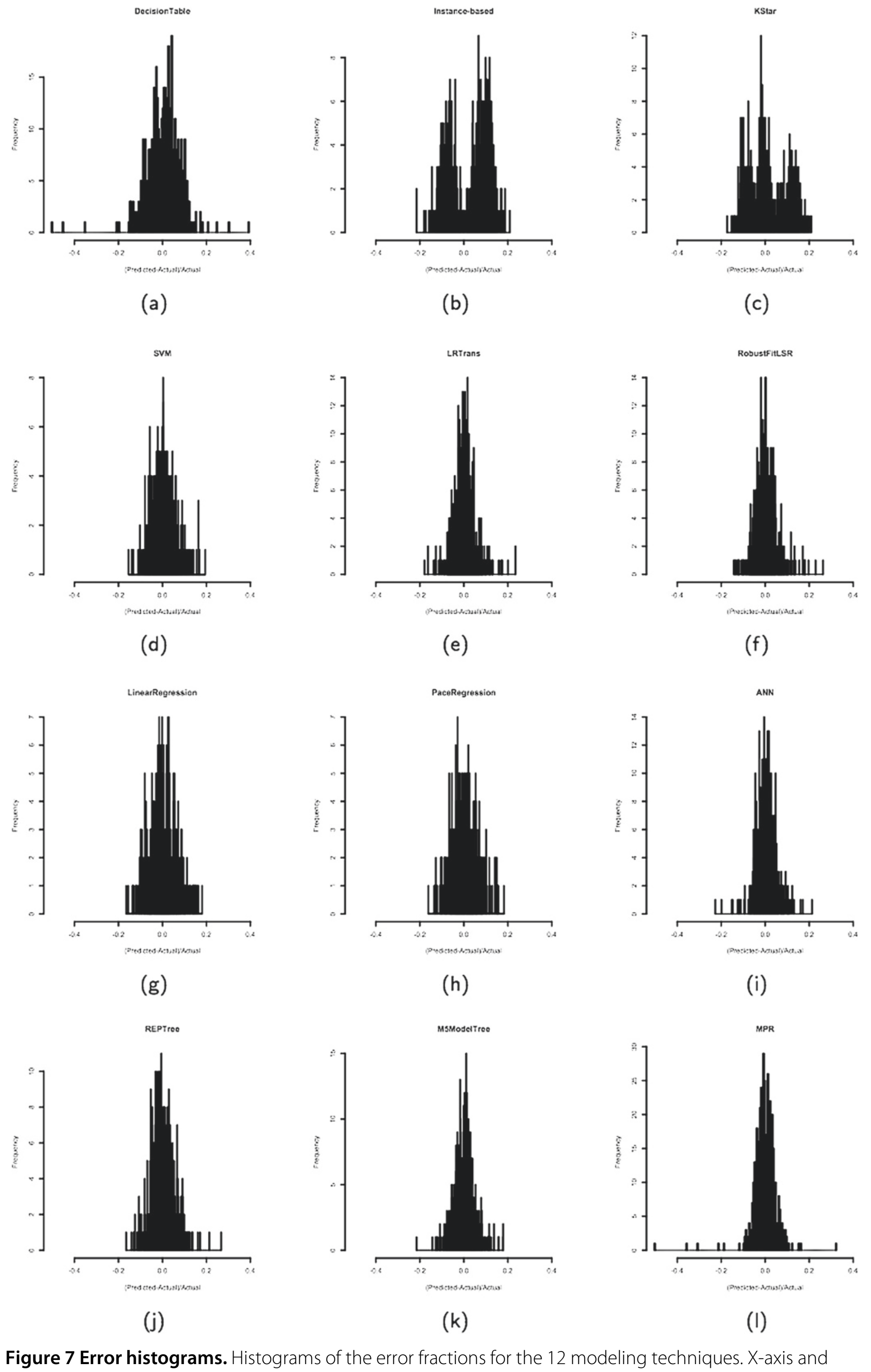

Figure 7 Error histograms. Histograms of the error fractions for the 12 modeling techniques. X-axis and Y-axis denote the error fraction bins and the frequency respectively. **a)** Decision Table; **b)** Instance-based; **c)** KStar; **d)** Support Vector Machines; **e)** Regression with Transformed Terms; **f)** RobustFit Regression; **g)** Linear Regression; **h)** Pace Regression; **i)** Artificial Neural Networks; **j)** Reduced Error Pruning Trees; **k)** M5 Model Trees; **l)** Multivariate Polynomial Regression.

errors. Thus, the methods that result in bimodal distribution of errors or the ones with significant peaks in higher error regions are not so good even though their reported R^2 may be reasonable.

The general opinion in data mining community about predictive modeling is that it is more helpful to know about a set of well performing techniques for a given problem rather than identifying a single winner. We have thus examined 12 different techniques for predictive modeling of fatigue strength, and it is shown that a number of different approaches produce highly reliable linkages. In particular, neural networks, decision trees, and multivariate polynomial regression were found to achieve a high R^2 value of greater than 0.97, which is significantly better than what has been previously reported in the literature. This is also shown by narrow distribution of errors. It is very encouraging to see that despite the limited amount of data available in this dataset, the data-driven analytics models were able to achieve a reasonably high degree of accuracy.

Although the main contribution of this paper is to present an end-to-end framework for exploring predictive materials informatics, and its application on NIMS data is a specific example of the application of the framework, it is nonetheless important for completeness to discuss some of the limitations of the proposed framework's specific application on the NIMS dataset. Since the data used in this study is very small compared to the typical amounts of data used in data mining studies in other domains, we believe that the obtained high accuracy is but an encouragement to use more data (possibly combine data from heterogenous sources) to further validate the results and/or making the model more robust. One possibility would be to add structure information to the data, which may ease the application of the developed models to actionable materials design, as structure information is what is primary responsible for the resulting properties. Another limitation of the NIMS data used in this study is the significantly different number of data instances corresponding to the different types of steels. Hence the predictive models, which are developed over the entire data may not be highly accurate for all steel types, which is also evident from the scatter plots. Possible approaches to deal with this imbalanced data distribution are discussed in the next section.

Conclusions

Materials Informatics, steeped in modern data analytics and advanced statistics, is fast emerging as a key enabler for accelerated and cost-effective development of new and improved materials targeted for advanced technologies. One of the core challenges addressed by this nascent field is the successful mining of highly reliable, quantitative, linkages capturing the salient connections between chemical compositions, processing history, and the final properties of the produced material. These linkages can provide valuable guidance to future effort investment with tremendous potential for cost-savings.

In this paper, we have tried to critically explore the viability of extracting such linkages from open access databases. As a specific example, we have focused on extracting reliable linkages between chemical compositions, processing history, and fatigue strength of a class of steels using data available from the open access materials database hosted by Japan's National Institute for Materials Science (NIMS). In this study, a range of advanced data analytics techniques, typically involving a combination of feature selection and regression methods, have been successfully employed and critically evaluated for the problem of fatigue strength prediction of different grades of steels.

There are several directions of future work that can stem from the present research. From the data analytics point of view, ensemble predictive modeling can be used to combine the results from multiple predictive models using same and/or different techniques built on different random subsets of the training data. Apart from this, since the scatter plots showed that different techniques can work well for different steel types, we can also try hierarchical predictive modeling, where we first try to classify the input test instance into one of the three grades of steel, and subsequently use the appropriate model(s) for that grade to predict the fatigue strength. From the materials science point of view, it would be good to explore the use of additional input features that may be easily measurable like some mechanical properties. Methods for using grouped variables representing each processing step could be of significant utility as well. It would also be extremely valuable to add structure information to the data, which may be able to give more actionable insights for materials design, as structure is very closely linked to property. Finally, the analytics framework developed and used in this paper can be used for building prediction models for other desired target properties, such as % Elongation, the data for which is also available in the NIMS dataset.

Availability of supporting data

The raw data used in this study was obtained from the publicly available NIMS MatNavi dataset, and was preprocessed as described in the paper. The preprocessed data is available as Additional file 1 of this paper.

Additional file

Additional file 1: Preprocessed NIMS MatNavi Data used in this study.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AA, PDD, AC applied one or more of the data analytics techniques used in this work. ANC provided data mining guidance and expertise; GPB and SRK provided materials science guidance and expertise. AA led the writing of the paper with inputs from all other authors. All authors read and approved the final submitted manuscript.

Acknowledgements
Tbe authors are grateful to NIMS to make the raw data on fatigue steel strength publicly available. This work is supported in part by the following grants: NSF awards CCF-0833131, CNS-0830927, IIS-0905205, CCF-0938000, CCF-1029166, ACI-1144061, and IIS-1343639; DOE awards DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, DESC0005340, and DESC0007456; AFOSR award FA9550-12-1-0458.

Author details
[1]Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA. [2]Tata Research Development and Design Centre, Tata Consultancy Services, Pune, Maharashtra, India. [3]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA. [4]Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

Received: 24 December 2013 Accepted: 17 March 2014
Published: 3 April 2014

References

1. Committee on Integrated Computational Materials Engineering N. R. C. (2008) Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security. http://www.nap.edu/openbook.php?record_id=12199.

2. National Science and Technology Council (2011) Materials genome initiative for global competitiveness. Technical report, National Science and Technology Council. http://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf.
3. Kalidindi SR, Niezgoda SR, Salem AA (2011) Microstructure informatics using higher-order statistics and efficient data-mining protocols. JOM - J Minerals, Met Mater Soc 63(4): 40–41
4. Rajan K (2005) Materials informatics. Materials Today 8(10): 38–45
5. Hey T, Tansley S, Tolle K (2009) The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, 1st edition. ISBN: 0982544200, URL: http://research.microsoft.com/en-us/collaboration/fourthparadigm/.
6. Linden G, Smith B, York J (2003) Amazon.com recommendations: item-to-item collaborative filtering. Internet Comput IEEE 7(1): 76–80
7. Mobasher B (2007) Data mining for web personalization. In: Brusilovsky P, Kobsa A, Nejdl W (eds) The adaptive web. Lecture Notes in Computer Science, vol. 4321. Springer-Verlag, Berlin, Heidelberg, pp 90–135
8. Zhou Y, Wilkinson D, Schreiber R, Pan R (2008) Large-scale parallel collaborative filtering for the netflix prize. In: Proceedings of the 4th International Conference on Algorithmic Aspects in Information and Management. AAIM '08, Springer, Berlin, Heidelberg, pp 337–348
9. Das AS, Datar M, Garg A, Rajaram S (2007) Google news personalization: Scalable online collaborative filtering. In: Proceedings of the 16th International Conference on World Wide Web. WWW '07, ACM, New York, NY, USA, pp 271–280
10. URL: Walmart is making big data part of its DNA. Bigdata startups, 2013, http://www.bigdata-startups.com/BigData-startup/walmart-making-big-data-part-dna/.
11. King M (2012) URL: Data Mining the TARGET way. http://www.slideshare.net/ipullrank/datamining-the-target-way.
12. Rajan K, Suh C, Mendez P (2009) Principal component analysis and dimensional analysis as materials informatics tools to reduce dimensionality in materials science and engineering. Stat Anal Data Min 1: 361–371
13. Suh C, Rajan K (2005) Virtual screening and qsar formulations for crystal chemistry. QSAR & Comb. Sci 24(1): 114–119
14. Nowers JR, Broderick SR, Rajan K, Narasimhan B (2007) Combinatorial methods and informatics provide insight into physical properties and structure relationships during ipn formation. Macromol Rapid Commun 28: 972–976
15. Gadzuric S, Suh C, Gaune-Escard M, Rajan K (2006) Extracting information from the molten salt database. Metallogr Mater Trans A 37(12): 3411–3414
16. George L, Hrubiak R, Rajan K, Saxena SK (2009) Principal component analysis on properties of binary and ternary hydrides and a comparison of metal versus metal hydride properties. J Alloys Compounds 478(1–2): 731–735
17. Singh S, Bhadeshia H, MacKay D, Carey H, Martin I (1998) Neural network analysis of steel plate processing. Iron-mak Steelmak 25: 355–365
18. Fujii H, MacKay D, Bhadeshia H (1996) Bayesian neural network analysis of fatigue crack growth rate in nickel base superalloys. ISIJ INT 36: 1373–1382
19. Hancheng Q, Bocai X, Shangzheng L, Fagen W (2002) Fuzzy neural network modeling of material properties. J Mater Process Technol 122(2–3): 196–200
20. Gopalakrishnan K, Ceylan H, Kim S, Khaitan SK (2010) Natural selection of asphalt mix stiffness predictive models with genetic programming. ANNIE Int Eng Syst Artif Neural Netw 20: 10
21. Gopalakrishnan K, Manik A, Khaitan SK (2006) Runway stiffness evaluation using an artificial neural systems approach. Int J Electrical Comput Eng 1(7): 496–502
22. Wen YF, Cai CZ, Liu XH, Pei JF, Zhu XJ, Xiao TT (2009) Corrosion rate prediction of 3c steel under different seawater environment by using support vector regression. Corrosion Sci 51(2): 349–355
23. Rao BV, Gopalakrishna SJ (2009) Hardgrove grindability index prediction using support vector regression. Int J Miner Process 91(1–2): 55–59
24. Gautham BP, Kumar R, Bothra S, Mohapatra G, Kulkarni N, Padmanabhan KA (2011) More Efficient ICME through Materials Informatics and Process Modeling. In: Proceedings of the 1st World Congress on Integrated Computational Materials Engineering (ICME) (eds J. Allison, P. Collins and G. Spanos). John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9781118147726.ch5
25. Dieter GE (1986) Mechanical Metallurgy. Mc Graw-Hill Book Co. 3rd edition, ISBN: 0-07-016893-8 26. Deshpande PD, Gautham BP, Cecen A, Kalidindi S, Agrawal A, Choudhary (2013) Application of Statistical and Machine Learning Techniques for Correlating Properties to Composition and Manufacturing Processes of Steels. In: 2nd World Congress on Integrated Computational Materials Engineering. John Wiley & Sons, Inc., pp 155-160. ISBN: 9781118767061
26. Deshpande PD, Gautham BP, Cecen A, Kalidindi S, Agrawal A, Choudhary (2013) Application of Statistical and Machine Learning Techniques for Correlating Properties to Composition and Manufacturing Processes of Steels. In: 2nd World Congress on Integrated Computational Materials Engineering. John Wiley & Sons, Inc., pp 155-160. ISBN: 9781118767061
27. URL: National Institute of Materials Science. http://smds.nims.go.jp/fatigue/index_en.html.
28. Weher E, Allen EL (1977) An introduction to linear regression and correlation. (a series of books in psychology.) w. h. freeman and comp., San Francisco 1976. 213 s., tafelanh., s 7.00. Biom J 19(1): 83–84
29. Wang Y (2000) A new approach to fitting linear models in high dimensional spaces. Technical report, University of Waikato, URL: http://books.google.com/books?id=Z0OntgAACAAJ.
30. Wang Y, Witten IH (2002) Modeling for optimal probability prediction. In: Proceedings of the Nineteenth International Conference on Machine Learning. Morgan Kaufmann Publishers, pp 650–657
31. URL: Robust Fit Regression, Mathworks. http://www.mathworks.in/help/stats/robustfit.html.
32. Aha DW, Kibler D (1991) Instance-based learning algorithms. Machine Learning Vol. 6. Kluwer Academic Publishers, Boston, pp 37–66
33. Cleary JG, Trigg LE (1995) K*: An instance-based learner using an entropic distance measure. In: Proceedings of the 12th International Conference on Machine Learning. Morgan Kaufmann Publishers, pp 108–114
34. Kohavi R (1995) The power of decision tables. In: Proceedings of the 8th European Conference on Machine Learning. ECML '95, Springer-Verlag, London, UK, pp 174–189

35. Vapnik VN (1995) The nature of statistical learning theory. Information Science and Statistics Series. Springer-Verlag New York, Inc., New York, NY, USA. ISBN: 0-387-94559-8
36. Bishop C (1995) Neural Networks for Pattern Recognition. 1st edition. Oxford University Press, USA. ISBN: 0198538642
37. Fausett L (1994) Fundamentals of Neural Networks. 1st edition. Prentice Hall, Pearson, New York. ISBN: 0133341860
38. Witten IH, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques. The Morgan Kaufmann Series in Data Management Systems, 2nd edition. Morgan Kaufmann Publishers. ISBN: 0120884070
39. Wang Y, Witten IH (1997) Induction of model trees for predicting continuous classes. In: Proc European Conference on Machine Learning Poster Papers, Prague, Czech Republic, pp 128–137
40. Quinlan JR (1992) Learning with continuous classes. In: 5th Australian Joint Conference on Artificial Intelligence. World Scientific, pp 343–348
41. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA
42. R Development Core Team (2011) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.
43. MATLAB (2010) Version 7.10.0 (R2010a). The MathWorks Inc., Natick, Massachusetts
44. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: An update. SIGKDD Explorations Newsletter 11(1): 10-18. doi:10.1145/1656274.1656278, URL: http://doi.acm.org/10.1145/1656274.1656278.

doi:10.1186/2193-9772-3-8
Cite this article as: Agrawal *et al.*: **Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters.** *Integrating Materials and Manufacturing Innovation* 2014 **3**:8.