Predicting Steel Fatigue Strength using Ensemble Modeling

Yadu Krishna Sunil

Intro

The engineering problem to be studied is the prediction of fatigue strength of steel, based on its chemical composition, processing parameters, and mechanical properties. Accurate prediction of fatigue strength is necessary for designing steel components that are resilient to cyclic stress and can withstand long periods of use without failure. This problem involves determining if having knowledge of the parameters allows for prediction of the steel fatigue strength, and if so, which parameters are the most important.

Accurately predicting fatigue strength is important to ensuring the safety and reliability of designs, safeguarding against failures. An understanding of how various factors influence fatigue strength allows engineers to optimize material selection and processing techniques. Additionally, fatigue strength predictions can mitigate over-engineering, reducing the need for safety margins well beyond what is necessary. This approach not only streamlines design processes but also leads to cost savings in material use and maintenance throughout a component's lifecycle.

There is existing research that examines the implementation of various predictive modeling techniques to estimate steel fatigue strength. One such study by Agrawal et al. evaluated 12 different models on its capability to predict fatigue strength and were able to achieve high levels of accuracy for models like Multivariate Polynomial Regression, M5 Model Trees, and Reduced Error Pruning Trees, with R^2 values above 0.97 [1]. Most techniques were able to separate out the three grades of steel within the dataset, however different techniques tend to perform better for different grades, a concept illustrated in Figure 1 below.

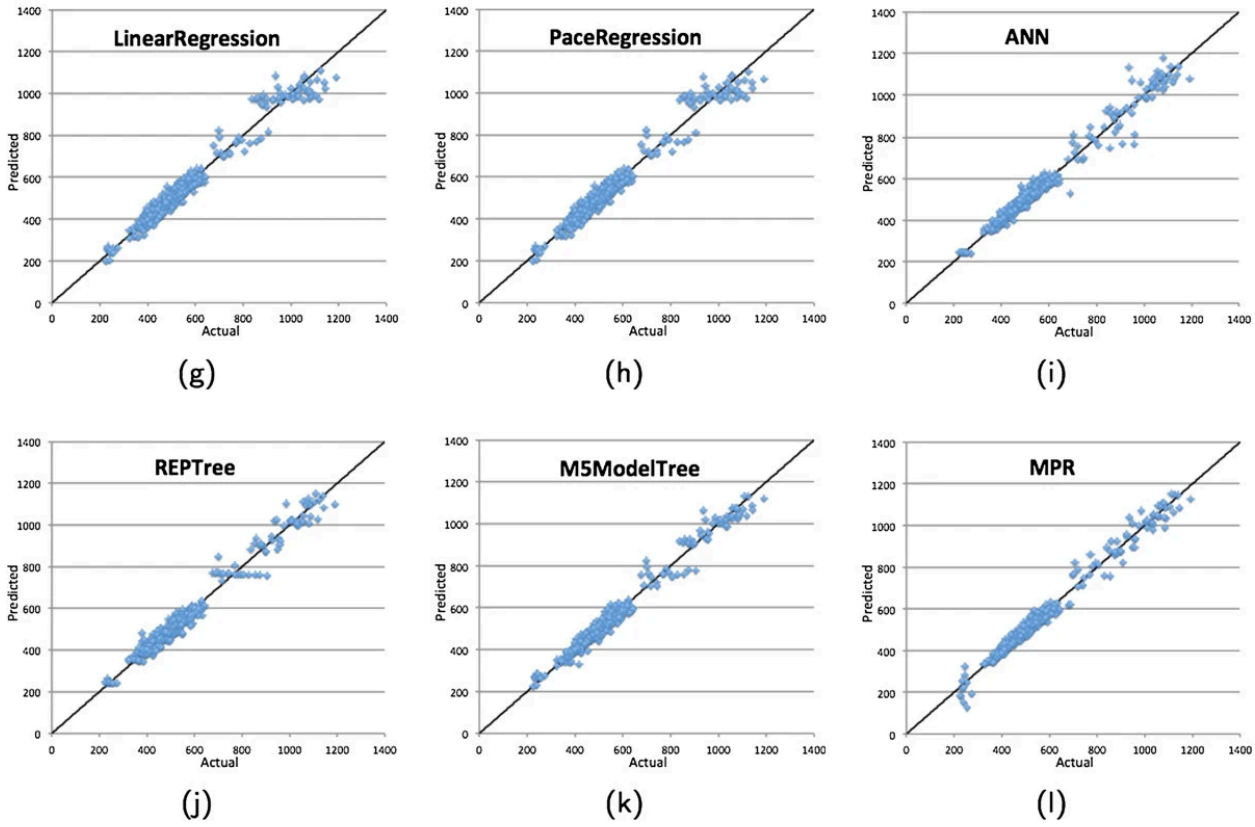

Figure 1: Scatterplots depicting 6 of the tested modeling techniques, with the x-axis and y-axis denote the actual and predicted fatigue strength values, respectively. g) Linear Regression; h) Pace Regression; i) Artificial Neural Networks; j) Reduced Error Pruning Trees; k) M5 Model Trees; l) Multivariate Polynomial Regression. [1]

This variation in prediction performance for different grade steels presents a significant challenge for accurately estimating fatigue strength. This study aims to address this challenge by implementing ensemble predictive modeling. Ensemble methods, which combine predictions from multiple models, offer a robust approach to improve accuracy and mitigate the variability in performance between different steel grades. Additionally, this study will explore advanced feature selection techniques to identify the most predictive features across all grades, further enhancing model performance.

Methodology

The dataset for this study is obtained from Kaggle, titled "Steel Fatigue Strength Prediction." This dataset contains 437 samples, each with 25 features that detail the chemical composition, heat treatment, and mechanical processing parameters of different steel samples, along with their corresponding fatigue strength. The data is composed of three steel grades, 371 carbon and low alloy steels, 48 carburizing steels, and 18 spring steels [2]. All features are numerics, so label encoding is not performed. A stratified data splitting strategy is used to ensure each steel type is proportionally represented in both the training and testing sets. Each steel grade is isolated and split into an 80-20 train-test split, and recombined to make an overall train-test split.

Advanced feature selection techniques are explored to identify the most predictive features for estimating fatigue strength. Agrawal et al. uses all 25 features to create their models [1]. Reducing the feature set can lead to a simpler model. This reduces the time for collecting all these features and the training time for the models. One of the most popular techniques is Recursive Feature Elimination (RFE) [3]. An advanced version of this, RFECV will be used in this study. This technique implements RFE with cross-validation and automatically finds the optimal number of features to keep [4]. The performance of the resulting feature set will be compared to BorutaSHAP which can perform better than RFE in some cases [5]. The feature selector estimators will be varied as well to find the optimal feature set. To evaluate the performance of the feature sets, the study uses Leave-One-Out Cross Validation (LOOCV) and R2 values. LOOCV is useful for small datasets as it maximizes use of data for training. The model used for prediction is Linear Regression. These choices are made because in the Agrawal et al. study, the code is not provided, so to compare the performance of the feature sets, Linear

Regression is a standard in the sklearn library, and thus the new feature sets are benchmarked on the same process used by Agrawal et al. The goal for this part of the study is to reduce the feature dimensionality as much as possible while keeping the R2 as close as possible to the value from the study. Finally, the performance of the feature set is evaluated on the test set as well.

With the reduced feature set, the study proceeds to create an ensemble model to address the variability in model performance across steel grades. This study evaluates a wide range of popular and high-performance models [6][7]. Each model is run with the default parameters to identify baseline performance. To tune the model, GridSearchCV with LOOCV is employed, sweeping a wide range of parameters specific to each model. The tuned model is then evaluated on the test set and the MSE of each steel grade is found as well. These values are compared to identify the best performing models, which are then used as inputs to an ensemble model.

To create the ensemble model, averaging and stacking are tested. Ensemble methods involve a Level 0 and a Level 1. The former trains different models on the same dataset, and outputs predictions. Level 1 generalizes the predictions made by different models to get the final output [8]. Averaging involves taking either the mean or the weighted average of the output prediction sets from Level 0. Stacking involves training a new model, often referred to as a meta-learner or meta-model, to effectively combine the predictions from Level 0 models. This meta-model is trained on the output predictions of the base models, using them as input features to learn the optimal way to blend these predictions to improve accuracy and reduce bias or variance in the final predictions [8].

Several models are evaluated for Level 1 of the stack while varying the combination of inputs to the model. Each trial is characterized by the R^2 value, the MSE, and the MSE of each steel grade. The best model is chosen by careful consideration of the trade-offs among these

metrics, aiming for a model that not only achieves a balance between high R^2 and low MSE but also demonstrates consistent performance across all steel grades.

Results

The initial step of the feature selection process involved examining the correlation matrix, depicted in Figure 2 below, which reveals several highly correlated features. Such features provide redundant information that does not enhance model performance and may unnecessarily extend training duration. Identifying and eliminating these extraneous features is crucial for the development of a streamlined regression model.

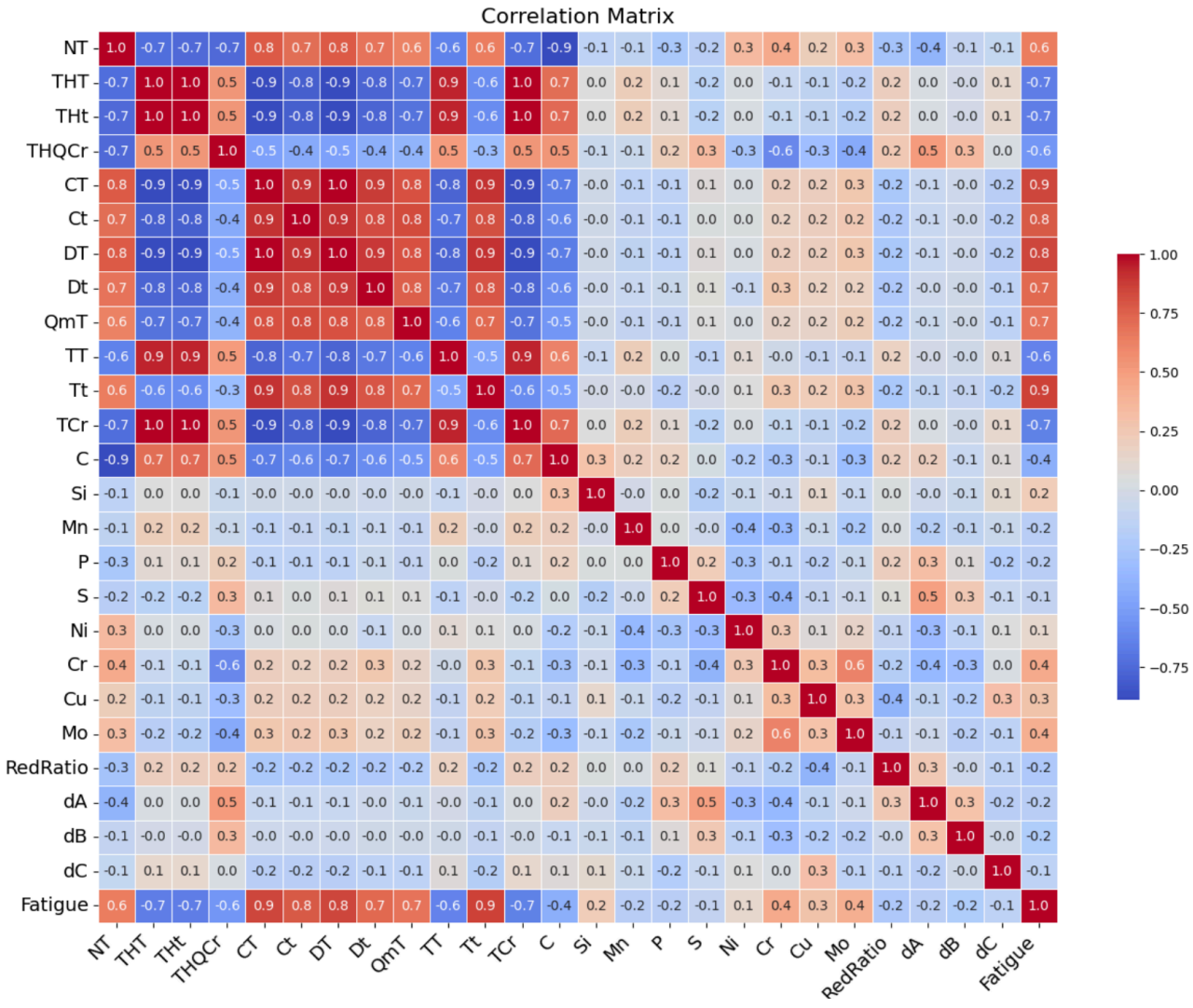

Figure 2: The correlation matrix shows the relationships between the features. Several values near 1 shows that there are many highly correlated features.

To eliminate the redundant features, several estimator and feature selector combinations were tested. The results of the LOOCV are shown below in Table 1.

Table 1: The results from testing various estimators with BorutaSHAP and RFECV

Estimator	**BorutaSHAP R^2**	**BorutaSHAP Features Kept**	**RFECV R^2**	**RFECV Features Kept**
XGB Regressor	0.7972	9	0.7958	10
Ridge	N/A	N/A	0.9598	23
LASSO	N/A	N/A	0.9611	25
SVM Linear Kernel	N/A	N/A	0.9594	12
LGBM Regressor	0.7025	10	0.8195	16
Cat Boost Regressor	0.9583	18	0.9599	23
Random Forest Regressor	0.9517	15	0.9611	23

The RFECV feature selector with the linear kernel Support Vector Machine estimator performed the best, requiring only 12 out of the 25 features to achieve a R^2 value of 0.9594. Agrawal et al.'s implementation of Linear Regression used all 25 features to achieve an R^2 value of 0.96, which is the same value achieved in this study as well. A difference of 0.063% is negligible, meaning the 12 features shown in Table 2 have enough predictive power to estimate fatigue strength.

Table 2: The 12 features selected by the RFECV method

Abbreviation	**Details**
THQCr	Cooling Rate for Through Hardening
CT	Carburization Temperature
Ct	Carburization Time
DT	Diffusion Temperature

QmT	Quenching Media Temperature (for Carburization)
TT	Tempering Temperature
Tt	Tempering Time
C	% Carbon
Si	% Silicon
Ni	% Nickel
Cr	% Chromium
Mo	% Molybdenum

To verify the efficacy of this feature subset, the Linear Regression model was trained on the entire training set, and predictions were made using the withheld test set. When comparing the metrics of R^2, fractional mean absolute error, root mean squared error, and standard deviation of error in Table 3, the values are all very similar between this study and Agrawal et al., indicating this feature subset is reliable.

Table 3: The results of various metrics with the reduced and full feature set

Metric	**Reduced Feature Set**	**Full Feature Set Agrawal et al**
$\mathbf{R^2}$	0.96	0.96
$\mathbf{MAE_f}$	0.05	0.05
$\mathbf{RMSE_f}$	0.06	0.06
$\mathbf{SDE_f}$	0.07	0.04

The following scatter plot in Figure 3 shows the relationship between the actual(x-axis) and the predicted fatigue strength values (y-axis) from the linear regression model. If the model's predictions were perfect, all points would lie exactly on the red line of slope 1.

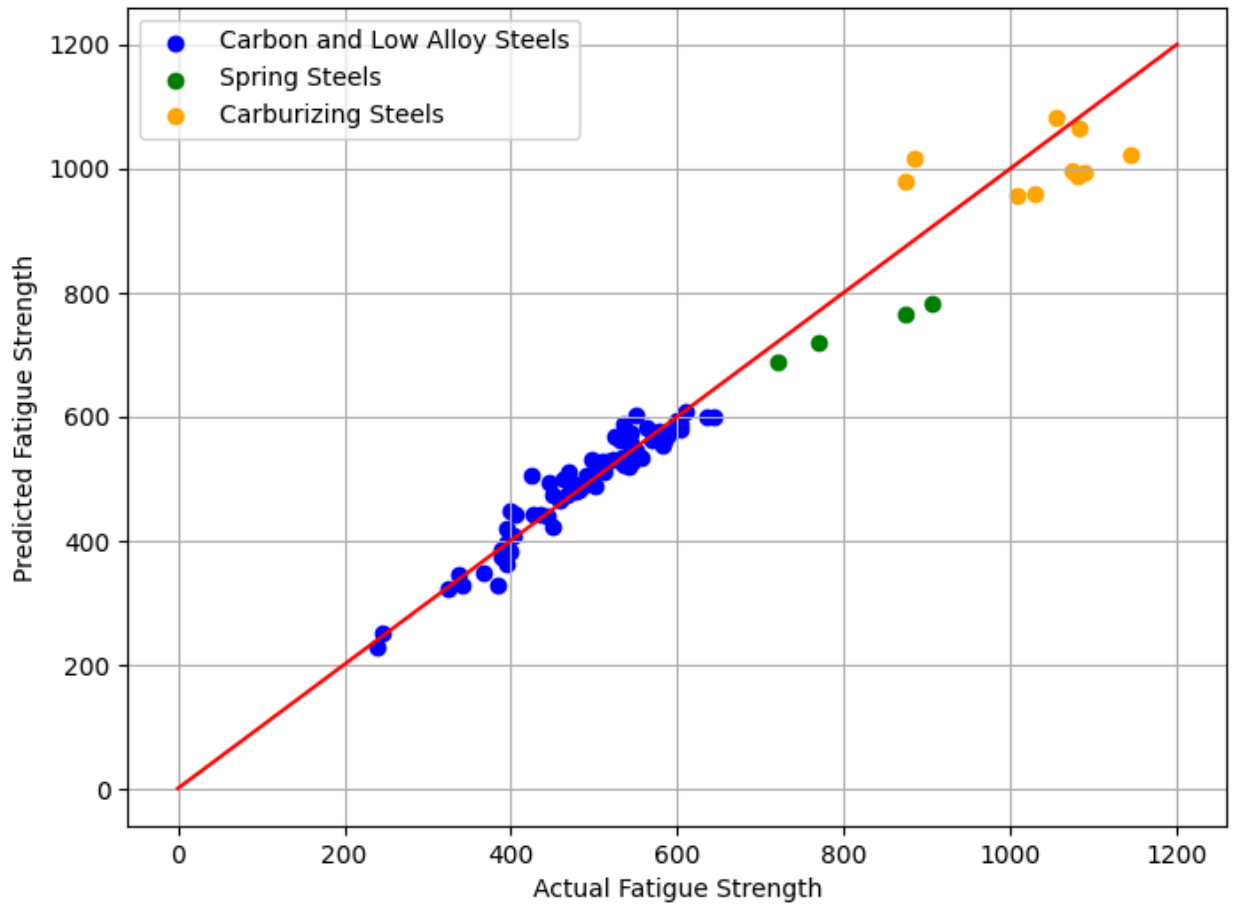

Figure 3: A scatter plot of the predicted vs actual fatigue strength values. The colors correspond to steel grade.

The separation of the steel grades is also visualized by Figure 3, and shows increased variance for spring and carburizing steels. To compare the performance of the reduced feature set to the full feature set, the error fractions are visualized in Figure 4 below. These plots show once again that even with the reduced feature set, the spread of data is virtually the same.

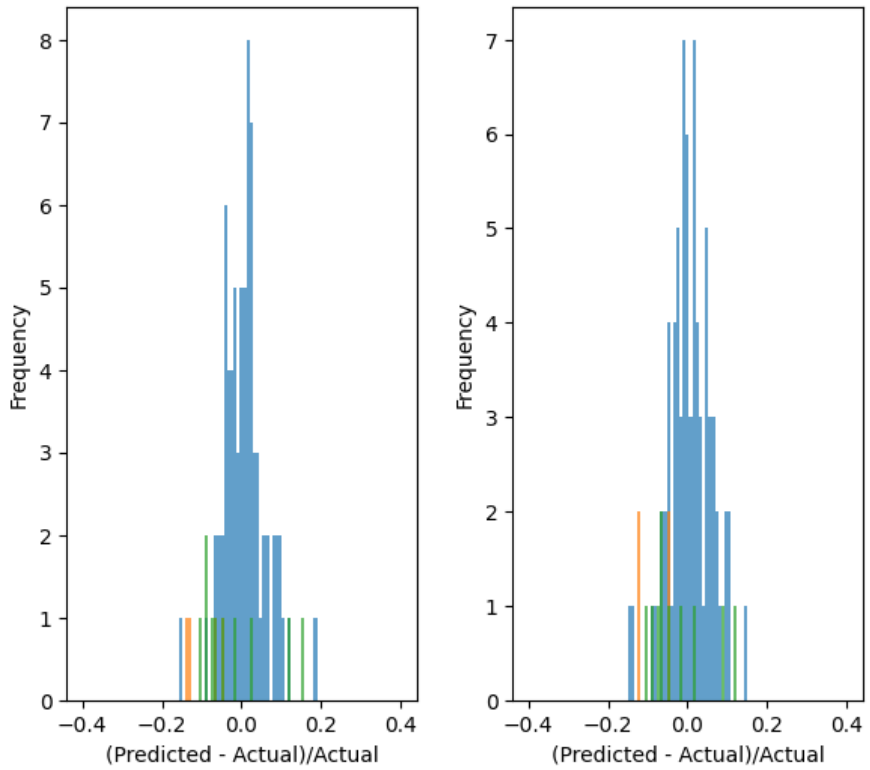

Figure 4: The error fractions of the reduced feature set (left) and the full feature set (right). Blue, green, and orange correspond to carbon and low alloy steels, spring steels, and carburizing steels, respectively.

The first component of the study has been achieved, and the dimensionality of the feature set has been reduced. However, as seen in Figure 3 and 4, the model has a higher standard deviation for high ranges of fatigue strength. To combat this, the next step is to implement ensemble predictive modeling. Table 4 shows the various Level 0 models that were tested, and the results from their best performing parameter set from GridSearchCV.

Table 4: The results of various estimators with the reduced set, also characterizing their performance across steel grades. The best case for each parameter is highlighted in green.

Level 0 Model	**Training CV Score R^2**	**Test R^2**	**Test MSE**	**Carbon and Low Alloy Steel MSE**	**Spring Steel MSE**	**Carburizing Steel MSE**
XGB	0.981	0.973	1066.18	300.29	15894.5	879.00

KNN	0.953	0.949	2011.30	547.584	9706.5	9911.10
LGBM	0.981	0.966	1347.76	416.21	16856.9	2130.68
CatBoost	0.980	0.976	942.95	324.85	10262.6	1850.85
SVM	0.941	0.926	2889.42	1028.40	1647.94	17343.66
MLP	0.923	0.925	2921.04	1746.61	9626.6	9047.04
ElasticNet	0.961	0.955	1766.74	609.54	7961.5	7967.87
GPR	0.933	0.937	2485.22	561.14	11647.1	13252.05
RANSAC	0.957	0.937	2455.09	1002.85	2745.37	12830.74
Linear	0.962	0.956	1709.54	594.16	7709.8	7674.81

Overall, most models seem to perform the best on carbon and low alloy steels while struggling with the other two. This result is expected as 84% of the dataset consists of carbon steels, making it easier to predict. All models also appear to generalize well to the test set, given the low differences between the training CV and test R^2 values, with a maximum discrepancy of 0.02 seen for RANSAC. CatBoost provides the best results overall, with the highest test R^2 and MSE values. The discrepancy of just 0.004 in R^2 when comparing to the training CV score also suggests the model is not overfitting. However, when evaluating the MSE per steel grade, it is evident that XGB excels at both the carbon and carburizing steels, while SVM performs the best for spring steels. As inputs to the ensemble model, these three models were chosen, with XGB and SVM providing the basis for all trials due to their best in grade performances. Additionally, LGBM was included due to its competitive test R^2, third only to CatBoost and XGB, and RANSAC was chosen for its proficiency with spring steel MSE, second to SVM.

Ensemble modeling is first tested by simply averaging the results from XGB, SVM, and CatBoost. The results, seen in the last row of Table 5 below, are promising, however it is only slightly better than the predictions of the best Level 0 model, CatBoost, for spring steels, and

worse in the other two categories. Testing with different weights on the averaging did not lead to improved results.

Stacking is implemented with sklearn's StackingRegressor. The input Level 0 estimators all have XGB and SVM present, and additions of CatBoost, LGBM, and/or RANSAC. Various Level 1 estimators are evaluated, with the results displayed in Table 5.

Table 5: The results of various Level 0 and 1 estimators and also characterizing their performance across steel grades. The best case for each parameter is highlighted in green.

Stack	Level 1 Model	Level 0 Models	Training CV Score R^2	Test R^2	Test MSE	Carbon and Low Alloy Steel MSE	Spring Steel MSE	Carburizing Steel MSE
1	Linear	SVM XGB	0.999	0.975	977.04	290.614	14150.95	855.70
2	Linear	SVM XGB CatBoost	0.998	0.977	915.20	281.09	12595.38	998.96
3	Linear	SVM XGB CatBoost LGBM	0.998	0.974	1031.47	310.60	13619.53	1402.72
4	Linear	SVM XGB CatBoost RANSAC	0.998	0.976	936.05	280.78	12973.12	1035.71
5	XGB	SVM XGB	0.989	0.973	1064.72	364.54	8794.35	3224.17
6	XGB	SVM XGB CatBoost	0.992	0.976	952.97	336.62	8810.30	2431.71
7	XGB	SVM XGB CatBoost	0.993	0.976	969.33	334.89	9356.88	2372.66

		LGBM						
8	XGB	SVM XGB CatBoost RANSAC	0.986	0.971	1135.85	295.50	8251.15	3841.95
9	KNN	SVM XGB	0.992	0.981	720.16	348.82	6860.34	1048.39
10	KNN	SVM XGB CatBoost	0.992	0.979	825.68	315.32	8501.71	1582.43
11	KNN	SVM XGB CatBoost LGBM	0.993	0.971	1140.44	355.04	14685.61	1612.88
12	KNN	SVM XGB CatBoost RANSAC	0.991	0.978	850.48	342.18	8501.71	1602.18
13	LGBM	SVM XGB	0.986	0.975	969.68	517.26	8866.83	1203.97
14	LGBM	SVM XGB CatBoost	0.986	0.975	978.56	560.27	8403.73	1145.72
15	CatBoost	SVM XGB	0.993	0.975	979.6	377.13	10149.08	1830.40
16	CatBoost	SVM XGB CatBoost	0.986	0.975	978.56	560.27	8403.73	1145.72
17	Averaging	SVM XGB CatBoost	N/A	0.974	1010.67	350.27	8381.65	3015.26

From Table 5 it is evident that there is no clear-cut best model. The Stack 9 KNN model performs the best overall, with by far the lowest MSE of 720.16 and the best test R^2 of 0.981.

Interestingly, it only needed SVM and XGB to make its predictions, which is reasonable considering these two models deliver the strongest performance across the different steel grades.. However, this Stack is not a complete improvement over Level 0 CatBoost, even after sweeping to find the optimal k of 5. It is better at spring and carburizing steels, but has marginally worse performance for carbon steels. The performance is visualized below by Figure 5.

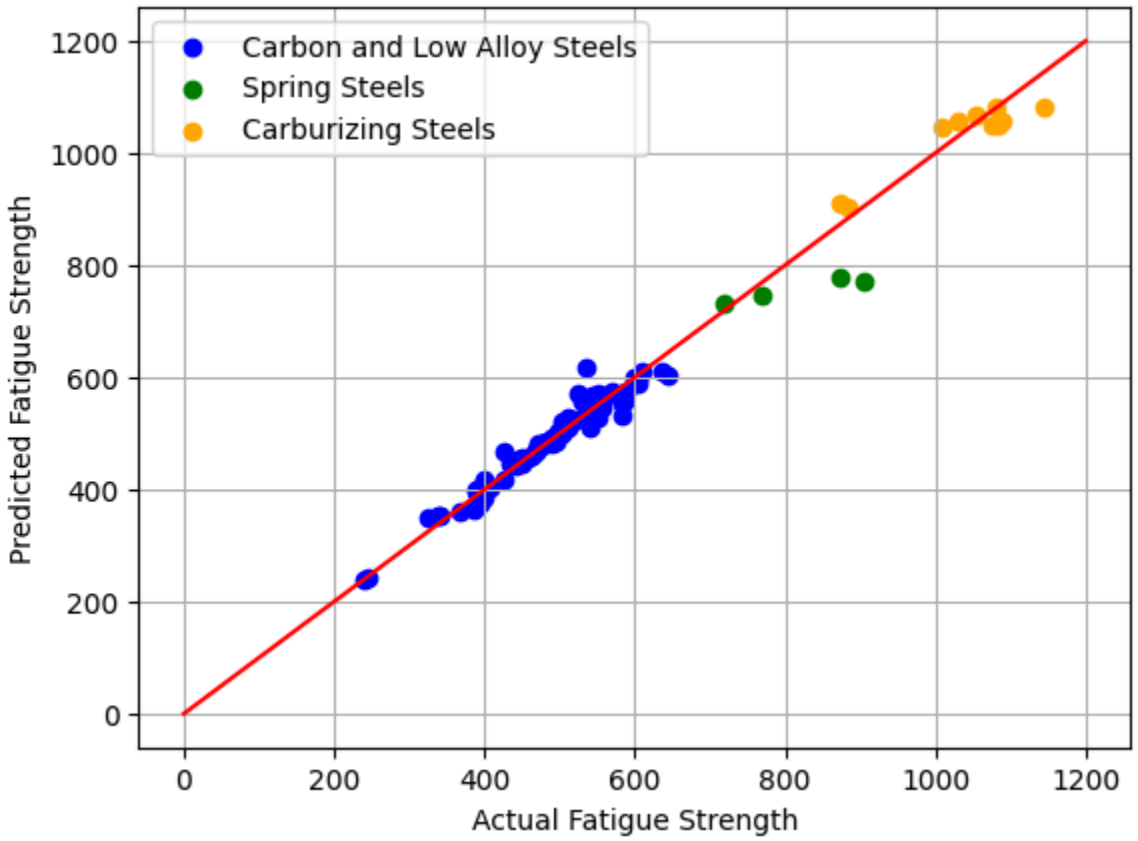

Figure 5: A scatter plot of the predicted vs actual fatigue strength values for Stack 9. The colors correspond to steel grade.

Figure 5 shows the improvement in predictions across all steel grades when compared to Figure 3, with linear regression. The overall minimization of error is further visualized by Figure 6 below. When compared to Figure 4, it is clear that the spread of data is much less across all steel grades.

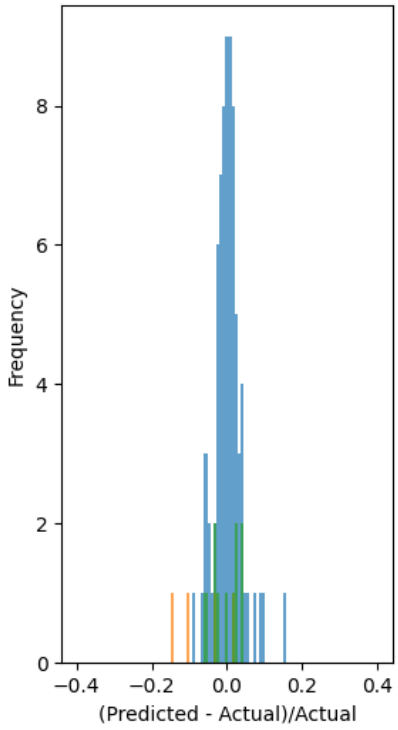

Figure 6: The error fractions of Stack 9 predictions. Blue, green, and orange correspond to carbon and low alloy steels, spring steels, and carburizing steels, respectively.

For carbon and carburizing steels, Stacks 4 and 1 respectively identified new optimal values that surpassed the performance of the Level 0 models. Both Stacks used linear regression for Level 1, but a different combination of models for level 0. No Stack reached a lower MSE than the result of the Level 0 SVM model for spring steel. The challenge in predicting spring steel fatigue strength aligns with expectations given the small sample set for this steel grade, which inherently limits the learning potential for the models.

Based on these results, the recommended optimal model would depend on the use case. Overall, Stack 9 provides the best performance across all steel grades. However, if the steel grade is already known, then to predict the fatigue strength the best performing model for the specific grade can be chosen. The recommended optimal models are Stack 4, Stack 1, and the

Level 0 SVM model for carbon and low alloy steels, carburizing steels, and spring steels, respectively.

However, these recommendations are predicated on the assumption that the dataset used is representative of the broader population. Given the relatively small size of the dataset, particularly for carburizing and spring steels, the accuracy and generalizability of these models might be constrained. Enhancing the robustness and applicability of these findings would likely benefit from repeating this analysis with a substantially larger dataset. The feature reduction conducted in this study simplifies future data collection by identifying fewer, yet more impactful, features. This not only reduces the complexity of data gathering but also reduces model training times. This study lays a solid groundwork for future research, setting a streamlined path for expanding and refining the predictive models with more extensive data.

Future Work

As mentioned earlier, the accuracy and generalizability of the models evaluated in this study could be significantly improved by replicating the analysis with a larger dataset. Additionally, a more exhaustive search for optimal parameters for each model at both Level 0 and Level 1 could further enhance performance.

Another promising direction involves the development of a hierarchical model that first classifies a sample based on its steel grade and then applies a predictive model specifically trained for that grade. This approach could leverage the distinct characteristics of each steel type to improve prediction accuracy. However, it was not pursued in this study due to the limited sample sizes for spring steel, which had only 18 samples, and carburizing steel, with just 48 samples. The small size of these subsets would not provide sufficient data to train robust classifiers or specialized predictive models effectively.

References

[1] Zhuang, Chao. “Steel Fatigue Strength Prediction.” Kaggle, 20 Mar. 2024, www.kaggle.com/datasets/chaozhuang/steel-fatigue-strength-prediction/data.

[2] Agrawal, Ankit, et al. “Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters.” Integrating Materials and Manufacturing Innovation, vol. 3, no. 1, 3 Apr. 2014, pp. 90–108, https://doi.org/10.1186/2193-9772-3-8.

[3] Hotcerts. “The 5 Feature Selection Algorithms Every Data Scientist Should Know.” Medium, Medium, 20 July 2023, hotcerts.medium.com/the-5-feature-selection-algorithms-every-data-scientist-should-know-c50b63db0ad4.

[4] T., Bex. “Powerful Feature Selection with Recursive Feature Elimination (RFE) of Sklearn.” Medium, Towards Data Science, 8 Apr. 2023, towardsdatascience.com/powerful-feature-selection-with-recursive-feature-elimination-rfe-of-sklearn-23efb2cdb54e.

[5] Yves-Laurent Kom Samo, PhD. “Boruta(SHAP) Does Not Work for the Reason You Think.” The Productive Machine Learning Engineer, The Productive Machine Learning Engineer, 3 May 2022,blog.kxy.ai/boruta-shap-is-as-bad-as-rfe/index.html#:~:text=To%20use%20a%20classification%20analogy,a%20benchmark%2C%20see%20this%20post.

[6] Polzer, Dominik. “7 of the Most Used Regression Algorithms and How to Choose the Right One.” *Medium*, Towards Data Science, 7 Aug. 2022, towardsdatascience.com/7-of-the-most-commonly-used-regression-algorithms-and-how-to-choose-the-right-one-fc3c8890f9e3.

[7] Faressayah. “XGBoost vs Lightgbm vs CatBoost vs Adaboost.” *Kaggle*, Kaggle, 8 Oct. 2023, www.kaggle.com/code/faressayah/xgboost-vs-lightgbm-vs-catboost-vs-adaboost.

[8] V, Shyam Sundar. “Improve Your Predictive Model’s Score Using a Stacking Regressor.” Analytics Vidhya, 30 Aug. 2022, www.analyticsvidhya.com/blog/2020/12/improve-predictive-model-score-stacking-regressor/.