# Data-Driven Indication of Flooding in an Industrial Debutanizer Column

Martin Mojto[a,*], Karol L'ubušký[b], Miroslav Fikar[a] and Radoslav Paulen[a]

[a]*Slovak University of Technology in Bratislava, Bratislava, Slovakia*
[b]*Slovnaft, a.s., Bratislava, Slovakia*
*martin.mojto@stuba.sk*

## Abstract

The profitability and sustainability of process industries are affected by the performance of each unit involved. A key measure of a unit's performance is based on whether it operates in a desired production window or whether it trips into an abnormal condition. In this contribution, we study flooding of industrial distillation columns. We aim to improve the performance of an industrial debutanizer column by designing a data-driven flooding indicator. The design of the indicator consists of three steps; (a) the data treatment, (b) a priori labeling, and (c) indicator design. The prior knowledge about flooding within the column is used to design a reference indicator. This knowledge is either unused or fully exploited during the design of the indicator. We compare various design methods and show the potential of data-driven approaches for flooding indication.

**Keywords:** Debutanizer Column, Flooding Indicator, Soft Sensor, Subset Selection

## 1. Introduction

Flooding is an undesired phenomenon in industrial distillation columns. It occurs when the liquid level rises above a tray, because of foaming or excessive downcomer fill-up (King, 2016). This state causes a significant loss in tray separation efficiency and, hence, plant profitability. Early detection (prediction) of flooding is thus crucial for a profitable and sustainable plant.

Several works dealt with the problem of flooding detection. They considered correlation of the flooding effect with internal process variables, especially the pressure difference (drop) across the column (Peiravan et al., 2020) and the time derivative of the pressure drop (Pihlaja, 2008). Industrial experts use these results and combine them often with an insight into the principal trigerring cause of flooding, creating a tailored solution for each column. The effort of creating a tailored solution could be saved by the use of machine learning (ML) approaches.

Several ML approaches (Mojto et al., 2021; Oeing et al., 2021; Fuentes-Cortés et al., 2022) were employed to aid decision making in industrial columns. A subset of unsupervised ML approaches, such as $k$-means clustering (Forgy, 1965) or principal component analysis (PCA) (Pearson, 1901), consider no prior knowledge about the model outcome. The supervised ML techniques, on contrary, use knowledge about the desired outcome for training. The representative methods are subset (feature) selection (SS) (Smith, 2018) or support vector machine (SVM) (Boser et al., 1992).

This paper investigates the design of data-driven flooding indicators for an industrial debutanizer. Performance of indicators designed via data-driven approaches (unsupervised and supervised ML)

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

is assessed by comparison against the reference indicator. The reference indicator (considered as ground truth) is designed according to the industrial specifications and knowledge about flooding.

## 2. Problem Statement

Flooding indication is essentially a binary classification problem. We aim to design an indicator $\mathbf{I}$ that assigns a categorical label $\hat{y}$ according to the classification model (classifier) $f(x)$ as:

$$\hat{y} = \begin{cases} +1 \text{ (flooding)}, & \text{if } f(x) \geq 0, \\ -1 \text{ (normal operation)}, & \text{if } f(x) < 0, \end{cases} \tag{1}$$

where $x \in \mathbb{R}^{n_\mathrm{p}}$ represents a subset (sparse representation, $n > n_\mathrm{p}$) of all online plant measurements $\xi \in \mathbb{R}^{n}$ at one time instant. In this contribution, we consider a linear classifier in the form:

$$f(x) = w^\intercal x + w_0, \tag{2}$$

where $w \in \mathbb{R}^{n_\mathrm{p}}$ represents a classifier normal vector and $w_0$ is a classifier off-set.

### *2.1. Industrial Debutanizer Column*

We study a debutanizer (distillation) column that is a part of the FCC unit of the refinery Slovnaft, a.s. in Bratislava, Slovakia. The column separates the C4/C5 fraction into the C4-fraction-rich distillate product and the C5-fraction-rich bottom product. The column contains 40 trays.

The available dataset involves the measurements from January 2019 to April 2021 (28 months). The input variables are recorded every minute by online sensors, yet their 30-minute moving average values are considered in this study. Overall, the dataset involves 34,297 measurements. The measurements from two plant shutdowns (May – July 2019 and December 2020) are excluded.

The following 41 input variables are directly measured (by online sensors) at the column:

$$\begin{aligned} \xi = (\, & vo_\mathrm{B}, vo_\mathrm{D,1–3}, vo_\mathrm{R}, vo_\mathrm{reb,h}, T_\mathrm{col,1–5}, T_\mathrm{B,1–2}, T_\mathrm{D,1–3}, T_\mathrm{F}, T_\mathrm{reb,h,1–2}, Q_\mathrm{con}, \\ & p_\mathrm{col}, p_\mathrm{D,1–4}, p_\mathrm{df,col}, p_\mathrm{con}, F_\mathrm{R}, F_\mathrm{B}, F_\mathrm{D,1–4}, F_\mathrm{F}, F_\mathrm{reb,h,1–2}, L_\mathrm{reb,1–2}, L_\mathrm{con,1–3}\,), \end{aligned} \tag{3}$$

where *vo*, $T$, $Q$, $p$, and $F$ stand for a valve opening, temperature, heat input, pressure, and flow rate, respectively. Indices B, D, F, R, col, con, reb, and df represent a bottom section, distillate section, feed section, reflux section, column section, condenser section, reboiler section, and cross-column difference, respectively. Note that exact location of sensors cannot be disclosed due to the confidentiality reasons. The input set of directly measured variables is extended with important ratios ($F_\mathrm{R}/F_\mathrm{F}$, $F_\mathrm{B}/F_\mathrm{F}$, $Q_\mathrm{con}/F_\mathrm{F}$) and pressure compensated temperatures ($PCT_\mathrm{B}$, $PCT_\mathrm{D}$).

The studied debutanizer column usually operates within the desired operating regime. At times, however, the operating conditions within the unit induce flooding. The envisioned low-cost solution to the flooding problem is to design a reliable indicator. The key aspect of this approach is that the designed indicator is not only used for monitoring the plant condition, but it can communicate directly with the advanced process controller that can provide a fast response.

The dataset does not contain any direct indication of flooding that could be used to label the data. However, it is possible to attribute flooding occurrence to the increased values of $p_\mathrm{df,col}$, $F_\mathrm{R}$, $F_\mathrm{reb,h,2}$, and $T_\mathrm{col,4}$ and decreased values of $T_\mathrm{reb,h,1}$. We use this knowledge to design the reference indicator to provide the ground truth of the flooding indicator for our study.

## 3. Methodology

Data-driven indicators are designed using unsupervised ($\mathbf{I}^\mathrm{Uns}$) and supervised ($\mathbf{I}^\mathrm{Sup}$) ML approaches. For training ($\mathbf{I}^\mathrm{Sup}$-type indicator) and testing, the ground truth is provided by the aforementioned reference indicator resulting from industrial knowledge about debutanizer flooding.

### *3.1. Indicator Design*

The design procedure of the data-driven indicator consists of three sequential steps:

1. Data processing (data filtering, data treatment, distribution to training/testing dataset).
2. A priori labeling of the training dataset (only applied for $\mathbf{I}^{\text{Uns}}$-type indicators).
3. Training of a classifier (calculation of the $f(x)$ parameters on the labeled training dataset).

After the standardization of the data set (removing the mean and scaling all the variables to unit variance), the aim of the data treatment (the 1st step) is to reduce the number of outliers. Due to the non-ideal (yet close normal) noise distribution within the industrial dataset, the minimum covariance determinant (MCD) approach (Hubert and Debruyne, 2010) is applied. The outlier detection is performed using the $F$-distribution, retaining data with 99.9999 % probability. The high probability value follows from the need to eliminate only the most deviated measurements while maintaining the data representing the flooding, which can be otherwise seen as outliers.

It is optional to smoothen the dataset by filtering out the high-frequency noise that does not represent slower effects of flooding. Subsequently, as flooding is characterized by the changes of the process variables, we extend the dataset (here, 46 variables) by time differences of each variable:

$$\Delta\xi_i(k) = \xi_i(k) - \xi_i(k-1), \qquad \forall i = \{1,2,\ldots,n\}, \tag{4}$$

where $k$ is a time instant. The resulting dataset considers both, the original dataset and time differences, i.e., 92 variables in this study. Effectively, we assign $\xi \leftarrow (\xi,\ \Delta\xi)$ in this step.

The 2nd step, applied to label the data for $\mathbf{I}^{\text{Uns}}$-type approaches, is performed by $k$-means clustering (Forgy, 1965) with the elbow method to determine the optimal number of clusters. The clusters with a low cardinality but large distance between the cluster center and the dataset mean are considered to represent the debutanizer flooding.

The training phase needs to choose an appropriate indicator input space ($\mathbb{R}^{n_\text{p}}$) among all the process variables and their time differences. The methods used in this study are:

1. Industrial patent by (Pihlaja, 2008), which exploits $\Delta p_{\text{df, col}}$ only (referred to as $\mathbf{I}_{\text{pat}}$).
2. Industrial experience (specific to the studied debutanizer) using $\Delta p_{\text{df,col}}$, $\Delta F_{\text{R}}$, $\Delta F_{\text{reb,h,2}}$, $\Delta T_{\text{col,4}}$, and $\Delta T_{\text{reb,h,1}}$ (referred to as $\mathbf{I}_{\text{ref}}$).
3. PCA approach (Pearson, 1901) that chooses a number principal components that explain at least 95 % of variance in the dataset (referred to as $\mathbf{I}_{\text{PCA}}$).
4. SS approach (Smith, 2018), which determines the best subset of input variables via cross-validation and comparison of different structures with $n_\text{p} = \{1,2,\cdots,5\}$ (referred to as $\mathbf{I}_{\text{SS}}$).

The finalization of the training phase designs a linear classifier (see Eq. (2)) based on the chosen input structure ($x \in \mathbb{R}^{n_\text{p}}$). To this end, we use support vector machines (SVM) (Boser et al., 1992).

### *3.2. Performance Assessment*

The outcome of an indicator can fall into four categories: true positive (TP) and false positive (FP), when flooding is indicated correctly and incorrectly, respectively, and, vice versa, true negative (TN) and false negative (FN), for indicating of normal operation. We use some well-known normalized performance criteria for the designed indicators:

$$\text{AC} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}, \quad \text{PR} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \quad \text{RC} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \quad \text{F1} = \frac{2\times\text{PR}\times\text{RC}}{\text{PR}+\text{RC}}, \tag{5}$$

where AC (accuracy) is a measure of how often the classifier makes the correct prediction, PR (precision or correctness) is a measure of how precisely is the true prediction achieved, RC (recall or sensitivity) is a measure of how actual observations are predicted correctly. F1-score (F1) is a harmonic mean between PR and RC. In industrial conditions, it is much more important to warn about the potential of flooding and thus low value of FN (high RC) is preferred.

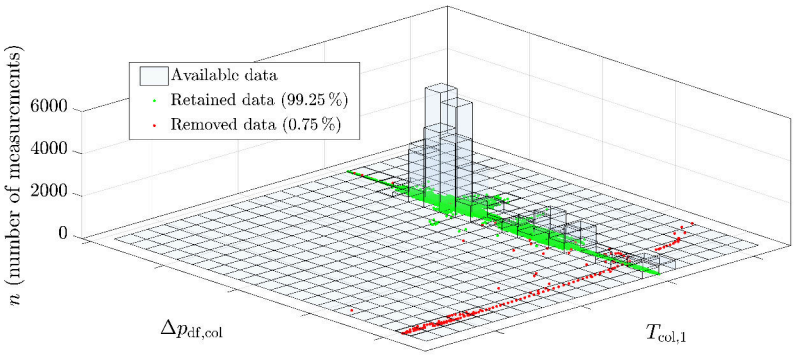


Figure 1: Histogram of two variables from the debutanizer dataset treated by the MCD method.

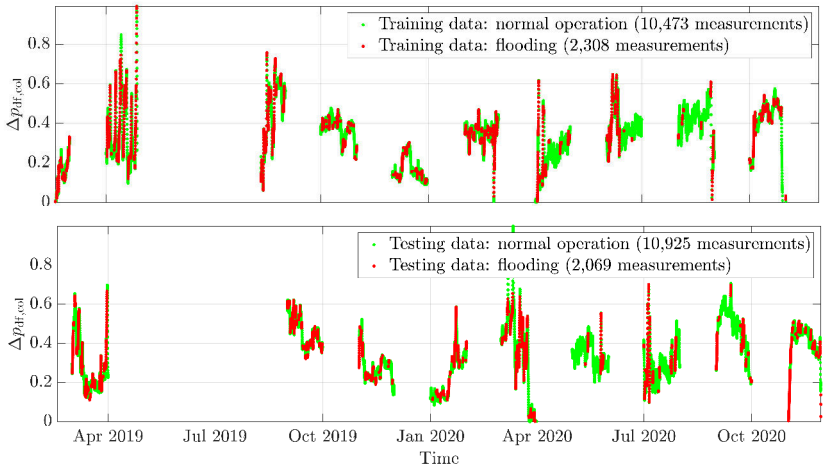


Figure 2: Visualization of training and testing datasets and ground truth labels.

## 4. Results

The results of data treatment using the MCD method are shown in Figure 1. The data values are anonymized for confidentiality reasons. As desired, only the most deviated measurements (0.75 %) are considered as outliers, and the rest of the measurements (99.25 %) is retained for further analysis. The dataset is further smoothened by filtering using a 10th-order low-pass Butterworth filter with a cut-off frequency of 0.028 mHz (with zero-phase distortion).

To guarantee fairness of indicator assessment, we distribute the retained data chronologically on an alternating monthly basis into the training and testing datasets (see Figure 2). From the entire dataset (25,775 measurement points), 12,781 and 12,994 points are assigned to the training and testing dataset, respectively. Figure 2 illustrates the training-testing data division together with (ground truth) labels assigned based on industrial experience with the reference indicator.

### *4.1. Training of Data-Driven Indicators*

Design of the data-driven flooding indicators for the debutanizer column is conducted via MATLAB based on the methods from Section 3.1. MATLAB built-in routines for *k*-means clustering, PCA, and SVM are exploited. We design indicators based on unsupervised ML ($\mathbf{I}^{\mathrm{Uns}}_{\mathrm{pat}}$, $\mathbf{I}^{\mathrm{Uns}}_{\mathrm{ref}}$, $\mathbf{I}^{\mathrm{Uns}}_{\mathrm{PCA}}$,

Table 1: The comparison of the true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), accuracy (AC), precision (PR), recall (RC), and complexity (no. of input variables $n_\mathrm{p}$, no. of principal components $n_\mathrm{pc}$) of the designed data-driven indicators on the testing dataset.

| ML method | Unsupervised learning | | | Supervised learning | | | |
|---|---|---|---|---|---|---|---|
| Structure | $\mathbf{I}^{\mathrm{Uns}}_{\mathrm{pat}}$ | $\mathbf{I}^{\mathrm{Uns}}_{\mathrm{ref}}$ | $\mathbf{I}^{\mathrm{Uns}}_{\mathrm{PCA}}$ | $\mathbf{I}^{\mathrm{Sup}}_{\mathrm{pat}}$ | $\mathbf{I}^{\mathrm{Sup}}_{\mathrm{ref}}$ | $\mathbf{I}^{\mathrm{Sup}}_{\mathrm{PCA}}$ | $\mathbf{I}^{\mathrm{Sup}}_{\mathrm{SS}}$ |
| TP | 1,784 | 1,704 | 618 | 1,192 | 2,031 | 1,720 | 2,029 |
| FP | 3,828 | 2,823 | 3,147 | 1,358 | 4 | 168 | 0 |
| TN | 7,097 | 8,102 | 7,778 | 9,567 | 10,921 | 10,757 | 10,925 |
| FN | 285 | 365 | 1,451 | 877 | 38 | 349 | 40 |
| AC | 68.3 | 75.5 | 64.6 | 82.8 | 99.7 | 96 | 99.7 |
| PR | 31.8 | 37.6 | 16.4 | 46.7 | 99.8 | 91.1 | 100 |
| RC | 86.2 | 82.4 | 29.9 | 57.6 | 98.2 | 83.1 | 98.1 |
| F1 | 46.5 | 51.7 | 21.2 | 51.6 | 99 | 86.9 | 99 |
| $n_\mathrm{p}/n_\mathrm{pc}$ | 1 | 5 | 17 | 1 | 5 | 17 | 2 |

$\mathbf{I}^{\mathrm{Uns}}_{\mathrm{SS}}$) and supervised ML ($\mathbf{I}^{\mathrm{Sup}}_{\mathrm{pat}}$, $\mathbf{I}^{\mathrm{Sup}}_{\mathrm{ref}}$, $\mathbf{I}^{\mathrm{Sup}}_{\mathrm{PCA}}$, $\mathbf{I}^{\mathrm{Sup}}_{\mathrm{SS}}$). A main difference between these approaches is the use of the $k$-means algorithm to classify the data (used for $\mathbf{I}^{\mathrm{Uns}}$ indicators).

A key success factor of unsupervised ML is an appropriate data labeling. The results indicate that, unsurprisingly, the best results are obtained when the $k$-means clustering is performed on a dataset with reduced dimensionality (e.g., one variable for $\mathbf{I}^{\mathrm{Uns}}_{\mathrm{pat}}$ indicator or seventeen principal components determined for $\mathbf{I}^{\mathrm{Uns}}_{\mathrm{PCA}}$), with appropriate input structure. The clustering method reveals 4–5 clusters out of which 1–2 clusters are selected to represent flooding. This result suggest that merging of steps 1 and 2 mentioned in Section 3.1 is a sensible approach to successful indicator design. For this reason, we can expect PCA-based approaches to give inferior performance overall. Also, we exclude $\mathbf{I}^{\mathrm{Uns}}_{\mathrm{SS}}$ from further assessment as its performance would suffer from the inappropriate data labelling. A much more complicated design method (iterating over design steps 1–3 from Section 3.1) would be needed to construct a useful indicator.

The performance assessment of the designed indicators on the testing dataset is shown in Tab. 1, taking into account the so-called confusion matrix elements (i.e., TP, FP, TN, and FN) and performance criteria (i.e., AC, PR, RC, and F1). The complexity of designed indicators is represented by the number of principal components $n_\mathrm{pc}$ for PCA-based approach and by the number of input variables $n_\mathrm{p}$ for the rest of approaches. We can directly see that the supervised ML approaches outperform the unsupervised ones when we compare similar structures. The only exception appear to be the RC criterion when evaluated for $\mathbf{I}_{\mathrm{pat}}$ indicator. There are two reasons for this performance drop: 1. RC is given up in training for the AC and precision as the dataset is more populated with data points of normal operation; 2. the industrial data labels indicate flooding based on other variables than pressure (the sole input to $\mathbf{I}_{\mathrm{pat}}$ indicator) and thus $\mathbf{I}_{\mathrm{pat}}$ indicator falls short in terms of model (input) adequacy (some extra input variables would explain flooding better). Note that, the first reason can be remedied by a modification to SVM objective and some proper tuning, which, however, is beyond the scope of this study.

Among the $\mathbf{I}^{\mathrm{Uns}}$-type approaches, it is interesting that, although the structure of the reference indicator is optimal, the highest RC criterion (low number of FN) is achieved by $\mathbf{I}^{\mathrm{Uns}}_{\mathrm{pat}}$. Of course, this is paid off by worse accuracy as the classifier indicates flooding wrongly (high FP) more often overall. The PCA-based indicator appears to be the least effective (worst in all criteria). This is attributed to the aforementioned inappropriate labelling in high dimensions.

Unlike for the unsupervised learning approaches, the performance of the $\mathbf{I}^{\mathrm{Sup}}_{\mathrm{PCA}}$ indicator is suffi-

cient. It also appears that the $\mathbf{I}^{\text{Sup}}_{\text{PCA}}$ is more efficient compared to the $\mathbf{I}^{\text{Sup}}_{\text{Pat}}$ indicator viewed by each performance criterion. The highest efficacy among supervised learning approaches is achieved for $\mathbf{I}^{\text{Sup}}_{\text{ref}}$ and $\mathbf{I}^{\text{Sup}}_{\text{SS}}$ indicators. These approaches already consider or can find the best possible input structure. It is noteworthy that $\mathbf{I}^{\text{Sup}}_{\text{SS}}$ achieves the best performance (almost 100 % in all performance criteria) using a very simple structure. This effectively tells that the reference structure is overly complicated (some inputs are redundant) and that it is possible to indicate flooding with data from just two sensors. It is also a very interesting result as it allows the industrial practitioners to concentrate efforts regarding sensor maintenance towards smaller subset of online sensors. Surprisingly, pressure is not among the inputs selected for the best indicator. The input structure involves reflux flow $\Delta F_{\text{R}}$ and the time difference of heating medium flowin the reboiler $\Delta F_{\text{reb,h,2}}$, which are both part of the reference indicator structure. It is possible that the two selected flow rates are measured with better precision and that they do not involve high-frequency fluctuations as pressure measurements do. These results need, of course, further validation in an industrial setup as the reference indicator (ground truth) involves the the same input variables as the best indicators found.

## 5. Conclusions

This contribution is focused on the design of a data-driven flooding indicator for an industrial debutanizer column. As the ground truth, a reference indicator is used that is designed according to the industrial knowledge and observations of the debutanizer flooding. The effectiveness of unsupervised and supervised machine learning approaches was evaluated by considering various input structures. The results showed that the unsupervised learning approach can provide sufficient flooding indicators if appropriate input variables are used. The supervised learning approaches achieved higher effectiveness compared to unsupervised learning approaches, resulting from the direct usage of reference indicator labels. The results of supervised learning approaches revealed that the most accurate estimate of the debutanizer flooding is provided by the reflux flow rate and heating medium flow in the reboiler. Future work might involve design of indicators for the chosen performance criterion in a multi-objective fashion. It would also be possible to design an unsupervised approach that would be capable of choosing the optimal indicator input structure much like the structure achieved with supervised learning in this study.

## References

B. E. Boser, I. M. Guyon, V. N. Vapnik, 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. COLT '92. Association for Computing Machinery, New York, NY, USA, pp. 144–152.

E. Forgy, 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics 21, 768–769.

L. F. Fuentes-Cortés, A. Flores-Tlacuahuac, K. D. P. Nigam, 2022. Machine learning algorithms used in pse environments: A didactic approach and critical perspective. Ind. Eng. Chem. Res. 61 (25), 8932–8962.

M. Hubert, M. Debruyne, 2010. Minimum covariance determinant. WIREs Computational Statistics 2 (1), 36–43.

M. King, 2016. Process Control: A Practical Approach. Wiley.

M. Mojto, K. Ľ ubušký, M. Fikar, R. Paulen, 2021. Data-based design of inferential sensors for petrochemical industry. Computers & Chemical Engineering 153, 107437.

J. Oeing, L. M. Neuendorf, L. Bittorf, W. Krieger, N. Kockmann, 2021. Flooding prevention in distillation and extraction columns with aid of machine learning approaches. Chemie Ingenieur Technik 93 (12), 1917–1929.

K. Pearson, 1901. 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 2 (11), 559–572.

H. Peiravan, A. R. Ilkhani, M. J. Sarraf, 2020. Preventing of flooding phenomena on vacuum distillation trays column via controlling coking value factor. SN Applied Sciences 2 (10), 1–11.

R. K. Pihlaja, 6 2008. Detection of distillation column flooding. https://patents.google.com/patent/US8216429B2/en.

G. Smith, 2018. Step away from stepwise. Journal of Big Data 5, 32.