Can Machine Learning and Explainable Artificial Intelligence Help to Improve an Expert Model for Predicting Thermomechanical Fatigue?

Stefan Glaser, Thomas Seifert, and Daniela Oelke

Offenburg University, Germany
firstname.lastname@hs-offenburg.de

Abstract. Machine learning (ML) models are increasingly used for predictive tasks, yet traditional data-based models relying on expert knowledge remain prevalent. This paper examines the enhancement of an expert model for thermomechanical fatigue (TMF) life prediction of turbine components using ML. Using explainable artificial intelligence (XAI) methods such as Permutation Feature Importance (PFI) and SHAP values, we analyzed the patterns and relationships learned by the ML models. Our findings reveal that ML models can be trained on TMF data, but integrating domain knowledge remains crucial. The study concludes with a proposal to further refine the expert model using insights gained from ML models, aiming for a synergistic improvement.

Keywords: explainable artificial intelligence (XAI), TMF life prediction

1 Introduction

Predictive models using machine learning (ML) are increasingly applied across various fields. At the same time, data-based models themselves are not new in many of these areas. There are already models that have been manually created by experts using domain knowledge and, where necessary, adapted to the available data using statistical methods. Both approaches have their pros and cons. ML models have the advantage of uncovering unknown relationships, but they can also inherit data biases, especially with limited training data. Expert models, on the other hand, have the advantage that they can be based on known and proven correlations.

This paper explores improving a TMF expert model with an ML model by training it on data of the turbine-blade nickelbase superalloy MAR-M247. We evaluate the ML models using explainable artificial intelligence (XAI) and exploratory data analysis to understand the patterns and relationships it learns, aiming to improve the expert model effectively, thus combining the strengths and weaknesses of both approaches.

This work differs from related previous work in that the focus is not on training the best possible ML model (see e.g. [1,2,3]), but on gaining new insights by comparing the approaches of the expert model and the ML model.

2 Domain background

High-temperature turbine components are exposed to start-up and shut-down cycles in which the material experiences high thermal and mechanical loadings. These cyclic loadings result in a progressive damage of the material and after a certain number of cycles to failure of the material. Different mechanisms are responsible for damage as fatigue crack growth (fatigue damage) that can be enhanced by creep of the material at higher temperatures (creep damage). For the assessment of the life of the turbine components, a model is required that allows engineers to predict the number of such start-up and shut-down cycles that the component can be used without failure. The number of cycles to failure N_f is the key result of models for TMF life assessment.

For the development of a TMF model and the fitting of the model to a certain material, different material tests are performed in which material specimens are cycled to failure using different load conditions, respectively. In isothermal low-cycle fatigue (LCF) tests, the specimens are exposed to constant amplitude cyclic mechanical loads at constant temperature, so that N_f is measured for different mechanical strain ranges and operation-relevant temperatures. In TMF tests, constant amplitude cyclic mechanical as well as cyclic thermal loads are applied to the specimen. Beside different mechanical strain and temperature ranges, different phase angle between the mechanical strain and the temperature are considered in the tests, since contributions of fatigue and creep damage and, thus, N_f strongly depends on the applied stress-temperature history the material is exposed to. In both tests, LCF and TMF, hold times at maximum load are often introduced as well. Since these material tests are time-consuming and expensive, usually the number of tests is rather small considering the number of factors influencing the TMF life.

Different features can be evaluated from the data of each test and used to generate a correlation to the number of cycles to failure N_f. We divide the available features into two categories: simple features and expert features (Table 1). Simple features are features that do not require any domain knowledge to create. They are either direct measurement results or simple combinations of features such as ratios or differences. The simple features include, for example, the maximum and minimum mechanical strain and the resulting mechanical strain range as well as maximum, minimum stress and the resulting stress range and stress ratio. When creating expert features, on the other hand, domain knowledge is required or additional information is included that is not part of the simple features.

In this work, a fracture-mechanics based model for TMF life prediction is considered that provides expert features. The first expert feature is the damage parameter, e.g. [4,5],

$$D_\mathrm{TMF} = \left(1.45\frac{\Delta\sigma_\mathrm{eff}^2}{\sigma_\mathrm{CY}E} + 2.4\frac{\Delta\sigma\Delta\varepsilon^\mathrm{p}}{\sigma_\mathrm{CY}\sqrt{1+3n'}}\right)F_\mathrm{creep} \tag{1}$$

that includes besides the simple features also elastic, plastic and creep properties of the material (Table 1) and that allows the contributions of fatigue and creep damage to N_f to be evaluated. The effective stress range $\Delta\sigma_\mathrm{eff}$ is depending on the stress ratio R according to the empirical equation [6],

$$\Delta\sigma_\mathrm{eff} = \left(0.35 + (2.2-R)^{-2}\right)\Delta\sigma, \tag{2}$$

and the function controlling creep damage $F_\mathrm{creep} \geq 1$ is a function of the stress-temperature history in the loading cycle. Using a fatigue crack growth law, the correlation

$$N_\mathrm{f} = \frac{A}{D_\mathrm{TMF}^B} \tag{3}$$

can be derived. The parameters A and B are fitted based on experimental data. This results in a linear relation in $log_{10} - log_{10}$ scale:

$$log_{10}\left(N_\mathrm{f}\right) = log_{10}\left(A\right) - B\;log_{10}\left(D_\mathrm{TMF}\right) \tag{4}$$

In the following, we refer to this calculation of N_f as the expert model or D_TMF regressor. Further expert features are the damage portions due to fatigue ($0 \leq D_\mathrm{fat} \leq 1$) and creep ($0 \leq D_\mathrm{creep} \leq 1$) adding up to a final damage of 1, i.e. $D_\mathrm{fat} + D_\mathrm{creep} = 1$. The latter is excluded because it is 100 % correlated to D_fat in the data.

Feature	Type	Name	Further information
maximum mechanical strain ε_max	simple	*MeStrainMax*	
minimum mechanical strain ε_min	simple	*MeStrainMin*	
maximum stress σ_max	simple	*StressMax*	
minimum stress σ_min	simple	*StressMin*	
stress range $\Delta\sigma$	simple	*DeltaStress*	$\Delta\sigma = \sigma_\mathrm{max} - \sigma_\mathrm{min}$
stress ratio R	simple	*R*	$R = \sigma_\mathrm{min}/\sigma_\mathrm{max}$
plastic strain range $\Delta\varepsilon^\mathrm{p}$	simple	*DeltaPlStrain*	$\Delta\varepsilon^\mathrm{p} = \varepsilon^\mathrm{p}_\mathrm{max} - \varepsilon^\mathrm{p}_\mathrm{min}$ $\varepsilon^\mathrm{p} = \varepsilon - \sigma/E$
maximum temperature T_max	simple	*TempMax*	
minimum temperature T_min	simple	*TempMin*	
hold time t_hold	simple	*HoldTime*	at maximum stress in LCF tests or temperature in TMF tests
heat time t_heat	simple	*HeatTime*	in TMF tests (0 s in LCF tests)
phase angle ϕ	simple	*PhaseFactor*	phase shift of cyclic mechanical and thermal load in TMF tests
damage parameter D_TMF	expert	*DTMF*	includes simple features and material properties
damage portion due to fatigue D_fat	expert	*Dfat*	$D_\mathrm{fat} = N_\mathrm{f,fat}/N_\mathrm{f} = 1/F^B_\mathrm{creep}$
Young's modulus E	expert	*E*	elastic property
cyclic yield stress σ_Cy	expert	*SigmaCY*	plastic property
cyclic hardening exponent n'	expert	*ROHardCy*	plastic property

Table 1: Used features; $N_\mathrm{f,fat}$ indicates the calculated fatigue life neglecting creep damage ($F_\mathrm{creep} = 1$)

The D_TMF model contains only a low number of material properties to be determined on the basis of experimental data. To this end, fatigue tests at different strain amplitudes and different temperatures are required and additionally some fatigue tests including hold times are necessary to calibrate the creep-function F_creep. Basically, the properties for a certain material can be determined based on a set of approx. 25 fatigue tests. The material properties of the D_TMF model for the nickel-base superalloy MAR-M247, determined on the basis for 22 test, are available from [6]. For this paper, available data for MAR-M247 from 12 literature sources (research reports and scientific papers, see Acknowledgement) is used. This database comprises 185 isothermal LCF tests and 117 TMF tests, both in the temperature range between 20 and 1150 °C and for mechanical strain ranges from 0.0007 to 0.014. The number of cycles to failure are in the range from 17 to 1 820 000 cycles. The fatigue life data is complemented by stress-strain data which is obtained for all tests using a time- and temperature-dependent plasticity model calibrated to MAR-M247 [6].

3 Can a ML model learn the relationships from the data?

In order to answer the question of whether and how well ML is able to recognize the correlations in the data, various ML models were trained on the data. Among others, ensemble methods such as Random Forest (RF), AdaBoost and XGBoost were tested, but also classic ML methods such as a Support Vector Regression (SVR), Decision Trees, Linear Regression or a K-Nearest-Neighbor (kNN) approach. Only a small multilayer perceptron (MLP) could be used as a neural network, as there was not enough training data available for a deep learning approach. All learning algorithms received scaled data. All prediction results (whether generated by the D_{TMF} regressor or an ML model) were post-processed so that values < 1 are set to 1. This procedure takes into account the fact that there can be no negative lifetime, but also ensures that the logarithm in the error function can be calculated (see below).

The evaluation was carried out with a 5-fold cross-validation in which the logarithmic Root Mean Squared Error (RMSL10E) was used as the evaluation measure:

$$RMSL10E = \sqrt{\sum_{i} (log_{10}(N_{\mathrm{f}i}) - log_{10}(\hat{N}_{\mathrm{f}i}))^2} \tag{5}$$

where $N_{\mathrm{f}i}$ represents the number of cycles to failure of data instance i as measured in the experiment and $\hat{N}_{\mathrm{f}i}$ is its corresponding predicted value. The reason for choosing a logarithmic error measure is that, from a domain perspective, the error factor is more important than the absolute value of the error. This is also the reason why we train all ML models on predicting the logarithm of the number of cycles to failure rather than the number of cycles to failure directly.

The D_{TMF} regressor achieves an RMSL10E score of 0.496 when the material properties given in [6] are applied to the data of all literature sources. All, or when using simple features all but two ML-based approaches, were able to achieve better results than the D_{TMF} regressor. For all further analyses, we will work with the RF (expert features: 0.292 / simple features: 0.297) and the SVR (0.324 / 0.333) which were the best approaches for both the expert features and the simple features. From this initial experiment, it can be concluded that it is possible to train a machine learning model on the data of all literature sources even if expert knowledge has not already been incorporated into the ML model via the features provided.

4 Exploration of the ML models with methods of XAI

Our goal is to examine whether the existing expert model can be further improved by exploring the patterns and relationships found by the data-based models in the data. Such analysis could lead to a deeper understanding of the problem and thus also advance development and research in this area. In the following, two methods of Explainable Artificial Intelligence (XAI) are used to make the functionality of the ML model generated by the RF transparent. Where necessary, additional exploratory analyses are carried out to gain an even better understanding of the existing relationships or to verify or falsify assumptions made with the help of the XAI methods.

4.1 XAI method 1: Permutation Feature Importance

Permutation Feature Importance (PFI) is a model-agnostic, global XAI technique. It measures the contribution of a feature to the performance of a model. This value is determined as follows: For a single feature, the values observed in the dataset are randomly swapped. The model is then queried with the new instances created in this way. The performance achieved is compared with the performance achieved on the unchanged instances. If the score hardly changes, this indicates that the feature was not important for the model's decision. Conversely, a significant deterioration in the prediction means that the model relies on this feature [7]. Note that PFI is not a feature selection method. It does not judge the predictive power of a feature for the task, but rather measures how important the feature is for the particular model being analyzed. When interpreting the results, it must be considered that unrealistic data instances can be created in the permutation step if feature dependencies exist. Moreover, the presence of correlated features may result in the observed importance being distributed among multiple features, which can prove challenging to interpret [8].

Figure 1 shows the results of the PFI for the models that were calculated on simple features. For the RF model, the feature *DeltaPlStrain* is by far the most important. It is followed by *StressMin* and - again at a distance - *MeStrainMax* and *HeatTime*. In case of the SVR, *TempMax* is reported as most important feature followed by *StressMin*, *MeStrainMin* and *DeltaPlStrain*. The notable discrepancies in the relative significance of the features between the SVR and RF models can be attributed to the inherent differences in their learning algorithms. While the RF model employs an iterative, greedy approach, considering each feature independently, the SVR model adopts a more comprehensive, global perspective on the data.

From a domain perspective, the importance of features representing the plastic and mechanical strain range is understandable. In simple fatigue life prediction models, these features are applied as in the Coffin-Manson model that correlates the plastic strain range (*DeltaPlStrain*) with the number of cycles to failure or the Coffin-Manson-Basquin model that correlates the mechanical strain range (*DeltaMeStrain*) with the number of cycles to failure. While these simple models result in a good description of isothermal lower temperature data, additional information is required to account for temperature-dependency. With the further important features, namely the minimum stress (*StressMin*) and the maximum temperature (*TempMax*), the ML model seems to acquire information on temperature-dependency. The lower the absolute value of the minimum stress, the higher the temperature at minimum stress.

For the RF model based on expert features (see Figure 2), D_{TMF} is by far the most important feature. It is followed at a far distance by *Dfat*, *StressMin* and *R*. In case of the SVR model, *Dfat* is reported as most important feature, tightly followed by D_{TMF} and at some distance again *StressMin*.

From a domain perspective it is not surprising that the features related to the damage parameter and the damage mechanisms are considered important. D_{TMF} is seen as particularly important showing that expert knowledge has actually been used to create a feature that has a high information content with regard to the TMF life prediction. The damage portion due to fatigue (*Dfat*) is also important. For the RF, the PFI plot almost gives the impression that D_{TMF} alone is sufficient for the prediction. However, additional experiments showed that it is not possible to achieve good results with a RF model if only the D_{TMF} value is provided. Still it can be concluded that the RF considers it a very important feature.

Again, the minimum stress (*StressMin*) appears as an important feature. This is interesting from a domain perspective, as the stress information is already considered in the D_{TMF} feature (which is the basis for the expert model) via the stress range (*DeltaStress*) $\Delta\sigma = \sigma_{\max} - \sigma_{\min}$ and the stress ratio (*R*) $R = \sigma_{\min}/\sigma_{\max}$.

None of the ML model considers the hold time (*HoldTime*) as important feature. Hence, effects of hold times, i.e. reduced life with longer hold times at higher temperatures due to creep damage, are not described. Indeed, only around 5 % of the tests in the experimental database contain hold times that show an effect in the fatigue live. While for the calibration of the expert model only a few (a minimum of two) tests with different hold times affecting the number of cycles to failure is required, modelling with ML may require more information for a reasonable description of the data. Hence, the training of the ML model requires a database with different tests than what is required for the calibration of the expert model. For the calibration of the ML model more tests with hold time are necessary such that the hold time becomes an important feature.

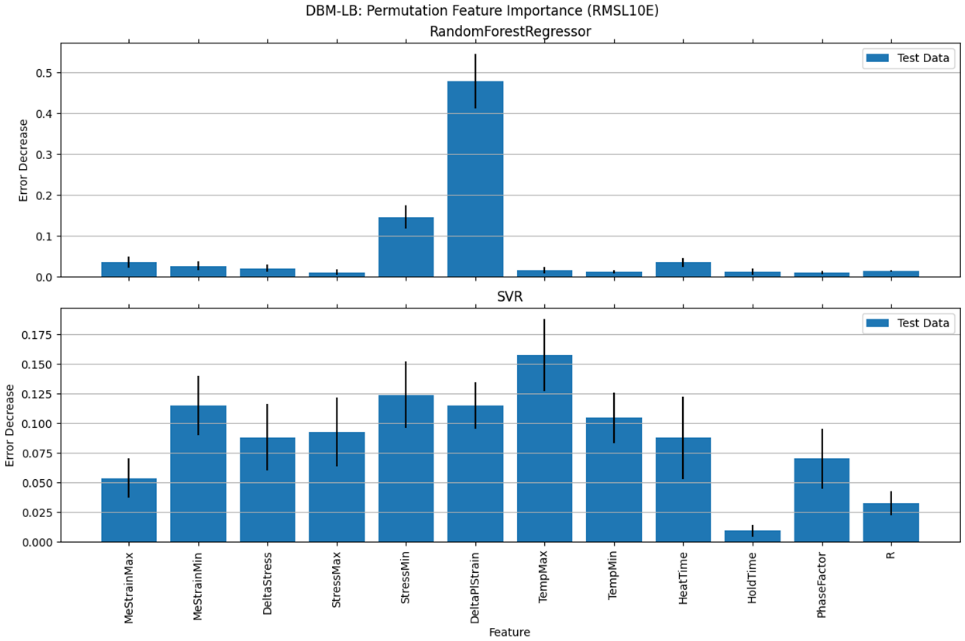

Fig. 1: Permutation Feature Importance (test data), Simple Features

4.2 XAI method 2: SHAP (SHapley Additive exPlanations) values

SHAP is an approximation method for calculating Shapely values. The methodology for calculating Shapely values has its origins in the field of game theory. The aim of the method is to determine the contribution of each player to the overall result. When applied to an ML model, the goal is to measure the contribution of each feature to the deviation of the predicted value from the expected value. In a regression task, the expected value is the average value of the target variable in the training data set [9].

To estimate the contribution of a feature to the overall result, a new model is trained for each possible subset of features. The deviation of the prediction quality with and without the feature gives an indication of its importance and, as a weighted sum of the marginal contributions, gives the contribution of the respective feature for the instance under consideration [9].

Since calculating Shapely values is a very computationally intensive process, several approximation methods have been developed to speed it up. We used KernelSHAP for

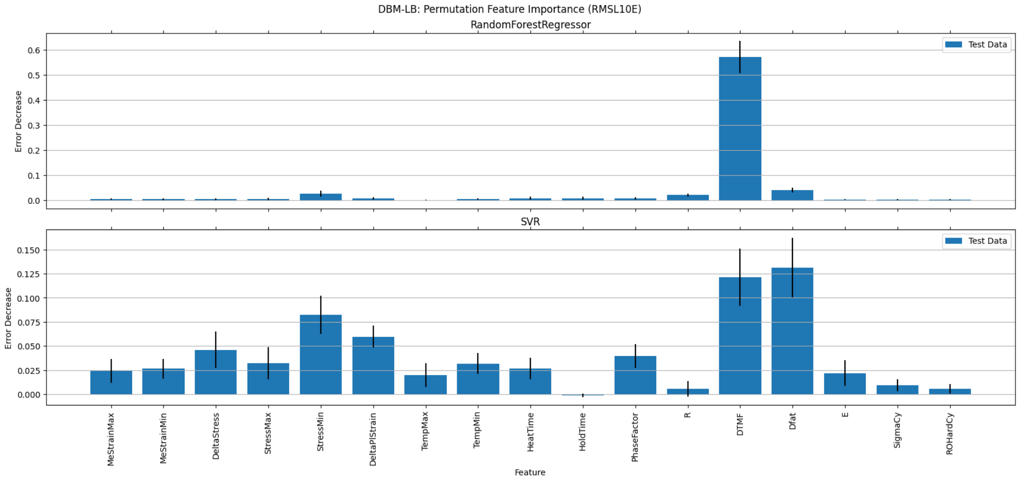

Fig. 2: Permutation Feature Importance (test data), Expert Features

the SVR models and TreeSHAP for the analysis of our RF models. TreeSHAP can only be applied to tree-based models and has the advantage of being significantly faster and handling feature dependencies well [8]. Nevertheless, in this paper we will only report the results for the SVR. The hierarchical structure of the tree-based models in the RF has the effect that some features are only applied to a small subset of the dataset. As a consequence, the results of the plots are difficult to interpret because it is unclear whether the described effect applies to all data or only to a small subset.

SHAP is a local method of XAI that makes a separate statement for each instance about how important each feature was to the decision made. In summary visualizations such as the Beeswarm plot (see Figure 3), each feature is represented as a line, with dots marking the SHAP value of a single data instance for that feature. Color indicates the value of the feature, making it possible to analyze the effect of the feature value on the target value. In this way, it is possible to get an impression of the global importance of the features - while retaining the advantage that the sometimes large differences for the individual instances are not lost by the formation of aggregation values.

According to the beeswarm plot for the SVR model based on simple features (Figure 3), a higher *StressMin* results in a higher life. This is especially true for tests with low plastic strain range and, thus, low D_{TMF} values, where only the stress range and the stress ratio are relevant. A higher minimum stress indicates a lower stress range resulting in a higher number of cycles to failure. However, higher *StressMin* can also reduce the life if the same stress range is applied and higher stress ratios occur.According to the SHAP analysis, lower values of the stress ratio R result in lower fatigue lives. From a domain perspective, the opposite should be true. Interestingly, this is only a global effect. The SHAP dependence plot (not shown here due to space limitations) for *StressMin* colored by R shows that higher R values tend to reduce fatigue life as expected, at least for higher *StressMin* values. However, the observed global effect again indicates that the database for the training of the ML model needs to be modified compared to the calibration of the expert model (i.e. more tests with different stress ratios). The XAI evaluation considers each feature separately. Changing *StressMin* (as well as the *StressMax*) should also change *DeltaStress* and stress ratio R that are not considered in the XAI evaluation. This could lead to uncertainties in the evaluation.

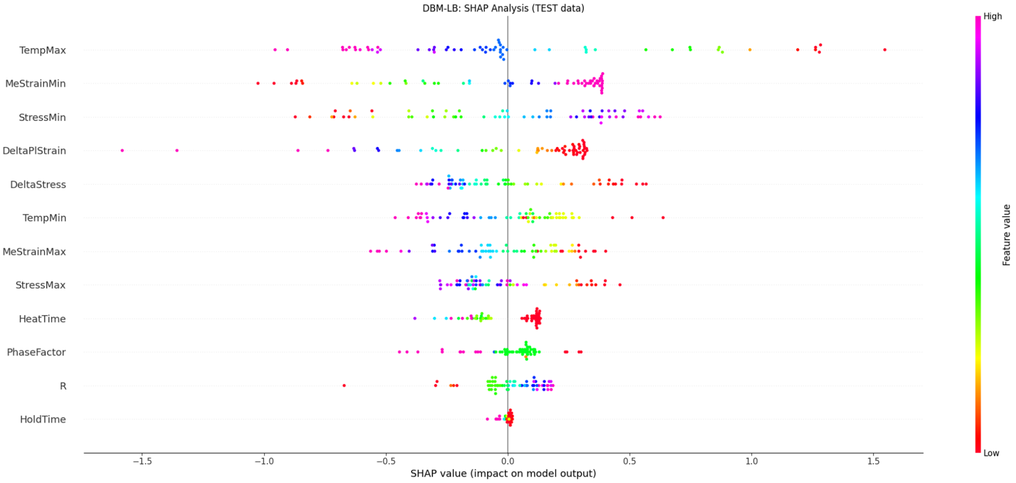

Fig. 3: SHAP beeswarm plot, SVR (test data), Simple Features

The SHAP plot for the expert features (not shown) did not provide additional insights and is therefore not shown here.

5 Conclusion

In summary, it can be said that the ML models were able to represent the data well. The analysis with the XAI methods showed that many of the relationships that the ML models detected are plausible from a domain perspective. Where this was not the case, this could often be explained by the fact that there was a bias in the data, as the experiments did not cover all possible cases or at least did not provide sufficient data for all cases. One lesson learned is thus that different data is required for training an ML-based model than is the case for a data-based expert model.

On the other hand, the analysis also revealed the limitations of the XAI methods. In particular, the fact that both methods assume that there are no dependencies between features limited the analysis, as there are many such dependencies in our dataset.

The principal trends identified by XAI are also predicted by the expert model in this way. This can also be seen as confirmation of the model derived by experts from the data. On the other hand, the additional use of further features by the ML model has also raised the question of whether this could be a key to further improvements of the expert model. It is possible that oversimplified assumptions are made when deriving the expert model, for example when calculating the effective stress range from the stress ratio or when assuming power-law hardening of the material via the hardening exponent n'. However, a direct ”hint” as to how the expert model should be modified cannot be derived from the XAI results.

In this first analysis, we did not succeed in identifying the relationships which the ML models discovered in a way that they could be directly used to improve the expert model. Therefore, as a next step we intend to the train a boosting model that directly builds on the D_{TMF} regressor and could give hints on what to add.

Acknowledgement

The authors acknowledge Dr.-Ing. Christoph Schweizer of the Fraunhofer Institute for Mechanics of Materials IWM for providing digitized LCF and TMF material data from literature sources.

References

1. Srinivasan, V., Valsan, M., Bhanu Sankara Rao, K., Mannan, S., Raj, B.: Low cycle fatigue and creep–fatigue interaction behavior of 316l(n) stainless steel and life prediction by artificial neural network approach. International Journal of Fatigue **25**(12) (2003) 1327–1338
2. Liu, Y., Wu, J., Wang, Z., Lu, X.G., Avdeev, M., Shi, S., Wang, C., Yu, T.: Predicting creep rupture life of ni-based single crystal superalloys using divide-and-conquer approach based machine learning. Acta Materialia **195** (2020) 454–467
3. Zhang, X.C., Gong, J.G., Xuan, F.Z.: A deep learning based life prediction method for components under creep, fatigue and creep-fatigue conditions. International Journal of Fatigue **148** (2021) 106236
4. Seifert, T., Riedel, H.: Mechanism-based thermomechanical fatigue life prediction of cast iron. part i: Models. International Journal of Fatigue **32**(8) (2010) 1358–1367
5. Seifert, T., Schweizer, C., Schlesinger, M., Möser, M., Eibl, M.: Thermomechanical fatigue of 1.4849 cast steel – experiments and life prediction using a fracture mechanics approach. International Journal of Materials Research **101**(8) (2010) 942–950
6. Serrano, L., Scholz, A., Müller, F., Berger, C., Schweizer, C., Schlesinger, M., Eckmann, S.: Tmf crack behavior: Crack behavior under anisothermal loading conditions – calculation method for nickel-based alloys (in german). Technical report, Forschungsvereinigung Verbrennungskraftmaschinen e.V., project Nr. 959, issue 945 (2011)
7. scikit-learn User Guide: Permutation feature importance `https://scikit-learn.org/stable/modules/permutation_importance.html`, accessed: 25.7.2024.
8. Molnar, C.: Interpretable machine learning - a guide for making black box models explainable (2024) `https://christophm.github.io/interpretable-ml-book/`, accessed: 18.9.2024.
9. Cooper, A.: Explaining machine learning models: A non-technical guide to interpreting shap analyses (2021) `https://www.aidancooper.co.uk/a-non-technical-guide-to-interpreting-shap-analyses/`, accessed: 25.7.2024.