npj | Materials Degradation

www.nature.com/npjmatdeg

ARTICLE OPEN

Data analytics approach to predict high-temperature cyclic oxidation kinetics of NiCr-based Alloys

Jian Peng[1], Rishi Pillai[1], Marie Romedenne[1], Bruce A. Pint[1], Govindarajan Muralidharan[1], J. Allen Haynes[1] and Dongwon Shin[1]✉

Although of practical importance, there is no established modeling framework to accurately predict high-temperature cyclic oxidation kinetics of multi-component alloys due to the inherent complexity. We present a data analytics approach to predict the oxidation rate constant of NiCr-based alloys as a function of composition and temperature with a highly consistent and well-curated experimental dataset. Two characteristic oxidation models, i.e., a simple parabolic law and a statistical cyclic oxidation model, have been chosen to numerically represent the high-temperature oxidation kinetics of commercial and model NiCr-based alloys. We have successfully trained machine learning (ML) models using highly ranked key input features identified by correlation analysis to accurately predict experimental parabolic rate constants (k_p). This study demonstrates the potential of ML approaches to predict oxidation kinetics of alloys over wide composition and temperature ranges. This approach can also serve as a basis for introducing more physically meaningful ML input features to predict the comprehensive cyclic oxidation behavior of multi-component high-temperature alloys with proper constraints based on the known underlying mechanisms.

npj Materials Degradation (2021)5:41 ; https://doi.org/10.1038/s41529-021-00188-z

INTRODUCTION

Ni-based alloys used at high temperatures are required to have both good mechanical properties and oxidation resistance. These alloys obtain their oxidation resistance by forming adherent protective chromia and/or alumina scales[1,2]. Currently, the evaluation of alloys' high-temperature oxidation still heavily relies on experimental investigations, which is costly and time-consuming. Hence, developing the capability to predict the high-temperature oxidation kinetics of multi-component alloys is highly desirable and of great interest in many applications involving extreme environments. Although computational approaches using conventional analytical and physics-based simulations have made significant advancements in investigating high-temperature oxidation, they are usually only applicable for a given length/time scale[3–6] or specific oxidation-related phenomena[7–11].

On the other hand, improvements in the machine learning (ML) approach have propelled the use of data analytics to assist the discovery of materials and the prediction of properties[12–19]. It possesses many advantages in handling complex multi-component alloys and offers the potential to extract insights from complex experiments or synthetic datasets generated from physics-based simulations. Thus, modern data analytics approaches can be considered an alternate and/or complementary tool to accelerate the design and development of materials with reduced cost and risk[12–14,20–22].

Recently, data analytics approaches have been successfully applied to predict the mechanical properties of multi-component high-temperature alloys[18,22–27]. While high-temperature oxidation also has scientific and practical importance, to the best of the authors' knowledge, very limited effort has been made to predict the oxidation kinetics of complex multi-component alloys by ML. A preliminary attempt has been made by Bhattacharya et al.[28] to build ML models to predict the high-temperature oxidation kinetics, described by the parabolic rate constant (k_p), of Ti alloys between 550 and 750 °C. From a statistical perspective, a good agreement between the predicted and experimental k_p was achieved. However, their dataset was collected from multiple literature sources; thus, the data may exhibit mixed oxidation mechanisms due to the difference in experimental protocols (i.e., atmospheres and isothermal or cyclic exposures)[26]. Moreover, their work focused only on building predictive models without analyzing the relationship between input features and k_p.

Herein, we present a data analytics framework consisting of correlation analysis of a consistently measured and well-curated oxidation dataset followed by ML to predict the cyclic oxidation kinetics of NiCr-based alloys, as illustrated by the workflow shown in Fig. 1. The oxidation kinetics of studied alloys was numerically represented by k_p from two representative oxidation models. Correlation analysis was performed using two algorithms (Pearson's correlation coefficient (PCC)[29] and maximal information coefficient (MIC)[30] to quantitatively evaluate the strength of correlation between input features (i.e., alloying compositions/oxidation temperature) and k_p, and hence rank the input features. The performance of five widely used ML models, i.e., linear regression (LR)[31], Bayesian ridge regression (BR)[32,33], *k*-nearest neighbor regression (NN)[34], random forest regression (RF)[35], and support vector machines (SVM) regression[36], in accurately predicting the k_p as a function of the number of top-ranking input features is evaluated and discussed in this paper.

RESULTS AND DISCUSSION

Numerical representation of cyclic oxidation data

Extensive experimental investigations of the cyclic oxidation kinetics of 25 model NiCr alloys and four commercial Ni-based alloys (i.e., Nimonic 80 A, Nimonic 90, René 41, and Haynes 282) have been performed in both dry air and wet air (air + 10% water vapor) atmospheres at Oak Ridge National Laboratory (ORNL). Figure 2 shows examples of the cyclic oxidation behavior of select NiCr alloys[37,38] in the dataset. The alloy composition ranges, experimental conditions, and details of the dataset are

[1]Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. ✉email: shind@ornl.gov

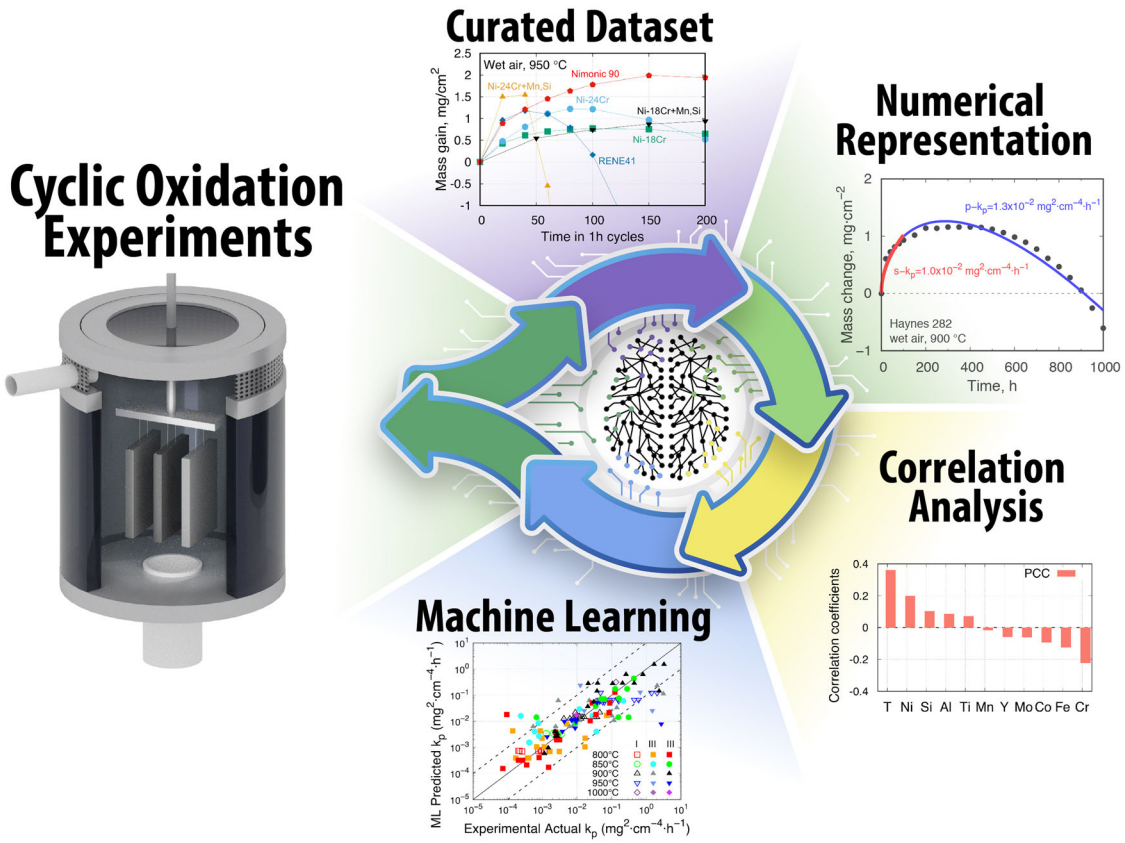

Fig. 1 Workflow of the present study. Raw experimental data from cyclic oxidation experiments were fitted with select oxidation models to represent their oxidation kinetics, resulting in a dataset. Next, correlation analysis was performed to evaluate the strength of correlation between input features and k_p. Then, the performance of various ML models was evaluated.

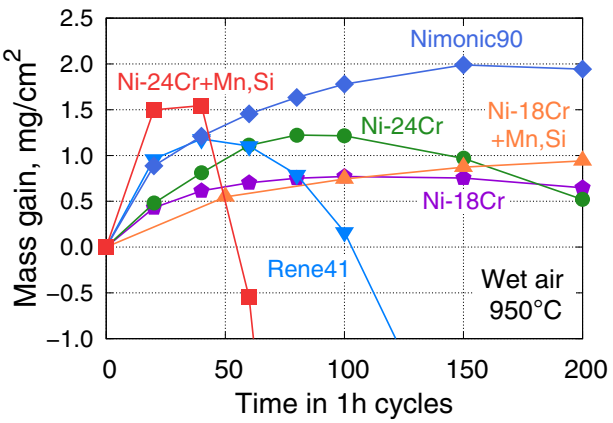

Fig. 2 Snapshot of ORNL cyclic oxidation dataset. Mass gain of select alloys as a function of oxidation time in 1 h cycles.

summarized in Table 1. This consistently measured oxidation and well-curated dataset, collected over years of experiments at ORNL, enabled exploring the potential of ML to predict the oxidation kinetics of multi-component high-temperature alloys. However, raw cyclic oxidation data, which is given as a $2 \times N$ matrix (cycle vs. mass change) per alloy, cannot be directly used in ML because singular numeric values are needed as the target for the regression-type ML training. Hence, it is necessary to find a proper numerical representation for the oxidation kinetics of each alloy so that the correlation between the input features and the target, as well as the capability of ML in modeling the high-temperature oxidation kinetics, can be evaluated.

High-temperature oxidation of multi-component alloys is a complex dynamic process, which requires concurrent consideration of both thermodynamics and diffusion kinetics[2]. As shown in Fig. 2, the addition of certain minor alloying elements can significantly alter the mass change behavior of some alloys during oxidation. Thus, identifying a proper numerical representation for such complex phenomena is the key to building reliable ML models to predict the high-temperature oxidation kinetics of NiCr-based alloys. Mathematically, any mass change curve can be fitted with an arbitrary polynomial. However, parameters obtained from such regression without any physical meaning will not allow gaining insights into the oxidation mechanisms and compare alloy oxidation performance. Thus, we have focused on identifying widely accepted oxidation models with a robust formalism to determine physically meaningful parameters to represent the oxidation kinetics of the studied alloys.

As illustrated in Fig. 3, two oxidation models, i.e., a simple parabolic law ($\Delta m = \sqrt{k_p t}$, where Δm is the mass gain and t is the oxidation time; *s*-k_p hereafter)[39,40], and a statistical cyclic oxidation model (*p*-k_p hereafter)[41], were adopted in this work. Both models have been widely used to interpret the oxidation kinetics of numerous alloys[42–48]. In this study, the s-k_p model evaluated the mass change data up to 100 h to exclusively represent growth rates with the minimized influence of mass loss. The *p*-k_p model[41], consisting of two parameters, i.e., p (discrete oxide spallation probability) and k_p, evaluated the complete curve, thus, enabling us to assess the capability of the ML approach to predict oxidation kinetics involving both oxide growth and loss processes. It is anticipated that the derived k_p will be a good numerical representation of our raw experimental cyclic oxidation data and can serve as proper target property for analyzing the oxidation

Table 1. List of alloying elements with composition range in wt% and experimental conditions, and details of the datasets.

Compositions of model alloys	Ni (41.7-86), Cr (14-25), Al (0-3.9), Co (0-16.2), Fe (0-40.1), Mn (0-1.5), Mo (0-9.8), Si (0-0.5), Ti (0-3.3), Y (0-0.1)
Commercial alloys	Haynes 282 (H282), Nimonic 80 A (N80), Nimonic 90 (N90), Rene 41 (R41)
Temperature, °C	800, 850, 900, 950
Atmospheres	Dry air and wet air (air + 10% water vapor)
Cycles of oxidation	Up to 2000 cycles (1 h/cycle)
Number of data points in each dataset	dry air (s-k_p): 57, dry air (*p*-k_p): 57 wet air (s-k_p): 151, wet air (*p*-k_p): 151
Input features	[elements]: Elemental composition in wt% T: Oxidation temperature
Target	Parabolic rate constant (k_p)

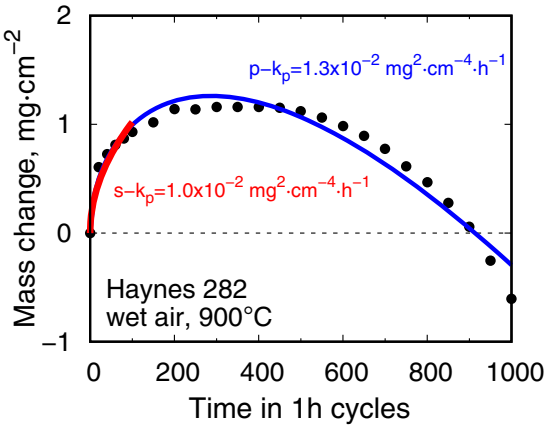

Fig. 3 Schematic diagram of the fitted mass change curve. The mass gain curve of Haynes 282 in wet air at 900 °C in a 1 h cycle was fitted with the s-k_p and *p*-k_p models, respectively.

kinetics by the data analytics approach. Since the oxidation mechanisms of NiCr-based alloys in dry and wet air are different[49,50], the complete experimental data after being fitted by each oxidation model were divided into two datasets based on the oxidation atmosphere. Finally, four datasets, *s*-k_p, and *p*-k_p in dry and wet air were generated from the existing experimental data for further analysis in this study (see Table 1).

Correlation analyses

We have used MIC and PCC methods to quantitatively analyze the correlation between all input features (alloy compositions and oxidation temperature from Table 1) and k_p, respectively. While PCC can evaluate the strength of the positive and negative, but only linear relationships between two variables. The MIC method can identify the strength of both linear and nonlinear relationships, but only the magnitude without a sign. Both approaches are expected to provide insights into the correlation between input features and k_p from differing statistical aspects, which may inspire alloy design experts to generate alloy hypotheses[23,26]. Moreover, correlation analysis can facilitate the training of high-fidelity ML models using highly ranked features. For the PCC analysis, a positive correlation between a feature and k_p here implies that the increase of the feature will increase k_p, i.e., a higher oxidation rate, and vice versa. Figure 4 shows the PCC analysis results for the *p*-k_p of dry and wet air datasets as examples. The remaining correlation analysis results are presented in Supplementary Figs. 1 and 2. It is worth mentioning that although as the base element, Ni content is the balance of remaining alloying elements, we still use Ni content as an input feature since ML algorithms are only solving a mathematical problem and do not know such a correlation. The oxidation temperature (T) was identified as the feature having the most substantial impact on k_p in all cases. It is encouraging that correlation analysis replicated the community's existing

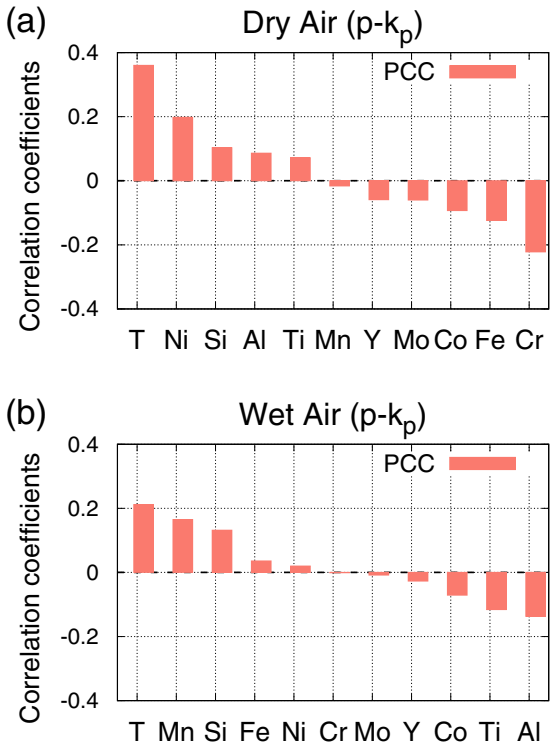

Fig. 4 Correlation analysis of the training dataset. Quantified correlation scores of input features (elemental compositions and oxidation temperature) and k_p determined for the p-k_p model using Pearson correlation coefficient at difference atmosphere: (**a**) dry and (**b**) wet air.

knowledge: oxidation temperature is typically the factor that most strongly affects oxidation rate. PCC also correctly identified that temperature has the most positive correlation with k_p, consistent with the fact that the oxidation rate increases with increasing temperature.

It can also be observed in Fig. 4 and Supplementary Fig. 1 that chromium (Cr) as the major alloying element in the studied alloys was identified to have the most negative correlation with k_p, which is consistent with the fact that the Cr addition is beneficial to the formation of an external chromia solid-state diffusion barrier that slows the oxidation reaction as it increases in thickness[51,52]. Most of the alloys in this dataset were designed to form protective chromia scales. Aluminum (Al) is expected to further decelerate the growth rate of the chromia scale for NiCr-based alloys[53]; thus, the negative coefficient of Al in wet air (Fig. 4b) is reasonable. Fe is known to form fast-growing oxides[54,55]. Its second strongest positive correlations with k_p in both dry and wet air (Fig. 4) are in accordance with this understanding. Similar correlations were also observed for Cr

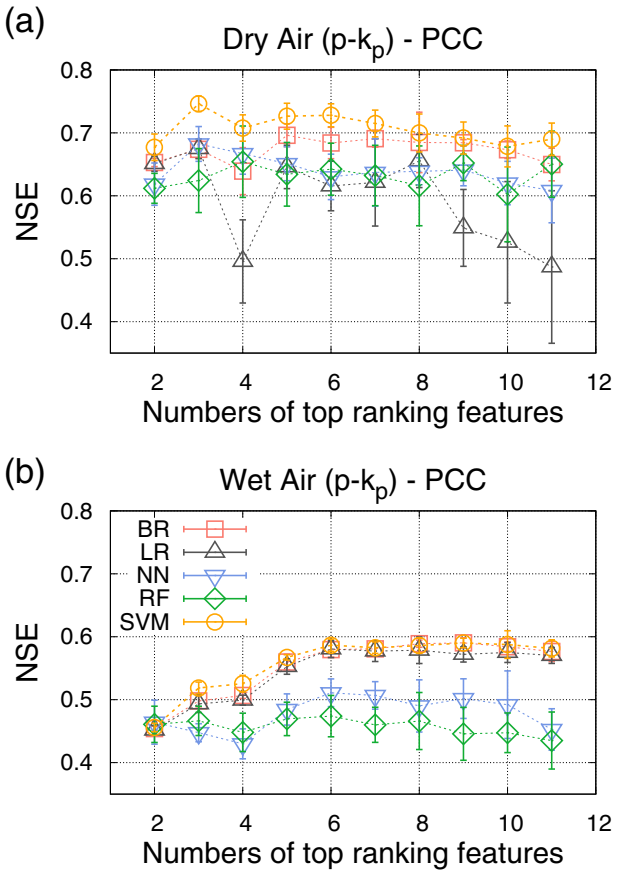

Fig. 5 Performance of trained ML models. NSE of five trained ML models (BR: Bayesian ridge regression, LR: linear regression, NN: nearest neighbor, RF: random forest, and SVM: support vector machines regression) as a function of the number of top-ranking features based on the rankings of |PCC| (the absolute values of PCC in Fig. 4) in (**a**) dry and (**b**) wet air. Each model was trained ten times for a given set of features to determine the averaged accuracy and its standard deviation (error bar).

and Fe in the dry and wet air (s-k_p) datasets, as shown in Supplementary Fig. 1. Other than identifying linear correlation as PCC does, MIC can also identify non-linear correlations. As presented in Supplementary Fig. 2, features like T, Cr, Fe, Mo, Mn, and Ni in dry air and T, Al, Fe, and Cr in wet air that have a significant impact on the oxidation kinetics of NiCr-based alloys were generally identified as high-ranking features by the MIC analysis.

Correlations that seem to contradict known oxidation mechanisms were also observed. For example, Ni exhibits a negative correlation with k_p, meaning that Ni addition will decelerate the oxidation rate. However, correlation analysis results also heavily depend on the nature of the dataset, which means that the rankings from correlation analysis here do not necessarily all make intuitive sense and mean causation[23]. In addition, high-temperature oxidation kinetics of multi-component alloys is controlled by chemical compositions and temperature and by factors including, but not limited to, microstructure and its evolution during oxidation, as well as oxide scale spallation/evaporation[2]. Although our dataset was collected over a number of years from differing studies at ORNL and could be considered a big dataset from a material science/alloy oxidation perspective, it is still small from a data science perspective. Thus, the dataset does not consistently cover the ideal composition ranges and test conditions to fully facilitate data analytics for the classes of alloys included. Certainly, this limitation could most likely be said of any existing materials dataset, which includes complex alloy compositions, structures, and behavior. However, the correctly identified correlation in Fig. 4, Supplementary Figs. 1 and 2 still clearly demonstrated the value of correlation analysis in applying data analytics to materials science. The intention of performing correlation analysis here is not only to gain insights into the impact of input features on k_p but also to have a numerical basis for the selection of input features to train ML models for understanding their influence on the performance of ML models.

Machine learning

Based on the ranking of features from correlation analyses, five widely used ML models, i.e., LR, BR, NN, RF, and SVM, have been trained using different numbers of top-ranking features to evaluate their performances. For models trained with the dry and wet air p-k_p datasets, Fig. 5 presents their average performance, represented by the Nash and Sutcliffe coefficient of efficiency (NSE)[56], and corresponding standard deviation as a function of the numbers of top-ranking features based on the rankings of |PCC| (the absolute values of PCC in Fig. 4). The remaining results are presented in Supplementary Figs. 3 and 4. Since NSE alone could not fully qualify the performance of these models, another metric, i.e., coefficient of determination (R^2 COD), and a band showing the goodness-of-fit were adopted to facilitate the assessment of ML models, as presented in Fig. 6.

NSE here was mainly used to compare the performance of these models. As shown in Fig. 5 and Supplementary Fig. 3, the performance of ML models trained with dry air datasets shows a similar trend to those with wet air datasets. In most cases, increasing the top-ranking features from 2 to ~4 or ~6 increased the NSE of these models markedly. Intriguingly, considering more features did not considerably improve the performance of the obtained models. The NSE of the models trained with the wet air datasets is relatively lower than their dry air datasets counterparts. This trend can be attributed to the increased variation of measured oxidation kinetics of alloys studied in wet air exposure. For the dry air dataset, the maximum NSE of all models is between 0.6 and 0.75. Among the considered models, SVM has the highest NSE (>0.7) in all cases (Fig. 5 and Supplementary Figs. 3 and 4). It reaches the maximum NSE when using the top 3 to 5 features. Afterward, it gradually decreases with an increasing number of features, revealing that in this study including features with a ranking lower than the 5th is not beneficial to the performance of ML models. For the wet air dataset, the maximum NSE of all models is only between 0.45 and 0.60. In general, the top four to six features in this study were required to achieve an NSE of 0.5 or higher. SVM still exhibits the highest accuracy among considered models. As discussed in the previous section, most of these features are experimentally identified features that significantly impact the oxidation of NiCr-based alloys.

In all cases, the performance of these ML models evaluated with NSE is generally in the order of SVM>BR>RF≈NN>LR (in dry air) and SVM>BR≈LR>NN>RF (in wet air). Results show that these models are sensitive to the number of features considered in training but to different extents. SVM performs similarly to the linear-based models BR and LR, particularly for those trained with the wet air datasets. A linear kernel function (assuming the input features and k_p is in a linear relationship) was adopted for the present SVM models after hyperparameter tuning. Considering more than the top six to eight features significantly degraded the performance of LR for the dry air dataset, whereas this was not the case for the wet air dataset. This could be caused by the smaller data volume of the dry air dataset than the wet air dataset. BR uses probability distributors to formulate LR, rather than point estimates used by LR and identifies the posterior distribution for the model parameters. Thus, it is believed that BR is more tolerant of overfitting[32,33]. This is why, in this case, the NSE of the BR model

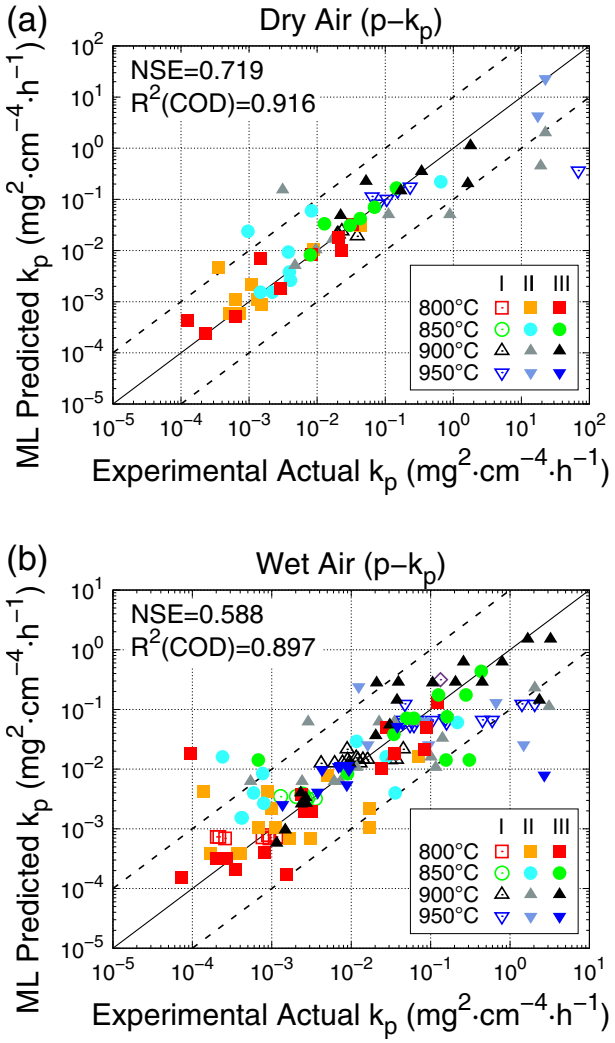

Fig. 6 ML parity plots. Experimental and SVM-predicted p-k_p of three groups of alloys (I: commercial alloys, II: NiCr binary alloys, and III: NiCr alloys with additions) at different oxidation temperatures are compared for (**a**) dry and (**b**) wet air. The region between the dashed lines indicates one order of magnitude fidelity range as an acceptable deviation.

does not notably decrease with an increasing number of features. RF in this study generally requires the inclusion of the top four features to obtain the NSE of >0.65 (for dry air datasets) and >0.5 (for wet air datasets); after that, its performance was almost independent of the number of considered features.

The observed trend can be understood as follows: as an ensemble learning method, RF assigns different importance to each feature during model training; thus, less critical features would have less or even no contribution to its performance[57]. Therefore, the performance of RF is not sensitive to the number of features considered in training after it reaches the maximum accuracy. NN hinges on the assumption that similar things are close to each other and simply outputs the average value of data points in *k*-nearest neighbors[58]. The performance of NN usually degrades as the dimensionality of data increases significantly[59], because its distance measure, i.e., the Euclidean distance, becomes less representative in a higher-dimensional space. However, as shown in Fig. 5, Supplementary Figs. 3 and 4, when the number of features increases, accompanying the increase of the dimensionality of data, the accuracy of NN does not decrease significantly. This finding indicates that the dimensionality of data

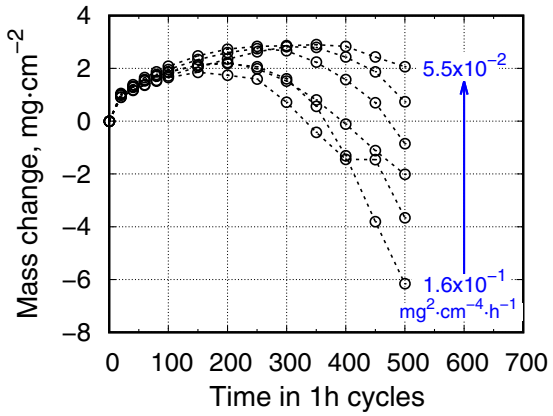

Fig. 7 Oxidation of Nimonic 80 A at 950 °C in wet air in 1 h cycle. The k_p determined by the *p*-k_p model largely varies from six tests under the same condition.

in this study does not induce a significant adverse effect on the performance of NN. Since SVM possesses the best performance among all considered ML models, its performance does not significantly degrade with increasing numbers of features.

SVM models trained with all 11 features in our datasets are believed to be the most promising ML models for predicting the k_p of NiCr-based alloys since they exhibit relatively good performance and concurrently consider essential features from a high-temperature oxidation perspective. Comparisons between the experimental and predicted k_p by the SVM models trained with all 11 features are presented in Fig. 6 (for *p*-k_p) and Supplementary Fig. 5 (for s-k_p). While the accuracy of trained ML models evaluated with NSE is around 0.7, it is encouraging that most of the predicted values are within the acceptable deviation range. In addition, the other metric R^2 COD for both dry and wet air (*p*-k_p) is close to 0.9 or higher, which strongly indicates that our ML models can efficiently capture the trend of cyclic oxidation response of NiCr-based alloys as a function of alloy compositions and oxidation temperature.

In practice, the same alloy tested under the same condition may exhibit significant test-to-test variations, particularly in wet air testing. For example, the k_p from the *p*-k_p model of Nimonic 80 A at 950 °C in wet air from six tests varied significantly (Fig. 7). Thus, dashed lines, as the acceptable deviation, representing one order of magnitude difference between the predicted and experimental k_p are superimposed in Fig. 6 and Supplementary Fig. 5. The predicted k_p correctly captures the experimental data trend, and most of the data points lie between the dashed lines. This indicates good agreement between experimental and predicted k_p was achieved from a high-temperature oxidation perspective for the multi-component commercial alloys (group I). This conclusion is further supported by the high R^2 COD of ≥0.9 in both cases. Data in Fig. 6b are slightly more scatted than those in Fig. 6a, attributing to the increased spallation probability and thus more considerable test-to-test variation in wet air exposure.

Although a satisfactory agreement between the experimental and predicted k_p was achieved, more work is required in the future. Firstly, the performance of ML models depends not only on what features have been considered but also on the nature and repartition of the dataset used for training; thus, further detailed analysis of this aspect is required. Secondly, k_p alone could not fully capture the complete oxidation kinetics of alloys. Other properties, such as spallation probability and oxidation lifetime, should also be concurrently considered as target properties to be modeled/predicted. Thirdly, besides the alloy composition and temperature considered in this study as input features, the high-temperature oxidation kinetics are also affected by factors like microstructure, oxide scale spallation, alloy surface depletion of

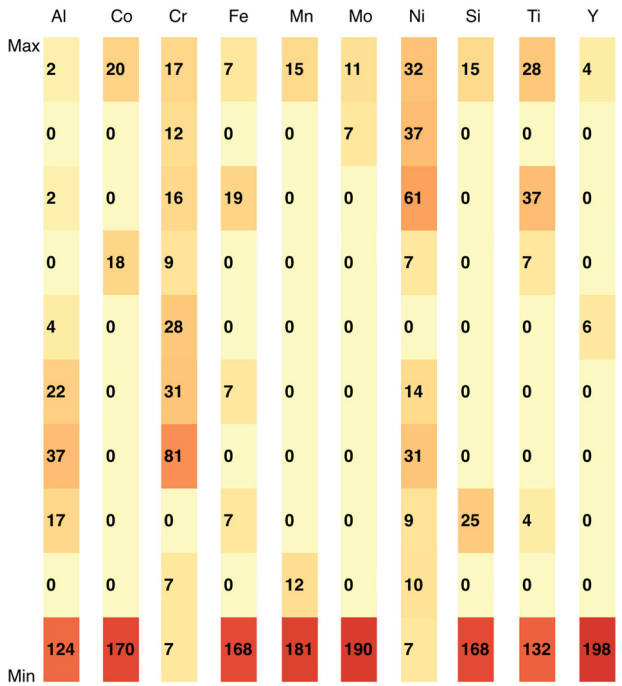

Fig. 8 Data distribution of the current dataset. The concentration range of each element, i.e., the difference between the maximum and minimum content, is evenly divided into ten segments. The number in each segment represents how many data points are located in each cell. For example, a cell with 0 means that there is no data in this range. The degree of data aggregation in each segment is represented by color.

elements that are selectively oxidizing, and oxide evaporation in wet air atmosphere. Features that can well represent these additional factors and ML frameworks that can incorporate these factors as much as possible will be highly conducive to establishing high-fidelity surrogate models.

Additionally, material datasets commonly suffer from the uneven data distribution. While the present dataset can be regarded as one of the largest and the most comprehensive cyclic oxidation dataset for high-temperature NiCr-based alloys, many gaps are identified. A highly desirable alloy dataset to apply data analytics would have an even distribution of data points of all elements without major gaps in a high-dimensional space. The movement toward such an ideal alloy dataset can be obtained by exercising a design of experiments (DOE)[60] campaign to fill key gaps in the existing dataset(s). However, the dataset used in the current study is biased to some extent and has a number of gaps, as shown in Fig. 8. This is attributed to the fact that the objective of past empirical alloy research has been mainly focused on identifying alloys with superior properties over existing ones. Consequently, the alloy composition variation strategy has been heavily geared toward improving properties, more often ending up in biased and narrow boundary conditions from a data analytics perspective. It is evident that alloy data with poor properties can serve as equally useful learning input for ML models; however, an effort to strategically collect such data to fill the gaps has been overlooked, not only by us but likely throughout the materials community. Hence, a strategy that can identify the most important gaps in an existing dataset for alloy data analytics and efficiently augment a small amount of experimental data needs to be developed to leverage legacy data, particularly for complex materials.

In summary, we demonstrated a workflow of applying a data analytics approach to predict the cyclic oxidation kinetics of NiCr-based alloys using a highly consistent and well-curated experimental dataset collected over many years and numerous different studies at ORNL. The oxidation data in dry and wet air as a function of alloy compositions and oxidation temperatures (800–950 °C) were used to train ML models. Identifying a proper numerical representation was the critical procedure for successfully predicting the high-temperature oxidation kinetics of NiCr-based alloys. Oxidation kinetics of alloys was represented by the k_p from properly selected oxidation models, i.e., a simple parabolic growth law for data up to 100 h (s-k_p) and a statistical cyclic oxidation model for the complete curve (p-k_p, up to 2000 h), respectively.

Correlation analysis and training of ML models (i.e., BR, LR, NN, RF, and SVM regression) were performed to quantitatively evaluate the inter-relationship between input features and k_p. The influence of top-ranking features on the performance of considered ML models was also identified. The trained ML models were able to predict k_p from both s-k_p and p-k_p models satisfactorily. The typical test-to-test variation in experimental high-temperature oxidation data was believed to be the primary source for the discrepancy between the ML-predicted and experimental k_p. The current results demonstrated the potential of the ML approach to predict the oxidation kinetics of multi-component NiCr alloys with wide composition and temperature ranges. We anticipate that this data analytics method can be applied to predict the oxidation kinetics of other classes of high-temperature alloys.

METHODS

Experimental procedure

All specimens with the dimensions of ~$10 \times 20 \times 1.5$ mm^3 were ground to a 600-grit finish and cleaned ultrasonically in acetone and methanol prior to cyclic oxidation experiments. A 1-h-cycle consisting of a 60 min exposure at temperature and 10 min cooling in automated cyclic rigs[61] was adopted in this work. The specimens were exposed at temperatures between 800 and 950 °C in flowing dry and wet air (air + 10% water vapor), respectively, with a gas flow rate of 500 cm^3 min^{-1} for up to 2000 h.

Correlation analysis and machine learning

The correlation between the input features and k_p was evaluated by PCC[62] and MIC[30], respectively. PCC considers the strength of the linear relationship between two variables, while MIC can identify the strength of both linear and nonlinear relationships. The correlation coefficient of PCC is between −1 and 1, where 1 indicates a total positive linear correlation, −1 represents a complete negative/reciprocal linear correlation, and 0 indicates no correlation. The correlation coefficient of MIC ranges between 0 and 1. Coefficient values close to 1 or −1 indicate the strongest correlation between the variables and target property.

Five representative ML models were used: LR[31], BR[32,33], NN[34], RF[35], and SVM regression[36]. A brief introduction to these models is provided below. For details, please refer to the corresponding literature. LR models the relationship between input and output variables by fitting a linear equation. Parameters of the LR model are fitted to minimize the residual sum of squares between input and output data. BR is another LR model. It formulates an LR using probability distributors rather than point estimates that LR does. The output is assumed to be drawn from a probability distribution instead of being estimated as a single value. *k*-NN outputs the average of the values of given data points in the *k*-nearest neighbors. Since the function is approximated only locally, it only uses a subset of the relevant dataset. RF is an ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees. SVM can be used for both classification and regression problems. For a classification problem, SVM constructs a set of hyperplanes in high-dimensional space to distinctly classifies the data points. For a regression problem, SVM is generalized by introducing an ε-insensitive region (ε-tube) around the function. This tube reformulates the regression problem to find a function exhibiting the least deviation from the obtained targets across all training data while balancing model complexity and prediction error.

These ML models were trained with different numbers of top-ranking features based on the ranking from MIC or |PCC| (the absolute value of

PCC) to explore the influence of these features on the performance of ML models. The hyperparameters for each model were tuned up to 10,000 iterations using the randomized optimization approach to search for the optimal parameters. Each model was trained ten times for a given set of features to determine the averaged accuracy and its standard deviation. All features have the same weight in ML training regardless of their rankings determined by correlation analysis. Both correlation analysis and training of ML models were performed with the Python-based open-source frontend, Advanced data SCiEnce toolkit for Non-Data Scientists (ASCENDS)[58,63], which is available via GitHub (https://github.com/ornlpmcp/ASCENDS). The performance/accuracy of the ML models was quantified by NSE[56] and R^2(COD), calculated without an intercept as a supplementary metric. The NSE is calculated as follows:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2} \quad (1)$$

where O_i and $\overline{O}$ are the observed values and their mean, respectively. P_i is the model predicted value. N is the number of data in the dataset. The k-fold cross-validation approach[64] with $k = 5$ was used in ML training. This approach randomly divided the input data into k groups. One group (i.e., unseen data) was withheld during training, and the remainder k-1 groups were used to train the ML model. Then the unseen dataset was used to evaluate the accuracy of models.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

CODE AVAILABILITY

Codes used in this work are available via GitHub (https://github.com/ornlpmcp/ASCENDS).

Received: 1 May 2021; Accepted: 13 July 2021;
Published online: 09 August 2021

REFERENCES

1. Birks, N., Meier, G. H. & Pettit, F. S. *Introduction to the High Temperature Oxidation of Metals.* (Cambridge Univ. Press, 2006).
2. Young, D. J. *High Temperature Oxidation and Corrosion of Metals.* 2nd edn (Elsevier, 2016).
3. Das, N. K. & Shoji, T. Early stage oxidation of Ni–Cr binary alloy (111), (110) and (100) surfaces: a combined density functional and quantum chemical molecular dynamics study. *Corros. Sci.* **73**, 18–31 (2013).
4. Ohler, B., Prada, S., Pacchioni, G. & Langel, W. DFT simulations of titanium oxide films on titanium metal. *J. Phys. Chem. C.* **117**, 358–367 (2013).
5. Hong, K.-H., Kim, J. H., Chang, K. & Kwon, J. The role of Cr on oxide formation in Ni-Cr alloys: a theoretical study. *Comput. Mater. Sci.* **142**, 185–191 (2018).
6. Qi, J., Xu, H., Liang, Z., Lu, P. & Zhou, C. The role of Cr atom in the early steam oxidation of Fe-based alloys: an atomistic simulation. *Mater. Corros.* **72**, 465–473 (2020).
7. Tedmon, C. S. The effect of oxide volatilization on the oxidation kinetics of Cr and Fe-Cr Alloys. *J. Electrochem. Soc.* **113**, 766–768 (1966).
8. Vaché, N., Cadoret, Y., Dod, B. & Monceau, D. Modeling the oxidation kinetics of titanium alloys: review, method and application to Ti-64 and Ti-6242s alloys. *Corros. Sci.* **178**, 109041 (2020).
9. Chyrkin, A. et al. Modeling carbide dissolution in alloy 602 CA during high temperature oxidation. *Corros. Sci.* **96**, 32–41 (2015).
10. Pillai, R. et al. External alpha-Al_2O_3 scale on Ni-base alloy 602 CA - Part II: microstructural evolution. *Corros. Sci.* **127**, 27–38 (2017).
11. Chyrkin, A. et al. Modelling compositional changes in nickel base-alloy 602 CA during high temperature oxidation. *Mater. High. Temp.* **32**, 102–112 (2014).
12. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).
13. Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 1–36 (2019).
14. Ramakrishna, S. et al. Materials informatics. *J. Intell. Manuf.* **30**, 2307–2326 (2019).
15. Bock, F. E. et al. A review of the application of machine learning and data mining approaches in continuum materials mechanics. *Front. Mater.* **6**, 110 (2019).
16. Meredig, B. et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 094104 (2014).
17. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 2810 (2013).
18. Pei, Z., Yin, J., Hawk, J. A., Alman, D. E. & Gao, M. C. Machine-learning informed prediction of high-entropy solid solution formation: beyond the Hume-Rothery rules. *npj Comput. Mater.* **6**, 1–8 (2020).
19. Lee, S., Peng, J., Shin, D. & Choi, Y. S. Data analytics approach for melt-pool geometries in metal additive manufacturing. *Sci. Technol. Adv. Mat.* **20**, 972–978 (2019).
20. Alberi, K. et al. The 2019 materials by design roadmap. *J. Phys. D: Appl. Phys.* **52**, 013001 (2018).
21. Rajan, K. Materials informatics. *Mater. Today* **8**, 38–45 (2005).
22. Peng, J. et al. Uncertainty quantification of machine learning predicted creep property of alumina-forming austenitic alloys. *JOM* **73**, 164–173 (2021).
23. Shin, D., Yamamoto, Y., Brady, M. P., Lee, S. & Haynes, J. A. Modern data analytics approach to predict creep of high-temperature alloys. *Acta Mater.* **168**, 321–330 (2019).
24. Verma, A. K. et al. Mapping multivariate influence of alloying elements on creep behavior for design of new martensitic steels. *Metall. Mater. Trans. A* **50**, 3106–3120 (2019).
25. Chang, Y.-J., Jui, C.-Y., Lee, W.-J. & Yeh, A.-C. Prediction of the composition and hardness of high-entropy alloys by machine learning. *JOM* **71**, 3433–3442 (2019).
26. Peng, J., Yamamoto, Y., Hawk, J. A., Lara-Curzio, E. & Shin, D. Coupling physics in machine learning to predict properties of high-temperatures alloys. *npj Comput. Mater.* **6**, 141 (2020).
27. Shen, C. et al. Physical metallurgy-guided machine learning and artificial intelligent design of ultrahigh-strength stainless steel. *Acta Mater.* **179**, 201–214 (2019).
28. Bhattacharya, S. K., Sahara, R. & Narushima, T. Predicting the parabolic rate constants of high-temperature oxidation of Ti alloys using machine learning. *Oxid. Met.* **94**, 205–218 (2020).
29. Sedgwick, P. Pearson's correlation coefficient. *BMJ* **345**, e4483 (2012).
30. Reshef, D. N. et al. Detecting novel associations in large data sets. *Science* **334**, 1518–1524 (2011).
31. Freedman, D. A. *Statistical Models: Theory and Practice* (Cambridge Univ. Press, 2009).
32. MacKay, D. J. Bayesian interpolation. *Neural Comput.* **4**, 415–447 (1992).
33. Tipping, M. E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **1**, 211–244 (2001).
34. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **46**, 175–185 (1992).
35. Barandiaran, I. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 832–844 (1998).
36. Awad, M. & Khanna, R. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers.* (Apress, 2015).
37. Pint, B. A., Haynes, J. A. & Armstrong, B. L. Performance of advanced turbocharger alloys and coatings at 850–950 °C in air with water vapor. *Surf. Coat. Technol.* **215**, 90–95 (2013).
38. Pillai, R., Romedenne, M., Haynes, J. A. & Pint, B. A. Oxidation behavior of candidate NiCr alloys for engine exhaust valves: part I—effect of minor alloying elements. *Oxid. Met.* **95**, 157–187 (2021).
39. Pieraggi, B. Calculations of parabolic reaction rate constants. *Oxid. Met.* **27**, 177–185 (1987).
40. Hindam, H. & Whittle, D. Microstructure, adhesion and growth kinetics of protective scales on metals and alloys. *Oxid. Met.* **18**, 245–284 (1982).
41. Poquillon, D. & Monceau, D. Application of a simple statistical spalling model for the analysis of high-temperature, cyclic-oxidation kinetics data. *Oxid. Met.* **59**, 409–431 (2003).
42. Pint, B. A., Terrani, K. A., Brady, M. P., Cheng, T. & Keiser, J. R. High temperature oxidation of fuel cladding candidate materials in steam–hydrogen environments. *J. Nucl. Mater.* **440**, 420–427 (2013).
43. Kuner, M. C., Romedenne, M., Fernandez-Zelaia, P. & Dryepondt, S. Quantitatively accounting for the effects of surface topography on the oxidation kinetics of additive manufactured Hastelloy X processed by electron beam melting. *Addit. Manuf.* **36**, 101431 (2020).
44. Pint, B. A., Pillai, R., Lance, M. J. & Keiser, J. R. Effect of pressure and thermal cycling on long-term oxidation in CO_2 and supercritical CO_2. *Oxid. Met.* **94**, 505–526 (2020).
45. Raffaitin, A., Monceau, D., Andrieu, E. & Crabos, F. Cyclic oxidation of coated and uncoated single-crystal nickel-based superalloy MC2 analyzed by continuous thermogravimetry analysis. *Acta Mater.* **54**, 4473–4487 (2006).

46. Gheno, T., Monceau, D. & Young, D. J. Kinetics of breakaway oxidation of Fe–Cr and Fe–Cr–Ni alloys in dry and wet carbon dioxide. *Corros. Sci.* **77**, 246–256 (2013).
47. Monceau, D. & Poquillon, D. Continuous thermogravimetry under cyclic conditions. *Oxid. Met.* **61**, 143–163 (2004).
48. Dryepondt, S., Turan, J., Leonard, D. & Pint, B. A. Long-term oxidation testing and lifetime modeling of cast and ODS FeCrAl alloys. *Oxid. Met.* **87**, 215–248 (2017).
49. Zurek, J. et al. Growth and adherence of chromia based surface scales on Ni-base alloys in high-and low-p_{O2} gases. *Mater. Sci. Eng., A* **477**, 259–270 (2008).
50. Pint, B. A. Addressing the role of water vapor on long-term stainless steel oxidation behavior. *Oxid. Met.* **95**, 335–357 (2021).
51. Wallwork, G. The oxidation of alloys. *Rep. Prog. Phys.* **39**, 401–485 (1976).
52. Giggins, C. & Pettit, F. Oxidation of Ni-Cr alloys between 800 degrees and 1200 degrees C. *Trans. Met. Soc. AIME* **245**, 2495–2507 (1969).
53. Giggins, C. S. & Pettit, F. S. Oxidation of Ni-Cr-Al alloys between 1000 degrees and 1200 degrees C. *J. Electrochem. Soc.* **118**, 1782–1790 (1971).
54. Stott, F. H. Influence of alloy additions on oxidation. *Mater. Sci. Technol.* **5**, 734–740 (1989).
55. Croll, J. E. & Wallwork, G. R. The high-temperature oxidation of iron-chromium-nickel alloys containing 0–30% chromium. *Oxid. Met.* **4**, 121–140 (1972).
56. Nash, J. E. & Sutcliffe, J. V. River flow forecasting through conceptual models part I —A discussion of principles. *J. Hydrol.* **10**, 282–290 (1970).
57. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
58. Peng, J., Lee, S., Williams, A., Haynes, J. A. & Shin, D. Advanced data science toolkit for non-data scientists–A user guide. *CALPHAD* **68**, 101733 (2020).
59. Geenens, G. Curse of dimensionality and related issues in nonparametric functional regression. *Stat. Surv.* **5**, 30–43 (2011).
60. Dean, A., Voss, D. & Draguljić, D. *Design and Analysis of Experiments* Vol. 1 (Springer, 1999).
61. Pint, B. A., Tortorelli, P. F. & Wright, I. G. Effect of cycle frequency on high-temperature oxidation behavior of alumina-forming alloys. *Oxid. Met.* **58**, 73–101 (2002).
62. Pearson, K. & Lee, A. On the laws of inheritance in man: I. Inheritance of physical characters. *Biometrika* **2**, 357–462 (1903).
63. Lee, S., Peng, J., Williams, A. & Shin, D. ASCENDS: advanced data SCiENce toolkit for non-data scientists. *J. Open Source Softw.* **5**, 1656 (2020).
64. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning* Vol. 112 (Springer, 2013).

ACKNOWLEDGEMENTS

This research was sponsored by the Department of Energy, Energy Efficiency and Renewable Energy, Vehicle Technologies Office, Propulsion Materials Program. This research used resources of the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory, which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC05-00OR22725. The authors thank C. Layton for his support for using CADES resources and G. Garner for assisting the oxidation experiments. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US Government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

AUTHOR CONTRIBUTIONS

D.S. conceived the study. R.P., M.R., and B.A.P. curated and prepared the dataset. J.P. performed correlation analysis and machine learning training. Commercial alloys relevant to the exhaust valve environment were selected by G.M., B.A.P., and J.A.H. All authors analyzed the data. J.P. and D.S. drafted the manuscript. All authors reviewed and edited the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41529-021-00188-z.

Correspondence and requests for materials should be addressed to D.S.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.