Computational Materials Science 237 (2024) 112896

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Full Length Article

Manufacturing process encoding through natural language processing for prediction of material properties

Ana P.O. Costa [a,*], Mariana R.R. Seabra [c], José M.A. César de Sá [a,b], Abel D. Santos [a,b]

[a] Department of Mechanical Engineering, University of Porto, s/n, R. Dr. Roberto Frias, Porto, 4200-465, Porto, Portugal
[b] INEGI, Institute of Science and Innovation in Mechanical and Industrial Engineering, Campus da FEUP, R. Dr. Roberto Frias 400, Porto, 4200-465, Porto, Portugal
[c] CFUL - centro de filosofia da universidade de Lisboa, Alameda da Universidade, Lisboa, 1600-214, Lisboa, Portugal

ARTICLE INFO

Dataset link: https://doi.org/10.17632/ysndxjs9gp.1

Keywords:
Manufacturing processing
Stainless steel
Machine learning
Natural language processing
Material design

ABSTRACT

Knowledge of manufacturing processes is crucial to determine the final properties of a material, thus this work focuses on analyzing the relationship between final properties, chemical composition, and manufacturing process through data analysis. Furthermore, techniques of natural language processing are used to encode the manufacturing process as input in the neural network. The work consisted of two main parts: firstly, the relevant data was gathered, cleaned-up, and analyzed using statistical and probabilistic methods, K-means, and Principal Components Analysis (PCA), and secondly, a model was developed to predict elongation, yield, and tensile strength. Fully Connected Neural Network (FCNN) algorithms were used to build the aforementioned model. In addition, in order to avoid overfitting and evaluate the model dropout function and K-fold cross-validation are incorporated. Results demonstrated reasonable accuracy in elongation, yield, and tensile strength. Two cases of mechanical properties prediction of stainless steel alloy were presented, first, an existing alloy that was not in the training and test set, and second a suggestion of a new stainless steel alloy, which combines good YS, UTS, and Pitting resistance equivalent number (PREN). Additionally, it was considered an example from the TRIP family to showcase the tool's versatility across various steel types.

1. Introduction

Manufacturing processes are crucial for the final characteristics of the material, as different thermal cycles, forming operations, and so forth, may result in different mechanical properties. Designing new alloys requires a comprehensive understanding of chemical composition and how varying element percentages can affect the final material properties, as well as an understanding of the manufacturing process. For decades, the design of new alloys has faced challenges and costly trial and error methods, which involve fully characterizing the physical and mechanical properties of the new materials, for instance, as described by Cann et al. [1] and WS [2]. As a consequence, there is a great amount of data available in the field of alloy metallurgy.

Steel industry is one of the oldest and the largest in the field of metallurgy, having produced a wide range of alloys, with good parameters for strength, corrosion, and many others. However, even these days, industries are still looking to design new alloys to attend to a variety of demands, e.g. biocompatibility combined with strength, producing high entropy alloys or magnetism combined with manufacturability as highlighted by Sidhu et al. [3], Yurchenko et al. [4] and Li et al. [5], respectively. These works exemplify the current demand for new alloys and, in fact, there are many ways in which chemical elements may be combined to produce new alloys that have not been attempted so far. Nevertheless, in the classical try-and-error method, it is not affordable in cost and time to combine and produce all these possibilities and to test which one is better.

The development of computational tools brings the possibility of solving this issue through numerical methods such as Machine Learning (ML), [6–9]. The use of these methods requires, to some extent, the definition of input and an expected output. In some cases, the relevant dataset is huge and complex, making the process of choosing the inputs difficult and compromising the expected results. Additionally, the amount of data may increase computational cost and time. These problems can be solved by resorting to data analysis, that supports the use of big data, and addresses imbalanced data issues [10,11].

The relevant data for alloy design may be divided into three main groups, namely chemical composition, material properties (mechanical, thermal, electrical, etc.), and information about the manufacturing process. By understanding the relationships between various parameters

* Corresponding author.
E-mail address: apcosta@fe.up.pt (A.P.O. Costa).

https://doi.org/10.1016/j.commatsci.2024.112896
Received 16 November 2023; Received in revised form 1 February 2024; Accepted 19 February 2024
Available online 26 February 2024

and mechanical properties, researchers and engineers can optimize the composition and processing of alloys to achieve specific performance goals. There is a diversity of ways to correlate and predict mechanical properties from material databases, besides a diversity of machine learning models that can be more appropriate for determined use. Nevertheless, according to the knowledge of the authors, up to the present moment, information on the manufacturing process has been either handled in a simplified way or not considered in most works on alloy design using ML.

Chemical composition and material properties comprise information usually available in numerical form, e.g. volume percentage of each chemical element in a certain alloy, and thus ready or nearly ready to be analyzed through numerical methods, including ML methods. On the contrary, the information on manufacturing process is often qualitative, posing an additional challenge to the use of the aforementioned numerical methods. As a result, existing works often focus on a restricted selection of manufacturing processes. For instance, Merayo et al. [12] focuses solely on tempered aluminum alloys, and employs a fully connected feedforward network of one input layer, three hidden layers, and one output layer to predict yield strength and ultimate tensile strength from chemical composition and Brinell hardness. Xu et al. [13] works with a database constituted by 76 rolled alloys and 36 extruded alloys to predict the yield strength (YS), ultimate tensile strength (UTS), and tensile elongation (EL) of a new extruded alloy. Other researchers, e.g. Conduit et al. [14] and Seabra and Costa [15], do not incorporate this information in their models, limiting their capabilities.

Therefore, the first goal of this work is to develop a suitable methodology to encode manufacturing information in a quantitative way and incorporate it with the other two types of information. The proposed strategy aims at handling various manufacturing processes in a systematized way in a unified model. In fact, when researchers pre-select a manufacturing process (e.g. Xu et al. [13]) there is a background assumption that clustering the available data according to such processes is the most adequate way to find correlations between chemical composition and material properties. Later in this work, it is shown that this is not always the case and thus an alternative approach needs to be developed.

Due to its qualitative nature, information related to manufacturing processes is typically presented in text format, with inherent difficulties in processing and interpretation. In Computer Science, Natural Language Processing (NLP) emerges as a solution to address this challenge [16]. NLP techniques enable machines to adapt and interpret text, capturing the context, and performing thorough analysis. By utilizing NLP, machines can effectively extract and provide to the user valuable insights from textual data. In this work, this leads to a more comprehensive understanding of the relationship between manufacturing processes and mechanical properties in the context of alloy design and development.

The methodology proposed in this paper offers significant contributions to the design of new steel alloys by addressing data-driven models to comprehend the intricate relationship between chemical composition, manufacturing processing, and resulting mechanical properties. Additionally, it establishes a new NN model to predict the mechanical properties, adequately including the manufacturing process information through natural language processing. The full methodology unfolds in three stages: firstly, it is shown through statistical analysis that the relation between the manufacturing process and mechanical properties does not exhibit a preferential direction across chemical compositions; secondly, a PCA and K-means clustering is applied to show that clustering the information according to manufacturing processes is not optimal and, finally, drawing from these analyses, a predictive tool is built based on NN, including dropout function and K-fold Cross-validation to avoid overfitting. Moreover, two strategies of natural language processing are compared and the efficiency of the proposed model is tested in three case studies. Two of the case studies concern duplex stainless steels, as data from a previous project was available [17,18]. To illustrate the generality of the tool for different grades of steel an example of the TRIP family is also considered.

This paper is organized as follows. In Section 2 PCA, K-means, NN, dropout, k-cross-validation, and Natural language processing are presented, including theoretical aspects and relevant implementation details. Next, the database used in this work is characterized in Section 3. The results obtained in the three main stages of development of the present methodology are presented and commented on in Section 4. This section also features two case studies which validate the model, namely the prediction of some mechanical properties of an alloy that was not part of the training data and the development of a new steel alloy. Finally, in Section 5 the final conclusions are outlined, and future work is proposed.

2. Methods

This section features an overview of the essential numerical techniques employed in this work: PCA, K-means clustering, and NN including dropout and k-fold cross-validation. Special attention is given to natural language processing techniques, which are essential to handle information on manufacturing processes.

2.1. PCA

Principal Component Analysis (PCA) stands as a prevalent technique applied in various fields, including data analysis and dimensionality reduction. This method aimed at optimizing the variance through the orthogonal projection of data onto a lower-dimensional linear space. According to Hotelling [19], this space, referred to as the principal subspace, allows for a concise representation of the original data while retaining its essential characteristics. Another perspective on PCA posits it as a linear projection that minimizes the average projection cost, quantified as the mean squared distance between data points and their corresponding projections — a concept initially introduced by Pearson [20], and revisited by Bishop and Nasrabadi [8].

PCA has practical applications in different areas, such as image processing and engineering. Furthermore, PCA can also be applied to visualize high-dimensional data in a low-dimensional space, making it easier to understand and interpret. In this paper, PCA is used with K-means clustering to compress data, extract essential information, and visualize high-dimensional data. In particular, it is evaluated how far pre-clustering data according to manufacturing process, chemical composition or material properties captures interesting regions in the data space.

In short, PCA may be executed by adhering to the following steps, as delineated in the study of [21]. For more details and alternative ways of executing PCA, interested readers may resort to the works of Vidal et al. [22] and Quarteroni et al. [23].

Step 1 Data normalization: Adjust the data to the same scale if the dataset exhibits varying scales.

Step 2 Calculate the covariance matrix: Calculate the covariance between variables using the formula:

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{1}$$

Representing the covariance matrix as:

$$\begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix} \tag{2}$$

Step 3 Calculate the eigenvectors and eigenvalues of the covariance matrix:

$$\Sigma v = \lambda v \tag{3}$$

where the covariance matrix is represented by Σ, the eigenvector by v, and its corresponding eigenvalue by λ.

$$det(\Sigma - \lambda I) = 0 \tag{4}$$

This second equation represents the determinant of the matrix $(\Sigma - \lambda I)$ equal to zero, where I is the identity matrix.

Step 4 Selecting components and creating a feature vector:

When aiming to diminish eigenvalues, arrange eigenvectors in descending order. This way, the larger eigenvalues will be linked to the main components that elucidate the largest amount of data variance. In order to determine the quantity of crucial elements to retain, it is crucial to compute the variance ratio explained by each eigenvalue.

Step 5 Transform the data: To modify the data based on the selected principal components, simply multiply the original data matrix by the chosen principal components matrix. This will produce a new matrix with reduced dimensionality, noticeable by the fewer columns in the resulting matrix when compared to the original data matrix.

$$Y = X * P \tag{5}$$

where X is the normalized matrix and P is the chosen principal component matrix.

2.2. K-means

K-means clustering is a widely used method for analyzing data in an unsupervised manner. It is a clustering algorithm that sorts a dataset into k clusters, with each data point assigned to the cluster whose centroid is closest to it, as pointed by Arthur and Vassilvitskii [24] and Bishop and Nasrabadi [8]. The objective of k-means clustering is to decrease the sum of squared distances between each data point and its designated centroid, which is initially selected randomly. This can be expressed in mathematical form as:

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{6}$$

where J is the objective function, K is the number of clusters, x is a data point, C_i is the set of data points belonging to cluster i, and μ_i is the centroid of cluster i.

The k-means algorithm aims to find the values of C_i and μ_i that minimize the objective function J. This is achieved through the iterative assignment and recalculation steps described above.

K-means clustering has a wide range of applications, including market segmentation, customer profiling, image segmentation, and anomaly detection. In material science, K-means clustering can be applied to identify different groups or clusters of alloys based on their chemical compositions and material properties, for example. This can be useful for researchers in materials science to better understand the relationships between chemical compositions and properties and/or to identify potential new materials with desired properties. In this research paper, the k-means algorithm is used to elucidate the interdependencies among chemical composition, mechanical properties, and manufacturing processes. The objective is to enhance cluster identification for optimization purposes and potentially improve the predictive modeling capabilities.

For instance, a study [25] applied k-means clustering to a database of steel alloys to identify different groups based on their chemical compositions and mechanical properties. The authors found that the k-means clustering algorithm was able to successfully identify different clusters of steel alloys based on their chemical compositions and mechanical properties. The authors also used principal component analysis (PCA) to reduce the dimensionality of the data and improve the performance of the clustering algorithm. Here, it is intended to go one step further by incorporating the manufacturing process information and using the clustering results in an inferential tool.

An essential consideration in k-means clustering is the number of clusters to use, or K. Smaller values of K produce bigger clusters, whereas bigger values of K produce smaller clusters. Here, the elbow technique has been employed, whose additional details may be found in the Appendix. The overall algorithm employed is given by the following steps:

step 1: Choose the number of clusters K.

step 2: Assign each data point to the nearest centroid.

step 3: Recalculate the centroids by computing the mean of all data points assigned to each centroid.

step 4: Iteratively execute steps 2 and 3 until either the centroids achieve stability or a predetermined maximum iteration threshold is attained.

2.3. Neural network

Neural Networks have become an increasingly popular model in machine learning. The most common approach involves supervised learning, whereby classified training data is utilized to recognize patterns between inputs and outputs. These networks are inspired by the architecture of the human brain, given their layers of nodes or neurons that can discern correlations between features in a dataset, as demonstrated in Fig. 1.

A potential limitation of this method is its lack of transparency. The manner in which the neurons establish connections to achieve the intended outcomes is not always evident, which may prevent the evaluation and enhancement of the dataset. Consequently, streamlining and enhancing results may prove to be a demanding endeavor.

In the context of this work, the fully connected neural network is used. Hence, an overview of this method is presented in the following paragraphs.

The FCNN has all neurons and layers connected, and this relation can be expressed by:

$$\mathsf{x}_i^k = \phi\left(\sum_j \mathsf{x}_j^{k-1} \mathsf{A}_{ij}^{k-1} \right) \tag{7}$$

where the x^k and x^{k-1} are vectors representing nodal values and bias of neurons in layers k and $k-1$, respectively, A_{ij}^{n-1} is the weight matrix, and $\phi(\mathsf{x})$ is the activation function.

The activation function plays a crucial role in determining whether or not a neuron is activated, thus introducing non-linearity to the neural network. Its primary function is to transform the summed weighted input of a node into an output value that can be passed on to the next hidden layer or used as output, as indicated in the research of Szandała [26]. With numerous activation functions to choose from, determining the optimal choice is not straightforward. However, after a thorough evaluation of the benefits of various activation functions, Relu was selected for its good performance and prevention of exploding gradients during backpropagation, as in the research of Seabra and Costa [15].

The Relu is defined as:

$$\phi(\mathsf{x}) = \begin{cases} \mathsf{x}\,, & \mathsf{x} > 0 \\ 0\,, & \mathsf{x} \leq 0 \end{cases} \tag{8}$$

During training, the weight matrices $\mathsf{A}ij$ hold the parameters that need to be adjusted. In this study, we utilize the Gradient Descent technique with the backpropagation algorithm, according to Rojas [27] to fine-tune these matrices. Essentially, the weights and biases are modified using the formula below:

$$\mathsf{w}^{t+1} = \mathsf{w}^t - \alpha \nabla_{\mathsf{w}} \mathsf{E}(\mathsf{w}^t) \tag{9}$$

where $\mathsf{E}(\mathsf{w})$, represents the error between the neural network's output and the desired output.

In this study, we utilized vector w^t and w^{t+1} to represent the weights and biases of the neural network at time instants t and $t+1$, respectively. The gradient of the error function concerning the weights and biases vector is denoted as $\nabla_{\mathsf{w}}\mathsf{E}(\mathsf{w})$, and a learning rate of 0.03 was employed. To evaluate the error, we employed a Loss function and specifically chose the Smooth L1 Function due to its ability to minimize sensitivity

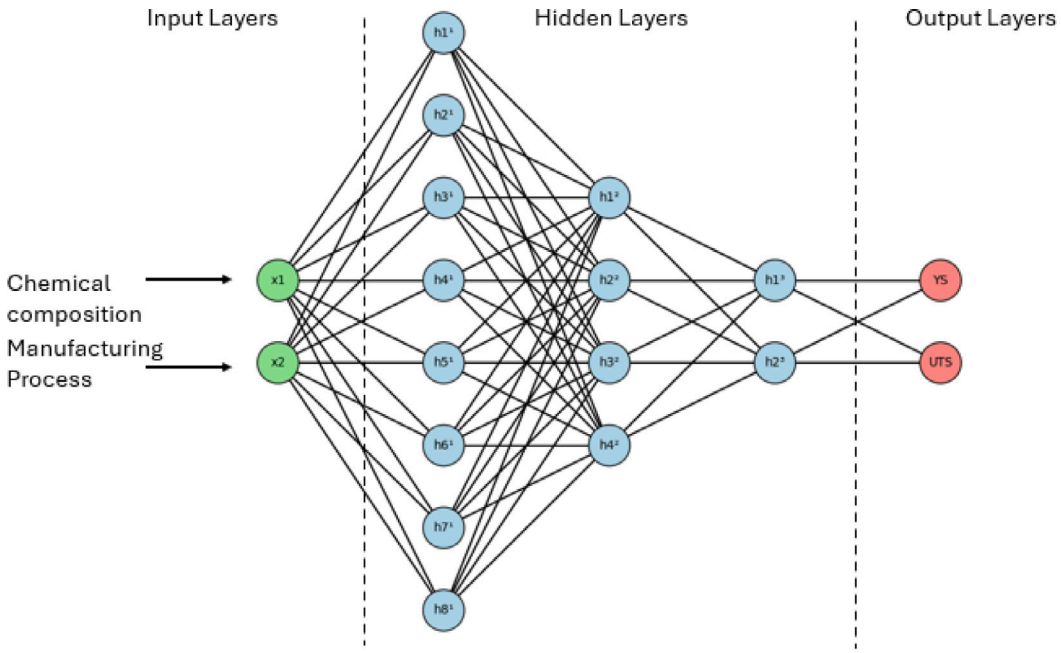


Fig. 1. Simplified representation of the neural network employed in this study. The original architecture comprises 256 input neurons, 3 hidden layers, and 2 output neurons designated YS and UTS. This reduced scheme illustrates the network's fully connected structure and functional layers.

to outliers and prevent explosive gradients. This function is defined as follows:

$$\begin{cases} \frac{0.5(y_i - y_p)^2}{\delta} & if \quad |y_i - y_p| < \delta \\ |y_i - y_p| - 0.5\delta\ , & otherwise \end{cases} \tag{10}$$

where y_i and y_p are the vectors containing the actual output and the predicted output, and δ is a hyperparameter, which was set equal to the learning rate, $\delta = \alpha = 0.03$.

2.4. Avoiding overfit

Overfitting is a common issue in machine learning algorithms. It happens when a model fits so well the training data that it loses the capacity to accommodate unseen data. Such ability to accommodate new data is referred to as the generalization of the model. Several signs could be observed in the model with overfitting: high training accuracy combined with poor performance on the test dataset, significant differences between training and validation performance, wrong training loss, fluctuations, or irregular behavior instead of decreasing the loss consistently during training. Some techniques can be applied to avoid overfitting, such as dropout, and k-cross validation, which are used in this research.

2.4.1. K-fold cross validation

K-fold cross-validation, a Monte Carlo method, is a technique used to assess how the predictive model is generalizable and avoids overfitting, as indicated by Berrar [28] in his study. It consists of regrouping the available learning set partitioned into k subsets of approximately equal size. These subsets are called 'fold'. The segmentation is accomplished by randomly sampling cases from the learning set without replacement. Hence, the model is trained using a $k - 1$ subset. There is no intersection in this method. Performance is assessed on the remaining subset, known as the validation set. Up to each of the k subsets has served as the validation set, this process is repeated.

The cross-validated performance is the average of the k performance measurements across the k validation sets. The error can be expressed by the equation:

$$\epsilon = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f_{-k}(xi)) \tag{11}$$

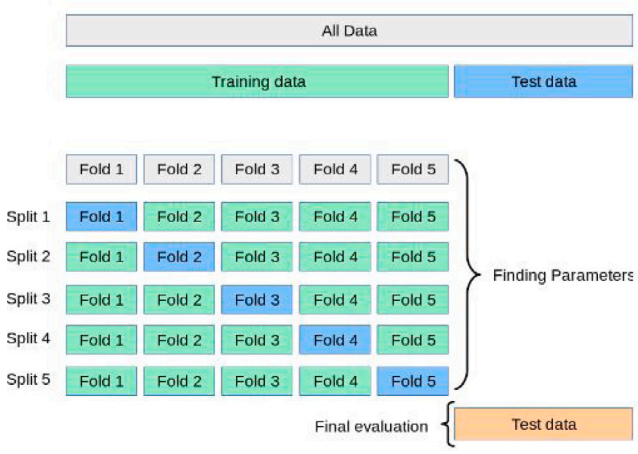


Fig. 2. Summary k-fold crossing validation [29].

where ϵ represents the cross-validated estimate of the predicted error, f_{-k} denotes the model that was trained on all but the k subset of the learning set.$\hat{y_i} = f_{-k}(xi)$ is the predicted or estimated value for the real class label, y_i, of case x_i, which is an element of the kth subset.

A summary of the process of k-fold crossing validation is present in Fig. 2, which shows a five-fold, train, and test set.

2.4.2. Dropout

Dropout is a regularization method that is introduced into the training dataset, according to the research paper of El Korchi and Ghanou [30] and in agreement with the research of Srivastava [31]. It performs parallel training of several neural networks with various architectures. That is, during training some number of layer outputs are randomly ignored, hence the prior layer becomes a new one with a different number of nodes and connectivity.

It is introduced in the feedforward operation, between hidden layers through a vector of Bernoulli random variables that has a probability p of 0 to 1. This vector multiplies element-wise with the output layer

to create the reduced output and is used as input to the next layer. Therefore, Eq. (7) becomes:

$$x_i^k = \phi\left(\sum_j x_j^{k-1}\tilde{A}_{ij}^{k-1}\right) \tag{12}$$

where

$$\tilde{A}_{ij}^{k-1} = A_{ij}^{k-1} r_{ij}^{k-1} \tag{13}$$

and r_{ij}^{k-1} are the independent Bernoulli random variables, each of which has a probability p of being 1. The outputs of the layer were thinned by sampling the Bernoulli vector and multiplying it elementally. The subsequent layer then receives the thinned outputs as input. Each layer receives this procedure. In essence, this is sampling smaller sub-networks.

2.5. Metrics

In this study, three specific measures are employed to assess the NN performance: the root mean square error (RMSE), the mean absolute error (MAE), and the R squared score (R2 score).

RMSE is given as:

$$\text{RMSE} = \sqrt{\frac{\sum(y_i - y_p)^2}{n}} \tag{14}$$

where y_i and y_p are vectors representing the actual and projected outputs, respectively, and n is the number of samples. The root mean square error (RMSE) is a measure of how concentrated the data is around the best-fit line. MAE, on the other hand, is the average deviation of all samples:

$$\text{MAE} = \frac{|y_i - y_p|}{n} \tag{15}$$

The R2 Score measures the percentage of correct predictions given by the model:

$$\text{R2 Score} = 1 - \frac{\sum(y_i - y_p)^2}{\sum(y_i - \bar{y}_i)^2} \tag{16}$$

where $\bar{y}_i$ is the mean of all real values. These three metrics are utilized to evaluate the RF and NN for a regression problem and identify the best model, that can be used for feature prediction.

2.6. Natural language processing

Natural Language Processing (NLP) is an important field of artificial intelligence. It encompasses a group of computational techniques for the analysis and representation of human languages, permitting computer programs to handle human language as it is spoken and written.

Neural networks can be enhanced in their predictive capabilities through the incorporation of natural language information through various NLP techniques. This paper leverages three such techniques — tokenization, information extraction, and encoding

Tokenization

Tokenization is the process of breaking down a character sequence into smaller units called tokens, often accompanied by the removal of certain characters like punctuation. It helps convert unstructured textual data into structured data by breaking sentences into individual components that can be analyzed using mathematical models, as demonstrated in the study by Rai and Borah [32].

For example, let us consider the phrase "Natural Language Processing is an important field of artificial intelligence". When tokenized, the output would be: "Natural", "Language", "Processing", "is", "an", "important", "field", "of", "artificial", "intelligence".

An instance where tokenization is used in this work is in the flow chart illustrated in Fig. 3, which provides a summary of the combined techniques of tokenization, information extraction, and encoding.

Information extraction

Data mining research uses a variety of techniques to extract useful information from source documents. This information extraction encompasses identifying specific pieces of information in a text document from written texts or speech transcripts, and converting them into structured representations, as evidenced by the research conducted by Ji [33].

In other terms, it makes an effort to draw insightful conclusions from vast amounts of raw data by choosing pertinent entities from unstructured datasets and then classifying them appropriately. An example is shown in Fig. 3, which summarizes the process of creating a historic record of the manufacturing part from a determined material. In the figure, MPF represents the manufacturing process, Temp. represents the temperature, Dim. represents dimension, and Geo. geometry.

Encoding

There are many techniques to encode natural language into computer language. Label, One Hot encoding, Bag-of-Words, Term Frequency–Inverse Document Frequency, and Word2Vec are some of the techniques used, to name a few.

Label encoding is a simple technique that transforms human language into computer language by assigning a unique numerical label to each category, as explained in [29]. However, this method has limitations as the machine cannot relate the meaning and the context of the phrase, unlike other sophisticated encoding methods like word embeddings. The main issue with using label encoding in neural networks is that the algorithm cannot distinguish between the numerical difference of encoded values and other variables. This means the algorithm may assign higher importance to encoded values than to numerical variables. For example, if we label encode the phrase "Natural Language Processing", it would be represented as follows:
Natural [1]
Language [2]
Processing [3]

Meaning that an inferential model may attribute higher importance to the word *Processing*.

One-hot encoding stands as a fundamental technique within the realm of natural language processing, providing a means to numerically represent categorical variables. This involves creating binary value vectors that align with each distinct category found within a given dataset, as elucidated by Chen [34]. An example is using the words "Natural Language Processing":

Natural [1,0,0] language [0,1,0] Processing [0,0,1]

On the opposite side of label encoding, one hot encode does not actively influence the weight the NN is going to give to the encoded input.

Other more advanced methods include word embedding techniques, which have the capability to capture contextual information, for instance, Bag-of-Words (BoW), which attempts to create a frequency-based vector representation by counting the occurrence of each word in a given text, as highlighted in the research of Zhao and Mao [35]. Another example is the Word2Vec technique, as emphasized in the research conducted by Church [36], which involves representing every word in a predetermined vocabulary using a vector of numerical values. Nevertheless, in this project, two straightforward techniques will be utilized: label encoding and one-hot encoding.

While label encoding is not traditionally associated with natural language processing, it will be employed to assess the impact of integrating NLP techniques into the neural network model. These techniques were chosen for their easy comprehensibility and implementation, which simplifies the working process and result interpretation. The selection of these techniques is based on specific advantages, particularly the strengths of one-hot encoding in comparison to alternative encoding methods. Additionally, its compatibility with a wide range of machine-learning algorithms makes it a versatile option for input encoding.

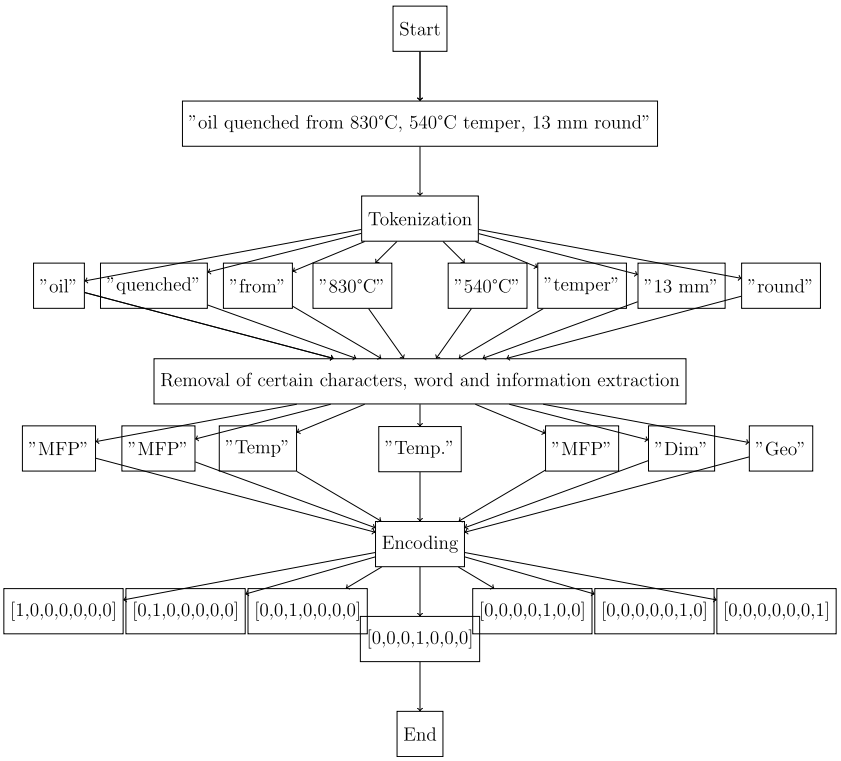


Fig. 3. Flowchart: An example of tokenization, information extraction, and encoding in a manufacturing process information of steel.

3. Dataset

A database was built to perform the proposed work, containing information on several steel alloys. The information gathered includes the chemical composition, mechanical properties, and manufacturing processes of these alloys. Concerning the mechanical properties, data on hardness, yield, tensile strength, and elongation were collected. Regarding the manufacturing processes, information such as the process, temperature, dimensions, and time of the manufactured part was compiled. For chemical compositions, information on twenty-six different chemical elements, namely Cr, Al, B, Co, Mo, Ni, Ti, Zr, C, Fe, Pd, Mn, P, Si, N, Cu, Nb, Se, Ta, W, V, and S, was gathered. All information resulted in a total of approximately thirteen thousand data points. The data was extracted from Refs. [37] and [38]. Furthermore, the database was cleaned and organized.

4. Results and discussion

The main goal of this paper is to build an inferential tool that can be used to discover new steel alloys according to some target properties and manufacturing process. The main innovation in relation to other existing tools based on ML is the way in which the information regarding the manufacturing processes is handled. The success of any ML technique depends, among other factors, on the adequacy of the technique to the data, therefore, in the first part of this section, a preliminary statistical analysis of the database of steel alloys and corresponding manufacturing processes is performed. Next, it is shown through PCA and K-means clustering that pre-grouping data according to manufacturing process may not yield optimal predictive results, followed by the predictive model, based on NN. This section closes with the presentation of three case studies, namely, the inference of material properties of two alloys which were not present in the training data and an example of a design of a new alloy.

4.1. Pre-processing and statistical analysis

Firstly, a statistical analysis was run to identify the characteristics of manufacturing processes present in the working database, as well as to understand the frequency of those processes, which is crucial to further design an inferential tool avoiding underfitting and overfitting. Figs. 4, 5(a), 5(b) and 5(c) show the quantity of each manufacturing process, the dimension of the specimens employed and their temperature, when applicable. In fact, many alloys experienced more than one type of manufacturing processing in this dataset. This analysis considered just the first two processes of each alloy, as only a few ones comprised a third process. Some entries also contained information on the duration of the manufacturing process. Again, as only a few alloys contained this information, it was not used in this preliminary analysis. However, the predictive model presented in Section 4.3 was built using all information of processing, such as specimen dimension, temperature, duration, and types of manufacturing processing. Therefore, the most frequent processes in the database are “annealed”, “aged”, “quenched”, “cold drawn”, “normalized”, and “as received”, as displayed in Fig. 4. “As received” in this database is assigned to alloys that do not have clear manufacturing processes.

Fig. 5 complement the information presented in Fig. 4. In Fig. 5(a) is possible to see that, there is a high quantity of alloys that do not use a second manufacturing processing, a total of 68.2%. For alloys that used the second manufacturing processing, the most frequent are “tempered”, “stress relieved” and “aged”. Observing the information of dimension and temperature, Figs. 5(b) and 5(c), parts with dimensions less or equal to 50 mm are most commonly found in the database, addressing temperature, there is more information about the process with greater or equal to 800 °C.

Next, it is inspected if any of the pairs chemical composition-mechanical properties, manufacturing process-chemical composition, or manufacturing process-mechanical properties, has the potential to be replaced by some sort of linear relation.

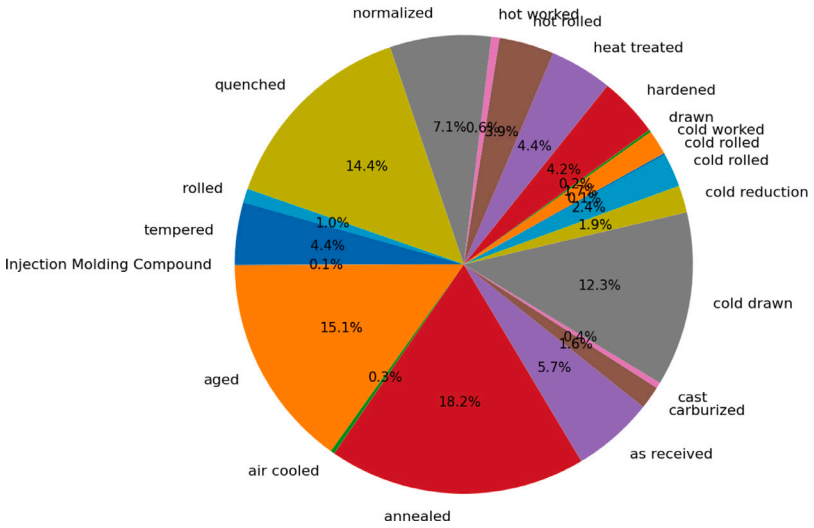
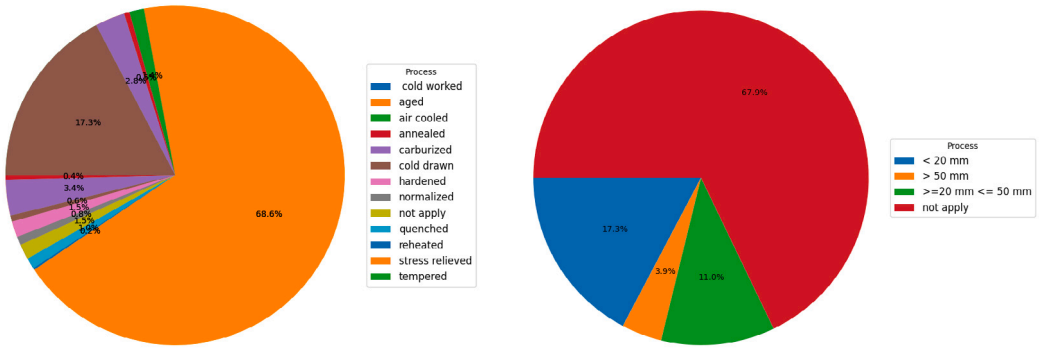
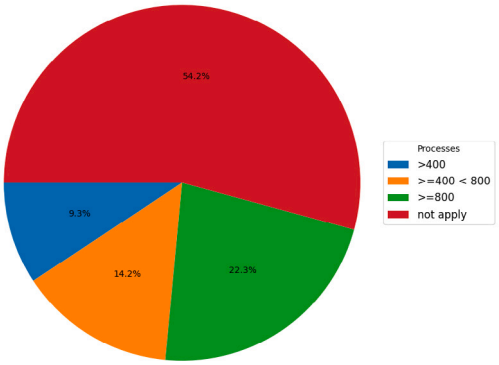


Fig. 4. Manufacturing process distribution.



(a) Second process

(b) Dimension



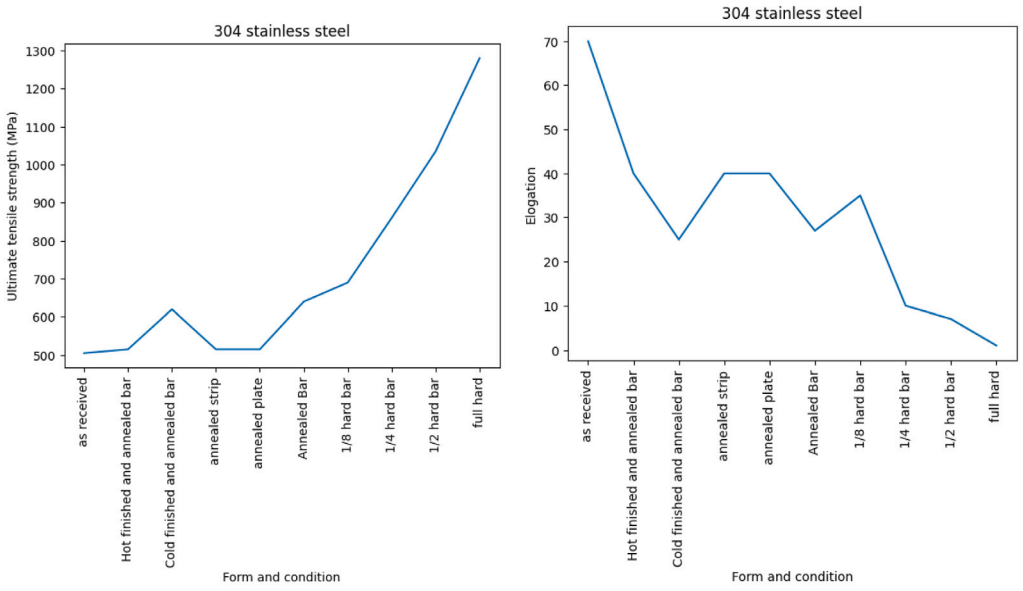
(c) Temperature

Fig. 5. Manufacturing process distribution, a-Second process, b-Dimension of manufacturing parts, c- Temperature of manufacturing parts Distribution (°C).

Two alloys were randomly picked from the database, and the values of UTS and Elongation were plotted on graphs, to evidence the impact of the manufacturing processing on mechanical properties. Fig. 6 shows the UTS and Elongation versus the form and condition of alloy AISI 304, Stainless steel, and Fig. 7, shows the UTS and Elongation versus the form and condition of alloy AISI 1030.

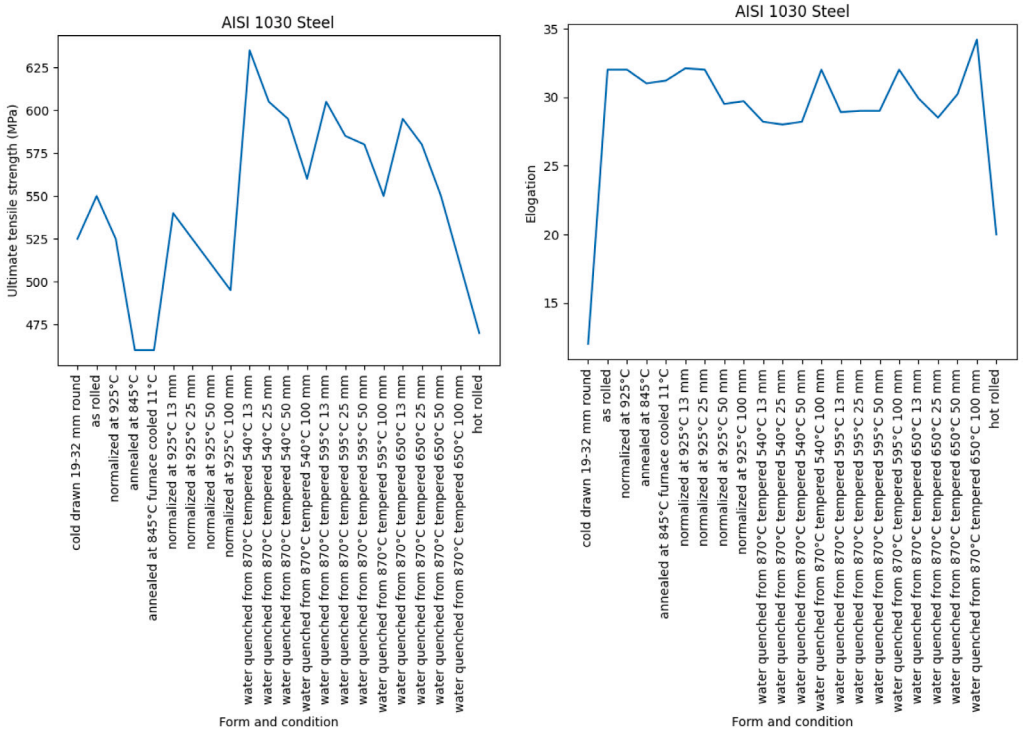
As expected, Figs. 6 and 7 show that changing the manufacturing process, the values of mechanical properties can have a drastic change. Moreover, the comparison between the two alloys also shows that the manufacturing process does not impact the final properties in a preferential way, and thus, the pair chemical composition-manufacturing process cannot be replaced by information on mechanical composition alone, when targeting a certain mechanical property. This analysis may be extended to the other pairs, as illustrated by the heatmap in Fig. 8 and the graphs in Figs. 9 and 10.

The heatmap in Fig. 8 exposes potential linear relations between the twenty-six chemical elements and four mechanical properties. This map brings some insights into how a change in chemical composition is



(a) Ultimate tensile strength (UTS) (b) Elongation

Fig. 6. UTS and Elongation versus the form and condition of alloy AISI 304, Stainless steel.



(a) Ultimate tensile strength (UTS) (b) Elongation

Fig. 7. UTS and Elongation versus the form and condition of alloy AISI 1030.

related to a change in mechanical properties. In this analysis, a *classical* Pearson method[1] is used to calculate the correlation between variables. In this case, values close to 1 mean a perfect linear relation, and −1 a perfect inverse relation, between 0 and 1, a positive linear relationship is observed, and between 0 and −1 an inverse linear relationship is observed. When zero appears in the heatmap means that there is not any linear relationship between the two variables, however, there may exist a non-linear relation between them.

In this case, there is a high value relating mechanical properties vs. mechanical properties, namely, yield strength vs. ultimate tensile strength, and vs. hardness, and for yield strength vs. elongation is verified an inverse linear relation. When observing the chemical

[1] in this work the classical version of the Pearson method was employed. Although there are extensions of the method to inspect possible non-linear relations, here the correlation presented only quantifies the intensity of a linear relationship between two variables.

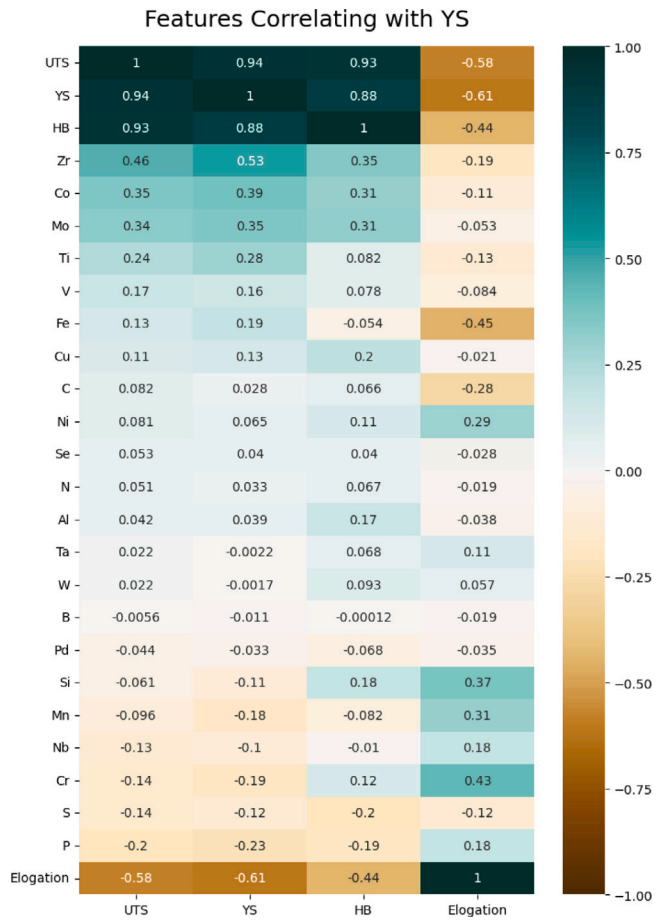


Fig. 8. Heatmap showing potential linear relations between mechanical properties and chemical composition, concerning the full database of steel alloys.

composition vs. mechanical composition a relevant linear relationship is observed between the Zr to UTS with the value of 0.53, and for elongation, Fe has a relevant inverse linear relation, and Cr presents a linear relation with elongation of 0.43.

For a better understanding of graphs were plotted in order to highlight some eventual relations between (i) mechanical properties and chemical composition, and (ii) mechanical properties vs. mechanical properties. The graph of Zr vs. UTS, Fig. 9(a), does not show a clear relationship between the variables, however, it still be possible to affirm that in all alloys that were added Zr, the ultimate tensile strength does not present values below 800 MPa in this database. In terms of Fe Vs. elongation, Fig. 9(b) as the data is sparse is difficult to confirm a linear relation, despite that it is possible to see on the graph, that when the Fe is low a high value of elongation is observed. Finally, the analysis of Cr Vs. Elongation, Fig. 9(c) faced the same problem as the previously mentioned graphs, and thus, due to dispersed data, it is not possible to draw any conclusion, regarding any straightforward relation.

In terms of mechanical properties, some relations are confirmed as UTS and YS have an almost perfect linear relation, which is also verified in hardness UTS, and YS. Elongation otherwise, has a clear inverse tendency, as for high values of UTS, YS, and hardness, it is possible to see low values of elongation, Figs. 10(a)–10(f). Graphs in Fig. 10 show some limiting tendencies that may be further explored when searching for new alloys, e.g., finding an alloy with improved elongation for high UTS. In particular, the combination of ML models with Pareto optimization allows the identification of optimal trade-offs between competing properties to support decisions during the design process. For instance, it can pinpoint chemical composition or processes that achieve an ideal compromise between yield strength and elongation, as illustrated in Fig. 11.

4.2. PCA and K-means clustering

The statistical analysis in the previous section showed that the effect of the manufacturing process on final mechanical properties is non-trivial across chemical compositions. Nevertheless, it is still legitimate to wonder whether clustering the data according to manufacturing processes or according to another intuitive criterion such as chemical composition or mechanical properties captures interesting regions in the data space. In this section, this evaluation is done by resorting to PCA and K-means clustering.

Clustering according to manufacturing process

In this first case, data comprising chemical compositions and mechanical properties is compressed into two main components through PCA and then, the clustering algorithm is applied. Fig. 12, shows the clustering by manufacturing processing and actual clustering determined by the algorithms. In this case, the label encoder is used to correlate the clusters with the manufacturing processes based on converting categorical information into numerical labels. These numerical labels are then used to associate the clusters with specific manufacturing processes. As it can be observed, the clusters found by the algorithm do not correlate well with a specific manufacturing process, meaning that manufacturing process simpliciter does not determine if one alloy is related to another.

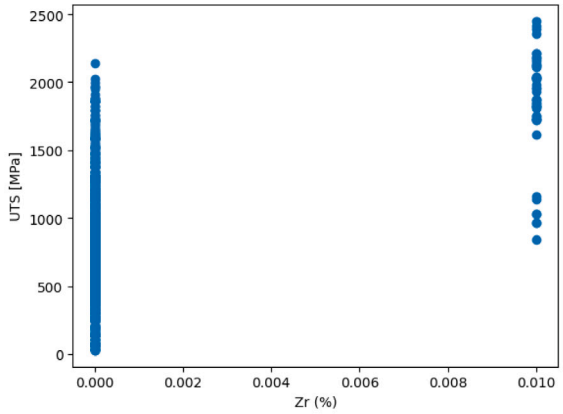
Clustering according to mechanical properties

The second analysis attempts to establish a relationship between manufacturing processes and chemical composition and uses prevalent UTS or elongation information to identify clusters. To prevent bias in the results of PCA and incorrect interpretation, a one-hot encoder was used instead of label encoding to process the manufacturing process. As previously referred to in Section 2.5, label encoding often introduces a disproportionate amount of variance and increased dimensions, as the encode variables are given by $x = \{0, 1, \ldots, n\}$, where $x \in \mathbb{N}$. On the other hand, one hot encode uses 0 to indicate that a process is absent and 1 to indicate that a process is present, and thus it does not privilege numerically any of the processes. Further justification for the use of one hot encoding is given in Section 4.3, where significant differences can be observed in the training of NN. Results are displayed in Fig. 13 where similarly to the first analysis, information regarding UTS or elongation was still label encoded.

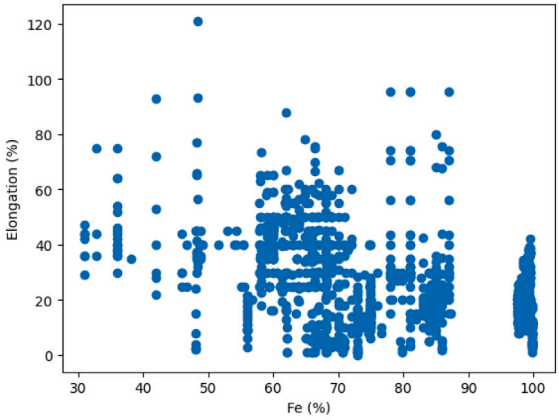
Once again, there is no evident association between each cluster and a certain mechanical property. Although one of the clusters is solely associated with UTS (orange region), UTS alone is not enough to distinguish between regions of the data space, as it is also present in the green region. As a result, pre-clustering data according to properties is also not optimal. Although the other mechanical properties such as yield strength or hardness, just to mention two examples, could be used, results are similar to those displayed.

Clustering according to chemical elements

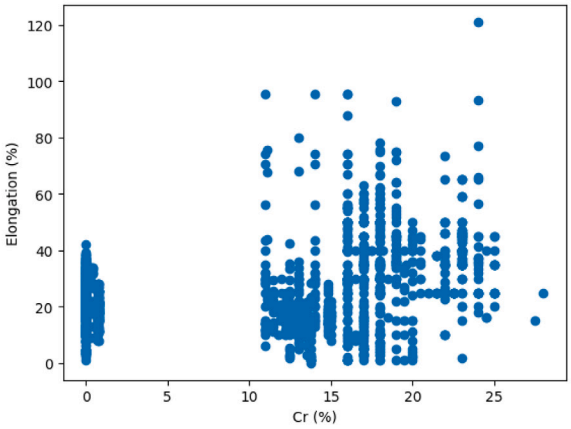
In this third trial, it is inspected if the presence of a particular chemical element is significant in discerning regions in the data space. The analysis follows on similar grounds as in the other two cases, however, here the PCA components are determined from the data corresponding to mechanical properties and manufacturing processes. Clusters are compared with the predominant element in each alloy besides Fe. Again, one-hot encode is used to process the manufacturing processing information, and label encode is used to correlate categorical data of chemical composition with the cluster. Results are displayed in Fig. 14, where some clear clusters can be identified, but they do not correlate with a dominant chemical element.



(a) Correlation between ultimate tensile strength (MPa) and chemical element Zr (%).



(b) Correlation between elongation (%) and chemical element Fe(%).



(c) Correlation between elongation (%) and chemical element Cr(%).

Fig. 9. Correlation between mechanical properties and chemical composition, a- ultimate tensile strength (MPa) and chemical element Zr (%); b- Correlation between elongation (%) and chemical element Fe(%); c-elongation (%) and chemical element Cr(%).

The above analysis suggests that optimal clustering should integrate information from chemical composition, mechanical properties, and manufacturing processes in a unified way. As a result, new main PCA components were calculated using all the information in the database. Then, K-means clustering was applied to the first 2 components and to the first 3 components, generating the 2D clusters and 3D clusters, respectively. 2D results for 5 clusters are displayed in Fig. 15, while 3D results for 4, 5, and 6 clusters are displayed in Fig. 16.

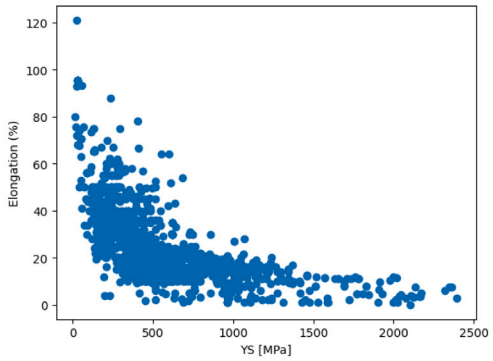
As previously referred, an elbow method was used to pick the right number of clusters, which is the same for 2D and 3D clusters. For completeness and comparison purposes, a plot with a lower number of clusters, and a plot with a higher number of clusters are also shown in Fig. 15. As a final conclusion, the right way to correlate data is more complex than pre-picking clusters, by manufacturing process, prevalent element in chemical composition, or prevalent mechanical properties, as has been practiced previously in the literature. The most adequate way to group the data is by 5 distinct clusters with a combination of manufacturing process, material properties, and chemical composition.

Clusters found by K-means applied to predictive models may allow adjustment of the ideal structure, and potentially improve the performance of the NN model when compared to a single model using the entire data, as it can focus on different patterns. On the opposite side, some of the clusters may not contain enough data to find better results. Hence in this work, clusters are separated and tested in NN models, and the hyperparameters were adjusted to achieve the best results in predicting the mechanical properties. The results of this topic will be presented and discussed in the next subsection.

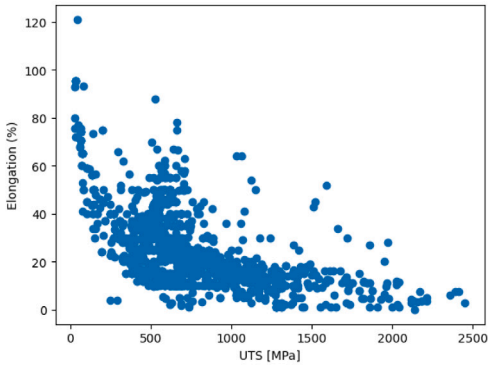
4.3. Inference through NN

A Neural Network model with four hidden layers following the description given in Section 2.3 was built. Input consisted of chemical composition and manufacturing processing was built. The manufacturing processing, which is textual data, was treated using label encoder and one hot encoding, according to Section 2.6. Output for general validation of the model consisted of YS and UTS. In the case studies, elongation was also considered. To achieve optimized results dropout and K-fold Cross-validation are also added. All information on testing parameters, such as neurons, dropout, and k-fold cross-validation and metrics are presented in Tables 1 to 6. Results concerning label encoder for YS and UTS are presented in Tables 1 and 2, respectively. While results concerning one hot encoding are presented in Tables 3 and 4. It can be easily noticed that one hot encoder surpasses the performance of label encoder and thus, the performance of individual folds in k-fold cross-validation is displayed for this method in Tables 5 and 6. While both techniques can be applied in neural networks, one-hot encoding offers a more transparent representation of categorical data while circumventing the introduction of misleading ordinal relationships.

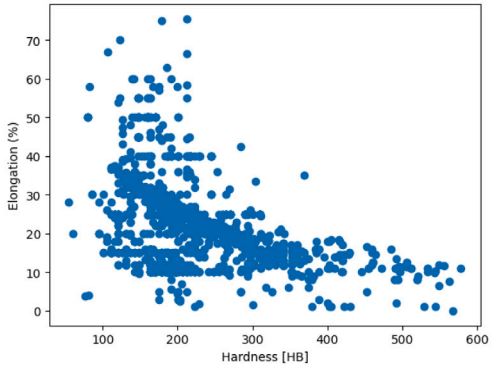
Dropout in the generalization model does not seem to present an improvement in the results, however, when k-folds are analyzed it is possible to see a bit of improvement in some folds. Some results from K-cross-validation are presented to be noticeable in these differences, as in the Tables 3, and 4, the results of the model are presented as a mean of all folds. The best results from yield and ultimate tensile



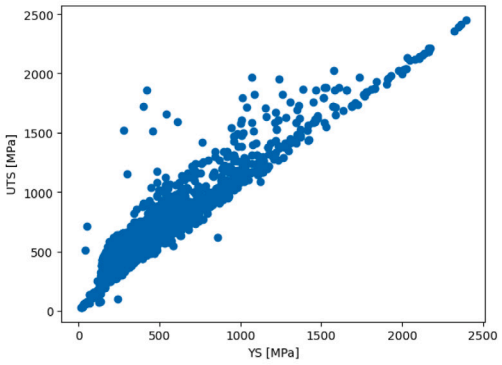
(a) Correlation between elongation (%) and YS [MPa]



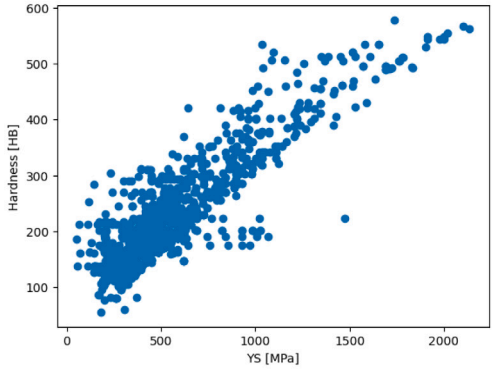
(b) Correlation between elongation (%) and UTS [MPa]



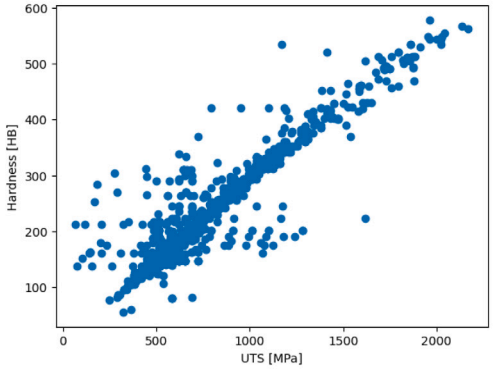
(c) Correlation between elongation (%) and Hardness [HB]



(d) Correlation between UTS [MPa] and YS [MPa]



(e) Correlation between Hardness [HB] and YS [MPa]



(f) Correlation between Hardness [HB] and UTS [MPa]

Fig. 10. Correlation between mechanical proprieties, a-elongation (%) and YS [MPa], b-elongation (%) and UTS [MPa], c- elongation (%) and Hardness [HB], d- UTS [MPa] and YS [MPa], e- Hardness [HB] and YS [MPa], f-Hardness [HB] and UTS [MPa].

strength are presented in the inference graphs, Figs. 17 and 18, both have a NN structure with 4 hidden layers, and 240 neurons, dropout in three layers, of 0.25 and represent the second k-fold.

Structures for individual K-means clusters were also tested. Tables 7 and 8 show the best-performing structures, for each cluster, and Tables 9 and 10 display the results in the prediction of yield strength and UTS, respectively. Some clusters have outstanding results, such as Cluster 5, and Cluster 4, and other has reasonable results, such as Cluster 1, 2, and 3. This can be explained by the unbalanced distribution of data. Cluster 4 and 5 concentrate approximately 50% of data, and the other 50% is distributed by the other clusters. The predictive results may improve with the incorporation of k-means clusters, however, it is necessary to have a higher and more balanced distribution of a dataset.

4.4. Ablation study

To assess the impact of including manufacturing process data in neural network (NN) models, an ablation study was conducted using an optimized NN architecture. This architecture comprises three hidden layers, 240 hidden neurons, and three dropout layers with a dropout rate of 0.25. The model utilized 4 K-folds for validation and a one-hot encoder for processing manufacturing process data (see Tables 11 and 12).

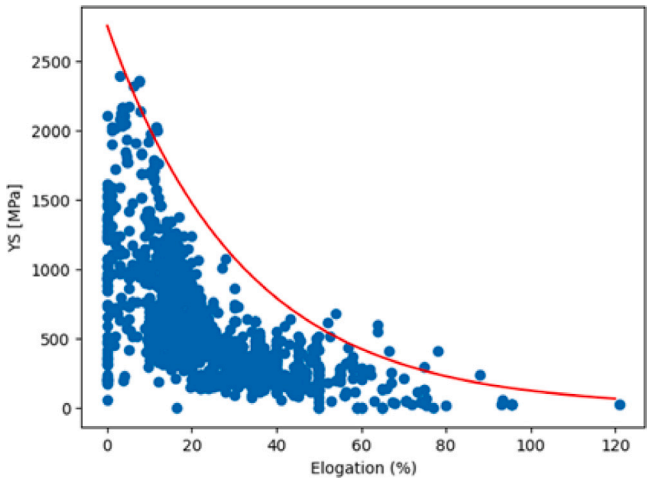


Fig. 11. This scatter plot illustrates the relationship between yield strength (YS) and elongation percentage, revealing the interplay between material strength and ductility. As elongation percentage increases, yield strength decrease, indicating a trade-off between these properties. This visualization underscores the potential of using machine learning models and Pareto optimization for advanced material design, aiming to optimize the balance between strength and ductility for material development.

Table 1
Evaluation of NN performance through various metrics in predicting **Yield Strength** with **label encoder** for manufacturing processes input. This table showcases a range of NN models, each characterized by its architecture: the configuration of hidden layers/total neurons, alongside the deployment of dropout (number of hidden layers employed and value) within these layers. The influence of varying K-folds in the training process is also examined.

NN structure	Dropout	K-fold	Training loss (MPa)	RMSE (MPa)	MAE (MPa)	R2 Score
3/240	3/0.25	4	77	195	119	0.74
3/240	–	4	83	196	121	0.73
3/240	1/0.25;2/0.5	4	77	166	1117	0.73
3/240	3/0.25	2	66	207	130	0.71
3/240	3/0.25	6	77	194	116	0.74
3/287	3/0.25	4	94	202	126	0.72
3/120	3/0.25	4	74	184	114	0.76

Table 2
Evaluation of NN performance through various metrics in predicting **Ultimate Tensile Strength** with **label encoder** for manufacturing processes input. Each varying architecture is characterized by hidden layers/total neurons, alongside the deployment of dropout (number of hidden layers employed and value) within these layers. The influence of varying K-folds in the training process is also examined.

NN structure	Dropout	K-fold	Training loss (MPa)	RMSE (MPa)	MAE (MPa)	R2 Score
3/240	3/0.25	4	81	236	131	0.64
3/240	–	4	83	303	152	0.32
3/240	1/0.25;2/0.5	4	79	251	143	0.57
3/240	3/0.25	2	73	226	139	0.67
3/240	3/0.25	6	90	211	130	0.69
3/287	3/0.25	4	75	196	133	0.59
3/120	3/0.25	4	96	224	144	0.67

The study reveals that integrating manufacturing process information significantly enhances NN performance in predicting both yield and ultimate tensile strength. Notably, the addition of manufacturing process data markedly reduces training loss and RMSE, while substantially improving the R2 score. This improvement indicates a more precise and reliable model, better capturing the complex interplay between input variables and mechanical properties. These findings also shed light on the varying sensitivities of different mechanical properties to manufacturing processes. While the improvement is more pronounced in predicting yield strength, the notable enhancement in predicting ultimate tensile strength underscores the nuanced impact of manufacturing techniques on material properties.

Table 3
Evaluation of NN performance through various metrics in predicting **Yield Strength** with **one hot encoder** for manufacturing processes input. The different range of NN models is characterized by the configuration of hidden layers/total neurons, alongside the deployment of dropout (number of hidden layers employed and value) within these layers. The influence of varying K-folds in the training process is also examined.

NN structure	Dropout	K-fold	Training loss (MPa)	RMSE (MPa)	MAE (MPa)	R2 Score
3/240	3/0.25	4	29	179	101	0.78
3/240	–	4	30	179	104	0.78
3/240	1/0.25;2/0.5	4	33	186	105	0.76
3/240	3/0.25	2	29	200	115	0.73
3/240	3/0.25	6	30	189	102	0.74
3/287	3/0.25	4	32	185	105	0.76
3/120	3/0.25	4	37	208	109	0.70

Table 4
Evaluation of NN performance through various metrics in predicting **Ultimate Tensile Strength** with **one hot encoder** for manufacturing processes input. This table showcases a range of NN models, each characterized by its architecture: the configuration of hidden layers/total neurons, alongside the deployment of dropout (number of hidden layers employed and value) within these layers. The influence of varying K-folds in the training process is also examined.

NN structure	Dropout	K-fold	Training loss (MPa)	RMSE (MPa)	MAE (MPa)	R2 Score
3/240	3/0.25	4	35	201	108	0.74
3/240	–	4	36	196	111	0.74
3/240	1/0.25;2/0.5	4	34	196	110	0.74
3/240	3/0.25	2	32	159	135	0.61
3/240	3/0.25	6	35	205	108	0.73
3/287	3/0.25	4	38	205	112	0.72
3/120	3/0.25	4	44	202	112	0.73

Table 5
Evaluation of NN performance in Predicting **Yeld Strength** with **one hot encoder** for encoding manufacturing processes input in different folds. This table showcases different folds of k-fold cross-validation to NN structure with 4 hidden layers, and 240 neurons, using k-fold cross-validation with 4 folds and without inclusion dropout.

K-fold	Training loss (MPa)	RMSE (MPa)	MAE (MPa)	R2 Score
1	44	193	108	0.70
2	36	213	118	0.68
3	26	198	108	0.79
4	42	183	108	0.79

Table 6
Evaluation of NN Performance in predicting **Ultimate Tensile Strength** with **one hot encoder** for encoding manufacturing processes input in different folds. This table showcases different folds of k-fold cross-validation to NN structure with 4 hidden layers, and 240 neurons, using k-fold cross-validation with 4 folds and with inclusion of dropout of 0.25 in three layer.

K-fold	Training loss (MPa)	RMSE (MPa)	MAE (MPa)	R2 Score
1	35	174	105	0.79
2	40	194	98	0.82
3	30	243	115	0.61
4	34	194	112	0.75

4.5. Case-studies

To prove the efficiency of the method three cases are exposed. In the first and second case, existent steel alloys that were not a part of the training and test set are used to validate the method. In the third case, a new combination and percentage of elements is proposed attempt to ally good elongation, UTS, and Pitting resistance equivalent number (PREN).

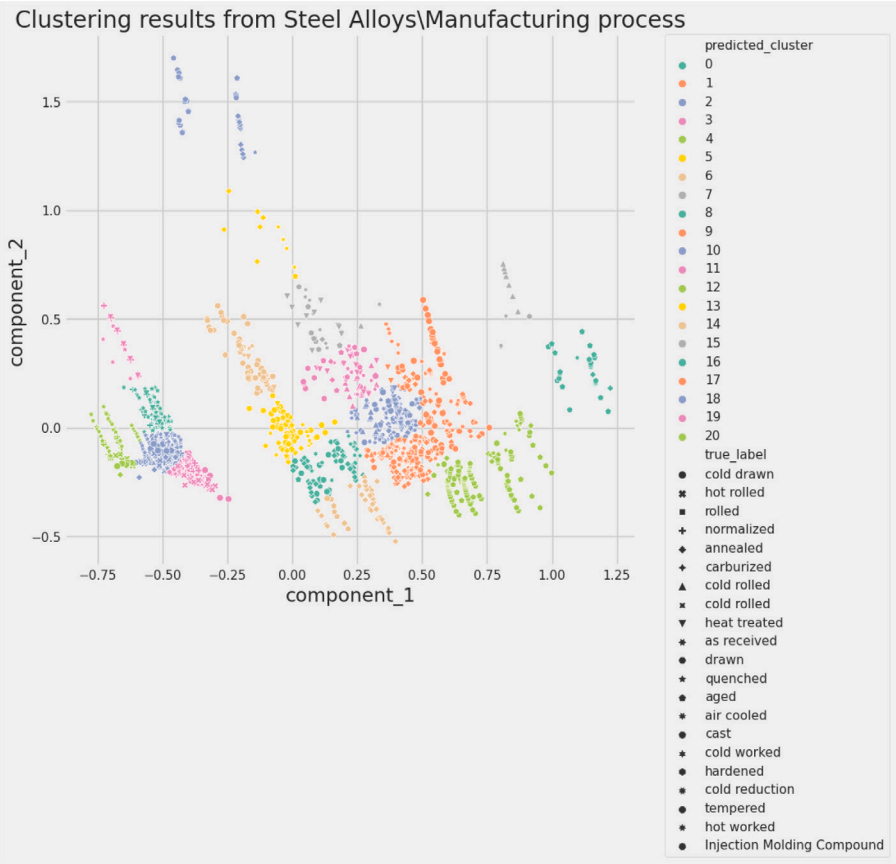


Fig. 12. Clustering results from Steel/Manufacturing process: cluster by manufacturing processing (shapes) and comparison with actual clustering determined by the algorithms (colors).

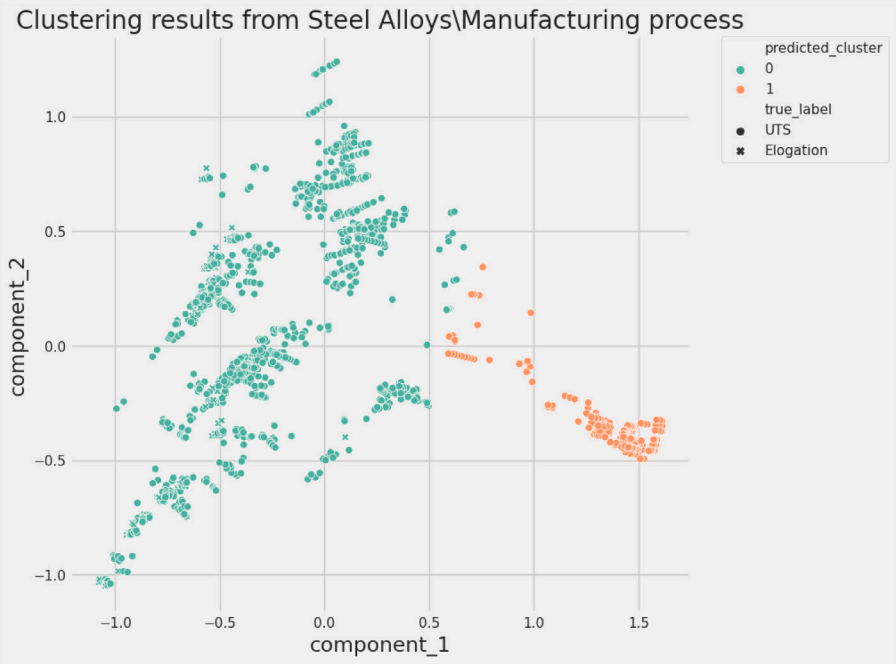


Fig. 13. Clustering results from Steel/Manufacturing process: cluster by Mechanical properties (UTS and Elongation) (shapes) with actual clustering determined by the algorithms (colors).

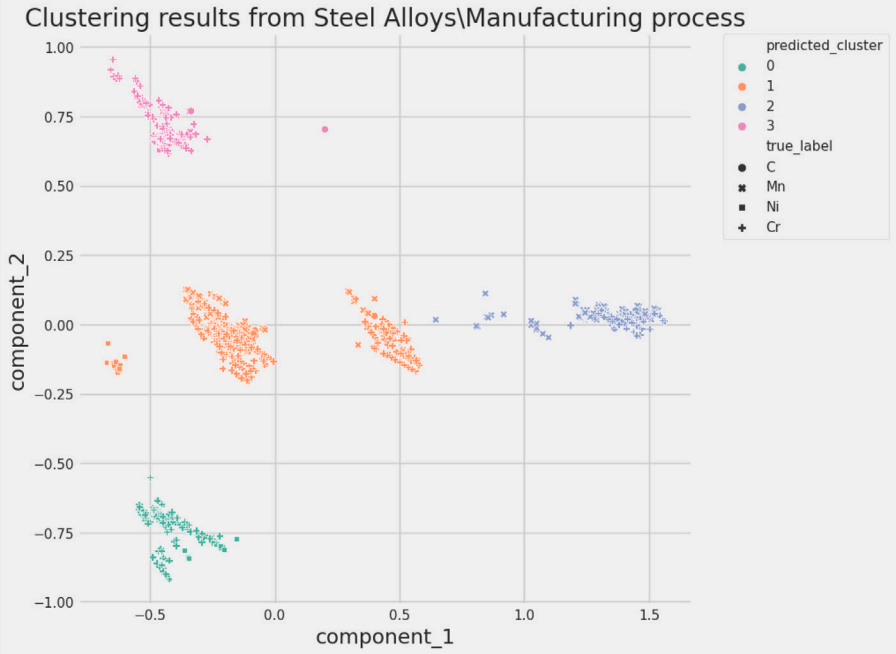


Fig. 14. Clustering results from Mechanical properties/Manufacturing process. Cluster by Chemical composition (C, Mn, Ni and Cr) compared with actual clustering determined by the algorithms.

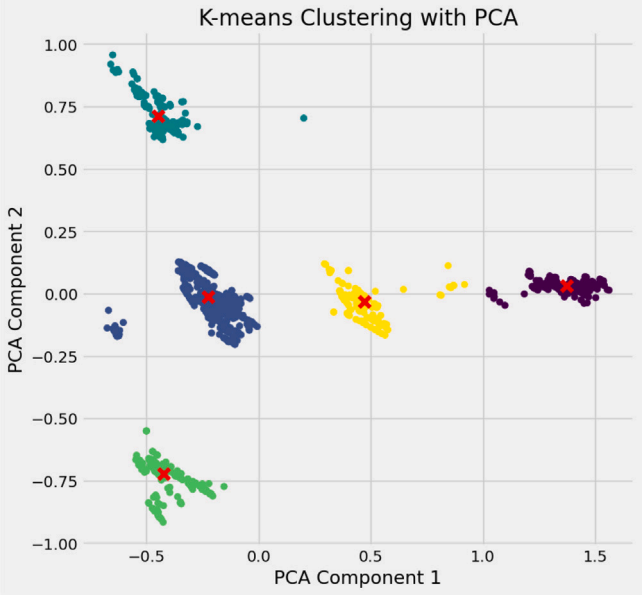


Fig. 15. Clustering results from Steel/Manufacturing process, five different clusters determined by the algorithms.

4.5.1. Super Duplex Stainless Steel absent from the training set

Cast Super Duplex Stainless Steel (SDSS) 25Cr-7Ni-Mo-N was chosen to validate the NN. This alloy was not part of the training set or inference set of the developed NN. Therefore, the best-performing networks, in this case, for yield, and ultimate tensile strength represented by the network containing 3 layers with a dropout of 0.25 were used. The chemical composition presented in Table 13, and the manufacturing process, in this case, casting, were set as input. Calculated values for the yield strength and tensile strength are displayed in Table 14 and compared with the actual values of $YS = 549$ MPa and $TS = 907$ MPa.

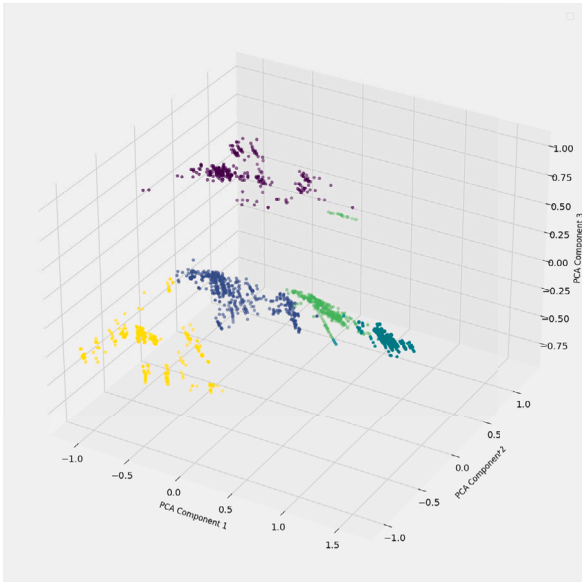
Table 7
Neural Network (NN) Architectures for Predicting Yield Strength across Various Clusters. This table presents a detailed NN model, each tailored to predict Ultimate Tensile Strength in specific clusters. The models are distinguished by their unique architecture, including the configuration of hidden layers and total number of neurons, the implementation of dropout strategies (specified by the number of layers and dropout rate), and the variance in K-folds during the training phase. Additionally, the use of One Hot Encoding for categorizing manufacturing process inputs is highlighted. The table also outlines the training/test split percentages, providing a comprehensive overview of each NN model's design and its targeted application for yield strength prediction in distinct clusters.

	NN structure	Dropout	K-fold	Training/test split (%)
Cluster 1	4/240	–	–	85/15
Cluster 2	4/120	–	4	–
Cluster 3	4/240	3/0.25	–	85/15
Cluster 4	4/120	3/0.25	–	80/20
Cluster 5	4/240	–	4	–

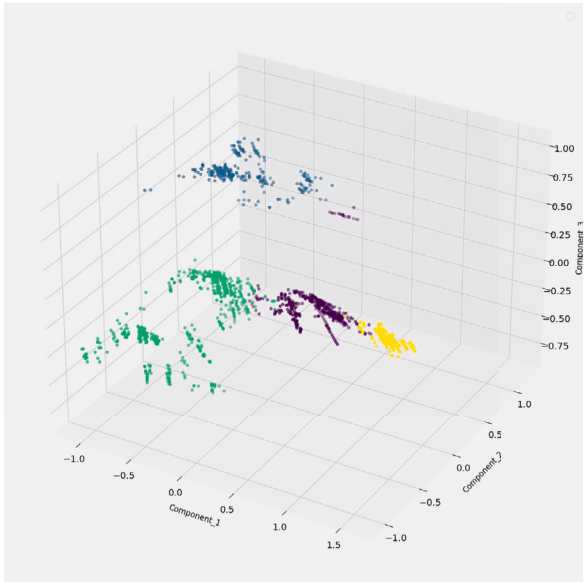
Table 8
Neural Network (NN) Architectures for Predicting Ultimate Tensile Strength across Various Clusters. This table presents a detailed NN model, each tailored to predict Ultimate Tensile Strength in specific clusters. The models are distinguished by their unique architecture, including the configuration of hidden layers and total number of neurons, the implementation of dropout strategies (specified by the number of layers and dropout rate), and the variance in K-folds during the training phase. Additionally, the use of One Hot Encoding for categorizing manufacturing process inputs is highlighted. The table also outlines the training/test split percentages, providing a comprehensive overview of each NN model's design and its targeted application for ultimate tensile strength prediction in distinct clusters.

	NN structure	Dropout	K-fold	Training/test split (%)
Cluster 1	4/240	3/0.25	–	85/15
Cluster 2	4/120	–	4	–
Cluster 3	4/240	–	–	85/15
Cluster 4	4/120	1/0.25	–	80/20
Cluster 5	4/240	–	4	–

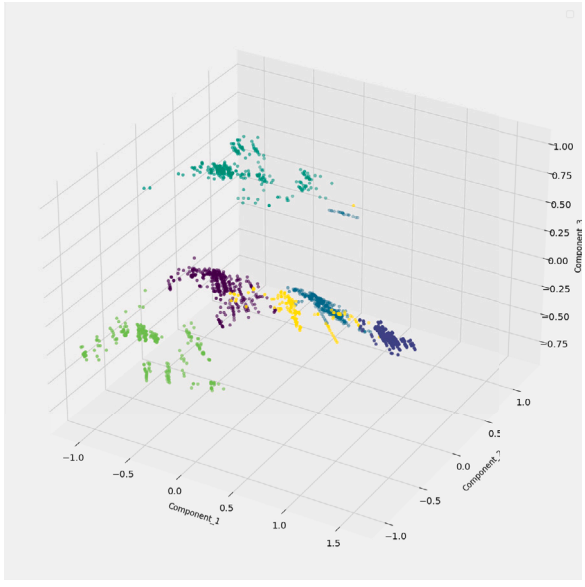
The actual values were experimentally determined in a tensile test, whose details are provided in Seabra and Costa [15]. The predictive model exhibits a reasonable accuracy.



(a) Five clusters



(b) Four clusters



(c) Six clusters

Fig. 16. 3D Clustering results from all information of process, chemical, and mechanical properties, clustering determined by the algorithms.

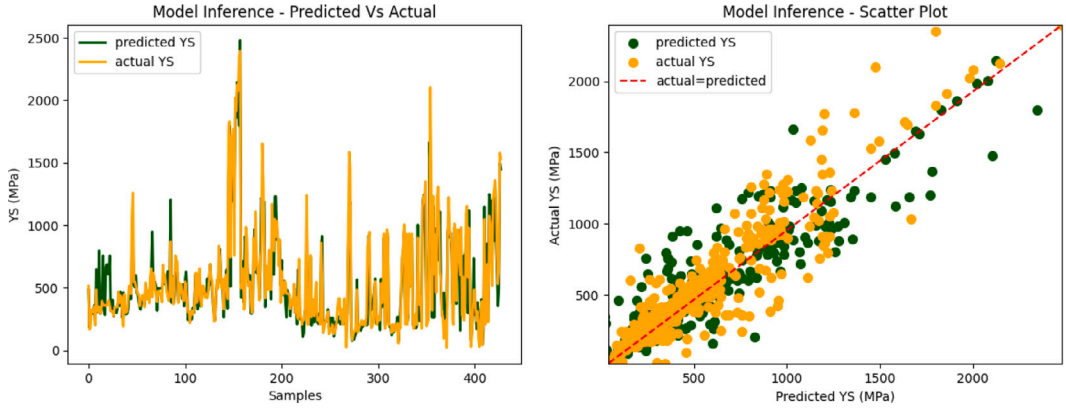


Fig. 17. Yield Strength results from 4/240 NN structure, dropout 3/0.25, second k-fold, with training loss of 29 MPa, RMSE: 158 MPa, MAE: 98 MPa.

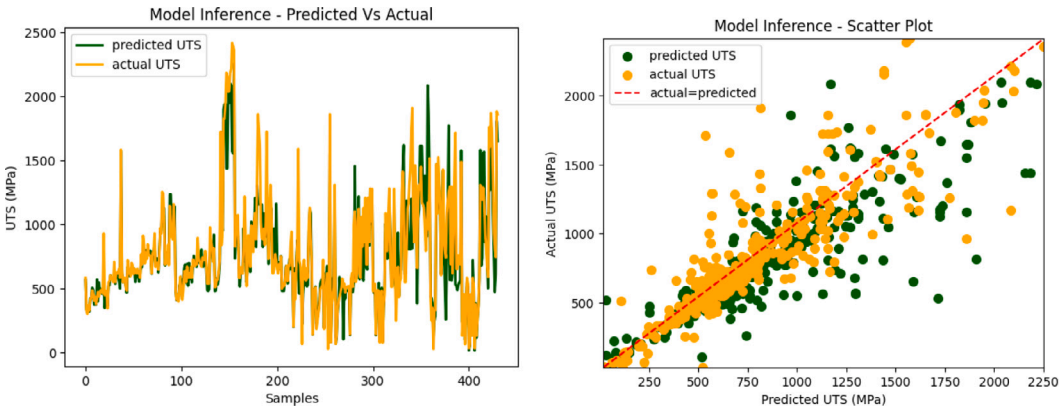


Fig. 18. Ultimate Tensile Strength results from 4/240 NN structure, dropout 3/0.25, second k-fold, with training loss of 40 MPa, RMSE: 194 MPa, MAE: 98 MPa.

Table 9
Comparative Performance of Neural Network in Predicting Yield Strength: This table presents the results of a neural network model using One Hot Encoder for categorical encoding of manufacturing processes. The model's performance is evaluated across five distinct clusters based on training loss, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) Score.

	Training loss (MPa)	RSME (MPa)	MAE (MPa)	R2 Score
Cluster 1	18	102	62	0.64
Cluster 2	43	240	150	0.70
Cluster 3	49	151	107	0.7
Cluster 4	32	65	33	0.93
Cluster 5	16	191	113	0.88

Table 10
Comparative Performance of Neural Network in Predicting Ultimate Tensile Strength: This table presents the results of a neural network model using One Hot Encoder for categorical encoding of manufacturing processes. The model's performance is evaluated across five distinct clusters based on training loss, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) Score.

	Training loss (MPa)	RSME (MPa)	MAE (MPa)	R2 Score
Cluster 1	23	104	58	0.62
Cluster 2	43	240	147	0.68
Cluster 3	63	121	86	0.7
Cluster 4	29	52	31	0.94
Cluster 5	18	205	114	0.83

Table 11
Performance metrics of the NN for yield strength prediction. The NN, used in the ablation study, consists of three hidden layers with 240 neurons in total, a dropout rate of 0.25 across three layers, and utilizes 4-fold cross-validation. A one-hot encoder is employed for encoding manufacturing process data.

Input	Training loss (MPa)	RMSE (MPa)	MAE (MPa)	R2 Score
Chemical composition	132	247	161	0.58
Chemical composition + manufacturing process	29	179	101	0.78

Table 12
Performance metrics of the NN for ultimate tensile strength prediction. The NN, used in the ablation study, consists of three hidden layers with 240 neurons in total, a dropout rate of 0.25 across three layers, and utilizes 4-fold cross-validation. A one-hot encoder is employed for encoding manufacturing process data.

Input	Training loss (MPa)	RMSE (MPa)	MAE (MPa)	R2 Score
Chemical composition	143	272	177	0.51
Chemical composition + Manufacturing process	35	201	108	0.74

Table 13
Chemical composition of the 25Cr-7Ni-Mo-N SDSS (wt%).

C	Si	Mn	Ni	Mo	Cr	Cu	W	N
0.02	0.7	0.7	7.9	3.8	25.3	0.9	0.7	0.2

Table 14
Calculated values of the yield strength and the tensile strength of SDSS 25Cr-7Ni-Mo-N and relative error concerning the actual value, which was experimentally determined.

	Actual	Predicted	Relative errors %
Yield strength (MPa)	549	610.31	10
Tensile strength (MPa)	907	821.91	9.4

Table 15
Chemical composition of the TRIP steel (wt%).

C	Si	Mn	Al	P	S
0.18	0.53	1.95	1.46	0.02	0.005

4.5.2. TRIP steel absent from the training set

The Neural Network (NN) was also tested on Transformation-Induced Plasticity (TRIP) steel with the composition Fe-0.18C-0.53Si-1.95Mn-1.46Al. This material was not included in the original training or inference datasets of the NN. The best-performing networks, previously successful with Super Duplex Stainless Steel (SDSS), were selected for this analysis. These networks comprised 3 layers with a dropout rate of 0.25. The specific chemical composition, detailed in Table 15, along with the complex manufacturing processes, were provided as input data for the NN. The manufacturing processes involved varied heat treatments for different sheet samples, as described in the work of [39] and subsequently. Predicted values for the yield strength and tensile strength of the TRIP steel are presented in Table 16, alongside their comparison with the experimental values.

The TRIP steel sheets labeled No. 1 to No. 8, underwent distinct heat treatment processes. Sheets No. 1, No. 2, and No. 3 were quenched to varying temperatures and then isothermally partitioned at 400 °C for different durations before air-cooling. Specifically, No. 1 was treated at 935–402–400 °C for 1 min, No. 2 at 907–447–400 °C for 2 min, and No. 3 at 903–440–400 °C for 20 min. Sheet No. 4 was quenched and cooled in an asbestine pit, while No. 5 underwent air-cooling post-quenching. Sheets No. 6 and No. 7 experienced isothermal partitioning at 400 °C after quenching to different temperatures, followed by air-cooling. No. 6 was treated at 913–520–400 °C for 5 min, and No. 7 at

Table 16
Predicted and actual Yield Strength of TRIP steel after various heat treatments, and relative error concerning the actual value.

Sheets	Actual YS (MPa)	Predicted YS (MPa)	Relative errors %
No. 1	1000	950	5
No. 2	790	717	9.2
No. 3	900	886	1.5
No. 4	825	783	5.1
No. 5	600	523	12.8
No. 6	605	595	1.65
No. 7	890	768	13.7
No. 8	900	885	1.67

Table 17
Predicted and actual Ultimate tensile strength of TRIP steel after various heat treatments, and relative error concerning the actual value.

Sheets	Actual UTS (MPa)	Predicted UTS (MPa)	Relative errors %
No. 1	1200	1320	10
No. 2	990	985	0.5
No. 3	1050	1013	3.5
No. 4	1100	976	11.3
No. 5	950	1065	12.1
No. 6	1050	1179	12.3
No. 7	1150	1174	2
No. 8	1350	1339	0.8

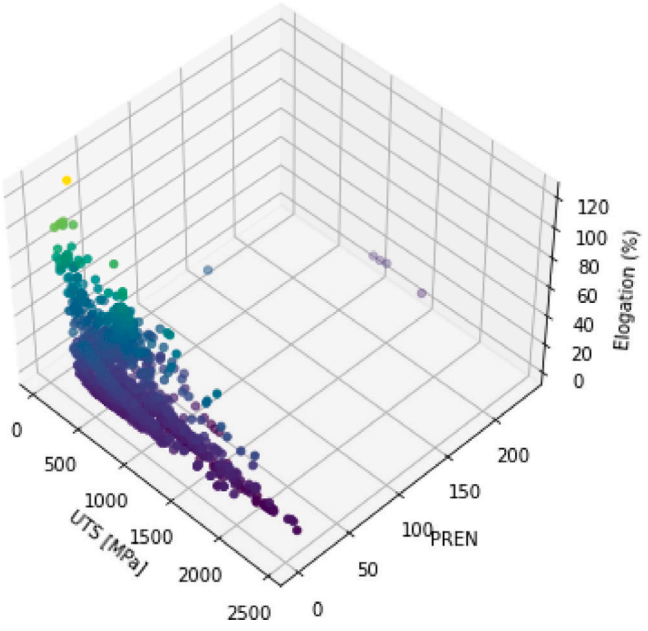


Fig. 19. An analysis is provided by the graph on elongation, ultimate tensile strength (UTS), and pitting resistance equivalent number (PREN).

894–294–400 °C for 5 min. Lastly, sheet No. 8 was directly quenched from 901 °C to room temperature.

The predictive capabilities of the NN for these varying heat treatment processes and their impact on the mechanical properties of TRIP steel were then assessed. Predicted values for yield strength and tensile strength, compared with the experimentally determined values, are displayed in Tables 16 and 17.

The observed variations in relative error across different processing methods could be attributed to the nature of the dataset. As indicated by cluster analysis, certain manufacturing processes and steel grades are better represented in the database than others, and thus yield richer and more informative data compared to others. This study illustrates the NN proficiency in predicting the mechanical properties of TRIP steel subjected to various heat treatments, demonstrating its potential as a valuable tool in materials science and engineering.

4.5.3. Designing a new steel alloy

A novel alloy of the duplex family is proposed, aimed at optimizing the balance between Yield and Tensile Strength, elongation, and PREN. Consequently, the NN model was extended in order to predict elongation. It was constructed utilizing the same architectural parameters from the best results for Yield and Tensile Strength. Hence, an NN structure with 4 hidden layers and 240 neurons, and 4 k-fold cross-validation, were used to build the elongation predicted model. Results are shown in Table 18.

In addition to this, the PREN factor is also taken into account within the optimization problem framework. As a result, the equation used for calculation is as follows:

$$PREN = \%Cr + 3.3(\%Mo + 0.5\%W) + 16\%N \quad (17)$$

A preliminary examination of elongation, ultimate tensile strength (UTS), and pitting resistance equivalent number (PREN) has already been conducted. To better understand their behavior, Graphs have been generated for these variables, facilitating the observation of their interactions (see Fig. 19).

A new combination of elements of alloys is proposed to consider the same casting procedure as in the initial case, which was subjected to testing in order to identify potentially promising alloys.

Following some compositional adjustments, a novel combination of element proportions was successfully attained. This new composition holds the potential for favorable values in ultimate tensile strength (UTS), yield strength (YS), and pitting resistance equivalent number (PREN). Detailed results are presented in Table 19.

Results from the predictive models and PREN equation are displayed in Table 20.

An improvement in the performance of the material is observed by adjusting the elements as Mo, and Cr, as confirmed by studies like those by Chail and Kangas [40] and Francis and Byrne [41]. However, these elements enhance the formation of the sigma phase, as Llorca-Isern et al. [42] presented in their work, and it may cause some prejudice in the mechanical properties if the thermal cycles are not well controlled. Therefore, although this alloy represents great potential to improve the performance of super duplex stainless steel, it should be further studied.

4.6. Procedures to enhance the accuracy of ML model

After identifying promising alloy candidates through a machine learning (ML) algorithm, a two-tiered approach is proposed to enhance the accuracy of ML predictions for alloy selection. The first phase, Preliminary Small-Scale Experimental Testing, focuses on collecting empirical data on material properties. This phase includes conducting hardness tests to quickly evaluate mechanical properties and performing Microstructure Observation (MO) to understand the material's microscopic features. Subsequently, the process advances to the second phase, Integration with Advanced Multiscale Modelling, as shown in work of [17,18]. The aim here is to refine ML predictions using the empirical data from the preliminary tests. This involves integrating the test results with multiscale modeling techniques, allowing for the simulation of alloy behavior under different conditions. The result of this integration significantly improves predictive accuracy by correlating empirical data with theoretical models. This approach not only enhances the reliability of predictions but also ensures that the alloy selection process is more aligned with real-world conditions and requirements.

Table 18
Evaluation of Neural Network (NN) Performance in Predicting Elongation: This table presents various NN models, each distinguished by its architecture, including the configuration of hidden layers and total neurons, as well as the application of dropout strategies (both the number and value of dropouts in hidden layers). Additionally, the impact of different K-fold values in the training process is explored. These factors collectively define the uniqueness of each NN model in predicting elongation, with the manufacturing processes encoded using One Hot Encoder.

NN structure	Dropout	K-fold	Training loss (MPa)	RMSE (MPa)	MAE (MPa)	R2 Score
3/240	–	4	0.66	8.8	5	0.65

Table 19
Chemical composition of the 32Cr-7Ni-Mo-N SDSS (wt%).

C	Si	Mn	Ni	Mo	Cr	W	N
0.02	0.7	0.7	7.9	7	32	0.7	0.2

Table 20
Calculated values of the yield strength and the tensile strength of SDSS 32Cr-7Ni-Mo-N and relative error.

	Predicted	Expect error
Yield strength (MPa)	1125	±101
Tensile strength (MPa)	1309	±108
Elongation (%)	19	±5
PREN	59	–

5. Conclusion

Steel industry features a wide range of manufacturing processes which may generate a variation in the grain size, structure, or cooling rate of material, and therefore they impact the final properties and performance of mechanical parts. This work aimed at creating an improved performance inferential tool that could accurately predict the final properties of materials based on their chemical composition and manufacturing process information, resorting to machine learning.

Manufacturing information generally is available in a qualitative form, contrasting with the quantitative form in which chemical composition and mechanical properties are available. Natural language processing, in particular one hot encoding, successfully allowed the integration of all data with significant advantages. Firstly, it allowed to evaluation of whether clustering data according to the manufacturing process was optimal or not, and secondly, it allowed the incorporation of the manufacturing information in an inferential tool based on NN.

In fact, analysis of the data confirmed that modifying the manufacturing process can greatly impact the mechanical properties. Furthermore, through k-means clustering, it was shown that the best way to correlate the data is more complex than simply choosing clusters based on the manufacturing process, prevalent element in chemical composition, or prevalent mechanical properties, as previously done in the literature.

At the inference level, it was possible to attest to the superiority of one hot encoding in relation to label encoding in predicting Yield Strength and Ultimate Tensile strength. Moreover, the ablation study showed a significant improvement in results with the inclusion of the manufacturing information versus its absence. Additionally, clusters discovered using k-means applied to the predictive NN model helped optimize the NN structure and enhanced its performance when compared to a single model for the entire dataset. Nevertheless, performance was poorer in the smaller clusters, which indicates that the database should be enlarged with alloys belonging to those clusters. In terms of overfitting prevention, the outcomes for the entire dataset show reasonable values for R2, MSE, and MAE, before and after the regularization of data through the dropout function. K-fold cross-validation yields better results with a higher number of k-folds.

The accuracy of the model and its potential for alloy discovery was tested in three case studies. In the first and second cases reasonable to very good accuracy was obtained against experimental results. Such results together with the clustering results suggest that lower performance cases were due to under underrepresentativity of certain manufacturing processes and steel grades in the database and thus, the database should be complemented as future work. In the third case, a new alloy of the Super Duplex Stainless Steel family was developed, which, nevertheless, requires further study in terms of possible precipitation of secondary phases during manufacturing.

On the downside, predictive errors could reach 10% which is significant from an industrial standpoint. However, the main goal of the inferential tool developed is to spot the most promising candidates for a certain application. Results from the inferential tool may be further combined with Finite Element and multiscale analysis to narrow down the selection. In addition, small-scale experiments such as hardness and microstructure observation (MO), which are quicker and less resource-intensive, can be conducted in small samples of the best candidates. Notwithstanding, some further strategies are also undersight as future work, namely, convolutional neural networks, integration with CALPHAD, and the development of a user-friendly graphical interface that streamlines the discovery of novel steel alloys.

CRediT authorship contribution statement

Ana P.O. Costa: Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mariana R.R. Seabra:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **José M.A. César de Sá:** Writing – review & editing, Supervision. **Abel D. Santos:** Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ana P.O. Costa reports financial support was provided by Foundation for Science and Technology. Mariana R.R. Seabra reports financial support was provided by University of Lisbon Center for Philosophy of Sciences. Ana P.O. Costa reports a relationship with Foundation for Science and Technology that includes: funding grants. Mariana R.R. Seabra reports a relationship with University of Lisbon Center for Philosophy of Sciences that includes: funding grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data are available at https://doi.org/10.17632/ysndxjs9gp.1.

Acknowledgments

The first author gratefully acknowledges the funding of the doctoral grant 2021.05067.BD by Fundação para a Ciência e Tecnologia, through the national budget and Community budget from the European Social Fund (ESF). The second author gratefully acknowledges the financial support from the Lancog Group of the Centro de Filosofia da Universidade de Lisboa (CFUL), under the project UIDP/00310/2020.

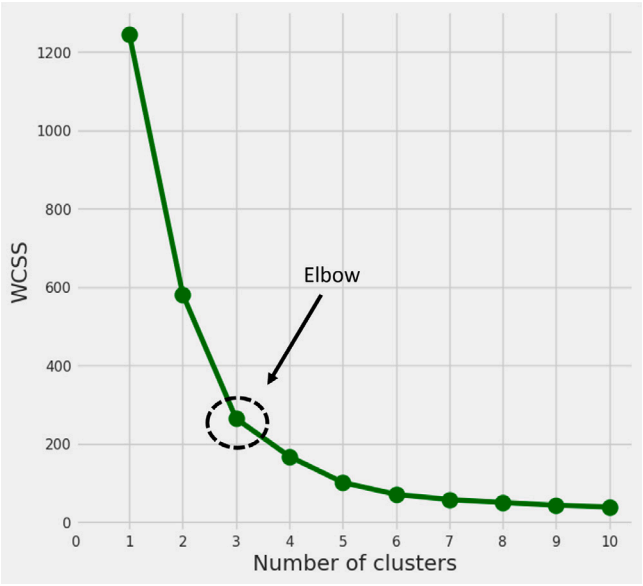


Fig. 20. Summary k-fold crossing validation [29].

Appendix

Selection of the number of clusters in K-means clustering

There are two main techniques that can be used to determine the number of clusters: the elbow technique and the silhouette method. These techniques were reviewed by Kodinariya et al. [43]. The elbow method involves plotting the sum of squared distances (WCSS) between each data point and its assigned centroid as a function of K and identifying the point where the curve starts to level out, Fig. 20 provides an example of how to determine the number of clusters by the elbow method. The silhouette method assigns a silhouette score to each data point, in order to compare how similar it is to other clusters in comparison to its own cluster. For each K value, the overall silhouette score is calculated and the K value with the highest score is selected. In this study, we have opted to use the elbow method.

References

[1] J.L. Cann, A. De Luca, D.C. Dunand, D. Dye, D.B. Miracle, H.S. Oh, E.A. Olivetti, T.M. Pollock, W.J. Poole, R. Yang, et al., Sustainability through alloy design: Challenges and opportunities, Prog. Mater. Sci. 117 (2021) 100722.
[2] World Steel in Figures 2019, World Steel Association, 2019, pp. 3–30.
[3] S.S. Sidhu, H. Singh, M.A.-H. Gepreel, A review on alloy design, biological response, and strengthening of β-titanium alloys as biomaterials, Mater. Sci. Eng. C 121 (2021) 111661.
[4] N. Yurchenko, E. Panina, S. Zherebtsov, N. Stepanov, Design and characterization of eutectic refractory high entropy alloys, Materialia 16 (2021) 101057.
[5] H. Li, A. Wang, T. Liu, P. Chen, A. He, Q. Li, J. Luan, C.-T. Liu, Design of Fe-based nanocrystalline alloys with superior magnetization and manufacturability, Mater. Today 42 (2021) 49–56.
[6] Y. Juan, Y. Dai, Y. Yang, J. Zhang, Accelerating materials discovery using machine learning, J. Mater. Sci. Technol. 79 (2021) 178–190.
[7] S.K. Kauwe, J. Graser, R. Murdock, T.D. Sparks, Can machine learning find extraordinary materials? Comput. Mater. Sci. 174 (2020) 109498.
[8] C.M. Bishop, N.M. Nasrabadi, Pattern Recognition and Machine Learning, Vol. 4, Springer, 2006.
[9] E. Alpaydin, Machine Learning, MIT Press, 2021.
[10] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284, http://dx.doi.org/10.1109/TKDE.2008.239.
[11] A. Mozaffari, M. Emami, A. Fathi, A comprehensive investigation into the performance, robustness, scalability and convergence of chaos-enhanced evolutionary algorithms with boundary constraints, Artif. Intell. Rev. 52 (2019) 2319–2380.
[12] D. Merayo, A. Rodríguez-Prieto, A.M. Camacho, Prediction of mechanical properties by artificial neural networks to characterize the plastic behavior of aluminum alloys, Materials 13 (22) (2020) 5227.
[13] X. Xu, L. Wang, G. Zhu, X. Zeng, Predicting tensile properties of AZ31 magnesium alloys by machine learning, Jom 72 (2020) 3935–3942.
[14] B. Conduit, N.G. Jones, H.J. Stone, G.J. Conduit, Design of a nickel-base superalloy using a neural network, Mater. Des. 131 (2017) 358–365.
[15] M. Seabra, A. Costa, Material model calibration using machine learning: a comparative study, Eur. J. Comput. Mech. (2022).
[16] K. Chowdhary, Fundamentals of Artificial Intelligence, Springer, 2020.
[17] A. Costa, R. Sousa, L. Ribeiro, A. Santos, J.C. de Sá, Multiscale modeling for residual stresses analysis of a cast super duplex stainless steel, Mater. Des. Appl. III (2021) 47–63.
[18] A.P. Costa, M.R. Seabra, A.D. Santos, L.M. Ribeiro, J.M.C. de Sá, Experimental and numerical multiscale characterization of a super duplex stainless steel 25Cr-7Ni-Mo-N, Mater. Today Commun. 33 (2022) 104903.
[19] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (6) (1933) 417.
[20] K. Pearson, LIII. On lines and planes of closest fit to systems of points in space, Lond. Edinb. Dublin Philos. Mag. J. Sci. 2 (11) (1901) 559–572.
[21] L.I. Smith, A tutorial on principal components analysis, 2002.
[22] R. Vidal, Y. Ma, S. Sastry, Generalized principal component analysis, IEEE Trans. Pattern Anal. Mach. Intell. 27 (12) (2005) 1–15.
[23] A. Quarteroni, G. Rozza, et al., Reduced Order Methods for Modeling and Computational Reduction, Vol. 9, Springer, 2014.
[24] D. Arthur, S. Vassilvitskii, k-means++: The Advantages of Careful Seeding, Technical Report, Stanford, 2006.
[25] M. Wenzlick, O. Mamun, R. Devanathan, K. Rose, J. Hawk, Data science techniques, assumptions, and challenges in alloy clustering and property prediction, J. Mater. Eng. Perform. 30 (2021) 823–838.
[26] T. Szandała, Review and comparison of commonly used activation functions for deep neural networks, in: Bio-Inspired Neurocomputing, Springer, 2021, pp. 203–224.
[27] R. Rojas, The backpropagation algorithm, in: Neural Networks, Springer, 1996, pp. 149–182.
[28] D. Berrar, Cross-validation, 2019.
[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[30] A. El Korchi, Y. Ghanou, DropWeak: A novel regularization method of neural networks, Procedia Comput. Sci. 127 (2018) 102–108.
[31] N. Srivastava, Improving Neural Networks with Dropout, Vol. 182, University of Toronto, 2013, p. 7.
[32] A. Rai, S. Borah, Study of various methods for tokenization, in: Applications of Internet of Things: Proceedings of ICCCIOT 2020, Springer, 2021, pp. 193–200.
[33] H. Ji, Information Extraction, Springer US, Boston, MA, 2009, pp. 1476–1481, http://dx.doi.org/10.1007/978-0-387-39940-9_204.
[34] W. Chen, Learning with Scalability and Compactness, Washington University in St. Louis, 2016.
[35] R. Zhao, K. Mao, Fuzzy bag-of-words model for document representation, IEEE Trans. Fuzzy Syst. 26 (2) (2017) 794–804.
[36] K.W. Church, Word2Vec, Nat. Lang. Eng. 23 (1) (2017) 155–162.
[37] M. LLC, MatWeb: Online materials information resource, 2022, http://www.matweb.com.
[38] M. Bauccio, et al., ASM Metals Reference Book, ASM international, 1993.
[39] X. Tan, H. He, W. Lu, L. Yang, B. Tang, J. Yan, Y. Xu, D. Wu, Effect of matrix structures on TRIP effect and mechanical properties of low-C low-Si Al-added hot-rolled TRIP steels, Mater. Sci. Eng. A 771 (2020) 138629.
[40] G. Chail, P. Kangas, Super and hyper duplex stainless steels: structures, properties and applications, Procedia Struct. Integr. 2 (2016) 1755–1762.
[41] R. Francis, G. Byrne, Duplex stainless steels—alloys for the 21st century, Metals 11 (5) (2021) 836.
[42] N. Llorca-Isern, H. López-Luque, I. López-Jiménez, M.V. Biezma, Identification of sigma and chi phases in duplex stainless steels, Mater. Charact. 112 (2016) 20–29.
[43] T.M. Kodinariya, P.R. Makwana, et al., Review on determining number of Cluster in K-Means Clustering, Int. J. 1 (6) (2013) 90–95.