Research Article

Evaluation of Imputation Methods for Handling Missing Data in Mechanical Materials Datasets

Earl Yin

Walnut High School, 400 Pierre Rd, Walnut, CA 91789, USA

ABSTRACT

Mechanical design materials are integral to advancing modern technologies due to their diverse mechanical properties. These properties are crucial in determining the material's suitability for various engineering applications. However, research on mechanical materials often encounters missing data, which can lead to biased results and reduced statistical power. While several imputation methods exist to handle missing data, there is a lack of focused studies evaluating their performance in the context of mechanical materials. To address this gap, a comprehensive dataset was obtained, and 10% of the original data for ultimate tensile strength (Su) and yield strength (Sy) were intentionally deleted. Four imputation methods—mean imputation, random fill, regression imputation, and k-nearest neighbors (KNN) imputation—were employed to restore the missing data. The performance of these methods was evaluated using Pearson's correlation, multiple linear regression, and permutation feature importance. The results showed that mean and KNN imputation methods provided the closest match to the original data, while regression imputation also performed well with minor deviations. Random fill was the least reliable method. These findings provide guidance on selecting appropriate imputation techniques for mechanical materials datasets, ultimately improving the robustness of future research.

Keywords: Mechanical Design Materials, Missing Data, Mechanical Properties, Imputation Methods, Performance Evaluation

INTRODUCTION

Machine design materials play a critical role in advancing modern technologies across various industries. These materials, which include steel, brass, aluminum, and a wide range of alloys, are selected based on their unique mechanical properties that suit specific engineering applications. Key properties such as yield strength, elastic modulus, shear modulus, and tensile strength determine how these materials respond under stress and strain during operation. The diverse mechanical characteristics of these materials allow for their use in different applications, from high-strength steel in construction and automotive industries to lightweight alloys in aerospace engineering. Understanding the mechanical behavior of materials is essential for optimizing performance, ensuring safety, and enhancing the durability of machine components, making material selection a vital aspect of the design process.

Previous research has extensively explored the mechanical properties of materials due to their critical

Corresponding author: Earl Yin, E-mail: earlyin060807@gmail.com.

Received October 17, 2024**; Accepted** October 30, 2024
https://doi.org/10.70251/HYJR2348.241222

importance in machine design and structural applications. Studies have focused on analyzing the behavior of various materials under different stress conditions to understand how properties like yield strength, tensile strength, elastic modulus, and shear modulus affect performance and durability. For instance, research on steel alloys has shown how variations in carbon content and heat treatment processes can significantly alter mechanical properties, enhancing both strength and ductility for specific applications (1). Similarly, investigations into aluminum alloys have highlighted their lightweight characteristics combined with sufficient tensile strength, making them ideal for aerospace and automotive industries where weight reduction is crucial (2). Additionally, studies on composite materials have demonstrated the potential for superior mechanical properties, such as increased toughness and fatigue resistance, compared to traditional metals (3). These investigations underscore the importance of understanding and optimizing material properties to improve performance, safety, and efficiency in engineering designs.

A common existing in the above-mentioned studies is the requirement for high-quality, complete datasets to draw accurate and reliable conclusions. Unfortunately, many research efforts are hindered by missing data, which can occur due to experimental limitations, equipment failures, human error, or insufficient sample sizes. Missing data can significantly impact the validity of research, as incomplete datasets may lead to biased results or reduced statistical power. As a result, numerous studies have been dedicated to addressing the issue of missing data and developing effective methods to handle it. For example, Rubin's seminal work on missing data mechanisms introduced key strategies such as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), which have been foundational in guiding imputation methods (4). Additionally, modern approaches like multiple imputation, introduced by Schafer and Olsen (5), and machine learning-based methods such as k-nearest neighbors and regression imputation have been applied to restore data integrity and improve the reliability of research conclusions. These methods, aimed at filling gaps in datasets, have become essential in fields where missing data is a frequent obstacle, ensuring that research findings remain robust and accurate despite incomplete information.

Although a wide range of methods to handle missing data exist, there remains a significant gap in studies that rigorously evaluate the performance of these methods, particularly in the field of mechanical materials. To address this gap, the author obtained a comprehensive mechanical materials dataset, encompassing various mechanical properties, and intentionally deleted portions of the data to simulate missing values. These missing data were then imputed using several alternative methods. The performance of each imputation method was evaluated from multiple perspectives, including accuracy, preservation of data relationships, and impact on model performance. The primary aim of this paper is to provide a detailed assessment of these imputation techniques in the context of mechanical materials, offering insights into the most reliable methods for restoring incomplete datasets. The expected benefit of this study is to guide researchers in selecting the most appropriate imputation methods for mechanical property datasets, ultimately improving the reliability and robustness of research in this field.

Data Description

The dataset used for this research is available on Kaggle, thanks to the efforts of Nawale (6). The original dataset includes detailed information on various material-related properties, as outlined in Table 1.

Table 1. The Variables Included in the Material Dataset

No.	Variables	Definitions
1	ID	Unique Identification code for the Material
2	Name	Material Name (e.g., Steel SAE 1015)
3	Su	Ultimate Tensile Strength in MPa
4	Sy	Yield Strength in MPa
5	E	Elastic Modulus in MPa
6	G	Shear Modulus in MPa
7	mu	Poisson's Ratio in Units of Length
8	Ro	Density in Kg/m^3
9	A5	Elongation at Break or Strain as a Percentage
10	Heat_Treat	Heat Treatment Method
11	BHN	Brinell Hardness Number in Microhardness Units
12	pH	Pressure at Yield in MPa
13	Desc	Description of the Material
14	HV	Vickers Hardness Number
15	Std	The applicable consensus standards for products (e.g., The American National Standards Institute – ANSI)

In Table 1, the first eight variables have complete observations for all 1,552 entries (ID, Name, Su, Sy, E, G, μ, and ρ). However, given the objectives of this study, the variables ID and Name were considered less relevant for analysis and thus excluded. Consequently, six key variables—Su, Sy, E, G, μ, and ρ—were retained for evaluation. To introduce missing data and simulate real-world conditions, the Su and Sy variables were randomly reduced by 10%. These variables were selected due to their relatively even distribution and narrower range compared to the others, providing a more robust platform for evaluating different methods of handling incomplete data. Additionally, eight observations were removed due to the lack of specific values for Sy, where only the maximum value was specified without precise measurements. Ultimately, the final dataset consists of 1,544 observations. Detailed information for both the complete and partial datasets used in this study is presented in Table 2.

Table 2 presents a comparative analysis of the complete and partial datasets used in the study. The complete dataset includes 1,552 observations, while the partial dataset reflects the same variables after the removal of 10% of data for the Su and Sy variables and the exclusion of eight observations with incomplete Sy values, resulting in 1,544 observations. For both datasets, the key variables Su, Sy, E, G, μ, and ρ are presented with their respective means, standard deviations, and range (maximum and minimum values). The comparison shows minimal differences between the complete and partial datasets, with slight variations in the means and standard deviations of Su and Sy due to the introduced missing data. Notably, the other variables (E, G, μ, and ρ) remain unchanged, as they were fully observed in both datasets. This table serves to illustrate the impact of the missing data on the statistical properties of the variables selected for analysis.

METHODS

In this research, two types of methods are employed. The first type focuses on handling missing data, while the second type is used to compare the statistical performance differences between the complete and partial datasets. Although numerous approaches exist for addressing missing data, four methods were deliberately selected for this study: mean imputation, K-Nearest Neighbors

Table 2. Descriptive Statistics of Both Complete and Partial Datasets

Complete Dataset				
Variable	**Mean**	**Standard Deviation**	**Max.**	**Min.**
Su	574.32	326.95	2220.00	69.00
Sy	387.76	290.04	2048.00	28.00
E	164356.87	56201.22	219000.00	73000.00
G	85627.85	125650.62	769000.00	26000.00
mu	0.30	0.02	0.35	0.20
Ro	6925.02	2119.58	8930.00	1750.00
Partial Dataset				
Variable	**Mean**	**Standard Deviation**	**Max.**	**Min.**
Su	569.47	325.33	2220.00	69.00
Sy	391.79	293.59	2048.00	28.00
E	164356.87	56201.22	219000.00	73000.00
G	85627.85	125650.62	769000.00	26000.00
mu	0.30	0.02	0.35	0.20
Ro	6925.02	2119.58	8930.00	1750.00

Notes: 1. Max.=Maximum value; Min.=Minimum value. 2. See Table 1 for the definition of Variables.

(KNN) imputation, random imputation, and regression imputation. The first method, mean imputation, represents a central tendency approach, filling in missing values with the mean of the available data. KNN imputation, a machine learning method, uses local data patterns to predict and fill in missing values based on the nearest neighbors. Regression imputation applies simple statistical regression models to estimate the missing values. Lastly, random imputation fill the missing data by using some of the randomly selected data, which tends to serve as the base imputation one. The details of both types of methods are discussed in the following subsections.

Missing Data-Handling Methods

Mean Imputation. Mean imputation is a widely used method for handling missing data, where the missing values are replaced by the arithmetic mean of the observed values for a given variable. This approach assumes that the missing data is missing at random and that the central tendency of the data can adequately represent the missing values. The imputed value, $\hat{x}$ for a missing data point is calculated using the equation (7):

$$\hat{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_j \quad 1)$$

where x_j represents the observed data points, and n is the total number of observed values for that variable. Mean imputation is simple and computationally efficient, making it useful in scenarios where quick estimations are required. However, it can underestimate the variability in the dataset and potentially lead to biased statistical estimates, as it does not account for the inherent uncertainty associated with missing values. Despite these limitations, mean imputation remains a common technique due to its ease of implementation.

KNN Imputation. K-Nearest Neighbors (KNN) imputation is a machine learning-based method for handling missing data, leveraging the structure and proximity of the data to estimate missing values. In this method, the missing value for a particular data point is imputed by examining the k-nearest observations (neighbors) that are most similar to the incomplete data point, based on a predefined distance metric such as Euclidean distance. For a given missing value , the imputed value is calculated by taking the weighted average of the nearest neighbors, as expressed by the equation (8):

$$\hat{x}_i = \frac{1}{k}\sum_{j \in N_i} x_j \quad 2)$$

where N_i represents the set of k-nearest neighbors for the i-th data point, and x_j is the observed value of the j-th neighbor. The algorithm assumes that similar data points share similar values, thus making the nearest neighbors a good reference for estimating missing data. One of the key advantages of KNN imputation is that it preserves the inherent relationships and patterns within the dataset. However, it can be computationally expensive, especially for large datasets, and its accuracy depends heavily on the choice of k and the distance metric used.

Regression Imputation. Regression imputation is a statistical method that predicts and fills in missing data by using the relationships between the variables in the dataset. In this method, a regression model is built with the observed data to predict the missing values based on the available predictors. In this study, missing values for Su (ultimate strength) and Sy (yield strength) are imputed using the predictors E (modulus of elasticity), G (shear modulus), μ (Poisson's ratio), and ρ (density). The regression imputation can be formulated as (9):

$$\hat{S}_u, \hat{S}_y = \beta_0 + \beta_1 E + \beta_2 G + \beta_3 \mu + \beta_4 \rho + \varepsilon \quad 3)$$

where β_0 is the intercept, β_1, β_2, β_3, β_4 are the regression coefficients, and ϵ represents the error term. This method provides estimates for the missing values by leveraging the linear relationship between the dependent variables S_u and S_y and the predictor variables E, G, μ, and ρ. Regression imputation is advantageous as it accounts for the interdependencies between variables, producing more accurate estimates compared to simpler methods like mean imputation. However, its accuracy depends on the strength of the underlying relationships between the variables.

Random Imputation. Aside from the above three methods, random imputation is an alternative one used to fill in missing data by randomly selecting observed values from the complete dataset to replace the missing entries. This approach assumes that the missing data is missing completely at random, meaning that the probability of data being missing is independent of both observed and unobserved data. The general equation for random imputation can be expressed as (10):

$$X_{imputed} = X_{observed} \text{ for a randomly selected value from } X_{observed} \quad 4)$$

Where $X_{imputed}$ represents the imputed value and $X_{observed}$ is the set of observed values from which a random value is selected. This method does not rely on relationships between variables, and as a result, it can introduce more

variability and potential bias, particularly if the data is not missing completely at random. Although random imputation maintains the distribution of observed data, it can weaken relationships between variables, as the imputed values are not based on the actual structure of the data. This can lead to lower predictive accuracy and weaker performance in model evaluations, as reflected by the lower correlation with the original dataset in the current study. Thus, random imputation is often considered less reliable and serve as the base imputation method in the present research.

Performance Evaluation Method

Once the missing data are created by using various missing data-handling methods, their performances are then needed to be assessed from different perspectives. For more reliable results, the assessment methods include three alternative ones which are correlation analysis of individual variables, multiple linear regression, and feature importance ranking. The details are presented in order as follows.

Correlation Analysis. To assess how closely the simulated Su and Sy data match the actual Su and Sy data, Pearson's correlation analysis is conducted. This method is selected due to the numerical nature of both datasets and its ability to quantify linear relationships between them. Pearson's correlation coefficient (denoted as *r*) measures the strength and direction of a linear relationship between two variables, with values ranging from -1 to 1. A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 suggests no linear correlation.

The formula for Pearson's correlation coefficient is (11):

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} \quad 5)$$

Where:

- X and Y are simulated and actual values for Su and Sy, respectively.
- $\bar{X}$ and $\bar{Y}$ are the mean values of X and Y.

Before applying Pearson's correlation, the assumptions of the test are considered: (1) both variables should be continuous and normally distributed, (2) the relationship between the variables should be linear, and (3) the absence of outliers is assumed as they can disproportionately affect the correlation coefficient. The correlation between the actual and simulated Su and Sy data is calculated for each missing data-handling method. To facilitate interpretation, the results are presented as a correlation coefficient matrix plot, enabling a clear visual comparison of the correlations for different methods.

Multiple Linear Regression Analysis. In addition to the correlation analysis between simulated and actual Su and Sy, linear regression models are developed for both the simulated and actual Su and Sy data. The models utilize common predictors, including E (modulus of elasticity), G (shear modulus), μ (Poisson's ratio), and ρ (density). The primary objective of this modeling is to evaluate how closely the simulated data approximates the actual data in terms of predictive performance.

Linear regression is a statistical method used to model the relationship between a dependent variable (Su or Sy) and one or more independent variables (E, G, μ, and ρ). The general form of a simple linear regression model is (12):

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon \quad 6)$$

Where:

- Y is the dependent variable (Su or Sy),
- β_0 is the intercept,
- $\beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients associated with each predictor variable $X_1, X_2, \ldots, X_n$,
- ε is the error term, representing the deviation of the actual values from the predicted ones.

To ensure the validity of the linear regression models, several key assumptions are checked which include linearity, independence, homoscedasticity and normality.

Permutation Feature Importance Analysis via Neural Network. To rank the importance of features for predicting Su and Sy, a simple feedforward neural network is employed using the Keras framework. The feature importance is computed using permutation feature importance, a model-agnostic method that evaluates how much the neural network's performance deteriorates when the values of a specific feature are randomly shuffled. This approach allows for the assessment of each feature's contribution to the model's predictions, as the greater the decrease in model performance after shuffling a feature, the more important that feature is deemed to be.

The neural network consists of an input layer, one or more hidden layers with nonlinear activation functions (such as ReLU), and an output layer. The network is trained to minimize the loss function (e.g., mean squared error, MSE) between the predicted and actual values of Su and Sy. The general structure of the model can be represented as (13):

$$\hat{y} = f(W * X + b) \quad 7)$$

Where:

- $\hat{y}$ is the predicted output (Su or Sy),
- X is the input matrix containing the features (e.g., E,G, μ, and ρ),
- W is the weight matrix learned during training,
- b is the bias term,
- f is the activation function applied at each layer.

Permutation feature importance is calculated by first measuring the baseline performance of the trained model using a suitable evaluation metric, such as MSE. For each feature Xi , the values are randomly shuffled, and the model's performance is re-evaluated. The feature importance score $I(X_i)$ for feature X_i is computed as:

$$I(X_i) = MSE_{shuffuled} - MSE_{baseline} \quad 8)$$

Where:

- $MSE_{shuffled}$ is the model's performance (MSE) after shuffling feature X*i*,
- $MSE_{baseline}$ is the model's baseline performance before shuffling any features.

A larger difference between $MSE_{shuffled}$ and $MSE_{baseline}$ indicates that the model's performance is highly dependent on that feature, signifying greater importance. This method assumes that the model is well-trained and the features are not highly correlated (multicollinear), which could otherwise affect the accuracy of the feature importance rankings. Using this approach, the features E, G, μ, and ρ can be ranked in terms of their relative importance for predicting Su and Sy, providing valuable insights into which variables have the strongest influence on the neural network's predictions. This information is essential for understanding the underlying relationships in the data and optimizing the model for better performance.

RESULTS

Further Comparison between Complete and Partial Datasets

Based on the previous discussion, 10% of the original Su and Sy data were removed. Before evaluating the performance of the simulated data using different imputation methods, the originally complete and partial datasets for Su and Sy are first compared visually using histograms, as shown in Figures 1 and 2.

Figure 1 presents histogram plots for Su from both the complete and partial datasets. In the complete dataset (left plot), the distribution is fairly smooth, with a peak around lower Su values (between 250 and 500) and a gradual tapering as Su increases. In contrast, the partial dataset (right plot) shows a more clustered distribution, with a sharper peak at the lower Su values and a more pronounced decline at higher values. This suggests that while the overall shape of the distribution is preserved, the missing data may have resulted in a slight compression of variability in the Su values, particularly for larger values.

Figure 2 presents histogram plots for Sy from both the complete and partial datasets. The complete dataset (left plot) shows a similar pattern to Su, with a peak around lower Sy values (between 250 and 500) and a smooth tapering as Sy increases. The partial dataset (right plot),

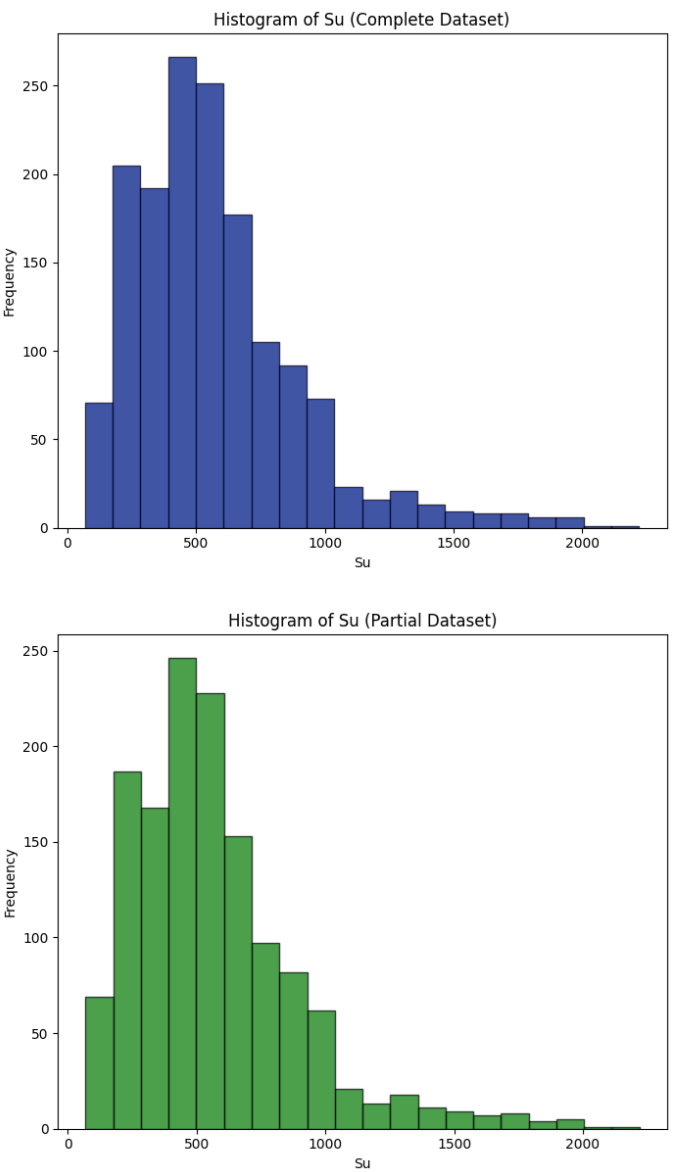

Figure 1. Histogram Plots of Su for both Complete and Partial Datasets.

however, shows a sharper peak and a more abrupt drop-off at higher Sy values. This pattern indicates that, while the partial dataset generally captures the central tendency of Sy, the tails of the distribution—particularly at higher values—may not be as well represented in the complete dataset.

Overall, in both figures, the partial dataset exhibits a more concentrated distribution around the lower values, with less variability in the higher range compared to the complete dataset. These phenomena suggest that while the partial data preserves well the central tendencies of Su and Sy, some deviations exist in the representation of higher values, where the incomplete may exhibit less variability.

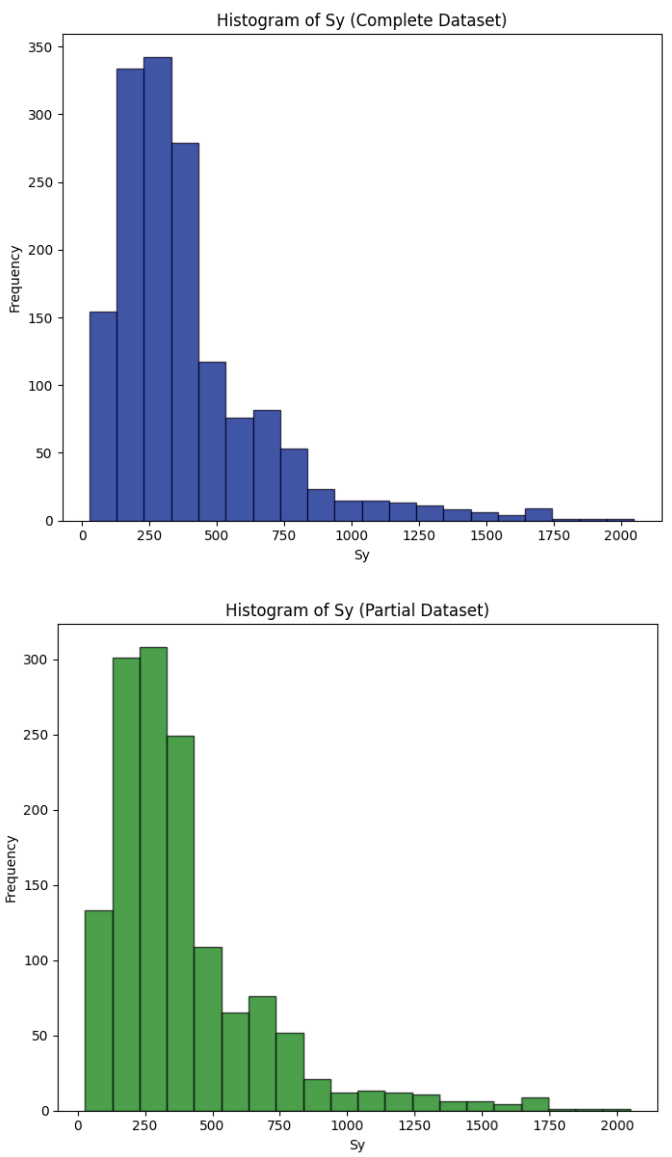

Figure 2. Histogram Plots of Sy for both Complete and Partial Datasets.

Correlation Analysis Results of various Imputation Methods

Pearson's correlation analysis is the first method selected to evaluate the performance of the three imputation methods used to fill the 10% missing data for Su and Sy. Pearson's correlation coefficient, which measures the strength of the linear relationship between two datasets, allows for a quantitative comparison of the imputed data against the original. To facilitate easy visual comparison, a correlation matrix plot was generated based on the correlations between the original data and each of the four imputation methods (mean, random, regression and k-nearest neighbors) for both Su and Sy. The results are displayed in Figures 3 and 4.

Figure 3, which shows the correlation matrix for Su across the different datasets, demonstrates that the k-nearest neighbors (KNN) imputation method achieved the highest correlation with the original Su data, with a correlation coefficient of 0.99. This indicates that the KNN method closely mirrors the original dataset. The regression imputation method also performed well, achieving a correlation of 0.96 with the original Su data. On the other hand, the random fill method exhibited a

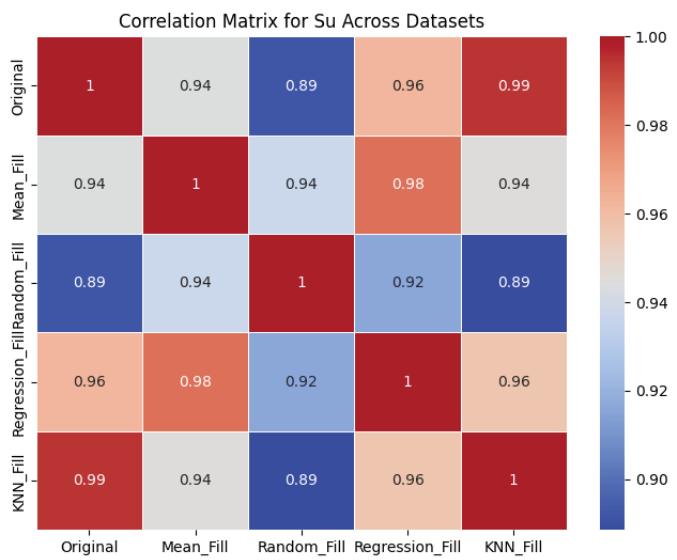

Figure 3. Correlation Matrix Plots of Su for Original and Other Datasets.

lower correlation of 0.89, indicating that it had the least similarity to the original data. The mean imputation method scored a correlation of 0.94, placing it between the random and regression methods in terms of performance.

Figure 4 presents the correlation matrix for Sy across the same datasets. Similar to Su, the KNN method achieved a perfect correlation with the original Sy data, registering a coefficient of 1.0, indicating that it most accurately replicated the original distribution. The regression method followed closely with a correlation of 0.99, demonstrating its ability to effectively fill the missing data for Sy. The mean imputation method performed reasonably well with a correlation of 0.96, while the random fill method again showed the lowest correlation at 0.91, signifying more noticeable deviations from the original dataset.

Overall, the correlation matrix plots highlight that the KNN imputation method consistently provided the closest match to the original Su and Sy data across both figures, followed by the regression imputation method. The mean imputation method also performed satisfactorily, though it showed slightly lower correlations. The random fill method, however, displayed the weakest performance in replicating the original data, as indicated by the lower correlation coefficients. These findings suggest that the choice of imputation method has a significant impact on the ability to preserve the original data's relationships, with KNN and regression imputation methods being the most reliable in this case.

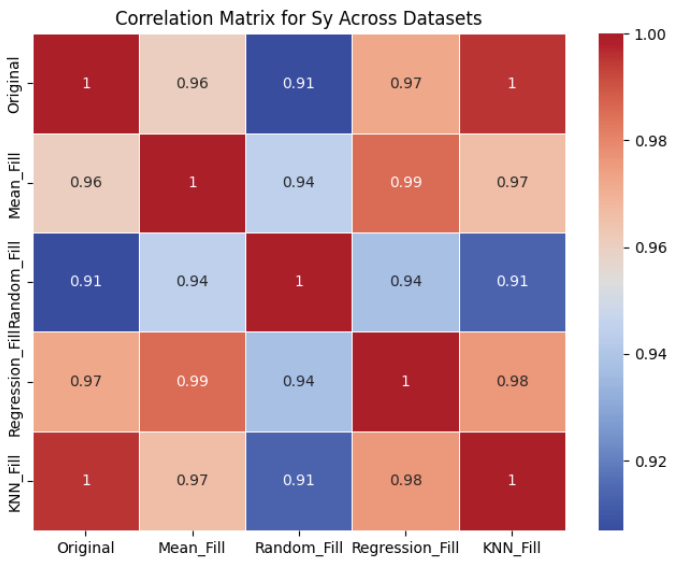

Figure 4. Correlation Matrix Plots of Sy for Original and Other Datasets.

Regression Analysis Results of various Imputation Methods

The impact of different imputation methods on model performance was further evaluated using a multiple linear regression model. The modeling results for both the original and simulated data for Su and Sy, under different imputation methods (mean, random, regression, and k-nearest neighbors), are presented in Table 3. This table includes the coefficients and p-values for each variable (E, G, μ, and ρ) across the various datasets, allowing for a detailed comparison of how closely the simulated data replicates the relationships observed in the original dataset.

For Su, the regression results indicate that, across all imputation methods, the variable μ (Poisson's ratio) consistently contributes the most to the model, with coefficients ranging from 361.7 to 538.8, depending on the imputation method used. The original dataset's coefficient for μ is 377.2, and the mean imputation method yields the closest match, with a coefficient of 376.5, followed by the regression imputation method with a coefficient of 383.3. On the other hand, the random fill method produces the largest deviation, with a μ coefficient of 538.8, indicating that this method is the least reliable in terms of reproducing the original data's structure. The p-values for μ across all methods remain relatively high, suggesting limited statistical significance for the prediction of Su. Variables E, G, and ρ consistently have coefficients and p-values of 0.0, indicating that they do not significantly contribute to the prediction of Su, regardless of the imputation method used.

For Sy, similar trends are observed. The variable μ again shows the most significant contribution to the model, with coefficients ranging from 148.4 (original dataset) to 259.3 (random fill). As with Su, the mean imputation method yields the closest match to the original dataset, with a coefficient of 192.6, followed by the regression imputation method at 230.4. The random fill method results in the largest discrepancy, with a coefficient of 259.3, while the k-nearest neighbors (KNN) imputation method produces a coefficient of 207.5, falling between the mean and regression methods. However, the p-values for μ remain relatively high across all methods, indicating that μ is not statistically significant for predicting Sy under these models. Similar to Su, the other variables—E, G, and ρ—show no significant influence, with coefficients and p-values of 0.0 across all imputation methods.

In end, the mean imputation method provides the closest match to the original dataset for both Su and Sy, followed by the regression imputation method. The random

Table 3. Linear Regression Model Results for Original and Simulated Datasets

Linear Model Results for Su										
	Original		**Mean_Fill**		**Random_Fill**		**Regress._Fill**		**KNN_Fill**	
Var.	**Coef.**	**P**	**Coef.**	**P**	**Coef.**	**P**	**Coef.**	**P**	**Coef.**	**P**
const	-107.0	0.3	-62.8	0.5	-116.0	0.3	-110.1	0.3	-90.3	0.4
E	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
G	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
mu	377.2	0.2	376.5	0.2	538.8	0.1	383.3	0.2	361.7	0.2
Ro	0.0	0.2	0.0	0.3	0.0	0.6	0.0	0.2	0.0	0.1
Linear Model Results for Sy										
	Original		**Mean_Fill**		**Random_Fill**		**Regress._Fill**		**KNN_Fill**	
Var.	**Coef.**	**P**	**Coef.**	**P**	**Coef.**	**P**	**Coef.**	**P**	**Coef.**	**P**
const	-17.3	0.9	7.8	0.9	-33.5	0.8	-44.6	0.7	-37.9	0.7
E	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
G	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
mu	148.4	0.6	192.6	0.5	259.3	0.4	230.4	0.4	207.5	0.5
Ro	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Notes: 1. Var.=Variable; Coef.=Coefficient; P=P-value. 2. See Table 1 for the definition of Variables. 3.constant represents the model intercept.

fill method, in contrast, consistently shows the largest deviations from the original data, making it the least reliable method. These results highlight the importance of selecting appropriate imputation techniques to ensure the accuracy of regression models when handling missing data.

Permutation Feature Importance Analysis Results of various Imputation Methods

To compare the performance of the simulated datasets (Mean Fill, Random Fill, Regression Fill, and KNN Fill) with the original dataset for Su and Sy, a simple feedforward neural network model using Keras was trained. The model consisted of two hidden layers: the first with 64 neurons and the second with 32 neurons. After training the model on both the original and simulated datasets, permutation feature importance was applied to evaluate how much the model's performance decreased when each feature was randomly shuffled. This provides insights into the importance of each feature in predicting Su and Sy.

In both Su and Sy feature importance rankings as shown in Figures 5 and 6, the mean imputation method and KNN Fill closely mirror the original dataset's feature importance structure, indicating that these methods effectively maintain the relationships between the input features and the target variables. Regression Fill also performs well but exhibits some minor shifts in importance. In contrast, the random fill method introduces significant distortions, with feature importance rankings deviating substantially from the original dataset. Therefore, when considering feature importance preservation, mean imputation and KNN Fill are more reliable methods compared to random fill, with regression fill providing a middle ground.

CONCLUSIONS

This study evaluated the performance of various imputation methods—mean imputation, random fill, regression imputation, and k-nearest neighbors (KNN) imputation—for handling missing data in mechanical materials datasets. After deleting 10% of the original data, these methods were assessed based on Pearson's correlation, multiple linear regression, and permutation

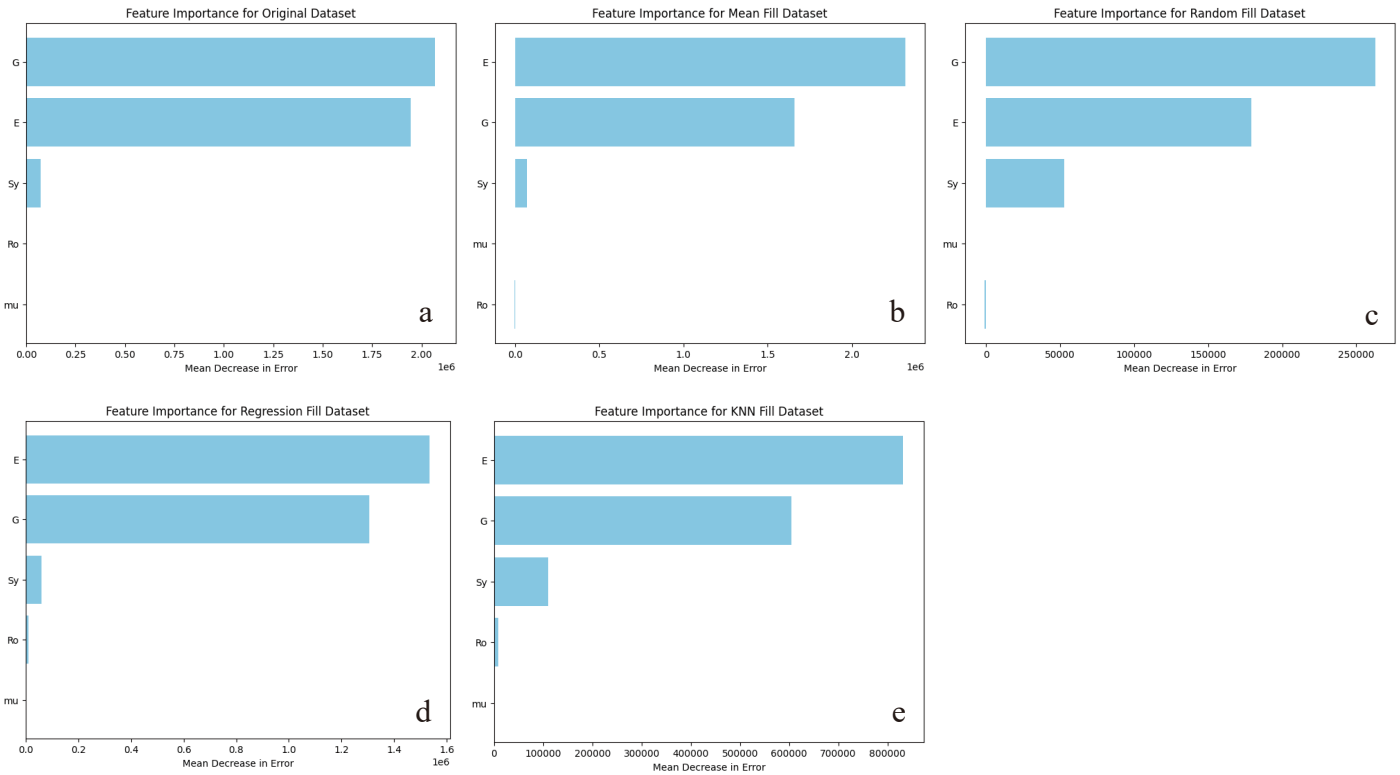

Figure 5. Feature Importance Ranking Plots of Su for Original and Simulated Datasets.

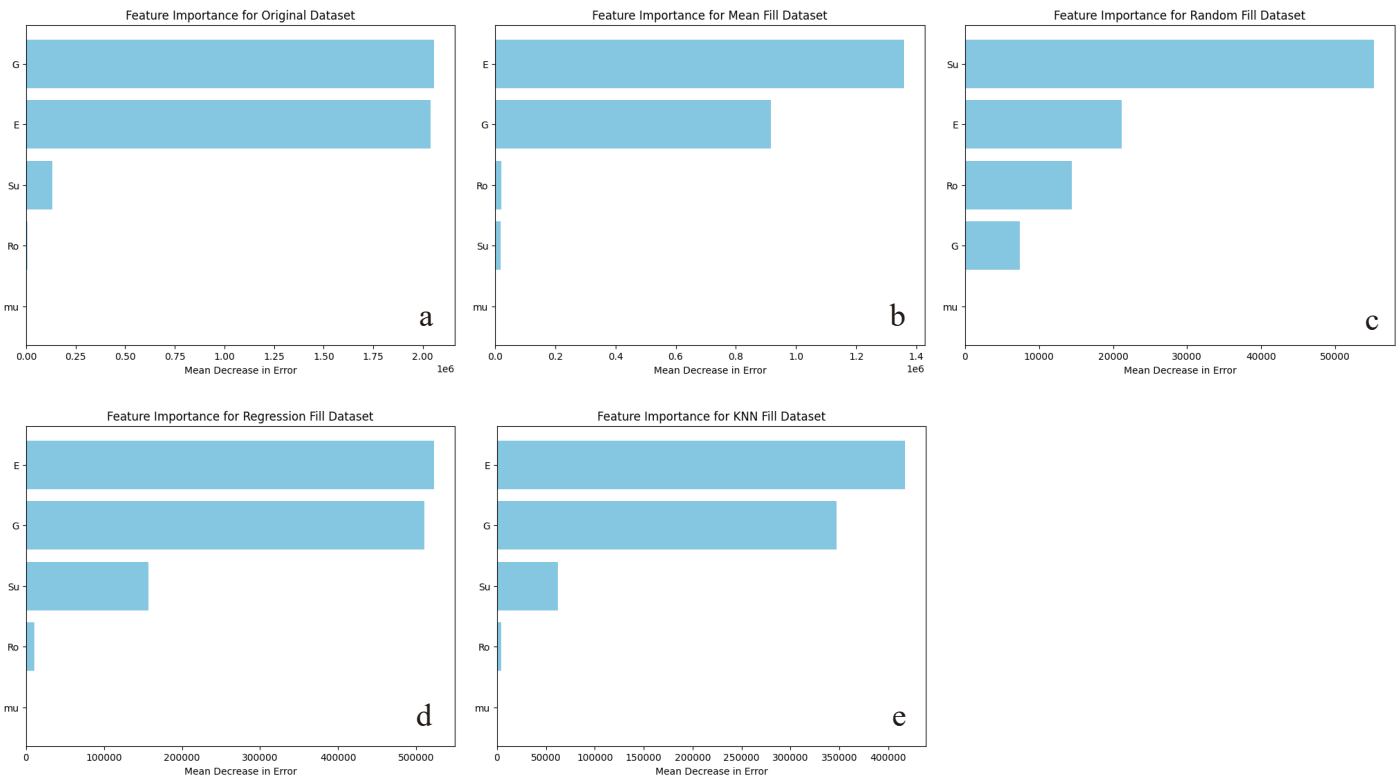

Figure 6. Feature Importance Ranking Plots of Sy for Original and Simulated Datasets.

feature importance to compare their ability to replicate the original dataset's structure. The following conclusions from the results can be obtained:

- Mean imputation consistently provided the closest match to the original dataset for both Su and Sy, performing well in terms of correlation and model performance, making it a reliable and computationally efficient method.
- KNN imputation achieved the highest correlation with the original data and preserved feature importance rankings, making it ideal for cases where high fidelity is critical, though it can be computationally intensive.
- Regression imputation offered a balance between accuracy and simplicity, performing well but with minor deviations compared to mean and KNN methods.
- Random fill was the least reliable method, introducing significant deviations from the original data and disrupting feature importance, and is not recommended for applications where accuracy is crucial.

The associated recommendations can also be made:

- Mean imputation is recommended for most applications due to its balance of simplicity and performance.
- KNN imputation should be used when preserving the original data structure is essential, despite its higher computational cost.
- Regression imputation is suitable for cases requiring a compromise between accuracy and computational efficiency.
- Random fill should be avoided in contexts where preserving data integrity is critical.

Future research should explore more advanced imputation techniques and evaluate the effects of different missing data patterns to further optimize data handling in mechanical materials research.

ACKNOWLEDGEMENT

The author gratefully acknowledges the dataset author for providing the data that made this research possible.

REFERENCES

1. Yadav P & Saxena KK. Effect of heat-treatment on microstructure and mechanical properties of Ti alloys: An overview. Materials Today: Proceedings. 2020; 26: 2546-2557. https://doi.org/10.1016/j.matpr.2020.02.541
2. Czerwinski F. Current trends in automotive lightweighting strategies and materials. Materials. 2021; 14 (21): 6631. https://doi.org/10.3390/ma14216631
3. Bhong M, Khan TK, Devade K, Krishna BV, Sura S, Eftikhaar HK & Gupta N. Review of composite materials and applications. Materials Today: Proceedings. 2023. https://doi.org/10.1016/j.matpr.2023.10.026
4. Rubin DB. Inference and missing data. Biometrika. 1976; 63 (3): 581-592. https://doi.org/10.1093/biomet/63.3.581
5. Schafer JL & Olsen MK. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate behavioral research. 1998; 33 (4): 545-571. https://doi.org/10.1207/s15327906mbr3304_5
6. Nawale P. Materials and their Mechanical Properties; 2023. https://www.kaggle.com/dsv/5411884; DOI: 10.34740/kaggle/dsv/5411884
7. Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U & Higgins PD. Comparison of imputation methods for missing laboratory data in medicine. BMJ open. 2013; 3 (8): e002847. https://doi.org/10.1136/bmjopen-2013-002847
8. Zhang S. Nearest neighbor selection for iteratively kNN imputation. Journal of Systems and Software. 2012; 85 (11): 2541-2552. https://doi.org/10.1016/j.jss.2012.05.073
9. Scheffer J. Dealing with missing data. 2002.
10. Kalton G & Kish L. Some efficient random imputation methods. Communications in Statistics-Theory and Methods. 1984; 13 (16): 1919-1939. https://doi.org/10.1080/03610928408828805
11. Sedgwick P. Pearson's correlation coefficient. Bmj. 2012; 345. https://doi.org/10.1136/bmj.e4483
12. Tranmer M & Elliot M. Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR). 2008; 5 (5): 1-5.
13. Kriegeskorte N & Golan T. Neural network models and deep learning. Current Biology. 2019; 29 (7): R231-R236. https://doi.org/10.1016/j.cub.2019.02.034
14. Altmann A, Toloşi L, Sander O & Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010; 26 (10): 1340-1347. https://doi.org/10.1093/bioinformatics/btq134