“Maturing the production standards of ultraporous structures for high density hydrogen storage bank operating on swinging temperatures and low compression” – MAST3RBoost

D1.5. Descriptor performance in Machine Learning

Due date of submission: 30/11/2022

Actual submission date: 24/02/2023

Funded by the European Union

TABLE OF CONTENTS

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

PROJECT INFORMATION

Project full title: Maturing the production standards of ultraporous structures for high density hydrogen storage bank operating on swinging temperatures and low compression

Acronym: MAST3RBoost

Call: HORIZON-CL4-2021-RESILIENCE-01

Topic: HORIZON-CL4-2021-RESILIENCE-01-17

Start date: 1st June 2022

Duration: 48 months

List of participants:

Number	Name of beneficiary	Acronym of beneficiary	Country
1	ENVIROHEMP	ENV	Spain
2	CONTACTICA	CTA	SPAIN
3	Consejo Superior de Investigaciones Científicas	CSIC	Spain
4	Spike Renewables Srl	SPIKE	Italy
5	EDAG Engineering GmbH	EDAG	Germany
6	Nanolayers	NANO	Estonia
7	FUNDACIÓN CIDETEC	CIDETEC	Spain
8	Leichtmetallkompetenzzentrum Ranshofen GmbH	LKR	Austria
9	University of Pretoria	UP	South Africa
10	Council for Scientific and Industrial Research	CSIR	South Africa
11	PSA	PSA	Portugal
12	TWI Ltd	TWI	UK
13	University of Nottingham	UoN	UK

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

DELIVERABLE DETAILS

Document Number:	D1.5
Document Title:	Descriptor performance in Machine Learning
Dissemination level	PU – Public
Period:	PR1
WP:	1
Task:	T1.3
Author:	Nanolayers OÜ
Abstract:	Using the database of pre-existing nano-porous materials provided by University of Nottingham, Nanolayers performed statistical analysis and applied conventional machine-learning techniques to train a kernel-ridge regression model that predicts a sample characterisation outputs from its synthesis parameters. Principal components analysis made it possible to identify the most important combinations of input features, and thus reducing the dimensionality of the input data. The final model accuracy is not quite good, especially when predicting quantities with large variability across samples. However, this is expected given the very small size of the database, and can only improve as more data becomes available during the project.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

1 Input Data

The initial database of pre-existing materials consists of 90 samples, and includes their manufacturing details, as well as the results of their performance characterisation measurements. The manufacturing parameters used as inputs in machine-learning (ML), providing a hopefully complete numerical description of each sample. These quantities are mostly the atomic composition of the precursor mixture used to synthesize the sample derived from elemental analysis of the precursors (Table 1).

Table 1: input quantities describing the samples.

Name	Units	Notes
O/C	adim.	precursor mixture atomic ratios O/C
H/C	adim.	precursor mixture atomic ratios H/C
N/C	adim.	precursor mixture atomic ratios N/C
rel_C_atomic	adim.	relative concentration of carbon
rel_O_atomic	adim.	relative concentration of oxygen
K/C	adim.	precursor mixture atomic ratios K/C
K/O	adim.	precursor mixture atomic ratios K/O
Activation_temperature	°C	temperature of pyrolysis
wt_ratio_porogen	adim.	porogen weight ratio
porogen	adim.	type of porogen, 0: $K_2C_2O_4$, 1: KOH

Performance measurements are used as output quantities that machine-learning models are trained to predict from the inputs. These are listed in Table 2.

Table 2: output quantities describing the performance of the samples.

Name	Units	Notes
ABET	m^2 / g	surface area (BET method)
Amicro	m^2 / g	micropore surface area (t-plot method)
Vsp	cm^3 / g	pore volume (single point method)
Vmicro	cm^3 / g	micropore volume (t-plot method)
rel_C_atomic	adim.	relative concentration of carbon
rel_O_atomic	adim.	relative concentration of oxygen
O/C	adim.	sample composition, atomic ratios O/C
H/C	adim.	sample composition, atomic ratios H/C
N/C	adim.	sample composition, atomic ratios N/C

The dataset can be found in tabular form, along with a detailed description of the data sources, in a LabCore notebook (https://labcore.nanolayers.com/notebookfluid/63761cd8e05f2a53fab2fdb7).

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

In order to apply machine-learning methods, the dataset was downloaded from the notebook and converted into two separate matrix representations for inputs and outputs. Each sample is represented by one row in the matrix, with its input/output features appearing on columns. It should be noted that despite having 90 samples in total, only 80 of them have all the input/output values: those with missing entries cannot be used, thus have to be discarded.

As is standard practice, the Z-score transformation is applied to the input matrix, which shifts each column around its mean and normalises it by its standard deviation. This ensures that the data is centred around zero and all the numbers have similar magnitude, so that input features that are numerically larger will not have any more significance than small ones. For example, temperature is always in the hundreds, and atomic ratios are always below unity, thus the former would always dominate any metric of comparison between sample descriptors. Applying the Z-score improves the robustness of the ML method and the performance of training algorithms. The mean and standard deviation of each column in the original dataset have to be recorded so that new sample descriptors can be normalised in the same way as the training data, otherwise ML predictions would be incorrect.

Principal component analysis (PCA) was applied to the input dataset in order to find the optimal input descriptor. This method replaces each input feature with *principal components*, i.e. linear combinations of the input features of the same sample. The linear transformation is determined by the covariance matrix of the inputs, so that the transformed features maximise the variance between samples.

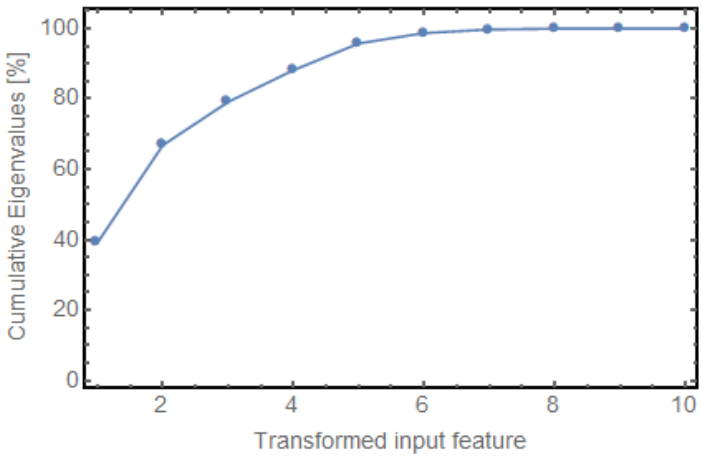

Figure 1: cumulative sum of the covariance matrix eigenvalues (normalised).

The eigenvectors of the covariance matrix make up the transformation matrix, while its eigenvalues can be interpreted as the information The eigenvalues of the covariance matrix represent the amount of variance in the data that each principal component is responsible for: loosely speaking, they quantify the information content of each component. The sum of all eigenvalues should add up to the total number of components, however, some principal components have much larger eigenvalues than others. Looking at the cumulative sum of the eigenvalues (Figure 1) reaveals that 95.7% of the data variance can be represented by just the first five principal components, while 98.7% is recovered with six. This means that it is possible to simplify the input data by ignoring the last four or five components, without losing too much information; doing so will allow for simpler ML models to be trained, increasing the ratio between data samples and fitting parameters, which often leads to more generalisable and robust models, without significant loss of accuracy.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

It should be noted that disregarding some of the principal components (PCs) is not necessarily equivalent to the elimination the original input features from the dataset. PCs are linear combinations of all the input features: each input feature, multiplied by appropriate coefficients, contributes to the definition of each PC. It is only possible to eliminate an input feature when its coefficients are negligible in all PCs that are intended to be used. However, even in such case it is not advisable to do so, because while the feature may seem unimportant in representing the variance of the dataset, this is only true for this dataset. The importance of each individual feature in the PCs will change when the dataset is extended with new samples.

2 Machine-Learning Model

Given the small dataset size (80 samples), neural networks should not really be considered, and the only sensible method to attempt is kernel-ridge regression (KRR).

The method can be intuitively seen as a statistical fit in the high-dimensional space of the inputs. Given any sample descriptor, a metric computes its distance to all the training descriptors, and the results are used with the training outputs to formulate the prediction. The first step in training a KRR model is to build the kernel matrix:

$$K_{ij} = d(M_i, M_j) - \alpha I_{ij}$$

where d is a metric function to estimate the distance between training samples M, α is the regularisation parameter, and I the identity matrix. In our study, the descriptor of a training sample M_i is a vector with 8 components: one for each input feature listed in Table 1. Thus, the simplest choice of metric is the Euclidean distance between the input vectors. The regularisation parameter α = 0.1 is used to avoid overfitting: this prevents the model from predicting the correct output for training inputs, and in general increases the prediction error, however, the model also becomes more generally applicable outside of the training data space. The training is completed with the inversion of the kernel matrix.

Given a set of input descriptors *X*, their outputs *Y* can be predicted from their distance from the training inputs *M*, and the output values of the training inputs Y_M. First, the distance matrix between the new inputs *X* and the training inputs *M* is constructed using the metric:

$$D_{ij} = d(X_i, M_j).$$

The following expression calculates the outputs for all inputs in *X* with simple matrix-matrix multiplications:

$$Y = Y_M \cdot K^{-1} \cdot D^T$$

Given the numerical complexity of matrix inversion and multiplication, KRR is generally the preferred method when dealing with small datasets (N ~ 1000), as training cost scales cubically with the size of the dataset, and prediction cost scales quadratically.

The accuracy of the trained model can be assessed by calculating the prediction error for a validation dataset of samples not included in the training set, for which the true outputs are known. This is obtained by randomly dividing the database into two sets: the training set takes 70% of the total data, while the remaining 30% is used for validation. The whole training and evaluation procedure is repeated 1000 times, with different random database splitting, in order to average the prediction error estimates.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

3 RESULTS

The accuracy of the KRR model is evaluated by the mean relative error of its predictions on the validation dataset, shown in Figure 2.

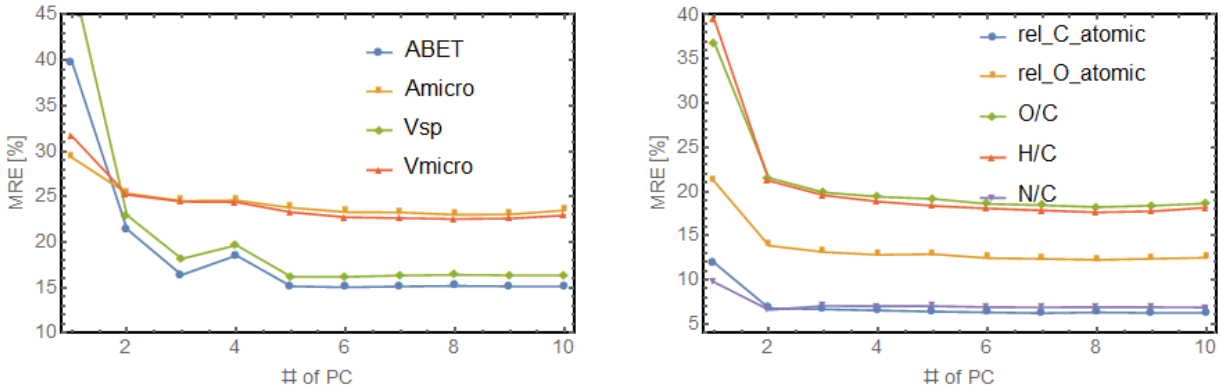

Figure 2: Mean relative error of the KRR method predictions of the output quantities, depending on the number of principal components included in the input. Values are averaged over 1000 runs with different random training/validation splits of the whole dataset.

As expected, including more principal components (PCs) in the inputs reduces the model prediction error, however, beyond the first six PCs there are no significant improvements. The error seems to even increase slightly when all PCs are used: this could be due to a combination of statistical noise as well as an actual loss of effectiveness of the Euclidean norm in representing the true distance between samples.

Overall, the largest error or about 23% is seen for pore volume and area obtained with the t-plot method (Amicro and Vmicro), while predictions of the same quantities from BET and single-point method (ABET and Vsp) have an error of about 15.5%; similar accuracy is seen for the O to C and H to C ratios in the sample composition, with an error of 18%. The relative O concentration is better predicted, with 12% error, while the relative C concentration and the N to C ratio in the sample are calculated with the best accuracy of about 6%.

As often happens with ML models, the reason for the wide range of prediction accuracy for the different output quantities is not really known, since the inner mechanism of the model is not transparent. However, there is a clear correlation between the model accuracy for a quantity and its variance in the dataset, shown in Figure 3.

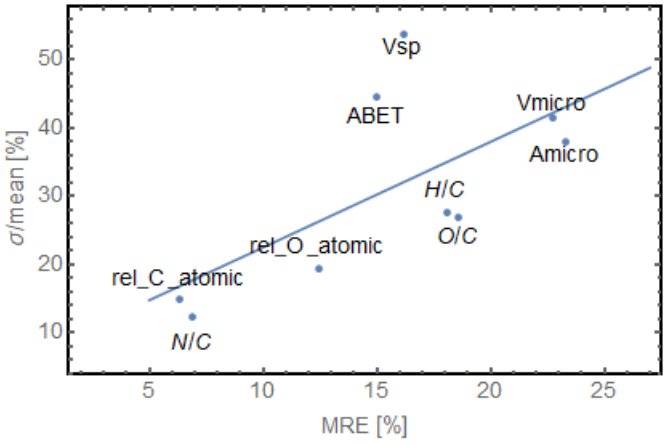

Figure 3: correlation between the model mean relative error on an output quantity and its standard deviation in the dataset. The standard deviation is expressed as percentage of its mean value. The fitting line gives $R^2 = 0.9$.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

Generally, the KRR model gives better accuracy on output quantities that vary less in the dataset, with a correlation factor of 0.99. One reasonable explanation is that the model has very little data available to learn the input-output relationship, and this becomes visible for output quantities with larger variability: there is not enough variety in the input data to infer such relationship accurately. Another possibility is that the input quantities themselves do not contain enough information about the physics and chemistry happening during the sample manufacturing process. However, the pore area from BET method and volume from single-point method exhibit large variance in the data, but their prediction error is smaller than expected from the simple correlation line. This is an encouraging result because these quantities are of major interest for materials optimisation, and it appears that our model was able to capture their relationship to the input features better than expected.

As is often the case in ML applications to materials science, the most likely source of error is the limited amount of data, usually orders of magnitude away from the size where ML methods operate comfortably. In this case there are only 80 samples (or less considering only 70% are used for training), and they are still too few for anything but preliminary investigations, even with KRR.

4 Conclusions & future work

In conclusion, while the dataset size is too small to produce reliable supervised ML models of the complex physics behind the sample fabrication from the precursor mixture, there is evidence that a standard KRR model can indeed approximate such process. Interesting porosity outputs (ABET and Vsp) vary significantly across samples, and yet show prediction errors lower than expected when compared to other output quantities. This suggests there is a significant correlation between pore area/volume and input features, and KRR was able to capture it. The accuracy of the model is expected to increase as more samples are tested and added to the dataset by experimental partners, and depending on the final dataset size, it might even be possible to test more advanced supervised ML methods, such as neural networks or genetic programming.

Should the model accuracy not increase significantly as more data is made available, the most likely cause would be the limited information content in the input features. In such case it will be necessary to include more physically or chemically insightful features, for example describing the molecular properties of the precursor mixture and porogen, or more details about the process. Currently, there is no indication this will be necessary, but the information limit may be masked by the small dataset size.

Unsupervised ML methods, PCA in this case, also proved to be quite important in the whole process. First it combined the input features into optimal principal components and revealed that only half of them could be discarded by the supervised ML model without significant information loss. Reducing the dimensionality of the problem improved the efficiency of KRR, making it possible to work with the limited dataset size. Indeed, even after extending the dataset with new measurements PCA should still be used to improve the efficiency of supervised learning.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.