Indian Institute of Technology Madras

Prediction of mechanical properties of Steel using ML

Group 13
Parth Sarathi Mandal
MM23M015

PS Harigovind
CE19B070

Department of Metallurgical and Materials Engineering

Introduction

The mechanical properties of steel, such as yield strength, tensile strength, and elongation, are crucial determinants of its performance and suitability for various applications across industries like construction, automotive, and manufacturing. Accurate prediction of these properties enables optimized material selection, improved product design, and cost savings. Traditionally, empirical models and constitutive equations have been employed to predict mechanical properties based on factors like chemical composition and processing conditions. However, these approaches often face limitations in capturing the complex, non-linear relationships between input features and target properties, especially for advanced steel grades with intricate microstructures.And the traditional measuring techniques are destructive which destroys the sample. With the advent of machine learning (ML) techniques, data-driven approaches have emerged as promising alternatives for material property prediction. ML models can learn and generalize from large datasets, potentially uncovering intricate patterns and relationships that are difficult to capture through traditional methods.

Literature Survey

Several studies have explored the application of ML algorithms, such as artificial neural networks (ANNs), support vector machines (SVMs), decision trees, and ensemble methods like random forests, for predicting mechanical properties of various materials, including steels. Some of them are:

(Sai et al., 2023) They investigated machine learning (ML) methodologies to predict the yield strength of neutron-irradiated F/M steels on On 460 datapoints. Popular ML algorithms, such as (RF), XGBoost, GBoost, and Support Vector Regression (SVR), were trained on a experimental dataset to understand the relationship between the input variables (e.g., irradiation dose, irradiation temperature, tensile test condition, heat treatment conditions and steels composition) and the output variable (yield strength). The XGBOOST algorithm achieved the highest PCC of 0.84 and the lowest RMSE of 90.47 MPa, followed by RF with a PCC of 0.77 and an RMSE of 106.80 MPa.[1]

(Diao et al., 2022) In this work, five machine learning algorithms are first employed to establish prediction models for different mechanical properties (tensile strength, fracture strength, Charpy absorbed energy, hardness, fatigue strength, and elongation) based on the collected carbon steels data. [2]

(Wang et al. 2023) initially employed six ML algorithms ANN, GPR ,SVM, random forests ,least squared boost tree (LSBT),KNN to predict the mechanical properties (UTS, YS, and TE) using the collected Q&T steel dataset from the Standard EN Steels handbook following data cleaning. Then the optimum GPR model was utilized. They achieved R2 score of 0.95 on GPR.[3]

(Lee et al., 2021) used 16 ML models to predict the yield and tensile strength for 5473 thermo-mechanically controlled processed (TMCP) steel alloys that were provided by Hyundai co. They achieved R2 score greater than 0.6.[4]

(Corsetti Silva & Pitz, 2020) used the neural network to predict the yield and tensile strength of steel dataset comprising of 312 datapoints on the

basis of composition only. They got R2= 0.84 for yield and 0.85 for tensile strength.[5]

Objectives

This study investigates the use of three different ML models – Linear Regression, Random Forest Regression, and XGBoost – for predicting the yield strength, tensile strength, and elongation of steel based on its composition. The primary objectives are:

1. To evaluate the performance of these ML models in predicting mechanical properties using appropriate evaluation metrics.
2. To compare the predictive capabilities of the models and identify the most effective approach for the given steel dataset.
3. To explore the impact of feature engineering techniques, such as the creation of computed features (e.g., carbon equivalents), on model performance.
4. To provide insights into the applicability and potential limitations of ML techniques for predicting mechanical properties of steel.

Methods

Dataset

The dataset used in this study consists of 312 samples of steel, each characterized by its chemical composition (weight percentages of elements were carbon, manganese, silicon, chromium, nickel, molybdenum, aluminium, vanadium, nitrogen, niobium, cobalt, tungsten, titanium).The target variables are the experimentally measured yield strength, tensile strength, and elongation values for each sample.The initail data consisted 13 input features of 312 samples of steel. The dataset was obtained from Matminer (a python library) and preprocessing steps such as handling missing values, outlier removal, and data normalization were performed before model training.

Fingerprinting of more Features

In addition to the raw input features representing chemical composition, computed features were derived based on domain knowledge and literature review. These include carbon equivalents (CE), which are empirical factors that estimate the hardenability and strength of steel based on its composition. Specifically, the following computed features were included:

Weight percentage of iron was calculated using 100-(total wt% of all elements)

- Carbon Equivalent (CEV): CEV = C + Mn/6 + (Cr + Mo + V)/5 + (Ni + Cu)/15
- Carbon Equivalent': C + Si/30 + Mn/20 + Cu/20 + Cr/20 + Mo/15 + V/10 + 5B
- The other properties were measured by measuring compositional weighted average of elemental properties.

After feature construction, total input features were 40.

Table 1: Features Constructed and formula used

S.No.	Features	Formula
1	**The atomic percentage of element**	$a_i = \frac{x_i/M_i}{\sum_i (x_i/M_i)}$
2	**Total valence electron concentration**	$\sum_i a_i VEC_i$
3	**Total atomic radii**	$\sum_i a_i r_i$
4	**Carbon equivalence**	**C + Mn/6 + (Cr + Mo + V)/5 +(Ni + Cu)/15**
5	**Carbon equivalence'**	**C + Si/30 + Mn/20 + Cu/20 + Cr/20 + Mo/15 + V/10 + 5B**
6	**Hardness**	$\sum_i c_i H_i$
7	**Melting point**	$\sum_i c_i M.P_i$
8	**Thermal Conductivity**	$\sum_i c_i T.C_i$
9	**Bulk Modulus**	$\sum_i c_i BM_i$
10	**Shear Modulus**	$\sum_i c_i SM_i$
11	**Youngs Modulus**	$\sum_i c_i YM_i$
12	**Density**	$\sum_i c_i VEC_i$

The inclusion of these computed features was motivated by their potential to capture the combined effects of multiple alloying elements on mechanical properties, thereby enhancing the feature representation and improving model performance.The physical properties off all the elements which were used in fingerprinting the features were taken from the website Periodictable.[6]

Machine Learning Models

Three different ML models were employed for predicting the mechanical properties:

Linear Regression: A simple yet interpretable model that assumes a linear relationship between input features and target variables. It serves as a baseline for comparison with more complex models.

Random Forest Regression: An ensemble learning method that combines multiple decision tree models to improve predictive accuracy and reduce overfitting. Random forests can capture non-linear relationships and are robust to outliers and noise in the data.

XGBoost (Extreme Gradient Boosting): A powerful ensemble technique that iteratively builds weak decision tree models and combines them in a boosting framework. XGBoost is known for its high predictive performance and efficient handling of sparse data.

Model Training and Evaluation

The dataset was split into training and test sets, with 80% of the data used for training the models and the remaining 20% for model evaluation and testing. Hyperparameter tuning was performed using grid search and cross-validation techniques to optimize the model parameters for each algorithm.Sci-kit learn library was used for model generation. The plots were generated using matplotlib library.For the data visualization, SweetViz library was used.

The performance of the models was evaluated using two widely adopted regression metrics:

Root Mean Squared Error (RMSE): This metric measures the square root of the average squared differences between the predicted and actual values, providing an estimate of the prediction error in the same units as the target variable.

R-squared (R^2): A statistical measure that represents the proportion of variance in the target variable that is explained by the model. Higher R^2 values (closer to 1) indicate better model fit and predictive performance.

Prediction on Candidate Set

To evaluate the predictive capabilities of the trained models, a candidate set consisting of 10,000 steel alloy data points with 40 input features was prepared. The models trained before feature elimination were used for prediction, as the feature elimination process did not consistently improve performance across all target variables.

The yield strength and elongation properties of the candidate set were predicted using the trained Ensemble models, as the ensemble models demonstrated superior performance compared to linear regression during the model evaluation phase.

The predicted values for yield strength and elongation on the candidate set can be utilized for various applications, such as rapid screening of potential steel alloy compositions, material selection to achieve desired mechanical properties.

Results and Discussion

The performance of the three machine learning models – XGBoost, Random Forest Regression, and Linear Regression – was evaluated using root mean squared error (RMSE) and R-squared (R^2) metrics for predicting the yield strength, tensile strength, and elongation of steel samples. The models were trained and tested on the steel dataset, with and without feature elimination techniques applied. Using the feature importance attribute in the sci-kit learn library for both algorithms, several features were removed that had importance less than the threshold of 10^{-3}

XGBoost

The XGBoost model demonstrated excellent predictive performance for yield strength and tensile strength, achieving R^2 values of 0.87 and 0.80, respectively, with corresponding RMSE

values of 122.92 and 132.10. For elongation prediction, the model exhibited a reasonable R^2 of 0.62 and an RMSE of 3.28. However, after applying feature elimination techniques, the model's performance slightly decreased for yield strength (R^2 = 0.86, RMSE = 126.06) and tensile strength (R^2 = 0.81, RMSE = 130.14), while elongation prediction suffered a more significant drop (R^2 = 0.42, RMSE = 4.03).

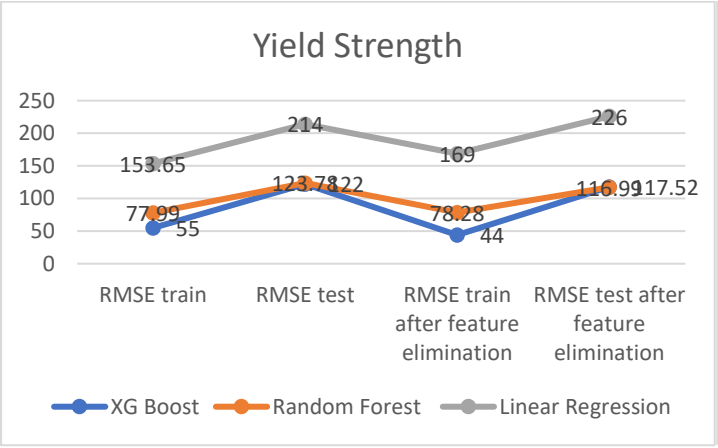
Random Forest Regression

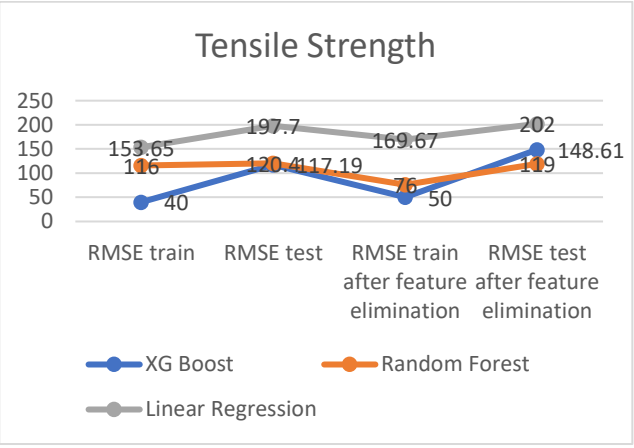
The Random Forest Regression model also displayed strong predictive capabilities, with R^2 values of 0.78 for yield strength (RMSE = 123.78), 0.86 for tensile strength (RMSE = 120.38), and 0.49 for elongation (RMSE = 3.46). After feature elimination, the model's performance improved slightly for yield strength (R^2 = 0.80, RMSE = 116.99) and tensile strength (R^2 = 0.86, RMSE = 118.95), but showed a marginal decrease in elongation prediction (R^2 = 0.46, RMSE = 3.53).

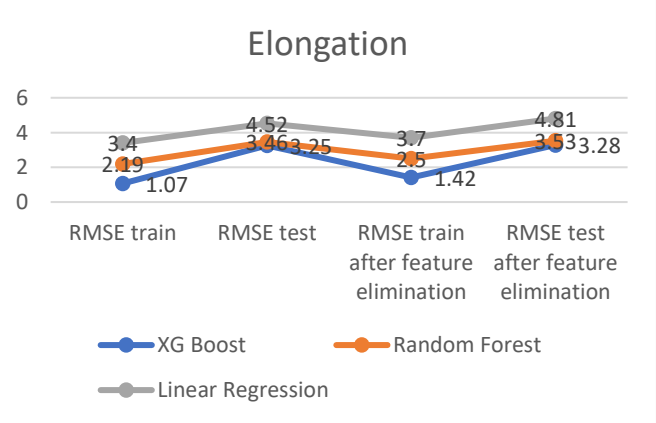
Linear Regression

As expected, the Linear Regression model exhibited the lowest overall performance among the three models. For yield strength prediction, the model achieved an R^2 of 0.54 and an RMSE of 214.22, while for tensile strength, it obtained an R^2 of 0.73 and an RMSE of 197.71. Elongation prediction was particularly challenging for the linear model, with an R^2 of 0.36 and an RMSE of 4.51.Applying feature elimination further degraded the model's performance, with R^2 values decreasing to 0.49 for yield strength (RMSE = 226.71), 0.72 for tensile strength (RMSE = 202.07), and 0.27 for elongation (RMSE = 4.81).

The parity plots between true and predicted values for yield strength, tensile strength, and elongation are included in Figures. These plots visually represent the deviations of the model predictions from the actual values, providing insights into the models' performance and potential biases or systematic errors.





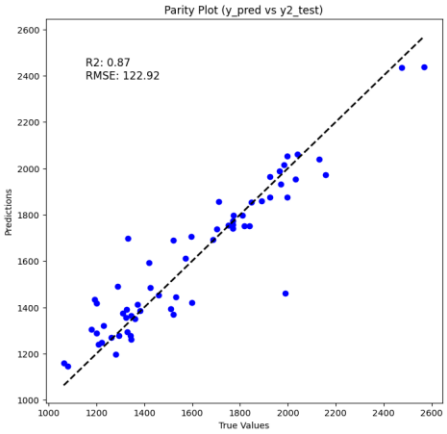


Line charts comparing the RMSE value of Train and Test set , before and after feature elimination of these 3 models on Yield strength, tensile strength and % elongation

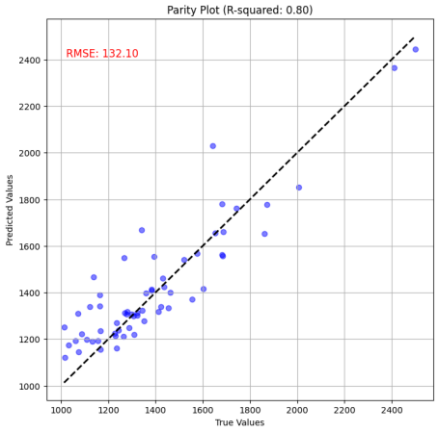
Parity plot

Parity plot is plotted between Predicted property and True property. The units of the Yield Strength and Tensile strength are in MPA while the elongation is in percentage.

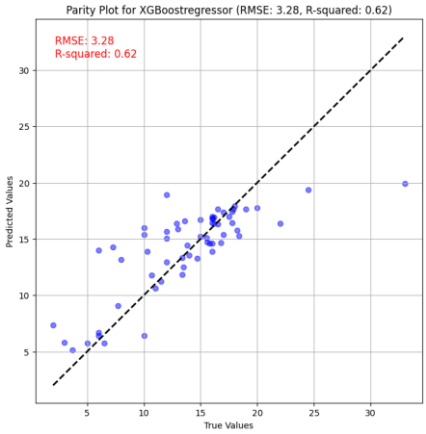
Results of XGBoost



Yield strength (MPA)

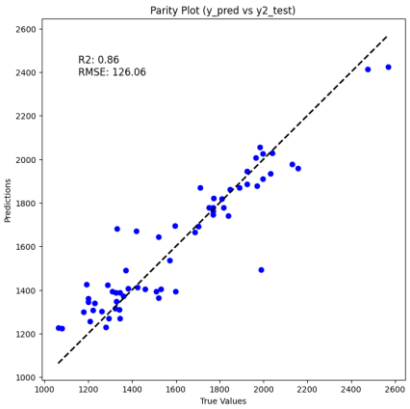


Tensile Strength (MPA)

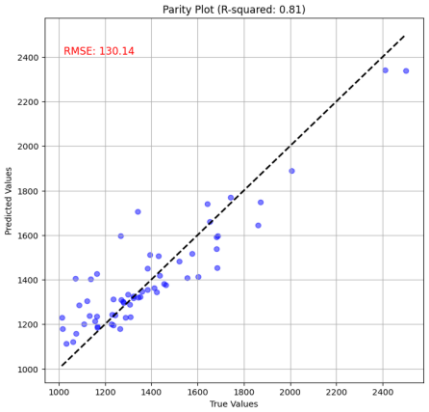


% Elongation

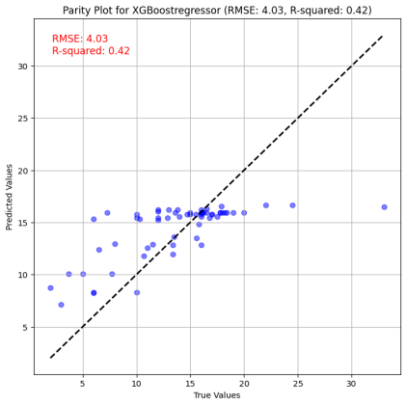
Results of XGBoost (After feature elimination)



Yield strength (MPA)

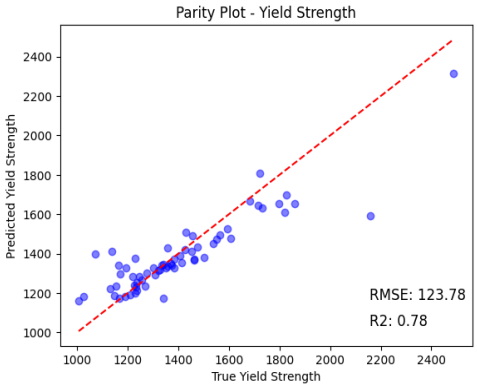


Tensile Strength (MPA)

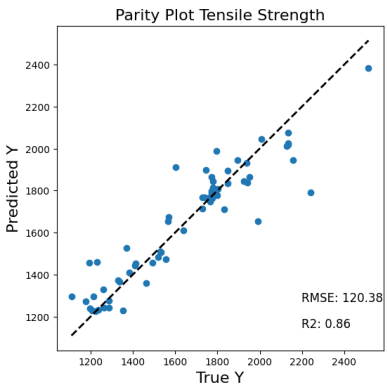


% Elongation

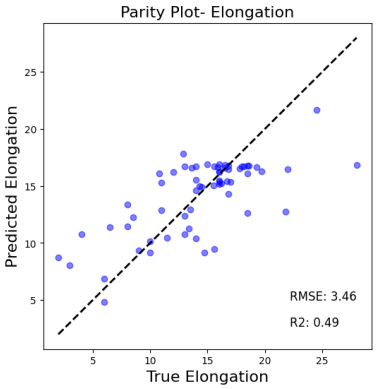
Results of Random Forest



Yield strength (MPA)

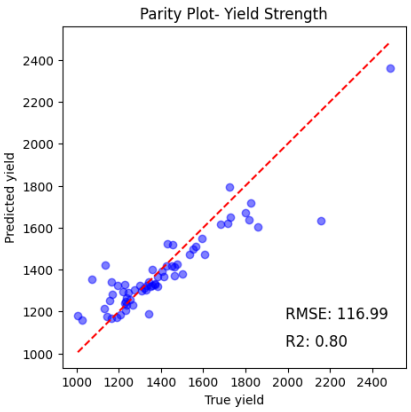


Tensile Strength (MPA)

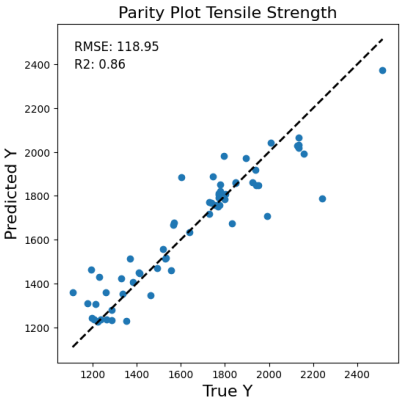


% Elongation

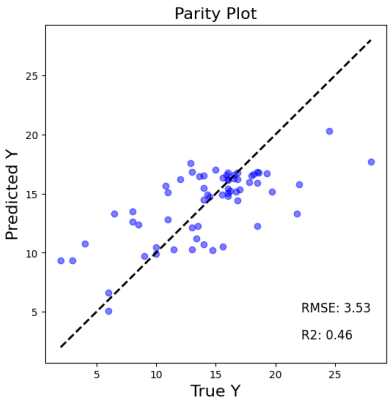
Results of Random Forest (After feature elimination)



Yield strength (MPA)

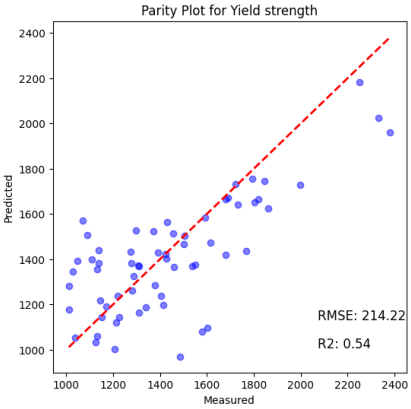


Tensile Strength (MPA)

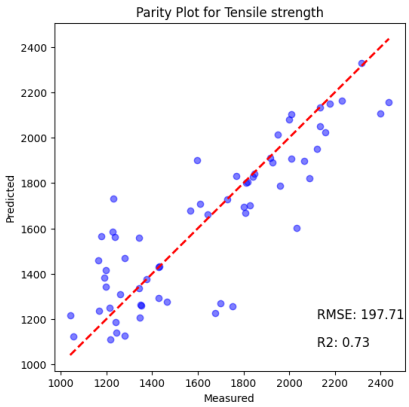


% Elongation

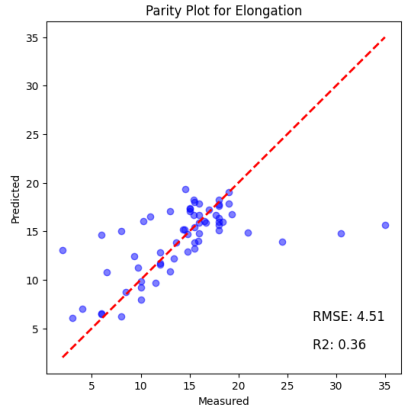
Linear Regression



Yield strength (MPA)

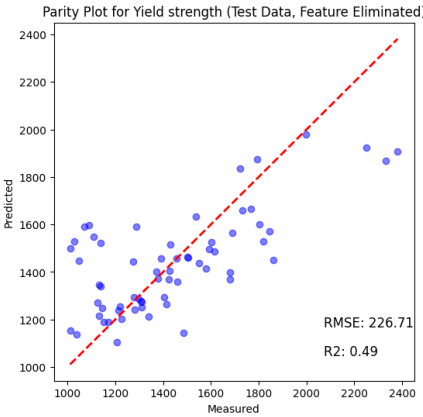


Tensile Strength (MPA)

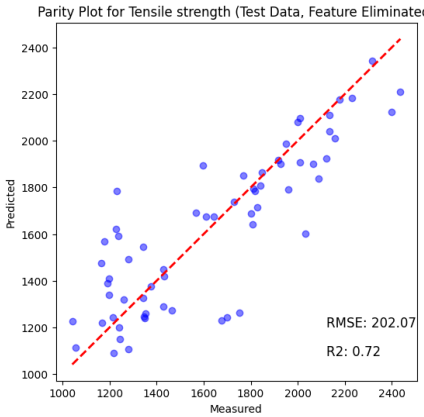


% Elongation

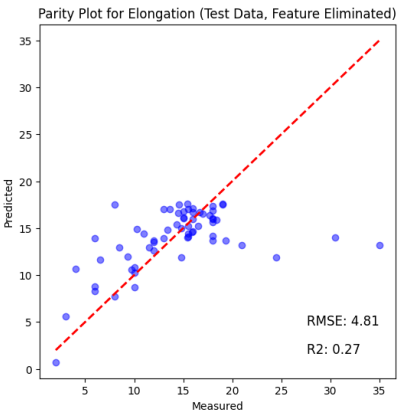
Linear Regression (After feature elimination)



Yield strength (MPA)



Tensile Strength (MPA)



% Elongation

Analysis and Discussion

The parity plots reveal interesting insights into the performance of the models. Both XGBoost and Random Forest Regression exhibit relatively low biases and systematic errors, with data points clustered around the diagonal line representing perfect prediction. However, some deviations and outliers are observed, indicating potential areas for further investigation and improvement.

The superior performance of the ensemble models (XGBoost and Random Forest) over linear regression can be attributed to their ability to capture non-linear relationships and handle complex interactions between input features. This is particularly important for predicting mechanical properties, which often have intricate dependencies on chemical.

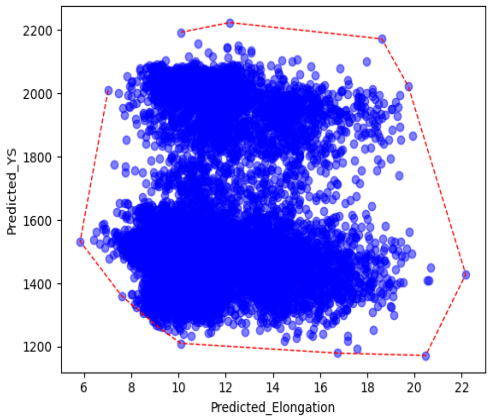
Interestingly, the feature elimination techniques did not consistently improve the model performance across all target variables. While minor improvements were observed for certain properties, the elongation prediction suffered a notable decline, particularly for the XGBoost model. This finding highlights the importance of the features and the need for domain knowledge in the feature engineering process.

It is worth noting that the computed features played a crucial role in enhancing the model performance by providing more informative representations of the chemical composition and its combined effects on mechanical properties.

Pareto Front Analysis and Composition Optimization

To further leverage the predictive capabilities of the trained models, a Pareto front analysis was performed by plotting the predicted yield strength against the predicted elongation for the candidate set

The Pareto front represents the optimal trade-off between these two conflicting mechanical properties, where improvements in one property (e.g., higher yield strength) often come at the expense of the other (lower elongation). By identifying the Pareto-optimal solutions, materials scientists and engineers can make informed decisions about the desired balance between strength and ductility based on the specific application requirements.



Parito front plot between Predicted yield strength and elongation

The Figure illustrates the Pareto front obtained by plotting the predicted yield strength and elongation values for the candidate set. Each data point on the Pareto front represents a unique steel alloy composition that offers the best possible combination of yield strength and elongation within the explored design space.

By leveraging the trained machine learning models and the Pareto front analysis, one can extract the compositions corresponding to the desired yield strength and elongation values. This capability enables rapid screening and optimization of steel alloy designs, significantly accelerating the material development cycle.

Conclusions

In this study, three machine learning models – XGBoost, Random Forest Regression, and Linear Regression – were employed to predict the yield strength, tensile strength, and elongation of steel based on its chemical composition and other features. The results demonstrated the effectiveness of machine learning, particularly ensemble methods like XGBoost and Random Forest Regression, in accurately predicting mechanical properties.

Key conclusions drawn from the study are:

1. The ensemble models (XGBoost and Random Forest Regression) outperformed the linear regression model, exhibiting higher R^2 values and lower RMSE values for predicting yield strength, tensile strength, and elongation.

2. The ability of ensemble models to capture non-linear relationships and handle complex interactions between input features contributed to their improved predictive performance.
3. Feature engineering, particularly the inclusion of computed features like carbon equivalents, valence electron concentration played a crucial role in enhancing model accuracy by providing more informative representations of the input data.
4. Feature elimination techniques did not consistently improve model performance across all target variables, indicating the need for careful feature selection and the incorporation of domain knowledge.
5. The developed machine learning models offer a data-driven approach that can complement or augment traditional methods for predicting mechanical properties of steel, enabling rapid screening and optimization of steel compositions.

Overall, this study highlights the potential of machine learning techniques to accurately predict mechanical properties of steel, offering a data-driven approach that can complement or augment traditional methods. The findings have practical implications for material selection, product design, and optimization in various industries relying on steel applications.

Future Work

While this study demonstrated promising results, there are several avenues for future research and improvements:

- Exploring other ML algorithms, such as gradient boosting machines, support vector machines, or deep learning techniques, and comparing their performance with the models evaluated in this study.
- Exploring the generation of additional descriptors or input features beyond the chemical composition and computed features used in this study. Incorporating a more comprehensive set of descriptors could potentially capture complex relationships more effectively, further improving the accuracy and robustness of the models, particularly for predicting the elongation property.
- Exploring active learning techniques, where the ML models can iteratively select the most informative data points for labeling and incrementally improve their predictive performance, potentially reducing the need for extensive experimental data collection.
- Extending the approach to predict other relevant material properties, such as fatigue life, corrosion resistance, or microstructural characteristics, by incorporating additional input features and adapting the ML models accordingly.
- Integrating the developed ML models into a decision support system or material design platform, enabling rapid screening and optimization of steel compositions to achieve desired mechanical properties.

Contributions from the Author

Parth Sarathi Mandal (MM23M015) – Data collection, data preprocessing,Fingerprinting feature, Model implementation- Linear regression and Random Forest, analysis, preparation of candidate set of 10000 datapoints, prediction on candidate set,parito front analysis and Report Writing.

PS Harigovind (CE19B070) – Data Visualisation, Used LightBGM,Extra tree regressor and XGBoost for best result- Finally selected XGBoost for model implementation, prediction of candidate set.

References

1. Corsetti Silva, G., & Pitz, D. B. (2020). Prediction of yield and tensile strengths for high-alloy steels from chemical composition: a data preprocessing approach. *XXVII Congresso Nacional de Estudantes de Engenharia Mecânica*. https://doi.org/10.26678/ABCM.CREEM2020.CRE2020-0002
2. Diao, Y., Yan, L., & Gao, K. (2022). A strategy assisted machine learning to process multi-objective optimization for improving mechanical properties of carbon steels. *Journal of Materials Science & Technology*, *109*, 86–93. https://doi.org/10.1016/j.jmst.2021.09.004
3. Lee, J.-W., Park, C., Do Lee, B., Park, J., Goo, N. H., & Sohn, K.-S. (2021). A machine-

learning-based alloy design platform that enables both forward and inverse predictions for thermo-mechanically controlled processed (TMCP) steel alloys. *Scientific Reports*, *11*(1), 11012. https://doi.org/10.1038/s41598-021-90237-z

4. Sai, N. J., Rathore, P., Sridharan, K., & Chauhan, A. (2023). Machine learning-based predictions of yield strength for neutron-irradiated ferritic/martensitic steels. *Fusion Engineering and Design*, *195*, 113964. https://doi.org/10.1016/j.fusengdes.2023.113964
5. Wang, S., Li, J., Zuo, X., Chen, N., & Rong, Y. (2023). An optimized machine-learning model for mechanical properties prediction and domain knowledge clarification in quenched and tempered steels. *Journal of Materials Research and Technology*, *24*, 3352–3362. https://doi.org/10.1016/j.jmrt.2023.03.215

6. https://periodictable.com