## Supplementary material

### Data Collection and Description

The dataset utilised in this study was carefully compiled from a broad range of academic publications and technical reports. It encompasses a variety of steel grades widely applied in structural engineering, integrating comprehensive chemical composition data alongside physical features and empirical formulas related to the martensitic transformation. Only samples with complete chemical composition data (in wt.%) and experimentally measured Mₛ values obtained via direct methods such as dilatometry and metallographic analysis were retained. Data derived from indirect estimation or from testing protocols that were not comparable were excluded to ensure consistency.

As shown in Fig. 1, the dataset includes the weight percentages (wt.%) of key alloying elements such as Carbon (C), Manganese (Mn), Silicon (Si), Chromium (Cr), Nickel (Ni), Molybdenum (Mo), Vanadium (V), Cobalt (Co), Aluminium (Al), Tungsten (W), Copper (Cu), Niobium (Nb), Titanium (Ti), Boron (B), and Nitrogen (N). In addition to elemental compositions, the dataset incorporates several physical features that capture the metallurgical complexity of the alloys, including valence electron concentration (VEC), atomic radius mismatch (Δr), mixing entropy (ΔSmix), and carbon equivalent (Ceq). Furthermore, nine empirical equations (Mₛ-Eq. 1 to Mₛ-Eq. 9) derived from prior studies are included to estimate the martensite start temperature (Mₛ) from various theoretical perspectives.

Table 1 presents a statistical summary of these variables, outlining their means, standard deviations, and ranges. The alloying elements exhibit a broad distribution across the dataset, reflecting the diversity of steel compositions. The physical features and empirical equations likewise span wide ranges, offering comprehensive metallurgical insights. The target variable, martensite start temperature (Mₛ in °C), exhibits substantial variation, underscoring the impact of chemical composition and microstructural characteristics on phase transformation behaviour.

This dataset, integrating chemical composition, physical features, and empirical formulas, provides a solid foundation for developing predictive models of the martensite start temperature. By leveraging these multidimensional inputs, the study aims to achieve accurate, data-driven predictions of Mₛ, thereby supporting alloy design and heat treatment optimisation for a wide range of engineering applications.



**Fig. 1.** Histograms displaying the distribution of variables, including element content (wt.%), physical features (VEC, Δr, ΔSmix, Ceq), empirical Mₛ equations (Mₛ-Eq.1 to Mₛ-Eq.9), and measured martensite start temperature (Mₛ in °C), with the x-axis representing the value range and the y-axis indicating frequency.

**Table 1**

Statistical summary of variable measurements (Composition elements in wt.%).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variables** | **Mean** | **Standard Deviation** | **Maximum** | **Minimum** |
| C | 0.3557 | 0.277874 | 2.25 | 0 |
| Mn | 0.734011 | 0.955322 | 10.24 | 0 |
| Si | 0.305651 | 0.416359 | 3.8 | 0 |
| Cr | 1.013848 | 2.173832 | 17.98 | 0 |
| Ni | 3.764478 | 7.858497 | 31.539848 | 0 |
| Mo | 0.241302 | 0.597067 | 8 | 0 |
| V | 0.069491 | 0.285754 | 4.55 | 0 |
| Co | 0.198916 | 1.451447 | 16.08 | 0 |
| Al | 0.023735 | 0.172055 | 3.00679 | 0 |
| W | 0.176067 | 1.294297 | 18.59 | 0 |
| Cu | 0.037803 | 0.185048 | 3.037432 | 0 |
| Nb | 0.009358 | 0.114047 | 1.983237 | 0 |
| Ti | 0.014066 | 0.156069 | 2.524636 | 0 |
| B | 0.000021 | 0.000274 | 0.006 | 0 |
| N | 0.025931 | 0.187757 | 2.65 | 0 |
| VEC | 6.799524 | 1.656987 | 10.122562 | 4 |
| ∆r | 0.105176 | 0.062169 | 0.244205 | 0 |
| ∆Smix | 8.645843 | 4.76354 | 16.612942 | 0 |
| Ceq | 0.996449 | 0.621363 | 4.75 | 0 |
| MS-Eq.1 | 285.530454 | 126.012217 | 539 | -558.328 |
| MS-Eq.2 | 261.577519 | 144.822477 | 531 | -535.5 |
| MS-Eq.3 | 286.780987 | 125.026072 | 539 | -535.4 |
| MS-Eq.4 | 281.165538 | 132.231309 | 561 | -722.3 |
| MS-Eq.5 | 290.30417 | 135.542032 | 560.5 | -587.225 |
| MS-Eq.6 | 281.28983 | 121.054375 | 521 | -497.44 |
| MS-Eq.7 | 309.299511 | 122.135486 | 550 | -469.5 |
| MS-Eq.8 | 311.288829 | 121.941985 | 550 | -450.75 |
| MS-Eq.9 | 340.58 | 166.724221 | 554 | -796 |
| MS (℃) | 294.597297 | 114.537356 | 665 | -119.85 |

**Data preprocessing**

Generally, the original data obtained is incomplete and may contain a significant number of missing values, outliers, and duplicate entries. If these unprocessed data are directly used to build a prediction model, the prediction accuracy of the model would be adversely affected [1, 2]. Missing data handling was carried out before normalisation. Records containing incomplete composition data or absent Mₛ measurements were removed entirely rather than imputed to avoid introducing estimation bias in a multi-source dataset. Duplicate entries, identified by matching both composition and Mₛ, were also removed. The preprocessing steps were conducted in the following sequence: data filtering based on completeness and consistency criteria, duplicate removal, outlier detection and removal, and min–max normalisation of all features and the target variable. This sequential approach ensured that extreme or anomalous values did not distort scaling parameters and that the training dataset was both statistically and physically representative.

Outliers can significantly affect the accuracy and robustness of prediction models [2]. Therefore, it is crucial to detect and treat outliers appropriately. In this study, boxplot analysis was conducted to identify outliers in the dataset. Fig. 2 illustrates the boxplots for various input variables, including chemical elements, irradiation conditions and Yield Stress. Each boxplot provides a visual representation of the distribution of values for a particular variable, highlighting the central tendency and spread of the data, as well as identifying any outliers.

From the boxplots, it was evident that several variables contained outliers. These outliers appear as individual points that lie significantly outside the range of the majority of the data. To address the presence of these outliers, the Interquartile Range (IQR) method was employed. The IQR method involves calculating the range between the first quartile (Q1) and the third quartile (Q3) of the data. Data points that lie below Q1 - 1.5IQR or above Q3 + 1.5IQR are considered outliers.

By applying the IQR method, outliers were systematically identified and removed from the dataset. This process ensured that the remaining data is more representative of typical conditions and reduced the likelihood of skewed analysis or model performance. The removal of outliers contributes to the overall robustness and accuracy of the predictive models developed in this study.

Normalisation was applied as a data preprocessing method in the current study, as shown in Eq. (1).

 $X\_{norm}=\frac{X\_{max}-X\_{min}}{X-X\_{min}}$ (1)

Where $X\_{norm}$ is the normalised value $X$ is the original value, $X\_{min}$ and $X\_{max}$ are the minimum and maximum values of the dataset, respectively.

In this work, all chemical composition variables, physical descriptors (VEC, Δr, ΔSₘᵢₓ, Ceq), empirical equation outputs (Mₛ-Eq.1 to Mₛ-Eq.9), and the target variable Mₛ were normalised before model training. We adopted the min–max scaling method, transforming each feature to the range [0, 1] according to Eq. (1). This approach preserves the relative distribution of each variable while eliminating differences in units and magnitude, ensuring that no single feature disproportionately influences model optimisation. Normalisation was applied after outlier removal to prevent extreme values from distorting the scaling parameters. The choice of min–max scaling, rather than z-score standardisation, was motivated by the heterogeneous nature of the dataset and the bounded ranges of many alloying elements, making the scaled values directly interpretable in terms of proportion of the observed range. This preprocessing step also improves the numerical stability of gradient-based learning algorithms and accelerates convergence during training.



**Fig. 2.** Boxplots representing the distribution and central tendency of variables, including element content (wt.%), physical features (VEC, Δr, ΔSmix, Ceq), empirical Mₛ equations (Mₛ-Eq.1 to Mₛ-Eq.9), and measured martensite start temperature (Mₛ in °C). The vertical axis denotes the value range, and the red line within each box indicates the median. Outliers are shown as individual points beyond the whiskers.

**Hyperparameter Optimisation**

To ensure optimal performance and fair comparison, hyperparameter tuning was conducted for each of the four developed models using the Optuna framework. The selected hyperparameters for each model are summarised below.

For the Transformer-based model, the hyperparameters include settings for each of the three feature branches (chemical composition, physical features, and empirical formulas) as well as the Transformer Fusion module. This structure allows independent feature extraction in each branch before combining via the Transformer architecture. The detailed hyperparameters for the Transformer model are presented in Table 2.

**Table 2**

Hyperparameters for the Transformer Fusion Model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter** | **Branch 1 (Composition)** | **Branch 2 (Physical Features)** | **Branch 3 (Empirical Formulas)** | **Transformer Fusion** |
| hidden units | 128 | 80 | 112 |  |
| depth | 2 | 1 | 1 |  |
| dropout | 0.226 | 0.124 | 0.278 |  |
| learning rate (lr) |  |  |  | 0.00071 |
| batch\_size |  |  |  | 64 |
| activation |  |  |  | ReLU |
| tf\_d\_model |  |  |  | 128 |
| tf\_heads |  |  |  | 4 |

For the two-multilayer perceptron (MLP) models, hyperparameter tuning was performed to identify the optimal settings for network architecture and regularisation parameters. The first MLP model is based solely on chemical composition features, while the second integrates chemical composition, physical features, and empirical formulas as concatenated inputs without branching. The selected hyperparameters for these two models are summarised in Table 3.

**Table 3**

Hyperparameters for Transformer MLP Models.

|  |  |  |
| --- | --- | --- |
| **Hyperparameter** | MLP-chemical model | MLP-combined model |
| **hidden\_layer\_sizes** | (100, 100) | (100, 100) |
| **activation** | relu | relu |
| **alpha** | 0.00158 | 0.0086 |
| **learning\_rate\_init** | 0.00132 | 0.00245 |
| **max\_iter** | 736 | 705 |

For the Lasso regression model, the original compositional features were expanded using a second-degree polynomial transformation, incorporating both squared and interaction terms. All features were standardised before model fitting. The regularisation strength (α) was optimised via a binary search strategy to ensure that the number of non-zero coefficients did not exceed 15, balancing model sparsity and predictive performance. To preserve physical interpretability, a fallback linear regression was used to manually retain cobalt-related terms if they were entirely removed by the Lasso process. The detailed configuration of the Lasso regression model is summarised in Table 4.

**Table 4**

Configuration for the Lasso regression model

|  |  |
| --- | --- |
| **Hyperparameter** | **Value** |
| polynomial\_degree | 2 (including squared and interaction terms) |
| interaction\_only | False |
| include\_bias | False |
| scaler | StandardScaler |
| lasso\_max\_iter | 30000 (fallback 50000) |
| alpha\_search\_range | [1e-5, 10] |
| alpha\_selection\_method | Binary search for ≤15 non-zero coefficients |
| max\_nonzero\_coefficients | 15 |

## References

1. K. Lakshminarayan, S.A. Harp, and T. Samad, Imputation of Missing Data in Industrial Databases, Applied Intelligence, 11 (1999) 259-275. [https://doi.org/10.1023/A:1008334909089](https://doi.org/10.1023/A%3A1008334909089).

2. R. Blessing, Outlier Treatment in Data Merging, J. Appl. Crystallogr., 30 (1997) 421-426. <https://doi.org/10.1107/S0021889896014628>.