# Input Structure Selection for Soft-Sensor Design: Does It Pay Off?

Martin Mojto
*Slovak University of Technology in Bratislava*
Bratislava, Slovakia
martin.mojto@stuba.sk

Karol Ľubušký
*Slovnaft, a.s.*
Bratislava, Slovakia
karol.lubusky@slovnaft.sk

Miroslav Fikar
*Slovak University of Technology in Bratislava*
Bratislava, Slovakia
miroslav.fikar@stuba.sk

Radoslav Paulen
*Slovak University of Technology in Bratislava*
Bratislava, Slovakia
radoslav.paulen@stuba.sk

***Abstract*—This contribution seeks to compare the efficiency of several soft-sensor design methods using the datasets from three different case studies, including two industrial examples. The selected set of design methods compared considers different principles to design soft sensors, consisting of three consequent stages: (a) data preprocessing with input domain analysis, (b) input structure selection, and (c) model training. The input domain analysis explores the potential of various nonlinear structures within the input dataset. The results obtained indicate that multivariate feature selection approaches have the highest efficiency. Moreover, the nonlinear soft sensors achieve higher accuracy compared to linear ones.**

***Index Terms*—feature selection, industrial processes, model training, soft sensors**

## I. Introduction

The monitoring of the process variables is a crucial aspect of sustainable operation. In fact, the key process variables (e.g., product quality) may directly affect the decisions of the operators or the advanced process controller in the industry. Monitoring the key process variables can be provided by soft (or inferential) sensors [1]–[4]. The principle of such a sensor is to estimate the desired variable according to the available measurements of other variables involved in the process. From a practical point of view, soft sensors find application in many industrial fields, such as petrochemical [3], pharmaceutical [4]–[8], or energy [9] industries.

Data-driven design of soft sensors is highly related to the quality of the available dataset. The raw industrial dataset usually involves outliers originating from abnormal operating conditions (e.g., plant shutdowns). Therefore, it is necessary to appropriately analyze and preprocess the data before they are used for the soft-sensor design. Recent research [10] showed an effective way of combining nonlinear soft sensors with outlier detection. In addition to the data treatment, the information capacity of the dataset can be significantly increased by exploring the potential of various nonlinear transformations of the input variables. The effective way of identifying these nonlinear transformations ensures the Automated Learning of Algebraic Models for Optimization (ALAMO) [11] approach. The abilities of ALAMO are explored by a new framework for the development of data-driven soft sensors [12].

The essential phase of the data-driven soft-sensor design is selecting the appropriate input structure. The best subset of input variables is selected to reduce the computational effort of the subsequent soft-sensor design while retaining the information content of the original dataset. The current research focus in this field indicates an increasing interest in feature selection (FS) approaches [13]. These approaches are well-suited to indicate the best subset of input variables based on specific objectives, such as classification [14] or cross-validation [15]. The least absolute shrinkage and selection operator (LASSO) [16] can also be used for variable selection. It can be combined with Principal Component Analysis (PCA) [17], [18]. The resulting approach not only allows for input variable selection but also for effective dimensionality reduction of the input dataset [19].

Once the input variables are selected, one designs the data-driven soft sensor by fitting the model parameters. According to the structure of the soft sensor, we distinguish between the approaches suitable for linear [3] and nonlinear [20] model structures. Recently, an adaptive soft sensor [21] based on the Gaussian process (GP) model, was designed for the debutanizer distillation column. Moreover, the advanced partial least square (PLS) algorithm has been developed to design soft sensors for industrial units [22].

This paper examines the performance of various approaches in variable selection and model training stages of soft-sensor design. Combinations of these approaches are applied to design soft sensors for three case studies, including two real industrial applications. The analysis of the soft-sensor performance includes consideration of both input structure complexity and prediction accuracy.

The structure of this paper is organized as follows. At first, the problem definition is introduced. Subsequently, the selected soft-sensor design methods are reviewed. Next, the description of the results is stated. Finally, the main findings and conclusions are summarized.

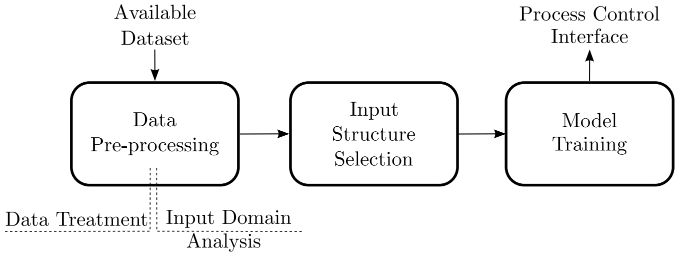

Fig. 1. The diagram of the soft-sensor development.

## II. Problem Definition

The procedure of the data-driven soft-sensor design is illustrated in Figure 1. It can be divided into three stages:

1) *Data Pre-processing:* The available (industrial) dataset is subjected to the data treatment analysis [10] to indicate systematic errors and outliers. Moreover, the input domain analysis [11], [12], focused on exploring the potential of nonlinear transformations, can be performed.
2) *Input Structure Selection:* The best subset of the input variables is selected to minimize the computation load of subsequent calculations while maintaining the information content of the dataset from data pre-processing.
3) *Model Training:* The model parameters are calculated to minimize the discrepancy between measured and estimated values within the training dataset.

The procedure for developing a soft sensor, as shown in Fig. 1, has its limitations. One such limitation is the non-transparent effect of data pre-processing (1st block) and subsequent input structure selection (2nd block) on the model training (3rd block) outcome. In this paper, we aim to address this drawback by comparing the performance of soft sensors while considering various approaches at different stages of the design procedure shown in Fig. 1.

Performance of the soft sensor is highly related to the quality of the design methodology. We briefly describe a few well-known data-driven approaches capable of solving particular challenges of soft-sensor design.

### A. Input Domain Analysis

The input dataset for the soft-sensor design consists of the particular number of input variables ($n_\text{v}$) and measurements ($n$). The original input dataset $X \in \mathbb{R}^{n \times n_\text{v}}$ should involve only the input variables directly measured by other sensors. The purpose of this phase of soft-sensor design is to enhance the original input dataset with new variables that can increase the potential of the input dataset to explain the output variable.

The straightforward way to enrich the input domain is to calculate additional input variables from the original variables using simple nonlinear functions (transformations) as follows:

$$X_\text{ad} = \left(x_i^2, x_i^3, x_i x_j, \ldots\right), \, \forall i, j = \{1, \ldots, n_\text{v}\}, \, i > j, \quad (1)$$

where $x$ is a particular input variable, $X_\text{ad} \in \mathbb{R}^{n \times n_\text{ad}}$ is a dataset with additional variables and $n_\text{ad}$ is a number of additional variables. The nonlinear transformations in (1) represent just a sample of the all possible configurations.

The nonlinear character of the additional variables should avoid the occurrence of linear dependencies within the joint dataset (original dataset with additional variables). The purpose of considering the nonlinear transformations is to linearize the possible nonlinear behavior of the estimated (output) variable. From a practical point of view, this can increase the predictive potential of the linear soft sensors, if nonlinear behavior is involved with the occurring phenomena [3]. The advantage of linear soft sensors is their transparent (simple) structure and low computational demands compared to nonlinear soft sensors. The similar principle of incorporating non-linear transformations of the independent variables is considered by the ALAMO approach [11].

It is reasonable to compare the performance of the soft-sensor designed according to the original dataset and the joint dataset (original dataset with additional variables) in order to properly indicate the contribution of the additional variables.

### B. Input Structure Selection

The purpose of the input structure selection is to effectively reduce the complexity (dimensionality) of the input datasets from the input domain analysis (see in Section II-A). The input dataset may contain variables with low informative content or redundant (linearly dependent) variables, which reduce the overall applicability of the dataset. Therefore, it is necessary to eliminate these variables or derive new informative variables before the dataset is used for the training of the soft sensors.

The popular and frequently used unsupervised learning method for reducing the dimensionality of datasets is the principal component analysis (PCA) [3]. This approach transforms the data into a new coordinate (principal component) space, allowing the explanation of most of the variance in the original space by a few variables from the principal component space. Depending on the selection of the principal components in the soft-sensor design, we can distinguish between the following representative approaches:

- PCA-1PC: Only the principal component, explaining the most variance, is selected for the soft-sensor design.
- PCA-Elb: The selection of the principal components is based on applying the Elbow method in the scree plot.

Another way of the input structure selection represents the family of the feature selection (FS) approaches [13]. These supervised learning approaches search for the best subset of variables regarding the pre-specified objectives. The popular FS approaches are Univariate Feature Selection (UFS) [23], Recursive Feature Elimination (RFE) [24] and Sequential Feature Selection (SFS) [23].

The use of an unsupervised-learning approach, such as PCA, is beneficial when the output variable is measured less frequently than the input variables. This is often the case in industrial soft-sensor design, where measuring the output (desired) variable is expensive and rare. In contrast, supervised-learning approaches like UFS, RFE, or SFS, incorporate output variable measurements when searching for the optimal input structure. This can be particularly advantageous when the output variable exhibits a complex, nonlinear nature.

### C. Model Training

The final step of soft-sensor development is the calculation of the model parameters given the training dataset. The computational load of this step is given by the quality (i.e., noise significance or linear dependencies) and complexity (i.e., number of variables and measurements) of the input dataset.

The design of industrial soft sensors is usually limited to structures linear in parameters. In this case, the following methods represent a reasonable choice:

- Ordinary Least Squares (OLS) minimizes the sum of squared errors between measured and estimated values.
- Partial Least Squares (PLS) [25], [26] reduces the set of input variables with a smaller set of uncorrelated components and performs OLS on these components.
- LASSO [27], [28] extends the objective function of OLS about $l_1$-penalization part reducing the model complexity.

The aforementioned methods can only learn the linear relationship between the input and output variables. To be able to learn nonlinear models, more advanced techniques are required. One of these techniques is Artificial Neural Networks (ANN) [29], [30]. The ANN structure consists of an input layer, hidden layers, and an output layer. Each layer has a specific number of neurons and one type of activation function. The activation function is the source of the nonlinearity within ANN. The disadvantage of using ANN for regression over linear approaches (OLS, PLS, and LASSO) is that it requires more data and reduces the transparency of the model structure.

Another way of designing models represents Gaussian Process (GP) [21], [31]. Unlike the previous methods, GP is a non-parametric approach, and therefore, it does not require the specification of any parameters to make the prediction. Non-parametric approaches should be used to provide a more accurate solution at a significantly higher computational load than parametric approaches. The main advantage of GP compared to other nonlinear structures (e.g., ANN) is the ability to provide uncertainty estimates.

## III. Case Studies

### A. Implementation Details

The presented design methods are implemented in Python 3.10.0. The approaches for input structure selection (see in Section II-B) are provided by *scikit-learn* and *mlxtend* libraries. The base for the model training approaches (see in Section II-C) are *scikit-learn* (OLS, PLS, LASSO), *keras* (ANN), and *GPy* (GP) libraries.

The design of the soft sensors is executed on the training dataset, while the testing (unseen) dataset is used for the subsequent comparison of the designed soft sensors. The training and testing datasets are assumed to be of equal size.

The number of available measurements from the studied industrial case studies limits the input domain analysis (see in Section II-A). To ensure a fair comparison, the following nonlinear transformations are considered in each case study:

$$X_{\text{ad}} = (x_i^2, x_i^3, \ln(x_i)), \quad \forall i = \{1, 2, \dots, n_{\text{v}}\}. \tag{2}$$

The studied ANN for the design of soft sensors in the following case studies consists of two layers, where the first layer involves a rectified linear unit activation function with 20 neurons and the second layer includes hyperbolic tangent activation functions with 20 neurons. The structure of the studied GP involves a squared-exponential kernel.

The complexity of the input dataset and the accuracy of the estimates determine the efficiency of the studied soft sensors. The representative criterion of the input dataset complexity is the number of input variables ($n_{\text{v}}$) selected by the concerned approaches (i.e., PCA-1PC, PCA-Elb, UFS, RFE, and SFS). Note that the number of input variables ($n_{\text{v}}$) stands for the number of principal components in the case of PCA-1PC and PCA-Elb. The accuracy of studied soft sensors is evaluated by the root mean squared error (RMSE) criterion.

### B. Soft-sensor Design for Pressure Compensated Temperature

The pressure compensated temperature ($PCT$) is a phenomenological variable frequently used in low-pressure distillation columns [3], [32]. It can be derived as a combination of the Antoine and Clausius-Clapeyron equations [1]:

$$\frac{1}{PCT} = \frac{R}{H_v} \ln\left(\frac{P}{P_{\text{ref}}}\right) + \frac{1}{T}, \tag{3}$$

where $H_v$ is the heat of vaporization, $R$ is the universal gas constant, $P_{\text{ref}}$ is the reference pressure, $P$ is the absolute pressure, and $T$ is the absolute temperature.

The parameters in the $PCT$ model are taken from [32]. The output variable to be estimated is $PCT$, while the input variables are $P$ and $T$. We consider the presence of noise in $PCT$ with $\sigma^2_{\text{noise}} = 1$. Subsequently, we generate 200 data points from the $PCT$ model, with 100 points drawn from $\mathcal{N}([10, 545], \text{diag}([1, 5]))$ and the remaining points drawn from $\mathcal{N}([3, 545], \text{diag}([1, 5]))$. We consider two different operating regimes for $PCT$ to incorporate more of its nonlinear behavior into the soft-sensor design. The data points are randomly divided into training and testing sets.

The results in Table I indicate the number of input variables $n_{\text{v}}$ and accuracy (RMSE) of designed soft sensors on the original $PCT$ dataset without nonlinear transformations. Table I provides the comparison of studied input structure selection approaches (i.e., PCA-1PC, PCA-Elb, UFS, RFE, and SFS) in combination with the model training approaches (i.e., OLS, PLS, LASSO, ANN and GP). The results presented in Table I are divided into training and testing datasets. The accuracy of both datasets indicates that the input dataset provided by PCA-1PC is insufficiently informative. The main cause of the low performance of PCA-1PC is that the first principal component explains only 76.6 % of the original dataset variance. This is expected since the variables within the input dataset are not linearly dependent. The input datasets provided by PCA-Elb, UFS, RFE, and SFS seem to be the same. The performance of these approaches is limited due to the simple input dataset (only two input variables) in this case study.

The accuracy on the testing dataset (see in Table I) indicates that the nonlinear soft sensors (ANN and GP) are more

TABLE I
COMPARISON OF THE COMPLEXITY ($n_{\mathrm{v}}$) AND ACCURACY (RMSE) OF THE SOFT SENSORS DESIGNED ON THE $PCT$ DATASET.

| Input Structure | $\boldsymbol{n}_{\mathrm{v}}$ | **RMSE** | | | | |
|---|---|---|---|---|---|---|
| | | **OLS** | **PLS** | **LASSO** | **ANN** | **GP** |
| | | *Training (Testing) Dataset* | | | | |
| PCA-1PC | 1 | 0.056<br>(0.084) | 0.056<br>(0.084) | 0.056<br>(0.084) | 0.032<br>(0.048) | 0.031<br>(0.044) |
| PCA-Elb | 2 | 0.048<br>(0.077) | 0.048<br>(0.077) | 0.048<br>(0.077) | 0.004<br>(0.02) | 0.003<br>(0.014) |
| UFS,<br>RFE, SFS | 2 | 0.048<br>(0.077) | 0.048<br>(0.077) | 0.048<br>(0.077) | 0.004<br>(0.018) | 0.003<br>(0.014) |

TABLE II
COMPARISON OF THE COMPLEXITY ($n_{\mathrm{v}}$) AND ACCURACY (RMSE) OF THE SOFT SENSORS DESIGNED ON THE EXTENDED $PCT$ DATASET.

| Input Structure | $\boldsymbol{n}_{\mathrm{v}}$ | **RMSE** | | | | |
|---|---|---|---|---|---|---|
| | | **OLS** | **PLS** | **LASSO** | **ANN** | **GP** |
| | | *Training (Testing) Dataset* | | | | |
| PCA-1PC | 1 | 0.064<br>(0.094) | 0.064<br>(0.094) | 0.064<br>(0.094) | 0.033<br>(0.048) | 0.034<br>(0.046) |
| PCA-Elb | 3 | 0.014<br>(0.027) | 0.014<br>(0.027) | 0.014<br>(0.027) | 0.003<br>(0.009) | 0.003<br>(0.006) |
| UFS | 8 | 0.005<br>(0.009) | 0.011<br>(0.021) | 0.005<br>(0.009) | 0.003<br>(0.009) | 0.003<br>(0.005) |
| RFE | 7 | 0.005<br>(0.009) | 0.008<br>(0.015) | 0.005<br>(0.009) | 0.003<br>(0.01) | 0.003<br>(0.005) |
| SFS | 4 | 0.005<br>(0.008) | 0.008<br>(0.015) | 0.005<br>(0.008) | 0.003<br>(0.014) | 0.003<br>(0.005) |

accurate compared to the linear ones (OLS, PLS, and LASSO). This is due to the nonlinear character of the estimated variable. On the other hand, the accuracy of ANN and GP with any input structure selection approach (except PCA-1PC) on the training dataset (Table I) suggests the overfitting of these approaches. This observation is confirmed by the increased discrepancy between the accuracy of these approaches on the training and testing datasets. Furthermore, the complexity of the ANN and GP models (Section III-A) increases the likelihood of overfitting. We can conclude that the combination of GP with any input structure selection approach (except PCA-1PC) is the best option for the soft sensor design on the $PCT$ dataset, considering the accuracy of the soft sensors.

Subsequently, we extend the original dataset with the nonlinear transformations from (2). The results using this dataset are shown in Table II. The low accuracy of the soft sensors designed according to PCA-1PC confirms that the provided input dataset is not sufficiently informative. The first principal component explains 71.3 % of the extended dataset variance. The explained variance is decreased compared to the original dataset, which is caused by extra variance from the nonlinear transformations. According to Table II, it seems that the rest of the designed soft sensors achieved higher accuracy compared to the results in Table I. We can see that the highest accuracy is achieved by GP in combination with any of the feature selection approaches. However, the nonlinear soft sensors (ANN and GP) seem to be more overfitted compared to the linear soft sensors (OLS, PLS, and LASSO). We can indicate this by comparing the accuracy of the soft sensors on the training dataset (Table II). The increased tendency of ANN and GP to overfit the models is caused by the complexity (see in Section III-A) of designed structures. The linear soft sensors (OLS and LASSO) achieve nine times better accuracy compared to the original dataset (Table I), which can compete with the accuracy of the nonlinear soft sensors (ANN, GP). The accuracy of PLS-designed soft sensors (Table II) is lower than that of OLS and LASSO. It appears that PLS further reduces the complexity of the input dataset provided by UFS, RFE, and SFS at the expense of accuracy.

The comparison of the results from the original $PCT$ dataset (Table I) and the extended $PCT$ dataset (Table II) leads us to conclude that the consideration of the nonlinear transformations can be beneficial for the soft-sensor design, especially for the linear soft sensors. The higher accuracy of the nonlinear soft sensors is supported by the generated dataset from the nonlinear $PCT$ model. Moreover, the dataset involves Gaussian noise with a small variance, which increases the potential of the complex models. It appears that the performance of the soft sensors, especially the nonlinear ones, is highly related to the quality of the available dataset. In order to validate the achieved conclusions and findings, we design soft sensors for two different (non-ideal) industrial datasets.

### C. Soft-sensor Design for Depropanizer Column

The depropanizer is a distillation column that separates the feed mixture of nine hydrocarbons C3–C5 into C3-fraction-rich distillate and C4/C5-fraction-rich bottom product [3]. This column is a part of the Fluid Catalytic Cracking unit in Slovnaft, a.s. in Bratislava, Slovakia.

The available industrial dataset involves measurements from December 2016 to October 2018 (22 months). The output variable (concentration of the main impurity in the bottom product) is measured only once every four days, and therefore, there are 165 measurements available for the soft-sensor design. The original input dataset consists of 18 variables, of which 16 are directly measured by online sensors and two are derived as ratios (nonlinear transformation) of two input variables. The available measurements are chronologically distributed among the training and testing datasets.

The complexity ($n_{\mathrm{v}}$) and accuracy (RMSE) of designed inferential sensors on the depropanizer dataset are compared in Table III. The accuracy of the designed soft sensors by PCA-1PC is decreased on the testing dataset compared to other soft sensors (except RFE with OLS). This suggests that the provided input dataset by PCA-1PC is not enough informative as we could see in the previous case study (see in Table I). Moreover, the first principal component explains only 49.5 % of the original dataset variance. The accuracy of the soft sensors designed by PCA-Elb is improved compared to PCA-1PC. We can see that the combinations of PCA-Elb with OLS, PLS, and LASSO outperform the other approaches on

TABLE III

Comparison of the complexity ($n_{\mathrm{V}}$) and accuracy (RMSE) of the soft sensors designed on the depropanizer dataset.

| Input Structure | $\boldsymbol{n}_{\mathrm{V}}$ | **RMSE** | | | | |
|---|---|---|---|---|---|---|
| | | **OLS** | **PLS** | **LASSO** | **ANN** | **GP** |
| | | *Training (Testing) Dataset* | | | | |
| PCA-1PC | 1 | 0.162<br>(0.222) | 0.162<br>(0.222) | 0.162<br>(0.222) | 0.155<br>(0.212) | 0.158<br>(0.218) |
| PCA-Elb | 3 | 0.142<br>(0.168) | 0.142<br>(0.168) | 0.142<br>(0.168) | 0.132<br>(0.182) | 0.141<br>(0.17) |
| UFS | 14 | 0.112<br>(0.12) | 0.113<br>(0.117) | 0.12<br>(0.143) | 0.104<br>(0.135) | 0.115<br>(0.137) |
| RFE | 9 | 0.113<br>(0.245) | 0.117<br>(0.128) | 0.116<br>(0.127) | 0.112<br>(0.152) | 0.118<br>(0.135) |
| SFS | 3 | 0.118<br>(0.111) | 0.118<br>(0.111) | 0.118<br>(0.111) | 0.116<br>(0.111) | 0.118<br>(0.111) |

TABLE IV

Comparison of the complexity ($n_{\mathrm{V}}$) and accuracy (RMSE) of the soft sensors designed on the VGH dataset.

| Input Structure | $\boldsymbol{n}_{\mathrm{V}}$ | **RMSE** | | | | |
|---|---|---|---|---|---|---|
| | | **OLS** | **PLS** | **LASSO** | **ANN** | **GP** |
| | | *Training (Testing) Dataset* | | | | |
| PCA-1PC | 1 | 0.114<br>(0.168) | 0.114<br>(0.168) | 0.114<br>(0.168) | 0.094<br>(0.192) | 0.086<br>(0.235) |
| PCA-Elb | 6 | 0.089<br>(0.152) | 0.089<br>(0.152) | 0.089<br>(0.152) | 0.078<br>(0.15) | 0.064<br>(0.17) |
| UFS | 4 | 0.08<br>(0.126) | 0.08<br>(0.126) | 0.08<br>(0.122) | 0.079<br>(0.124) | 0.056<br>(0.154) |
| RFE | 4 | 0.107<br>(0.205) | 0.115<br>(0.184) | 0.107<br>(0.205) | 0.114<br>(0.186) | 0.073<br>(0.268) |
| SFS | 3 | 0.079<br>(0.104) | 0.079<br>(0.104) | 0.079<br>(0.104) | 0.08<br>(0.102) | 0.063<br>(0.123) |

the testing dataset (Table III). The performance of soft sensors designed by UFS and RFE is comparable (except for RFE with OLS) on the testing dataset. These similarities come from the complex input structure suggested by these approaches. The highest accuracy of the soft sensors is achieved with the input structure provided by SFS. Moreover, SFS determines a less complex input structure than UFS and RFE.

According to the results in Table III, the combination of SFS with any of the approaches for model training (i.e., PCA-1PC, PCA-Elb, UFS, RFE, and SFS) is the best option for the depropanizer column. Note that the RMSE values of SFS on the testing dataset (Table III) are not the same. This effect is caused by rounding the RMSE values to three decimal places, but the most accurate approach is GP. Nevertheless, the soft sensors designed by OLS, PLS, and LASSO possess a linear structure that is less complex and more transparent compared to ANN and GP. This leads us to conclude that SFS with OLS, PLS, or LASSO is the most efficient for the depropanizer dataset. Moreover, this proves the conclusion from Section III-B that the nonlinear transformations within the input dataset can increase the efficacy of the linear soft sensors. On the other hand, the performance of the nonlinear soft sensors is significantly decreased compared to the $PCT$ dataset (Section III-B) resulting from the non-ideal character (e.g., non-Gaussian noise) of the depropanizer dataset.

### D. Soft-sensor Design for Vacuum Gasoil Hydrogenation Unit

The Vacuum Gasoil Hydrogenation (VGH) unit is an essential part of the Slovnaft, a.s. refinery, which consists of high-pressure reaction and low-pressure fractionation sections [3]. The desired soft sensor should estimate the concentration of the particular product from the product fractionator (distillation column) located in the low-pressure fractionation.

The industrial dataset available spans from January 2018 through December 2019 (24 months). The output variable is measured approximately once per day, resulting in 621 measurements available for soft-sensor design. The input dataset comprises 35 variables, 33 of which are measured by online sensors from both sections of the VGH unit. Two variables represent nonlinear transformations of the input variables, which are two $PCT$s (as seen in Section III-B) derived for the product fractionator. The available measurements are chronologically assigned to the training and testing datasets.

The results in Table IV compare the performance of the studied variable selection approaches (i.e., PCA-1PC, PCA-Elb, UFS, RFE, and SFS). We can see that the PCA-1PC approach does not provide a sufficiently informative input dataset, which causes the low accuracy of all designed soft sensors on the testing dataset. Moreover, the first principal component explains only 38.6 % of the original dataset variance. The input structure of PCA-Elb involves five extra principal components compared to PCA-1PC. The accuracy of designed soft sensors is increased on the training dataset but relatively low on the testing dataset. Therefore, it appears that PCA-Elb provided too complex input structure, causing the overfitting of the designed soft sensors. It appears that RFE selected the same number of input variables as UFS. Nevertheless, the soft sensors designed by RFE are less accurate compared to the UFS ones. This suggests that RFE indicated inappropriate input variables. The results in Table IV show that the SFS approach provided a less complex yet more efficient input structure compared to UFS and RFE.

The comparison of the studied model training approaches (i.e., OLS, PLS, LASSO, ANN, and GP) is presented in Table IV. The results show that the most accurate combination is SFS with ANN, but the accuracy of linear soft sensors (OLS, PLS, and LASSO) can also compete with ANN. Therefore, selecting the best soft sensor for the VGH unit requires a compromise between complexity and accuracy. The similar accuracy of the linear and nonlinear soft sensors confirms that nonlinear transformations within the input dataset can increase the performance of linear soft sensors. However, the performance of nonlinear soft sensors is decreased compared to those designed on the $PCT$ dataset (Section III-B). This suggests that the quality of the data from VGH is lower compared to $PCT$. The accuracy of soft sensors designed by GP indicates overfitting, as they achieve high accuracy on the training dataset but lower accuracy on the testing dataset.

## IV. Conclusions

The contribution investigated the effectiveness of combining various approaches in the soft-sensor development procedure. Several soft sensors were designed and compared using datasets from petrochemical case studies. The increasing complexity of the considered case studies established the validity and applicability of the findings. In general, nonlinear soft sensors achieved higher accuracy on the testing dataset compared to linear ones. Linear soft sensors could compete with nonlinear ones if nonlinear transformations of the original input variables were incorporated into the input dataset. Additionally, the performance of nonlinear approaches was significantly affected by the presence of noise, especially in real industrial case studies. The input structure selection analysis indicates that unsupervised learning approaches provided a less effective input structure compared to supervised learning approaches. This was caused by the nonlinear nature of the output variables in the considered petrochemical case studies.

## Acknowledgment

This research is funded by the Slovak Research and Development Agency under the projects APVV-21-0019 and APVV SK-FR-2019-0004, by the Scientific Grant Agency of the Slovak Republic under the grants VEGA 1/0691/21 and VEGA 1/0297/22, and by the European Commission under the grant no. 101079342 (Fostering Opportunities Towards Slovak Excellence in Advanced Control for Smart Industries).

## References

[1] M. King, *Process Control: A Practical Approach*. John Wiley & Sons Ltd., 2011.

[2] F. Curreri, S. Graziani, and M. G. Xibilia, "Input selection methods for data-driven soft sensors design: Application to an industrial process," *Information Sciences*, vol. 537, pp. 1–17, 2020.

[3] M. Mojto, K. L'ubušký, M. Fikar, and R. Paulen, "Data-based design of inferential sensors for petrochemical industry," *Computers & Chemical Engineering*, vol. 153, p. 107437, 2021.

[4] M. Jia, D. Xu, T. Yang, Y. Liu, and Y. Yao, "Graph convolutional network soft sensor for process quality prediction," *Journal of Process Control*, vol. 123, pp. 12–25, 2023.

[5] L. Hua, C. Zhang, W. Sun, Y. Li, J. Xiong, and M. S. Nazir, "An evolutionary deep learning soft sensor model based on random forest feature selection technique for penicillin fermentation process," *ISA Transactions*, 2022.

[6] K. Qiu, J. Wang, X. Zhou, R. Wang, and Y. Guo, "Soft sensor based on localized semi-supervised relevance vector machine for penicillin fermentation process with asymmetric data," *Measurement*, vol. 202, p. 111823, 2022.

[7] A. S. Rathore, S. Nikita, and N. G. Jesubalan, "Digitization in bioprocessing: The role of soft sensors in monitoring and control of downstream processing for production of biotherapeutic products," *Biosensors and Bioelectronics: X*, vol. 12, p. 100263, 2022.

[8] L. Li, N. Li, X. Wang, J. Zhao, H. Zhang, and T. Jiao, "Multi-output soft sensor modeling approach for penicillin fermentation process based on features of big data," *Expert Systems with Applications*, vol. 213, p. 119208, 2023.

[9] F. Zhang, N. Li, L. Li, S. Wang, and C. Du, "A local semi-supervised ensemble learning strategy for the data-driven soft sensor of the power prediction in wind power generation," *Fuel*, vol. 333, p. 126435, 2023.

[10] Q. Liu, M. Jia, Z. Gao, L. Xu, and Y. Liu, "Correntropy long short term memory soft sensor for quality prediction in industrial polyethylene process," *Chemometrics and Intelligent Laboratory Systems*, vol. 231, p. 104678, 2022.

[11] A. Cozad, N. V. Sahinidis, and D. C. Miller, "Learning surrogate models for simulation-based optimization," *AIChE Journal*, vol. 60, no. 6, pp. 2211–2227, 2014.

[12] J. Ferreira, M. Pedemonte, and A. I. Torres, "Development of a machine learning-based soft sensor for an oil refinery's distillation column," *Computers & Chemical Engineering*, vol. 161, p. 107756, 2022.

[13] H. Liu, *Feature Selection*. Boston, MA: Springer US, 2010, pp. 402–406.

[14] P. Wang, B. Xue, J. Liang, and M. Zhang, "Feature selection using diversity-based multi-objective binary differential evolution," *Information Sciences*, vol. 626, pp. 586–606, 2023.

[15] M.-R. Yang and Y.-W. Wu, "A cross-validated feature selection (CVFS) approach for extracting the most parsimonious feature sets and discovering potential antimicrobial resistance (AMR) biomarkers," *Computational and Structural Biotechnology Journal*, vol. 21, pp. 769–779, 2023.

[16] J. Liang, C. Wang, D. Zhang, Y. Xie, Y. Zeng, T. Li, Z. Zuo, J. Ren, and Q. Zhao, "VSOLassoBag: a variable-selection oriented lasso bagging algorithm for biomarker discovery in omic-based translational research," *Journal of Genetics and Genomics*, 2023.

[17] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, 1901.

[18] D. Meng, Q. Zhao, and Z. Xu, "Improve robustness of sparse PCA by L1-norm maximization," *Pattern Recognition*, vol. 45, no. 1, pp. 487–497, 2012.

[19] S. Sharifzadeh, A. Ghodsi, L. H. Clemmensen, and B. K. Ersbøll, "Sparse supervised principal component analysis (SSPCA) for dimension reduction and variable selection," *Engineering Applications of Artificial Intelligence*, vol. 65, pp. 168–177, 2017.

[20] J. Zheng, L. Ma, Y. Wu, L. Ye, and F. Shen, "Nonlinear dynamic soft sensor development with a supervised hybrid CNN-LSTM network for industrial processes," *ACS Omega*, vol. 7, no. 19, pp. 16 653–16 664, 2022.

[21] S. Yamakage and H. Kaneko, "Design of adaptive soft sensor based on bayesian optimization," *Case Studies in Chemical and Environmental Engineering*, vol. 6, p. 100237, 2022.

[22] W. Shao, W. Han, Y. Li, Z. Ge, and D. Zhao, "Enhancing the reliability and accuracy of data-driven dynamic soft sensor based on selective dynamic partial least squares models," *Control Engineering Practice*, vol. 127, p. 105292, 2022.

[23] J. Hua, W. D. Tembe, and E. R. Dougherty, "Performance of feature-selection methods in the classification of high-dimension data," *Pattern Recognition*, vol. 42, no. 3, pp. 409–424, 2009.

[24] I. R. Subramanian, J. Weston, S. D. Barnhill, and V. N. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.

[25] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, III, "The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, no. 3, pp. 735–743, 1984.

[26] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001.

[27] F. Santosa and W. W. Symes, "Linear inversion of band-limited reflection seismograms," *SIAM Journal on Scientific and Statistical Computing*, vol. 7, no. 4, pp. 1307–1330, 1986.

[28] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 273–282, 2011.

[29] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.

[31] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 11 2005.

[32] M. Mojto, K. L'ubušký, M. Fikar, and R. Paulen, "Support vector machine-based design of multi-model inferential sensors," in *32nd European Symposium on Computer Aided Process Engineering*, ser. Computer Aided Chemical Engineering, L. Montastruc and S. Negny, Eds. Elsevier, 2022, vol. 51, pp. 1045–1050.