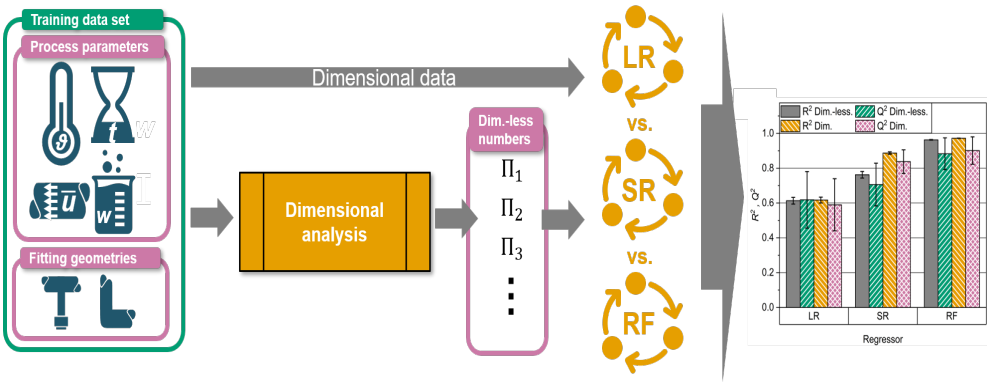
Graphical Abstract

Development of a Soft Sensor for Fouling Prediction in Pipe Fittings using the Example of Particulate Deposition from Suspension Flow

Niklas Jarmatz, Wolfgang Augustin, Stefan Scholl, Alberto Tonda, Guillaume Delaplace



Highlights

Development of a Soft Sensor for Fouling Prediction in Pipe Fittings using the Example of Particulate Deposition from Suspension Flow

Niklas Jarmatz, Wolfgang Augustin, Stefan Scholl, Alberto Tonda, Guillaume Delaplace

- Innovative approach applying Machine learning to predict particulate fouling in food processing
- Deduction of seven dimensionless numbers supporting the model performance
- Generation of a robust soft sensor based on a Random Forest regressor
- Plant operators can apply the presented method to comparable processes and problems

Development of a Soft Sensor for Fouling Prediction in Pipe Fittings using the Example of Particulate Deposition from Suspension Flow

Niklas Jarmatz[a], Wolfgang Augustin[a,1], Stefan Scholl[a], Alberto Tonda[b,d], Guillaume Delaplace[c]

[a] *Technische Universität Braunschweig, Institute for Chemical and Thermal Process Engineering (ICTV), Langer Kamp 7, 38106, Braunschweig, Germany*
[b] *UMR 518 MIA-PS, INRAE, Université Paris-Saclay, 22 place de l'Agronomie, Palaiseau, 91120, France*
[c] *University of Lille, CNRS, INRAE, Centrale institut, UMR 8207 - UMET - Unité Matériaux et Transformations, Lille, 59000, France*
[d] *UAR 3611 Institut des Systémes Complexes Paris Ile-de-France, CNRS, 113 rue Nationale, Paris, 75013, France*

Abstract

Fouling is the unwanted accumulation of material on a processing surface which is an especially problematic issue in the food industry. Characterizing or predicting fouling through traditional methods or models is a challenge due to the complexity of fouling mechanisms. Machine Learning techniques can overcome this challenge by creating models for prediction directly from experimental data. Unfortunately, the results can be hard to interpret depending on the algorithm.

Here, a soft sensor is generated from an extensive data set to predict the fouling of a model particle material system. This is performed inside two different pipe fittings, an inaccessible and accessible fitting (e.g., for sensor measurements). Additionally, Dimensional Analysis is conducted to identify the correlations responsible for fouling while keeping descriptors with physical meaning. The resulting dimensionless numbers are further processed by three machine learning algorithms: Linear Regression, Symbolic Regression, and Random Forest.

Email address: w.augustin@tu-braunschweig.de (Wolfgang Augustin)

The soft sensor generated using a Random Forest outperformed the other two regressors for the dimensional ($Q^2 = 0.90 \pm 0.08$) and for the dimensionless data ($Q^2 = 0.88 \pm 0.09$). The parameter *time* and *particle mass fraction* were determined to be most influential. Furthermore, seven dimensionless numbers were obtained allowing a reduced experimental design.

Keywords: Food processing, Fouling, Cleaning, Sustainability, Machine learning, Dimensional analysis

---

1. Introduction

Fouling and cleaning issues

Fouling has a significant negative environmental and economic impact on numerous production processes due to its impact on heat transfer and fluid flow [1, 2, 3, 4]. Fouling also results in food safety hazards [5], product contamination [6], and/or aroma carry-over [7]. In complex material systems, a considerable number of potential fouling components (proteins, salts, carbohydrates, fats) exist and could, depending the operating conditions, result in different or even coexisting fouling mechanisms. For instance, thermal processing of milk (e.g., pasteurization) can lead to polymer fouling (proteins) and crystallization fouling (milk salts). Additional types of fouling that may occur are particulate, corrosion or bio-fouling [8]. Due to the prevalence of fouling in food processing facilities, better understanding of fouling mechanisms and prediction of the presence of fouling would have significant implications on the development of strategies to reduce or prevent fouling.

Fouling and the associated cleaning steps for a production plant have significant potential for optimization, as the cleaning steps are usually designed with considerable safety margins and result in cleaning processes which are 'too long, too hot and employ too much chemical' [9]. These 'unspecific cleaning protocols' result from the fact that foodstuffs are prone to hygienic risks and are often treated in multi-component production plants [10]. Hence, the significant potential for the improvement of these processes can be accomplished through a better mechanistic understanding of fouling and online prediction of the fouling events. These improvements would enable the establishment of highly optimized and tailored cleaning steps with the help

of sensor data [11]. With these measures, the resource consumption (e.g., energy, water, and chemicals) of cleaning processes could be reduced, while benefiting food product safety and quality. Even though physical sensors enable the prediction of the fouling status online, their use in the field of food processing is restricted due to the diversity of foulants, lack of sensitivity and non-invasive measurement techniques. Furthermore, the need for knowledge of the fouling mechanisms that is required for proper interpretation of the sensor signal is also important. Therefore, so called 'soft sensors', which are digital sensors employing numerical predictive models, are an attractive solution for fouling prediction and mitigation.

Particulate and heat exchanger fouling

To manage the risk of fouling in food processing, hygienic design principles and standards have been established to prevent hazardous process designs and constructions [12, 13]. For particulate fouling, these standards focus on the drainability and the prevention of dead zones in the equipment [14]. Dead zones can induce particulate fouling in both heat transfer equipment and also in the surrounding, transfer piping of a production plant. If there are particles present in the process medium, fouling in both locations are possible even though the cause for the particulate fouling does differ. Particulate fouling is influenced by the bulk gaseous or liquid phase. Particles from the liquid phase, which is the focus of this work, can further be classified by their size: colloidal particulate fouling has a particle size ranging from nanometers to the lower micrometer scale. Conversely, the deposition of larger particles is assigned to sedimentation fouling. Generally speaking, for the flow conditions applied in industrial processes, small particles ($d_p < 1\ \mu m$) are preliminary transported by diffusion, while larger particles ($d_P > 20\ \mu m$) are highly influenced by inertia [15]. Notably, for situations with a high particle concentration, particles can form agglomerates consisting of multiple particles that may be transported like single large particles [16].

Particulate fouling in heat exchangers can be heat induced due to dead zones in non-ideal geometries which lack a proper mixing of the fluid and therefore are overheated compared to other areas of the heat exchanger [17]. Previous work has focused on the influence of the particle concentration on deposition inside heat exchangers [4]. Here, the particle deposition is influenced by the overall fluid flow. For particulate fouling in heat exchangers, the change in

the rate of fouling with time follows an asymptotic trend[18, 8, 4].

In contrast to apparatuses, like heat exchangers, the focus of this work lies on pipe fittings which do not induce a strong temperature gradient between the surface and bulk fluid but highly influence the flow pattern. The deflection of the fluid flow is therefore the main driving force for sedimentation fouling. Even though particulate fouling is common, especially in cooling water applications in the form of sand, mud or dust [8], the actual particle mass fraction in the bulk fluid is often unknown or highly variable. Particulate fouling in food processes has numerous sources and represents a significant hazard for the corresponding equipment. In the dairy industry, the driving factor for fouling is the aggregation of whey proteins which is heat induced. Here, aggregates with calcium ions are formed having a particle size of $d_P = 0.5\ \mu m$ [19]. Furthermore, a hygienic risk for the food industry are biofilms which are formed by microorganisms. Here, single cells, agglomerates of cells, or whole bits of a biofilm can cause a formation of a new biofilm elsewhere [20]. Due to the variation in the size of these particles, it is challenging to characterize and predict the flow and fouling behavior.

Anti-fouling approaches

Extensive efforts have been made in the past decades to develop tools and methods to better understand fouling mechanisms, while establishing and optimizing strategies for the mitigation of fouling. A very prominent strategy is the modification of the surface that is prone to deposition. Here, the measures depend on the process, the fouling mechanism, and surface material (reviewed in [21, 22, 23]). Other strategies attempt to optimize equipment or flow design, e.g., the heat exchanger design, the application of spiral inserts into tubes, or the use of reverse or pulsating flow [24, 25, 26]. To develop strategies for fouling prevention, the system has to be investigated and well understood [27]. Investigations in the lab or production plant to understand the effect of different processing parameters on fouling are usually followed by calculations and modelling of varying complexity. Simple established methods regarding the determination of fouling in a heat exchanger, which is prone to fouling by heat sensitive material, rely on the calculation of the change in pressure drop or the thermal fouling resistance during the production time [28, 29, 30, 31]. More complex approaches use statistical methods to link the degree of fouling with the heat exchanger performance [32] or generate dynamic models to quickly detect changes in the

fouling status of an apparatus [33, 34]. Other studies even enable the local modeling of fouling in heat exchangers from experimental data [35]. Even though the aforementioned approaches allow the description and modeling of fouling, they require a significant amount of experimental data which can be expensive and limited in application to the material system employed in investigations. Recent improvements in computing power and the availability of sensors in the process industry promote promising new approaches for fouling modelling and prediction.

Dimensionless numbers

In addition to the use of sensor data, the application of Dimensionless numbers (DNs) represents another methodology for the establishment of predictive fouling correlations, which are independent of the mechanism of fouling. DNs are algebraic expressions where all variables are physical parameters with a base dimension or combination of base dimensions (e.g., length, time, mass) while the resulting quantity is dimensionless. The application of DNs have several advantages, such as the resulting equation describing a problem can be simplified, scale estimates might be possible and the amount of necessary experiments can be reduced [36]. Furthermore, through the direct use of important physical quantities, underlying physical relations are evident in DNs which promotes the ease of scale-up in further steps.

For complex tasks in the area of chemical and food process engineering (e.g, reaction kinetics, fluid dynamics in complex geometries, etc.) it is often necessary to build a preliminary predictive model instead of relying on fundamental transport phenomenon as a source of modelling. For example, a failure in computational fluid dynamics where the transport and conversion of species are not fully described by equations can occur [37]. Well-known DNs, such as the Reynolds, Froude, Archimedes, Bodenstein or Damköhler number, are only a small selection of the possibilities that DNs have in this research field. Sritham *et al.* established a mathematical approach by applying Dimensional analysis (DA) to model the prediction of soymilk fouling in plate heat exchangers [38]. DNs have also been applied for the prediction of the fouling mass in a plate heat exchanger during thermal processing of a whey protein solution. In this case, DNs incorporate the kinetics of the protein denaturation reaction, the temperature profile along the flow path within the plate heat exchanger, the Reynolds number, and the molar ratio of calcium and β-lactoglobulin [39] or the mass prediction [40, 41]. Deponte *et al.*

applied DNs to model the cleaning times for the three different model soils, namely starch, gelatin and egg yolk [42]. For particles in fluids, Reynolds Re and Archimedes number Ar were applied to describe the relevant flow types in pipes [43] and to characterize motion mechanisms of single particles [44].

Applications of Artificial intelligence

Large improvements in computing capacity and the optimization of the applied algorithms has led to an incredible boost of new applications for Artificial intelligence (AI) . One subfield of AI in particular has become a focal point in recent years due to the direct benefits it provides to end users, e.g., voice and handwriting recognition, autonomous driving, or user profiling for personalized advertising. Machine learning (ML) is a branch of AI where predictive models are automatically derived from raw data. In this study, *supervised* ML is considered, in which the training data includes the correct value of the target variable to be predicted. Furthermore, *regression* algorithms enable the generation of models that predict continuous values. These models are characterized by their *capacity*, that is their ability of learning functions of a given complexity. In general, capacity is inversely proportional to *interpretability*, that is the human ability to make sense of the behavior of a model. For example, linear regression is limited to the approximation of linear functions, but experts can easily interpret a linear model by analyzing the weight it assigns to each problem variable (low capacity, high interpretability). Conversely, large neural networks can deliver good predictions for extremely complex problems, but they can contain billions of weights and are thus complete black boxes which are even im- pervious to the sense-making of domain experts (high capacity, low interpretability). Furthermore, ML models with a high capacity are more sensitive to *overfitting*, a phenomenon where a model memorizes patterns that only exist in the training data. As a result of overfitting, the model shows poor generalization, or in other words, poor performance on unseen (test) data. On the contrary, models with a capacity lower than the minimum capacity needed to model a problem result in *underfitting* and therefore exhibiting low efficiency on all data [45].

Despite a rich collection of theoretical studies on the topic [46], both evaluating the capacity of a specific model and estimating the minimal capacity necessary to describe a regression problem are still open issues in the ML field. The choice of the most appropriate ML algorithm for a problem is

further complicated by the possibility of choosing different *hyperparameters*, which are user- defined values for the settings of an algorithm. In real-world applications, it is common to try different algorithms, evaluate their performance with a k-fold cross-validation, and let a stakeholder select the model that represents the most suitable compromise between performance and human interpretability.

There are many examples of the successful application of ML techniques in the field of process engineering. With specific relevance to this work, ML has been applied in the prediction of fouling in heat exchangers using ML techniques like Deep Learning [47], Symbolic Regression [34] or Gaussian Process Regression, Decision Trees, and Support Vector Regression [48]. Yet, the risk associated with rapid evolution of such ML tools and their introduction to a new research community includes the generation of models that are, e.g., highly overfitted and therefore cannot be generalized well to unseen data points. For example, Kim *et al.* [49] applied an artificial neural network to predict the fouling of particles in a pipe, while obtaining extremely accurate results. However, the methodology employed in the paper uses a single training-test split for the data, so the robustness of the results is questionable; in such cases, ML practitioners would perform at least a cross-validation to obtain a more reliable estimate of the mean and the standard deviation of the algorithm's performance.

The comparison of different ML models is not straightforward. Since most algorithms tend to overfit the data that was used for training, it is advisable to split the data: 1) a *training set*, used for training the algorithm, and 2) a *test set*, unseen during training, that can be used to assess the generalization of the algorithms. A common approach is to shuffle the available data and randomly split it into a training and test set. However, since the split is random, it might accidentally under- or over-estimate the performance of the algorithms. For this reason, it is advisable to perform a k-fold cross-validation, a procedure where the data is randomly divided into k-parts, and the training/ test procedure is repeated k-times, always leaving one part of the data for test and training on $k-1$. A k-fold cross-validation results in a mean performance and an associated standard deviation that can be informative to compare different algorithms. Inclusion of a standard deviation is helpful because in some cases two performances might not be statistically separable even if their means differ. The cross-validation process is still affected

by randomness, but the impact is mitigated by the repeated evaluations.

An advantageous application of ML in the field of process engineering is its use in the generation of soft sensors. Generally speaking, soft sensors take an arbitrary amount of input parameters (sensors, lab data, expert knowledge) to estimate a target value (physical quantity, quality level, etc.). The performance of a soft sensor can be improved by updating simple estimations or calculations with ML algorithms. There is a wide application of soft sensors in chemical, pharmaceutical, and food industries (reviewed in [50]). Despite this widespread use, soft sensors are often application-specific, thus their transferability to other applications can be challenging.

Selected approach

In this work, various ML algorithms are applied and compared for the prediction of particulate fouling in pipe fittings, commonly found in food processing environments. The ML algorithms are then grouped to constitute a robust soft sensor. Before applying the ML algorithms to the data set, a selection of input variables (fluid flow domain, thermophysical properties of the fluid medium, and process parameters) were grouped together in DNs to propose concise semi-empirical correlations between relevant parameters. The goal was that they are not devoid of physical meaning and allow the application of the soft sensor or the established correlations for sizing of real world problems. A model particulate material system was investigated and used for the generation of experimental data. This well-established soda lime glass particle system exhibits beneficial properties including a relatively small particle size ($d_{p,50} = 3.14\ \mu m$), monodispersity, and chemical inertness. Application of this model system led to reproducible results and proved to be beneficial for further modelling and calculations [51, 52, 53]. Experimental investigations of the particulate fouling of two pipe fittings were used to create a soft sensor which was then trained with an extensive data set for one fitting (pipe socket) and a reduced data set for the other fitting (pipe bend). This mimics the situation in a real processing environment: the fouling status of a fitting which is directly accessible by sensors (classically pipe sockets) is estimated by analysis of the online sensor data. It is important to note that integral fouling detection techniques like the determination of the pressure drop would not be able to detect this kind of deposition due to the dead space inside the pipe socket. Here, recent advances in methods for the direct detection of fouling could be applied, namely the use of

temperature measurements in a pipe fitting for the detection of particulate fouling as reported by Jarmatz *et al.* [54] . This study presents an innovative methodology and the results for the development of a soft sensor to predict the fouling of a particulate material system as a pre-step before the inclusion of the mentioned online temperature measurement into the model.

2. Material and Methods

2.1. Generation and preparation of the training data

The applied data set was taken from [53] who performed a comprehensive parameter screening for the investigation of particulate fouling in a selection of prominent pipe fittings. The data set was provided for the reported analysis after the experiments were completed and therefore not tailored for the processing. This mimics an arbitrary record of process data that is further analyzed. The trials included the variation of the fitting type, the fitting orientation in the three-dimensional space, and the pipe diameter or size of the fitting. Furthermore, process parameters such as particle mass fraction, time, volumetric flow rate, and temperature were systematically varied. Fouling was quantified by the mass of fouling deposit in the fitting at the end of each experiment. For better comparison of the experiments, the dimensionless Reynolds number Re (Equation 1) was kept constant for most of the screening experiments (except for the variation of the volumetric flow rate). For the investigated system, the Re is influenced by the variation of the volume flow $\dot{V}$ at a constant pipe diameter and the temperature variation (fluid density ρ and viscosity η are temperature dependent). Since the volume fraction Φ of the particle suspension is relatively low ($\Phi_{max} \leq 0.02$), the fluid can be classified as *Newtonian* and therefore the Re be applied [55, 56, 57].

$$Re = \frac{d_{in} \cdot \rho_{Fl} \cdot \overline{u}}{\eta_{Fl}} \tag{1}$$

The experimental setup consisted of the circulation of the tempered particle suspension through the corresponding fitting as indicated in Figure 1.

The complete screening data set consists of 1452 total data point of which 510 were selected for the data analysis presented in this study. The data was selected based on the following criteria:

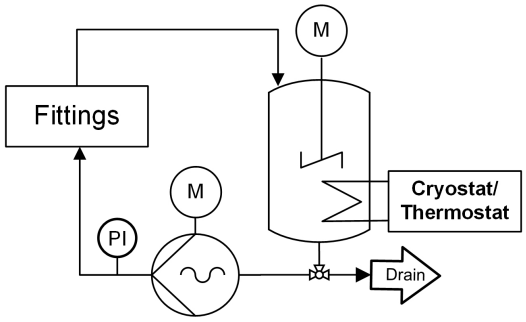


Figure 1: Process flow diagram of the test rig used to generate the investigated data set [53].

1. Data only for the fittings 'pipe socket' and 'pipe bend' were considered to include one fitting accessible to a potential sensor (measuring fitting) and one that is not (target fitting), respectively.
2. Extreme values (either positive or negative) that highly exceed the maximal or minimal value of the other fitting were excluded to prevent problems in the generation of the models.
3. Data only for the pipe diameter of $d_{in} = 6\ mm$ were included for both fittings to exclude a scaling effect to mimic a process plant where the size of the equipment does not change regarding the equipment in question (including experimental data respecting larger diameters is planned for future work).

The data set for the variation of time, particle mass fraction, and Re is visualized for the pipe socket and pipe bend in Figure 2.

The data cloud is most dense at the lower end of the investigated parameter range. This is due to the initial parameter screening which was gradually extended by the inclusion of different lab equipment that allowed for an investigation of a wider parameter range. Starting from there, the process parameters were increased continuously to measure the response of the system. For each data point at least three replicates were recorded but not shown due to overlap. The parameter range for the data set is summarized in Table 1. Based on the second criteria for data selection discussed previously, the data sets for the pipe socket and pipe bend were aligned with respect to the minimum and maximum fouling mass, respectively. This results in a difference in the minima and maxima for each pipe fitting.

For the training of the algorithms the set of the *dimensional* data points

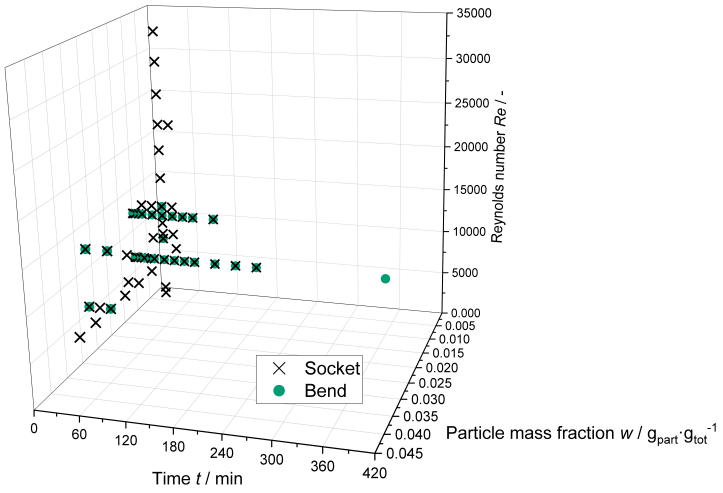


Figure 2: Overview of the combined data set for the socket and bend plotted for the three parameters Reynolds number, particle mass fraction and time.

Table 1: Parameter range of the processed data set for the two fittings socket and bend.

	Socket		Bend	
Parameter	min.	max.	min.	max.
Particle mass fraction w / $g_{part} \cdot {g_{tot}}^{-1}$	0.005	0.04	0.01	0.03
Time t / min	1	180	1	360
Temperature ϑ / $°C$	20	55	20	55
Volume flow $\dot{V}$ / $L \cdot min^{-1}$	0.75	9.5	0.97	3.75
Reynolds number Re / $-$	2,600	33,500	6,700	13,200

consists of the features shown in Table 1, as well as the fitting type, the inner surface area A_{in}, the fluid density ρ_{Fl}, fluid dynamic viscosity η_{Fl}, the fluid kinematic viscosity ν_{Fl}, and a geometric factor ($\alpha = 180°$ for the pipe socket and $\alpha = 90°$ for the pipe bend). This factor describes the straight pipe flow for the pipe socket and the angular flow of the pipe bend. The target parameter was the obtained soil mass.

2.2. Dimensional analysis

For the generation of the DNs, the method developed by [58] and the underlying rules described by [37] were applied:

1. Establishment of a relevance list containing the input and target variables,

2. Screening of the physical quantities for their dimensions in SI-units,
3. Application of the Buckingham Π theorem [59],
4. Generation of the dimensionless numbers,
5. Rearrangement of the determined numbers.

For the generation of the dimensionless numbers, the Buckingham Π theorem was applied as indicated in Equation 2 where p is the number of dimensionless numbers, n the amount of process variables and k the number of physical dimensions [59].

$$p = n - k \tag{2}$$

To the best of the authors' knowledge, this is the first attempt to answer the question of how particulate fouling for defined process conditions can be predicted in pipe fittings. After establishing a relevance list (see Table 2), the further analysis led to the dependencies stated in Equation 3.

Table 2: Relevance list of parameters for particulate fouling in pipe fittings.

Category	Parameters
Process parameters	Average flow velocity $\overline{u}$
	Time t
Geometrical parameters	Inner pipe diameter d_{in}
	Gravitational acceleration g
Particle properties	Particle diameter d_P
	Particle density ρ_P
	Particle mass fraction $\frac{m_P}{m_P+m_{Fl}}$
Fluid properties	Fluid density ρ_{Fl}
	Fluid dynamic viscosity η_{Fl}
	Fluid kinematic viscosity ν_{Fl}
Deposition properties	Deposit mass m_d

$$m_d = f\left(d_p,\ \rho_P,\ \rho_{Fl},\ \eta_{Fl},\ \overline{u},\ \frac{m_P}{m_P + m_{Fl}},\ t,\ g,\ d_{in},\ G_1\right) \tag{3}$$

The causal variables in Equation 3 are composed of i) *common* and *specific* geometrical parameters of the pipe fitting; ii) thermophysical properties

of the bulk fluid, including the density (ρ_{Fl}) and dynamic viscosity (η_{Fl}); and iii) properties of the particles, including mean average diameter (d_P) and the density (ρ_P) . Since the particles are exposed to gravity, the gravitational acceleration (g) is also considered, even though it is constant for all experiments. The *common* geometrical parameters are the diameter of the pipe fitting d_{in} (characteristic length perpendicular to the flow axis) and the specific pipe fitting geometrical parameter G_1, which refers to the *specific* geometrical parameters. These parameters are required to fully define the shape of the socket and bend. These parameters are not generic to all pipe fittings. The term $m_P \cdot (m_P + m_{Fl})^{-1}$ equals the mass fraction of particles in the fluid phase and also impacts the flow. Applying ρ_{Fl}, $\overline{u}$ and d_{in} as relevant physical variables, a first set of dimensionless numbers governing the target variable is chosen (see Equation 4).

$$\frac{m_d}{\rho_P \cdot {d_{in}}^3} = F_1 \left(\frac{\rho_P}{\rho_{Fl}} \ , \ \frac{d_P}{d_{in}} \ , \ \frac{m_P}{m_P + m_{Fl}} \ , \ \frac{\eta_{Fl}}{\rho_{Fl} \cdot \overline{u} \cdot d_{in}} \ , \ \frac{t \cdot \overline{u}}{d_{in}} \ , \ \frac{g \cdot d_{in}}{\overline{u}^2} \ , \ {G_1}^* \right) \quad (4)$$

Note that in Equation 4, the parameter ${G_1}^*$ refers to the set of dimensionless numbers which would be obtained by non-dimensionalizing G_1. Equation 4 represents a pertinent dashboard showing the commands responsible for the variations for the target variable. As explained in [60] and [37], the choice of scaling variables and recombination is user-dependent because the user decides which influence needs to be emphasized and, consequently, the resulting DNs have different interpretations. These choices do not affect the content of the experimental data set, but affect only the form of their presentation. Rearrangements are often motivated to i) produce DNs whose physical meaning is well established and accepted, ii) eliminate fractional exponents, iii) eliminate a physical quantity of a DN to obtain a new one independent of it or to isolate a physical quantity within a single DN of the set. Rearrangement of Equation 4 results in a classical set of DNs in Equation 5 which control sedimentation of particles when a solid/liquid suspension circulates in a pipe fitting. This set of DNs are processed and presented as features to the ML algorithms.

$$\frac{m_d}{\rho_P \cdot d_{in} \cdot A_{in}} = F_2 \left(\frac{\rho_P - \rho_{Fl}}{\rho_{Fl}} \, , \, \frac{d_P}{d_{in}} \, , \, \frac{m_P}{m_P + m_{Fl}} \, , \, Re = \frac{\rho_{Fl} \cdot \overline{u} \cdot d_{in}}{\eta_{Fl}} \, , \right.$$
$$\left. \frac{t \cdot \overline{u}}{d_{in}} \, , \, Ar = \frac{(\rho_P - \rho_{Fl}) \cdot g \cdot {d_P}^3}{\eta_{Fl}}, \, {G_1}^* \right) \quad (5)$$

2.3. Machine learning techniques

In this study, three, representative, albeit not exhaustive, subset of techniques commonly used in real-world applications, including linear models, free-form equation-based models, and ensembles of decision trees are presented. These techniques were selected after performing preliminary runs with several ML algorithms (details reported in Appendix A). The performance of all algorithms is evaluated using a 10-fold cross-validation scheme.

A classical quality metric for assessing a ML algorithm's performance for regression problems is the mean squared error between the model's prediction and observed values; however, the scale of the data can differ between data sets, thus practitioners favor the coefficient of determination R^2 which is defined as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (6)$$

where y_i is the observed value for the i-th data sample, $\hat{y}_i$ is the ML model's prediction for the i-th sample, and $\bar{y}$ is the mean value for all observed values. A R^2 value close to 1.0 indicates good performance of a model, while a R^2 value close to 0.0 (or negative) indicates poor performance. Following classical naming conventions in this study, R^2 will be referred to as the coefficient of determination computed on the training set, and Q^2 the coefficient of determination computed on the test set. For most ML algorithms, $R^2 > Q^2$ due to overfitting on the training data. Thus, Q^2 is a more reliable value to compare algorithms.

Some, but not all, ML algorithms require data to be normalized and pre-processed for maximum efficiency. Therefore, a classical pre-processing step

for all algorithms is applied. First, the values of the only categorical feature `Fitting_type`, whose original values are strings which signify the difference between bend and pipe socket, are changed to numerical values with physical meaning, namely the angle in the bend (90) and the pipe socket (180) in degrees. Second, during the cross-validation process, a standard normalization is applied to all features in the problem that learns the normalization on the training set and applies it to the test set. For each feature, the normalization removes the mean and scales its values to unit variance. Given the limited amount of available samples, no hyperparameter tuning was performed for the ML algorithms, as a proper hyperparameter tuning process requires a validation set which necessitates a further split in the dataset. Moreover, this work is a proof of concept and the performance of the algorithms was acceptable even without extensive tuning (see Table A.4.)

The code for the experiments is implemented in `python`, relying upon the ML package `scikit-learn` [61] and for the Symbolic regression (SR) algorithm, the compatible package `PySR` [62] was implemented. For all the algorithms, default hyperparameter values from the library were applied, except when specified. The code is freely available on a GitHub repository (the link will be disclosed after peer review).

2.3.1. Linear regression

Linear regression (LR) [63] is one of the first and most popular approaches for creating ML models. The algorithm fits a linear model with weights $w = (w_1, w_2, ..., w_p)$, where p is the number of features, to minimize the difference between the observed targets in the training data and the predictions of the model. This solves the problem in the form, $\min ||Xw - y||_2^2$, where X is the matrix containing the feature values for each sample and y is the vector with the observed values in the training set. Compared to other ML algorithms, LR has a low capacity, but is highly interpretable for humans.

2.3.2. Symbolic Regression

The term SR defines a class of algorithms that perform a stochastic exploration of the search space consisting of all free-form equations to find the equation that best approximates the training set. SR techniques are based on a stochastic optimization algorithm called Genetic programming (GP) [64], where candidate solutions are described as binary trees. Intermediate nodes encode mathematical operations (e.g. addition, subtraction, multiplication,

logarithm...) and leaves/terminals of the tree encode problem features or constant values. In theory, SR can deliver equations of arbitrary shape, so this algorithm typically creates models of higher capacity than LR. On the other hand, the interpretability of SR models can be lower, because it might be hard to make sense of extremely complex equations, for example involving logarithmic or trigonometrical operators.

SR has a considerable number of hyperparameters. For example, μ is the size of the archive of GP trees kept in memory at each iteration, while λ is the number of new GP trees created with each iteration. For the experiments reported in this study, these values were set at $\mu = 500$, $\lambda = 500$ after a few trial runs and a stop condition after 100 iterations. All other hyperparameters use the default values from the `PySR` package. Since SR does not have a default method for extracting the relative feature importance from the final solution, the feature importance was evaluated by assessing the relative frequency of appearance of each feature among the best candidate solutions found in the archive at the last iteration.

2.3.3. Random forest

Decision trees (DTs) are popular, human-readable predictive models that are capable of approximating non-linear functions as piecewise constant values. DTs can be hand-crafted by experts or automatically obtained from data using ML algorithms [65]. Despite their desirable properties in ML, DTs have relatively low *capacity* and thus only deliver satisfying results for simple functions. Ensembles of DTs are an approach devised to address this issue, making it possible to tackle more complex functions by relying upon a set of DTs instead of just one. In an ensemble, each DT processes the input sample separately and obtains a single constant output value with an averaging of the output values of all trees to produce the final answer of the ensemble. Several techniques have been proposed to create effective ensembles, but one of the simplest and most effective remains Random forest (RF) [66]. In RF, each DT in the ensemble is created iteratively resulting in a deterministic ML algorithm, while the training data shown to the algorithm is randomly selected and sets of samples and features at each iteration are removed. RF ensembles are fast to train and surprisingly effective, even when compared to more complex ensemble creation techniques, such as Boosting [67]. RF models boast a high capacity, and the algorithm is known to be effective for a variety of real-world problems; on the other hand, it has the lowest interpretability among the other techniques considered in this study because it is

impossible for human experts to interpret the predictions of tens or hundreds of trees. At best, RF models can return the relative importance of a feature, based on the frequency of the appearance of a specific variable among the splits of the trees in the ensemble. RF has a few hyperparameters that can be set by a user before the start of the training process and have an influence on the algorithm's performance. Notably, these hyperparameters are the number of trees in the ensemble and the metric used to determine the splits in the trees. In the following experiments, all hyperparameters for RF have the default values of the `scikit-learn` package.

3. Results and Discussion

3.1. Deduced dimensionless numbers

According to the Buckingham Π theorem and the fundamental rules described by [37], seven DNs were calculated while Π_1 to Π_6 refer to input parameters for the model and Π_7 is the target variable (see Equation 7 to 13). The DN Π_1 is equal to the well-known Reynolds number Re, which is defined as the ratio of the inertial forces to viscous forces for a fluid element, that characterizes the flow regime [60]. The inclusion of Re in this set of DN is reasonable since the fluid flow and the corresponding shear forces play an important role for the formation and removal of fouling [24]. Furthermore, since the density and dynamic viscosity of the fluid is temperature dependent, the Re takes the variation of temperature into account, while keeping this dimensionless number constant. This was done by Jarmatz *et al.* while generating the training data set which is further processed in this study [53]. The detection of DN $\Pi_2 = m_P \cdot (m_P + m_{Fl})^{-1}$ is expected, as it is widely accepted that the particle mass fraction impacts the fluid flow and ultimately the fouling outcome. Π_3 is referred to the Time number (TimeNo) that is a dimensionless measure of the experimental duration (time t). Another prominent DN is the Archimedes number Ar which is a dimensionless number used in fluid mechanics to quantify the ratio of buoyancy effects to viscous effects. The Archimedes number emphasizes that a solid particle immersed in a liquid undergoes a vertical upward thrust equal to the weight of the volume of liquid displaced by the particle and simultaneously a vertical downward thrust due to gravitational acceleration [36]. It is important to note that despite the fact that the density of a fluid is temperature dependent, this fact is not very pronounced for solids and can therefore be assumed constant in the narrow temperature range applied here [68].

$$\Pi_1 = Re = \frac{\rho_{Fl} \cdot \overline{u} \cdot d_{in}}{\eta_{Fl}} \tag{7}$$

$$\Pi_2 = w = \frac{m_P}{m_{Fl} + m_P} \tag{8}$$

$$\Pi_3 = t \cdot \frac{\overline{u}}{d_{in}} \tag{9}$$

$$\Pi_4 = Ar = \frac{(\rho_P - \rho_{Fl}) \cdot g \cdot {d_P}^3}{\rho_{Fl} \cdot {\nu_{Fl}}^2} \tag{10}$$

$$\Pi_5 = \frac{(\rho_P - \rho_{Fl})}{\rho_{Fl}} \tag{11}$$

$$\Pi_6 = \frac{d_P}{d_{in}} \tag{12}$$

$$\Pi_7 = \frac{m_d}{\rho_P \cdot d_{in} \cdot A_{in}} \tag{13}$$

The DN Π_5 isolates the buoyancy correction term ($\Delta\rho \cdot \rho^{-1}$) from the Ar and is also a function of the temperature. DN Π_6 describes the ratio between the particle and the pipe diameter. Since these diameters are constant throughout this study, this DN is not further processed but will be addressed in further studies when the associated parameters are varied. The DN Π_7 is the sole target number that quantifies fouling. It refers to the dry mass of the deposited particles m_d to the density of the particles ρ_P, the inner diameter d_{in} and the inner surface area A_{in}. With this comprehensive set of DN the model training for the fouling prediction was conducted applying ML algorithms.

3.2. Model generation using machine learning

To create a robust and reliable soft sensor, the DNs generated in section 3.1 as well as the dimensional process parameters (see section 2.1) were used to generate separate models for direct comparison. For this purpose, a total of 42 ML algorithms were investigated in terms of their predictive performance (for a full overview, see the supplemental information in Appendix A). The following detailed analysis is carried out for the three different data sets (pipe socket, pipe bend, and the combination of bend and socket fittings) mentioned above and described in section 2.1. For the three data sets and the 42 algorithms, the models were trained, tested, and evaluated by comput-

ing the R^2 and Q^2. During model training, it was investigated whether the generated set of DNs (Π_1 ... Π_5) contain the same information for the model as the dimensional process parameters. It was shown that a diverse result is obtained regarding the differences between the calculated values for R^2 and Q^2 regarding the data sets of dimensional and dimensionless numbers. This indicates that for some algorithms there is a high variation between the R_2 and Q_2 when comparing the dimensionless with the dimensional data set as input, while for other algorithms there is no or a low difference between the R_2 and Q_2. However, this analysis demonstrates that most of the algorithms extract the same amount of information from the dimensionless data set as from the dimensional, original data set (for details, see supplemental information Appendix B). To evaluate the screened algorithms for the development of a soft sensor to predict fouling, three algorithms that differ considerably in levels of capacity and interpretability were selected for evaluation as shown in Table 3.

Table 3: Selection properties for the three target algorithms.

Algorithm	Capacity	Interpretability
Linear regression	Low	High
Symbolic regression	Medium	Medium
Random forest	High	Low

When multiple algorithms performed similarly, with respect to R^2 and Q^2, and also fit in the same preference category mentioned above, the algorithm with more interpretability (and less capacity) was chosen. Figure 3 shows the results for the algorithms LR, SR and RF for the three data sets: 1) pipe socket, 2) pipe bend, and 3) the combination of socket and bend for the two approaches of dimensional as well as dimensionless data. For the pipe socket (3a), all three algorithms perform quite similarly with respect to the R^2 and Q^2. The overall performance of the RF regressor is a little higher, while there not much difference between the dimensionless and dimensional data set. This result was expected since the data set for the pipe socket is quite exhaustive and Jarmatz *et al.* reported mostly linear correlations for the experimental data (see [53]). These correlations lead to a similar result during the cross-validation for the three data sets. The Q^2 value exhibits a larger standard deviation than the R^2 which results from the calculation of the two values. As explained in section 2.3, R^2 is a measure for the goodness of the fit, while Q^2 is a measure of the goodness of prediction. Furthermore,

a low difference between R^2 and Q^2 indicate a good model which is true for all models associated with the pipe socket.

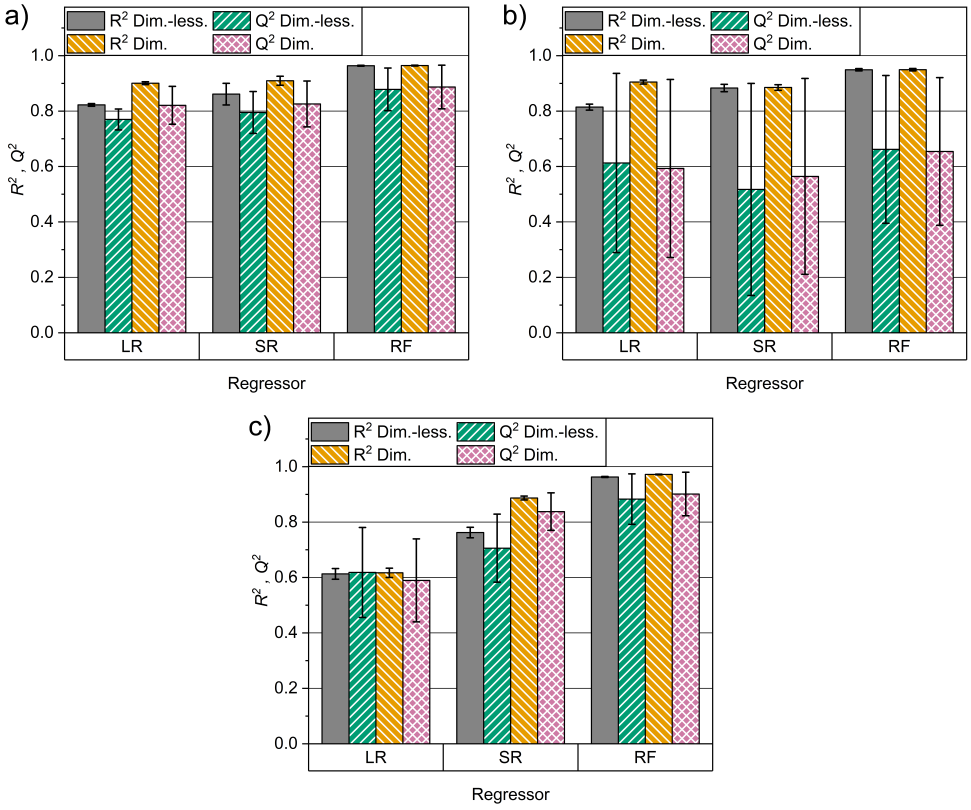


Figure 3: Overall model performance indicated by R^2 and Q^2 for the three data sets a) the sole pipe socket, b) the sole pipe bend and c) the combination of socket and bend for the dimensional and dimensionless data set. LR = Linear regression, SR = Symbolic regression and RF = Random forest regressor.

The situation is different for the data set of the pipe bend (Figure 3b). The values for R^2 are similar for the three algorithms and similar in magnitude to the results for the pipe socket. In contrast to the pipe socket, the results for the pipe bend illustrated that the goodness of prediction (Q^2) is lower for all three models and the standard deviation of the cross validation is significantly higher. This result is also not surprising since the data set of the pipe bend is considerably smaller than that for the pipe socket and also less distributed over the range of the process parameters investigated. This naturally leads to a decrease in the predictive power of the model and

results in a higher standard deviation in the cross validation process. Apart from the reduction in the predictive power, the overall performance is again higher for the RF compared to the two other algorithms. Interestingly, the results are quite different for the combined data set of the socket and the bend (see Figure 3c). Here, all three chosen algorithms exhibit a different result regarding the goodness of fit and prediction. For the LR model, all values are reduced to approximately $R^2 = 0.6$ and $Q^2 = 0.6$ for the dimensional as well as dimensionless data set. The standard deviation for the goodness of fit for the LR model exhibits the highest standard deviation compared to the other models in this data set. This result indicates that a simple linear regression is unable to map the complex data set of two independent fittings with a different geometry. The SR leads to different values for R^2 and Q^2 for the dimensionless (similar to the LR) and the dimensional (similar to the results of the RF). It is hypothesized that the use of DNs results in a lack of information for this tool. The overall performance of the RF regressor is high with respect to the goodness of fit and prediction and comparable to the result of the data sets of each pipe fitting in isolation. Furthermore, the respective use of dimensional parameters or DNs does not lead to a significant variation in model performance. This is advantageous since DNs enable scaling and a potential reduction in the number of experiments due to the replacement of a higher number of dimensional parameters with less DNs. Therefore, a similar model performance makes it possible to apply the DNs without a loss of information for the model training. Since the RF tool is known to be a powerful regressor, the similarity in the values of R^2 and Q^2 is expected. Furthermore, the RF shows the highest R^2 and Q^2 for the analysis of all three data sets, which is not surprising, as ensembles of DTs are known to outperform techniques operating on tabular data sets, such as the other ones considered in this study [69]. Furthermore, ensembles of DTs are more suited to handle categorical variables, such as the variable which defines the difference between the pipe socket and bend samples, when compared to models based on equations.

3.3. Feature importance

Apart from the overall performance of the model, it is also important to analyze how the algorithms computed the results and what importance is assigned to each input parameter. For this purpose, the three selected algorithms were compared in terms of the importance they assign to the independent features that serve as input variables for the model building.

When comparing the values of feature importance, it is worth noting that each algorithm employs a different strategy for assessing feature importance. Thus, the absolute values assigned to each feature cannot be compared across algorithms, but what is important is the relative feature importance for the same algorithm.

Figure 4 summarizes the results for the analysis of the feature importance for the dimensionless data set for the a) socket, b) bend, and c) combination of socket and bend. For the socket data set, the LR identifies the Ar (Π_4) and the Density number (DenNo) (Π_5) as the features with the highest impact on the target DN (Deposition number (DepNo)). The fitting type and the Re have a small, negative effect on the target (e.g., an increase of the Re for instance reduces the DepNo). Conversely, an increase in the Particle mass fraction (PMF) or the TimeNo lead to higher fouling. The effects of the PMF and time were already reported for the experimental screening data by [53]. The Ar and the DenNo are related by the temperature of the suspension. Since Jarmatz *et al.* could not identify a significant impact of a fluid temperature change on the deposition mass, the feature importance from the LR model should be questioned. One possible explanation is that the density of the particles ρ_P is used for the calculation of the Ar, DenNo as well as the target (DepNo). This might lead an overestimation in the importance of the DN Π_4 and Π_5. Another possibility is that the real impact of the features can only be assessed by non-linear relationships. This is further confirmed by the results obtained by the SR algorithm.

Figure 4b shows the results for the feature importance assigned by SR. The obtained results confirm the observations made by [53] regarding the significant influence of experimentation time and the PMF on the deposition. Also, the fitting type and Re are assigned a high importance. Interestingly, SR exploits the information contained in DenNo, but not in Ar. This may be explained by the fact that the two features contain similar information. Nevertheless, both SR and RF agree in assigning the highest importance to experimentation time and DenNo. Since these two algorithms have a higher capacity, it is hypothesized that LR was unable to capture the correct relationship.

Compared to the feature importance of the other algorithms, the feature importance of RF closely follows the experimental observations made by [53] (see Figure 4c). The TimeNo followed by the PMF were identified as the features with the highest importance . The relatively high rating of the fitting type is expected due to the difference in geometry of the fittings. A

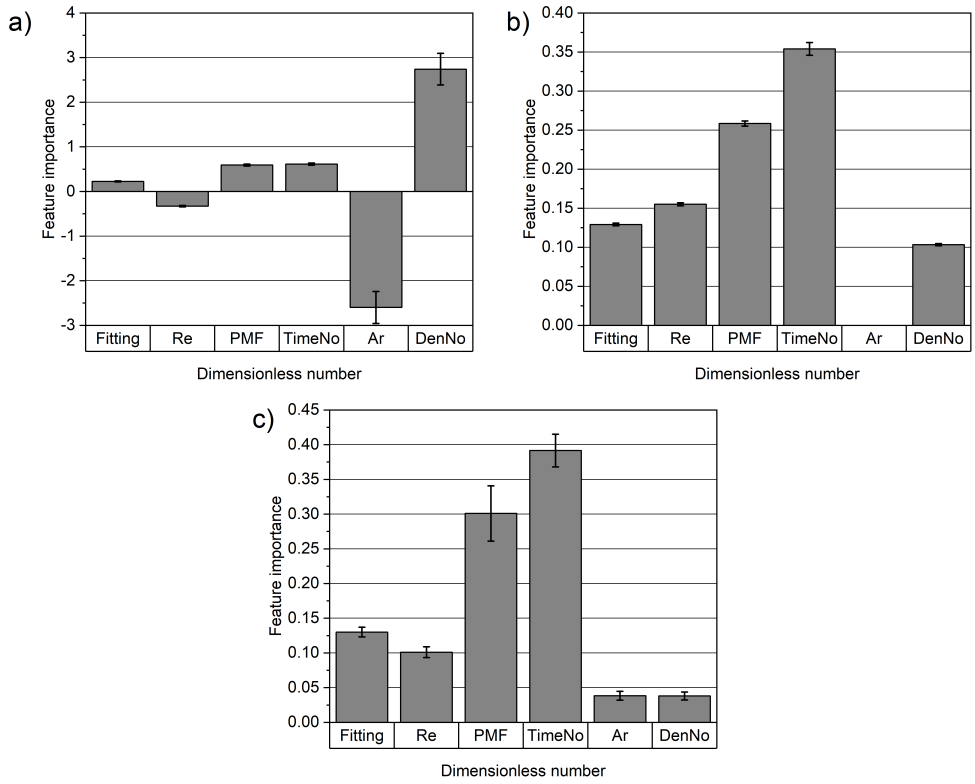


Figure 4: Calculated feature importance by a) LR, b) SR and c) RF applying the combined data set of socket and bend for dimensionless numbers.

comparably lower influence of the Re and therefore mainly the volume flow is supported by a direct comparison to the experimental data. A low and additionally extremely similar result for the Ar and DenNo is comprehensible, since the temperature (and therefore the difference in fluid density) did not have a large effect on the experimental data. Since the same parameters are used for the calculation of the Ar and DenNo in this study (with exception of the kinematic viscosity of the fluid which additionally occurs in the Ar), this internal quality control confirms that the model generates the same results from features that contain very similar information. This illustrates that the RF is able to extract the same information from the dimensionless as well as dimensional data set (see Figure 3), while also applying equal importance to the two mentioned DNs. This indicates that a combined application of DA and ML techniques allow the reduction of variables which might lead to reduced experimental effort.

3.4. Soft sensor performance

After establishing the overall performance of the models, the next logical step is to check how well they can predict unseen data to evaluate possible applications in real world processes. The values of Q^2 have already been compared for all algorithms in section 3.2. Figure 5 shows the observed vs. predicted plots for the three compared techniques, LR, SR, and RF, for all three data sets. The data points are in general less scattered for the SR (Figure 5b) compared to the LR (Figure 5a). For the LR and SR data sets, there is a dense distribution of the data at low values of the target variable Π_7, while a wider distribution for data at higher values of the the target variable Π_7 . This is coherent with the results discussed above regarding the overall model performance (see Figure 3), where the values for R^2 and Q^2 are higher for LR than SR with respect to the dimensionless data sets. For the RF results, the higher values of R^2 and Q^2 (Figure 3) compared to the LR and SR algorithms correspond to a different scattering pattern of the data points in the observed vs. predicted plot (5c). The data points for the RF are much closer to the orthogonal line than for the LR and SR algorithms indicating the best theoretical fit in the RF model. This is important because all data sets illustrate that the majority of experiments led to comparably low fouling while the density of the data points decreases significantly when the obtained fouling mass of the experiments increases. The RF appears to be more suitable for the investigated case to replicate this issue in the prediction. It is important to note that the data set employed in this study was not explicitly designed and generated to be processed by DA and ML. Therefore, the uneven distribution of data outlined by the observed vs. predicted plot is an example of real process data that can be effectively analyzed by this methodology.

With respect to a future industrial applications of this soft sensor, this comparison shows that the predictive power of RF algorithms outperforms that of LR and SR models. Despite the reduced accessibility of the fouling prediction at the target position (pipe bend fitting), RF provides higher performance in terms of regression and classification applications than LR and SR. It is advantageous for operators that the RF obtains similar results for the dimensionless and dimensional data because this indicates that the chosen algorithms is quite robust. Additionally, less training data (and therefore necessary process sensors) is required with the application of DN. This holds significant importance in food processing environments since inaccuracies specially in terms of fouling (and the corresponding, necessary cleaning)

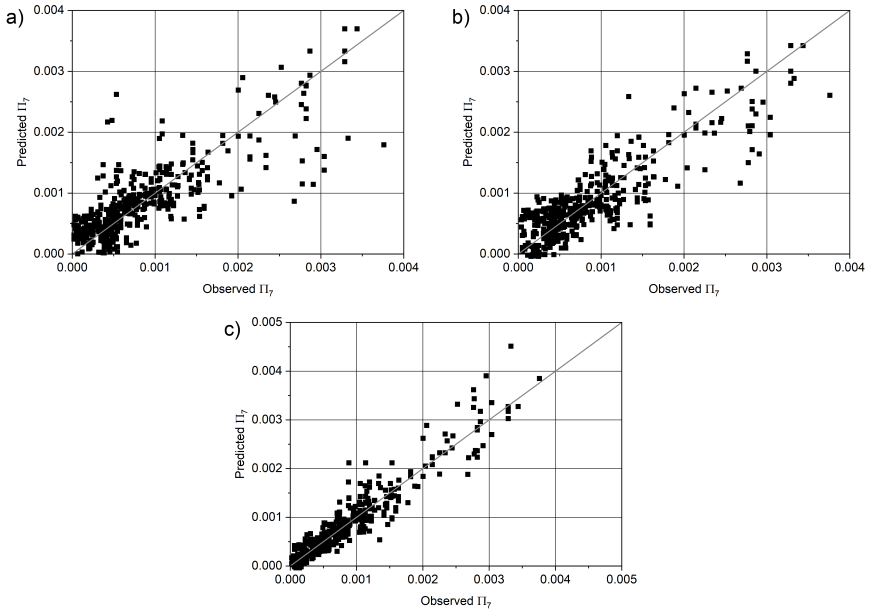


Figure 5: Measured vs. predicted plots for the combination of socket and bend for a) LR, b) SR and c) RF applying the set of dimensionless numbers.

lead to an increased risk for contamination or hygiene issues.

4. Conclusions

In this study, a comprehensive and innovative approach was presented for the development of a soft sensor that predicts particulate fouling in pipe fittings with the application of DA for the training of ML algorithms. The concept is applicable for plant operators that need an approach for quick training of robust ML algorithms with historical data for training, testing and validation. This study shows that a variety of different algorithms can be used to process the investigated data set while obtaining similar results.

The analysis focuses on the evaluation of the predictive performance of three well established ML algorithms. The results show a difference between the LR and SR compared to the RF with respect to the overall model performance (R^2, Q^2) and the obtained feature importance of the input parameters. The RF outperforms the LR and the SR, especially in the application of DNs as input features. This is a significant advantage for the use of real-world

data since dimensionless descriptors reduce the number of parameters that need to be processed. Furthermore, this hybrid approach combining DN and ML simplifies the problem while maintaining underlying physical relations as features for the ML algorithms. This increases the chances of generating a powerful and transferable model which can be applied to other systems. The applied RF regressor proved to be very suitable to generate a superordinated model for the data set with data from both pipe fittings, as illustrated by having a predictive performance with similar R^2 and Q^2 as that calculated for the sole pipe socket. The results for the combined set of fittings outperform the values for the sole bend fitting indicating that the algorithm can manage the combination of two data sets despite the reduced data for the pipe bend.

In further research, the presented approach should be tested on different data sets, and ideally with a more complex fouling system. While promising, the model results here are unlikely to be easily transferred to pipe shapes unseen by the ML algorithms, so more diversified data is needed. Furthermore, future work should focus on increasing the scale of the equipment to better emulate the scale of industrial applications. Future steps will extend the proposed methodology to predict whey protein fouling in a plate heat exchanger. This will lead to a comprehensive toolbox that operators can deploy on their specific fouling problem. A second area of future work is further investigation of the application of DA to reduce the number experiments through the use of screening experiments. Ultimately, future investigations should include the use of the soft sensor developed here in combination with online temperature measurement in the pipe socket to predict the fouling status of the pipe bend. Additionally, more sophisticated ML algorithms, such as Long-Short-Term Memory Networks, could also be employed to evaluate the fouling prediction as a time-series forecasting problem which may better describe the dynamic nature of the fouling phenomena; however, ML algorithms employing time series require large volumes of high-quality data to produce satisfying results. Therefore, the real-world applicability of these more complex ML approaches ultimately rests on the availability of better data sets.

Acronyms

AI Artificial intelligence

DA Dimensional analysis

DenNo Density number

DepNo Deposition number

DN Dimensionless number

DT Decision tree

GP Genetic programming

LR Linear regression

ML Machine learning

PMF Particle mass fraction

RF Random forest

SR Symbolic regression

TimeNo Time number

Nomenclature

Dimensionless numbers

Ar	Archimedes number	–
Re	Reynolds number	–

Greek letters

α	Geometric parameter	–
Δ	Difference	–
η	Dynamic viscosity	$kg \cdot m^{-1} \cdot s^{-1}$
λ	Number of new trees created at each iteration	–

μ	Size of the tree kept in memory at each iteration	–
ν	Kinematic viscosity	$m^2 \cdot s^{-1}$
Φ	Volume fraction	–
Π	Dimensionless number	–
ρ	Density	$kg \cdot m^{-3}$
ϑ	Temperature	$°C$

Latin letters

A	Area	m^2
d	Diameter	m
G	Specific pipe fitting geometrical parameter	–
g	Gravitational acceleration	$m \cdot s^{-2}$
k	Number of parts for the cross validation	–
k	Number of physical dimensions	–
L	Length	m
m	Mass	g
n	Number of process variables	–
p	Number of dimensionless numbers	–
p	Number of features for linear models	–
Q^2	Predictive relevance	–
R^2	Coefficient of determination	–
SS_{res}	Residual sum of squares	–
SS_{tot}	Total sum of squares	–
t	Time	min

$\overline{u}$	Average flow velocity	$m \cdot s^{-1}$
$\dot{V}$	Volume flow	$L \cdot min^{-1}$
w	Particle mass fraction	$g_{part} \cdot g_{tot}{}^{-1}$
w	Weight for linear models	–
X	Matrix containing feature values for each sample	–
y	Vector with observed values in the training data set	–
$\hat{y}_i$	Model prediction for the i-th sample	–
$\overline{y}_i$	Mean value over all observed values	–
y_i	Observed value for the i-th data sample	–

Sub- and superscripts

*	Dimensionless
Fl	Fluid
in	Inner
P	Particle
tot	Total

Acknowledgments

The authors like to thank Holly Huellemeier for the proofreading of the document.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that they have no conflict of interest regarding the work reported in this article.

Tweetable abstract

An innovative approach of a successful soft sensor generation for the prediction of particulate fouling in pipe fittings was established by a combination of Machine learning and Dimensional analysis

Appendix A. Preliminary runs of ML algorithms

In a series of preliminary experiments, all regression algorithms available in the `scikit-learn` Python package were run on the whole dataset. The algorithms were selected using the `sklearn.utils.all_estimators(type_filter="regressor")` option. Four more algorithms were added to the set: `gplearn.SymbolicRegressor`, `XGBoostRegressor`, `LightGBMRegressor`, and `CatBoostRegressor`. All algorithms that returned errors were then discarded, leading to the results presented in Table A.4.

Table A.4: Results of the preliminary experiments, using all regressors on the dimensionless data set.

Regressor name	**R2**	**Q2**	**MSE_train**	**MSE_test**
ARDRegression	0.6124 +/- 0.0193	0.6192 +/- 0.1623	0.3876 +/- 0.0193	0.3808 +/- 0.1623
AdaBoostRegressor	0.7689 +/- 0.0207	0.7267 +/- 0.0664	0.2311 +/- 0.0207	0.2733 +/- 0.0664
BaggingRegressor	0.9603 +/- 0.0017	0.8749 +/- 0.0884	0.0397 +/- 0.0017	0.1251 +/- 0.0884
BayesianRidge	0.6125 +/- 0.0192	0.6190 +/- 0.1620	0.3875 +/- 0.0192	0.3810 +/- 0.1620
DecisionTreeRegressor	0.9642 +/- 0.0016	0.8804 +/- 0.0910	0.0358 +/- 0.0016	0.1196 +/- 0.0910
ElasticNet	0.0493 +/- 0.0196	0.0446 +/- 0.0176	0.9507 +/- 0.0196	0.9554 +/- 0.0176
ElasticNetCV	0.6121 +/- 0.0193	0.6206 +/- 0.1617	0.3879 +/- 0.0193	0.3794 +/- 0.1617
ExtraTreeRegressor	0.9642 +/- 0.0016	0.8804 +/- 0.0910	0.0358 +/- 0.0016	0.1196 +/- 0.0910
ExtraTreesRegressor	0.9642 +/- 0.0016	0.8804 +/- 0.0910	0.0358 +/- 0.0016	0.1196 +/- 0.0910
GaussianProcessRegressor	0.9634 +/- 0.0015	0.8805 +/- 0.0905	0.0366 +/- 0.0015	0.1195 +/- 0.0905
GradientBoostingRegressor	0.9270 +/- 0.0057	0.8469 +/- 0.0879	0.0730 +/- 0.0057	0.1531 +/- 0.0879
HistGradientBoostingRegressor	0.9134 +/- 0.0074	0.8411 +/- 0.0722	0.0866 +/- 0.0074	0.1589 +/- 0.0722
KNeighborsRegressor	0.9404 +/- 0.0028	0.8572 +/- 0.0890	0.0596 +/- 0.0028	0.1428 +/- 0.0890
KernelRidge	0.6126 +/- 0.0192	0.6187 +/- 0.1619	0.3874 +/- 0.0192	0.3813 +/- 0.1619
Lars	0.6130 +/- 0.0193	0.6183 +/- 0.1623	0.3870 +/- 0.0193	0.3817 +/- 0.1623
LarsCV	0.6127 +/- 0.0195	0.6187 +/- 0.1629	0.3873 +/- 0.0195	0.3813 +/- 0.1629
Lasso	0.0000 +/- 0.0000	-0.0000 +/- 0.0000	1.0000 +/- 0.0000	1.0000 +/- 0.0000
LassoCV	0.6123 +/- 0.0193	0.6201 +/- 0.1618	0.3877 +/- 0.0193	0.3799 +/- 0.1618
LassoLarsCV	0.6127 +/- 0.0195	0.6187 +/- 0.1629	0.3873 +/- 0.0195	0.3813 +/- 0.1629
LassoLarsIC	0.6126 +/- 0.0192	0.6184 +/- 0.1619	0.3874 +/- 0.0192	0.3816 +/- 0.1619
LinearRegression	0.6130 +/- 0.0193	0.6183 +/- 0.1623	0.3870 +/- 0.0193	0.3817 +/- 0.1623
LinearSVR	0.5887 +/- 0.0212	0.6124 +/- 0.1848	0.4113 +/- 0.0212	0.3876 +/- 0.1848
MLPRegressor	0.9392 +/- 0.0054	0.8740 +/- 0.0813	0.0608 +/- 0.0054	0.1260 +/- 0.0813
NuSVR	0.9291 +/- 0.0037	0.8666 +/- 0.0785	0.0709 +/- 0.0037	0.1334 +/- 0.0785
OrthogonalMatchingPursuit	0.3222 +/- 0.0285	0.3619 +/- 0.2314	0.6778 +/- 0.0285	0.6381 +/- 0.2314
OrthogonalMatchingPursuitCV	0.6126 +/- 0.0192	0.6184 +/- 0.1619	0.3874 +/- 0.0192	0.3816 +/- 0.1619
PassiveAggressiveRegressor	0.1629 +/- 0.3850	0.1768 +/- 0.3988	0.8371 +/- 0.3850	0.8232 +/- 0.3988
RANSACRegressor	0.4332 +/- 0.1258	0.4839 +/- 0.2229	0.5668 +/- 0.1258	0.5161 +/- 0.2229
RadiusNeighborsRegressor	0.8454 +/- 0.0101	0.7839 +/- 0.1259	0.1546 +/- 0.0101	0.2161 +/- 0.1259
RandomForestRegressor	0.9628 +/- 0.0019	0.8857 +/- 0.0875	0.0372 +/- 0.0019	0.1143 +/- 0.0875
Ridge	0.6126 +/- 0.0192	0.6187 +/- 0.1619	0.3874 +/- 0.0192	0.3813 +/- 0.1619
RidgeCV	0.6118 +/- 0.0192	0.6203 +/- 0.1622	0.3882 +/- 0.0192	0.3797 +/- 0.1622
SGDRegressor	0.6122 +/- 0.0193	0.6195 +/- 0.1608	0.3878 +/- 0.0193	0.3805 +/- 0.1608
SVR	0.9295 +/- 0.0035	0.8671 +/- 0.0786	0.0705 +/- 0.0035	0.1329 +/- 0.0786
TheilSenRegressor	0.5068 +/- 0.0188	0.5244 +/- 0.1367	0.4932 +/- 0.0188	0.4756 +/- 0.1367
TransformedTargetRegressor	0.6130 +/- 0.0193	0.6183 +/- 0.1623	0.3870 +/- 0.0193	0.3817 +/- 0.1623
PySRRegressor	0.7622 +/- 0.0190	0.7059 +/- 0.1227	0.2378 +/- 0.0190	0.2941 +/- 0.1227
XGBRegressor	0.9642 +/- 0.0016	0.8805 +/- 0.0912	0.0358 +/- 0.0016	0.1195 +/- 0.0912
LGBMRegressor	0.9133 +/- 0.0063	0.8376 +/- 0.0727	0.0867 +/- 0.0063	0.1624 +/- 0.0727
CatBoostRegressor	0.9635 +/- 0.0016	0.8812 +/- 0.0912	0.0365 +/- 0.0016	0.1188 +/- 0.0912

Appendix B. Complete results of the Machine learning analysis

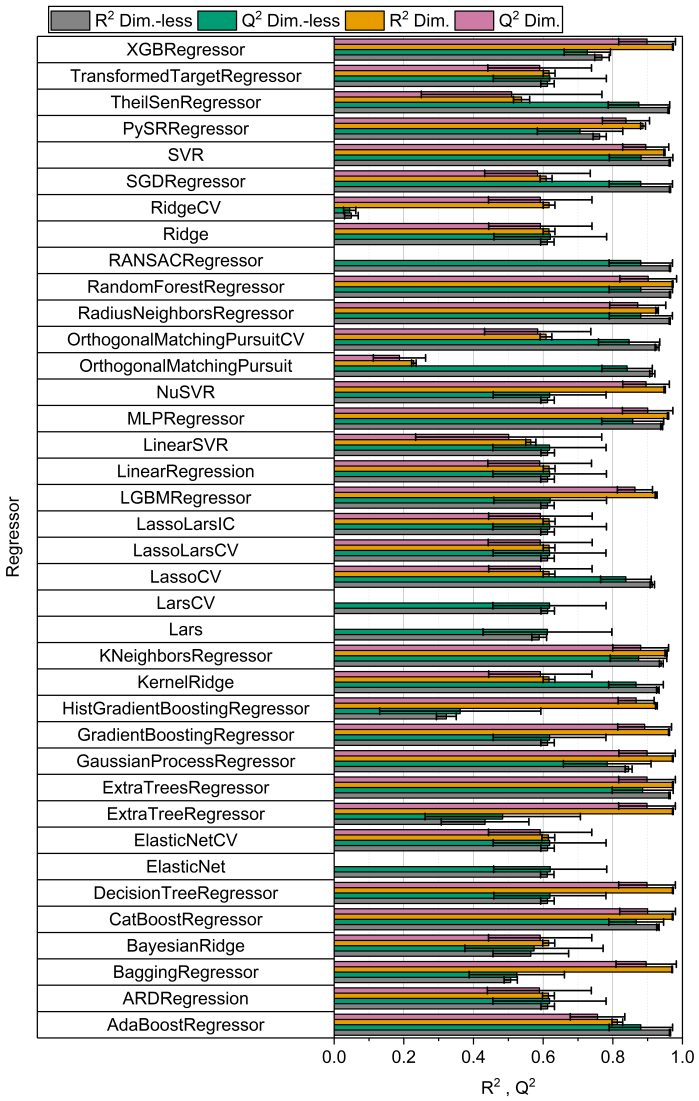


Figure B.6: Overview of ML results for all samples with all algorithms.

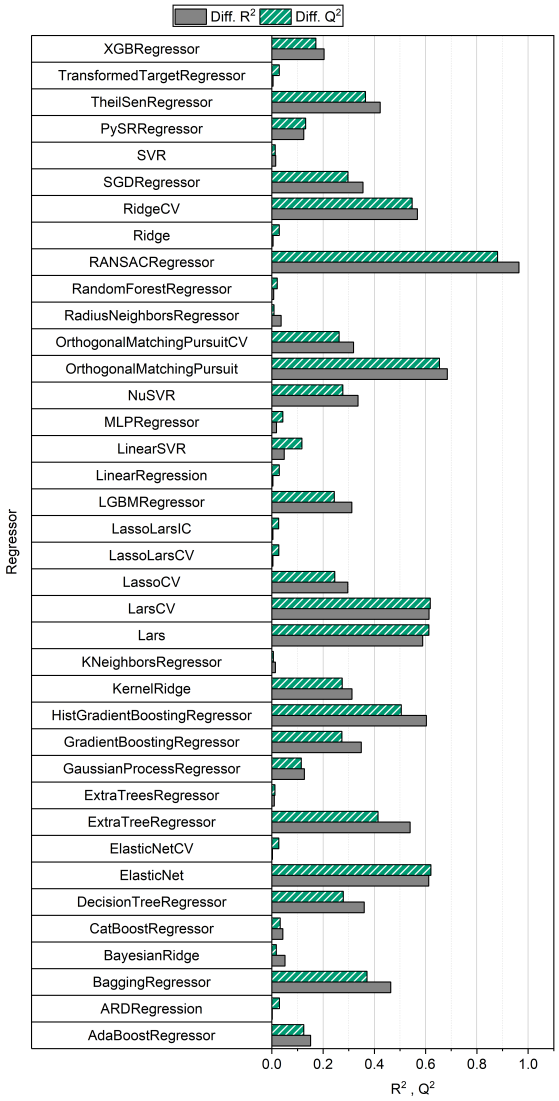


Figure B.7: Difference between dimensionless and dimensional ML results for all samples with all investigated algorithms.

References

[1] D. Butterworth, Design of shell-and-tube heat exchangers when the fouling depends on local temperature and velocity, Applied Thermal Engineering 22 (7) (2002) 789–801. doi:10.1016/S1359-4311(02)00025-X.

[2] N. Epstein, Thinking about heat transfer fouling: A 5 × 5 matrix, Heat Transfer Engineering 4 (1) (1983) 43–56. doi:10.1080/01457638108939594.

[3] H. Müller-Steinhagen, M. R. Malayeri, A. P. Watkinson, Fouling of heat exchangers-new approaches to solve an old problem, Heat Transfer Engineering 26 (1) (2005) 1–4. doi:10.1080/01457630590889906.

[4] H. Müller-Steinhagen, F. Reif, N. Epstein, A. P. Watkinson, Particulate fouling during boiling and non-boiling heat transfer, in: 8th Int. Heat Transfer Conference, Begel House Inc., Hemisphere, 1986, pp. 2555–2560.

[5] A. P. Hasting, Monitoring of Fouling, Cleaning, and Disinfection in Closed Processing Plants, in: Handbook of Hygiene Control in the Food Industry: Second Edition, Elsevier Ltd, 2016, pp. 663–672. doi:10.1016/B978-0-08-100155-4.00043-1.

[6] S. Flint, P. Bremer, J. Brooks, J. Palmer, F. A. Sadiq, B. Seale, K. H. Teh, S. Wu, S. N. Md Zain, Bacterial fouling in dairy processing, International Dairy Journal 101 (2020) 104593. doi:10.1016/j.idairyj.2019.104593.

[7] S. Elss, S. Kleinhenz, P. Schreier, Odor and taste thresholds of potential carry-over/off-flavor compounds in orange and apple juice, LWT - Food and Science Technology 40 (10) (2007) 1826–1831. doi:10.1016/j.lwt.2006.12.010.

[8] Gesellschaft VDI, VDI-Wärmeatlas, 11th Edition, Springer Berlin Heidelberg, Wiesbaden, 2013.

[9] P. J. Fryer, K. Asteriadou, A prototype cleaning map: A classification of industrial cleaning processes, Trends in Food Science and Technology 20 (6-7) (2009) 255–262. doi:10.1016/j.tifs.2009.03.005.

[10] D. I. Wilson, Fouling during food processing – progress in tackling this inconvenient truth, Current Opinion in Food Science 23 (2018) 105–112. doi:10.1016/j.cofs.2018.10.002.

[11] D. I. Wilson, Challenges in cleaning: Recent developments and future prospects, Heat Transfer Engineering 26 (1) (2005) 51–59. doi:10.1080/01457630590890175.

[12] G. Hauser, Hygienegerechte Apparate und Anlagen für die Lebensmittel-, Pharma- und Kosmetikindustrie, 1st Edition, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2008.

[13] F. Moerman, P. Rizoulières, F. A. Majoor, Cleaning in place (CIP) in food processing, in: Hygiene in Food Processing: Principles and Practice: Second Edition, Woodhead Publishing Limited, 2013, pp. 305–383. doi:10.1533/9780857098634.3.305.

[14] D. Nikoleiski, Principles of hygienic design, Journal of Hygienic Engineering and Design 1 (1) (2012) 15–18.

[15] S. Beal, Deposition of Particles in Turbulent Flow on Channel or Pipe Walls, Nuclear Science and Engineering 40 (1970) 1–11.

[16] R. Blöchl, H. Müller-Steinhagen, Influence of particle size and particle/fluid combination on particulate fouling in heat exchangers, The Canadian Journal of Chemical Engineering 68 (4) (1990) 585–591. doi:10.1002/cjce.5450680408.

[17] H. Müller-Steinhagen, Heat transfer fouling: 50 years after the Kern and Seaton model, Heat Transfer Engineering 32 (1) (2011) 1–13. doi:10.1080/01457632.2010.505127.

[18] M. M. Awad, Impact of Flow velocity on Surface Particulate Fouling-Theoretical Approach, Journal of American Science 8 (9) (2012) 1545–1003.

[19] J. Visser, T. J. Jeurnink, Fouling of heat exchangers in the dairy industry, Experimental Thermal and Fluid Science 14 (4) (1997) 407–424. doi:10.1016/S0894-1777(96)00142-2.

[20] J. Verran, Biofouling in food processing: Biofilm or biotransfer potential?, Food and Bioproducts Processing 80 (4) (2002) 292–298. doi:10.1205/096030802321154808.

[21] L. Schnöing, W. Augustin, S. Scholl, Fouling mitigation in food processes by modification of heat transfer surfaces: A review, Food and Bioproducts Processing 121 (2020) 1–19. doi:10.1016/j.fbp.2020.01.013.

[22] C. Françolle de Almeida, M. Saget, G. Delaplace, M. Jimenez, V. Fierro, A. Celzard, Innovative fouling-resistant materials for industrial heat exchangers: a review, Reviews in Chemical Engineering 39 (1) (2023) 71–104. doi:doi:10.1515/revce-2020-0094.

[23] M. Saget, C. F. de Almeida, V. Fierro, A. Celzard, G. Delaplace, V. Thomy, Y. Coffinier, M. Jimenez, A critical review on surface modifications mitigating dairy fouling, Comprehensive Reviews in Food Science and Food Safety 20 (5) (2021) 4324–4366. doi:10.1111/1541-4337.12794.

[24] H. Müller-Steinhagen, M. R. Malayeri, A. P. Watkinson, Heat exchanger fouling: Mitigation and cleaning strategies, Heat Transfer Engineering 32 (3-4) (2011) 189–196. doi:10.1080/01457632.2010.503108.

[25] L. Bouvier, I. Fargnier, S. Lalot, G. Delaplace, Effect of swirl flow on whey protein fouling and cleaning in a straight duct, Journal of Food Engineering 242 (2019) 115–123. doi:10.1016/j.jfoodeng.2018.08.024.

[26] W. Augustin, T. Fuchs, H. Föste, M. Schöler, J. P. Majschak, S. Scholl, Pulsed flow for enhanced cleaning in food processing, Food and Bioproducts Processing 88 (4) (2010) 384–391. doi:10.1016/j.fbp.2010.08.007.
URL http://dx.doi.org/10.1016/j.fbp.2010.08.007

[27] M. Lalande, F. Rene, J. P. Tissier, Fouling and its control in heat exchangers in the dairy industry, Biofouling 1 (3) (1989) 233–250. doi:10.1080/08927018909378111.

[28] O. Gudmundsson, S. Lalot, J. Thorsen, Comparison of Fouling Detection Methods Using Experimental Data, in: Proceedings of 10th interna-

tional Conference on Heat Exchanger Fouling and Cleaning, Budapest, 2013, pp. 429–436.

[29] F. Schlüter, L. Schnöing, H. Zettler, W. Augustin, S. Scholl, Measuring Local Crystallization Fouling in a Double-Pipe Heat Exchanger, Heat Transfer Engineering 41 (2) (2020) 149–159. doi:10.1080/01457632.2018.1522084.

[30] W. Liu, X. D. Chen, R. Jeantet, C. André, S. Bellayer, G. Delaplace, Effect of casein/whey ratio on the thermal denaturation of whey proteins and subsequent fouling in a plate heat exchanger, Journal of Food Engineering 289 (2021) 110–175. doi:10.1016/j.jfoodeng.2020.110175.

[31] L. A. Scudeller, P. Blanpain-Avet, T. Six, S. Bellayer, M. Jimenez, T. Croguennec, C. André, G. Delaplace, Calcium chelation by phosphate ions and its influence on fouling mechanisms of whey protein solutions in a plate heat exchanger, Foods 10 (2) (2021). doi:10.3390/foods10020259.

[32] D. K. Mohanty, P. M. Singru, Numerical method for heat transfer and fouling analysis of a shell and tube heat exchanger using statistical analysis, Korean Journal of Chemical Engineering 29 (9) (2012) 1144–1150. doi:10.1007/s11814-012-0003-6.

[33] E. Diaz-Bejarano, F. Coletti, S. MacChietto, A Model-Based Method for Visualization, Monitoring, and Diagnosis of Fouling in Heat Exchangers, Industrial and Engineering Chemistry Research 59 (10) (2020) 4602–4619. doi:10.1021/acs.iecr.9b05490.

[34] S. Alhuthali, G. Delaplace, S. Macchietto, L. Bouvier, Whey protein fouling prediction in plate heat exchanger by combining dynamic modelling, dimensional analysis, and symbolic regression, Food and Bioproducts Processing 134 (2022) 163–180. doi:10.1016/j.fbp.2022.05.009.

[35] F. Schlüter, W. Augustin, S. Scholl, Application of experimental data to model local fouling resistances, Heat and Mass Transfer 58 (2022) 29–40. doi:10.1007/s00231-021-03094-x.

[36] M. C. Ruzicka, On dimensionless numbers, Chemical Engineering Research and Design 86 (8) (2008) 835–868. doi:10.1016/j.cherd.2008.03.007.

[37] G. Delaplace, K. Loubière, F. Ducept, R. Jeantet, Dimensional analysis of food processes, Elsevier, 2015.

[38] E. Sritham, N. Nunak, E. Ongwongsakul, J. Chaishome, G. Schleining, T. Suesut, Development of Mathematical Model to Predict Soymilk Fouling Deposit Mass on Heat Transfer Surfaces Using Dimensional Analysis, Computation 11 (4) (2023). doi:10.3390/computation11040083.

[39] Y. Gu, L. Bouvier, A. Tonda, G. Delaplace, A mathematical model for the prediction of the whey protein fouling mass in a pilot scale plate heat exchanger, Food Control 106 (2019). doi:10.1016/j.foodcont.2019.106729.

[40] M. Khaldi, P. Blanpain-Avet, R. Guérin, G. Ronse, L. Bouvier, C. André, S. Bornaz, T. Croguennec, R. Jeantet, G. Delaplace, Effect of calcium content and flow regime on whey protein fouling and cleaning in a plate heat exchanger, Journal of Food Engineering 147 (C) (2015) 68–78. doi:10.1016/j.jfoodeng.2014.09.020.

[41] J. Petit, T. Six, A. Moreau, G. Ronse, G. Delaplace, β-lactoglobulin denaturation, aggregation, and fouling in a plate heat exchanger: Pilot-scale experiments and dimensional analysis, Chemical Engineering Science 101 (2013) 432–450. doi:10.1016/j.ces.2013.06.045.
URL http://dx.doi.org/10.1016/j.ces.2013.06.045

[42] H. Deponte, M. Helbig, N. Gottschalk, W. Augustin, S. Scholl, Dimensional Analysis of Cleaning-In-Place Processes for fouled organic material in Food Processes, in: 10th International Conference on Fouliing and Cleaning in Food Processing, Lund, Sweden, 2018.

[43] H. Kalman, Role of Reynolds and Archimedes numbers in particle- fluid flows, Reviews in Chemical Engineering 38 (2) (2022) 149–165.

[44] E. Rabinovich, H. Kalman, Incipient motion of individual particles in horizontal particle-fluid systems: B. Theoretical analysis, Powder Technology 192 (3) (2009) 326–338. doi:10.1016/j.powtec.2009.01.014.

[45] C. M. Bishop, N. M. Nasrabadi, Pattern recognition and machine learning, Vol. 4, Springer, 2006.

[46] V. N. Vapnik, A. Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory of Probability & Its Applications 16 (2) (1971) 264–280. doi:10.1137/1116025.

[47] S. Sundar, M. C. Rajagopal, H. Zhao, G. Kuntumalla, Y. Meng, H. C. Chang, C. Shao, P. Ferreira, N. Miljkovic, S. Sinha, S. Salapaka, Fouling modeling and prediction approach for heat exchangers using deep learning, International Journal of Heat and Mass Transfer 159 (2020) 120112. doi:10.1016/j.ijheatmasstransfer.2020.120112.

[48] S. Hosseini, A. Khandakar, M. E. Chowdhury, M. A. Ayari, T. Rahman, M. H. Chowdhury, B. Vaferi, Novel and robust machine learning approach for estimating the fouling factor in heat exchangers, Energy Reports 8 (2022) 8767–8776. doi:10.1016/j.egyr.2022.06.123.

[49] D. H. Kim, T. I. Zohdi, R. P. Singh, Modeling, simulation and machine learning for rapid process control of multiphase flowing foods, Computer Methods in Applied Mechanics and Engineering 371 (2020) 113286. doi:10.1016/j.cma.2020.113286.

[50] Y. Jiang, S. Yin, J. Dong, O. Kaynak, A Review on Soft Sensors for Monitoring, Control, and Optimization of Industrial Processes, IEEE Sensors Journal 21 (11) (2021) 12868–12881. doi:10.1109/JSEN.2020.3033153.

[51] R. Kasper, H. Deponte, A. Michel, J. Turnow, W. Augustin, S. Scholl, N. Kornev, Numerical investigation of the interaction between local flow structures and particulate fouling on structured heat transfer surfaces, International Journal of Heat and Fluid Flow 71 (2018) 68–79. doi:10.1016/j.ijheatfluidflow.2018.03.002.

[52] H. Deponte, L. Rohwer, W. Augustin, S. Scholl, Investigation of deposition and self-cleaning mechanism during particulate fouling on dimpled surfaces, Heat and Mass Transfer 55 (12) (2019) 3633–3644. doi:10.1007/s00231-019-02676-0.

[53] N. Jarmatz, W. Augustin, S. Scholl, Comprehensive Parameter Screening for the Investigation of Particulate Fouling in Pipe Fittings, Chemie-Ingenieur-Technik 95 (5) (2023) 708–716. doi:10.1002/cite.202200208.

[54] N. Jarmatz, W. Augustin, S. Scholl, Generation Of Experiental Data For Model Training To Optimize Fouling Prediction, Heat and Mass Transfer (2023). doi:10.1007/s00231-023-03393-5.

[55] S. Mueller, E. W. Llewellin, H. M. Mader, The rheology of suspensions of solid particles, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 466 (2116) (2009) 1201–1228. doi:10.1098/rspa.2009.0445.
URL https://doi.org/10.1098/rspa.2009.0445

[56] T. B. Lewis, L. E. Nielsen, Viscosity of Dispersed and Aggregated Suspensions of Spheres, Transactions of the Society of Rheology 12 (3) (1968) 421–443. doi:10.1122/1.549114.

[57] D. Chan, R. L. Powell, Rheology of suspensions of spherical particles in a newtonian and a non-newtonian fluid, Journal of Non-Newtonian Fluid Mechanics 15 (2) (1984) 165–179. doi:10.1016/0377-0257(84)80004-X.

[58] H. Deponte, A. Tonda, N. Gottschalk, L. Bouvier, G. Delaplace, W. Augustin, S. Scholl, Two complementary methods for the computational modeling of cleaning processes in food industry, Computers and Chemical Engineering 135 (2020) 106733. doi:10.1016/j.compchemeng.2020.106733.

[59] E. Buckingham, On Physically Similar Systems; Illustrations of the Use of Dimensional Equations, Physical Review 4 (4) (1914) 345–376. doi:10.1103/PhysRev.4.345.

[60] F. M. White, Fluid mechanics, 7th Edition, McGraw-Hill, 2011.

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[62] M. Cranmer, Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl (2023). arXiv:2305.01582, doi:10.48550/arXiv.2305.01582.

[63] K. Pearson, Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 187 (1896) 253–318.
URL http://www.jstor.org/stable/90707

[64] J. R. Koza, Genetic programming as a means for programming computers by natural selection, Statistics and Computing 4 (1994) 87–112. doi:10.1007/BF00175355.

[65] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, Classification and regression trees, CRC press, 1984.

[66] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[67] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.

[68] H. D. Baehr, K. Stephan, Wärme- und Stoffübertragung, 9. Auflage, Springer Vieweg (2016).

[69] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on tabular data? (2022). doi:10.48550/ARXIV.2207.08815.