Uncertainty Quantification and Bayesian Active Learning for Rupture Life Prediction in Ferritic-Martensitic Steels

Osman Mamun (mamun.che06@gmail.com)
Pacific Northwest National Laboratory
M.F.N. Taufique
Pacific Northwest National Laboratory
Madison Wenzlick
National Energy Technology Laboratory
Jeffrey Hawk
National Energy Technology Laboratory
Ram Devanathan
Pacific Northwest National Laboratory

Research Article

Keywords: framework. uncertainty, model, probabilistic, developed, creep

Posted Date: September 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-887257/v1

Uncertainty Quantification and Bayesian Active Learning for Rupture Life Prediction in Ferritic-Martensitic Steels

Osman Mamun[1*], M.F.N Taufique[1], Madison Wenzlick[2, 3], Jeffrey Hawk[2], and Ram Devanathan[1]

[1]Energy and Environment Directorate, Pacific Northwest National Laboratory, Richland, USA

[2]Materials Performance Division, National Energy Technology Laboratory, 1450 Queen Avenue SW, Albany, OR 97321, USA

[3]Leidos Research Support Team, 1450 Queen Avenue SW, Albany, OR 97321, USA

[*]Corresponding author

Abstract

Three probabilistic methodologies are developed for predicting the long-term creep rupture life of $9-12\ wt\%\ Cr$ ferritic-martensitic steels using their chemical and processing parameters. The framework developed in this research strives to simultaneously make efficient inference along with associated risk, i.e., the uncertainty of estimation. The study highlights the limitations of applying probabilistic machine learning to model creep life and provides suggestions as to how this might be alleviated to make an efficient and accurate model with the evaluation of epistemic uncertainty of each prediction. Based on extensive experimentation, Gaussian Process Regression yielded more accurate inference ($Pearson\ \ correlation\ coefficent > 0.95$ for the holdout test set) in addition to meaningful uncertainty estimate (i.e., coverage ranges from 94 – 98% for the test set) as compared to quantile regression and natural gradient boosting algorithm. Furthermore, the possibility of an active learning framework to iteratively explore the material space intelligently was demonstrated by simulating the experimental data collection process. This framework can be subsequently deployed to improve model performance or to explore new alloy domains with minimal experimental effort.

Introduction

Advanced ultra-supercritical power plants require increased steam temperature and pressure, for higher efficiency and lower carbon emissions, to comply with environmental regulations. Without building new power plants this can be achieved by increasing the operating temperature and pressures of the existing power plants. To design ferritic-martensitic steels (the most common cost-effective alloys used in power plants today) in the 9-12 *wt% Cr* range that can withstand these higher operating temperatures and pressures, a predictive model needs to be developed that can reliably and accurately predict the lifetime of an alloy and its interrelation to the factors that govern the mechanical and chemical properties of the alloys[1–3]. In this effort, machine learning has lent itself quite well to develop models with unprecedented accuracy with inference time orders of magnitude shorter than the traditional first-principles density functional theory[4–6], Monte Carlo simulations[7], molecular dynamics[8], or phase field modeling[9–11]. In a recent article, it was shown that a Gradient Boosting Algorithm[12] can be used to efficiently predict the rupture life[13] and rupture strength[14] of 9-12 *wt% Cr* ferritic-martensitic steels and austenitic stainless-steels. However, oftentimes the prediction is unreliable for low confidence modeling, where the data is either scarce and highly non-linear or the uncertainty associated with the prediction is not available. Specifically, long-term creep rupture life exhibits uncertainty that ranges from several years to decades[15]. For proper reliability analysis, it is imperative to quantify the uncertainty associated with each data point, i.e., Epistemic uncertainty. It should be noted that there are two types of uncertainty associated with data sets: 1) Aleatoric uncertainty (i.e., resulting from the intrinsic nature of the data generation process), and 2) Epistemic uncertainty (i.e., resulting from the limitation of the model itself). In this article, three probabilistic machine learning approaches were used to make an efficient prediction with epistemic uncertainty estimate,

namely; 1) Quantile Regression, 2) Natural Gradient Boosting, and 3) Gaussian Process Regression.

As a general rule the time span between the initial discovery of a novel material (i.e., the idea for a new alloy and the research to support further development) and integrating it into existing infrastructure can take more than 20 years[16]. By using artificial intelligence and machine learning, the screening process of identifying new candidate alloys that need to be synthesized and tested in a laboratory can be sped up, thereby greatly reducing the time span of the entire process. In this pursuit, the uncertainty estimate, from the probabilistic machine learning model, of the individual data points yet to be explored, can guide the candidate selection process via Bayesian Optimization in order to optimize for a certain desired property, e.g., longer creep rupture life or high yield stress. Another utility of the uncertainty estimate is to improve the existing machine learning model performance by acquiring data points that lead to maximum information gain (or greatest minimization of the entropy). Active learning has been demonstrated to be very effective in the design of experiments for the classification problem[17]; however, its application to the regression problem is still limited, owing to the sequential nature of the algorithm. In classification, only the decision boundary needs to be learned which implies using only a fraction of data points to build a reliable model. On the other hand, the regression problem requires learning over the whole data range, leading to a very slow data acquisition rate in a sequential learning fashion. In this research, a batch-mode, pool-based active learning framework has been demonstrated where data can be acquired in a parallel manner at each iteration by selecting candidates from several clusters (i.e., learned using unsupervised techniques).

Probiabilistic Machine Learning Algorithms

Quantile Regression

Quantile regression forests are a non-parametric way of estimating the conditional quantiles of high dimensional predictor variables[18,19]. For this approach Y is defined as the response variable while X represents an array of predictor variables. In standard formalism, the regression analysis provides an estimate $\hat{\mu}(x)$ of the conditional mean ($E(Y \mid X = x)$) for the response variable, which is obtained through *learning* the parameter of a regression model that minimizes the expected squared error loss:

$$E(Y \mid X = x) = \arg\min_z E\{(Y - z)^2 \mid X = x\} \quad (1)$$

However, the conditional mean only captures the point estimate of the response variable. It does not provide any information about the distribution of the response variable at a certain point. The conditional distribution of Y being smaller than y given the predictor variable $X = x$ is given by,

$$F(y \mid X = x) = P(Y \leq y \mid X = x) \quad (2)$$

The associated loss function to optimize this relation is given by the following equation,

$$\mathcal{L} = \begin{cases} \alpha\,|y_i - \hat{y}_i| & if\ (y_i - \hat{y}_i) \geq 0 \\ (\alpha - 1)|y_i - \hat{y}_i| & if\ (y_i - \hat{y}_i) < 0 \end{cases} \quad (3)$$

Here, α ranges from 0 to 1 depending on the percentile that one wishes to achieve. To get the 95% prediction interval, two models are fit with $\alpha = 0.025$ and 0.975. In this work, in addition to the prediction interval, the median response variable is also computed using $\alpha = 0.5$. For quantile regression, Gradient Boosting Decision Tree (GBDT) is used which was the best performing non-probabilistic model in a previous study[13]. GBDT is an ensemble of weak decision tree models that iteratively fit data to minimize the error made in the previous iteration[12].

Natural Gradient Boosting Regression

Natural Gradient Boosting (NG Boosting) algorithm is another probabilistic GBDT based model; however, unlike quantile regression, or conditional mean estimation, it learns the full probability distribution by construction[20]. The key idea in NG Boosting is to use the natural

gradients instead of the regular gradients, thus allowing the algorithm to fit a probability distribution over the outcome space, conditioned on the predictor variables. The algorithm consists of three distinct components: 1) Base learner, 2) Parametric distribution, and 3) Scoring rule. The base learner used in the algorithm is the GBDT. Instead of making a point estimate, a probability distribution *learns* by learning the parameter of the distribution, e.g., the mean and standard deviation in the case of Gaussian distribution. The scoring rule is selected in such a way that the forecasted probability distribution gets a high score if it matches the true distribution with respect to the observation y. In the algorithm, negative log-likelihood is used as the scoring rule.

Gaussian Process Regression

Gaussian Process Regression (GPR) is a flexible class of non-parametric models[21], where the non-linearity in the data can be modeled by incorporating different basis functions within the kernel to compute the covariance matrix. It can also include user-defined prior functions to take advantage of domain knowledge (which is crucial when data are scarce). The GPR prior is an infinite-dimensional, multi-variate distribution that can be completely described by a mean function $\big(m(x)\big)$ and a covariance function $k(x, x')$,

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{4}$$

The posterior predictive distribution is then obtained by computing the likelihood from the observed data within a Monte-Carlo framework. However, it can also be posed as an optimization problem to minimize the analytical negative log-likelihood. Once the learning process is complete, the posterior distribution not only provides a point estimate for the prediction but also a standard deviation of the prediction.

Active Learning

Active learning is a machine learning method that is used for optimal design of experiments by iteratively guiding the selection of the next unexplored data points to be acquired by using a suitable acquisition function[17]. These selected new data, when added to the already explored training data, will yield the maximum improvement in the model performance thus leading to a better model with fewer data points acquired. In this work, a pool-based Bayesian active learning method based on GPR as the base learner is used which selects the most useful samples from a pool of unlabeled samples. However, the traditional pool-based sequential active learning is not suitable for an iterative exploration of 9 - 12 *wt% Cr* ferritic-martensitic steel space, as each experiment is expensive and can take several days to years to complete. Instead of choosing one sample at each iteration, a batch mode is adapted which enables selecting multiple samples at each iteration. In Figure 1, a schematic of the active learning loop is shown to illustrate the inner working of the active learning cycle adapted in this work. The algorithm used for active learning for batch mode in this study is the same as the standard approach; except, during the query stage, the unpooled data are clustered using kmeans algorithm. At this point the variance reduction approach (i.e., selecting the most uncertain sample) is used to select one sample from each cluster. This approach is not only faster than the traditional sequential active learning, but it is also better in terms of *informativeness* (which represents the ability of the sample to reduce the generalization error in the adopted model and ensures lower uncertainty in the subsequent iteration), and diversity in the samples (meaning the data is the most dissimilar to the data already present in the training set). Since the data are clustered into several groups, each selected sample is diverse from another, and the variance reduction approach ensures that the samples are also very informative.

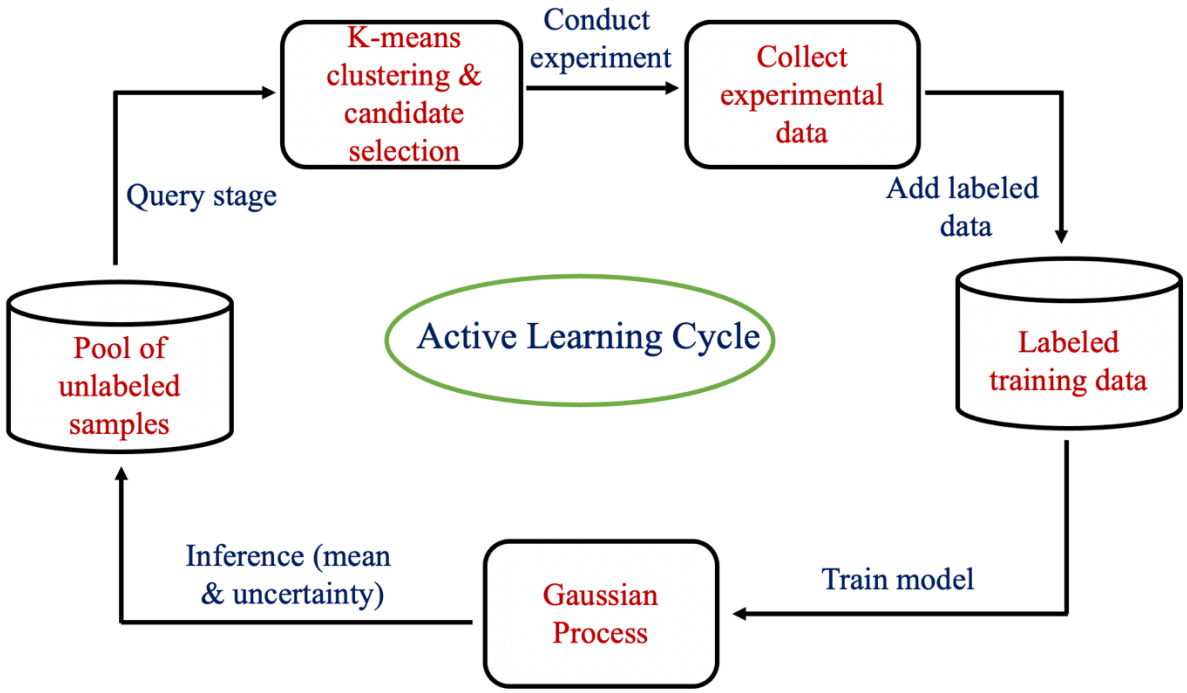

Figure 1: Active learning workflow of a batch-mode, pool-based method. In the query stage, candidates are selected from the pool of unlabled samples based on K-means algorithms and variance reduction approach. The selected samples are tested in the laboratory and added to the training data to make refined inference on the pool of unlabeled samples.

Methods

The dataset for 9 – 12 *wt% Cr* ferritic-martensitic steel was collected and compiled by the National Energy Technology Laboratory (NETL) as part of the Extreme Environment Materials (eXtremeMAT or XMAT), a DOE-funded project that seeks to reduce the time to bring new alloys to commercial readiness while at the same time providing the modeling framework necessary to describe its entire life in operation at all length scales. In the interest of brevity, the interested readers should refer to the article[13] that developed a non-probabilistic machine learning model for rupture life prediction for a detailed description of the dataset (n=875) and data preprocessing approach. The creep rupture life, or the target variable, is always positive, resulting in inherent

constraints that need to be incorporated into the machine learning algorithms. It is a particularly severe problem for probabilistic modeling as the response variable is allowed to have any values with some certain probabilities, albeit very small for values outside the 95% confidence interval. Due to this freedom in probabilisitc modeling, it is crucial to log-transform the data so that the predicted response variable is always positive. In Figure 2, the distribution of response variables in their original space and their transformed space is shown. This transformation also makes the data more normally distributed which helps tackle the heteroscedasticity problem to some extent.

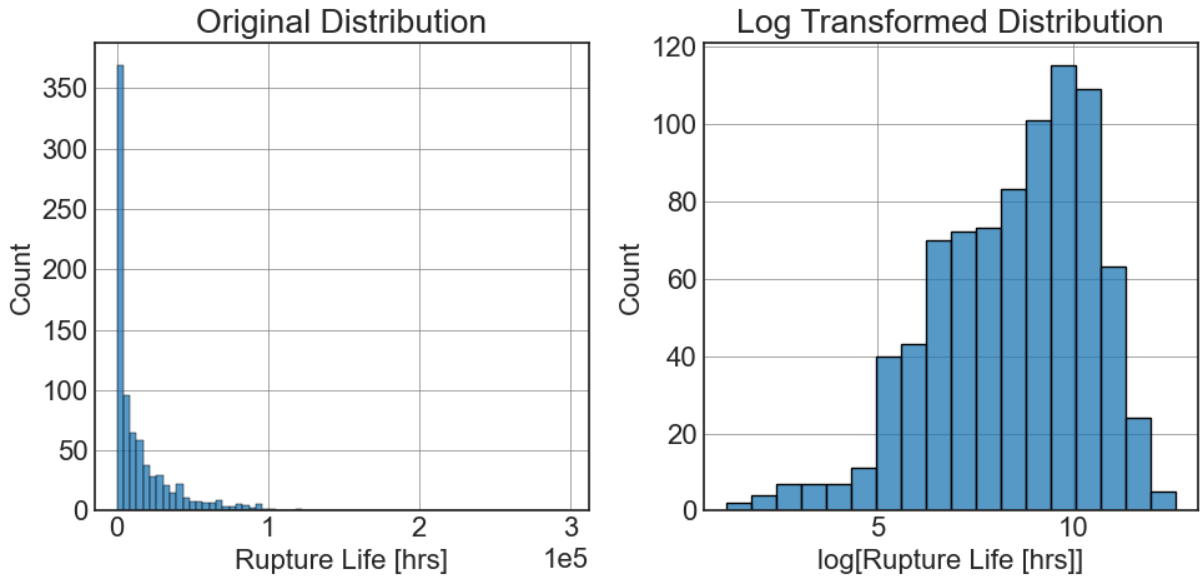

Figure 2: The distribution of rupture life in both original and transformed space.

For Quantile regression and NG Boosting, the predicted response variable was computed by taking the exponential of the prediction for all the quantiles. However, for GPR the mean and standard deviation of the log-transformed prediction of indivdual points must be transformed with the following formula to achieve the mean and standard deviation in the original space.

$$\mu_y = e^{(\mu_{\log(y)} + \frac{\sigma_{\log(y)}^2}{2})} \tag{5}$$

$$\sigma_y = \sqrt{\mu_y^2 \times \sigma_{\log(y)}^2} \quad (6)$$

Where $\mu_{\log(y)}$ and $\sigma_{\log(y)}$ are mean and standared deviation of the prediction in the log-transformed sapce, and μ_y and σ_y are the mean and standard deviation of the prediction in the original space, respectively. For Quantile Regression, the CatBoost[22] python package was used with scikit-learn API. For NG Boosting, the NGBoost[20] python package was used. To perform GPR, scikit-learn[23] python package was used. The code to reproduce the machine learning models and active learning results can be obtained from https://github.com/mamunm/uncertainty_quantification_creep_life.

To quantify the overall model performance for prediction, a five-fold, cross-validation scheme was used where in each iteration 80% data were used to train the model and 20% data were used to test the model performance. Pearson correlation coefficient (PCC) was used to measure the predictive performance and coverage (i.e., the fraction of data that lie within two (2) standard deviations) was used to quantify the reliability of the uncertainty estimate.

For active learning, 20% data were used as test data and then 20% of the remaining data were used as the initial training data. Kmeans clustering algorithm is used to fit the remaining data into 10 clusters. In each iteration, 10 new data points are added to the training data by selecting 1 point each from the 10 clusters. The correlation coefficient of the holdout test set is then calculated to determine the accuracy of the model at each iteration. The active learning technique is compared to a baseline random sample addition to indicate the relative efficiency of the active learning model.

Results and Discussion

Figure 3 illustrates the actual rupture life and the predicted rupture life with the 95% confidence interval for prediction on the hold-out test set. PCC for training and testing data is 0.980 ± 0.006 and 0.890 ± 0.050, respectively, with the coverage ranging from 95.34 ± 0.77 % and 80.74 ± 3.10 %, respectively. The accuracy of the model is quite satisfactory as discussed in[13]. In addition to that, the model can now provide a 95% prediction interval that contains about 80% of the test dataset. There are several problems associated with quantile regression that questions the reliability of the model for real-world applications.

The first and obvious problem is the lower prediction interval ($|y_{median} - y_{lower}|$) is greater than the higher prediction interval ($|y_{higher} - y_{median}|$) in the high rupture life area. Due to the inherent nature of the loss function of the quantile regression, it overestimates the lower prediction interval for this dataset as there are significantly more data in that region. Another problem is that in a few instances the predicted median is not within the prediction interval. Since three models are being optimized for median and two (2) standard deviations in the interquartile range, each model optimizes the parameters irrespective of the other models. As a result, inconsistency appears in the prediction interval and predicted median. The third and final problem is the confidence interval is not well-calibrated which questions the reliability of the uncdertainty estimate. For a well-calibrated model, 95% of the data should fall within two (2) standard deviations which is satisfied for the training data but only80.74 ± 3.10 % of the testing data are contained within two standard deviation leading to under-estimation of the confidence interval.

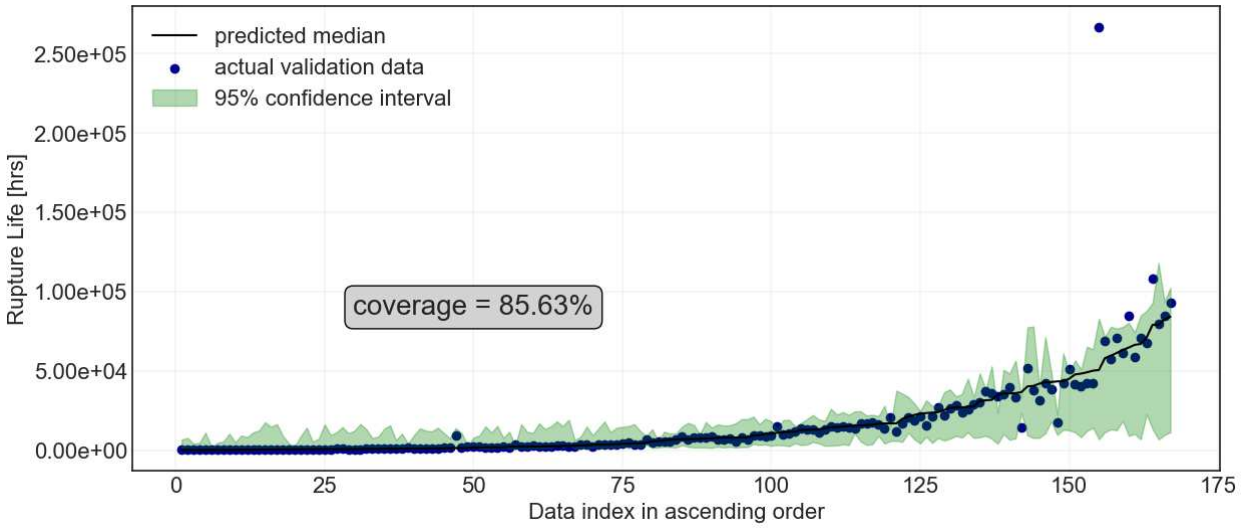

Figure 3: Actual and Quantile Regression predicted creep rupture life of ferritic-martensitic steels in ascending order of the predicted creep rupture life. The green area indicates the 95% prediction interval.

In Figure 4, the actual and NG Boosting predicted rupture life along with the prediction interval is shown for the hold-out test set. The PCC for the training and testing set is 0.980 ± 0.002 and 0.920 ± 0.030, respectively, with the coverage ranging from 98.47 ± 0.28 % and 84.08 ± 2.03 %, respectively. Surprisingly, the model prediction is more accurate than the non-probabilistic Gradient Boosting approach. Furthermore, the uncertainty estimate is better behaved than the quantile regression. The prediction interval always contains the mean and the uncertainty is higher where the data is scarce, leading to a more faithful model that can be used for an iterative exploration of the materials space. However, the uncertainty is not well-calibrated as evident by this coverage of the testing data, i.e., 84.08 ± 2.03 %.

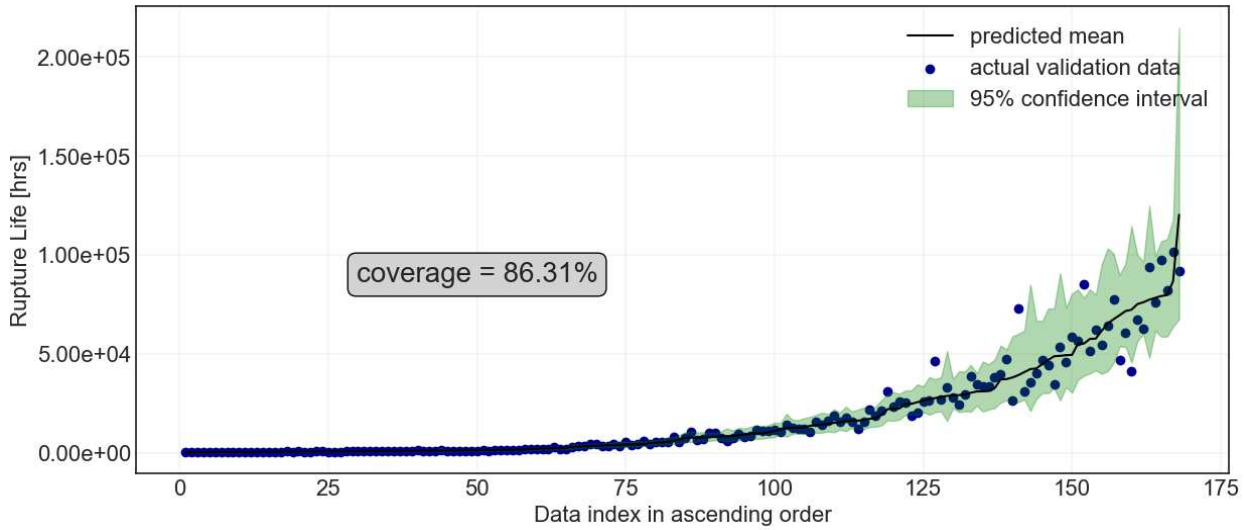

Figure 4: Actual and NG Boosting predicted creep rupture life of ferritic-martensitic steels in ascending order of the predicted creep rupture life. The green area indicates the 95% prediction interval.

Next, the GPR model was trained with an additive kernel consisting of a Matern kernel, a white kernel, and a dot product kernel. The Matern kernel models the short-range interaction present in the data. The white kernel adds noise to the diagonal elements of the kernel matrix to make it more robust and generalizable to noise. The dot product kernel models any linear trends between the data points. In Figure 5, the kernel matrix is illustrated for the training data. As expected, there is a high correlation between adjacent data points but the correlation slowly fades with increasing distance. Also, the data seem to be clustered naturally into about six (6) distinct clusters.

Next, Figure 6 illustrates the actual and predicted creep rupture life for the GPR along with the 95% prediction interval. The PCC for training and testing was set at 0.990 ± 0.001 and 0.970 ± 0.020, respectively, with the coverage ranging from 99.25 ± 0.18 % and 96.40 ± 1.52 %, respectively. Not only is the GPR highly accurate compared to the other two algorithms,

but it also has very high coverage which means more than 96% of the data lies within the 95% prediction interval, as it should be for a well-calibrated uncertainty estimation.

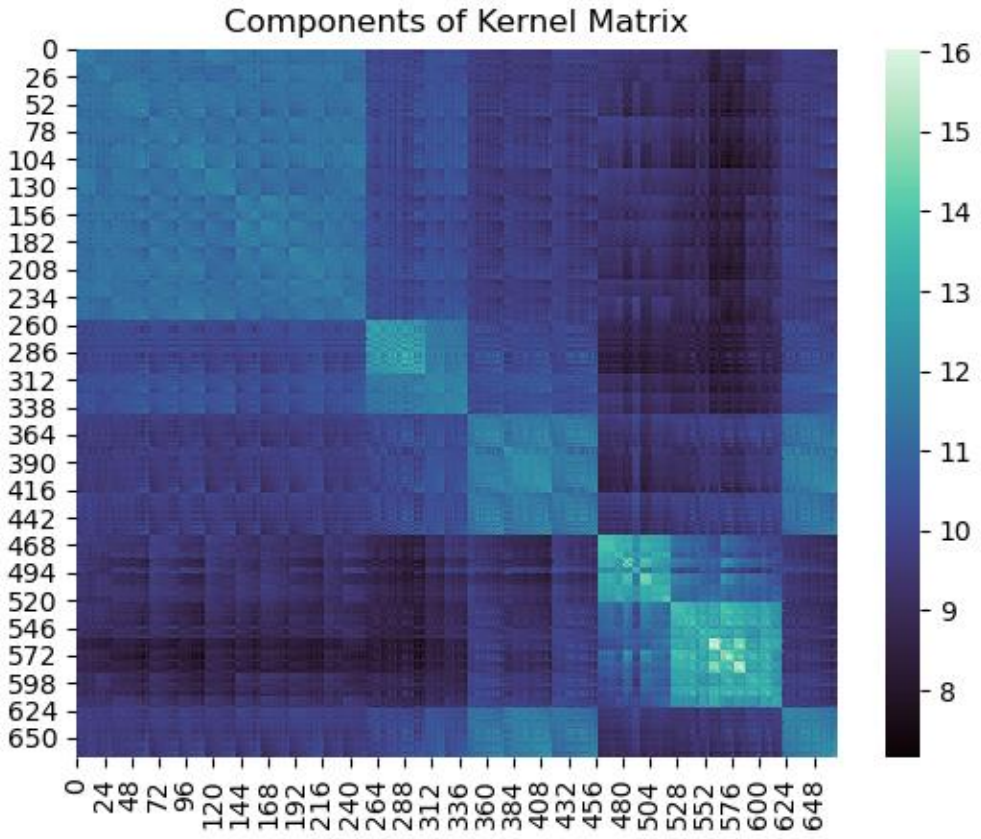

Figure 5: Componenets of kernel matrix for the additive kernel used in this study. The x and y axis represents the index of the samples and the color bar shows the magnitude of the kernel function.

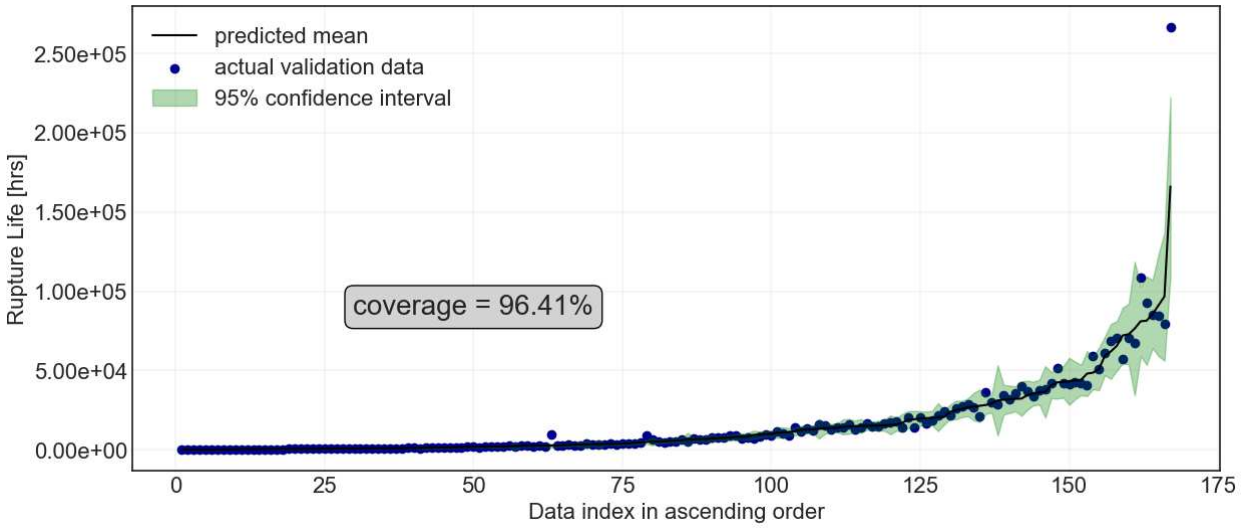

Figure 6: Actual and Gaussian Process predicted creep rupture life of ferritic-martensitic steels in ascending order of the predicted creep rupture life. The green area indicates the 95% prediction interval.

Active Learning for Iterative Exploration

The data used in this study were collected over 30 years of concerted efforts from government and non-government initiatives[14]. The overarching goal of probabilistic machine learning is to make reliable predictions for unknown alloys as well as to accelerate the data collection process to improve the model performance. This process exhaustively explores a material's space with minimal experimental effort. To this end, the uncertainty of prediction provides valuable information about the knowledge gap present in the dataset, and by acquiring the data with the highest uncertainty, significant improvement in the model's performance can be achieved. In Figure 7, the correlation coefficient for the hold-out test data for both the random and the active data collection processes are shown. Not surprisingly, the active learning process can readily find the data points that lead to the most improvement in test set performance. . This technique therefore can be used to design the experimentation necessary both to improve the model performance and to optimize the desired alloy properties. By incorporating this framework into

the experimental/computational data collection process, significant improvement in materials discovery can be achieved.

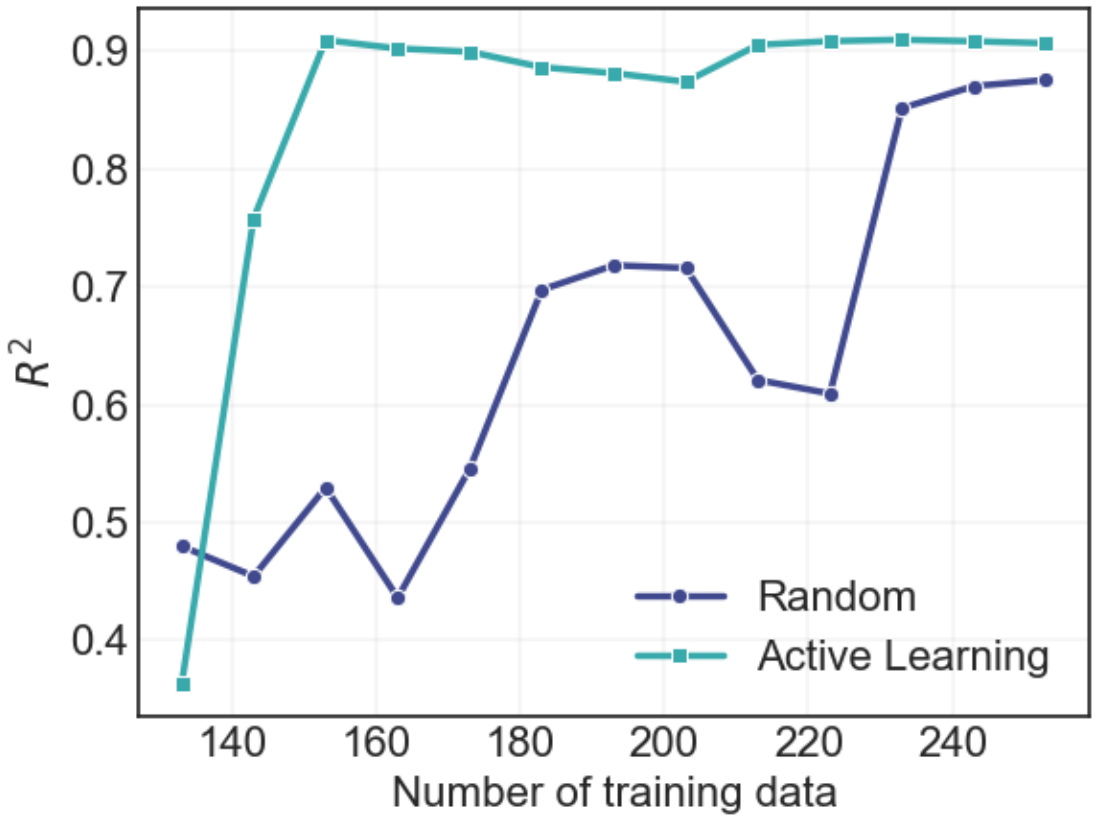

Figure 7: Correlation coefficient of the hold-out test set for both random data acquisiton and active learning algorithm.

Conclusion

In this work, three approaches are described to determine the uncertainty associated with machine learning prediction of creep rupture life of ferritic-martensitic steels for each data point. The advantages and disadvantages of each approach was described with the accuracy of the prediction based on coverage of the 95% prediction interval. From this effort the GPR was

identified as the best algorithm for both accurate inference and uncertainty estimation. Implementing this uncertainty estimation technique in the machine learning workflow will enable researchers to estimate the variance in the predicted values and to understand the risk inherent in the model. This technique will further help to ensure model interpretability by providing a 95% confidence interval for the predicted values. Finally, a simulated design of experiments was used to iteratively collect data using an active learning framework to accelerate data collection for reliable and informative data acquisition. Implementing this technique will enable researchers to more efficiently explore the alloy design space, and optimize the experimentally collected data points for improving the model predictions. These methods will ultimately improve model predictions to optimize the desired properties of the alloys.

Associated content

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the U.S. Department of Energy, Office of Fossil Energy, eXtreme environment MATerials (XMAT) consortium. This research used resources of the Pacific Northwest National Laboratory, which is supported by the U.S. Department of Energy.

Disclaimer

This manuscript was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process

Bibliography

1. Klueh, R. L. & Nelson, A. T. Ferritic/martensitic steels for next-generation reactors. *J. Nucl. Mater.* **371**, 37–52 (2007).
2. Klueh, R. L. *et al.* Ferritic/martensitic steels – overview of recent results. *J. Nucl. Mater.* **307**–**311**, 455–465 (2002).
3. Bischoff, J. *et al.* Corrosion of ferritic–martensitic steels in steam and supercritical water. *J. Nucl. Mater.* **441**, 604–611 (2013).
4. Hohenberg, P. & Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **136**, B864–B871 (1964).
5. Kohn, W. & Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
6. Mamun, O., Winther, K. T., Boes, J. R. & Bligaard, T. A Bayesian framework for adsorption energy prediction on bimetallic alloy catalysts. *npj Comput. Mater.* **6**, 177 (2020).
7. Rahman, A. Correlations in the Motion of Atoms in Liquid Argon. *Phys. Rev.* **136**, A405–A411 (1964).
8. Alder, B. J. & Wainwright, T. E. Studies in Molecular Dynamics. I. General Method. *J. Chem. Phys.* **31**, 459–466 (1959).
9. Chen, L.-Q. Phase-field models for microstructure evolution. *Annu. Rev. Mater. Res.* **32**, 113–140 (2002).
10. Boettinger, W. J., Warren, J. A., Beckermann, C. & Karma, A. Phase-Field Simulation of Solidification. *Annu. Rev. Mater. Res.* **32**, 163–194 (2002).
11. Steinbach, I. Phase-field models in materials science. *Model. Simul. Mater. Sci. Eng.* **17**, 73001 (2009).
12. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
13. Mamun, O., Wenzlick, M., Sathanur, A., Hawk, J. & Devanathan, R. Machine learning augmented predictive and generative model for rupture life in ferritic and austenitic steels. *npj Mater. Degrad.* **5**, 20 (2021).

14. Mamun, O., Wenzlick, M., Hawk, J. & Devanathan, R. A machine learning aided interpretable model for rupture strength prediction in Fe-based martensitic and austenitic alloys. *Sci. Rep.* **11**, 5466 (2021).

15. Hossain, M. A. & Stewart, C. M. A probabilistic creep model incorporating test condition, initial damage, and material property uncertainty. *Int. J. Press. Vessel. Pip.* **193**, 104446 (2021).

16. Romanov, V. N. Deep-freeze graph training for latent learning. *Comput. Mater. Sci.* **199**, 110757 (2021).

17. Kumar, P. & Gupta, A. Active Learning Query Strategies for Classification, Regression, and Clustering: A Survey. *J. Comput. Sci. Technol.* **35**, 913–945 (2020).

18. Meinshausen, N. & Ridgeway, G. Quantile regression forests. *J. Mach. Learn. Res.* **7**, (2006).

19. Koenker, R. & Hallock, K. F. Quantile Regression. *J. Econ. Perspect.* **15**, 143–156 (2001).

20. Duan, T. *et al.* NGBoost: Natural Gradient Boosting for Probabilistic Prediction. *CoRR* **abs/1910.03225**, (2019).

21. Rasmussen, C. E. Gaussian processes in machine learning. in *Summer school on machine learning* 63–71 (Springer, 2003).

22. Dorogush, A. V., Ershov, V. & Gulin, A. CatBoost: gradient boosting with categorical features support. *arXiv Prepr. arXiv1810.11363* (2018).

23. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).