Open Ceramics 15 (2023) 100383

Contents lists available at ScienceDirect

Open Ceramics

journal homepage: www.sciencedirect.com/journal/open-ceramics

Investigation of process influences on the amount of single-fiber siliconization in C/C–SiC samples by machine-learning methods

Tobias Lehnert [a,*], Bernhard Heidenreich [a], Dietmar Koch [b]

[a] Institute of Structures and Design, German Aerospace Center Stuttgart, Pfaffenwaldring 38-40, 70569, Stuttgart, Germany
[b] Institute for Materials Resource Management, University of Augsburg, Am Technologiezentrum 8, 86159, Augsburg, Germany

ARTICLE INFO

Handling Editor: Dr P Colombo

Keywords:
Ceramic matrix composites
Machine-learning
Artificial intelligence
Process optimization

ABSTRACT

Ceramic Matrix Composites are an interesting option for high-temperature combustive environments as often encountered in aerospace applications. In the past a lot of research was conducted in order to find the right process parameters for optimal performance of these materials. The mechanical properties of CMCs are vastly dependent on their microstructure. Therefore, a lot of past research focused on finding correlations between process parameters and microstructure of CMCs, most of which was based on empirical trial and error methods.

In this paper we use several data-driven, probabilistic machine-learning models to quantify the microstructural composition of C/C–SiC based on the process parameters and choice of raw materials. As a ground truth 123 samples of C/C–SiC with varying process parameters and microstructures were used. The predictive capabilities of the models were demonstrated by the use of the R^2 metric. By this analysis density in siliconized state as well as open porosity and mass change during siliconization proved to be the parameters with the highest impact on microstructural formation. If siliconization was taken out of the equation the porosity in CFRP state and fiber type were found to be the most influential factors.

1. Introduction

CMCs have gained a lot of relevance over the last decades in many application fields where high operating temperatures have to be expected [1]. Due to their complex nature including the vastly inhomogeneous microstructure which involves fibers, pores and other inclusions, processing and machining of CMCs until today poses a big challenge. The high complexity of the processing and machining steps leads to a high variance in important mechanical properties of the resulting components compared to their metallic counterparts. A lot of empirical research has been conducted in order to improve quality and reproducibility of CMC components and to identify the reasons for the dispersion. In Naskar et al. [2] the mechanical properties such as flexural strength and ductility of oxide CMCs were found to be strongly dependent on the viscosity of the infiltrates, the number of infiltrations and the sintering temperatures. Friess et al. [3] investigated the influence of different precursors, C-fibers and annealing process parameters on the thermophysical properties of C/C–SiC. The results suggested that the thermal conductivity as well as spectral emissivity were dominated by the fiber selection, whereas the specific heat capacity was influenced by fiber and matrix properties. Furthermore, annealing was found to have beneficial impact on the thermal conductivity if done in C/C state prior to siliconizing. Also, properties of components manufactured by novel methods such as 3D printing show a high dependency on process parameters. This is demonstrated in Zhu et al. [4] where mechanical performance and microstructure of carbon fiber reinforced silicium-carbide (C/SiC) parts are significantly affected by the 3D printing process parameters. Li et al. [5] as well as Krenkel [6] examined the effects of machining of CMCs on their surface integrity and microstructural quality. Here rotary ultrasonic machining was shown to yield high quality results. The importance of process parameters on mechanical properties is not only restricted to CMCs but includes most other material engineering fields. Moses et al. [7] use an empirical approach to predict the effects of stir casting parameters on the ultimate tensile strength of aluminum matrix composites.

Over the recent years computational methods have gained a lot of importance due to the steep rising of artificial intelligence (AI) and the progress in hardware resources [8]. A recent study investigated the correlation of several microstructural parameters and tensile strength of CMC samples by a neural network [9]. The underlying data was taken from already published scientific papers and included a high percentage of missingness. Ghayour et al. [10] used an AI modelling approach to

* Corresponding author.
E-mail address: tobias.lehnert@dlr.de (T. Lehnert).

https://doi.org/10.1016/j.oceram.2023.100383
Received 21 December 2022; Received in revised form 4 May 2023; Accepted 31 May 2023
Available online 1 June 2023

predict Vickers hardness of ceramic samples based on additive contents, sintering temperature, -time and -pressure. Aggour, Grupta et al. [11] showed that a type of deep neural network can be used to successfully characterize optical photomicrographs of CMC components.

Although the current increase of computational approaches to material design is not to be overlooked, the use in the world of CMCs is still limited. Existing studies mainly focus on other composite materials or have only small amounts of data to feed into their machine-learning algorithms. In this paper we use various machine-learning algorithms to describe the correlation of process parameters and resulting microstructure of C/C–SiC. A data pool of 123 CMC samples which include selected raw materials, process parameters and SEM images of constant magnification serves as a basis for the AI models.

2. Dataset and preprocessing

As a ground truth a dataset consisting of $n = 123$ C/C–SiC samples was used which were manufactured and documented at German Aerospace Center (DLR) in Stuttgart (Germany) via the liquid silicon infiltration (LSI) process. The associated processing steps are shown in Fig. 1.

Fig. 2 gives an overview over the most common fiber/matrix combinations in the data pool which will be relevant for later discussion. For every sample at least one representative SEM image with magnification of 100× was present from which the phase area shares and therefore the carbon conversion ratio (CCR) was determined, which is defined below. For every sample 29 columns were created which featured all the documented properties such as the type of resin and fiber used, porosity and density after every production step, mass changes, processing temperatures, -pressures and -times, fiber volume content and many more.

As can be seen in Fig. 2Figure 2 the investigated fiber types mainly consisted of HTA and T800 fibers. A small number of samples also featured T1000 or YS90 fibers. The most common precursors used were JK60, MF43, XP60, MF13 and MF88 which made up more than 95% of the dataset. Additionally, some water-based precursors such as PF7554 and PF0433 were present. The samples were prepared via different manufacturing processes which included autoclave, RTM (resin transfer molding), hot-press and winding. Furthermore, the processing parameters for the same manufacturing route also differed amongst the samples; for example some RTM processes were run with different temperatures than others. Due to the wide range of different raw materials and manufacturing parameters, very different C/C–SiC materials can be produced, which differ greatly in their microstructure and properties.

2.1. Carbon conversion ratio (CCR)

In order to train a probabilistic model on the data, one or more dependent variables had to be selected which quantify the microstructure corresponding to the process parameters. In this case the CCR was chosen which is defined as the percentage of carbon in an SEM image which gets converted into silicon carbide during siliconization as described in *Eq. (1)*. As CCR directly correlates with the carbon (C) content in the image, the latter could also have been chosen as dependent variable. Nevertheless, CCR was slightly preferred because it combines information of carbon as well as silicon-carbide (SiC) contents.

$$CCR = \frac{A_{SiC} \bullet K}{A_C + A_{SiC} \bullet K}$$

$$with\ K = \frac{V_C}{V_{SiC}} = \frac{6,53 \frac{cm^3}{mol}}{12,45 \frac{cm^3}{mol}} = 0,52 \qquad \text{Eq (1)}$$

Where A_{SiC} and A_C describe the sizes of SiC- and C-areas in the SEM image and V_{SiC} and V_C describe the molar volumes of SiC and C.

The microstructure of C/C–SiC can show various degrees of single fiber siliconization (SFS), as depicted in Fig. 3. Composites with low amounts of SFS exhibit greater damage tolerance and quasi-ductile failure behavior whereas high amounts of SFS tend to embrittle the material [13]. The CCR now quantifies the images in a way that the higher the CCR value is the more SFS is present and vice-versa.

Since most applications at DLR aimed for C/C–SiC components with a low amount of single-fiber-siliconization (corresponding to XB-structures and low CCR) the dataset is imbalanced [13,14]. Fig. 4 shows a histogram of the CCR-distribution with a bin-size of 5 throughout the dataset. A Shapiro-Wilk test [15] yielded a p-value of $p = 2 \bullet 10^{-8}$ which means that the data is not normally distributed as also easily observable in the histogram. From the histogram it is obvious that roughly ¾ of the samples have a CCR below 20.

For samples where more than one image was available, mean and standard deviation were calculated for the CCR. This applied to 65% of the dataset. The mean relative standard deviation of these samples was 11.6%, which shows that the CCR is relatively constant over different SEM-images of the same sample. Hence the CCR can be regarded as a robust measure for capturing the amount of SFS in a sample.

2.2. Encoding

Machine-learning models cannot work with alphabetical data such as

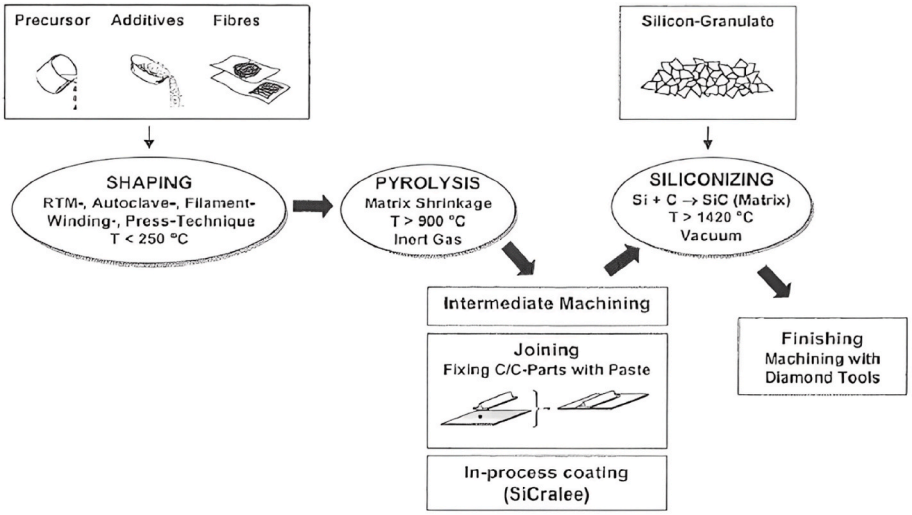

Fig. 1. Process flow chart for production od C/C–SiC via Liquid Silicon Infiltration (LSI) [12].

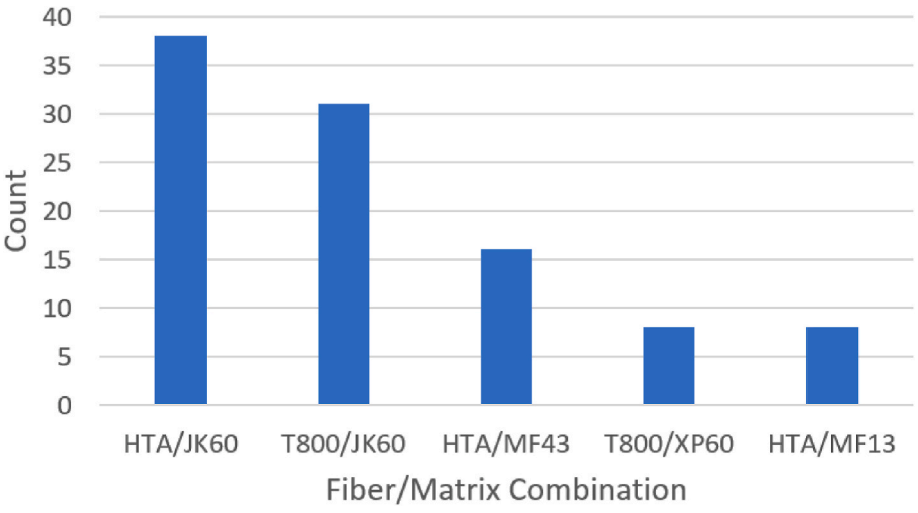

Fig. 2. Investigated fiber/matrix combinations; only combinations with more than 5 members are shown.

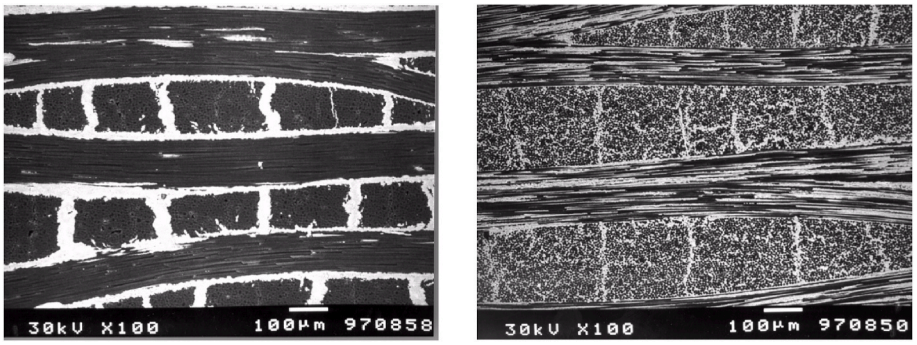

Fig. 3. Images of XB-microstructure (left) which relates to lower CCR values and XD-microstructure (right) which relates to higher CCR values [13].

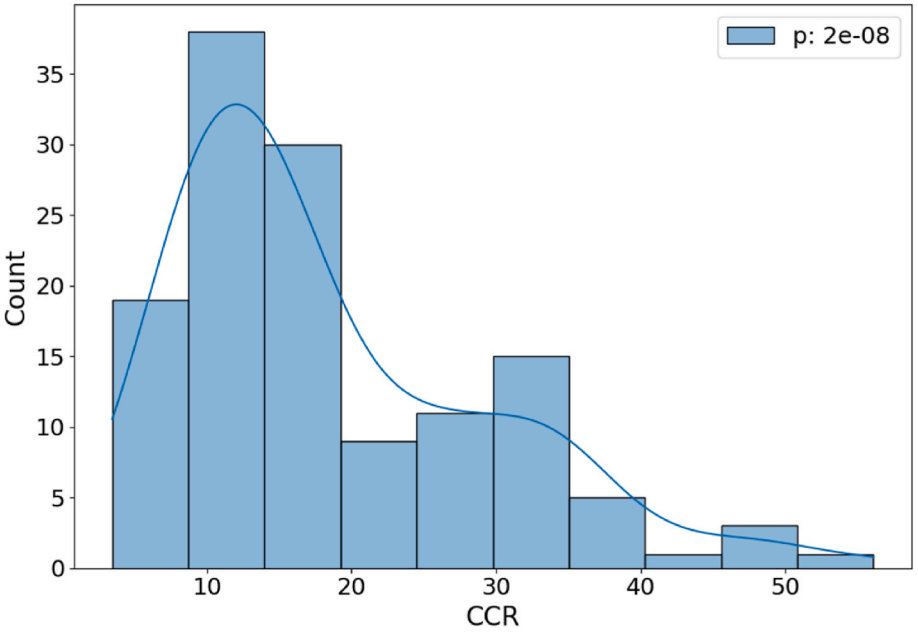

Fig. 4. Histogram of CCR-distribution throughout the dataset with a p-value of p«0.05 and a bin size of 5.

resin or fiber names which are designated as categorical variables. Since a lot of properties in the dataset contained such non-numerical data, all categorical variables were encoded by one-hot-encoding as described in Ref. [16].

2.3. Randomness in data splitting

During the splitting process as well as during the model training and hyperparameter selection there is always randomness involved. For example, each 80:20 split of the data set leads to a different outcome if the samples are randomly drawn. In order to make different algorithms and preprocessing steps more comparable to one another the same random seeds were used when comparing two different methods to eliminate randomness effects due to 'easy' or 'difficult' datasets.

2.4. Imputation

The aforementioned dataset contained a lot of missing values, which poses a challenge for machine-learning algorithms. In literature a variety of different methods are suggested to deal with such cases, the simplest of which would be to drop the missing data. An alternative, less wasteful method is to use imputation techniques to make educated guesses about the missing data. But in order to do so it is necessary to determine the reason for its missingness. In van Buuren [17] missingness is categorized in three different groups: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In case of this study the data is classified as MAR because although not all measurements were always made for every sample of the dataset (randomly distributed) some measurements were generally performed less frequently than others because of their higher expense which provides some kind of pattern. For missing data of type MCAR and MAR multiple imputation methodology can be used [17]. Since dropping data with missing entries would have left over only a small fraction of the dataset, two different imputation methods were chosen and compared: mean univariate imputation as well as iterative multivariate imputation. Mean imputation is as simple as filling the missing values by the mean of each respective column. The iterative imputation is more complex and calculates a similarity between each sample by using the k nearest features for which k has to be determined. Nearness is determined by calculating the absolute correlation coefficient between each parameter pair. Then the missing values are imputed by a weighted estimate based on sample similarity and iteratively re-calculating the similarities. The underlying estimator used was Bayesian Ridge regression [18]. A comparison for mean imputation and iterative imputation for an identical random seed is shown in Fig. 5. In both cases the nearest 5 features were used to impute the missing values by the ridge regression model. As an example, the density distribution in siliconized state was chosen, where blue markers denote measurements and red markers denote imputed values. It can be observed that iterative imputation provides more realistic outcomes for missing values and leads to better model accuracy as discussed later.

2.5. Splitting the data

The dataset was split into training- and testing sets by a common ratio of 80:20. Because the CCR distribution is imbalanced and heavily shifted to lower CCR values, stratified splitting was investigated in order to preserve class proportions as suggested in Farias et al. [19] and then compared to random splitting. In this case "classes" refer to bins of CCR-values which were created in 5% steps from CCR = 0% to CCR = 40% and one bin for CCR = 40% to the maximum CCR. A histogram of splitting with versus splitting without stratification is shown in Fig. 6 where blue bins denote training data and red bins denote testing data. Stratification was found to be beneficial for model accuracy as discussed later.

2.6. Cross validation

After splitting the data into training- and testing sets, the training data was further split by the same 80/20 ratio during 5-fold cross validation (CV), leaving 20% of the samples for validation in each run. During CV the optimal hyperparameters for each respective model were chosen. In the end the optimal model was evaluated by the testing set. The whole procedure is outlined in Fig. 7. After evaluation the model with the best parameters was trained again on the whole training dataset.

3. Developed models

Four different algorithms were trained and compared against each other on the dataset which was summed up in Table 1. The mean squared error was used as regression criteria for all algorithms.

3.1. Accuracy measure

All models were trained to minimize the coefficient of determination R^2 which can take values between $(-\infty, 1]$ and describes a measure for the predictive capabilities of the model in regression tasks. R^2 was chosen because it is more informative than statistical rates like RMSE, MAE or MSE [20]. The coefficient of determination states how much percent of the variance of the dependent variable can be explained by the independent variables and is calculated by Eq. (2). Here S_d denotes the declared scatter, S_r the residual scatter and S_{tot} the total scatter.

$$R^2 = \frac{S_d^2}{S_{tot}^2} = 1 - \frac{S_r^2}{S_{tot}^2} = 1 - \frac{\sum_{i=1}^{m} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{m} (\bar{y} - y_i)^2} \quad \text{Eq (2)}$$

There are three different cases:

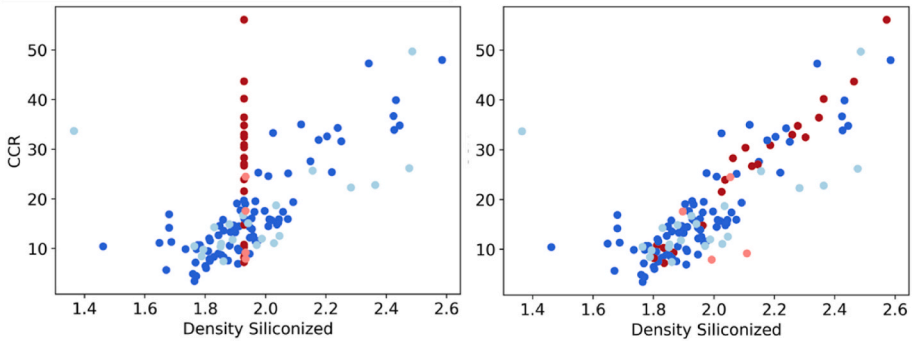

Fig. 5. Comparison of mean imputation (left) and iterative imputation (right) for the same random seed; dark blue: measurement (training set), light blue: measurement (testing set), dark red: imputed (training set), light red: imputed (testing set). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

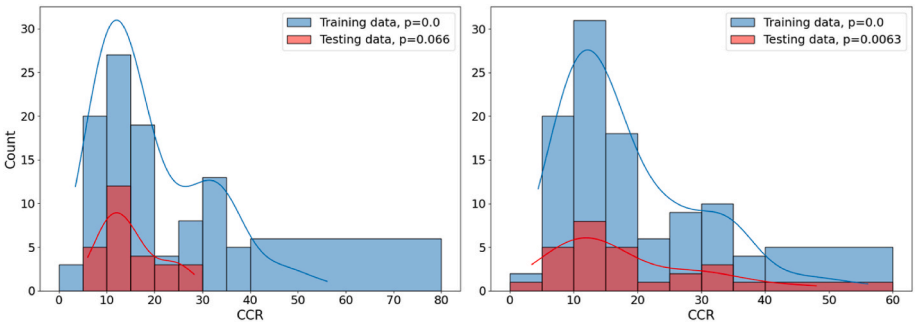

Fig. 6. Comparison of CCR-histograms for a 80%/20% Train/Test-split of the data using random sampling (left) and stratified sampling (right).

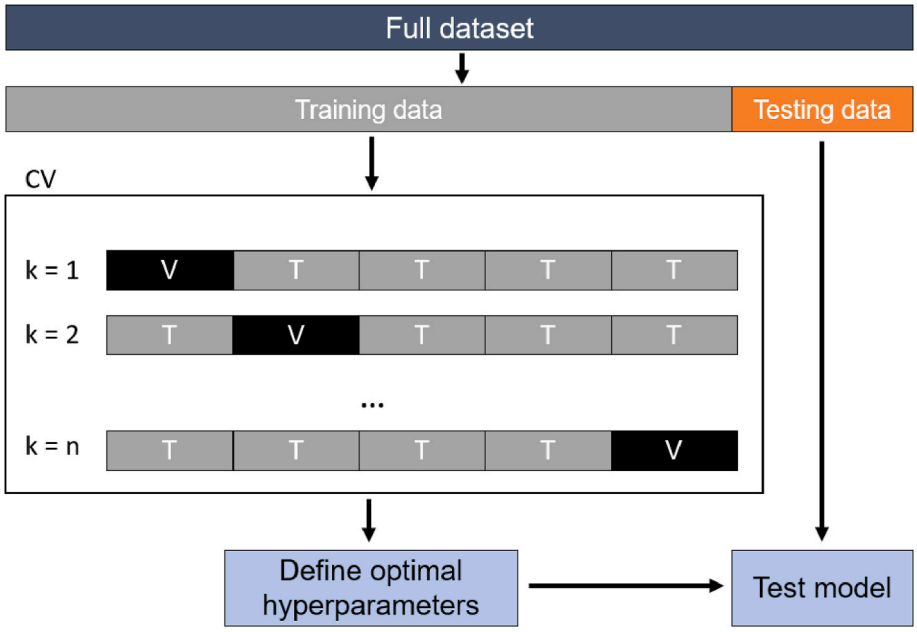

Fig. 7. Procedure of finding optimal hyperparameters and determining model accuracy; CV: T denotes training data, V denotes validation data.

Table 1
Selection of models which were used for predictions on the dataset.

Model Type	Model Algorithm
Decision Tree	Optimized CART
Random Forest	Optimized CART
Lasso Regression	Modified linear regression
Artificial Neural Network	Multi-Layer Perceptron

- $1 > R^2 > 0$: In this case the model explains some percentage of the variance of the dependent variable around its mean. Higher values are always better.
- $R^2 = 0$: The model explains none of the variance of the dependent variable around its mean
- $R^2 < 0$: The model performs more poorly than a horizontal line whereas the latter would be equivalent of always guessing the mean of the dependent variable

3.2. Decision tree

Despite their simplicity, decision trees (DT) are widely used in machine learning in geosciences and material engineering. This is mostly attributed to their easy interpretability compared to other machine-learning algorithms such as artificial neural networks as well as their low computational cost and the usability for classification as well as regression problems. There are several different algorithms to assemble a DT such as CART, C4.5 and CHAID which differ in the way they grow a tree. Generally, a DT contains a sequence of hierarchically organized conditions which are applied from the root node to the individual leaf nodes. The data is recursively split and each split evaluated for its purity by regression. As criterion for a split, either gini impurity or information gain are often times used for classification problems and mean squared error for regression problems. The splitting process is repeated until a stopping criterion is reached. After induction of the tree, a pruning process is applied in order to increase the generalization capability [21, 22].

3.3. Random forest

A random forest describes a set of decision tree predictors which can be used for classification as well as regression problems. Compared to single tree models, random forests are less prone to overfitting and have increased generalization capabilities. The diversity between the trees in a random forest is gained by growing each tree on a bootstrapped subset n_i of the whole dataset n. The bootstrapped datasets are generated by randomly resampling the data with replacement. Through this procedure roughly one third of the samples are not included in the training set which are called Out-of-Bag samples (OOB) and are used to estimate the feature importance. Moreover, only a random selection X_i of all the independent variables X is chosen as candidates for splitting criteria at the

tree nodes. Fig. 8 describes the procedure for growing a random forest [23–25].

3.4. Artificial neural network

Artificial neural networks (ANN or NN) have gained a lot of relevance over the last decade because of fast increasing computer hardware. They are classified as deep learning, a branch of machine learning, and are inspired by the human brain which resulted in a lot of shared terminology with neuroscience. ANN consist of several layers of neurons which perform simple calculations and connections between them of different "connection strengths" or "weights" indicating to what extend a signal is amplified or diminished. Each neuron receives input from and sends signals to many other neurons of the network. Furthermore, an activation function is applied in every neuron acting as a threshold to determine if a neuron fires or not. On its own a single neuron is not very powerful; the strength of the whole system stems from the interaction of many neurons connected in the right way [26,27].

3.5. Lasso regression

Least absolute shrinkage and selection operator regression, or LASSO regression in short, is a variant of linear regression which trades of bias for a better expected overall prediction and thus is less prone to overfitting [28]. Unlike normal linear regression LASSO regression tries to minimize the sum of the squared residuals plus a penalty term as shown in:

$$\sum_{i=1}^{n}\left(y_i-\sum_j x_{ij}\beta_i\right)^2+\lambda\sum_{j=1}^{p}|\beta_i| \qquad \text{Eq (3)}$$

Where λ is a model tuning parameter which has to be optimized by the algorithm. As λ increases bias increases but variance decreases and vice versa [29].

3.6. Hyperparameter optimization

Hyperparameters are defined as purely external parameters and thus are independent of the dataset. Nevertheless, they influence the model accuracy and should be optimized in that regard as further explained in Refs. [30–32].

For all above mentioned models the optimal hyperparameters were found using a randomized grid search over 200 combinations of predetermined ranges of each hyperparameter. This method yields good results and is computationally more efficient than an exhaustive grid search because it does not go over all the possible combinations [33,34]. The randomized search was used in tandem with 5-fold cross validation and the parameter ranges for each model type are listed in Table 2. Thus, all in all 1000 estimators were trained for each algorithm, or 4000 in total. For ANN the data was standardized during preprocessing.

3.7. Determination of important features

Out of the 29 documented columns only a small selection correlated well with the resulting CCR. To reduce model training time and counteract overfitting, an independent RF model was used to rank the relative importance of the features before model selection. The importance calculation is based on mean impurity reduction within each tree of the random forest, where the impurity is measured by least squares method. For example, the importance-value of feature 'A' is equal to the mean of all importance-values for feature 'A' between all trees which contain feature 'A'. The sum of all the relative importance yields 1 or 100%, respectively. The formula for measuring the feature importance in random forests is given in Eq. (*4*). Here M denotes the number of trees in the forest, φ_m the m-th tree (for m = 1, …, M), $p(t)$ the proportion of samples reaching node t, j_t the identifier of the variable used for splitting node t and $i(t)$ the impurity measure which in this case was RMSE [35].

Table 2
Optimized hyperparameters per algorithm; numbers in brackets indicate the range of candidates for the parameter space. Out of all possible combination a fixed amount of 200 per algorithm were trained.

Algorithm	Optimized Hyperparameters
DT	maximum features: [1–7], maximum depth: [4–11], minimum samples for split: [2,6,10,14,18,22,26], minimum samples per leaf: [1,3,5,7,9,11], splitter: [best, random]
RF	number of trees: [50, 60, 70, 80, 90, 100, 110, 120, 130], maximum features: [1–7], maximum depth: [4–11], minimum samples for split: [2,6,10,14,18,22,26], minimum samples per leaf: [1,3,5,7,9,11], bootstrap: [with, without]
ANN	activation: [logistic, reLu, tanh], solver: [lbfgs, adam, sgd], hidden layers: [(5, 5), (15, 15), (5, 5, 5), (50, 50, 50)], epochs: [100, 200, 300, 400, 500], learning rate: [constant, adaptive], initial learning rate: [0.001, 0.004, 0.007, 0.01], alpha: [0.0001, 0.0004, 0.0007, 0.001], momentum: [0.3, 0.6, 0.9]
LR	lambda: [0.01, 0.05, 0.1, 0.3, 0.5, 0.8, 1]

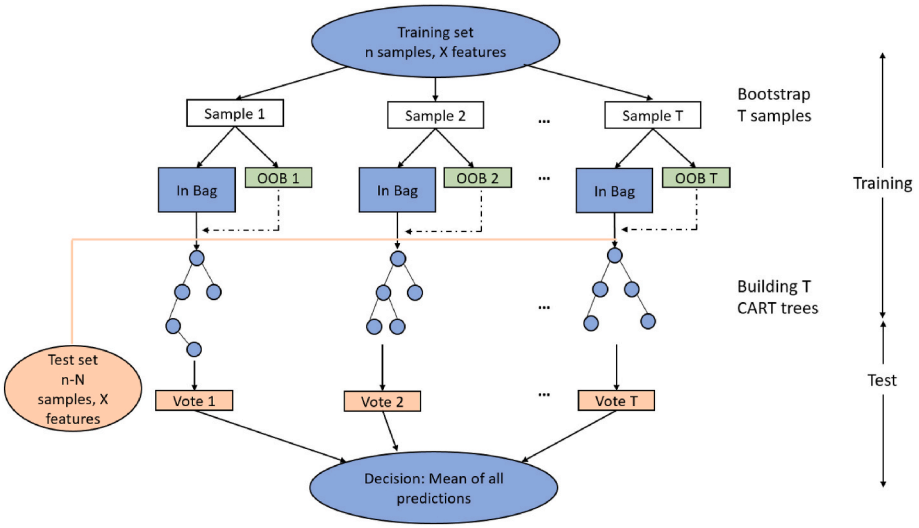

Fig. 8. Procedure of growing a random forest [23].

$$Imp(X_j) = \frac{1}{M}\sum_{m=1}^{M}\sum_{t\in\varphi_m} 1(j_t = j)[p(t)\Delta i(s_t, t)] \quad \text{Eq (4)}$$

4. Results and discussion

All the above explained models and methods were used and compared against each other for their suitability to predict the CCR of samples from the test set.

4.1. Effects of random and stratified sampling

In general stratification lead to more reliable model accuracy measures and lower standard deviations when repeatedly calculating accuracy. A reason for this can be made visible by plotting the CCR values of datapoints which ended in trainings- and test sets over their density in siliconized state as shown in Fig. 9. Density was used because it was by far the most influential parameter which the model used to predict CCR. Blue dots show samples from the training set and red dots samples from the test set.

Fig. 9 shows that stratification results in more representative splits than random drawing, in a way that samples from all over the CCR spectrum are included in the test- and training sets respectively. By using a random split no CCR values above 33 were present in the test set in this particular case which makes the reported model accuracy less meaningful.

Another way to prove the difference between random and stratified splitting is by comparing the mean, minimum, maximum and standard deviation of CCR values in the respective test sets calculated from 5 splits as shown in Table 3. It can be seen that random splitting leads to a high variance in mean CCR-values in test sets between each split; within the 5 splits, the mean CCR in the test set ranged between 15.0 and 21.3. For stratified splits the mean only ranged between 17.6 and 18.3. Since stratified splitting produced more reliable and representative model accuracy measures, it was preferred over random splitting.

4.2. Effects of mean and iterative imputation

Analogously the effect of imputation was examined using a similar set-up. During 15 model training processes using mean and iterative imputation respectively the R^2 of a random forest was tracked using an identical random seed for splitting and model creation. For iterative imputation the process was continued until a deviation tolerance of $\varepsilon =$ 1% between two consecutive steps was met or a maximum of 20 iterations was reached.

The models which used iterative imputation not only provided more realistic distributions (see Fig. 5) but also achieved slightly better R^2 scores, as shown in Table 4. For the iterative process the similarity between samples is calculated based on a given number of features n_f, which therefore can be understood as a hyperparameter. The best model accuracy was found for $n_f = 5$ which also fits the observed trends best when plotting the datapoints. Therefore, iterative imputation was preferred over mean imputation.

Table 3
Comparison of mean and standard deviation of CCR in test sets calculated from 5 splits using random sampling and stratified sampling.

	Random Splitting	Stratified Splitting
Mean CCR Interval (Test)	[15, …21.3]	[17.6, …18.3]
Mean ± Std. in Interval	18.1 ± 2.4	18.0 ± 0.3
R^2	0.54	0.56

Table 4
Mean model accuracy for models with mean imputation and iterative imputation for missing values.

Mean Imputer R^2	Iterative Imputer R^2
0.55 ± 0.13	0.60 ± 0.12

4.3. Most important features

Feature Selection was based results from 20 different test/train splits, thus a mean importance and standard deviation could be calculated. To determine an appropriate threshold value for inclusion or exclusion of features, an artificial feature was added to the dataset which contained random numbers and thus showed no correlation to CCR. Since the relative importance was also determined for the random feature, it could be used as a guideline in a way that any features with importance values in the range of or lower than its importance could safely be excluded from the model. This led to the decision that only features with a relative importance of 5% or higher were passed on to the AI-model which resulted in only 3 of the initial 29 features to be selected as shown in Fig. 10.

The determination of relative feature importance concluded that density in siliconized state, mass change during siliconization and open porosity in siliconized state were the 3 most important features to predict CCR. Out of these, density was by far the most important characteristic. The high correlation between density in siliconized state and CCR was expected because microstructures with high degrees of SFS absorb more silicon during siliconization.

Although this trend makes a lot of sense, its information value is rather low. The earlier in the production line a correlation between a parameter and the CCR can be drawn the higher the potential for saving resources or labor time. Predicting CCR from a sample in siliconized state thus is rather unspectacular because all the labor- and energy intense production steps have already been done.

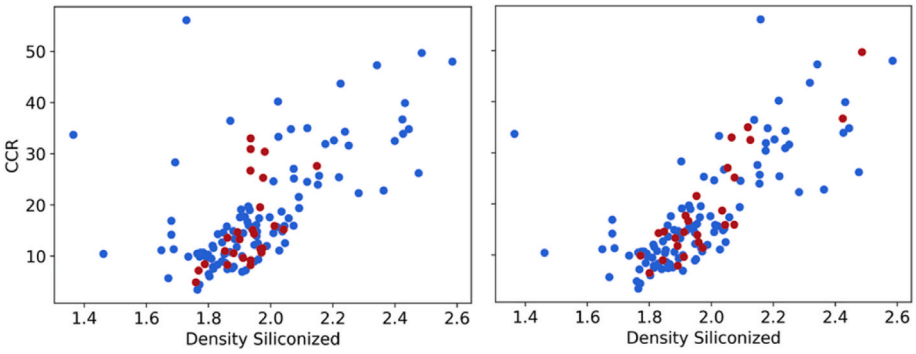

Fig. 9. Comparison of a random split (left) and a stratified split (right) of samples into testing (red) and training data (blue). The random split does not yield CCR values over 33 in the test set whereas the stratified split does. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

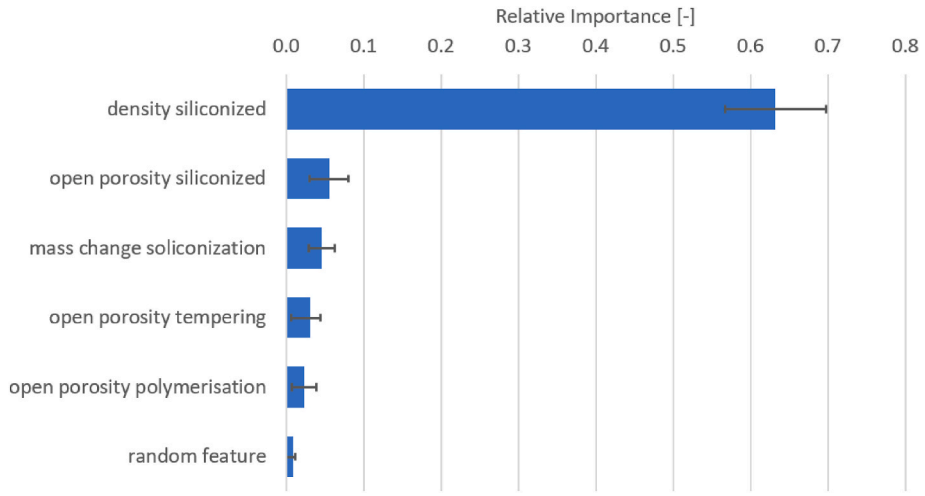

Fig. 10. Relative importance and standard deviation of the 5 most important features for predicting CCR; a random feature was added for comparison.

By excluding all measurements taken during siliconization and re-training the model on the new dataset, the most important features could be determined for this case analogously (see Fig. 11). The resulting most important feature was now porosity in tempered state, not so closely followed by porosity in polymerized state, porosity in pyrolyzed state and the information weather the used fiber was of type T800 or not. All other features fell below the 5% threshold which was again set as a criterion for a feature to be included in the model.

It has to be mentioned that by excluding siliconization the determined importance values showed a lot more dispersion and the models trained from this dataset generally achieved lower R^2 scores (see Table 5). Nevertheless, correlations of CCR with processing steps as early as polymerization (CFRP state) could be drawn. This is also in accordance with some findings in literature where the appropriate processing of the composite in CFRP state prior to pyrolysis is regarded as the crucial step in the production of XB-C/C–SiC [36].

Fig. 12 shows a plot of CCR over both porosity in polymerized as well as in tempered state. By looking at porosity alone, a linear correlation with CCR could be drawn, which is characterized by high dispersion. Upon closer inspection the scatter could be reduced by simultaneous separation of the data points into fiber types T800 and HTA. Here HTA fibers clearly show higher CCR values for higher porosity whereas this trend is much flatter for T800 fibers. Furthermore, HTA fibers generally yield higher CCR values than T800 fibers. This trend was also picked up by the algorithm as the information wether the fiber is of type T800 or not was the 4th most important feature when excluding siliconization.

Table 5
Comparison of achieved mean R^2 score for the best suited model algorithm (RF) and n = 20 repetitions in case of including and excluding siliconization measures from the dataset.

R^2 (RF, Test Set, including siliconization)	R^2 (RF, Test Set, excluding siliconization)
0.62 ± 0.16	0.49 ± 0.17

Further splitting into resin types did not provide additional insight with XP60 being the only exception (only 4 most prominent resin types shown in Fig. 12). For XP60 the trend of higher CCR values for higher porosity did not apply.

A possible reason for the general difference in CCR of samples with HTA or T800 fibers could be a difference in fiber matrix bonding strength. The correlation between interface strength and microstructure was already proven in Brandt et al. [37], Schulz [38] as well as Schulte-Fischedick [39] where a weaker interface between fibers and matrix was shown to produce XD-microstructure (and thus high CCR) whereas a stronger interface benefits the formation of XB-structures (and thus low CCR). Although, the measurement of interface strengths

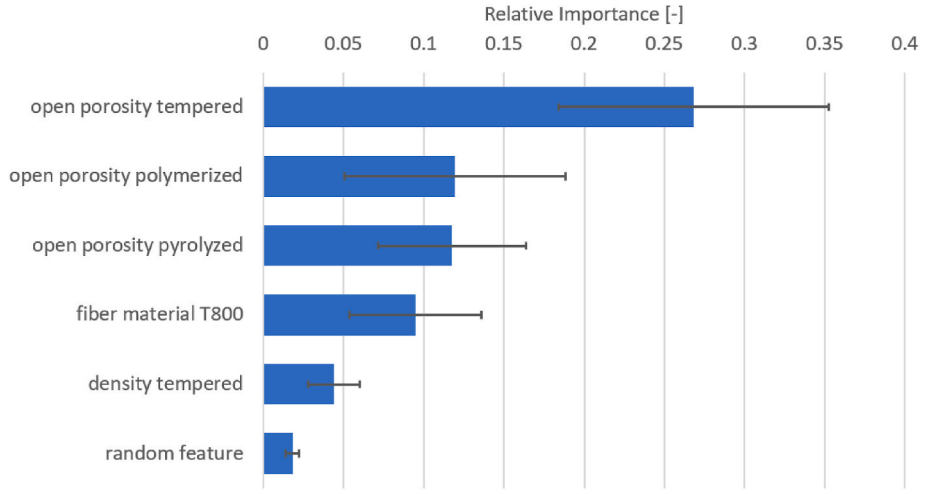

Fig. 11. Relative importance and standard deviation of the 5 most important features for predicting CCR by leaving out all measurements taken during siliconization; a random feature was added for comparison.

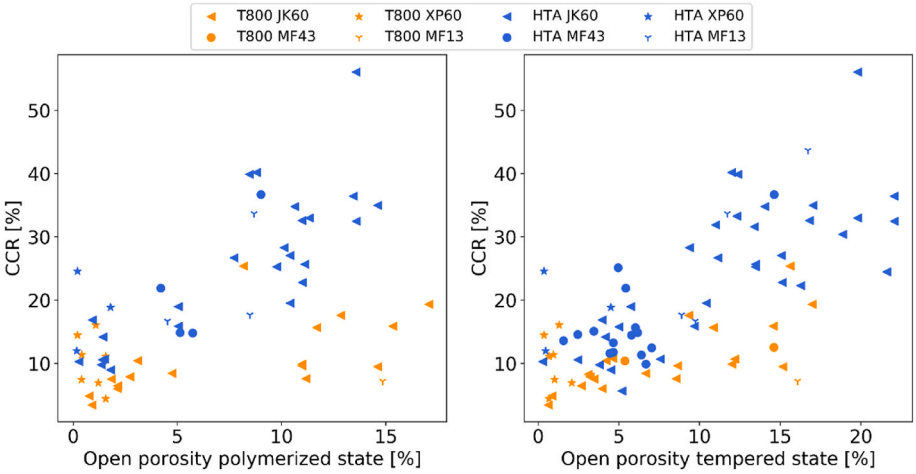

Fig. 12. CCR over open porosity in polymerized and tempered state grouped by fiber type (color-coded) and resin type (symbol-coded). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

of HTA and T800 fibers to prove this hypothesis exceeded the scope of this work and no applicable information was found in literature.

Porosity itself was not dependent of the fiber type but in some cases of the resin type used. Fig. 13 shows CCR over porosity grouped by the 4 most frequently used resin types of this work. It can be seen that XP60 led to the least amount of porosity (e' ~ 1%) followed by MF43 (e' ~ 4%). JK60 and MF13 both led to unpredictable kinds of porosity apparently depending on other properties which could not be determined in this work.

4.4. Comparison of model accuracy

The coefficient of determination R^2 was calculated for every algorithm on the test set after the optimal hyperparameters were chosen during CV. This process was then repeated 20 times with varying test/train splits in order to provide a mean and standard deviation for R^2. Fig. 14 provides an overview over the results separated by model algorithm.

Three observations can be made in Fig. 14:

- Out of the four tested algorithms RF performed best given the underlying data with a score of $R^2 = 0.62 \pm 0.16$ if data from all manufacturing steps was used
- The standard deviation of all models was comparably large
- DT and NN models showed a larger gap between R^2 score during CV and final R^2 score on the test set which indicates overfitting

The high standard deviation is an indicator for high variance in the underlying data. Moreover, the number of samples is comparably small considering the high number of parameters, whilst simultaneously containing a lot of missing values. This also means that the imputation method plays an important role for the results.

The better performance of simpler models such as random forest and lasso regression compared to the neural network is probably due to the limited amount of data, as traditional ML methods tend to outperform deep learning in these scenarios [31]. RF also have greater generalization capabilities compared to DT which explains the better performance and the lower drop in accuracy between cross validation and test set.

5. Conclusion

In this paper the influence of production parameters as well as choice of raw materials on the microstructure formation of C/C–SiC samples was investigated by machine-learning methods. The goal was to find the most important parameters from a given selection which lead to C/C–SiC with XB- or XD microstructure. As ground truth 123 samples with varying manufacturing parameters were used. A lot of missing data was present and thus imputed by an iterative approach. After preprocessing, four different supervised machine-learning algorithms were trained on the dataset and compared using R^2 as accuracy metric from which RF performed best.

Within the scope of this study the most relevant factors for either receiving high or low amounts of single fiber siliconization could be

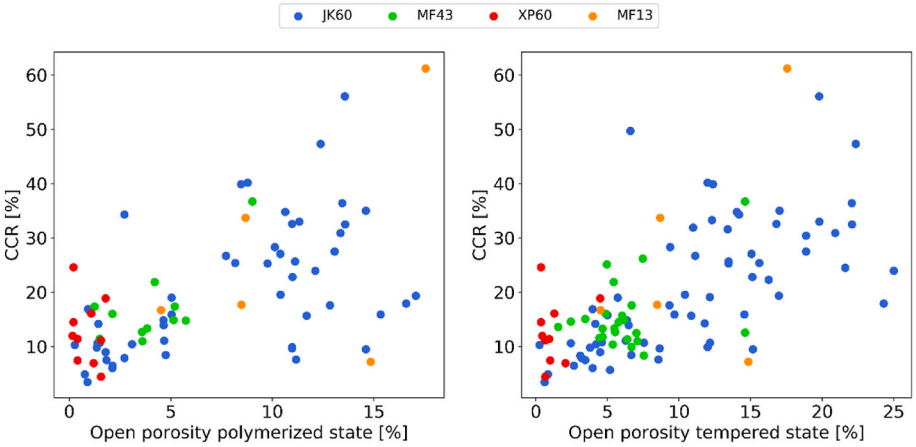

Fig. 13. Received porosity by using different resins; left: polymerized state, right: tempered state.

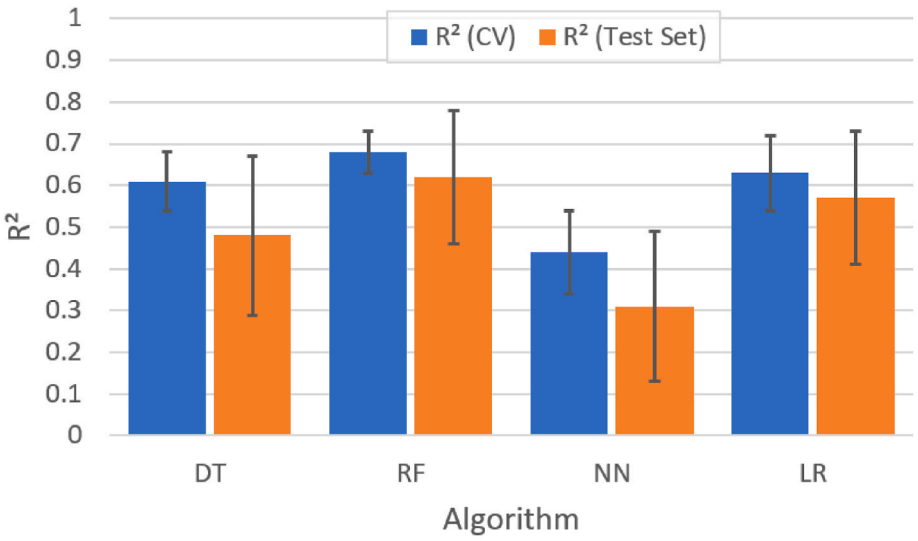

Fig. 14. Coefficient of determination for the best model for different algorithms; DT = decision tree, RF = random forest, NN = artificial neural network, LR = lasso regression.

determined by a model intrinsic method. For that, the C/C–SiC microstructures were quantified by introducing CCR in order to be able to feed this information into the models.

Generally, it can be said that the evaluation of feature importance done by the model is more reliable the greater its accuracy is. Due to the mediocre accuracy of $R^2 = 0.62$ achieved by the best models stemming from the high dispersion in the data, the determination of feature importance has to be evaluated carefully through subsequent investigations. If only data from CFRP state was used, the mean model accuracy for predicting CCR dropped to $R^2 = 0.49$.

Observed trends by the algorithms were:

- Data gained during or after siliconization was the most important for predicting CCR. Especially a high density correlated well with high CCR values, followed by open porosity and mass change during siliconization.
- If siliconization was taken out of the equation, porosity in tempered, polymerized and pyrolyzed state were the most important features to predict CCR.
- T800 fibers generally benefited lower CCR values compared to HTA fibers for the same amount of open porosity in CFRP state.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

List of 29 processing parameters used to train the models:

Density in polymerized state, density in tempered state, density in pyrolyzed state, density in siliconized state, porosity in polymerized state, porosity during tempering, porosity in siliconized state, porosity in pyrolyzed state, mass change during pyrolysis, mass change during siliconization, duration of polymerization, duration of tempering, maximum temperature during polymerization, maximum temperature during tempering, maximum temperature during pyrolysis, maximum temperature during siliconization, number of siliconizations, number of pyrolyses, precursor, fiber-pretreatment, desizing of fibres, fiber volume content in CFRP state, fiber-material, fiber architecture, fiber density, manufacturing method for CFRP, fiber orientation, sample thickness, geometry from which the sample was cut.

References

[1] W. Krenkel (Ed.), Ceramic Matrix Composites – Fiber Reinforced Ceramics and Their Applications, Wiley-VCH, Weinheim, 2008.
[2] M.K. Naskar, M. Chatterjee, A. Dey, et al., Effects of processing parameters on the fabrication of near-net-shape fibre reinforced oxide ceramic matrix composites via sol–gel route, Ceram. Int. 30 (2) (2004) 257–265, https://doi.org/10.1016/S0272-8842(03)00097-X.
[3] M. Frieß, W. Krenkel, R. Brandt, et al., Influence of process parameters on the thermophysical properties of C/C-SiC, in: W. Krenkel, R. Naslain, H. Schneider (Eds.), (Hrsg.): High Temperature Ceramic Matrix Composites, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG, 2001, pp. 328–333.
[4] W. Zhu, H. Fu, Z. Xu, et al., Fabrication and characterization of carbon fiber reinforced SiC ceramic matrix composites based on 3D printing technology, J. Eur. Ceram. Soc. 38 (14) (2018) 4604–4613, https://doi.org/10.1016/j.jeurceramsoc.2018.06.022.
[5] Z.C. Li, Y. Jiao, T.W. Deines, et al., Rotary ultrasonic machining of ceramic matrix composites: feasibility study and designed experiments, Int. J. Mach. Tool Manufact. 45 (12–13) (2005) 1402–1411, https://doi.org/10.1016/j.ijmachtools.2005.01.034.
[6] O. Gavalda Diaz, G. Garcia Luna, Z. Liao, et al., The new challenges of machining Ceramic Matrix Composites (CMCs): review of surface integrity, Int. J. Mach. Tool Manufact. 139 (2019) 24–36, https://doi.org/10.1016/j.ijmachtools.2019.01.003.
[7] J.J. Moses, I. Dinaharan, S.J. Sekhar, Prediction of influence of process parameters on tensile strength of AA6061/TiC aluminum matrix composites produced using stir casting, In: Trans. Nonferrous Metals Soc. China 26 (6) (2016) 1498–1511, https://doi.org/10.1016/S1003-6326(16)64256-5.
[8] J.S. Huang, J.X. Liew, A.S. Ademiloye, et al., Artificial intelligence in materials modeling and design, In: Arch. Comput. Methods Eng. 28 (5) (2021) 3399–3413, https://doi.org/10.1007/s11831-020-09506-1.
[9] G.A. Xiang, L.I. Guanghui, T.A. Rong, et al., Using deep neural networks to predict the tensile property of ceramic matrix composites based on incomplete small dataset, in: IOP Conference Series: Materials Science and Engineering, vol. 647, 2019, 12004, https://doi.org/10.1088/1757-899X/647/1/012004. Heft 1.
[10] H. Ghayour, M. Abdellahi, M. Bahmanpour, Artificial intelligence and ceramic tools: experimental study, modeling and optimizing, In: Ceram. Int. 41 (10) (2015) 13470–13479, https://doi.org/10.1016/j.ceramint.2015.07.138.
[11] K.S. Aggour, V.K. Gupta, D. Ruscitto, et al., Artificial intelligence/machine learning in manufacturing and inspection: a GE perspective, In: MRS Bull. 44 (7) (2019) 545–558, https://doi.org/10.1557/mrs.2019.157.
[12] M. Patel, K. Saurabh, V.V.B. Prasad, et al., High temperature C/C–SiC composite by liquid silicon infiltration: a literature review, In: Bull. Mater. Sci. 35 (1) (2012) 63–73, https://doi.org/10.1007/s12034-011-0247-5.
[13] M. Frieß, CMC with a Graded Lay-Up Manufactured via LSI-Process, CIMTEC 2010, 12th Int. Ceramics Congress, Montecatini Terme (Italy), 2010.
[14] J. Schulte-Fischedick, A. Zern, J. Mayer, et al., The morphology of silicon carbide in C/C–SiC composites, In: Mater. Sci. Eng. 332 (1–2) (2002) 146–152, https://doi.org/10.1016/S0921-5093(01)01719-1.
[15] G.S. Mudholkar, D.K. Srivastava, C. Thomas Lin, Some p-variate adaptations of the shapiro-wilk test of normality, In: Commun. Stat. Theor. Methods 24 (4) (1995) 953–985, https://doi.org/10.1080/03610929508831533.
[16] K. Potdar, C. Pai, T. Pardawala, A comparative study of categorical variable encoding techniques for neural network classifiers, in: International Journal of Computer Applications, 2017, https://doi.org/10.5120/ijca2017915495.

[17] S. van Buuren, Flexible Imputation of Missing Data, Chapman & Hall/CRC, Boca Raton, 2021.
[18] Q. Shi, M. Abdel-Aty, J. Lee, A Bayesian ridge regression analysis of congestion's impact on urban expressway safety, Accid. Anal. Prev. 88 (2016) 124–137, https://doi.org/10.1016/j.aap.2015.12.001.
[19] F. Farias, T. Ludermir, C. Bastos-Filho, Similarity Based Stratified Splitting: an Approach to Train Better Classifiers Ausgabe, 2020.
[20] D. Chicco, M.J. Warrens, G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, In: PeerJ. Computer science 7 (2021) e623, https://doi.org/10.7717/peerj-cs.623.
[21] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, et al., Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines, In: Ore Geol. Rev. 71 (2015) 804–818, https://doi.org/10.1016/j.oregeorev.2015.01.001.
[22] L. Breiman, J.H. Friedman, R.A. Olshen, et al., Classification and Regression Trees, Routledge, 2017.
[23] L. Guo, N. Chehata, C. Mallet, et al., Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests, ISPRS J. Photogrammetry Remote Sens. 66 (1) (2011) 56–66, https://doi.org/10.1016/j.isprsjprs.2010.08.007.
[24] W.-Y. Loh, Classification and regression trees, In: WIREs Data Min. Knowl. Discov. 1 (1) (2011) 14–23, https://doi.org/10.1002/widm.8.
[25] Segal, M.: Machine learning benchmarks and random forest regression. In: : CSF: Center for Bioinformatics and Molecular Biostatistics, S. 4-6.
[26] H. Kukreja, K. Shiruru, An introduction to artificial neural network, In: Int. J. Adv. Res. Innov. Ideas Educ. 1 (5) (2016) 27–30. Vol.1 Issue-5.
[27] J. Steinwendner, R. Schwaiger, Neuronale Netze programmieren mit Python, Rheinwerk Computing, 2019.
[28] J. Ranstam, J.A. Cook, LASSO regression, In: Br. J. Surg. 105 (10) (2018) 1348, https://doi.org/10.1002/bjs.10895.
[29] Encyclopedia of Statistical Sciences, A Wiley-Interscience publication, Wiley, New York, NY, 1982.
[30] B.H. Shekar, G. Dagnew, Grid search-based hyperparameter tuning and classification of microarray cancer data, in: 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), IEEE, Gangtok, India, 2019, pp. 1–8.
[31] O.R. Sanchez, M. Repetto, A. Carrega, et al., Evaluating ML-based DDoS detection with grid search hyperparameter optimization, in: 2021 IEEE 7th International Conference on Network Softwarization (NetSoft), IEEE, Tokyo, Japan, 2021, pp. 402–408.
[32] H. Alibrahim, S.A. Ludwig, Hyperparameter optimization: comparing genetic algorithm against grid search and bayesian optimization, in: 2021 IEEE Congress on Evolutionary Computation (CEC), IEEE, Kraków, Poland, 2021, pp. 1551–1559.
[33] P. Liashchynskyi, P. Liashchynskyi, Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS Ausgabe, 2019.
[34] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, In: J. Mach. Learn. Res. 13 (2012) 281–305 (2012).
[35] G. Louppe, Understanding Random Forests – from Theory to Practice, University of Liège. University of Liège Department of Electrical Engineering & Computer Science, Liège, 2015.
[36] M. Frieß, R. Renz, W. Krenkel, Hrsg, Graded Ceramic Matrix Composites by LSI-Processing, 2002.
[37] R. Brandt, M. Frieß, G. Neuer, Thermal conductivity, specific heat capacity, and emissivity of ceramic matrix composities at high temperatures, In: High. Temp. - High. Press. 35/36 (2) (2003) 169–177, https://doi.org/10.1068/htjr105.
[38] M. Schulz, Einfluss der Faser-Matrix-Anbindung auf die Ausbildung der Mikrorissstruktur bei der Herstellung von keramischen Faserverbundwerkstoffen im Flüssigsilizierverfahren, PhD. Universität Augsburg, Universität Augsburg, 2021.
[39] J. Schulte-Fischedick, Die Entstehung des Rissmusters während der Pyrolyse von CFK zur Herstellung von C/C-Werkstoffen, 2006. PhD-Thesis.