Jurnal Polimesin
Department of Mechanical Engineering
State Polytechnic of Lhokseumawe
http://e-jurnal.pnl.ac.id/polimesin
e-ISSN : 2549-1999 No. : 5 Month : October
p-ISSN : 1693-5462 Volume : 22 Year : 2024

Article Processing Dates: Received on 2024-06-02, Reviewed on 2024-09-15, Revised on 2024-10-09, Accepted on 2024-10-12 and Available online on 2024-10-30

Optimizing prediction of stainless steel mechanical property with random forest: a comparative study of feature selection methods

Maimuzar[1], Hendra[1], Syarif Khan[2], Desmarita Leni[3]*, Islahuddin[4]

[1]Teknik Mesin, Politeknik Negeri Padang, Padang, 25164, Indonesia
[2]AABCO Engineering, Machine Technology SDN BHD, Malaysia
[3]Teknik Mesin, Universitas Muhammadiyah Sumatera Barat, Bukittinggi, 26181, Indonesia
[3,4]Fakultas Farmasi, Sains dan Teknologi, Universitas Dharma Andalas, Padang, 25171, Indonesia
*Corresponding author: desmaritaleni@gmail.com

Abstract

Predicting the mechanical properties of stainless steel, such as Yield Strength (YS), Ultimate Tensile Strength (UTS), and Elongation (EL), requires many input variables, such as chemical composition, type of heat treatment, heating duration, and cooling method. However, the complexity and number of these variables can increase processing time and reduce model accuracy. This study aims to explore the impact of selecting the most influential input variables to improve prediction accuracy. It compared two feature selection techniques to enhance prediction accuracy: Recursive Feature Elimination (RFE), which systematically excludes less relevant features, and Information Gain (IG), which evaluates each variable's contribution to predictions. Both techniques were implemented using the random forest algorithm, chosen for its robustness in handling large datasets and its ability to capture complex interactions between variables. Parameter optimization was performed using a grid search. The analysis showed that the RFE-based model out performed both the IG-based model and the model without feature selection. In predicting YS, RFE identified 13 out of 21 influential variables, achieving a Mean Absolute Error (MAE) of 9.91, Root Mean Square Error (RMSE) of 14.20, and R-squared value of 0.89. For UTS, RFE identified 8 out of 21 variables, with an MAE of 12.89, RMSE of 16.97, and R-squared of 0.97. In predicting EL, RFE identified 14 out of 21 variables, with an MAE of 3.82, RMSE of 6.10, and an R-squared value of 0.85. The high R-squared values (> 0.85) across all properties indicate the model's strong predictive capabilities, supporting its practical use in stainless steel property prediction.

Keywords:
Optimization, stainless steel, mechanical properties, random forest.

1 Introduction

Stainless steel is a type of steel that is resistant to corrosion, oxidation, and other chemical reactions thanks to its additional metal content, especially chromium. Chromium forms an oxide layer that protects the steel surface, resulting in rust-resistance [1]. Stainless steel is commonly used in various construction projects, such as multi-story building structures, bridges, and household appliances [2]. In the construction field, stainless steel is often used to build structures that are resistant to corrosion and mechanical loads. Its tensile performance is crucial for ensuring adequate structural strength and durability in a variety of projects, including bridges and multistory buildings [3]. A good understanding of the tensile strength of stainless steels can help reduce the risk of premature material failure [4]. According to Morini (2019) [5], knowledge of the mechanical properties of a material is not only necessary to prevent premature failure of machines or industrial components but also for user safety aspects. The tensile strength of stainless steel is influenced by its microstructure, chemical elements, and heat-treatment temperature [6] [7]. Many complex factors influence the mechanical properties of stainless steel and are nonlinear, such as the chemical composition of chromium, nickel, and molybdenum, which can have a complex and nonlinear impact on the mechanical properties of stainless steel. Changes in the percentage of these chemical elements can have a disproportionate effect on the tensile strength, such as the Yield Strength (YS), Tensile Strength (TS), and Elongation (EL) [8], [9]. Heat-treatment processes, such as controlled heating and cooling, play a key role in changing the microstructure and mechanical properties [10]. Meanwhile, the austenitic or ferritic structure of stainless steel can provide different mechanical characteristics, including tensile properties [11]. Based on the complexity of the influence of chemical elements and heat treatment temperature on the mechanical properties of stainless steel, an effective and efficient method is required to comprehensively understand the mechanical properties of stainless steel.

The rapid development of computer technology in the field of materials science has encouraged experts and researchers to develop computational approaches for analyzing and solving various problems in the field of materials [12]. Machine learning is a method that uses data and algorithms to imitate the way humans learn. Machine learning allows computers to make decisions or predictions based on patterns found in the data. This method is one of the most popular topics in material science. This can be seen on the web of science, which states that almost 2000 papers were published on this topic in 2020 alone, compared to only approximately 400 papers in 2017 [13]. In materials science, machine learning is widely used to model material properties, design new materials, and optimize their mechanical properties. In models predicting the mechanical properties of materials such as alloy steel, it is generally necessary to select algorithms, set model parameters, and select features. Feature selection is the process of selecting the most relevant and important subset of features from a set of available features in a dataset. The main goal of feature selection is to improve model performance and computational efficiency by reducing data complexity without sacrificing accuracy [14]. By selecting the most informative features, the learning process can become more efficient and the model can overcome problems such as overfitting. Recursive Feature Elimination (RFE) and Information Gain (IG) are two methods used to select the input variables that have the most influence on the output variables [15]. For predicting the mechanical properties of stainless steel, a feature selection method is most appropriate for selecting the chemical elements and heat treatments that are most relevant to the mechanical properties of stainless steel. Understanding the influence of feature selection on the prediction of steel tensile strength has been the main focus of many previous studies, such as Jiang et al. (2020) [16], who conducted a comprehensive study on the use of RFE on a large steel wire dataset. Their findings highlight that RFE not only improves the prediction accuracy but also significantly reduces the risk of overfitting. This research confirms the success of RFE as an effective feature selection tool for dealing with the complexity of large and complex material datasets. In another study [17] on the prediction of the fatigue strength of alloy steels, it was found that

information gain makes a valuable contribution to improving the accuracy of fatigue strength prediction of alloy steels. This method proved effective in reducing the dimensionality of complex mechanical property datasets by focusing on features that have a significant impact on the fatigue strength of alloy steel. Ruiz et al. (2020) [18] investigated the use of RFE with various machine learning algorithm models in predictive modeling of the tensile strength of steel bars produced in an electric arc furnace. In this study, there were 97 features in the dataset during the steel-processing process. The research results show that RFE with random forest can determine the features that are most relevant to the output variable and provide the most accurate prediction results compared to other algorithm combinations. Based on the results of previous research, feature selection is an important step in improving prediction models for material mechanical properties. Therefore, this study was designed to compare the effectiveness of two feature selection methods, Recursive Feature Elimination (RFE) and Information Gain (IG), combined with the Random Forest algorithm to predict the mechanical properties of stainless steel. It is expected that the results of this study will provide a strong foundation for understanding the performance comparison between RFE and IG in the context of mechanical property prediction. These findings are expected to guide the selection of the most appropriate feature selection method to obtain optimal predictions of the mechanical properties of stainless steel.

2 Research Methods

This study compares two feature selection techniques combined with a random forest algorithm predictive model to predict the mechanical properties of stainless steel. Generally, to predict the mechanical properties of stainless steel, such as the Yield Strength (YS), Ultimate Tensile Strength (UTS), and Elongation (EL), many input variables are required, such as chemical elements, type of heat treatment, length of heating time, and cooling method. This large and complex number of variables can result in long processing times and reduce the accuracy of the prediction model. Therefore, this research compares two feature selection techniques commonly used in machine learning methods: Recursive Feature Elimination (RFE) and Information Gain (IG). The goal was to improve the prediction of the mechanical properties of stainless steel. The developed model was evaluated using three evaluation metrics that are often used to predict the mechanical properties of alloy steels [8], [17], and [19]: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared. The model was created using the Python programming language and run on the Google Colab framework. The stages of this research are shown in Fig. 1, which illustrates the stages in the analysis and development of a prediction model for the mechanical properties of stainless steel.

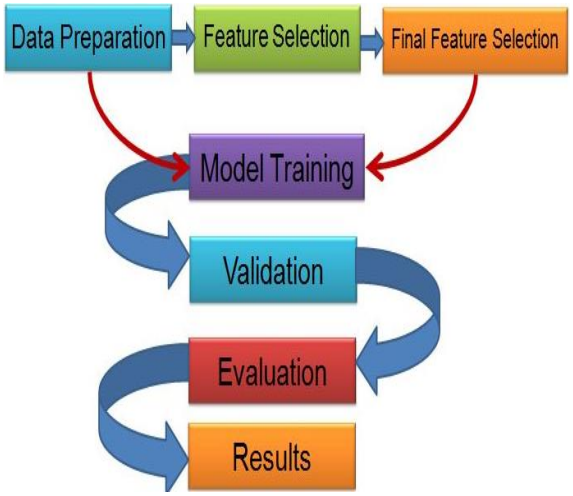

Fig. 1. Research scheme.

2.1 Datasets

The dataset used in this study was the tensile test results of several types of Austenitic Stainless Steel (ASS), such as SUS 304, SUS 316, SUS 321, SUS 347, and NCF 800H. This dataset consists of 2180 samples with austenitic stainless steel mechanical properties, alloy chemical elements, heat treatment temperatures, and cooling methods. The data obtained from the Creep Data Sheet of Steel (No. 4B, 5B, 6B, 14B, 15B, 26B, 27B, 28B, 32A, 42, and 45), which is a data source from NIMS MatNavi and BSCC High Temperature Data from the British Steelmakers Creep Committee [20]. The data were collected by the Material Algorithm Project (MAP) [21], which is a project carried out by the University of Cambridge that can be used for research and educational purposes.

2.2 Data Preprocessing

Data preprocessing is an initial stage in data analysis that aims to clean, prepare, and organize raw data so that they can be used more effectively in statistical analysis or modeling [22]. At this stage, the identification and handling of missing or invalid values in the dataset are carried out, such as replacing missing values, deleting irrelevant rows or columns, handling significant outliers, and changing the data format to suit analysis needs, such as by normalizing the data. Data normalization is a data preprocessing technique that aims to change the data scale to the same or equivalent range from 0 to 1 [23]. Data normalization was performed using the MinMaxScaler method. This normalization method is the most commonly used method in machine learning, and can be calculated using Eq. 1.

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X is the original data, while [24].

2.3 Random Forest Modeling with Feature Selection

At this stage, the random forest model is first trained using all dataset features to obtain an initial picture. Next, hyperparameter optimization is carried out with a grid search to find the best parameters in the Random Forest algorithm, which will be used to retrain the model. To increase the reliability of the model evaluation, K-fold cross-validation was applied with K = 10. The use of K-fold cross-validation with K = 10 provides a more stable estimate of model performance, minimizing the possibility of bias from one test that depends on partition-specific data [8]. After obtaining the best parameters, the next step is to apply two feature selection methods: Recursive Feature Elimination (RFE) and Information Gain (IG). In general, RFE works in the following ways: (1) initialize the machine learning model with all the features in the dataset, (2) calculate the importance of each feature in the model, (3) delete features that are not significant to the output variable, and (4) repeat steps 2 and 3 until the desired number of features is reached [25]. IG identifies features that have high informativeness regarding the target mechanical properties. IG can be calculated using Eq. 2.

$$Gain\ (A) = Entropy\ (s) \sum_{i=1}^{k} \frac{|s_i|}{|s|}\ x\ Entropy\ (s_i) \quad (2)$$

Where S is a dataset, A is an attribute, |Si| is a subset of the dataset S, where the value of attribute A is equal to the ith value, |S| is the number of all data samples, and entropy (Si) is the entropy of the subset s_i, which has an attribute value A equal to the i value [26].

2.4 Model Evaluation

Model evaluation is the process of assessing the performance of a machine learning model using certain metrics or indicators. The goal of model evaluation is to understand the extent to which the model can generalize and make accurate predictions on never-

before-seen data [27]. The random forest model with the best combination of feature selection was evaluated through external testing using new data. The results of this external testing were calculated using evaluation metrics such as the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2). The selection of these three metrics is designed to provide holistic and comprehensive information regarding the performance of models predicting the mechanical properties of alloy steels. This evaluation metric has also been used in previous research on predicting the mechanical properties of alloy steels [8], [14], and [23]. MAE provides an understanding of overall prediction accuracy; RMSE places emphasis on handling large errors; and R-squared provides an idea of the extent to which the model is able to explain variations in mechanical property data. By using these three metrics, model evaluation can be better carried out from various perspectives relevant to research goals and practice in the field. Model evaluation metrics can be calculated using the Eq. 3-Eq. 5 [8]:

1. Mean Absolute Error (MAE)

$$MAE = \frac{i}{N}\sum|y_i - z_i| \quad (3)$$

where i is the index of the data sample, N is the total number of samples, y_i is the actual value of the i-th data point, and z_i is the value predicted by the model for the i-th data point.

2. Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2} \quad (4)$$

where n is the number of data points used to test the model, $f(X_i)$ is the value predicted by the model for the i-th data point, and Y_i is the actual value for the i-th data point.

3. R-squared

$$R = \frac{\sum_{i}^{n}=1(f(X_i)-f(\bar{X}))(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(f(X_i)-f(\bar{X}))^2}\sqrt{\sum_{i=1}^{n}Y_i-\bar{Y})^2}} \quad (5)$$

where $f(X_i)$ is the predicted value of the dependent variable (Y) based on the independent variable (X) for the i-th observation, $f(\bar{X})$ is the average of all predicted values $f(X_i)$ across all observations, Y_i is the actual observed value of the dependent variable for the i-th observation, $\bar{Y}$ is the average of all observed values Y_i across all observations, and n is the total number of observations.

The selection of the best feature selection method between RFE and IG was determined based on an R-squared value above 0.8, as well as MAE and RMSE values below 20. These standards are used to ensure that the model achieves a high level of accuracy and can capture most of the data variability with minimal prediction error. Additionally, for the IG method, variables related to chemical composition and heat treatment with an information gain score below 0.1 are not used as inputs in the model training. This is because low-scoring variables tend to contribute insignificantly to predictions, so excluding them can reduce model complexity and improve the efficiency of the training process without sacrificing accuracy.

3 Results and Discussion

3.1 Datasets

This austenitic stainless steel dataset consists of 2180 samples; however, after data preprocessing, there were 1194 data samples that had unnecessary information, such as missing and invalid values, so they were not used in this research. In the original database, there are several other features, such as melting type, grain size, and product shape, but these data are incomplete and have a very low correlation with the mechanical properties of stainless steel; therefore, they were not used in this study. The data with complete and relevant information for this research only amount to 986 samples consisting of independent and dependent variables. The independent variables are chemical elements and heat treatment temperature, totaling 14 variables, as shown in Table 1, while the dependent variables are the mechanical properties of stainless steel, which consist of Yield Strength (YS), Ultimate Tensile Strength (UTS), and Elongation (EL).

Table 1. Austenitic Stainless Steel (ASS) dataset variables

Variables	Variables
Chromium (Cr, wt%)	Carbon (C, wt%)
Nickel (Ni, wt%)	Boron (B, wt%)
Molybdenum (Mo, wt%)	Phosphorus (P, wt%)
Manganese (Mn, wt%)	Sulfur (S, wt%)
Silicon (Si, wt%)	Solution treatment temperature (Ts, K)
Niobium (Nb, wt%)	Solution treatment time (ts, s)
Titanium (Ti, wt%)	Water-quenched or air-quenched

3.2 Data Preprocessing

This stainless steel dataset was normalized using the min–max scalar method. The goal was to change the values in the dataset such that they had a uniform range between 0 and 1. This method is often used in various cases of predicting the mechanical properties of alloy steels [23], [28], and [29], where this method can improve the performance of machine learning models. This is because normalization helps avoid large gradient problems that can arise when variables have different scales. Apart from that, basically, machine learning models pay more attention to variables on a larger scale and ignore variables on a smaller scale, so this method is very helpful in improving the performance of machine learning models. An illustration of normalization using the MinMaxScaler method as shown in Fig. 2.

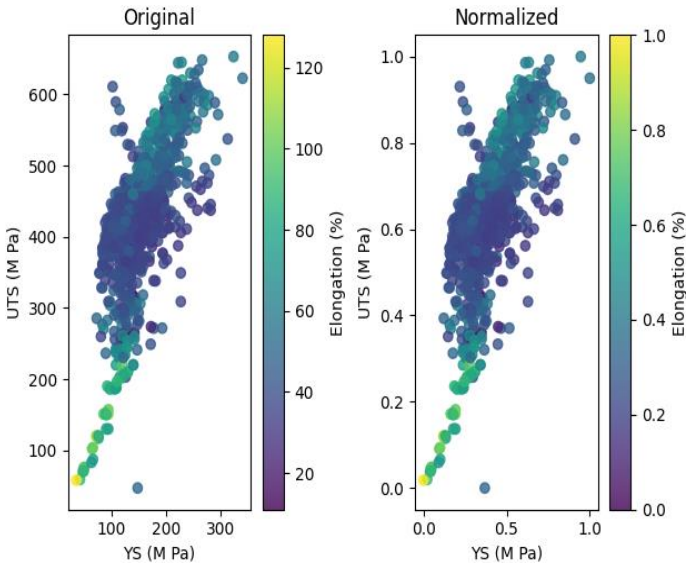

Fig. 2. Illustration of normalization using the MinMaxScaler method.

Based on Fig. 2, changes in the variable values of the mechanical properties of stainless steel can be seen, where the UTS initially had a value range from 100 to 600, YS 100 to 300, and EL 20 to 120. However, after being normalized using the MinMaxScaler method, the values of the three properties change mechanically to a range of 0 to 1. This normalization process was also applied to chemical elements and heat treatment. At this stage, the stainless-steel dataset was divided into two parts: training and external testing. The total amount of data that had been cleaned reached 986 samples, which were then reduced to 50 samples for external testing purposes. Therefore, 936 training data points were used. These training data are then divided back into

two parts: 80% for the training process and 20% for validation purposes.

Before training the model, it was necessary to carry out a correlation analysis to determine how the relationship between chemical elements and heat treatment affects the mechanical properties of stainless steel for each variable. In this study, correlation analysis was performed using Pearson correlation. Pearson's correlation is a statistical method used to measure the extent to which two variables are correlated or have a linear relationship with each other. This metric produces a correlation coefficient (r) ranging from -1 to 1. An r-value close to 1 indicates a perfect positive correlation, indicating that there is a positive linear relationship between the two variables. When one variable increases, the other also tends to increase. The r value is close to -1, indicating perfect negative correlation, which means that there is a negative linear relationship between the two variables. When one variable increases, the other tends to decrease. The r value is close to 0, which indicates that there is no linear relationship between the two variables [8]. The results of the data analysis of the stainless-steel tensile test are shown in Fig. 3.

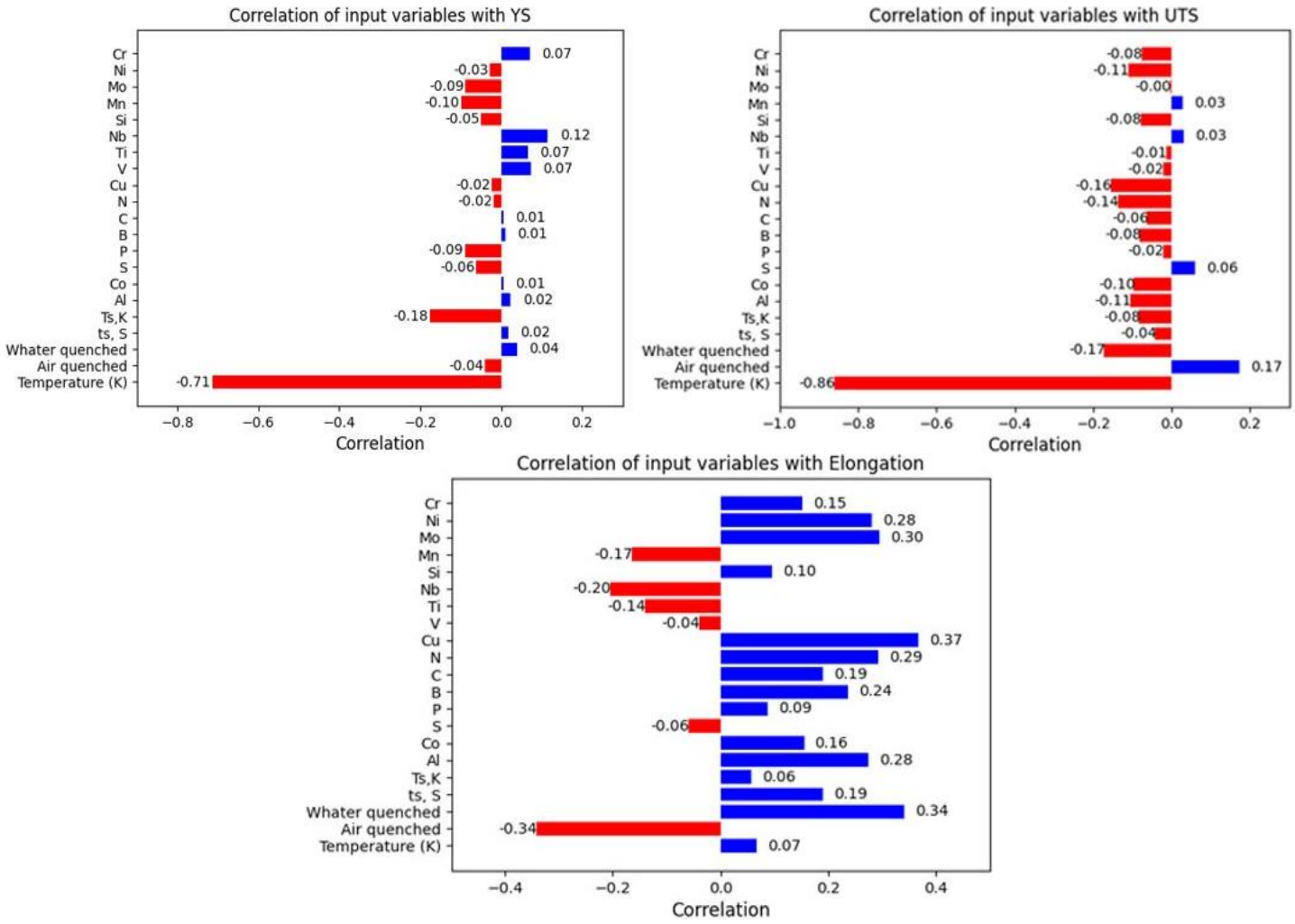

Fig. 3. Relationship between chemical elements and heat treatment with mechanical properties.

Based on the results of the Pearson correlation analysis, it can be concluded that temperature has a significant relationship with the mechanical properties of stainless steel. It was found that temperature has a very strong negative correlation with Yield Strength (YS) and Ultimate Tensile Strength (UTS), with correlation values of -0.71 and -0.86, respectively. This indicates that increasing the temperature during the heat treatment process causes a decrease in the strength and ductility of the material. On the other hand, Elongation (EL) shows a very weak positive correlation with temperature, with a correlation value of 0.07. The strong negative correlation between the temperature and YS and UTS is consistent with the structural phenomena and phase transformations that occur during the heat treatment of austenitic stainless steels [30]. Increasing the temperature during heat treatment can result in structural hardening, with the formation of new phases with different mechanical properties. As a result, high temperatures can reduce the strength of the material by affecting the distribution of atoms in the crystal or reducing the content of structural elements. In addition, heat treatment temperature can also affect phase transformations such as carbide deposition, which can reduce the ductility and elasticity of the material [31]. However, the very small positive correlation between EL and heat treatment temperature indicates that the influence of temperature on elongation is relatively low. This shows that changes in temperature tend to have minimal impact on the level of deformation or ductility of the material. Thus, understanding these relationships is important in planning appropriate heat treatment processes to meet specific needs in material applications.

Chemical elements such as copper (Cu) and nickel (Ni) have a positive correlation with Elongation (EL), which indicates that the higher the concentration of these chemical elements, the higher the elongation value. The research results of Niu et al. (2018) [32] regarding the influence of copper in enhancing the effect of Transformation-Induced Plasticity (TRIP) in stainless steel found that the addition of Cu had a significant impact on elongation in stainless steel. This research highlights the role of Cu in accelerating the kinetics of austenite reversion, namely the transformation of austenite back into its original form. Cu acts as a heterogeneous crystallizer and provides the necessary chemical conditions through interfacial segregation, ultimately enhancing the austenite formation. On the other hand, the Ni content in stainless steel can reduce the degree of martensitic transformation, which in turn can increase elongation and formability. Thus, understanding the role of Cu and Ni in the structure and

mechanical properties of stainless steel is an important factor in designing materials that have the desired performance. The implication is that controlling the concentration of these elements can be used as a strategy to modify the mechanical properties of materials with the aim of increasing their elasticity and deformability.

Cooling methods such as water quenching and air quenching play an important role in determining the mechanical properties of stainless steel, especially elongation. The results of the analysis using Pearson correlation show that elongation has a positive correlation with quenched water of 0.34 and a negative correlation with quenched water of -0.34. The increase in elongation that occurs due to the water quenching of stainless steel can be explained by the rapid cooling process. This rapid cooling prevents or minimizes the formation of undesirable phase precipitates such as carbides or nitrides, which can inhibit dislocation movement and reduce plastic deformation in the material. By preventing the formation of these phases, water quenching can help increase the plastic deformation of stainless steel [33]. Apart from that, water quenching can also produce a microstructure that is more homogeneous and free from structural imperfections. The rapid cooling process allows the atoms in the material to lock into more regular positions, thereby reducing the dislocations and structural defects. This can result in materials with better strength and higher deformability, which in turn contributes to increased elongation. On the other hand, cooling with air produces a slower cooling process compared to cooling with water, which results in the formation of a more complex and non-uniform microstructure. The microstructure formed may include undesirable phases or structural defects, such as carbides or nitrides formed during cooling. Therefore, the selection of an appropriate cooling method can significantly influence the mechanical properties of stainless steel, especially elongation [34].

3.3 Random Forest Modeling with Feature Selection

In model testing, cross-validation techniques are used to ensure model accuracy. Cross-validation is an evaluation technique employed to measure a model's performance by dividing the data into two parts: training data and testing data. In cross-validation, the training data is split into several different subsets or folds, and each subset is iterated as the testing data, while the remaining subsets are used as training data. In this modeling, 10-fold cross-validation is applied, where the data is divided into 10 different subsets or folds and iterated 10 times by selecting each subset in turn as the testing data and the remaining subsets as the training data. The evaluation results from each iteration will be averaged to obtain a more valid evaluation metric. Fig. 4 illustrates the use of 10-fold cross-validation.

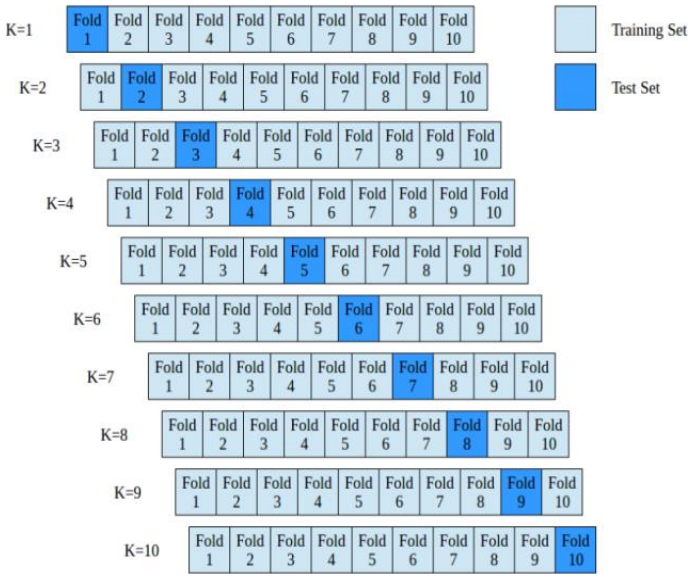

Fig. 4. Illustration of cross-validation.

In this stage, the modelling of mechanical properties of stainless steel is conducted by dividing the process into three distinct parts. In the first part, the Random Forest (RF) model is evaluated using all input variables available in the dataset. To enhance model performance, parameter optimization is performed via grid search to identify the optimal parameters for predicting mechanical properties. The grid search process specifically examines the number of estimators parameter, exploring values ranging from 5 to 100. This approach can mitigate overfitting; by employing a greater number of decision trees, the model becomes more robust and is better equipped to reduce the risk of overfitting. Overfitting arises when the model is excessively complex, resulting in an overly precise fit to the training data, which hinders its ability to make accurate predictions on previously unseen data [35].

In some cases, increasing the number of estimators can improve model performance, especially if the data has high complexity. However, it is also often found that there is a point where adding an estimator does not provide significant improvements and can even increase the computational load [36]. In this research, parameter searches were carried out separately for each mechanical property, namely Yield Strength (YS), Ultimate Tensile Strength (UTS), and Elongation (EL). A visualization of the process of searching for the best parameters for each mechanical property can be found in Fig. 5. Meanwhile, the best number of estimators obtained from the grid search as shown in Fig. 6.

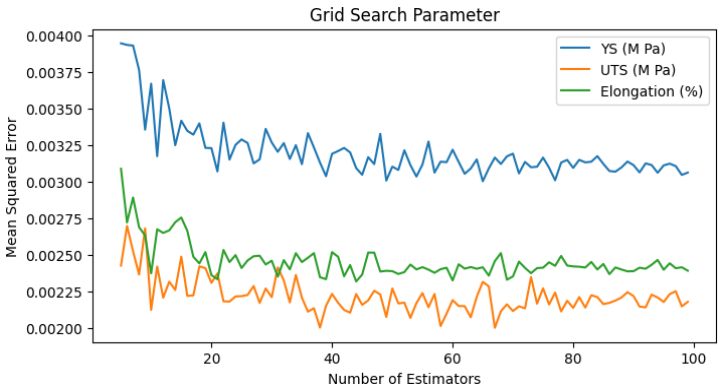

Fig. 5. Changes in MSE for each variation of number of estimators.

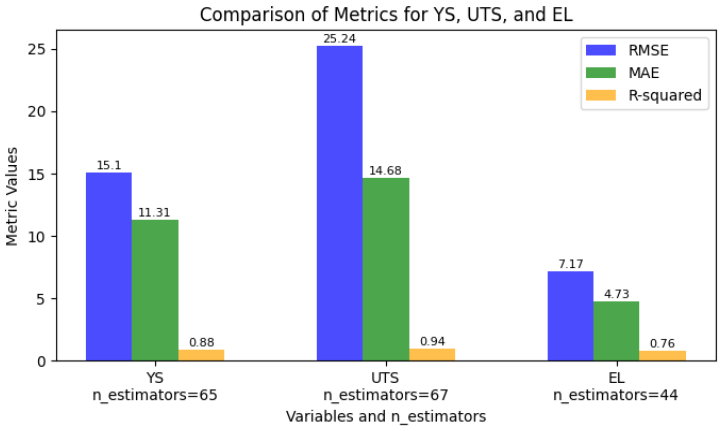

Fig. 6. Results of the best model parameters.

Fig. 5 presents an overview of the model's performance for each mechanical property, highlighting variations in the number of estimator values. As depicted in Fig. 6, the model exhibits its highest performance in predicting Ultimate Tensile Strength (UTS), achieving an R-squared value of 0.94 with 67 estimators. In contrast, Yield Strength (YS) attains its highest R-squared value of 0.88 with the same number of estimators. Elongation (EL) yields an R-squared value of 0.76 when the number of estimators is set to 44.

In the subsequent stage, following the testing of the model utilizing all features and identifying the optimal parameters, the next step involves integrating the best model identified through the Recursive Feature Elimination (RFE) feature selection method. During this process, the number of features employed ranged from 5 to 21. This experiment was conducted separately for the three mechanical properties of stainless steel. RFE plays a crucial role in identifying the features that most significantly influence the prediction of mechanical properties, thereby providing a deeper understanding of the dominant factors affecting model performance. The results of the evaluation metrics for the three mechanical properties of stainless steel are presented in Fig. 7.

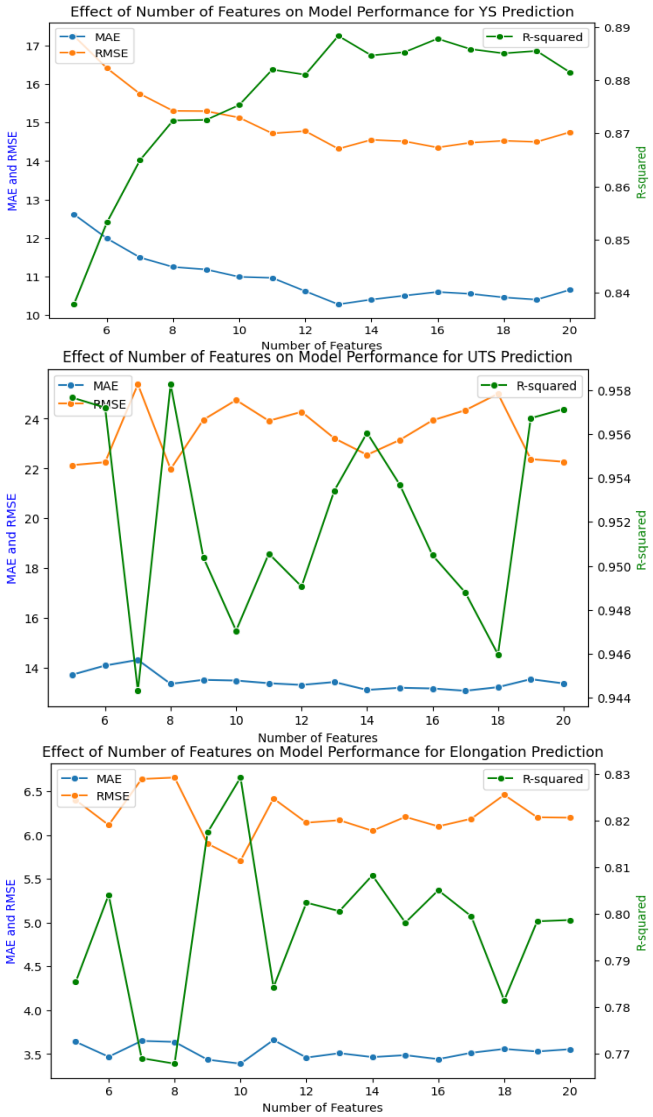

Fig. 7. Results of the evaluation metrics for the three mechanical properties of stainless steel.

Based on Fig. 7, it is evident that a higher number of features does not necessarily correlate with improved model performance. In the prediction of Yield Strength (YS), the model achieves the highest evaluation metrics with 13 features, resulting in a Mean Absolute Error (MAE) of 10.27, a Root Mean Squared Error (RMSE) of 14.31, and an R-squared value of 0.88. In contrast, the prediction of Ultimate Tensile Strength (UTS) demonstrates superior performance compared to YS prediction, with optimal results obtained using 8 features; this model yields an MAE of 13.37, an RMSE of 21.97, and an R-squared value of 0.95. Furthermore, in elongation prediction, the model exhibits the best performance with 19 features, achieving an MAE of 3.46, an RMSE of 6.04, and an R-squared value of 0.80. These findings align with research conducted by Probst and Boulesteix (2018) [37], which indicates that increasing the number of trees in a random forest ensemble can enhance prediction accuracy. This improvement is attributed to the model's ability to leverage a greater amount of information to generate more precise predictions. Another study conducted by Lin et al. (2022) [38] shows that increasing the number of trees in a random forest ensemble can improve the stability and generalization of the model, contributing to better performance in predicting various types of output. Therefore, the results of this study are consistent with previous findings that the number of estimators plays an important role in determining the performance of random forest models.

The variable names corresponding to each number of features selected using Recursive Feature Elimination (RFE) are presented in Table 2. The RFE results indicate that 13 out of 21 input variables significantly influence Yield Strength (YS). In contrast, the analysis for Ultimate Tensile Strength (UTS) yielded markedly different results, revealing that only 8 out of 21 input variables substantially affect UTS. Similarly, for Elongation (EL) predictions, the results were closely aligned with those for YS, with RFE identifying 14 input variables that notably impact EL. These discrepancies may arise from model complexity; the performance of a model can be compromised if it is either overly simplistic or excessively complex, necessitating the identification of an appropriate balance. The optimal number of features is contingent upon the inherent complexity of the relationship between input variables and the output [39]. This observation aligns with the findings of prior research, wherein RFE operates iteratively to eliminate features deemed less significant from the dataset. Features of lesser importance are systematically removed at each iteration, culminating in an optimal subset of features, which is subsequently evaluated for performance. The results demonstrate that the RFE model successfully selected 40 features from an initial pool of 754, achieving a feature reduction rate of 94.69% and an accuracy rate of 93.88% [40].

The subsequent phase of this modeling process involves the application of Information Gain (IG) to identify the input variables that most significantly influence the three mechanical properties of stainless steel. Information gain quantifies the extent to which input variables, such as chemical elements and heat treatments, contribute valuable information for improving predictions of these mechanical properties. IG employs the concept of entropy to assess how effectively a feature can mitigate uncertainty in predicting a target variable. Entropy serves as a measure of uncertainty or randomness within a dataset; thus, lower entropy reflects more ordered or structured data. In this study, chemical element and heat treatment variables with an information gain score below 0.1 were excluded from model training. The outcomes of feature selection utilizing IG are depicted in Fig. 8.

Table 2. Name and number of variables using RFE

Mechanical properties	Num features	Selected features
YS	13	Cr, Ni, Mo, Mn, Si, Nb, Ti, P, S, Al, ST temperature (K), ST time(s), and Temperature (K)
UTS	8	Cr, Ni, Mo, Ti, Cu, S, Al, and Temperature (K)
EL	14	Cr, Ni, Mo, Mn, Si, Ti, Cu, B, P, S, Al, ST temperature (K), ST time(s), and Temperature (K)

Based on Fig. 8, the results indicate that each mechanical property of stainless steel yields different outcomes. Five variables are classified as having minimal impact on Yield Strength (YS), specifically manganese (Mn), silicon (Si), vanadium (V), sulfur (S), and cobalt (Co). In contrast, for Ultimate Tensile Strength (UTS), three variables are deemed to have limited influence:

vanadium (V), carbon (C), and sulfur (S). Regarding elongation, two variables, vanadium (V) and carbon (C), are similarly identified as having negligible influence. The results of feature selection using Information Gain (IG) demonstrate that heat treatment variables—such as temperature, heating time, and cooling method—significantly affect the three mechanical properties of stainless steel. This finding markedly contrasts with the previous results obtained from feature selection using Recursive Feature Elimination (RFE), which identified eight elements that exerted the most influence on UTS, while recognizing only one heat treatment variable, namely temperature, as influential.

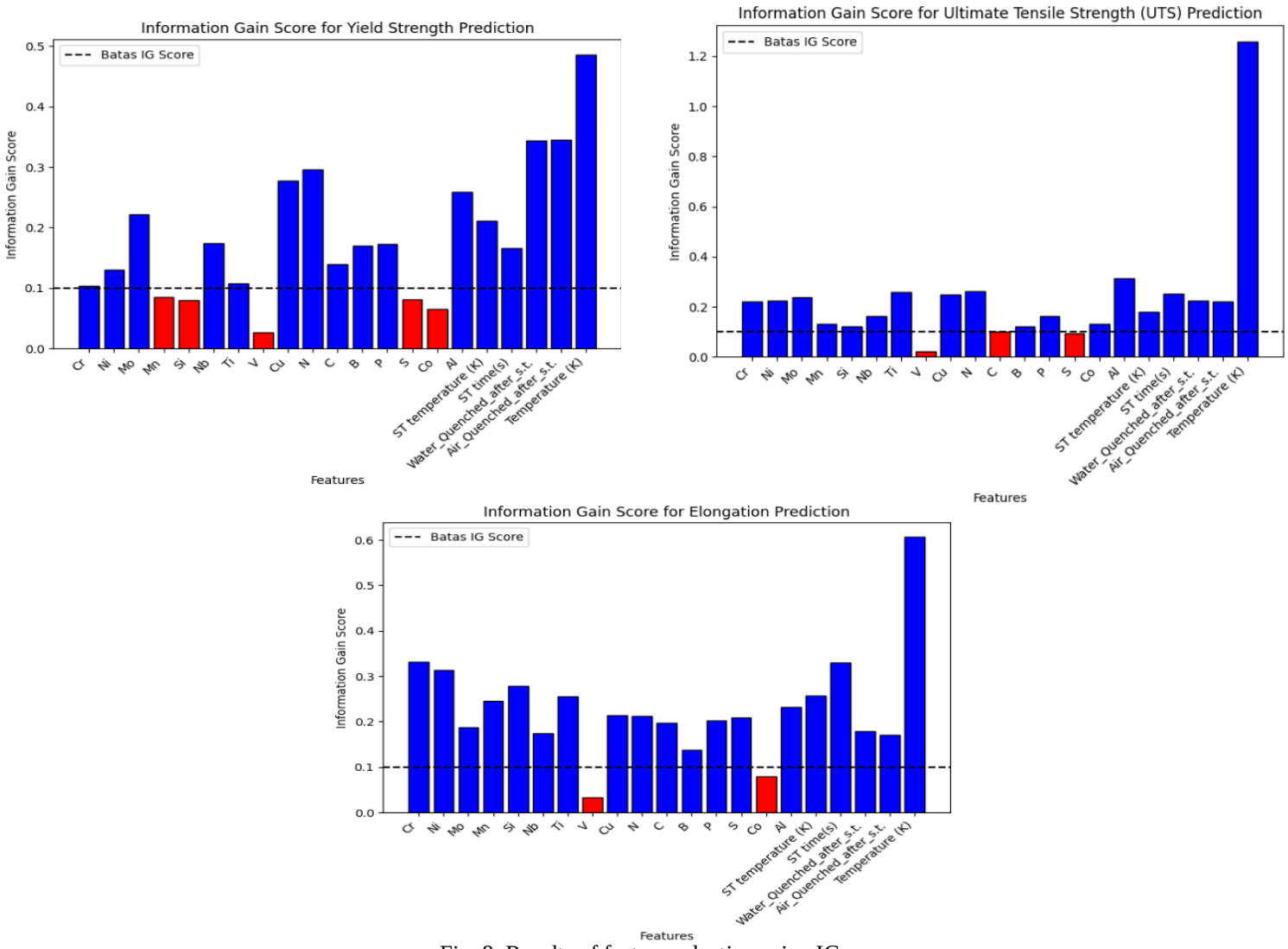

Fig. 8. Results of feature selection using IG.

The comparison results for these three treatments are available in Table 3. This comparison shows that feature selection using RFE produces better performance compared to IG and without feature selection. If analyzed in more detail, it can be seen that the combination of the random forest algorithm with RFE shows the best performance on the three mechanical properties of stainless steel. The most significant difference is seen in the R-squared value for elongation, where RFE is able to reach an R-squared value of 0.8. This figure is much higher compared to feature selection using IG, which only achieved an R-squared value of 0.78. On the other hand, without feature selection, the model gives the worst results, with an R-squared value of 0.76. This indicates that RFE helps improve the interpretability of the model by selecting the most relevant subset of features, thereby facilitating the understanding of the factors that contribute to the prediction. By reducing feature dimensions and preventing overfitting, models developed with Random Forest and RFE tend to have better performance [15].

Table 3. Comparison of model evaluation metrics

No feature selection				RFE				IG			
Mechanical properties	MAE	RMSE	R^2	Mechanical properties	MAE	RMSE	R^2	Mechanical properties	MAE	RMSE	R^2
YS	11.31	15.1	0.88	YS	10.272	14.319	0.888	YS	10.377	14.67	0.882
UTS	14.68	25.24	0.94	UTS	13.375	21.977	0.958	UTS	13.367	22.73	0.955
EL	4.73	7.17	0.76	EL	3.464	6.048	0.808	EL	3.59	6.346	0.788

3.4 Model Evaluation

Model evaluation is a critical process in model development that aims to assess the performance and accuracy of the model in predicting the mechanical properties of stainless steel. In this evaluation, various metrics are analyzed to determine the model's effectiveness in achieving its specific objectives. External testing is a key component of model evaluation that involves using a dataset that is distinct from the training dataset to assess the model's performance on previously unseen data [42]. By employing external testing, the reliability and generalizability of the model can be more rigorously evaluated, as it is assessed in scenarios that more closely resemble actual usage. Consequently,

evaluating the model through external testing helps ensure its capability to accurately predict the mechanical properties of stainless steel when presented with new, unseen data.

In this study, new data that is not included in the training dataset of 50 samples is utilized. The model being evaluated in this stage is the RFE model, which is optimized with the best parameters: 13 features with 65 estimates for Yield Strength (YS), 8 features with 67 estimates for Ultimate Tensile Strength (UTS), and 14 features with 44 estimates for Elongation (EL). Thus, the new data is aligned with the number of variables specified by the model. The results of the model testing using this new data are depicted in Fig. 9.

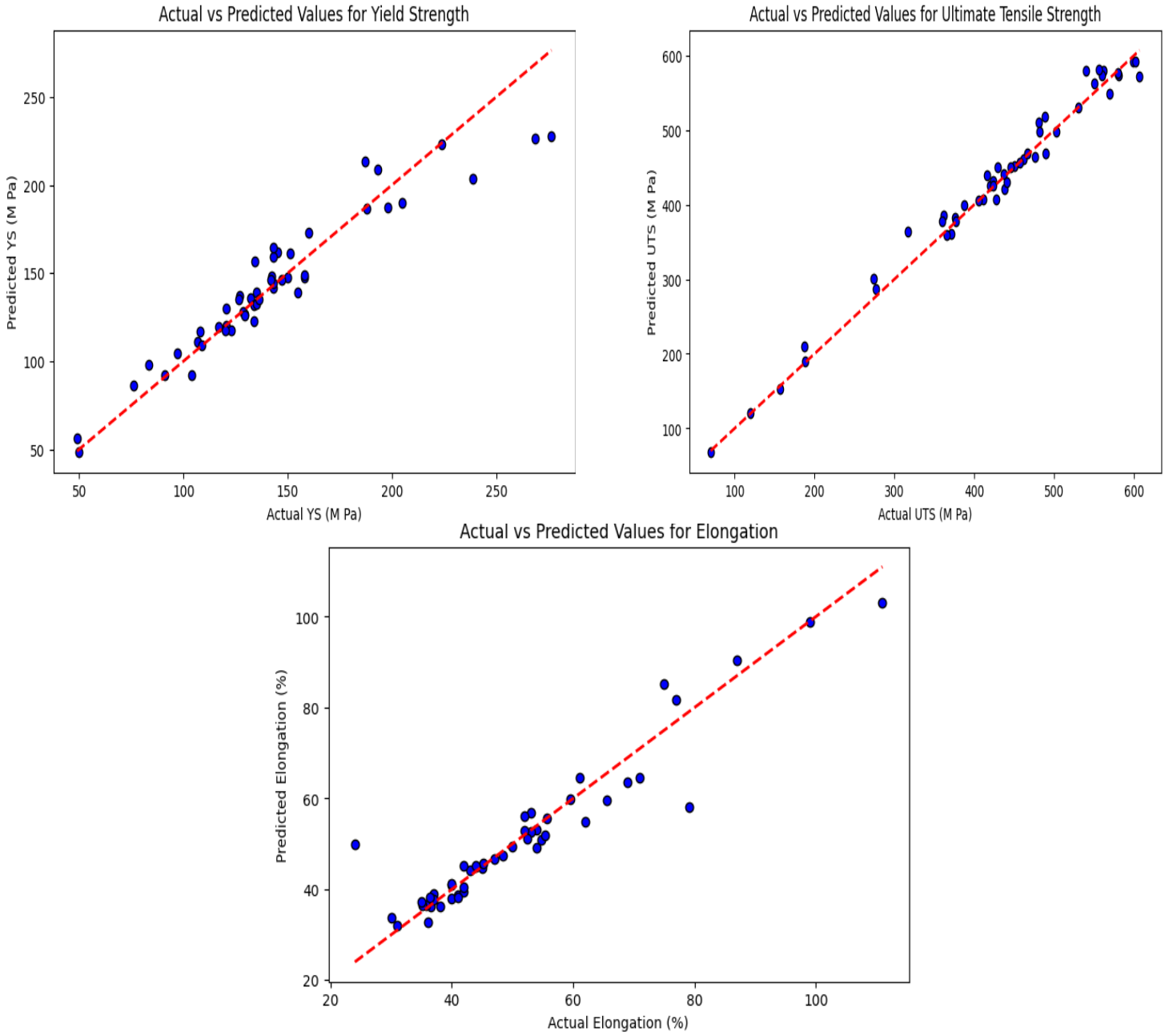

Fig. 9. Results of model testing using new data.

Based on Fig. 9, it can be seen that the random forest model with RFE is able to predict new data very well. The results of model evaluation using new data are not much different from the results of model training. This can be seen from the results of the evaluation metrics, which are not much different; namely, in the YS prediction, the MAE value was 9.91, the RMSE 14.20, and the R-squared 89. Meanwhile, for predicting the UTS value, the model showed better performance, namely with MAE values of 12.89, RMSE 16.97, and R-squared 9.7. while for the EL model, the MAE value was 3.82, RMSE 6.1, and R-squared 0.85. These results indicate that the model is able to predict the mechanical properties of stainless steel with a wide variety of data. By using this model, steel industry players can optimize production processes and improve product quality by predicting the mechanical properties of stainless steel with a high degree of accuracy. This can reduce the time and costs required for complex physical testing. In addition, accurate predictions of material mechanical properties can help industry design more efficient and safer products by adjusting manufacturing process parameters, selecting appropriate materials, and designing optimal product structures based on the resulting predictions. In the face of diverse data variations, these models can provide consistent and reliable predictions, enabling the industry to overcome the challenges often faced in processing non-uniform data. Overall, the results of this research provide a strong foundation for the steel industry to adopt a predictive approach in their production processes, with the potential to increase efficiency, reduce costs, and improve product quality.

4 Conclusion

This study compared three combinations of random forest models for predicting the mechanical properties of stainless steel: without feature selection, with Recursive Feature Elimination (RFE), and with Information Gain (IG).

1. The comparison showed that RFE significantly outperformed both IG and no feature selection. The RFE model was able to select the most influential input variables for each stainless steel mechanical property.
2. For Yield Strength (YS), RFE identified 13 of the 21 most influential variables, with an evaluation metric value of Mean Absolute Error (MAE) of 9.91, Root Mean Square Error (RMSE) of 14.20, and R-squared of 89.
3. For Ultimate Tensile Strength (UTS), the model identified 8 of the 21 most influential input variables, with an MAE value of 12.89, RMSE of 16.97, and R-squared of 9.7.
4. For Elongation (EL), this model identified 14 of the 21 most influential variables, with an MAE evaluation metric value of 3.82, RMSE 6.1, and R-squared 0.85.
5. Parameter optimization revealed the best performance with 65 estimators for YS, 67 for UTS and 44 for EL, emphasizing the importance of optimal parameter selection.
6. The results indicated that the use of the random forest model with RFE as a feature selection technique and optimal parameter selection can help the steel industry increase the accuracy of predicting the mechanical properties of materials, which in turn can support the development of more efficient and high-quality products. Thus, this research provides valuable guidance for industry in implementing a predictive approach to stainless steel development and production.

References

[1]. [1] Momeni, M. M., & Motalebian, M. (2021). Chromium-doped titanium oxide nanotubes grown via one-step anodization for efficient photocathodic protection of stainless steel. Surface and Coatings Technology, 420, 127304.

[2]. Rabi, MM (2020). Analysis and design of stainless steel reinforced concrete structural elements (Doctoral dissertation, Brunel University London).

[3]. Tabrizikahou, A., Kuczma, M., Łasecka-Plura, M., Farsangi, E.N., Noori, M., Gardoni, P., & Li, S. (2022). Application and modeling of Shape-Memory Alloys for structural vibration control: State-of-the-art review. Construction and Building Materials, 342, 127975.

[4]. Kumar, P., Jayaraj, R., Suryawanshi, J., Satwik, U.R., McKinnell, J., & Ramamurty, U. (2020). Fatigue strength of additively manufactured 316L austenitic stainless steel. Acta Materialia, 199, 225-239.

[5]. Morini, A. A., Ribeiro, M. J., & Hotza, D. (2019). Early-stage materials selection based on embodied energy and carbon footprint. Materials & Design, 178, 107861.

[6]. Wang, C., Zhu, P., Lu, Y. H., & Shoji, T. (2022). Effect of heat treatment temperature on microstructure and tensile properties of austenitic stainless 316L using wire and arc additive manufacturing. Materials Science and Engineering: A, 832, 142446.

[7]. Pan, M., Zhang, X., Chen, P., Su, X., & Misra, RDK (2020). The effect of chemical composition and annealing conditions on the microstructure and tensile properties of a resource-saving duplex stainless steel. Materials Science and Engineering: A, 788, 139540.

[8]. Leni, D. (2023). Prediction Modeling of Low Alloy Steel Based on Chemical Composition and Heat Treatment Using Artificial Neural Network. Polymachinery Journal, 21(5), 54-61.

[9]. Leni, D., Karudin, A., Abbas, M. R., Sharma, J. K., & Adriansyah, A. (2024). Optimizing stainless steel tensile strength analysis: through data exploration and machine learning design with Streamlit. EUREKA: Physics and Engineering, (5), 73-88.

[10]. Laleh, M., Sadeghi, E., Revilla, R.I., Chao, Q., Haghdadi, N., Hughes, A.E., ... & Tan, M.Y. (2023). Heat treatment for metal additive manufacturing. Progress in Materials Science, 133, 101051.

[11]. Vorontsov, A., Astafurov, S., Melnikov, E., Moskvina, V., Kolubaev, E., & Astafurova, E. (2021). The microstructure, phase composition and tensile properties of austenitic stainless steel in a wire-feed electron beam melting combined with ultrasonic vibration. Materials Science and Engineering: A, 820, 141519.

[12]. Luo, M., Zhou, G.Y., Shen, H., Wang, X.T., Li, M.C., Zhang, Z.H., & Cao, G.H. (2022). Effect of Tempering Temperature on Microstructure and Sulfide Stress Cracking of 125 Ksi Grade Casing Steel. Materials, 15(7), 2589.

[13]. Packwood, D., Nguyen, LTH, Cesana, P., Zhang, G., Staykov, A., Fukumoto, Y., & Nguyen, D.H. (2022). Machine learning in materials chemistry: An invitation. Machine learning with applications, 8, 100265.

[14]. Leni, D., Sumiati, R., Angelia, N., & Nofriyanti, E. (2023). The Influence of Heatmap Correlation-based Feature Selection on Predictive Modeling of Low Alloy Steel Mechanical Properties Using Artificial Neural Network (ANN) Algorithm. Journal of Energy, Materials, and Instrumentation Technology, 4(4), 152-162.

[15]. Liu, W., & Wang, J. (2021). Recursive elimination–election algorithms for wrapper feature selection. Applied Soft Computing, 113, 107956.

[16]. Jiang, X., Jia, B., Zhang, G., Zhang, C., Wang, X., Zhang, R., ... & Ma, H. (2020). A strategy combining machine learning and multiscale calculation to predict tensile strength for pearlitic steel wires with industrial data. Scripta Materialia, 186, 272-277.

[17]. Agrawal, A., & Choudhary, A. (2018). An online tool for predicting fatigue strength of steel alloys based on ensemble data mining. International Journal of Fatigue, 113, 389-400.

[18]. Ruiz, E., Ferreño, D., Cuartas, M., López, A., Arroyo, V., & Gutiérrez-Solana, F. (2020). Machine learning algorithms for the prediction of the strength of steel rods: an example of data-driven manufacturing in steelmaking. International Journal of Computer Integrated Manufacturing, 33(9), 880-894.

[19]. Narayana, PL, Lee, SW, Park, CH, Yeom, JT, Hong, JK, Maurya, AK, & Reddy, NS (2020). Modeling high-temperature mechanical properties of austenitic stainless steels by neural networks. Computational Materials Science, 179, 109617.

[20]. The British Steelmakers Creep Committee: BSCC High Temperature Data; The Iron and Steel Institute: London, UK, 1973.

[21]. Materials Algorithms Project. Available online: https://www.phase-trans.msm.cam.ac.uk/map (accessed on 11 April 2022).

[22]. Leni, D., Kesuma, D. S., Maimuzar, Haris, & Afriyani, S. (2024). Prediction of Mechanical Properties of Austenitic Stainless Steels with the Use of Synthetic Data via Generative Adversarial Networks. Engineering Proceedings, 63(1), 4.

[23]. 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, 2014, 3: 90-108.

[24]. Deepa, B., & Ramesh, K. (2022). Epileptic seizure detection using deep learning through min max scaler normalization. Int. J. Health Sci, 6, 10981-10996.

[25]. Jeon, H., & Oh, S. (2020). Hybrid-recursive feature elimination for efficient feature selection. Applied Sciences, 10(9), 3211.

[26]. Lin, X., Yang, F., Zhou, L., Yin, P., Kong, H., Xing, W., ... & Xu, G. (2012). A support vector machine-recursive feature elimination feature selection method based on artificial

contrast variables and mutual information. Journal of chromatography B, 910, 149-155.
[27]. Pham, BT, Jaafari, A., Avand, M., Al-Ansari, N., Dinh Du, T., Yen, HPH, ... & Tuyen, TT (2020). Performance evaluation of machine learning methods for forest fire modeling and prediction. Symmetry, 12(6), 1022.
[28]. Hu, M., Tan, Q., Knibbe, R., Wang, S., Li, X., Wu, T., ... & Zhang, MX (2021). Prediction of mechanical properties of wrought aluminum alloys using feature engineering assisted machine learning approach. Metallurgical and Materials Transactions A, 52(7), 2873-2884.
[29]. Xiong, J., Zhang, T., & Shi, S. (2020). Machine learning of mechanical properties of steels. Science China Technological Sciences, 63(7), 1247-1255.
[30]. Zhang, S., Jiang, Z., Li, H., Zhang, B., Fan, S., Li, Z., ... & Zhu, H. (2018). Precipitation behavior and phase transformation mechanism of super austenitic stainless steel S32654 during isothermal aging. Materials characterization, 137, 244-255.
[31]. Moniruzzaman, F.M., Shakil, S.I., Shaha, S.K., Kacher, J., Nasiri, A., Haghshenas, M., & Hadadzadeh, A. (2023). Study of direct aging heat treatment of additively manufactured PH13–8Mo stainless steel: role of the manufacturing process, phase transformation kinetics, and microstructure evolution. Journal of Materials Research and Technology, 24, 3772-3787.
[32]. Niu, M. C., Yang, K., Luan, J. H., Wang, W., & Jiao, Z. B. (2022). Cu-assisted austenite reversion and enhanced TRIP effect in maraging stainless steels. Journal of Materials Science & Technology, 104, 52-58.
[33]. Bleck, W., Guo, X., & Ma, Y. (2017). The TRIP effect and its application in cold formable sheet steels. Steel Research International, 88(10), 1700218.
[34]. Cronemberger, MER, Mariano, NA, Coelho, MF, Pereira, JN, Ramos, É. C., de Mendonça, R., ... & Maestrelli, S.C. (2014, December). Study of cooling rate influence on SAF 2205 duplex stainless steel solution
[35]. Probst, P., & Boulesteix, A.L. (2018). To tune or not to tune the number of trees in random forest. Journal of Machine Learning Research, 18(181), 1-18.
[36]. Probst, P., Wright, M.N., & Boulesteix, A.L. (2019). Hyperparameters and tuning strategies for random forests. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, 9(3), e1301.
[37]. Probst, P., & Boulesteix, A.L. (2018). To tune or not to tune the number of trees in random forest. Journal of Machine Learning Research, 18(181), 1-18.
[38]. Lin, S., Zheng, H., Han, B., Li, Y., Han, C., & Li, W. (2022). Comparative performance of eight ensemble learning approaches for the development of models of slope stability prediction. Acta Geotechnica, 17(4), 1477-1502.
[39]. Ibrahim, M. (2022). Evolution of Random Forest from Decision Trees and Bagging: A Bias-Variance Perspective. Dhaka University Journal of Applied Science and Engineering, 7(1), 66-71.
[40]. Lamba, R., Gulati, T., & Jain, A. (2022). A hybrid feature selection approach for Parkinson's detection based on mutual information gain and recursive feature elimination. Arabian Journal for Science and Engineering, 47(8), 10263-10276.
[41]. Wang, X., Guo, B., Shen, Y., Zhou, C., & Duan, X. (2019). Input feature selection method based on feature set equivalence and mutual information gain maximization. IEEE Access, 7, 151525-151538.
[42]. Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808.