Received: 3 June 2022 | Revised: 13 July 2022 | Accepted: 19 July 2022

DOI: 10.1111/ffe.13800

SPECIAL ISSUE ARTICLE

FFEMS Fatigue & Fracture of Engineering Materials & Structures WILEY

Comparison of machine learning and stress concentration factors-based fatigue failure prediction in small-scale butt-welded joints

Moritz Braun | **Leon Kellner**

Institute of Ship Structural Design and Analysis, Hamburg University of Technology, Hamburg, Germany

Correspondence
Moritz Braun, Institute for Ship Structural Design and Analysis, Hamburg University of Technology, Am Schwarzenberg Campus 4(C), Hamburg D-21073, Germany.
Email: moritz.br@tuhh.de

Abstract

Fatigue behavior of welded joints is significantly influenced by numerous factors, for example, local weld geometry. A representative quantity for the influence of the notch effect created by the local weld geometry is the *stress concentration factor* (SCF). Thus, SCFs are often used to estimate fatigue failure locations and fatigue strength; however, this simplifies the mutual effect of other influencing factors. Consequently, fatigue strength estimates for welded joints may deviate from experimental results. Machine learning techniques offer an alternative to traditional fatigue assessment approaches based on SCFs. This study presents a comparison of failure location predictions and number of cycles to failure for 621 fatigue tests of small-scale butt-welded joints. In addition, an understanding of importance and mutual influence of the factors is desired. We used *gradient boosted trees* in combination with the *SHapley Additive exPlanation* framework to identify influential features and their interactions.

KEYWORDS

explainable AI, fatigue life prediction, fatigue strength, gradient boosted trees, machine learning models, SHAP

1 | INTRODUCTION

The fatigue behavior of welded joints is a complex phenomenon driven by numerous factors, for example, local weld geometry, load level, or material strength. In principle, these factors can be grouped such that they are related to loading, geometric features of the welded connection, material properties, and environmental effects. In addition, these factors interact in a complex manner, which makes it difficult to separate effects. For example, geometrical features vary on the different sides of weld seams. Hence, local stress concentrations differ for the different weld transitions, yet current design practices primarily rely on simplifications such as idealized constant geometries instead of actual local weld geometries. Consequently, fatigue lifetime predictions often deviate significantly between model and experiment.

Studies have shown a significant influence of local weld geometry on the fatigue strength of welded joints, as it influences stress concentrations; see previous studies.[1–4] Recent studies show that fatigue cracks often initiate at locations with high stress concentrations.[5–7] A

representative quantity to describe the magnitude of local stress concentrations is the *stress concentration factor* (SCF),[8] which is the quotient of local notch stress and nominal stress.[9]

Due to recent advances in automated measurement techniques, virtually unrestricted measurements of local weld geometries have become feasible. Finite element (FE) simulations can make use of such data, for example, through exactly replicating the weld geometry, yet 3D simulations of long seam welds or even more complex 3D fatigue crack growth simulations are currently computationally too demanding to be feasible for fatigue life predictions. Machine learning (ML) techniques present an alternative to quickly process large amounts of multivariate data, for example, for different geometrical features; however, such techniques need to be verified and compared with state-of-the-art fatigue assessment approaches.

This study presents a comparison of ML techniques to analyze a large number of fatigue tests performed on small-scale butt-welded joint specimens with an assessment based on SCFs. The aim is to predict fatigue failure locations (one of the four weld toes) and fatigue strength (in terms of cycles to failure). To this end, we combined local weld geometry data obtained from high-resolution laser scans of butt joint specimens with the corresponding fatigue tests results and analyzed the data with a ML model based on *gradient boosted trees*. In addition, this study seeks to improve the understanding of driving factors for fatigue of welded joints. For this purpose, the *SHapley Additive exPlanation* (SHAP) framework was applied to assess the mutual influence of the various influencing factors on the prediction of fatigue failure locations and fatigue life and to rank them by impact.

For the comparison with state-of-the-art fatigue assessment, we also investigate whether the SCF is a useful quantity to predict fatigue failure locations. To achieve this, the weld geometry data were used to determine the distribution of SCFs along the four weld seams of each specimen. Next, the maximum SCF of each specimen is used to estimate failure locations. Finally, to determine whether the SCFs improve the accuracy of the ML model, tests are performed with and without using SCFs as input features.

This study is organized as follows. In Section 2, the two methods (ML and based on SCFs) to estimate fatigue failure locations and fatigue life are introduced, as well as the SHAP framework. Section 3 presents the fatigue test data and the measured weld geometry data. The results obtained by the two methods are described in Section 4. Finally, the main findings of this study are discussed in Section 5.

2 | METHODS

2.1 | Fatigue assessment by means of SCFs

With the development of weld geometry measurement systems, the characterization of stress concentrations at welded joints has seen significant progress; compare earlier research.[6,10–14] To determine SCFs at weld transitions, weld geometries are typically idealized as two-dimensional geometries, which are described by geometrical properties such as notch opening angle, notch radius, and weld height (see Figure 1). Based on such simplified descriptions of the weld geometry, it is possible to determine SCFs, for example, by means of parametric formulas, for example, Anthes et al.[15] An alternative is the direct calculation by FE simulations. Although the latter has proven to be more reliable, parametric formulas are still often applied to this day, yet various recent studies showed that this often leads to incorrect estimates of the actual SCFs at welded joints.[14,16–19] This is particularly true for complex weld shapes such as weld toes with undercuts; see Ottersböck

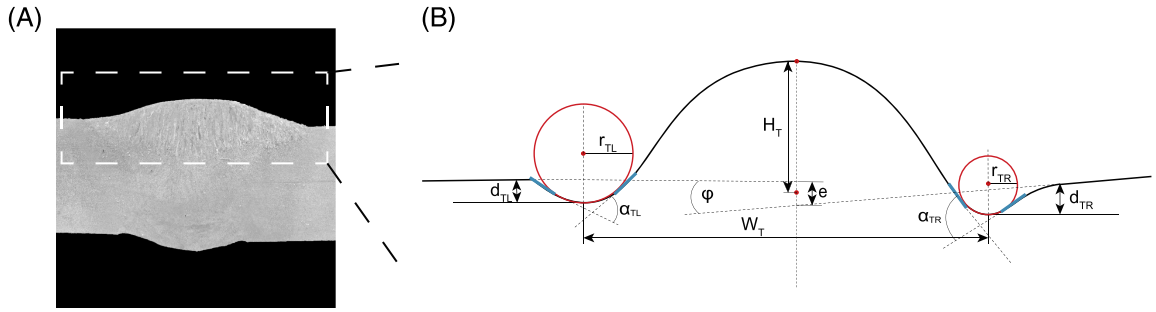


FIGURE 1 (A) Micrograph of a specimen with (B) schematic presentation of the top side geometry of butt-welded joints. Adopted from Braun et al.[23] [Colour figure can be viewed at wileyonlinelibrary.com]

et al.[17] In contrast, performing complex three-dimensional FE simulations or many two-dimensional simulations for various locations along weld seams is computationally demanding. Thus, novel methods focus on polynomial regression with coupling terms (PRC) and artificial neural networks (ANNs) to determine stress concentrations based on weld geometry measurements; see other studies.[20–22]

By performing large numbers of parametric FE simulations, Oswald and co-workers[20–22] showed that parametric SCF formulas sometimes lead to high errors. Additionally, parametric SCF formulas are typically limited to certain parameter combinations. Once these models (PRC and ANN) are set up, no additional time-consuming FE simulations are required. In principle, such ANNs are based on the same inputs as the parametric SCF formulas (typically geometrical parameters, for example, weld toe radius, flank angle, and plate thickness), one or two hidden layers, and one or two outputs of SCFs (i.e., for tension and bending loading).[11]

In the current study, a deep neural network (DNN) for butt joints with weld toe failure presented in Braun et al[9] is applied to determine the SCFs (for tension and bending loading) along the weld seams of the fatigue test specimens. A DNN is a subtype of an ANN with several layers. The SCFs are then used to estimate failure locations (one of the four weld toes) based on the determined maximum SCF of each specimen. In comparison, a ML model—recently presented in Braun et al.[23]—is applied to perform the same task. To determine whether the SCFs improve the accuracy of the ML model, tests are performed with and without using SCFs as input features.

2.2 | ML models

ML models are increasingly used to predict fatigue of components and parts.[24] Among ML methods, ANNs are the most common type for failure assessment[25]; however, investigations for welded joints are limited; see, for example.[23,25–29] Typical input variables for fatigue life prediction with ML were the respective load information (including nominal stress) as well as material, defect classes, and SCFs. For example, Karakaş and Tomasella,[30] as well as Kalayci et al[31] used information about the applied stress amplitudes, stress ratios, and SCFs to estimate a number of cycles to failure and to derive stress-life curves.

The features we used exhibit various nonlinear interactions. Moreover, the data contain nonphysical or missing values. This can make the application of classic statistical tools difficult, whereas ML models are especially useful under such circumstances.[32] Therefore, we used ML models to predict the fracture location and lifetime (i.e., the number of cycles to failure) of welded specimens (see Figure 2).

We employed gradient boosted trees for several reasons: Tree-based algorithms perform well on tabular data of this size.[33,34] Moreover, this type of model was successfully applied on similar data in an earlier study.[23] In particular, the XGBoost implementation[35] of gradient boosted trees was used due to its good integration with the explainability framework SHAP. The following hyperparameters were applied: 250 boosting iterations (i.e., the maximum number of trees), a learning rate of 0.07, a maximum number of leaves of 8, and a subsample size of 50%. Both models were applied with the same

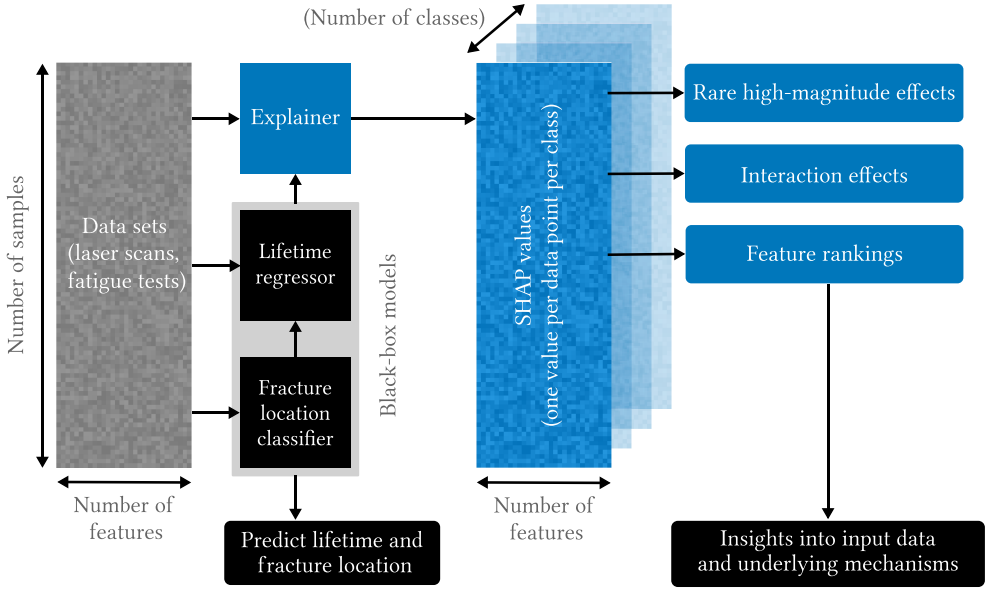


FIGURE 2 We combine two datasets to train gradient boosted decision trees to predict number of cycles to failure and fracture location of fatigue tests. The number of cycles regressor also receives the predicted fracture location from the classifier as input. Both models are combined with an additional explainer model to generate SHapley Additive exPlanation (SHAP) values. These can be used to understand the model structure. Based on Lundberg et al.[33] [Colour figure can be viewed at wileyonlinelibrary.com]

hyperparameters. The parameter values loosely follow the suggestions in previous studies.[36,37]

The data were split into training (80%) and test data (20%). To show the generalizability, four-fold cross-validation (CV) was used on the training data (which, for this purpose, is split again into per-fold training and test data).[38] The test data remain unseen by the algorithm until final validation. To prevent overfitting, for example, indicated by the detection of patterns in the noise of the data, early stopping was used. This means that the training stopped if the model performance did not increase over three rounds of training.

Two models were trained: first, a fracture location classifier which predicts on which of the four sides of the weld the failure crack initiated (top left/TL, top right/TR, bottom left/BL, bottom right/BR). The softmax function is used as an objective and the multiclass logloss as an evaluation metric for early stopping. Second, a regression model was used to predict the number of cycles to failure. In this case, squared error was used as an objective and root mean squared error as an evaluation metric for early stopping. For more details on these metrics and objectives, respectively, see, for example, earlier research.[35,37,39] In both cases, features whose values are unknown a priori are dropped before training. That is, the classifier did not have the number of cycles as input and the regressor did not have access to the fracture location.

2.3 | Explainability

To understand why a model has made a prediction is a key challenge, in particular for ML models. The more complex the patterns in the data, the more sophisticated the ML model must be to detect these patterns. Yet the more sophisticated a model is, the less explainable it is; that is, its decision rules are unclear. This black-box model character limits user trust and usefulness of models. It also makes it difficult to debug and improve models. Ultimately, black-box models impede knowledge discovery. Hence, the trade-off is not only between accuracy and complexity of the models; a third dimension is explainability.[40,41] Analogously, many data-driven studies are not only interested in the "best" model (i.e., with a high accuracy) but also seek to identify which features drive model predictions through, for example, feature importance rankings.[42,43]

In response to this, several tools under the umbrella term of *eXplainable AI* (XAI) have been developed. Different approaches exist;[44] here we employed an additional explanation model linked to the original ML model and the input data. More specifically, we used SHAP values which are based on Shapley values from game theory.[33,45,46] Imagine a game with a coalition of agents, where each agent contributes differently to the overall payoff. Shapley values consider the marginal contribution of an agent, that is, the difference in payoffs the coalition would receive with and without the agent. Since the coalition could be built according to different sequences, that is, agents could join or leave the game at any time, the marginal contribution is averaged over all sequences. Instead of payoffs and agents, SHAP values attribute model outcome to input features. They explain the difference between the expected value E, which a random model would predict, and the actual model prediction.

For the regression model, SHAP values have the same unit as the output. For the classification model, the SHAP values are the log-odds per class. SHAP unifies other XAI approaches[47,48] and, contrasting other more heuristic approaches, has a solid theoretical foundation.[49] Here, what is termed a SHAP *TreeExplainer* was used as an explanation model. It enables a fast and exact computation of SHAP values through exploitation of the structure of tree-based models.

2.4 | Data scrubbing

The general data processing is depicted in Figure 3. The raw data from two datasets were used; see Sections 2.1 and 3). Initially, the laser scan data were extended with SCFs for each slice. SCF values were replaced with *NaN* (not a number) when they were below the 5th percentile and above the 95th percentile with respect to all values. This was done since it is expected that there are some outliers in the laser scan measurements.[9] All other feature values were replaced with NaN if they were outside of realistic bounds.

Next, the fatigue test (butt joint) data include per-specimen test results and further measurements such as the temperature during tests. These features were extended with maximum and median SCF values per side (top left/right, bottom left/right) or per specimen (features not linked to any weld toe). All geometry data, for example, weld toe radii, were computed by taking the median of the laser scan data for each side. This median geometry data were also added to the butt joint dataset. The median was taken instead of the mean as it is less sensitive to outliers.[9,13]

Lastly, before training the models, invalid tests were removed from the data; for example, tests that were aborted without failure (run-out tests). Categorical features were replaced with numerical features with ordinal encoding instead of one-hot (also called one-of-K)

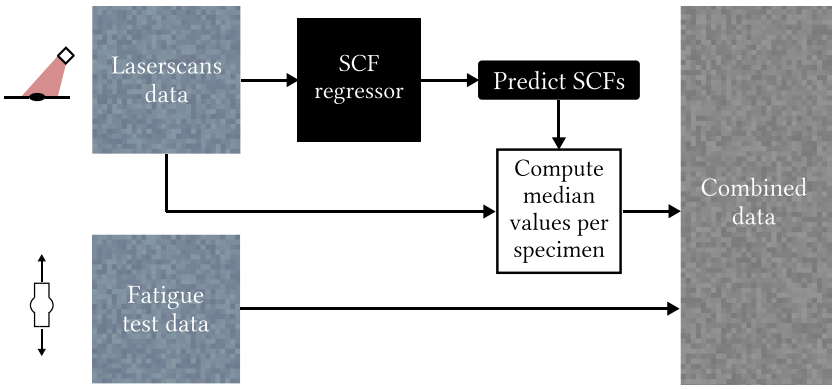


FIGURE 3 Two datasets are preprocessed and combined. The laser scans data describe the geometry through about 300 2D slices per specimen. These data are used to predict stress concentration factors (SCFs) per slice and (possible) fracture location. Lastly, median (and for SCFs also maximum) values are computed and combined with the fatigue test data [Colour figure can be viewed at wileyonlinelibrary.com]

encoding since this type of encoding was applied successfully in a previous study.[50]

3 | FATIGUE TEST DATA

In total, the data of 621 fatigue tests on small-scale butt-welded steel joints made by different fusion and laser welding methods have been collected from previous studies[51,52] and from unpublished results. All tests were performed uniaxially and force-controlled in the high-cycle fatigue regime. In addition, some tests were performed at subzero temperatures down to −50°C. Of all tests, about 10% are run-outs, which are marked with an arrow in Figure 4. To present the original data as obtained from the tests, no correction for misalignment effects or stress ratio (R) is applied. For the later fatigue assessment by the ML models, no corrections were applied either, as the goal is to account for this effect within the ML model directly.

As expected, the majority of data exceed the two design curves for butt-welded joints according to the recommendations of the International Institute of Welding.[53] FAT80 is the general fatigue design curve for butt joints, and FAT90 is permitted for high-quality butt joints made in shops in flat position.

The geometrical variations and misalignment in the investigation (see Figure 1) were determined prior to testing for all specimens. For the measurement of axial and angular misalignment, a test setup with dial gauges was used; see Braun,[50] and the local weld geometry was obtained by laser triangulation and postprocessing by means of the curvature method; see earlier research.[13,14] Figure 5 presents the medians of the measured geometrical and SCF parameters that are used as input for the ML models. Herein, the data are summarized as violin plots—showing the distribution density smoothed by a kernel density estimator.

The data are separated based on the different failure weld transitions or top and bottom sides. The median

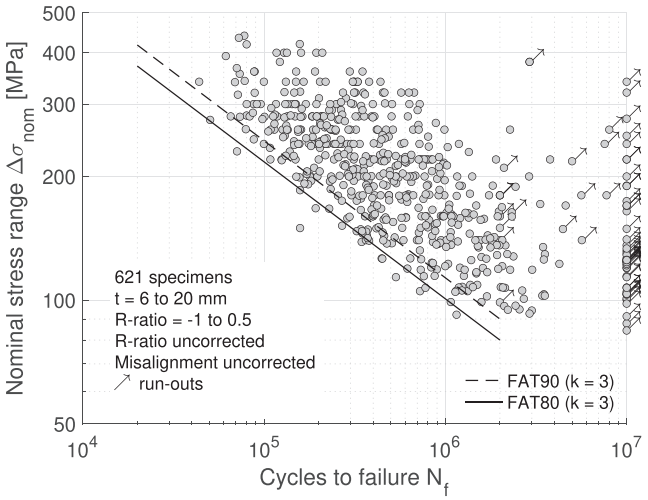


FIGURE 4 Nominal stress fatigue test results of small-scale butt-welded joints with fatigue design curves according to the recommendations of the International Institute of Welding[53]

weld toe radii were in almost all cases larger than 1 mm and only in a few cases larger than 4 mm. Interestingly, the smallest weld toe radii are observed for the bottom right weld toes. The majority of weld toe angles lie between 10° and 50°, with higher angles for the bottom weld toes. This is related to laser-hybrid welded joints, which typically have steeper weld flanks on the bottom side and the usage of the temporary root backing for the joints made by flux-cored arc welding; see Braun et al.[51] The median undercuts are mainly between 0 and 0.1 mm; however, a few specimens had quite deep undercuts up to 0.6 mm. The weld width and height are comparable for top and bottom sides. Due to the smaller weld toe radii and larger weld toe angles, it is not surprisingly that the median SCFs (for tension loading) are in average larger on the bottom than on the top side weld toes. In addition to the geometrical and SCF parameters, axial and angular misalignment of all joints was measured prior to testing. The ratio of axial misalignment to plate thickness e/t is smaller than 10% for all joints. In

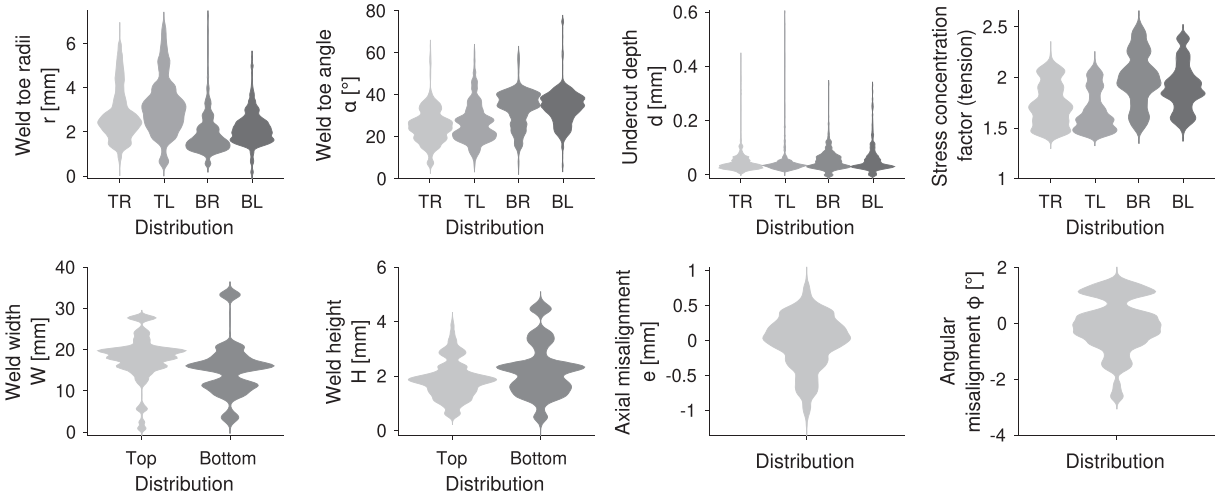


FIGURE 5 Distribution of the medians of the measured geometrical and stress concentration factor parameters that are used as input for the machine learning models

contrast, high variations are observed for angular misalignment with absolute values of angles up to 3°.

4 | RESULTS

4.1 | Performance of prediction based on SCF

Prior to using ML models, a prediction of fracture location based on SCF values was made. First, the maximum or median SCF was identified per location (out of the about 300 scan slices per specimen). Then, out of four possible locations, the fracture was predicted to be wherever the highest maximum or median SCF was computed. The accuracy (ACC) and Matthews correlation coefficient (MCC) are used to quantify the performance of the models. These metrics are calculated from the number of true positive predictions (TP), true negatives (TN), false positives (FP), and false negatives (FN). For further information, see Braun et al.[23]

The accuracy is the ratio of correct predictions and the number of all predictions.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Since the number of specimens belonging to each class (i.e., failure location) varies, the MCC was used additionally. Its value is only high if the classifier does well on the prediction of all classes. For binary classes, the MCC is defined as follows:

$$MCC = \frac{TP + TN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2}$$

For the median SCF, the accuracy and MCC were 0.30 and −0.01. For the maximum SCF, the accuracy and MCC were 0.30 and 0.02. The confusion matrices are given in the appendix, Figure A1. Considering that there are four classes to predict, an accuracy of 0.3 is only slightly better than a random guess. This can also be seen by the MCC, which, when close to zero, indicates that the prediction is no better than random.[54]

4.2 | Performance of ML models

One condition for meaningful analyses, for example, with respect to explainability, is a good performance of the models. The performance of the SCF ANN model has been confirmed in an earlier study.[9] The performance of the fracture location classifier was checked once for the SCF being known and once for it being unknown. A difference in performance (disregarding any subsequent explainability results) is a hint that the information of the SCF is indeed useful in predicting the fracture location.

Table 1 shows both the performance of the fracture location classifier as well as the ranks of the SCF features when included as input data. A higher rank indicates that the feature was important for the ML model's prediction, that is, that it was an influential feature. Four cases with different input features were assessed as follows: using no SCF values, maximum SCF values, median SCF

TABLE 1 Classifier performance and rankings with and without SCF features

Metric	No SCF	SCF max	SCF median	SCF max and median
Accuracy CV	0.72 ± 0.05	0.72 ± 0.02	0.72 ± 0.02	0.71 ± 0.03
MCC CV	0.61 ± 0.06	0.62 ± 0.03	0.61 ± 0.03	0.60 ± 0.04
Accuracy FM	0.72 ± 0.02	0.72 ± 0.04	0.72 ± 0.04	0.71 ± 0.04
MCC FM	0.62 ± 0.03	0.62 ± 0.06	0.61 ± 0.06	0.61 ± 0.06
K_{tTRmax} rank		15.2 ± 5.6		19.2 ± 7.4
K_{tTLmax} rank		15.0 ± 3.9		16.7 ± 5.2
K_{tBRmax} rank		9.1 ± 4.3		8.6 ± 3.1
K_{tBLmax} rank		15.5 ± 5.7		16.3 ± 4.0
K_{tTRmed} rank			19.7 ± 5.5	23.8 ± 6.7
K_{tTLmed} rank			12.8 ± 3.6	14.6 ± 4.3
K_{tBRmed} rank			27.5 ± 2.2	31.5 ± 2.3
K_{tBLmed} rank			16.5 ± 4.6	22.7 ± 5.3

Note. Numbers are mean ± standard deviation for $n = 10$.
Abbreviations: CV, cross-validation; FM, final model; MCC, Matthews correlation coefficient; SCF, stress concentration factor.

values, or maximum and median SCF values. All cases were run 10 times with different random seeds.

The first four rows (Table 1) give the accuracy and MCC during CV and for the final model (FM). The values of the performance metrics are similar for the CV and the FM. This shows that the model generalizes well to previously unseen samples. The overall performance is good with more than 70% correctly classified samples. The relatively high MCC shows that the model performs well on all classes.

From the fifth row on, the values indicate the rankings based on average SHAP values of each SCF-related feature. All SCF features are consistently ranked very low except for the maximum SCF value for the bottom right location (K_{tBRmax}). Based on this, further results were generated without including SCF features as inputs.

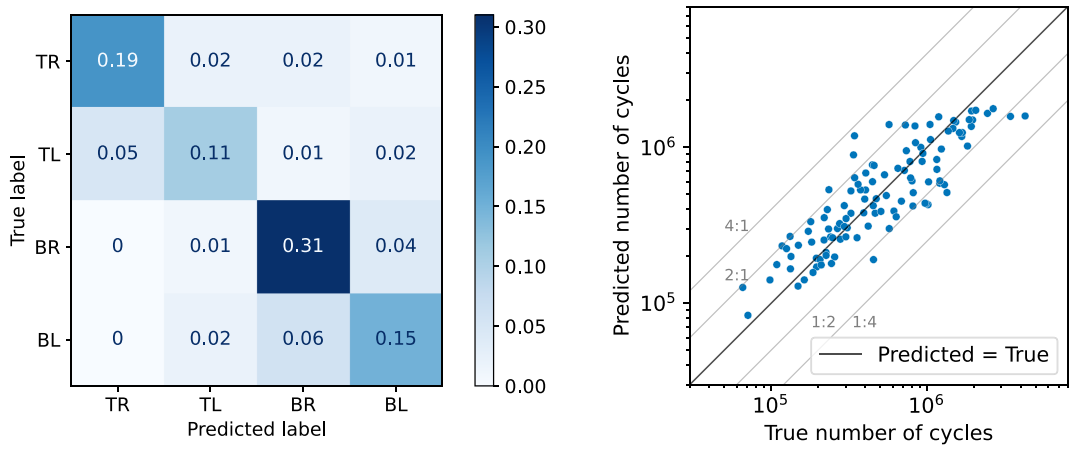
Next, Figure 6A shows the confusion matrix for the ML classifier model. It gives the fraction of samples in the test data per combination of true and predicted labels. For instance, the sum of top row values indicates the fraction of true TR samples (24%); however, the top left entry shows that the model only predicted the correct label for 19% of (all test) samples. Other entries in the top row indicate that, for example, the model incorrectly predicted a BL label when the true label was TR for 1% of the samples. Generally, the perfect model would only have (nonzero) entries on the diagonal and the fraction values would represent the distribution of labels in the test data. Here, there are nonzero off-diagonal entries, but they are very small. This means that there were some misclassified samples, but overall, most samples were correctly classified.

Table 2 shows both the performance of the number of cycles regression model as well as the ranks of the SCF features when included as input data. In this case, two new SCF features are introduced which indicate the SCF maximum or median value at the fracture location predicted by the classifier. Four cases were evaluated: either using no SCF features as inputs, using only the maximum or medium SCF value, respectively, and using both. Again, all cases were run 10 times with different random seeds. The SCF features are consistently ranked very low and do not affect overall model performance. Hence, further results are presented for the "no SCF" model only.

The "no SCF" lifetime model has an average mean absolute error (MAE) of 0.16 (logarithmized targets); however, the interpretation of average errors on transformed target values is not intuitive. Alternatively, Figure 6B shows a plot of the predictions on the test data plotted over the true values. This shows that the lifetime model performed well with no prediction off by more than four times the true number of cycles to failure (N_{f}).

4.3 | Explanation of results

Both the fracture location classifier and the lifetime prediction model performed well. This merits further analysis with the explanation models to gain insight into the data and underlying mechanisms. Based on the low impact of including SCF features as input data (Sections 4.1 and 4.2), these features were not used during the explanation of the results.



(A) (B)

FIGURE 6 Performance of machine learning (ML) models. (A) Confusion matrix indicating the true versus predicted fraction of fracture locations for the prediction with the classifier; the low off-diagonal values indicate good performance. (B) Plot of predicted versus true number of cycles to failure for the lifetime model [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Regressor performance and rankings with and without SCF features

Metric	No SCF	SCF max	SCF median	SCF max and median
RMSE CV	0.20 ± 0.01	0.21 ± 0.01	0.20 ± 0.01	0.21 ± 0.01
MAE CV	0.16 ± 0.01	0.16 ± 0.01	0.16 ± 0.01	0.16 ± 0.01
RMSE FM	0.21 ± 0.02	0.22 ± 0.02	0.21 ± 0.02	0.21 ± 0.02
MAE FM	0.16 ± 0.01	0.17 ± 0.02	0.16 ± 0.02	0.16 ± 0.01
K_{tFmax} rank		22.0 ± 2.1		24.0 ± 2.4
K_{tFmedian} rank			17.4 ± 4.0	19.2 ± 4.4

Note. Numbers are mean ± standard deviation for $n = 10$. K_{tFmax} and K_{tFmedian} are SCF values at the fracture location predicted by the classifier.
Abbreviations: CV, cross-validation; FM means final model; MAE, mean absolute error; RMSE, root mean square error; SCF, stress concentration factor.

4.3.1 | Explanation of fracture location classifier

The output of the fracture location classifier model is the log-odds per class or location, respectively. A higher value means a higher probability for fracture at that location. Hence, a higher SHAP value means a higher impact of the corresponding feature value toward the class we are currently looking at.

The overall ranking of features (i.e., the feature importance) is given in Figure 7A for the top right class. A high rank means including this feature in the input data had a high impact on the prediction; however, this ranking of average SHAP values conflates magnitude and prevalence. The beeswarm plots in Figure 7b give SHAP value and feature value per sample (i.e., one dot is one sample) in the same order as in the ranking in Figure 7A. This further clarifies the impact of features and can, for instance, reveal rare but high-magnitude effects.

For most features, Figure 7B shows a continuous impact from high to low feature values. For the angular misalignment, a high feature value (high misalignment) increases the log-odds and hence the probability of failure at this location and vice versa for low feature values. For the axial misalignment, this is reversed. Surprisingly, the undercut of the *bottom right* location d_{BR} is ranked third; however, the beeswarm plot reveals that a high undercut value decreases the chance of fracture at the top right location, for which the ranking was made. All follow-up features have a very low impact.

The ranking and beeswarm plots depict the impact of features separately. We furthermore seek to assess the interaction of features, that is, the amplification or damping of the impact of a primary feature by a secondary one (Figure 8). Such relations can be identified with SHAP dependence plots,[55] which enrich traditional partial dependence plots.[56] SHAP dependence plots show the SHAP value (the impact) of the primary feature on the

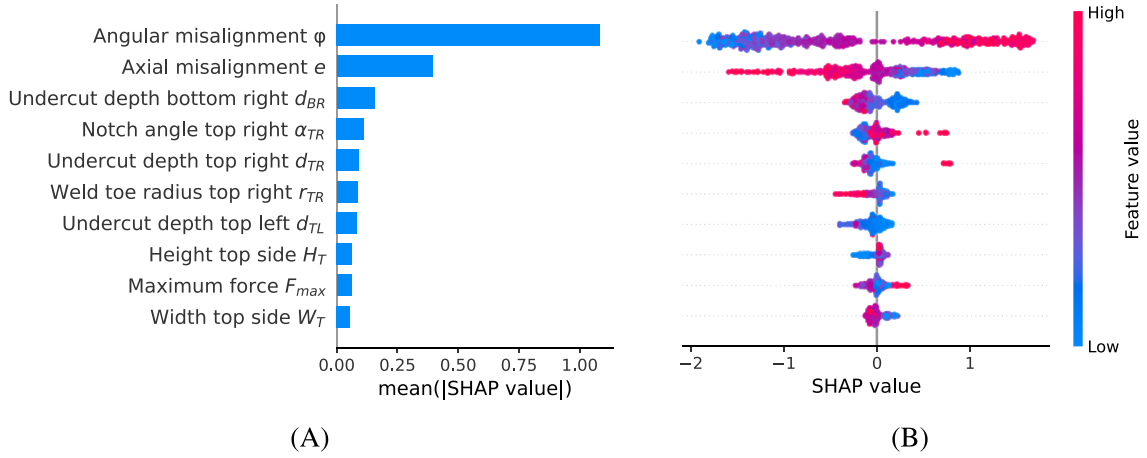


FIGURE 7 Explanation for fracture location classifier for class top right. (A) Average SHapley Additive exPlanation (SHAP) value indicating global feature importance. (B) Local explanation with beeswarm plots with vertical feature order as in the ranking. Each dot is one observation; its color corresponds to its feature value. The dot's position on the horizontal axis shows the impact the feature has on the model output for that observation. Multiple observations with similar impact pile up to indicate density [Colour figure can be viewed at wileyonlinelibrary.com]

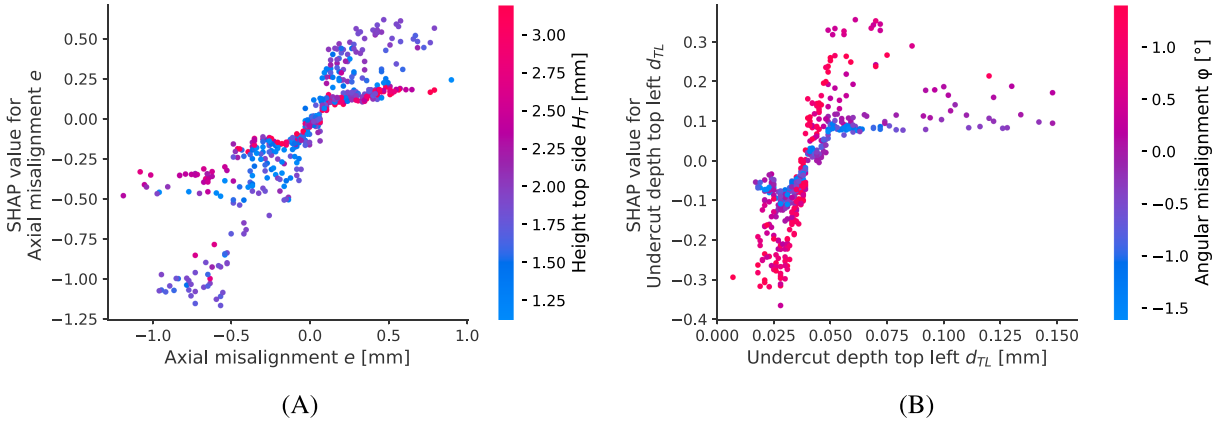


FIGURE 8 Interaction plots for fracture location classifier for the top left location. Each dot is one sample; its feature value is given on the horizontal axis and the corresponding SHapley Additive exPlanation (SHAP) value on the vertical axis. The dot's color corresponds to a secondary feature that possibly interacts with the primary feature. (A) Axial misalignment and height top side. (B) Undercut top left and angular misalignment [Colour figure can be viewed at wileyonlinelibrary.com]

vertical axis versus the feature value on the horizontal axis. This shows overall trends regarding the impact of the primary feature. In addition, one dot represents one sample and the dot's color is the value of a secondary feature. Clear vertical color patterns hint at interaction effects. Moreover, we are looking for inflection points which, when surpassed, strongly change a feature's impact.

In general, the number of possible interaction pairs is features × (features-1) × models × classes and therefore high, yet few clear interaction effects were found; a selection is depicted below. Figure 8A shows a dependence plot for axial misalignment and weld height on the top side as a secondary feature. An axial misalignment of zero has no impact, whereas a positive value pushes the prediction toward a higher probability of fracture at the top left location (and vice versa). Also, red dots tend to cluster around a SHAP value of zero; so if the height is large, the impact of the axial misalignment is dampened.

Figure 8B shows the same plot for the top left undercut and angular misalignment. Here, the majority of samples is between an undercut of 0.02 and 0.06 mm. In this range, an increase of undercut depth increases the probability of fracture at that location. This effect is amplified by larger angular misalignment values, which leads to tensile secondary bending stresses on the top side.

4.3.2 | Explanation of lifetime regressor

The lifetime regressor output is the logarithmized number of cycles to failure (N_f). The logarithm is a monotonously increasing function, so a higher $\log(N_f)$ means a higher N_f. Equivalently, a positive SHAP value means an impact toward a higher N_f. Before applying the explanation model, force-related features were excluded as they strongly correlate with stress-related features.

The overall ranking is given in Figure 9A. Stress-related features (amplitude and maximum value) are ranked highest, followed by the three geometry features of angular misalignment, height of weld top side, and axial misalignment (φ, H_T, and e). The beeswarm plots also show the high and continuous impact of the stress-related features. The beeswarm plots for φ and e agree well with the general understanding of misalignment effects. Large positive of negative values of φ and e lead to a reduction in number of cycles to failure (N_f), and close to zero values of φ and e (indicated by purple) result in higher number of cycles to failure (positive SHAP values). The weld heights on top and bottom sides (H_T, H_B) behave similarly as the stress-related features (large heights reduce umber of cycles to failure). The follow-up features have a comparatively low impact and will not be further discussed.

Figure 10 presents a SHAP dependence plot of the stress amplitude σ_a and the weld toe radius on the top left side. With increasing σ_a, the number of cycles to failure decreases. This effect is slightly amplified by small weld toe radii (see also Figure A2 in the Appendix).

5 | DISCUSSION

Overall, the proposed ML models are well-suited to predict the fatigue behavior of welded joints. The successful CV showed that the models are robust with respect to the input data. It is furthermore expected that they will generalize well to new data samples. The performance of the models was similar to a previous study[50] though the current dataset includes more samples; nonetheless, the current data also include more test types which can make it harder to achieve high accuracy. The ratio of samples versus the number of features is likely suboptimal. On the other hand, the good performance shows that the data include features that strongly influence fatigue behavior.

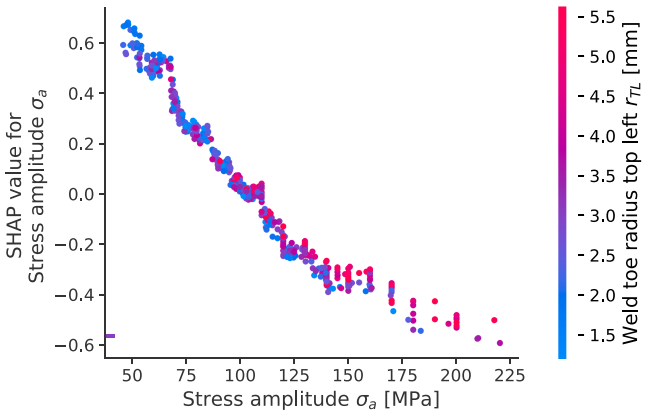


FIGURE 10 Interaction plot for lifetime regressor with stress amplitude as primary and top left weld toe radius as secondary feature [Colour figure can be viewed at wileyonlinelibrary.com]

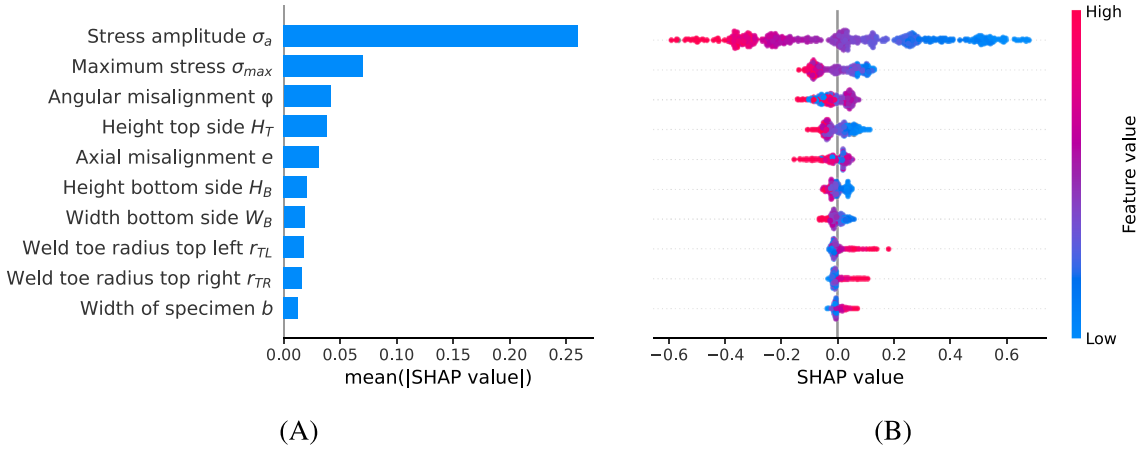


FIGURE 9 Explanation for lifecycle regression model. (A) Average SHapley Additive exPlanation (SHAP) value indicating global feature importance. (B) Local explanation with beeswarm plots [Colour figure can be viewed at wileyonlinelibrary.com]

Regarding the explanation model, the results show that the approach works well. SHAP values are well-suited to quantify feature impact. The explanation results are conclusive and mostly as expected. Some unclear trends in beeswarm plots could be due to interaction effects or correlations between features which may negatively impact explainability. Contrasting the earlier study,[50] SCF-related features were used by the ML models but shown to be almost irrelevant in the prediction of fatigue behavior. This aspect could be related to the fact that SCFs are determined based on other input features and therefore likely correlate with those other features.

Notwithstanding the good results, there are sources of error and uncertainty. From a ML perspective, typical error sources are bias and variance. The former is due to erroneous assumptions by the learning algorithm. Decision trees make almost no assumptions and are usually not susceptible to high bias. The latter originates from highly complex models. Here, tree ensembles are used where each tree is pruned to control variance.[37] In addition, the high model performance and successful CV make it unlikely that the models suffer from either of both error sources.

From a data perspective, there are two central issues: missing values and biased data. Few samples contain missing values, so this is not an issue. Biased features, however, cannot be ruled out as, for example, certain test types may exist more often in the data. This should not influence the models' accuracy but can limit their generalizability. It may also result in ranks that do not reflect reality since SHAP values can only attribute a high importance to features by contrasting values within one feature. A feature may be very influential, but if the feature value is not varied during tests, this influence cannot be revealed. In other words, there is no pattern to be detected by either ML or explanation model.

Generally, the results are in line with theoretical expectations. Not surprisingly, stress-related features (stress amplitude and maximum stress) have a high predictive power for the lifetime regressor but less so for the fracture location classifier. In contrast, the impact of axial (e) and angular (φ) misalignment makes sense w.r.t. fracture locations. From a structural mechanics' perspective, a positive angular misalignment increases secondary bending stresses on the top side and thereby increases the likelihood of failure at either top left or top right weld toes. Likewise, axial misalignment is responsible for tensile secondary bending stresses on either left or right side of a weld depending on the direction of misalignment.

Of the geometrical features, the undercut depth of the *bottom right* location d_{BR} is ranked third for the failure location classifier. This is related to the fact that a high undercut value at a different location than the one, for which the ranking was made, decreases the chance of fracture at the assessed location. All follow-up features have a very low impact.

Interestingly, the weld heights on top and bottom sides (H_T, H_B) have a high predictive power for the lifetime regressor. This agrees with the results of the first ML model (Braun et al.[23]). A high weld height corresponds to a steeper flank angle and thereby increases the local stress range. In contrast, the weld toe angles do not have a high impact on the model. The only feasible explanation is that a high weld height increases the local stresses at two weld toes at once. Hence, a higher impact is expected compared with a feature for a single weld toe. The follow-up features have a comparatively low impact; nevertheless, as the database is continuously extended, more mutual influences will probably be detected in the future due to more available training data.

Material strength related parameters like yield and ultimate tensile strength were not found to have any significant impact on the ML model. This also agrees with the general understanding of fatigue at welded joints. Interestingly, also, the stress ratio was not found to have a significant impact. In international standards and guidelines (e.g., Hobbacher[53]) it is typically assumed that welding residual stresses are relieved in small-scaled fatigue specimens. Consequently, an effect of stress ratio on fatigue behavior would be expected. This is not the case in the current study and indicates no general welding residual stress relief for all specimens. Comparing the S-N curves obtained from tests with different stress ratios, a difference in fatigue strength close to the fatigue limit is observed from some test series; see Braun et al.,[51,52] yet the impact of the feature is found to be small compared with other features. This makes sense w.r.t. to the larger difference in fatigue strength between test specimens made by different welding methods such as flux-cored arc welding and laser-hybrid welding. In other words, the difference in local weld geometry and misalignment parameters has a higher impact on fatigue strength than other influencing factors, that is, stress ratio, yield, and ultimate tensile strength, as well as test temperature. This does not mean that these factors are insignificant, but the other factors are more influential.

6 | CONCLUSIONS

The purpose of the current study was to tackle the complicated engineering issues of fatigue life assessment of welded joints by removing the need for complex three-dimensional FE calculations based on the actual measured weld geometry. To this end, a ML framework was

used to predict the number of cycles to failure and the fracture initiation locations in small-scale butt-welded joints. The created approach was compared with a traditional approach based on SCFs. The following conclusions are drawn from the investigation:

- The presented model is capable of accurate predicting the number of cycles to failure and the fracture initiation locations for more than 600 fatigue test specimens despite the wide range of features such as different welding methods, stress ratios, and test temperatures.
- To determine whether the SCFs improve the accuracy of the ML model, tests were performed with and without using SCFs as input features; however, this did not improve the predictions.
- Using the maximum SCF of each specimen to predict a fracture initiation at that location is only slightly more accurate than a random guess and far less accurate than the ML predictions.
- Explainable ML methods are a useful tool to predict the mutual influence of ML methods or to determine less influential features. By ranking features by impact, unexpectedly high or low impacts of specific features can be detected.
- The most influential parameters for the lifetime regressor of small-scale butt-welded joints are stress-related (stress amplitude and maximum stress) followed by misalignment-related (axial and angular misalignment), and the weld height on top and bottom sides.
- In contrast, axial and angular misalignment are the most influential parameters for the failure location classifier. Apart from the undercut depths on one of the four weld toes, geometrical features appear to be less impactful for the failure location.

ACKNOWLEDGMENT

There is nobody acknowledged for this study. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

DATA AVAILABILITY STATEMENT

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

NOMENCLATURE

$\alpha_{TL}, \alpha_{TR}, \alpha_{BL}, \alpha_{BR}$	Weld toe angles
φ	Angular misalignment
σ_a, σ_{max}	Stress amplitude, maximum stress
b	Width of specimen
$d_{TL}, d_{TR}, d_{BL}, d_{BR}$	Undercut depths
e	Axial misalignment
k	Slope of S–N curves
$r_{TL}, r_{TR}, r_{BL}, r_{BR}$	Weld toe radii
t	Thickness
F_a, F_{max}	Force amplitude and maximum
H_T, H_B	Height of weld on top and bottom sides
$K_{tFmax}, K_{tFmedian}$	SCF values at the fracture location
$K_{tTRmax}, K_{tTLmax}, K_{tBRmax}, K_{tBLmax}$	Maximum SCFs
$K_{tTRmed}, K_{tTLmed}, K_{tBRmed}, K_{tBLmed}$	Median SCFs
N_f	Number of cycles to failure
R	Stress ratio
T	Test temperature
W_T, W_B	Width of weld on top and bottom sides
ACC	Accuracy
ANN	Artificial neural network
BR, BL, TR, TL	Failure locations
CV	Cross-validation
DNN	Deep neural network
FAT	Reference fatigue strength
FE	Finite element
FL	Fracture location
FP, FN	False positive and false negative predictions
FM	Final model
MAE	Mean absolute error
MCC	Matthews correlation coefficient
ML	Machine learning
PRC	Polynomial regression with coupling terms
RMSE	Root mean square error
SCF	Stress concentration factor
TP, TN	True positive and true negative predictions
XAI	Explainable artificial intelligence

ORCID

Moritz Braun https://orcid.org/0000-0001-9266-1698

REFERENCES

1. Barsoum Z, Jonsson B. Influence of weld quality on the fatigue strength in seam welds. *Eng Fail Anal*. 2011;18(3):971-979.
2. Jonsson B, Samuelsson J, Marquis GB. Development of weld quality criteria based on fatigue performance. *Weld World*. 2013;55(11-12):79-88.
3. Schork B, Kucharczyk P, Madia M, et al. The effect of the local and global weld geometry as well as material defects on crack initiation and fatigue strength. *Eng Fract Mech*. 2018;198:103-122.

4. Åstrand E, Stenberg T, Jonsson B, Barsoum Z. Welding procedures for fatigue life improvement of the weld toe. *Weld World*. 2016;60(3):573-580.
5. Hultgren G, Barsoum Z. Fatigue assessment in welded joints based on geometrical variations measured by laser scanning. *Weld World*. 2020;64(11):1825-1831.
6. Hultgren G, Myrén L, Barsoum Z, Mansour R. Digital scanning of welds and influence of sampling resolution on the predicted fatigue performance: Modelling, experiment and simulation. *Metals*. 2021;11(5):822.
7. Madia M, Zerbst U, Th. Beier H, Schork B. The ibess model – elements, realisation and validation. *Eng Fract Mech*. 2017;198:171-208.
8. Haibach E. *Betriebsfestigkeit: Verfahren und daten zur bauteillauslegung*. 3rd ed. Berlin, Heidelberg, New York: Springer-Verlag; 2006.
9. Braun M, Neuhäusler J, Denk M, et al. Statistical characterization of stress concentrations along butt joint weld seams using deep neural networks. *Appl Sci*. 2022;12:6089.
10. Alam MM, Barsoum Z, Jonsen P, Kaplan AFH, Häggblad HÅ. The influence of surface geometry and topography on the fatigue cracking behaviour of laser hybrid welded eccentric fillet joints. *Appl Surf Sci*. 2010;256(6):1936-1945.
11. Braun M. Recent progress on geometrical and stress concentration characterization of welded joints. In: The 7th International E-Conference on Industrial, Mechanical, Electrical, and Chemical Engineering (ICIMECE 2021); 2021.
12. Liinalampi S, Remes H, Lehto P, Lillemae I, Romanoff J, Porter D. Fatigue strength analysis of laser-hybrid welds in thin plate considering weld geometry in microscale. *Int J Fatigue*. 2016;87:143-152.
13. Renken F, von Bock und Polach RUF, Schubnell J, et al. An algorithm for statistical evaluation of weld toe geometries using laser triangulation. *Int J Fatigue*. 2021;149:106293.
14. Schubnell J, Jung M, Le CH, et al. Influence of the optical measurement technique and evaluation approach on the determination of local weld geometry parameters for different weld types. *Weld World*. 2020;64(2):301-316.
15. Anthes RJ, Köttgen VB, Seeger T. Kerbformzahlen von stumpfstößen und doppel-t-stößen. *Schweißen und Schneiden*. 1993;45(12):685-688.
16. Dabiri M, Ghafouri M, Rohani Raftar HR, Björk T. Utilizing artificial neural networks for stress concentration factor calculation in butt welds. *J Constr Steel Res*. 2017;138:488-498.
17. Ottersböck MJ, Leitner M, Stoschka M. Characterisation of actual weld geometry and stress concentration of butt welds exhibiting local undercuts. *Eng Struct*. 2021;240:112266.
18. Pachoud AJ, Manso PA, Schleiss AJ. New parametric equations to estimate notch stress concentration factors at butt welded joints modeling the weld profile with splines. *Eng Fail Anal*. 2017;72:11-24.
19. Wang Y, Luo Y, Tsutsumi S. Parametric formula for stress concentration factor of fillet weld joints with spline bead profile. *Mater (Basel)*. 2020;13(20):4639.
20. Neuhäusler J, Rother K. Determination of notch factors for transverse non-load carrying stiffeners based on numerical analysis and metamodeling. *Weld World*. 2022;66(4):753-766.
21. Oswald M, Mayr C, Rother K. Determination of notch factors for welded cruciform joints based on numerical analysis and metamodeling. *Weld World*. 2019;63(5):1339-1354.
22. Oswald M, Neuhäusler J, Rother K. Determination of notch factors for welded butt joints based on numerical analysis and metamodeling. *Weld World*. 2020;64(12):2053-2074.
23. Braun M, Kellner L, Schreiber S, Ehlers S. Prediction of fatigue failure in small-scale butt-welded joints with explainable machine learning. *Proc Struct Integ*. 2022;38:182-191.
24. Chen J, Liu Y. Fatigue modeling using neural networks: A comprehensive review. *Fatigue Fract Eng Mater Struct*. 2022;45(4):945-979.
25. Kalayci CB, Karagoz S, Karakaş O. Soft computing methods for fatigue life estimation: A review of the current state and future trends. *Fatigue Fract Eng Mater Struct*. 2020;43(12):2763-2785.
26. Han YL. Artificial neural-network technology as a method to evaluate the fatigue life of weldments with welding defects. *Int J Press Vessel Pip*. 1995;63(2):205-209.
27. Karakas Ö. Estimation of fatigue life for aluminium welded joints with the application of artificial neural networks. *Materwiss Werksttech*. 2011;42(10):888-893.
28. Yang XH, Deng W, Zou L, Zhao HM, Liu JJ. Fatigue behaviors prediction method of welded joints based on soft computing methods. *Mater Sci Eng A*. 2013;559:574-582.
29. Zou L, Sun Y, Yang X. An entropy-based neighborhood rough set and pso-svrm model for fatigue life prediction of titanium alloy welded joints. *Entropy (Basel)*. 2019;21(2):117.
30. Karakaş O, Tomasella A. Fatigue life estimation of non-penetrated butt weldments in ligth metals by artificial neural network approach. *Materwiss Werksttech*. 2013;44(10):847-855.
31. Kalayci CB, Karagoz S, Karakaş O. Bee colony intelligence in fatigue life estimation of simulated magnesium alloy welds. *Int J Fatigue*. 2019;127:36-44.
32. Larrañaga P, Calvo B, Santana R, et al. Machine learning in bioinformatics. *Brief Bioinform*. 2006;7(1):86-112.
33. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67.
34. Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-Normalizing Neural Networks. In: Guyon I, Luxburg UV, Bengio S, et al., eds. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*: Neural Information Processing Systems Foundation, Inc. (NIPS); 2017:972–981.
35. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Krishnapuram B, Shah M, Smola A, Aggarwal C, Shen D, Rastogi R, eds. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. New York, USA: ACM Press; 2016:785-794.
36. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 2002;38(4):367-378.
37. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edition, corrected at 12th printing 2017, Springer Series in Statistics. New York, NY: Springer; 2009.
38. Refaeilzadeh P, Tang L, Huan L. Cross-validation. In: Liu L, Özsu M, eds. *Encyclopedia of Database Systems*: Springer US; 2009:532-537.

39. Géron A. *Hands-on Machine Learning With Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd ed. Sebastopol, CA: O'Reilly; 2019.
40. Kellner L. Analyzing the complexity of ice with explainable machine learning for the development of an ice material model. *Doctoral thesis*: Hamburg University of Technology, Hamburg, Germany; 2022.
41. Stender M. Data-driven techniques for the nonlinear dynamics of mechanical structures. *Doctoral thesis*: Hamburg University of Technology; 2020.
42. Kellner L, Stender M, von Bock und Polach F, Ehlers S. Predicting compressive strength and behavior of ice and analyzing feature importance with explainable machine learning models. *Ocean Eng*. 2022;255:111396.
43. Stender M, Adams C, Wedler M, Grebel A, Hoffmann N. Explainable machine learning determines effects on the sound absorption coefficient measured in the impedance tube. *J Acoust Soc Am*. 2021;149(3):1932.
44. Roscher R, Bohn B, Duarte MF, Garcke J. Explainable machine learning for scientific insights and discoveries. *IEEE Access*. 2020;8:42200-42216.
45. Chen H, Janizek JD, Lundberg S, Lee S-I. True to the Model or True to the Data? 2020.
46. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Neural Information Processing Systems Proceedings 2017; 2017:4768–4777.
47. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749-760.
48. Ribeiro MT, Singh S, Guestrin C. Why Should I Trust You? In: Krishnapuram B, Shah M, Smola A, Aggarwal C, Shen D, Rastogi R, eds. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. New York, USA: ACM Press; 2016:1135-1144.
49. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub; 2020.
50. Braun M. Assessment of fatigue strength of welded steel joints at sub-zero temperatures based on the micro-structural support effect hypothesis. *Doctoral thesis*: Hamburg University of Technology; 2021.
51. Braun M, Ahola A, Milaković A-S, Ehlers S. Comparison of local fatigue assessment methods for high-quality butt-welded joints made of high-strength steel. *Forces Mech*. 2021;6:100056.
52. Braun M, Kahl A, Willems T, Seidel M, Fischer C, Ehlers S. Guidance for material selection based on static and dynamic mechanical properties at sub-zero temperatures. *J Offshore Mech Arct Eng*. 2021;143(4):1-45.
53. Hobbacher AF. *Recommendations for Fatigue Design of Welded Joints and Components*. 2nd ed., IIW Collection: Springer International Publishing Switzerland; 2016.
54. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020; 21(1):6.
55. Lundberg SM, Erion GG, Lee S-I. Consistent Individualized Feature Attribution for Tree Ensembles; 2019.
56. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232.

How to cite this article: Braun M, Kellner L. Comparison of machine learning and stress concentration factors-based fatigue failure prediction in small-scale butt-welded joints. *Fatigue Fract Eng Mater Struct*. 2022;45(11): 3403-3417. doi:10.1111/ffe.13800

APPENDIX

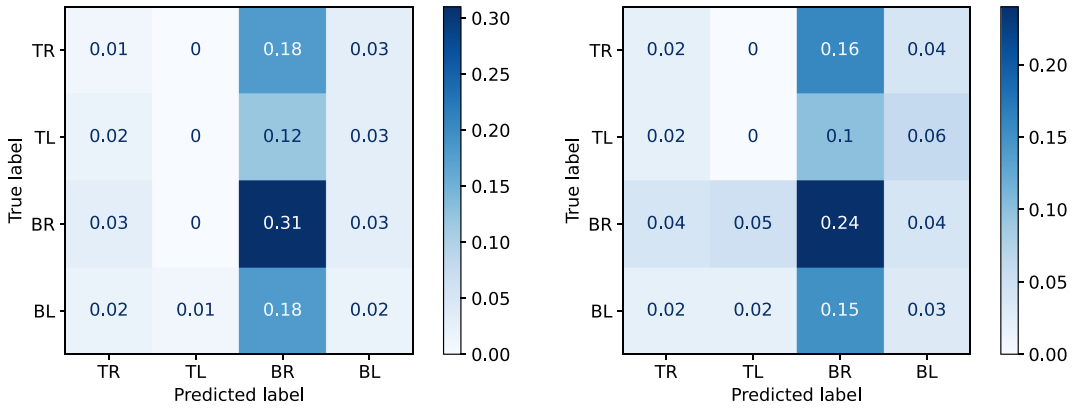


FIGURE A1 Confusion matrices for prediction based on stress concentration factor (SCF) values. (A) Based on maximum values, and (B) based on minimum values [Colour figure can be viewed at wileyonlinelibrary.com]

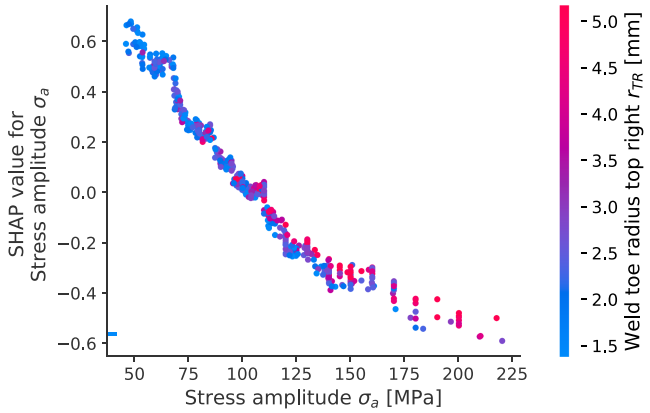


FIGURE A2 Interaction plot for lifetime regressor with stress amplitude as primary and top right weld toe radius as secondary feature [Colour figure can be viewed at wileyonlinelibrary.com]