Machine Learning based Models for Pressure Drop Estimation of Two-phase Adiabatic Air-Water Flow in Micro-finned Tubes: Determination of the Most Promising Dimensionless Feature Set

Behzad Najafi[1], Keivan Ardam, Andrej Hanušovský, Fabio Rinaldi, Luigi Pietro Maria Colombo

Dipartimento di Energia, Politecnico di Milano, Via Lambruschini 4, Milano 20156, Italy

Abstract

The present study is focused on determining the most promising set of dimensionless features and the optimal machine learning algorithm that can be employed for data-driven frictional pressure drop estimation of water (single-phase) and air-water mixture (two-phase) flow in micro-finned horizontal tubes. Accordingly, an experimental activity is first conducted, in which the frictional pressure drop of both water and air-water flows, at various flow conditions, is measured. Next, machine learning based pipelines are developed, in which dimensionless parameters are provided as features and the friction factor (for the single-phase case) and the two-phase flow multipliers (for the two-phase case) are considered as the targets. Next, the feature selection procedure is performed, in which the most promising set of features, while employing a benchmark algorithm, is determined. An algorithm optimization procedure is then performed in order to choose the most suitable algorithm (and the corresponding tuning parameters) that lead to the highest possible accuracy. Moreover, the state-of-the-art physical models are implemented and the corresponding accuracy, while being applied to the experimentally obtained dataset, is determined.

It is demonstrated that only 5 dimensionless features are selected (among 23 provided features) in the obtained pipeline developed for the estimation of the two-phase gas multiplier (in the extraction procedure of which, the single-phase friction factors are determined only using the Reynolds number and two geometrical parameters). Therefore, the latter procedures notably reduce the complexity of the model, while providing a higher accuracy (MARD of 6.72 % and 7.05 % on the training and test sets respectively) compared to the one achieved using the most promising available physical model

[1]behzad.najafi@polimi.it; Tel.: +39-02-2399-8518

(MARD of 15.21 %). Finally, through implementing the forward feature combination strategy on the optimal pipeline, the contribution of each feature to the achieved accuracy is shown and the trade-off between the model's complexity (number of features) and the obtained accuracy is presented. Thus, the latter step provides the possibility of utilizing an even inferior number of features, while achieving an acceptable accuracy. Moreover, since these pipelines will be made publicly accessible, the implemented models also offer a higher reproducibility and ease of use.

Keywords: Machine Learning, Pressure Drop Estimation, Two-Phase Flow, Feature Selection, Relative Feature Importance

---

1. Introduction

Modelling the behavior of multi-phase flows has become increasingly important in a wide variety of engineering applications as it facilitates system design and operation optimization of the units that include this type of flow [1]. The frictional pressure drop is one of the major characteristics of two-phase flows, the estimation of which is considered to be a complex problem as it depends on several parameters such as flow conditions, fluid properties, and geometry. The latter complexity is increased even further while utilizing micro-finned tubes (considering the influence of the number of employed fins and the corresponding geometrical properties). Although using micro-fins increases the frictional pressure drop, it enhances the heat transfer in tubes, through augmenting the heat exchange surface and accelerating the transition to a turbulent flow. Therefore, micro-finned tubes are widely employed, specifically in the air-conditioning and refrigeration industries [2].

Since the beginnings of the 20^{th} century, many studies have been carried out in this area and several models have been accordingly proposed in the literature for various applications [3]. In order to determine the frictional pressure gradient of two-phase flows in pipes, two distinct approaches are commonly employed. The first one is the so-called homogeneous methodology, in which the two-phase flow is considered as a single-phase one with physical properties that are defined by assigning suitable weights to the properties of individual phases. The pioneering models based on this approach were proposed by McAdams et al. [4], Beattie and Whalley [5], Awad and Muzychka [6], and Shannak [7]. The separated flow model is another approach, in which it is assumed that the two-phase pressure gradient is related to the pressure gradient corresponding to each phase that are taken into account separately. The pioneering work utilizing the separated flow model was proposed by Lockhart and

Nomenclature

Δp	Pressure drop $[Pa]$
$\frac{\Delta p}{\Delta z}$	Pressure gradient $[\frac{Pa}{m}]$
$\frac{Nl}{h}$	Normal litter per hour
AI	Artificial intelligence
ANN	Artificial neural network
CV	Cross-Validation
D_{int}	pipe internal diameter $[m]$
e	Mean fin height $[m]$
f	Friction factor $[-]$
Fr	Froude number $[-]$
G	Mass flux $[\frac{kg}{m^2\,s}]$
g	Gravitational acceleration $[\frac{m}{s^2}]$
G^*	Apparent mass flux $[\frac{kg}{m^2\,s}]$
J	Superficial velocity $\frac{m}{s}$
La	Laplace constant $[-]$
$MARD$	Mean absolute relative deviation [%]
ML	Machine learning
MRD	Mean relative deviation [%]
n	Number of fins $[-]$
p	Pressure $[Pa]$
Q	Volume flow rate $[\frac{m^3}{s}]$
Re	Reynolds number $[-]$
RF	Random Forest algorithm
S	Internal wetted perimeter $[m]$
SVM	Support Vector Machines
U	Phase velocity $\frac{m}{s}$
We	Weber number $[-]$
X	Lockhart-Martinelli parameter $[-]$
x	Average mass quality $[-]$
x_v	Average volume quality $[-]$
Y	Chisholm parameter $[-]$

Greek symbols

α	Void fraction $[-]$
β	Helix angle
Γ	Mass flow rate $[\frac{kg}{s}]$, non-dimensional parameter (Eq. 37)
μ	Dynamic viscosity $[Pa\,s]$
Ω	Cross section $[m^2]$
Φ^2	Two-phase flow friction multiplier $[-]$
ρ	Density $[\frac{kg}{m^3}]$
σ	Surface tension $[\frac{N}{m}]$
τ_w	Shear stress $[Pa]$

Subscripts

a	Accelerative
av	Average
b	Bulk
exp	Experimental value
f	Frictional
g	Gas, gravitational (Eq. 12)
go	Gas only
l	Liquid
lo	Liquid only
m	Micro-finned
ms	Manufacturer's specifications
$pred$	Predicted value
s	Smooth
tp	Two-phase
tt	Turbulent liquid, turbulent gas flow
tv	Turbulent liquid, laminar gas flow
vt	Laminar liquid, turbulent gas flow
vv	Laminar liquid, Laminar gas flow

Martinelli [8], based on an experimental analysis conducted on a circular pipe while utilizing two-phase mixtures of air with water, organic fluids, and various types of oils. The latter work was then extended in a study conducted by Chisholm [9], in which a simple model based on Lockhart-Martinelli charts, was introduced. Further improvement on the model proposed by Chisholm[9] was then carried out by Mishima and Hibiki [10] for viscous liquid and gas flows. Chisholm's Model was also extended by Sun

and Mishima [11], for air, water, CO_2, and refrigerant flows. Lockhart and Martinelli's methodology has also been utilized for estimating the pressure drop in non-straight pipes. In this context, Colombo et al. [12] proposed and implemented modifications to take into account the effect of centrifugal forces in steam-water flow passing through helical tubes, in which the model proposed in a previous study [13] was employed to estimate the single-phase laminar flow pressure drop. Another model based on this approach was proposed by Chisholm [14], which included the transformation of the Baroczy [15] plots. Muller-Steinhagen and Heck [16] also proposed a purely empirical model, obtained using a large dataset. Friedel [17] instead carried out an investigation, utilizing the separated flow model, while employing a large dataset and taking into account the effects of gravity and surface tension. For the specific case of refrigerant flows, other studies were conducted by Souza and Pimenta [18], Cavallini et al. [19] (condensation process), and Tran et al. [20].

Although empirical models have been widely employed for estimating the frictional pressure gradient in two-phase flows, owing to the non-linear behavior of the corresponding phenomena and its dependence upon several parameters, utilizing them does not necessarily result in achieving an acceptable accuracy. Accordingly, following the progress in the area of artificial intelligence, several data-driven methodologies and specifically machine learning (ML) based have been proposed for simulating the behavior of multi-phase flows. In this context, Artificial Neural Network (ANN) is one of the most commonly utilized methodologies for pressure drop estimation. ANN was employed for estimating the pressure drop of the oil-water-air mixture flowing through the horizontal and vertical tubes [21, 22]. This algorithm has also been employed for prediction of the pressure drop in R407C flow while undergoing evaporation inside horizontal smooth tubes [23], and non-Newtonian fluid passing through the piping components [24]. It has also been utilized in more recent studies for pressure drop estimation in CuO/(Ethylene glycol-water) nanofluid flows in car radiators [25] and non-azeotropic mixtures that are undergoing cryogenic forced boiling [26]. The group method data handling was integrated with an ANN to predict the frictional pressure drop occurring within mini-channel multi-port tubes while considering five different refrigerant fluids [27]. Different machine learning methods were utilized for predicting the pressure drop associated with R407C flow that is going through evaporation [28] and several fluids that are undergoing condensation in inclined smooth tubes [29]. Furthermore, employing the data obtained from an experimental study of horizontal gas-liquid two-phase flow in two medium-diameter pipes, back propagation neural networks based models were developed aiming at estimating the liquid holdup and pressure drop [30]. Genetic Algorithms were instead utilized in another

study [31] in order to optimize the artificial neural network models that were developed for predicting the steady-state two-phase pressure drop in pipes. Similar approaches have also been utilized in the area of chemical and petroleum industries such as predicting the bottom hole pressure in multi-phase flow in vertical oil production wells [32], estimating dew point pressure in retrograded condensate gas reservoirs [33], and monitoring the O2/CO/CO2 composition in oil production wells to evaluate the combustion quality through porous media [34]. In other investigations conducted in this area, least square support vector machine was employed for predicting CO2 solubility in brines [35], estimating the permeability and porosity of petroleum reservoirs [36], determining dew point pressure through condensate gas reservoirs [37], predicting the amount of dissolved calcium carbonate concentration throughout oil field brines [38], and determining condensate-to-gas ratio in retrograded condensate gas reservoirs [39]. Support Vector Machine (SVM) was also utilized for predicting the pressure drop of the air-water mixture, which is passing through the concentric annuli, and the corresponding performance was compared with the mechanistic models [40]. However, none of the previously conducted machine learning based studies that were aimed at estimating the pressure drop of two-phase flows, has been focused on optimizing the corresponding pipeline, including the selection of the most promising features and the optimal algorithm (among a large set of available algorithms and tuning parameters). Motivated by the latter research gap, the aim of the present study is developing optimal machine learning based pipelines for estimating the pressure drop in two-phase adiabatic flows, which are passing through micro-finned tubes. An experimental activity is accordingly conducted, in which the frictional pressure drop of both water and air-water flows, passing through this type of tubes at various flow conditions, is measured. Next, the physical models that are available in the literature, are implemented and the corresponding accuracy, while being applied to the experimentally obtained dataset, is determined. In the next step, machine learning based pipelines are implemented, in which dimensionless parameters are provided as features and the friction factor and two-phase flow multiplier (for the single-phase and two-phase case, respectively) are considered as targets.

The obtained dataset is then divided into training and testing sets, where the training set (employing the cross-validation method) is utilized in the feature selection and algorithm optimization steps, aiming at determining the most promising pipeline. The test set is instead utilized for determining the generalizability of the obtained pipeline. In the feature selection procedure, the most promising set of features, while employing a benchmark algorithm, are determined. The implemented pipelines are next optimized aiming at determining the machine learning algorithm and the corresponding tuning

parameters that lead to the highest possible accuracy.

The test set is then employed in order to assess the accuracy of the proposed pipeline while estimating the pressure drop in operating conditions for which it has not been trained or optimized. Finally, considering the trade-off between the model's complexity (in terms of the number of the employed features) and the obtained prediction precision, the accuracy of the optimized pipelines with lower number of features is progressively determined and presented. Accordingly, compared to the previously conducted studies focused on machine learning/artificial intelligence based estimation of pressure drop in two-phase flows, the present study offers the following contributions

- Conducting the feature selection step, while considering a large set of dimensionless parameters (using which enhances the generalisability of the pipelines), which permits reducing the number of features (thus, notably decreasing the computational cost), increasing the achieved accuracy, and improving the physical interpretability of the results.
- Conducting the algorithm optimization step, in which a large set of algorithms and wide range of corresponding tuning parameters are considered, aiming at maximizing the obtained accuracy.
- Demonstrating the trade-off between the number of utilized features and the achieved accuracy (employing the obtained optimal pipeline), which permits the user to utilize less complex models while achieving an acceptable accuracy.
- Providing the dataset and the optimized pipeline (in python programming language) as a publicly accessible tool, which enhances the reproducibility and ease of use of the proposed method.

2. Experimental activity and employed dataset

The frictional pressure drop measurement procedure was conducted on adiabatic streams of water (single-phase) and air-water mixture (two-phase) with various flow rates [41], while passing through horizontal micro-finned copper tubes. The characteristics of the utilized tubes are provided in the supplementary material.

2.1. Description of the laboratory set up

Fig. 1 shows the simplified layout of the employed laboratory setup. Water collected in the storage tank is pumped to the mixing section and passes through the temperature, pressure, and volumetric

flow rate measurement sections before streaming to the test section and being re-circulated through the loop. Three parallel rotameters, each of which is suitable for a different range of flow rates, are employed to measure the volumetric flow rate of water.

Air is compressed and injected into the flow loop using the building's auxiliary supply system and the corresponding temperature, pressure, and volumetric flow rate are measured and controlled. Similarly, the volumetric flow rate of the air is measured by a set of three parallel rotameters.

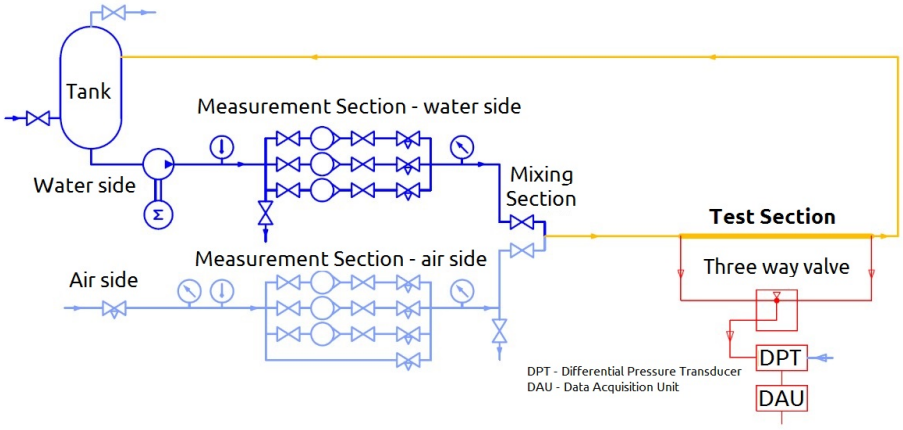


Figure 1: A schematic representation of the laboratory set-up utilized for conducting the experiments

In the test section, two types of copper micro-finned tubes are used. Two pressure taps are installed at the inlet and the outlet of the test section, which are connected to a differential pressure transducer (DPT). Two nylon tubes, filled with water, are utilized to create a connection between each pressure tap and the pressure transducer in order to hydraulically transmit the pressure variation. Utilizing a by-pass valve provides the possibility of measuring either the pressure drop between the two taps or the pressure difference between each tap and the environment, which helps to reduce the signal-to-noise ratio in some operating conditions. The DPT sends the electrical signals to a data acquisition unit (DAU), which works at a sampling frequency of 1 kHz with an acquisition time of 15 s.

Rotameters with flow rate ranges of 4-190, 85-850, and 400-4000 $\frac{Nl}{h}$ are utilized for air, while the rotameters employed for measuring the flow rate of water have ranges of 10-100 and 40-400 $\frac{Nl}{h}$. The uncertainty of all of the utilized rotameters is 3% on the observed value. The measurement range of the differential pressure transducer is instead 0-70 $[kPa]$ (with the uncertainty of 1.5% on the full scale). The measurement range of the manometer is 0-6 bar g with an uncertainty of 0.2 bar. The

measurement range of the thermometer is 5 to 120 [$^{\circ}C$] with an uncertainty of 1 degree. It is worth mentioning that, as the considered flow in the present study is adiabatic, the temperature measurement is only employed in the data processing procedure; therefore, the uncertainty of the thermometer has a negligible impact on the final results.

2.2. Measurement and data processing procedures for the single-phase case

2.2.1. Measurement procedure

The experiments were conducted at various volumetric flow rates by increasing it step-by-step in the range of 50 to 400 [$\frac{Nl}{h}$] with the step of 25 [$\frac{Nl}{h}$]. For each operating condition, the inlet and the outlet pressures along with the inlet temperature were measured. This procedure was repeated 10 times (averaged as a single point), resulting in 15 experimental data points for each of the micro-finned tubes.

2.2.2. Data processing procedure

In the data processing step, all of the fluid properties were determined at the arithmetic average of pressure (between the inlet and the outlet) and the inlet temperature. The difference between the measured pressures at the inlet and outlet was then utilized to determine the pressure drop across the test section. The pressure gradient was determined as the ratio of the pressure drop to the distance between the pressure taps (as the flow is fully developed).

Considering the employed volumetric flow rates, the ranges of mass flux for micro-finned tube 1 is determined to be 224.3 to 1794.5, which corresponds to Reynolds number between 2300 to 18396. For micro-finned tube 2 instead, the mass flux range is demonstrated to be 335.5 to 1779.1 [$\frac{kg}{(m^2\,s)}$] corresponding to the Reynolds number range of 3673 to 19592. The range of measured pressure gradient is 183.1 to 7050.9 [$\frac{Pa}{m}$] and 168.2 to 7852.3 [$\frac{Pa}{m}$] for micro-finned tubes 1 and 2 respectively. It is worth noting that, while analyzing the results of the experiment on micro-finned tube 2 with the lowest volume flow rate, it was observed that the fluctuations were excessive, because of which the corresponding 10 data-points were discarded.

2.3. Measurement and data processing procedures for the two-phase case

2.3.1. Measurement procedure

The experiments on the water-air mixture were conducted using different combinations of water and air volumetric flow rates. In this context, the flow rates in the range of 10 to 100 [$\frac{Nl}{h}$] with a step

of 10 $[\frac{Nl}{h}]$ were considered for water, while flow rates of 500 to 4000 $[\frac{Nl}{h}]$ with a step of 500 $[\frac{Nl}{h}]$ were employed for the air flow. Starting from the lowest water flow rate, the flow rate of air was progressively changed (following the above-mentioned values). The same procedure was performed for the next planned values of water's flow rate in order to explore the whole range. In each step, the steady-state was reached after iterative adjustment of the volumetric flow rates (using the regulation valves) in order to keep the pressure (in the measurement section) constant. Measurements in each step were repeated 5 to 10 times (depending on the observed fluctuations) and averaged as single data-points, which results in 74 measured data points for micro-finned tube 1 and 75 ones for micro-finned tube 2. Throughout the experiment, the pressure of air and water streams (both in the measurement section) was kept constant at 3 [bar] and 2.2 [bar] respectively. The temperature was measured and recorded at the beginning of each test, which was insignificantly variable between 19 to 21 $^{\circ}C$ for air and 22 and 25 $^{\circ}C$ for water.

2.3.2. Data processing procedure

The observed value using air flow meters corresponds to the normal conditions defined as $T_0 = 20$ $^{\circ}C$ and $P_0 = 101325Pa$. According to the manufacturer's specifications, through considering the temperature (T_{ms}) and pressure (p_{ms}) at the measurement section and the average pipe pressure (p_{av}), the actual volume flow rate is determined as:

$$Q = Q_0 \sqrt{\frac{p_{ms}\, p_0}{p_{av}^2} \frac{T_{ms}}{T_0}} \tag{1}$$

Considering the utilized volumetric flow rates, the air mass flux is determined to be in the range of 5.3 to 42.8 $[\frac{kg}{(m^2\, s)}]$, the water mass flux is between 44.4 and 444.8 $[\frac{kg}{(m^2\, s)}]$, while the mass quality range is 0.012 to 0.49 $[-]$. The measured pressure gradient is between 297.4 to 19720.9 $[\frac{Pa}{m}]$ for micro-finned tube 1 and 154.5 to 22979.3 $[\frac{Pa}{m}]$ for micro-finned tube 2.

3. Gas-liquid flow fundamentals

3.1. Characteristic quantities

The essential quantities that are needed in order to provide a complete description of two-phase (gas-liquid) flows are reported in Table 1.

Characteristic quantity	Symbol and equation	Description	Unit	Eq. No
Total mass flow rate	$\Gamma = \Gamma_g + \Gamma_l$	Sum of the mass flow rates of the gas and liquid phases	$[\frac{kg}{s}]$	(2)
Total volumetric flow rate	$Q = Q_g + Q_l$	Sum of the volume flow rates of both phases	$[\frac{m^3}{s}]$	(3)
Mass quality	$x = \frac{\Gamma_g}{\Gamma}$	The ratio between the mass flow rate of the gas phase to the total mass flow rate	$[-]$	(4)
Volume quality	$x_v = \frac{Q_g}{Q}$	The ratio between the mass flow rate of the gas phase to the total mass flow rate	$[-]$	(5)
Void fraction	$\alpha = \frac{\Omega_g}{\Omega}$	The ratio of the cross section area occupied by the gas phase to the total cross section area of the pipe	$[-]$	(6)
Phase velocity	$U_i = \frac{Q_i}{\Omega_i}$	The ratio of the phase volume flow rate to the cross section area occupied by the phase	$[\frac{m}{s}]$	(7)
Superficial velocity	$J_i = \frac{Q_i}{\Omega}$	The ratio of the phase volume flow rate to the total cross section area	$[\frac{m}{s}]$	(8)
Mass flux	$G_i = \frac{\Gamma_i}{\Omega_i}$	The ratio of the phase mass flow rate to the cross section area occupied by the phase	$[\frac{kg}{m^2\,s}]$	(9)
Apparent mass flux	$G_i^* = \frac{\Gamma_i}{\Omega}$	The ratio of the phase mass flow rate to the total cross section area	$[\frac{kg}{m^2\,s}]$	(10)
Bulk density	$\rho_b = \rho_g x_v + \rho_l(1-x_v)$ $= (\frac{x}{\rho_g} + \frac{1-x}{\rho_l})^{-1}$	Mixture density based on volume quality weighted average	$[\frac{kg}{m^3}]$	(11)

Table 1: Utilized characteristic quantities and the corresponding descriptions

3.2. Pressure gradient models

Based on the above-mentioned parameters, considering a flow element, one-dimensional momentum balance for a constant cross-section pipe can be employed to express the total pressure gradient (Eq. 12) in which the subscripts a, f, and g indicate accelerative, frictional, and gravitational components respectively. Additionally, τ_w is the shear stress, S is the wetted perimeter, g is the gravitational acceleration, and θ represents the duct inclination.

$$-\frac{dp}{dz} = -(\frac{dp}{dz})_a - (\frac{dp}{dz})_f - (\frac{dp}{dz})_g = G^2\frac{d}{dz}(\frac{x^2}{\rho_g\alpha} + \frac{(1-x)^2}{\rho_l(1-\alpha)}) + \tau_w\frac{S}{\Omega} + [\rho_g\alpha + \rho_l(1-\alpha)]gsin\theta \quad (12)$$

As the considered flow is adiabatic and horizontal and the mass quality is constant, gravitational and accelerative pressure gradients are absent [42]. The total pressure gradient is therefore only due to the friction; thus, Eq. 12 can be re-written as:

$$-\frac{dp}{dz} = \tau_w\frac{S}{\Omega} \quad (13)$$

3.2.1. Frictional pressure gradient for the single-phase flow

For the single-phase flow, frictional pressure gradient is obtained from the momentum balance, in which the wall shear stress τ_w is the only unknown that can be expressed in a dimensionless form using Fanning friction factor:

$$f = \frac{\tau_w}{\frac{1}{2}\rho U^2} = \frac{2\tau_w \rho}{G^2} \tag{14}$$

By integrating along the pipe axis and and replacing the latter in Eq. 13, the following correlation can be obtained.

$$\frac{\Delta p_f}{\Delta z} = \frac{2fG^2}{\rho D} \tag{15}$$

Where G represents the mass flux, U is the cross section average velocity of the flow, and Δp_f is the frictional pressure drop across a pipe length denoted as Δz. Thus, the friction factor can be experimentally obtained through the measurement of the frictional pressure drop using the following equation:

$$f_{exp} = \frac{\Delta p_{f,exp}}{\Delta z}.(\frac{2G^2}{\rho D})^{-1} \tag{16}$$

which in turn is a function of the Reynolds number:

$$Re = \frac{GD}{\mu} \tag{17}$$

For the particular case of laminar flow in smooth pipes ($Re < 2300$):

$$f = \frac{16}{Re} \tag{18}$$

For turbulent flow ($Re > 2300$) in smooth tubes instead, several models have been proposed in the literature aiming at extending the corresponding range of validity and improving prediction accuracy. Some of the most popular models in this area are reported below.

$$\text{Blasius [43]:} \quad \begin{array}{ll} f = 0.079\ Re^{-0.25} & \text{for } 2300 < Re < 20000 \\ f = 0.046 Re^{-0.2} & \text{for } Re > 20000 \end{array} \tag{19}$$

$$\text{Filonenko [44]:} \quad f = (1.58 \ln Re - 3.28)^{-}2 \quad \text{for } Re > 2300 \tag{20}$$

Nikuradse [45]: $$\frac{1}{\sqrt{f}} = 4\log\left(2Re\sqrt{f}\right) - 1.6 \quad \text{for } Re > 2300 \tag{21}$$

Fang et al. [46]: $$f = 0.0625[\log\left(\frac{150.39}{Re^{0.98865}}\right) - \frac{152.66}{Re}]^{-2} \quad \text{for } Re > 2300 \tag{22}$$

For the case of finned tubes, correlations for friction factor prediction have been proposed by Jensen and Vlakancic [47] and Haaland [48]. The correlation proposed by Jensen and Vlakancic [47] is as follows:

$$\frac{f}{f_{Filonenko}} = (\frac{l_c}{D})^{-1.25}(\frac{\Omega_s}{\Omega_m})^{1.75} - (\frac{0.0151}{f_{Filonenko}})[(\frac{l_c}{D})^{-1.25}(\frac{\Omega_s}{\Omega_m})^{1.75} - 1]\exp\left(-\frac{Re}{6780}\right)$$

$$\frac{l_c}{D} = 1 - 1.577(\frac{nsin\beta}{\pi})^{0.64}(\frac{2e}{D})^{0.53}(\frac{\pi}{n} - \frac{s}{D})] \quad \text{for } Re > 2300 \tag{23}$$

Where n is the number of fins, β is the fin helix angle, e is height of the fin, D is the internal diameter at the roots of the fins, s isthe mean fin thickness, l_c is the characteristic length and $\frac{\Omega_s}{\Omega_m}$ is the ratio between the cross sections of the smooth and micro-finned tube with the same internal diameter. The model proposed by Haaland [48] is instead as follows:

$$f = \frac{0.3086}{\{\log[\frac{6.9}{Re} + (\frac{e}{3.7D})^{1.11}]\}^2} \quad \text{for } Re > 2300 \tag{24}$$

Where e is height of the fin and D is internal diameter at the roots of the fins.

3.2.2. Two-phase frictional pressure gradient

Prediction of the frictional pressure drop for two-phase flow is more complicated than the single-phase, which is attributed to the corresponding higher complexity of motion. The phases that are flowing together demonstrate different arrangements (flow patterns), which strongly affect their mutual interaction and the interaction between each phase and the pipe. For engineering purposes, the so-called flow pattern independent models have been introduced in the literature aiming at simplifying the description of the above-mentioned behavior, in spite of their high degree of approximation. In this framework, two different approaches can be utilized:

1. **Homogeneous flow model**: The two-phase mixture is assumed as a single-phase fluid while considering averaged properties. This approach is suitable only if the slip ratio is equal to one (homogeneous flow, s = 1). The latter is approximately achieved if the velocity of the mixture is very high resulting in almost flat velocity and concentration profiles.

2. **Separated flow model**: The two-phase mixture is considered to be consisting of two independent single-phase streams, each of which is flowing separately. In this approach, the overall pressure gradient is determined by a suitable combination of the pressure gradients of single-phase streams. It is employed if the slip ratio is notably different from unity.

Details of the latter approaches are provided in the following section.

3.2.3. Homogeneous approach

In the homogeneous approach, two-phase flow frictional pressure gradient is determined as:

$$-\frac{dp_f}{dz} = \frac{2 f_{tp} G_{tp}^2}{\rho_b D} \tag{25}$$

in which f_{tp} is commonly evaluated as a function of a two-phase Reynolds number which involves the definition of an average two-phase dynamic viscosity μ_{tp}. Some of the most promising models [4–7], which have employed the homogeneous approach are listed below.

$$\text{McAdams et al. [4]:} \quad \frac{1}{\mu_{tp}} = \frac{x}{\mu_g} + \frac{1-x}{\mu_l} \tag{26}$$

Beattie and Whalley [5]:

$$\mu_{tp} = \mu_l(1-\beta)(1+2.5\beta) + \mu_g\beta$$
$$\beta = \frac{x}{x + (1-x)\frac{\rho_g}{\rho_l}} \tag{27}$$

Awad and Muzychka [6]:

$$\mu_{tp} = \mu_g \frac{2\mu_g + \mu_l - 2(\mu_g - \mu_l)(1-x)}{2\mu_g + \mu_l + (\mu_g - \mu_l)(1-x)} \tag{28}$$

Shannak [7]:

$$\mu_{tp} = \frac{\mu_g\, x + \mu_l\,(1-x)\,\frac{\rho_g}{\rho_l}}{x^2 + (1-x)^2\,\frac{\rho_g}{\rho_l}} \tag{29}$$

3.2.4. Separated flow approach

It is convenient to employ a dimensionless factor, which correlates the two-phase frictional pressure gradient to the single-phase one. These defined factors are called two-phase multipliers as (for gas-liquid flows) the two-phase frictional pressure gradient is higher than the single-phase one. In this context, four different two-phase multipliers that can be utilized are listed below.

1. The single-phase flow is the liquid flowing alone, i.e. at the liquid superficial velocity (l –liquid alone).
2. The single-phase flow is the liquid flowing with the total flow rate, i.e. at the mixture velocity (lo – liquid only).
3. The single-phase flow is the gas flowing alone, i.e. at the gas superficial velocity (g – gas alone)
4. The single-phase flow is the gas flowing with the total flow rate, i.e. at the mixture velocity (go – gas only)

Accordingly, the four different, but equivalent, factors are expressed as follows:

$$\Phi_l^2 = \frac{\frac{dp_f}{dz}}{(\frac{dp_f}{dz})_l},\quad \Phi_{lo}^2 = \frac{\frac{dp_f}{dz}}{(\frac{dp_f}{dz})_{lo}},\quad \Phi_g^2 = \frac{\frac{dp_f}{dz}}{(\frac{dp_f}{dz})_g},\quad \Phi_{go}^2 = \frac{\frac{dp_f}{dz}}{(\frac{dp_f}{dz})_{go}} \tag{30}$$

Similar to the previous part, some of the most promising predicting models, available in the literature [9–11, 14, 16–20, 49–51], in which the separated flow approach is utilized, are listed in Table 2.

Author/s and the corresponding reference	Equation	Eq. No
Chisholm [9]	$\Phi_l^2 = 1 + \frac{C}{X} + \frac{1}{X^2}$ $X^2 = \frac{(\frac{\Delta p}{\Delta z})_l}{(\frac{\Delta p}{\Delta z})_g}$ $C = 5$ for $vv, C = 10$ for tv $C = 12$ for $vt,\ C = 20$ for tt	(31)
Mishima and Hibiki [10]	$C = 21[1 - \exp(-0.319 D_{int})]$	(32)
Zhang et al. [49]	$C = 21[1 - \exp\Big(-\frac{a}{La}\Big)]$ $La = \frac{\sqrt{\frac{\sigma}{g(\rho_l - \rho_g)}}}{D_{int}}$ For gas and liquid $a = 0.647$ For vapor and liquid $a = 0.142$	(33)
Sun and Mishima [11]	For viscous flow: $C = 26(1 + \frac{Re_l}{1000})[1 - \exp\Big(\frac{-0.153}{0.8 + 0.27La}\Big)]$ For turbulent flow: $\Phi_l^2 = 1 + \frac{C}{X^{1.19}} + \frac{1}{X^2}$ $C = 1.79(\frac{Re_g}{Re_l})^{0.4} \sqrt{\frac{1-x}{x}}$ where $Re_g = \frac{G_{tp}\, x\, D_{int}}{\mu_g}$, $Re_l = \frac{G_{tp}(1-x)D_{int}}{\mu_l}$	(34)

Continue on the next page

Author/s and the corresponding reference	Equation	Eq. No
Chisholm [14]	$\Phi_{lo}^2 = 1 + (Y^2 - 1)\{B[x(1-x)]^{0.875} + x^{1.75}\}$ $Y^2 = \dfrac{(\dfrac{\Delta p}{\Delta z})_{go}}{(\dfrac{\Delta p}{\Delta z})_{lo}}$ if $0 < Y < 9.5$, $B = \begin{cases} 4.8 & G_{tp} \leq 500 \\ 2400 & 500 < G_{tp} < 1900 \\ \dfrac{55}{G_{tp}^{0.5}} & G_{tp} \geq 1900 \end{cases}$ if $9.5 < Y < 28$, $B = \begin{cases} \dfrac{520}{Y G_{tp}^{0.5}} & G_{tp} \leq 600 \\ \dfrac{21}{Y} & G_{tp} > 600 \end{cases}$ if $Y > 28$, $B = \dfrac{15000}{Y^2\, G_{tp}^{0.5}}$	(35)
Muller-Steinhagen and Heck [16]	$\Phi_{lo}^2 = Y^2 x^3 + (1-x)^{0.333}[1 + 2x(Y^2 - 1)]$	(36)
Souza and Pimenta [18]	$\Phi_{lo}^2 = 1 + (\Gamma^2 - 1)x^{1.75}(1 + 0.9524\,\Gamma\, X_{tt}^{0.4126})$ $\Gamma = (\dfrac{\rho_l}{\rho_g})^{0.5}\,(\dfrac{\mu_g}{\mu_l})^{0.125},\ X_{tt} = \dfrac{1}{\Gamma}(\dfrac{1-x}{x})^{0.875}$	(37)
Friedel [17]	$\Phi_{lo}^2 = (1-x)^2 + x^2 \dfrac{\rho_l\, f_{go}}{\rho_g\, f_{lo}} + \dfrac{3.24\, x^{0.78}(1-x)^{0.224}\, H}{Fr_{tp}^{0.045}\, We_{tp}^{0.035}}$ $H = (\dfrac{\rho_l}{\rho_g})^{0.91}\,(\dfrac{\mu_g}{\mu_l})^{0.19}\,(1 - \dfrac{\mu_g}{\mu_l})^{0.7}$ $We_{tp} = \dfrac{G_{tp}^2\, D_{int}}{\sigma\, \rho_{tp}},\ Fr_{tp} = \dfrac{G_{tp}^2}{g\, D_{int}\, \rho_{tp}^2},\ \dfrac{1}{\rho_{tp}} = \dfrac{x}{\rho_g} + \dfrac{1-x}{\rho_l}$	(38)
Cavallini et al. [19]	$\Phi_{lo}^2 = (1-x)^2 + x^2 \dfrac{\rho_l\, f_{go}}{\rho_g\, f_{lo}} + \dfrac{1.262\, x^{0.6987}\, H}{We_{go}^{0.1458}}$ $H = (\dfrac{\rho_l}{\rho_g})^{0.3278}\,(\dfrac{\mu_g}{\mu_l})^{-1.181}\,(1 - \dfrac{\mu_g}{\mu_l})^{3.477}$ $We_{go} = \dfrac{G_{tp}^2\, D_{int}}{\sigma\, \rho_g},$	(39)
Tran et al. [20]	$\Phi_{lo}^2 = 1 + (4.3\, Y^2 - 1)\{[x\,(1-x)]^{0.875}\, La + x^{1.75}\}$	(40)

Continue on the next page

Author/s and the corresponding reference	Equation	Eq. No
Haraguchi et al. [50]	$\Phi_g^2 = [1.1 + 1.3(\frac{X_{tt}\, G_{tp}}{\sqrt{g\, D_e\, \rho_g\, (\rho_l - \rho_g)}})^{0.35}]^2$ $D_e = \sqrt{\frac{4\Omega}{\pi}}$ $X_{tt} = (\frac{1-x}{x})^{0.9}\, (\frac{\rho_g}{\rho_l})^{0.5}\, (\frac{\mu_l}{\mu_g})^{0.1}$	(41)
Goto et al. [51]	$\Phi_g^2 = [1 + 1.64\, X^{0.79}]^2$	(42)

Table 2: Selected models that have utilized the separated flow approach

4. Methodology and implemented pipelines

Machine learning teaches the computer to solve problems by looking at hundreds or thousands of examples, learning from them, and then using that experience to solve the same problem in new situations. Two key machine learning settings are supervised learning and unsupervised learning [52]. The ultimate goal of supervised learning is to identify a function f that produces accurate predictions. Two classic supervised learning tasks are classification and regression. In classification, the output domain is a finite discrete set of categories, whereas in regression the output domain is a continuous value [52]. The most prominent example of unsupervised learning is data clustering. In clustering, the goal is to construct a function f that partitions an unlabeled dataset into different clusters.

In the present study, several supervised (regression) [52] machine learning based pipelines are implemented in order to determine the most promising one, which leads to the most accurate prediction of the pressure drop. In the first step, the dataset is divided into a training set (that is also used for validation through cross-validation) and a test set. The training set is employed for determining the optimal pipeline. The test set is instead utilized for assessing the accuracy of the model while being applied to a dataset for which it has not been tuned. At first, all of the possible features are provided to the random forest algorithm (described in section 4.3) and the resulting accuracy metrics are obtained for both the training set (through cross-validation) and the test set (after training the model on the training set).

In the second step, while employing random forest algorithm, a feature selection procedure is implemented (using the training set) in order to determine the most promising combination of features that

lead to the highest possible accuracy [53]. In this methodology, the features are progressively added, after being ordered based on their correlation (Pearson's correlation coefficient [54]) to the target, and the obtained accuracy is monitored. The set of features that results in the highest possible accuracy (lowest error) is then chosen as the initial most promising combination. Next, among the remaining features, the ones that increase the overall obtained error are discarded, while the remaining ones are sorted in a descending order based on the overall error obtained while adding them. The latter set of features is then placed before (at the beginning of) the previously determined initial promising combination. In the last step, features are progressively eliminated and the final set of features, using which results in the lowest error, is obtained.

Next, while providing the selected features as inputs, employing a pipeline optimization tool [55], a genetic algorithm [56–58] based optimization procedure is carried out (using the training set) to determine the most promising machine learning algorithm along with the corresponding tuning parameters based on the training set. Finally, in order to assess the generalizability of the optimal pipeline (that is obtained using the feature selection and algorithm optimization steps) its accuracy while being applied to the test set is evaluated. The Mean Relative Deviation (MRD) (Eq. 43), and the Mean Absolute Relative Deviation (MARD) (Eq. 44) are utilized for representing the error, among which MARD is considered as the key accuracy metric in order to choose the most promising models.

$$MRD = \frac{1}{N} \sum_{i=1}^{N} \frac{y_{i,pred} - y_{i,exp}}{y_{i,exp}} \quad [\%] \tag{43}$$

$$MARD = \frac{1}{N} \sum_{i=1}^{N} \frac{\mid y_{i,pred} - y_{i,exp} \mid}{y_{i,exp}} \quad [\%] \tag{44}$$

It is noteworthy to mention that the cross-validation strategy is employed for obtaining the latter accuracy metrics for feature selection and pipeline optimization on the training set. In the implemented cross-validation strategy, the training set is divided into 10 subsets (k = 10) and each unique subset once plays the validation set while the remaining subsets play the training set. The latter procedure provides more accurate estimate of out-of-sample accuracy.

Considering the fact that random forest is utilized as the benchmark algorithm, a brief description of this algorithm, and its underlying methodologies (decision trees and ensemble methods) is provided in the following sub-sections. Furthermore, a brief description of other algorithms, which are utilized in the determined optimal pipelines is also provided.

4.1. Decision trees

In decision trees, the value of target is estimated by learning simple decision based rules inferred from the features [59, 60].

4.2. Ensemble methods

In ensemble methods, the estimations of several estimators (each of which is developed using an algorithm) are put together. This approach helps to improve the generalizability of the model compared to single estimators. Two main families of ensemble methods, including averaging and boosting, are commonly utilized.

In the averaging approach (e.g. random forest and bagging), several independent estimators are created and the average of their prediction is provided as the output. The provided average prediction, owing to its lower variance, commonly has a better performance than any of the individual estimators.

In boosting methods (e.g. the gradient boosting), the estimators are created in a sequential manner with the objective of decreasing the bias in the final estimator. Thus, a powerful ensemble is created from several weak models [59]. High bias (the difference between the average prediction of the model and the correct value) oversimplifies the model which leads to a high error on training and test data. The model with high variance (variability of model prediction for a given data point) instead pays a lot of attention to training data and does not generalize on the data which it has not seen before. As a result, such models perform very well on training data but have high error rates on test data. There is a trade-off between a model's ability to minimize both bias and variance.

4.3. Random forest

In this algorithm, a set of decision tree-based estimators are built through introducing randomness in the corresponding creation procedure [59]. Each estimator (which is a tree [61]) is built using a sample that is extracted from the training set and a random portion of the features. The latter randomness results in a slight increment in the bias (in comparison with a single tree). However, owing to the employed averaging, the resulting reduction in the variance is dominant over the effect of the observed increase in the bias. Therefore, the overall performance of the model, compared to that of a single tree, is improved [59].

4.4. Gradient tree boosting

The key concept of boosting is to fit a sequence of weak learners on a dataset that is repeatedly modified. Each weak learner is a model that is marginally better than random guessing (as an example a small decision tree). In order to generate the final prediction, the predictions of all weak learners are combined through a weighted majority vote or summing. The modification that is conducted on the dataset is applying weights to each sample of the training dataset. These weights, which were all initialized as $\frac{1}{N}$ (N is the number of samples), are individually modified in the next iterations and the learning algorithm is applied to the modified data. At each iteration, the training samples that were incorrectly predicted, have their weights increased while the weights of those that were predicted correctly are reduced. Accordingly, samples that are difficult to predict are given a higher weight and thus a higher influence. Thus, each weak learner is enforced to focus on the samples which could not be correctly predicted by the previous weak learner in the sequence [59, 62, 63]. Gradient tree boosting, which is also called gradient boosted regression trees (GBRT), is a generalization of boosting to loss functions that are arbitrarily differentiable. In this ensemble method, the employed weak learners are decision trees of fixed sizes. Through implementing the abovementioned procedure, the learning process becomes a minimization problem, which is solved numerically [59, 64].

4.5. K nearest neighbors

In the k nearest neighbors algorithm, the average (weighted or uniform) of the values of k (a chosen value) neighbors is provided as the estimated output. The distance based method (in which each neighbor's weight is assigned based on the inverse of the corresponding distance (Manhattan or Euclidian) from the query point [65]) is the most commonly utilized method in this algorithm for determining the weighted average.

4.6. SGD regressor

Stochastic gradient descent is an optimization method in which the estimation of the gradient of the loss is performed each sample at a time and the model is accordingly updated along the way based on a schedule with decreasing strength. SGD regressor is a linear model that is fitted by attempting to minimize a regularized empirical loss using SGD [59]. The loss function can be written as:

$$E(w,b) = \frac{1}{n}\sum_{i=0}^{n} L(y_i, f(x_i)) + \alpha\, R(w) \tag{45}$$

where L is a loss function that measures model (mis)fit and R is a regularization term that penalizes model complexity; α is a non-negative hyperparameter that controls the regularization stength. The regularizer is a penalty added to the loss function that shrinks the model parameters towards the zero vector using either the squared euclidean norm L2 or the absolute norm L1 [66] or a combination of both (Elastic Net [59, 67]), which are represented in Eq. 46.

$$R(w) = \begin{cases} \frac{1}{2}\sum_{j=1}^{m} w_j^2 & L2\ norm \\ \sum_{j=1}^{m} |w_j| & L1\ norm \\ \frac{\alpha}{2}\sum_{j=1}^{m} w_j^2 + (1-\alpha)\sum_{j=1}^{m} |w_j| & Elastic\ Net \end{cases} \tag{46}$$

Where n is the number of data points and m is the number of input features.

4.7. Linear Support Vector Regressor

Support Vector Regressor (SVR) utilizes the Support Vector Machine (SVM) algorithm in order to predict a continuous variable. This algorithm's fundamental concept is to nonlinearly map the original data into a higher dimensional feature space [68]. Linear Support Vector Regression is similar to SVR with the difference that it has higher flexibility regarding the penalties and loss functions' choice and scales better to elevated numbers of samples [59].

4.8. Forward feature combination

By implementing the previously mentioned steps, the highest possible accuracy with the minimum number of features corresponding to the optimized pipeline is obtained. Further reduction in the number of features, although leads to a decrement in accuracy, can help to reduce the models' complexity and the computational cost. Therefore, there is a trade-off between the simplicity and accuracy of the model. Accordingly, for the case of two-phase flow, an additional procedure (called forward feature combination) is implemented in order to obtain optimized pipelines with lower number of features (even if resulting in a lower accuracy). The main benefit of providing the latter sub-optimal pipelines is that the user can select the most promising model based on the requirements and constraints of the case study including the available measurement devices and the allowable range of computational cost. In the forward feature combination procedure, the optimized pipeline (considering the training dataset) is first fitted with each of the individual selected features (as a single input) and the one which results in the lowest MARD is selected as the first feature. Then the model is trained using only two variables

(the first one is already chosen), where each of the remained features is provided as the second feature. Similarly, the one which leads to the lowest MARD is selected as the second feature. This procedure is continued until all of the selected features are added. Clearly, the result obtained in the last step is the same as the optimal pipeline provided with selected features. The obtained order of features is then provided to the algorithm in order to monitor the corresponding accuracy of the optimized pipeline on the test set.

5. Results and Discussions

5.1. Single-phase case

5.1.1. Accuracy of the standard physical models available in the literature

The empirical models, introduced in section 3.2.1, are first utilized in order to estimate the water flow's friction factor in the considered flow conditions and the obtained results are reported in the Table 3. As can be observed, the most accurate model (lowest MARD) is the one proposed by Haaland [48] (Eq. 24) resulting in an MARD of 11.91 %. Fig. 2 demonstrates the comparison between the experimentally obtained friction factor and the values estimated by the model proposed by Haaland [48].

Author/s and the corresponding reference	MRD [%]	MARD [%]
Blasius [43]	-19.36	24.29
Filonenko [44]	-18.82	24.61
Nikuradse [45]	-19.05	24.46
Fang et al. [46]	-20.61	25.52
Jensen and Vlakancic [47]	9.48	14.12
Haaland [48]	3.47	11.91

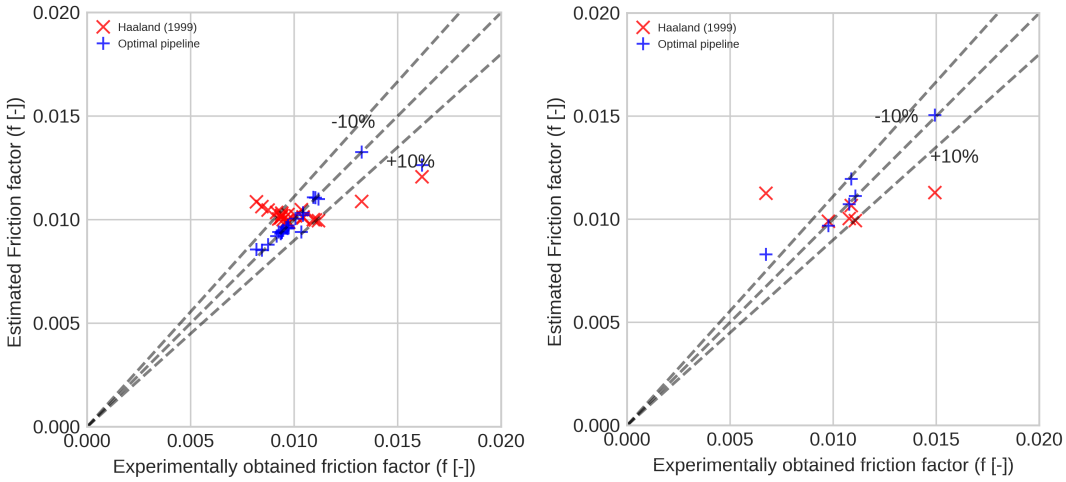
Table 3: Results of the standard physical models (the error between the predicted friction factor f [−] and the experimentally obtained values) for the single-phase case

5.1.2. Implemented machine learning based pipelines

For single-phase flow, the friction factor is taken into account as the target, while the non-dimensional parameters (Re, $\frac{1}{Re}$, $\frac{e}{D_{int}}$, and n) are provided as inputs. As was previously pointed out, the accuracy achieved using a random forest algorithm being provided with all features is first determined and the

feature selection based on the training set using this algorithm as the benchmark is implemented. Since, based on the obtained results, all of the features have been selected in the feature selection procedure, the latter two pipelines are identical and result in an MRD of 0.94 % and MARD of 3.59 % on the training set and an MRD of 6.47 % and an MARD of 18.20 % on the test set. Next, the pipeline optimization procedure is implemented in which the most promising algorithm and the corresponding optimal tuning parameters are optimized. K-nearest neighbors algorithm (with 2 neighbors and employing euclidean distance) is chosen as the optimal pipeline, while MinMaxScaler[1], Normalizer[2], and gradient boosting regressor as stacking estimator are utilized as three pre-processing steps. Expectedly, this optimal pipeline provide a higher accuracy as it results in an MRD of -1.26 % and an MARD of 2.15% on the training set and an MRD of 5.49 % and an MARD of 5.89 % on the test set.

Fig. 2 compares the friction factors estimated by the optimal pipeline with the experimental data and the predictions of the most accurate physical model (Haaland [48]). It can be observed that predictions of the optimal pipeline are notably more accurate than that of the mentioned most promising physical model.



(a) Training set (CV)

(b) Test set

Figure 2: Experimentally obtained friction factors compared to the values estimated by Haaland [48] and the estimations of the optimal pipeline

[1] A pre-processing step that scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.

[2] A pre-processing step that scales input vectors individually to unit norm

5.2. Two-phase case

5.2.1. Accuracy of the standard physical models available in the literature

Table 4 presents the results corresponding to the (pressure gradient) prediction accuracy of different models for the two-phase dataset. As can be observed in this table, the model proposed by Chisholm [9], with an MARD of 15.21 %, is determined to be the most accurate correlation. Employing the obtained Φ_l^2 and utilizing Eq. 30, the pressure gradient values are determined and are compared to the experimental data (Fig. 3).

Author/s and the corresponding reference	MRD [%]	MARD [%]
McAdams et al. [4]	-19.97	31.77
Bettiel & Whalley [5]	-36.70	41.15
Awad & Muzychka [6]	-6.01	25.90
Shannak [7]	43.22	43.50
Chisholm [9]	-3.44	15.21
Mishima & Hibiki [10]	2.08	18.10
Zhang et al. [49]	-2.56	19.37
Sun & Mishima [11]	-31.91	36.88
Chisholm [14]	135.77	135.77
Muller-Steinhagen & Heck [16]	97.97	97.97
Souza & Pimenta [18]	31.59	49.27
Friedel [17]	43.35	44.37
Cavallini et al. [19]	-83.90	83.90
Tran et al. [20]	-39.22	40.90
Haraguchi et al. [50]	48.99	50.72
Goto et al. [51]	-61.59	62.47

Table 4: Results of the standard physical models (the difference between the estimated pressure gradient ($\frac{dp}{dz}$ [$\frac{Pa}{m}$]) and the experimentally obtained values) for the two-phase case

5.2.2. Implemented machine learning based pipelines

For the two-phase flow case, two-phase multipliers (liquid, liquid only, gas, and gas-only) are taken into account as the targets, while dimensionless parameters are provided as inputs. Accordingly, four machine learning based pipelines were defined. For each target, the previously obtained optimal pipeline, defined in section 5.1.2, is utilized in order to calculate the single-phase friction factors. Similar to the single-phase case, for each target, all of the features are first provided to random forest algorithm. Next, for each pipeline, while utilizing random forest (RF) as the algorithm, feature selection procedure is implemented and the most promising set of features and the corresponding accuracy are obtained. Finally, while utilizing selected features, the optimal pipeline is determined for each case and the corresponding obtained accuracies for both the training set and the test set are compared. As demonstrated in Table 5, the most accurate pipeline is determined to be the optimal pipeline with the (Φ_g^2) as the target, which leads to an MARD of 6.72 % and 7.05 % on training set and test set, respectively.

Pipeline No.	Two-phase flow multiplier ($\Phi_i^2[-]$)	Pipeline	Training set (CV)		Test set	
			MRD %	MARD %	MRD %	MARD %
A	Φ_l^2	All features- Random Forest Regressor (RF)	4.08	12.07	7.62	14.84
B		Selected features– RF	4.68	12.04	8.67	15.47
C		Selected features– Optimal Pipeline	3.41	8.95	3.58	11.42
D	Φ_{lo}^2	All features- RF	3.28	10.85	6.55	13.23
E		Selected features– RF	3.33	10.68	6.82	13.46
F		Selected features– Optimal Pipeline	-0.24	6.78	-1.03	7.18
G	Φ_g^2	All features- RF	4.15	10.79	11.62	16.56
H		Selected features– RF	4.31	10.54	10.18	15.69
I		Selected features– Optimal Pipeline	1.26	6.72	-0.27	7.05
J	Φ_{go}^2	All features- RF	3.61	10.24	7.84	11.85
K		Selected features– RF	3.11	9.50	5.69	10.94
L		Selected features– Optimal Pipeline	1.83	8.44	3.32	10.88

Table 5: Employed machine learning based pipelines for the two-phase case and the corresponding obtained results

Table 6 lists all of the provided dimensionless features along with the ones which have been selected for the latter pipeline. It can be observed that only 5 features have been chosen as inputs, which notably reduces the complexity of the model. However, it should be noted that in the process of extracting the defined dimensionless target of the optimal pipeline (Φ_g^2), as expressed in Eq. 30, the single-phase

friction factors are first determined utilizing the corresponding features (Re, $\frac{1}{Re}$, $\frac{e}{D_{int}}$, and n). Thus, the reason behind the absence of geometrical parameters (relative height of fins ($\frac{e}{D_{int}}$) and number of fins (n)) among the selected features is that their influence is already accounted for in the obtained optimal pipeline for determination of single-phase friction factor (which is then utilized to extract Φ_g^2). Furthermore, it is also noteworthy that, although other features can have a correlation with the target, due to their physical correlation with the selected features, their influence is already represented by these (selected) features; thus, adding them does not improve the achieved accuracy, because of which these additional parameters are not chosen in the implemented feature selection procedure.

All features	Re_l	$\frac{1}{Re_l}$	Re_{lo}	$\frac{1}{Re_{lo}}$	Re_g	$\frac{1}{Re_g}$	Re_{go}	$\frac{1}{Re_{go}}$	X	Y
	f_l	f_{lo}	f_g	f_{go}	x	$1-x$	$\frac{1-x}{x}$	x_v	$1-x_v$	$\frac{1-x_v}{x_v}$
	$\frac{G_l^*}{G_g^*}$	n	$\frac{e}{D_{int}}$							
Selected features	xv	$\frac{1-x}{x}$	$\frac{1}{Re_g}$	Re_l	Re_{go}					

Table 6: All of the provided features and the selected ones for the two-phase case

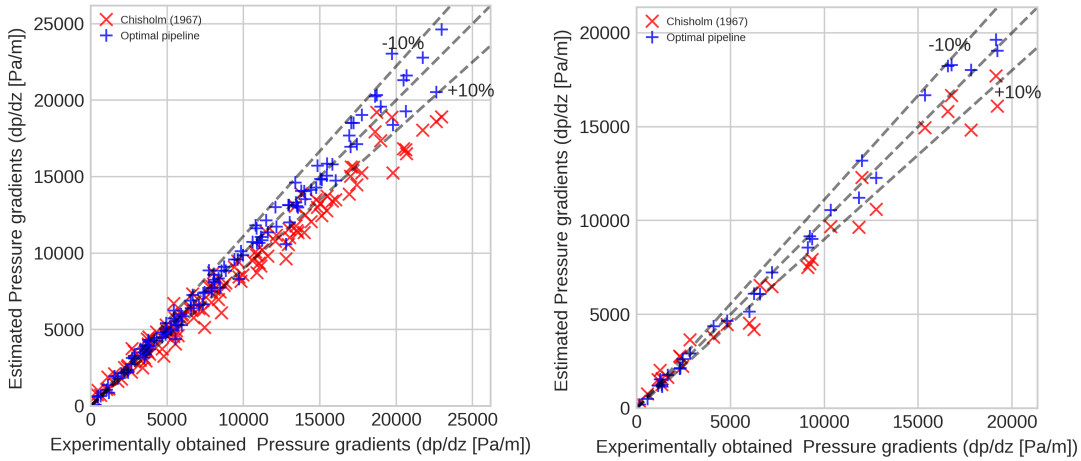
Table 7 represents the properties of the obtained optimal pipeline, which includes further feature processing along with the chosen estimators and the corresponding tuning parameters.

Optimal Pipeline Steps	Arguments	Definitions	Values
Step 1: StackingEstimator: estimator= GradientBoostingRegressor	alpha	The alpha-quantile of the huber loss function	0.75
	learning-rate	Tuning parameter for algorithm optimization	0.1
	loss	loss function to be optimized	ls
	max-depth	maximum depth of the individual regression estimators	4
	max-features	The number of features to consider when looking for the best split	0.8
	min-samples-leaf	The minimum number of samples required to be at a leaf node	3
	min-samples-split	The minimum number of samples required to split an internal node	3
	n-estimators	The number of boosting stages to perform	100
	subsample	The fraction of samples to be used for fitting the individual base learners	0.45
Step 2: StackingEstimator: estimator= SGDRegressor	alpha	Constant that multiplies the regularization term	0.0
	eta0	The exponent for inverse scaling learning rate	1
	fit-intercept	Intercept should be estimated	True
	l1-ratio	The Elastic Net mixing parameter	0.5
	learning-rate	Tuning parameter for algorithm optimization	invscaling
	loss	The loss function to be used	epsilon-insensitive
	penalty	The penalty to be used	elasticnet
	power-t	The exponent for inverse scaling learning rate	0.0
Step 3: StackingEstimator: estimator= RandomForestRegressor	bootstrap	Whether bootstrap samples are used when building trees	True
	max-features	The number of features to consider when looking for the best split	0.2
	min-samples-leaf	The minimum number of samples required to be at a leaf node	17
	min-samples-split	The minimum number of samples required to split an internal node	14
	n-estimators	The number of trees in the forest	100
Step 4: MinMaxScaler	-	-	-
Step 5: LinearSVR	C	Regularization parameter	20.0
	dual	Select the algorithm to either solve the dual or primal optimization problem	False
	epsilon	Epsilon parameter in the epsilon-insensitive loss function	0.001
	loss	The loss function to be used	squared-epsilon-insensitive
	tol	Tolerance for stopping criteria	1e-05

Table 7: Characteristics of determined optimal pipeline (pipeline I) for the two-phase case and the definition of each argument (provided by scikit-learn guidelines [59])

Fig. 3 compares the experimentally obtained pressure gradient values with the corresponding values estimated by the obtained optimal pipeline along with the estimations provided by the model proposed by Chisholm [9]. As can be observed, the optimal pipeline estimates the two-phase pressure drop with a notably higher accuracy compared to the most accurate empirical model available in the literature. Regarding the comparison with previously conducted studies focused on machine-learning based estimation of two-phase flow pressure drop, the studies conducted by Osgouei et al. [40], Qinghua et al. [30], and Chaari et al. [31] have only been identified to have similar objectives to those of the present work. The model proposed in the study carried out by Osgouei et al. [40], which was focused on

air-water flows' pressure drop estimation in horizontal annuli, was reported to result in an MARD of 7.462 % which is higher than the accuracy obtained in the present work (both for validation and tests sets), while the present study covers a wider range of pressure drop values. In the study performed by Qinghua et al. [30], that was dedicated to predicting the pressure drop of gas-liquid two-phase flow in two medium-diameter pipes, the utilized accuracy metrics differ from the ones employed in the present study (while the experimental values and the model predictions are neither available), because of which it is not possible to compare the results obtained in the present study and the results of the latter study. Finally, the investigation conducted by Chaari et al. [31], which was performed employing a large database containing experimental measurements spanning a wide range of flow conditions and pipe characteristics, has reported an MARD of 20.43 %. Considering the different range of operating conditions and characteristics of the dataset utilized in the latter study and that of the present work, in order to have a more reliable comparison, the authors have developed a neural network model with the same configuration as the one proposed in the latter study. Through applying the latter developed model on the dataset obtained in the present study, an MARD of 11.66 % on the validation set (training set evaluated through cross-validation) and 10.83 % on the test set, are obtained, which are both higher than the ones achieved using the optimal pipeline proposed in the present study. Moreover, since the obtained optimal pipeline is provided as a publicly accessible tool, the proposed approach can also offer a higher reproducibility and ease of use.



(a) Training set (CV)

(b) Test set

Figure 3: Experimentally obtained pressure gradient compared the estimations of the optimal pipeline (pipeline I) and the values estimated by Chisholm [9] model

Fig. 4 shows the accuracies achieved by the optimal pipeline (reported in Table 7), while implementing the forward feature combination procedure. The results demonstrate the contribution of each feature to the prediction accuracy of the model on the test set. It can be observed that, by utilizing only the first three parameters, an elevated accuracy can already be obtained. By adding more features, although the MARD is reduced, it results in an increment in the dimensionality of the model. Considering the latter trade-off, the results of the forward feature combination procedure provides the user with the possibility of choosing the most promising model considering the requirements and constraints of the project. As an instance, the latter results demonstrate that, although utilizing all the selected features leads to obtaining the lowest possible MARD, (considering the dataset's range of operating conditions) by employing only the first three features (xv, Re_l, $\frac{1}{Re_g}$) an elevated accuracy (MARD of less than 11%) can already be achieved.

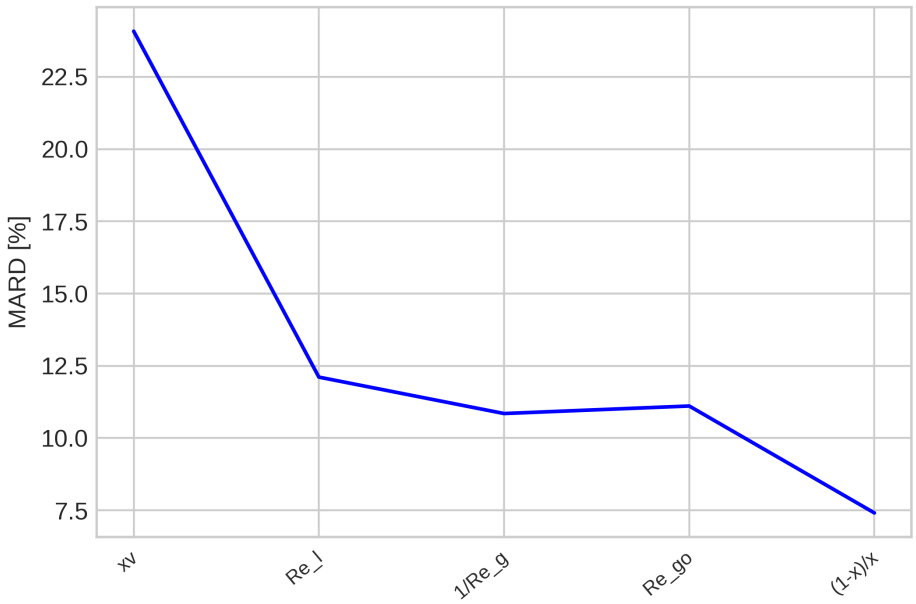
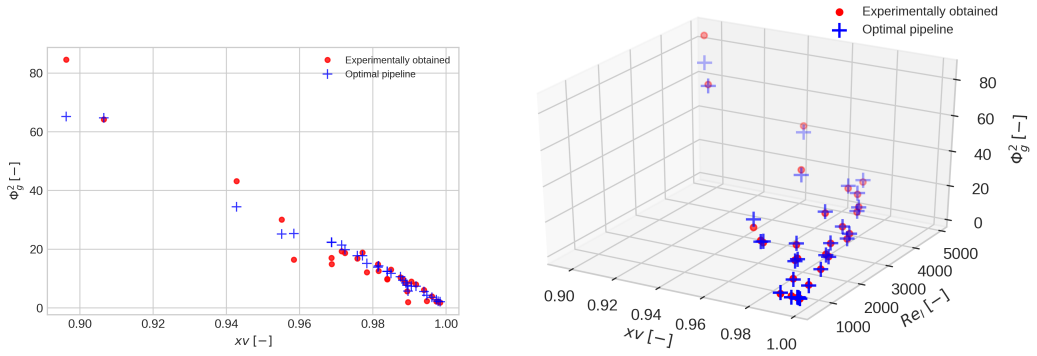


Figure 4: The accuracy achieved by the optimal pipeline on test set while being provided with lower number of features (results of forward feature combination)

Fig. 5a shows the experimental two-phase multiplier values of the test set along with the estimation of the optimal pipeline trained with only first feature (xv $[-]$) as input. Fig. 5b instead demonstrates the comparison between the test set's experimental data and the two-phase multiplier values estimated by the optimal pipeline that is trained with the all selected features, with respect to volume quality

xv [−] and Reynolds number of the liquid phase (Re_l [−]). It can be graphically observed that the accuracy of the optimal pipeline trained with all of the selected features is increased as it can estimate the real values better with respect to the estimations presented in Fig. 5a (using the model trained with only the first feature), specifically for those with higher values of Φ_g^2.



(a) Optimal Pipeline trained with first feature

(b) Optimal Pipeline trained with all selected feature

Figure 5: Experimentally obtained two-phase flow multiplier of test set compared to the estimations of the optimal pipeline, with respect to the selected features

6. Conclusion

In the present paper, the most promising set of dimensionless features (along with the contribution of each feature to the achieved accuracy) and the most suitable machine learning algorithm that can be employed for data-driven frictional pressure drop estimation of water (single-phase) and water-air mixture (two-phase) flow, in micro-finned horizontal tubes, were investigated. Accordingly, after implementing the corresponding machine learning based pipelines, in which a pool of dimensionless features was provided as inputs, a feature selection procedure was implemented in order to determine the set of features utilizing which results in the highest achieved accuracy. Next, an algorithm optimization procedure was utilized in order to determine the optimal machine learning algorithm and the corresponding tuning parameters. Furthermore, the state-of-the-art physical models, proposed in the literature, were implemented and the accuracy achieved utilizing the obtained optimal pipeline (for both single-phase and two-phase cases) was compared with the one obtained employing the determined most promising physical model.

For the single-phase case, it was demonstrated that by utilizing the obtained optimal pipeline (which employs all of the provided dimensionless features) an elevated accuracy (MARD of 2.15 [%] on the

training set and MARD of 5.89 % on the test set) can be obtained, which is notably lower than the one achieved using the most accurate available physical model (Haaland [48] with an MARD of 11.91 %). For the case of two-phase flow, the optimal pipeline was determined to be the one in which the gas two-phase multiplier is chosen as the target and 5 dimensionless parameters (xv , Re_l ,$\frac{1}{Re_g}$, Re_{go}, $\frac{1-x}{x}$ that are selected among 23 provided features) are utilized as input features. However, it should be noted that in the process of extracting the defined dimensionless target of the optimal pipeline (Φ_g^2), as expressed in Eq. 30, the single-phase friction factors are first determined utilizing the corresponding features (Re, $\frac{1}{Re}$, $\frac{e}{D_{int}}$, and n). Thus, the reason behind the absence of geometrical parameters (relative height of fins ($\frac{e}{D_{int}}$) and number of fins (n)) among the selected features is that their influence is already accounted for in the obtained optimal pipeline for determination of single-phase friction factor. Employing the latter pipeline an MARD of 6.72 % and 7.05 % on the training set and test set respectively, can be achieved that is notably more accurate than the most promising physical model (Chisholm [9] leading to an MARD of 15.21 %).

Finally, through implementing the forward feature combination strategy on the optimal pipeline, the contribution of each feature to the achieved accuracy was demonstrated and the trade-off between the complexity of the model (in terms of the number of features) and the obtained accuracy was presented. The obtained results demonstrated (considering the dataset's range of operating conditions) that employing only the first three features (xv, Re_l, $\frac{1}{Re_g}$) an MARD of less than 11% can be achieved. The latter results enable the user to choose the most promising model considering the requirements of the project and the corresponding constraints. Since the dataset and the obtained optimal pipelines along with the models trained with the reduced number of features will be made publicly accessible, the models that are proposed and implemented in the present work provide a higher reproducibility and ease of use compared to the state-of-the-art physical models, while providing a notably higher accuracy.

Appendix A. Online repository of the obtained optimal pipelines

The utilized processed dataset and the obtained optimal pipelines (including the most promising sets of features and the most suitable algorithm) are provided in an online repository (Link).

References

[1] M. Ishii and T. Hibiki, *Thermo-fluid dynamics of two-phase flow*. Springer Science & Business Media, 2010.

[2] L. Colombo, A. Lucchini, and A. Muzzio, "Flow patterns, heat transfer and pressure drop for evaporation and condensation of r134a in microfin tubes," *International Journal of Refrigeration*, vol. 35, no. 8, pp. 2150–2165, 2012.

[3] Y. Xu, X. Fang, X. Su, Z. Zhou, and W. Chen, "Evaluation of frictional pressure drop correlations for two-phase flow in pipes," *Nuclear engineering and design*, vol. 253, pp. 86–97, 2012.

[4] W. McAdams, W. Woods, and R. Bryan, "Vaporization inside horizontal tubes ii-benzene mixtures," in *ASME*, vol. 64, pp. 93–200, 1942.

[5] D. Beattie and P. Whalley, "A simple two-phase frictional pressure drop calculation method," *International journal of multiphase flow*, vol. 8, no. 1, pp. 83–87, 1982.

[6] M. Awad and Y. Muzychka, "Effective property models for homogeneous two-phase flows," *Experimental Thermal and Fluid Science*, vol. 33, no. 1, pp. 106–113, 2008.

[7] B. A. Shannak, "Frictional pressure drop of gas liquid two-phase flow in pipes," *Nuclear Engineering and Design*, vol. 238, no. 12, pp. 3277–3284, 2008.

[8] R. Lockhart and R. Martinelli, "Proposed correlation of data for isothermal two-phase, two-component flow in pipes," *Chem. Eng. Prog*, vol. 45, no. 1, pp. 39–48, 1949.

[9] D. Chisholm, "A theoretical basis for the lockhart-martinelli correlation for two-phase flow," *International Journal of Heat and Mass Transfer*, vol. 10, no. 12, pp. 1767–1778, 1967.

[10] K. Mishima and T. Hibiki, "Some characteristics of air-water two-phase flow in small diameter vertical tubes," *International journal of multiphase flow*, vol. 22, no. 4, pp. 703–712, 1996.

[11] L. Sun and K. Mishima, "Evaluation analysis of prediction methods for two-phase flow pressure drop in mini-channels," *International Journal of Multiphase Flow*, vol. 35, no. 1, pp. 47–54, 2009.

[12] M. Colombo, L. P. Colombo, A. Cammi, and M. E. Ricotti, "A scheme of correlation for frictional pressure drop in steam–water two-phase flow in helicoidal tubes," *Chemical Engineering Science*, vol. 123, pp. 460–473, 2015.

[13] J. De Amicis, A. Cammi, L. P. Colombo, M. Colombo, and M. E. Ricotti, "Experimental and numerical study of the laminar flow in helically coiled pipes," *Progress in Nuclear Energy*, vol. 76, pp. 206–215, 2014.

[14] D. Chisholm, "Pressure gradients due to friction during the flow of evaporating two-phase mixtures in smooth tubes and channels," *International Journal of Heat and Mass Transfer*, vol. 16, no. 2, pp. 347–358, 1973.

[15] C. Baroczy, "Systematic correlation for two-phase pressure drop.," in *Chem. Eng. Progr., Symp. Ser., 62: No. 64, 232-49 (1966).*, Atomics International, Canoga Park, Calif., 1966.

[16] H. Müller-Steinhagen and K. Heck, "A simple friction pressure drop correlation for two-phase flow in pipes," *Chemical Engineering and Processing: Process Intensification*, vol. 20, no. 6, pp. 297–308, 1986.

[17] L. Friedel, "Improved friction pressure drop correlation for horizontal and vertical two-phase pipe flow," *Proc. of European Two-Phase Flow Group Meet., Ispra, Italy, 1979*, 1979.

[18] A. Lobo de Souza and M. de Mattos Pimenta, "Prediction of pressure drop during horizontal two-phase flow of pure and mixed refrigerants," *ASME-PUBLICATIONS-FED*, vol. 210, pp. 161–172, 1995.

[19] A. Cavallini, G. Censi, D. Del Col, L. Doretti, G. A. Longo, and L. Rossetto, "Condensation of halogenated refrigerants inside smooth tubes," *Hvac&R Research*, vol. 8, no. 4, pp. 429–451, 2002.

[20] T. Tran, M.-C. Chyu, M. Wambsganss, and D. France, "Two-phase pressure drop of refrigerants during flow boiling in small channels: an experimental investigation and correlation development," *International Journal of Multiphase Flow*, vol. 26, no. 11, pp. 1739–1754, 2000.

[21] E.-S. A. Osman and M. A. Aggour, "Artificial neural network model for accurate prediction of pressure drop in horizontal and near-horizontal-multiphase flow," *Petroleum Science and Technology*, vol. 20, no. 1-2, pp. 1–15, 2002.

[22] A. Ebrahimi and E. Khamehchi, "A robust model for computing pressure drop in vertical multiphase flow," *Journal of Natural Gas Science and Engineering*, vol. 26, pp. 1306–1316, 2015.

[23] J. J. Garcia, F. Garcia, J. Bermudez, and L. Machado, "Prediction of pressure drop during evaporation of r407c in horizontal tubes using artificial neural networks," *International Journal of Refrigeration*, vol. 85, pp. 292–302, 2018.

[24] N. Bar, T. K. Bandyopadhyay, M. N. Biswas, and S. K. Das, "Prediction of pressure drop using artificial neural network for non-newtonian liquid flow through piping components," *Journal of Petroleum Science and Engineering*, vol. 71, no. 3-4, pp. 187–194, 2010.

[25] M. H. Ahmadi, M. Ghazvini, H. Maddah, M. Kahani, S. Pourfarhang, A. Pourfarhang, and S. Z. Heris, "Prediction of the pressure drop for cuo/(ethylene glycol-water) nanofluid flows in the car radiator by means of artificial neural networks analysis integrated with genetic algorithm," *Physica A: Statistical Mechanics and its Applications*, p. 124008, 2020.

[26] J. Barroso-Maldonado, J. Montañez-Barrera, J. Belman-Flores, and S. Aceves, "Ann-based correlation for frictional pressure drop of non-azeotropic mixtures during cryogenic forced boiling," *Applied Thermal Engineering*, vol. 149, pp. 492–501, 2019.

[27] A. López-Belchí, F. Illán-Gómez, J. M. Cano-Izquierdo, and J. R. García-Cascales, "Gmdh ann to optimise model development: Prediction of the pressure drop and the heat transfer coefficient during condensation within mini-channels," *Applied Thermal Engineering*, vol. 144, pp. 321–330, 2018.

[28] A. Khosravi, J. Pabon, R. Koury, and L. Machado, "Using machine learning algorithms to predict the pressure drop during evaporation of r407c," *Applied Thermal Engineering*, vol. 133, pp. 361–370, 2018.

[29] A. Zendehboudi and X. Li, "A robust predictive technique for the pressure drop during condensation in inclined smooth tubes," *International Communications in Heat and Mass Transfer*, vol. 86, pp. 166–173, 2017.

[30] W. Qinghua, Z. Honglan, L. Wei, Y. Junzheng, W. Xiaohong, and W. Yan, "Experimental study of horizontal gas-liquid two-phase flow in two medium-diameter pipes and prediction of pressure drop through bp neural networks," *International Journal of Fluid Machinery and Systems*, vol. 11, no. 3, pp. 255–264, 2018.

[31] M. Chaari, J. Ben Hmida, A. C. Seibi, A. Fekih, *et al.*, "An integrated genetic-algorithm/artificial-neural-network approach for steady-state modeling of two-phase pressure drop in pipes," *SPE Production & Operations*, 2020.

[32] M. A. Ahmadi and Z. Chen, "Machine learning models to predict bottom hole pressure in multi-phase flow in vertical oil production wells," *The Canadian Journal of Chemical Engineering*, vol. 97, no. 11, pp. 2928–2940, 2019.

[33] M. A. Ahmadi, M. Ebadi, and A. Yazdanpanah, "Robust intelligent tool for estimating dew point pressure in retrograded condensate gas reservoirs: application of particle swarm optimization," *Journal of Petroleum Science and Engineering*, vol. 123, pp. 7–19, 2014.

[34] M.-A. Ahmadi, M. Masumi, R. Kharrat, and A. H. Mohammadi, "Gas analysis by in situ combustion in heavy-oil recovery process: Experimental and modeling studies," *Chemical Engineering & Technology*, vol. 37, no. 3, pp. 409–418, 2014.

[35] M. Ali Ahmadi and A. Ahmadi, "Applying a sophisticated approach to predict co2 solubility in brines: application to co2 sequestration," *International Journal of Low-Carbon Technologies*, vol. 11, no. 3, pp. 325–332, 2016.

[36] M.-A. Ahmadi, M. R. Ahmadi, S. M. Hosseini, and M. Ebadi, "Connectionist model predicts the porosity and permeability of petroleum reservoirs by means of petro-physical logs: application of artificial intelligence," *Journal of Petroleum Science and Engineering*, vol. 123, pp. 183–200, 2014.

[37] M. A. Ahmadi and M. Ebadi, "Evolving smart approach for determination dew point pressure through condensate gas reservoirs," *Fuel*, vol. 117, pp. 1074–1084, 2014.

[38] M.-A. Ahmadi, A. Bahadori, and S. R. Shadizadeh, "A rigorous model to predict the amount of dissolved calcium carbonate concentration throughout oil field brines: side effect of pressure and temperature," *Fuel*, vol. 139, pp. 154–159, 2015.

[39] M. A. Ahmadi, M. Ebadi, P. S. Marghmaleki, and M. M. Fouladi, "Evolving predictive model to determine condensate-to-gas ratio in retrograded condensate gas reservoirs," *Fuel*, vol. 124, pp. 241–257, 2014.

[40] R. E. Osgouei, A. M. Ozbayoglu, E. M. Ozbayoglu, E. Yuksel, and A. Eresen, "Pressure drop estimation in horizontal annuli for liquid–gas 2 phase flow: Comparison of mechanistic models and computational intelligence techniques," *Computers & Fluids*, vol. 112, pp. 108–115, 2015.

[41] P. Vega-Penichet Domecq, "Pressure drop measurement for adiabatic single and two phase flows inside horizontal micro-fin tubes," 2018.

[42] S.-M. Kim and I. Mudawar, "Review of databases and predictive methods for pressure drop in adiabatic, condensing and boiling mini/micro-channel flows," *International Journal of Heat and Mass Transfer*, vol. 77, pp. 74–97, 2014.

[43] H. Blasius, "Das aehnlichkeitsgesetz bei reibungsvorgängen in flüssigkeiten," in *Mitteilungen über Forschungsarbeiten auf dem Gebiete des Ingenieurwesens*, pp. 1–41, Springer, 1913.

[44] G. Filonenko, "Hydraulic resistance in pipes," *Teploenergetika*, vol. 1, pp. 40–44, 1954.

[45] J. Nikuradse, *Strömungsgesetze in rauhen Rohren*. VDI-Verlag, 1933.

[46] X. Fang, Y. Xu, and Z. Zhou, "New correlations of single-phase friction factor for turbulent pipe flow and evaluation of existing single-phase friction factor correlations," *Nuclear Engineering and Design*, vol. 241, no. 3, pp. 897–902, 2011.

[47] M. K. Jensen and A. Vlakancic, "Technical note experimental investigation of turbulent heat transfer and fluid flow in internally finned tubes," *International Journal of Heat and Mass Transfer*, vol. 42, no. 7, pp. 1343–1351, 1999.

[48] S. E. Haaland, "Simple and explicit formulas for the friction factor in turbulent pipe flow," *Journal of Fluids Engineering*, vol. 105, no. 1, pp. 89–90, 1983.

[49] W. Zhang, T. Hibiki, and K. Mishima, "Correlations of two-phase frictional pressure drop and void fraction in mini-channel," *International Journal of Heat and Mass Transfer*, vol. 53, no. 1-3, pp. 453–465, 2010.

[50] H. Haraguchi, "Condensation heat transfer of refrigerant hfc134a, hcfc123 and hcfc22 in a horizontal smooth tube and a horizontal microfin tube," in *Proc. 30ˆ¡ th¿ National Heat Transfer Symp. of Japan, Yokohama*, pp. 343–345, 1993.

[51] M. Goto, N. Inoue, and N. Ishiwatari, "Condensation and evaporation heat transfer of r410a inside internally grooved horizontal tubes," *International Journal of Refrigeration*, vol. 24, no. 7, pp. 628–638, 2001.

[52] B. Najafi, S. Moaveninejad, and F. Rinaldi, "Data analytics for energy disaggregation: methods and applications," in *Big Data Application in Power Systems*, pp. 377–408, Elsevier, 2018.

[53] B. Najafi, P. Bonomi, A. Casalegno, F. Rinaldi, and A. Baricci, "Rapid fault diagnosis of pem fuel cells through optimal electrochemical impedance spectroscopy tests," *Energies*, vol. 13, p. 3643, Jul 2020.

[54] K. Pearson, "Vii. note on regression and inheritance in the case of two parents," *proceedings of the royal society of London*, vol. 58, no. 347-352, pp. 240–242, 1895.

[55] R. S. Olson, N. Bartley, R. J. Urbanowicz, and J. H. Moore, "Evaluation of a tree-based pipeline optimization tool for automating data science," in *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, pp. 485–492, ACM, 2016.

[56] B. Najafi, H. Najafi, and M. Idalik, "Computational fluid dynamics investigation and multi-objective optimization of an engine air-cooling system using genetic algorithm," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 225, no. 6, pp. 1389–1398, 2011.

[57] A. H. Mamaghani, B. Najafi, A. Casalegno, and F. Rinaldi, "Optimization of an ht-pem fuel cell based residential micro combined heat and power system: A multi-objective approach," *Journal of cleaner production*, vol. 180, pp. 126–138, 2018.

[58] T. Selleri, B. Najafi, F. Rinaldi, and G. Colombo, "Mathematical modeling and multi-objective optimization of a mini-channel heat exchanger via genetic algorithm," *Journal of Thermal Science and Engineering Applications*, vol. 5, no. 3, 2013.

[59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[60] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees. wadsworth int," *Group*, vol. 37, no. 15, pp. 237–251, 1984.

[61] B. Najafi, L. Di Narzo, F. Rinaldi, and R. Arghandeh, "Machine learning based disaggregation of air-conditioning loads using smart meter data," *IET Generation, Transmission & Distribution*, 2020.

[62] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[63] H. Drucker, "Improving regressors using boosting techniques," in *ICML*, vol. 97, pp. 107–115, 1997.

[64] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[65] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[66] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pp. 477–485, Association for Computational Linguistics, 2009.

[67] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[68] W.-C. Hong, "Rainfall forecasting by technological machine learning models," *Applied Mathematics and Computation*, vol. 200, no. 1, pp. 41–57, 2008.

View publication stats