



**Bruno Xavier Ferreira**

**Development of artificial intelligence models  
applied to the flow assurance problems in the oil  
and gas industry**

**Dissertação de Mestrado**

Dissertation presented to the Programa de Pós-graduação em Engenharia Química, de Materiais e Processos Ambientais of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia Química, de Materiais e Processos Ambientais.

Advisor: Prof. Brunno Ferreira dos Santos  
Co-Advisor: Prof. Vinicius Tadeu Kartnaller Montalvão

Rio de Janeiro  
September 2022



**Bruno Xavier Ferreira**

**Development of artificial intelligence models  
applied to the flow assurance problems in the oil  
and gas industry**

Dissertation presented to the Programa de Pós-graduação em Engenharia Química, de Materiais e Processos Ambientais of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia Química, de Materiais e Processos Ambientais. Approved by the Examination Committee.

**Prof. Brunno Ferreira dos Santos**

Advisor

Departamento de Engenharia Química e de Materiais – PUC-Rio

**Prof. Vinicius Tadeu Kartnaller Montalvão**

Co-Advisor

Instituto de Química – UFRJ

**Profa. Amanda Lemette Teixeira Brandão**

Departamento de Engenharia Química e de Materiais – PUC-Rio

**Prof. Tiago Dias Martins**

Instituto de Ciências Ambientais, Química e Farmacêuticas – UNIFESP

Rio de Janeiro, September the 16th, 2022

All rights reserved.

### **Bruno Xavier Ferreira**

He got a degree in Chemical Engineering from Pontifical Catholic University of Rio de Janeiro (PUC-Rio) in 2019. During the Master's degree, he studied the development of machine learning models to be applied in the oil and gas industry.

#### Bibliographic data

Ferreira, Bruno Xavier

Development of artificial intelligence models applied to the flow assurance problems in the oil and gas industry / Bruno Xavier Ferreira; advisor: Brunno Ferreira dos Santos ; co-advisor: Vinicius Tadeu Kartnaller Montalvão. – 2022.

172 f. : il. color. ; 30 cm

Dissertação (mestrado)—Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Química e de Materiais, 2022.

Inclui bibliografia

1. Engenharia Química e de Materiais – Teses. 2. Garantia de escoamento. 3. Incrustação inorgânica. 4. Medição de pH. 5. Redes neurais convolucionais. 6. Perceptrons multi-camadas. I. Santos, Brunno Ferreira dos. II. Montalvão, Vinicius Tadeu Kartnaller. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Química e de Materiais. IV. Título.

CDD: 620.11

To my family and friends, from this and other plans,  
for all the support.



## Acknowledgments

Firstly, I would like to thank my parents, Ana Luisa and Rildo (*in memoriam*), my brother Vitor and all the dogs I have had. All the love, support, and lessons you all shared with me during all my years of life made me the person I am now, and I can thank you enough for this. Also, I would like to thank my godmothers, Vera and Ceça, for their support and partnership and my aunt Ildete and uncle Zé Rosa for looking after me during my written shifts over the last six months.

To my advisors Brunno Ferreira dos Santos e Vinicius Kartnaller for the opportunities and guidance in these almost three years. But mainly, I would like to thank you both for your support, comprehension, and patience during these last six months. Without that, this document probably would not have been finished.

To my friends from PUC-Rio and before: Erick, Pedro Henrique, and Carlos. In addition to several others who helped me through the good and difficult phases, tests, and work we had over these years.

To Prof. João Cajaiba and the laboratory Núcleo de Desenvolvimento de Processos e Análises Químicas em Tempo Real (NQTR) for their collaboration for this work.

To the Prof. Amanda Lemetete Teixeira Brandão, Prof. Tiago Dias Martins, Prof. Roberto Bentes de Carvalho and Dr. Fabrício de Queiroz Venâncio for accepting to participate in my dissertation defense and for the understanding with the alterations of date.

For financial support, the Brazilian foundation National Council for Scientific and Technological Development (CNPq).

To PUC-Rio for providing excellent facilities to perform high-level research.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

## Abstract

Ferreira, Bruno Xavier; Santos, Brunno Ferreira dos (Advisor); Montalvão, Vinicius Tadeu Kartnaller (Co-Advisor). **Development of artificial intelligence models applied to the flow assurance problems in the oil and gas industry**. Rio de Janeiro, 2022. 172p. Dissertação de mestrado – Departamento de Engenharia Química e de Materiais, Pontifícia Universidade Católica do Rio de Janeiro.

A significant concern during oil and gas production is flow assurance to avoid loss of time and money. Due to production conditions changes (such as pressure and temperature), especially in the Brazilian pre-salt region, the solubility of the components of the crude oil (oil-gas-water) can decrease, resulting in the formation of deposits. The fouling is usually caused by wax, gas hydrate, and inorganic salt (scale). In this work, models were developed using Machine Learning strategies for scale formation monitoring and measuring process parameters associated with remediation of obstruction from other sources. First, models for the calcium carbonate scaling formation process in the presence of monoethylene glycol (typical gas hydrate inhibitor) were created using feedforward neural network architecture to predict the differential pressure ( $\Delta P$ ) one and five steps ahead, obtaining an  $R^2$  higher than 92.9% for the training and test group for both the prediction horizon. The second approach explored was the development of models for determining the pH in atmospheric and pressurized systems (up to 6.0 MPa) using image analysis. These models could be applied to control and monitor the Nitrogen Generation System, which can be used for different flow assurance problems, and its kinetics strongly depend on the system's pH value. This step initially created classification models for the system pH (2, 3, 4, 5, 6, 7, 8, 9, 10) using the Convolution Neural Networks (CNN), Support Vector Machine, and decision tree architectures. Also, CNN models were built to predict the pH in the range of 2-10.

## Keywords

Flow assurance; Scale; pH measurement; Convolution Neural Network; Multilayer perceptron.

## Resumo

Ferreira, Bruno Xavier; Santos, Brunno Ferreira dos (Advisor); Montalvão, Vinicius Tadeu Kartnaller (Co-Advisor). **Desenvolvimento de modelos utilizando inteligência artificial para problemas de garantia de escoamento na indústria de petróleo.** Rio de Janeiro, 2022. 172p. Dissertação de Mestrado – Departamento de Engenharia Química e de Materiais, Pontifícia Universidade Católica do Rio de Janeiro.

Uma preocupação significativa durante a produção de óleo e gás é a garantia de escoamento para evitar desperdício de tempo e dinheiro. Devido às mudanças nas condições durante a produção (como pressão e temperatura), principalmente na região do pré-sal brasileiro, a solubilidade dos componentes do petróleo bruto (óleo-gás-água) pode diminuir, resultando na formação de depósitos. A incrustação é geralmente causada por parafina, hidratos e sal inorgânico. Neste trabalho, foram desenvolvidos modelos utilizando estratégias de Aprendizado de Máquina para monitoramento da formação de incrustações inorgânicas e medição de parâmetros de processo associados com formas de remediação de obstruções de outras fontes. Primeiramente, foram criados modelos do processo de formação de incrustação de carbonato de cálcio na presença de monoetilenoglicol (inibidor de hidrato) usando a arquitetura de redes neurais *feedforward* prever o pressão diferencial um e cinco instantes à frente, obtendo um  $R^2$  superior a 92,9% para ambos os horizontes de predição. O segundo tópico explorado foi desenvolver modelos para determinação do pH em sistemas pressurizados (até 6,0 MPa) por meio de análise de imagens. Podendo ser aplicados no monitoramento de sistemas como Sistema Gerador de Nitrogênio, utilizado para remediar alguns problemas de incrustação, dado que sua cinética depende fortemente do pH do sistema. Foram criados modelos de classificação para o pH do sistema (2, 3, 4, 5, 6, 7, 8, 9, 10) usando Redes Neurais Convolucionais (CNN), Máquina de Vetor de Suporte e Árvores de Decisão. Além disso, modelos CNN foram construídos para prever o pH na faixa de 2-10.

## Palavras-chave

Garantia de escoamento; Incrustação inorgânica; Medição de pH; Redes Neurais Convolucionais; Perceptrons Multi-Camadas.

## Table of contents

1 Introduction	23
1.1 Objectives	24
1.2 Organization	25
1.3 References	27
2 Literature review	29
2.1 Oil and Gas Production	29
2.1.1 Flow assurance	30
2.2 Modeling in the oil and gas industry	35
2.3 pH meter techniques	36
2.3.1 pH meter using image processing	38
2.4 Soft sensors	39
2.4.1 Multi-Layer Perceptron (MLP)	41
2.4.2 Convolutional Neural Network (CNN)	43
2.4.3 Support Vector Machine (SVM)	45
2.4.4 Decision Tree (DT)	49
2.5 References	51
3 Development of Artificial Neural Networks models for the simulation of CaCO <sub>3</sub> scale formation process in the presence of monoethylene glycol (MEG) in Dynamic Tube Blocking Test equipment	59
3.1 Article	60
3.1.1 Introduction	61

3.1.2 Methodology	65
3.1.2.1 Experimental details	65
3.1.2.2 ANN database preparation	66
3.1.2.3 Artificial Neural Network optimization	68
3.1.2.4 Statistical performance evaluation	68
3.1.2.5 Sensitivity analysis	69
3.1.3 Results and Discussion	70
3.1.3.1 Evaluation of the ANN models	70
3.1.3.2 Validation of the MLP models	71
3.1.3.3 Sensitivity analysis	76
3.1.4 Conclusions	78
3.5 References	80
4 Development of machine learning models to measure the pH values using image analysis	86
4.1 Article Manuscript	87
4.1.1 Introduction	88
4.1.1.1 Modeling Strategies	89
4.1.1.1.1 Convolution Neural Networks (CNNs)	89
4.1.1.1.2 Support Vector Machines (SVMs)	91
4.1.1.3 Decision Trees (DTs)	92
4.1.2. Methodology	93
4.1.2.1 Case study: Pressurized reactor	93
4.1.2.2 Database preparation	95
4.1.2.3 Modeling strategies	96
4.1.2.3.1 CNN	97

4.1.2.3.2 SVM	98
4.1.2.3.3 Decision Tree (DT)	98
4.1.2.4 Statistical Performance Evaluation	99
4.1.3 Results and Discussion	101
4.1.3.1 Evaluation of the classification models	101
4.1.3.1.1 CNN classification models	101
4.1.3.1.2 SVM models	106
4.1.3.1.3 DT models	108
4.1.3.2 Evaluation of the prediction models	110
4.1.3.2.1 CNN prediction models	110
4.1.3.3 Validation of the models	113
4.1.3.3.1 Case study: Titration curve of a strong acid with a strong base	113
4.1.3.3.2 Case study: CO <sub>2</sub> -H <sub>2</sub> O equilibrium systems	115
4.1.4. Conclusions	117
4.5 References	120
5 Conclusions	124
6 Suggestions for future works	125
A Appendix - Supporting information of the article: Development of MLP artificial neural network models for the simulation of CaCO <sub>3</sub> scale formation process in the presence of monoethylene glycol (MEG) in a dynamic tube blocking test (TBT) equipment	126
B Appendix of the article: Development of MLP artificial neural	139

network models for the simulation of CaCO<sub>3</sub> scale formation process  
in the presence of monoethylene glycol (MEG) in a dynamic tube  
blocking test (TBT) equipment

C Supporting information of the article: Development of MLP  
artificial neural network models for the simulation of CaCO<sub>3</sub> scale  
formation process in the presence of monoethylene glycol (MEG) in a  
dynamic tube blocking test (TBT) equipment 142

D Supporting information of the article: Machine learning  
models for measurement of pH using a low-cost image analysis  
strategy 154

E Graphical Abstracts from the articles 171

F Databases and codes 172

## List of figures

Figure 2.1. Subdivisions of Oil and Gas Industrial operations (adapted from AALSALEM <i>et al.</i> , 2018).	30
Figure 2.2: Real cases of incrustations formation (a) inorganic (scale), (b) organic, (c) gas hydrate (Kartnaller, 2018; Hw Institute of Petroleum Engineering; Doelman, 2013; Irmann- Jacobsen e Hægland, 2014).	31
Figure 2.3: Scheme of a Dynamic Scale Loop (DSL) system used in a TBT experiment (adapted from KARTNALLER <i>et al.</i> , 2018).	33
Figure 2.4: Reaction rate constant as a function of pH (adapted from NGUYEN <i>et al.</i> , 2001).	34
Figure 2.5: Scenarios with good application potential in the context of the "Oil and Gas 4.0" era (adapted from LU <i>et al.</i> , 2020).	35
Figure 2.6: Scheme of physical and virtual space using a digital twin framework with five components (physical space, virtual space, connection between them, data, and service) (adapted from WANASINGHE <i>et al.</i> , 2020).	36
Figure 2.7: Experimental setup scheme (adapted from DE OLIVEIRA <i>et al.</i> , 2019)	38
Figure 2.8: Flow of soft sensor analysis and problems involved at each stage (adapted from FUNATSU, 2013).	39
Figure 2.9: Statistics on exiting relevant work applications in different fields (adapted from SUN and GE, 2021)	40
Figure 2.10: Multi-Layer Neural Network scheme (adapted from CHOJACZYK <i>et al.</i> , 2015).	41
Figure 2.11: Activation functions: (a) logsigmoid, (b) hyperbolic tangent , (c) linear (adapted from SOLEIMANI <i>et al.</i> , 2013).	42



Figure 2.12: Typical CNN structure (adapted from ZAN <i>et al.</i> , 2020).	44
Figure 2.13: Convolutional operation with a 3 x 3 convolutional kernel (adapted from YUAN <i>et al.</i> , 2020).	44
Figure 2.14: Max pooling operation with 2 x 2 size (adapted from YUAN <i>et al.</i> , 2020).	45
Figure 2.15: Linearly separable data with two dimensions and two classes (solid line – hyperplane separating the classes; dashed lines – margins of the hyperplane) (adapted from NOGUEIRA, 2021).	46
Figure 2.16: Representation of the input transformation from the Input space (right) to the Feature space (left) by the use of the kernels (adapted from CHAUHAN <i>et al.</i> , 2019).	48
Figure 2.17: Kernel functions behavior in classification with SVM (adapted from NOGUEIRA, 2021).	47
Figure 2.18: Multi-class SVM approaches (a) OvR (One-versus-Rest) and (b) OvO (One-versus-One) (dashed lines – hyperplanes) (adapted from NOGUEIRA, 2021).	49
Figure 2.19: Example of a general decision tree for classification (adapted from BARROS, 2014).	50
Figure 3.1: Scheme of a Dynamic Scale Loop (DSL) system used in a TBT experiment (adapted from KARTNALLER <i>et al.</i> , 2018).	66
Figure 3.2 Flowchart of the methodology.	69
Figure 3.3. Representation of the behavior of the experimental data of the six experiments of the second database and the respective predicted data for the output $\Delta P_{(t+1)}$ by the MLP model <code>logsig_7_purelin_1_trainbr</code> .	73
Figure 3.4. Representation of the behavior of the experimental data of the six experiments of the second database and the respective predicted data for the output $\Delta P_{(t+5)}$ by the MLP model <code>logsig_6_purelin_1_trainlm</code> .	75

Figure 3.5. Relevancy factor of both output variables $\Delta P_{(t+1)}$ (A) and $\Delta P_{(t+5)}$ (B).	76
Figure 3.6. Relative Importance (RI) of both output variables $\Delta P_{(t+1)}$ (A) and $\Delta P_{(t+5)}$ (B) calculated by the Garson method (GARSON, 1991).	77
Figure 4.1: CNN schematic representation.	90
Figure 4.2: SVM schematic representation, with a three classes problem using linear kernels (dashed lines)	91
Figure 4.3: DT schematic representation.	92
Figure 4.4: Experimental setup scheme (adapted from DE OLIVEIRA <i>et al.</i> , 2019)	94
Figure 4.5: Examples of the images presented in the dataset for each pH category.	96
Figure 4.6: Scheme of a confusion matrix (2x2) in a binary case	100
Figure 4.7: Flowchart of the methodology	101
Figure 4.8: Confusion matrix (CM) of the CNN_class_RGB_crop_model_4 for the training (A) and test (B) datasets.	105
Figure 4.9: Accuracy values for the CNN classification models in the neutralization	113
Figure 4.10: Accuracy values in the neutralization curve scenario for the classification models: SVM (A) and DT (B)	114
Figure 4.11: RMSE (A) and $R^2$ (B) values for the CNN predict models in the neutralization curve scenario.	115
Figure 4.12: Accuracy values for the CNN classification models in the equilibrium CO <sub>2</sub> -H <sub>2</sub> O system scenario	116
Figure 4.13: Accuracy values in the equilibrium CO <sub>2</sub> -H <sub>2</sub> O system scenario for the classification models: SVM (A) and DT (B).	116
Figure 4.14: RMSE (A) and $R^2$ (B) values for the CNN predict models in the equilibrium CO <sub>2</sub> -H <sub>2</sub> O system scenario	117

Figure C1. Regression plot between experimental versus the predicted values for the variables $\Delta P_{(t+1)}$ (A) and $\Delta P_{(t+5)}$ (B).	143
Figure C2. Comparison between normalized residuals of the prediction of the $\Delta P_{(t+1)}$ and $\Delta P_{(t+5)}$ variables for the training dataset (A) and test dataset (B)	144
Figure E.1: Graphical Abstract of the article: Development of artificial neural network models for the simulation of $\text{CaCO}_3$ scale formation process in the presence of monoethylene glycol (MEG) in a dynamic tube blocking test (TBT) equipment	171
Figure E.2: Graphical Abstract of the article: Machine learning models for measurement of pH using a low-cost image analysis strategy	171

## List of tables

Table 2.1. pH meter electrodes for pressurized system (Hanna Instruments, 2021; Ato, 2021; Winn-Marion Companies, 2021)	37
Table 3.1. Range of the experimental variables	67
Table 4.1. Number of imagens for each pH category in the training database	95
Table 4.2. Types of input tested in the different models developed.	96
Table 4.3: Hyperparameters tested in the CNN models	97
Table 4.4. Hyperparameters tested in SVM	98
Table 4.5: Hyperparameters tested in DT	98
Table 4.6: CNN classification models topologies for the best models – Part I	103
Table 4.7: CNN classification models topologies for the best models – Part II	104
Table 4.8: SVM topologies for the best models	107
Table 4.9: DT topologies for the best models	109
Table 4.10: CNN prediction models topologies for the best models	111
Table B.1: MLP topology models for the variables $\Delta P_{(t+1)}$ and $\Delta P_{(t+5)}$ .	140
Table C1: Performance values for each MLP topology for the experiment with 10 v/v% MEG.	145
Table C2: Performance values for each MLP topology for the experiment with 20 v/v% MEG.	146
Table C3: Performance values for each MLP topology for the experiment with 30 v/v% MEG.	147

Table C4: Performance values for each MLP topology for the experiment with 50 v/v% MEG.	149
Table C5: Performance values for each MLP topology for the experiment with 60 v/v% MEG.	150
Table C6: Performance values for each MLP topology for the experiment with 70 v/v% MEG.	151
Table C7: Optimized parameters (weight and bias) of the MLP <code>logsig_7_purelin_1_trainbr</code> used to predict the $\Delta P_{(k+1)}$ .	153
Table C8: Optimized parameters (weight and bias) of the MLP <code>logsig_6_purelin_1_trainlm</code> used to predict the $\Delta P_{(k+5)}$ .	153
Table D1: Performance values for all SVM models (PR = Precision, RC = Recall, ACC = Accuracy)	155
Table D2: Performance values for all DT models (PR = Precision, RC = Recall, ACC = Accuracy)	166

## List of Abbreviations

ACC	Accuracy
ANFIS	Adaptive Neuro-Fuzzy Inference System
AI	Artificial Intelligence
ANN	Artificial Neural Network
BP	backpropagation
BD	Big Data
CAP	coccolith-associated polysaccharide
$R^2$	coefficient of determination
HSV	color system Hue, Saturation, and Value
RGB	color system Red, Green, and Blue
CMIS	Committee Machine Intelligent System
CARG	Compound Annual Growth Rate
CM	Confusion Matrix
CNN	Convolutional Neural Network
DT	Decision Tree
DL	Deep Learning
DSL	Dynamic Scale Loop
ENN	Elman Neural Network
EOR	Enhanced Oil Recovery
FN	False negatives
FP	False positives
FFNN	Feedforward Neural Networks
FET	field effect transistor
GA	Genetic Algorithm
GPU	Graphic Processing Unit
HPLC	High Performance Liquid Chromatography
<i>tansig</i>	hyperbolic tangent (activation function) in MATLAB

<i>tanh</i>	hyperbolic tangents activation function in Python
LS-SVM	Least Square - Support Vector Machine
<i>purelin</i>	linear activation function in MATLAB
<i>logsig</i>	logsigmoid activation function in MATLAB
ML	Machine Learning
MSE	Mean Squared Error
MIC	Minimum Inhibitor Concentration
MEG	monoethylene glycol
MLP	Multi-Layer Perceptron
MLR	Multivariate Linear Regression
NGS	Nitrogen Generated System
OVA	One-versus-All
OvO	One-versus-One
OvR	One-versus-Rest
PR	Precision
PCA	Principal Component Analysis
RBF	Radial-Basis Function
RC	Recall
ReLU	Rectified Linear Unit activation function
RI	Relative Importance
<i>r</i>	relevancy factor
RMSE	Root Mean Squared Error
sigmoid	sigmoid function activation function in Python
SSE	Sum of Squared Errors
SVM	Support Vector Machine
THI	Thermodynamic Hydrate Inhibitors
TDNN	Time-delay Neural Network
TSS	Total Sum of Squares
<i>trainbr</i>	training algorithm Bayesian Regularization Backpropagation in MATLAB
<i>traingdx</i>	training algorithm Gradient Descent with Momentum and Adaptive Learning Rate Backpropagation in MATLAB
<i>trainlm</i>	training algorithm Levenberg-Marquardt Backpropagation in

	MATLAB
TEG	triethylene glycol
TN	True negatives
TP	True positives
TBT	Tube Blocking Test
WAT	wax appearance temperature



## List of Symbols

$f_{(aj)}$	Activation function
$\text{CaCO}_3$	Calcium carbonate
$\text{CO}_2$	carbon dioxide
$C_{\text{MEG}}$	Concentration of MEG
$C_{\text{HCO}_3^-}$	Concentration of the ions calcium
$C_{\text{Ca}^{2+}}$	Concentration of the ions carbonate
$\Delta P$	differential of pressure
$-\text{OH}$	hydroxyl group
$\varphi$	Kernel mathematical functions for the SVM models
$b_{ji}$	neuron bias
$x_i$	neuron input
$w_{ij}$	neuron weight
$\Delta P_{(t+5)}$	predict the differential of pressure five-step ahead
$\Delta P_{(t+1)}$	predict the differential of pressure one-step ahead
$k$	rate constant
$\Delta H_{\text{Rx}}$	Reaction enthalpy

*Forever – is composed of Nows*

**Emily Dickinson**

## Introduction

Petroleum reservoirs have complex compositions due to their formation process. Usually, they present three phases of the mixture formed by gas-oil-water, which are present inside the pores of the reservoir rock. The gas phase generally comprises small chain hydrocarbons and other gases, such as hydrogen sulfide (H<sub>2</sub>S) and carbon dioxide (CO<sub>2</sub>). The oil phase is formed by a diverse mixture of heavier hydrocarbon molecules, such as paraffin, aromatics, resins, and asphaltenes. The third phase contains water with different types of ions dissociated; that aqueous solution originated during the reservoir formation is called “formation water”. The three phases are mixed inside the reservoirs in equilibrium under high pressure and temperature conditions (KELLAND, 2014; NASIRI and JAFARI, 2017; ALADE *et al.*, 2020).

Associated with this diverse composition of the fluid mixture in the oil well, the operational conditions during the exploration, such as the change in the pressure and temperature during the transportation process in the pipelines, can provoke the precipitation, deposition, and agglomeration of solids in the pipelines and equipment (also called fouling). These flow assurance problems can be the result of different kinds of obstructions, the main ones being associated with the formation of gas hydrates, the precipitation of inorganic salts (scale), and the solidification of wax (MAGNINI *et al.*, 2019; MELCHUNA *et al.*, 2020; AMAR *et al.*, 2021).

The flow assurance problems result in great financial losses and safety problems. The fouling process is a complex subject that simultaneously involves kinetics, thermodynamics, and transport phenomena for understanding (FRENIER and ZIAUDDIN, 2008; ZHENG *et al.*, 2017; MELCHUNA *et al.*, 2020). These reasons led to several studies for understanding and monitoring the fouling formation (LEOPORINI *et al.*, 2019; ZAREI and BAGHBAN, 2017; LIM *et al.*, 2020), avoiding their appearance (SOUZA *et al.*, 2019), and

for treatments to unplug the pipelines once they are formed (RAMZI *et al.*, 2016).

Meanwhile, different kinds of Artificial Intelligence (AI) techniques are being used to model problems associated with the petroleum industry (RAHMANIFARD and PLAKSINA, 2019). Due to the vast amount of data generated for this industry and the complexity of some of the systems to be modeled and monitored, the strategy to create the AI models known as data-driven is very commonly applied. In this method, the models (also called black-box models) are created using only the process data. (MOHAMMADPPOR and TORABI, 2020; SHANG *et al.*, 2014).

Artificial Neural Networks (ANNs), inspired by the human brain neural arrangement, are a group of AI strategies commonly used to develop models. One of this set's most usually applied structures is the Multilayer Perceptron (MLP) topology. It is often composed of three layers: input layer, hidden layer, and output layer (KUMAR *et al.*, 2013; CHOJACZYK *et al.*, 2015; LI *et al.*, 2017). The Convolutional Neural Network (CNN) is another type of ANN but is frequently used for image classification since its usual topology has more than two hidden layers, making this also a Deep Learning (DL) technique (MADHAN *et al.*, 2021).

Other AI techniques have been used to develop data-driven models. For this work, it is interesting to highlight the Support Vector Machine (SVM) and the Decision Tree (DT). SVM is based on statistical learning theory and geometric distance interval maximization to give a solid generalization capability. The technique evolve to be used for multiclass classification problems (PENG *et al.*, 2018; CHAUHAN *et al.*, 2019). DT is another technique usually applied for classification problems. It is formed by combining a series of hierarchically organized binary tests (GEURTS *et al.*, 2009).

In this scenario, this work proposes to apply AI techniques to develop models that could be used to resolve different assurance problems.

## 1.1

### Objectives

This work aims to apply AI strategies to develop models that can be applied in future applications as soft sensors relate to flow assurance problems during the oil

production process. For that, two secondary goals were chosen.

First, to create models to monitor the fouling of inorganic salts (scaling) in the presence of a gas hydrate inhibitor by predicting the differential pressure ( $\Delta P$ ) reached inside a pipeline during the obstruction process. The next secondary aim was to create models to determine the pH in atmospheric and pressurized systems using image analysis, exploring the different types of AI, and developing classification and prediction models.

To reach the secondary objectives, the following specific targets were settled:

- Organize and pre-treat the databases;
- Create the MLP models to predict the  $\Delta P$  in different prediction horizons;
- Develop the classificatory models for the pH class (2, 3, 4, 5, 6, 7, 8, 9, 10);
- Develop the prediction models for the pH values in the range 2-10;
- Evaluate and optimize the models' hyperparameters;
- Analyze the models' performance parameters and choose the best ones;

## 1.2

### Organization of the Dissertation

The organization of this work was chosen to provide a better experience for the readers. This first chapter is composed of a brief introduction with the general motivation, the general and specific objectives, and a description of this study.

Chapter 2 presents a literature review of the flow assurance problems that arise during oil and gas production and the application of different kinds of AI models to monitor, detect and solve these problems with several approaches presented.

Chapter 3 presents the information about the first main objective, the MLP models to predict the scale formation process in the presence of a gas hydrate inhibitor (monoethylene glycol). The section is composed of a short introduction, followed by the manuscript of the article published in the scientific journal Energy

& Fuels.

Chapter 4 carries the content of the second main goal, creating the classification and prediction models for the pH measurement using image analysis. It follows a similar structure to the previous section but presents the manuscript of the article to be submitted to the scientific journal such as Computer and Chemical Engineering, Sensors and Chemical Engineering Research and Design.

Chapter 5 contains the general conclusions of this study, and in Chapter 6 the suggestions for future works are presented. The related references are placed at the end of each section.

The supporting information associated with the articles is presented in the Appendix section.

## 1.3

## References

ALADE, O. S., HASSAN, A., MAHMOUD, M., *et al.* "Novel Approach for Improving the Flow of Waxy Crude Oil Using Thermochemical Fluids: Experimental and Simulation Study", **ACS Omega**, v. 5, n. 8, p. 4313–4321, 2020. DOI: 10.1021/acsomega.9b04268. .

CHAUHAN, V. K., DAHIYA, K., SHARMA, A. "Problem formulations and solvers in linear SVM: a review", **Artificial Intelligence Review**, v. 52, n. 2, p. 803–855, 2019. DOI: 10.1007/s10462-018-9614-6.

CHOJACZYK, A. A., TEIXEIRA, A. P., NEVES, L. C., *et al.* "Review and application of Artificial Neural Networks models in reliability analysis of steel structures", **Structural Safety**, v. 52, n. PA, p. 78–89, 2015. DOI: 10.1016/j.strusafe.2014.09.002.

FRENIER, W. W., ZIAUDDIN, M. **Formation, removal, and inhibition of inorganic scale in the oilfield environment**. [S.l.], Society of Petroleum Engineers, 2008.

GEURTS, P., IRRTHUM, A., WEHENKEL, L. "Supervised learning with decision tree-based methods in computational and systems biology", **Molecular BioSystems**, v. 5, n. 12, p. 1593–1605, 2009. DOI: 10.1039/b907946g.

KELLAND, M. A. **Production Chemicals for the Oil and Gas Industry**. [S.l.], CRC Pres, 2014.

KUMAR, S., NAIYA, T. K., KUMAR, T. "Developments in oilfield scale handling towards green technology-A review", **Journal of Petroleum Science and Engineering**, v. 169, n. May, p. 428–444, 2018. DOI: 10.1016/j.petrol.2018.05.068.

LEPORINI, M., TERENCE, A., MARCHETTI, B., *et al.* "Experiences in numerical simulation of wax deposition in oil and multiphase pipelines: Theory versus reality", **Journal of Petroleum Science and Engineering**, v. 174, n. February 2018, p. 997–1008, 2019. DOI: 10.1016/j.petrol.2018.11.087.

LI, H., ZHANG, Z., LIU, Z. "Application of artificial neural networks for catalysis: A review", **Catalysts**, v. 7, n. 10, 2017. DOI: 10.3390/catal7100306.

LIM, V. W. S., METAXAS, P. J., STANWIX, P. L., *et al.* "Gas hydrate formation probability and growth rate as a function of kinetic hydrate inhibitor (KHI) concentration", **Chemical Engineering Journal**, v. 388, n. November 2019, p. 124177, 2020. DOI: 10.1016/j.cej.2020.124177.

MADHAN, E. S., KANNAN, K. S., RANI, P. S., *et al.* "A distributed submerged object detection and classification enhancement with deep learning", **Distributed and Parallel Databases**, n. 0123456789, 2021. DOI: 10.1007/s10619-021-07342-1.

MAGNINI, M., MATAR, O. K. "Fundamental Study of Wax Deposition in Crude

Oil Flows in a Pipeline via Interface-Resolved Numerical Simulations", **Industrial and Engineering Chemistry Research**, v. 58, n. 47, p. 21797–21816, 2019. DOI: 10.1021/acs.iecr.9b05250.

MELCHUNA, A., ZHANG, X., SA, J. H., *et al.* "Flow Risk Index: A New Metric for Solid Precipitation Assessment in Flow Assurance Management Applied to Gas Hydrate Transportability", **Energy and Fuels**, v. 34, n. 8, p. 9371–9378, 2020. DOI: 10.1021/acs.energyfuels.0c01203.

MOHAMMADPOOR, M., TORABI, F. "Big Data analytics in oil and gas industry: An emerging trend", **Petroleum**, v. 6, n. 4, p. 321–328, 2020. DOI: 10.1016/j.petlm.2018.11.001.

NASIRI, M., JAFARI, I. "Produced water from oil-gas plants: A short review on challenges and opportunities", **Periodica Polytechnica Chemical Engineering**, v. 61, n. 2, p. 73–81, 2017. DOI: 10.3311/PPch.8786.

RAHMANIFARD, H., PLAKSINA, T. "Application of artificial intelligence techniques in the petroleum industry: a review", **Artificial Intelligence Review**, v. 52, n. 4, p. 2295–2318, 2019. DOI: 10.1007/s10462-018-9612-8.

RAMZI, M., HOSNY, R., EL-SAYED, M., *et al.* "Evaluation of scale inhibitors performance under simulated flowing field conditions using dynamic tube blocking test", **International Journal of Chemical Sciences**, v. 14, n. 1, p. 16–28, 2016.

SHANG, C., YANG, F., HUANG, D., *et al.* "Data-driven soft sensor development based on deep learning technique", **Journal of Process Control**, v. 24, n. 3, p. 223–233, 2014. DOI: 10.1016/j.jprocont.2014.01.012.

SOUSA, A. L., MATOS, H. A., GUERREIRO, L. P. "Preventing and removing wax deposition inside vertical wells: a review", **Journal of Petroleum Exploration and Production Technology**, v. 9, n. 3, p. 2091–2107, 2019. DOI: 10.1007/s13202-019-0609-x.

ZAREI, F., BAGHBAN, A. "Phase behavior modelling of asphaltene precipitation utilizing MLP-ANN approach", **Petroleum Science and Technology**, v. 35, n. 20, p. 2009–2015, 2017. DOI: 10.1080/10916466.2017.1377233.

ZHENG, S., FOGLER, H. S., HAJI-AKBARI, A. "A Fundamental Wax Deposition Model for Water-in-Oil Dispersed Flows in Subsea Pipelines", **AIChE Journal**, v. 63, p. 4201–4213, 2017. DOI: 10.1002/aic.15750.



## 2

### Literature review

#### 2.1

##### Oil and Gas Production

An oil reservoir is formed under high pressure and temperature conditions where the organic material is transformed into hydrocarbons. Usually, it contains a three-phase mixture: oil, gas, and water. The gas phase is mainly composed of methane ( $\text{CH}_4$ ), but also other light hydrocarbons (such as ethane ( $\text{C}_2\text{H}_6$ ), propane ( $\text{C}_3\text{H}_8$ ), butane ( $\text{C}_4\text{H}_{10}$ )), and other compounds (hydrogen sulfide ( $\text{H}_2\text{S}$ ), carbon dioxide ( $\text{CO}_2$ ), water vapor and others). The crude oil phase is a complex mixture of organic compounds, mainly hydrocarbons. The latter phase is composed of the formation of water. This water may contain different kinds of ions such as  $\text{K}^+$ ,  $\text{Na}^+$ ,  $\text{Mg}^{2+}$ ,  $\text{Ca}^{2+}$ ,  $\text{Ba}^{2+}$ ,  $\text{Sr}^{2+}$ ,  $\text{Cl}^-$ ,  $\text{HCO}_3^-$ ,  $\text{SO}_4^{2-}$ , and others that naturally exist in the reservoir (RENPU, 2011; DEVOLD, 2006).

The oil and gas industry, Figure 2.1, is commonly divided into three sectors: upstream, midstream, and downstream. The upstream sector is responsible for searching for the oil and gas wells and then extracting their raw resources, bringing the oil and gas to the surface (AALSALEM *et. al*, 2018). The problems with fouling faced in this area were the motivation for this work.

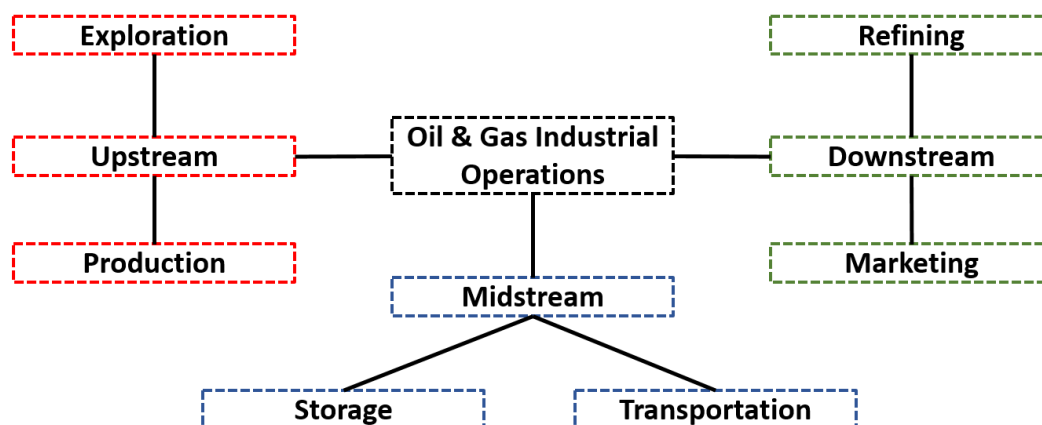


Figure 2.1: Subdivisions of Oil and Gas Industrial operations (adapted from AALSALEM *et. al.*, 2018).

### 2.1.1

#### Flow assurance

Flow assurance is a multidisciplinary area responsible for guaranteeing the transport of hydrocarbons in all the industry sectors, concerning the safety and economical parts. Its importance is even more significant in deepwater and ultra-deepwater scenarios due to the operational conditions (high pressure and low subsea temperatures) and the long distances associated with these environments (OLAJIRE, 2020; MELCHUNA *et al.*, 2020; DE OLIVEIRA AND GONÇALVES, 2012). With the beginning of the exploration of the well, the system is exposed to an abrupt change in its conditions. Directly, there is a reduction in the temperature due to the deepwater characteristic and the pressure, which continues to drop from the wellhead to the end of the pipelines. This causes several alterations in the equilibrium and saturation conditions of the several species present in the mixture that leaves the well, resulting in more favorable conditions and the appearance of solid deposits and obstructions along the pipelines during the production (BELL *et al.*, 2021; THEYAB, 2018; KARTNALLER, 2018).

The most common kinds of obstruction are associated with phase change or precipitation, caused by the formation of gas hydrate, organic fouling (asphaltene and wax), and inorganic fouling (scale), Figure 2.2 (MELCHUNA *et al.*, 2020; DE OLIVEIRA AND GONÇALVES, 2012).

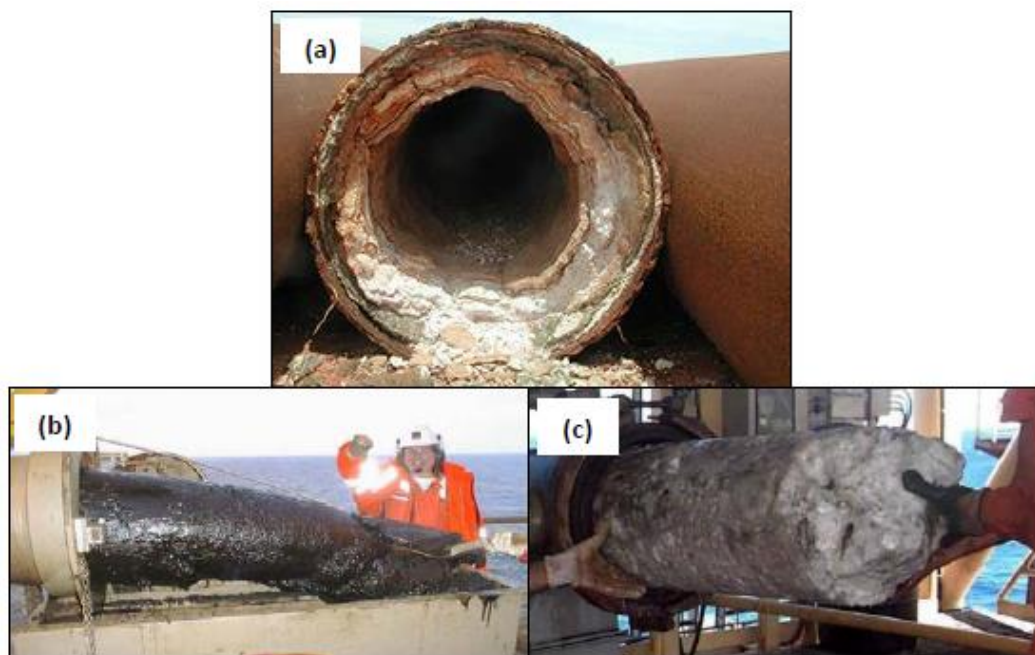


Figure 2.2: Real cases of fouling formations: (a) inorganic (scale), (b) organic, (c) gas hydrate (KARTNALLER, 2018; Hw Institute of Petroleum Engineering; Doelman, 2013; Irmann-Jacobsen e Hægland, 2014)

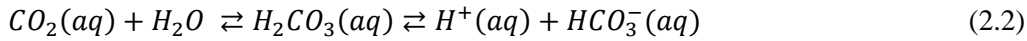
Gas hydrates are formed under the production and transportation conditions, with the combination of high pressure and low temperature, in which gas molecules, such as  $\text{CH}_4$ ,  $\text{C}_2\text{H}_6$ ,  $\text{C}_3\text{H}_8$ ,  $\text{CO}_2$ , and  $\text{H}_2\text{S}$ , are trapped in cages by hydrogen bonding with water through van der Waals forces, creating crystal-like solids (Figure 2.2(c)). Their formatting can lead to the shutting down of production, costing around \$1 million per day (QASIM *et al.*, 2019; NASIR *et al.*, 2020).

There are several approaches to avoid hydrate formation in pipelines, one of the most adopted is the injection of thermodynamic hydrate inhibitors (THIs), such as monoethylene glycol (MEG) (one of the most commonly used). The inhibitor changes the pressure and temperature conditions to the hydrate stability to values beyond the operation conditions (LIM *et al.*, 2020).

Another significant fouling problem is scale, caused by the inorganic deposition on the pipelines due to the exceeded solubility limit of one or more components, and the solution becomes saturated. That condition can be achieved by a change in the ionic composition, pH, pressure, temperature, partial pressure of  $\text{CO}_2$ , and other factors. The most common kinds of scales

found in the oilfield are calcium carbonate ( $\text{CaCO}_3$ ), calcium sulfate, barium sulfate, and strontium sulfate. To prevent scale formation, it can also be used chemicals inhibitors. (OLAJIRE, 2015; DYER and GRAHAM, 2002; KUMAR *et al.*, 2018).

The  $\text{CaCO}_3$  can be used as an example to show the effects of pressure variation and pH on their formation. For that, it is important to analyze the equilibrium equation of the  $\text{CO}_2$  solubilization on water, Eq. 2.1, and the equilibrium dissociation equation of its species, Eqs.2.2-3. The presence of the different species is directly related to the pH conditions, in which the low values favor the  $\text{CO}_2$ , and high pH values increase the predominance of the  $\text{CO}_3^{2-}$ . The relation between the presence of the species is dependent on the temperature conditions (KARTNALLER, 2018).



Another variable that has a significant influence on the equilibria is pressure. During the production of the fluids that come from the reservoir to the surface, there is a decrease in its pressure, which results in a reduction in the  $\text{CO}_2$  solubility due to the Le Chatelier principle. This change results in the exit of  $\text{CO}_2$  from the solution, which leads to an increase in pH and a rise in the  $\text{CO}_3^{2-}$  relative concentration, which can lead to the system reaching the supersaturation condition for the  $\text{CaCO}_3$ , Eq. 2.4 (KARTNALLER, 2018).

A concern with using the chemical formulation as an inhibitor for different fouling problems is how they affect the formation of the other types of fouling and the performance of other products. For example, this happens with gas hydrate chemical inhibitors, such as MEG, and how their interaction with the water molecules affects the equilibrium of the other system species. The direct effect is the increase in the activity of the ions that elevates the supersaturation ratio favoring the scale formation. However, the works of Kartnaller *et al.* (2018b) and Chao *et al.* (2020) show that the use of MEG as hydrate inhibitor results in an increase in the scaling time. The opposite

expected result because it minimizes the  $\text{CaCO}_3$  accumulation in the system. That shows the complexity of using these chemical products and why they need to be tested to verify their effect in real situations.

Specifically for scale management, a typical methodology to evaluate a commercial inhibitor before its application during production is the dynamic tube blocking test (TBT), Figure 2.3. It allows not only to verify if the product works but also to determine an inhibitor's minimum inhibitor concentration (MIC). These experiments can also be modified to study the inorganic salt morphologies. Also, these experiments result in an extensive data set that can be applied to modeling the scale formation, which is a complicated task since it happens in a complex system (SANTOS *et al.*, 2017; PAZ *et al.*, 2017; RAMZI *et al.*, 2016). An example of this kind of study is the works of Kartnaller *et al.* (2018) and Chao *et al.* (2020), which study the effect of the MEG, a hydrate inhibitor, on scale formation, using the  $\text{CaCO}_3$  as a model of the salts.

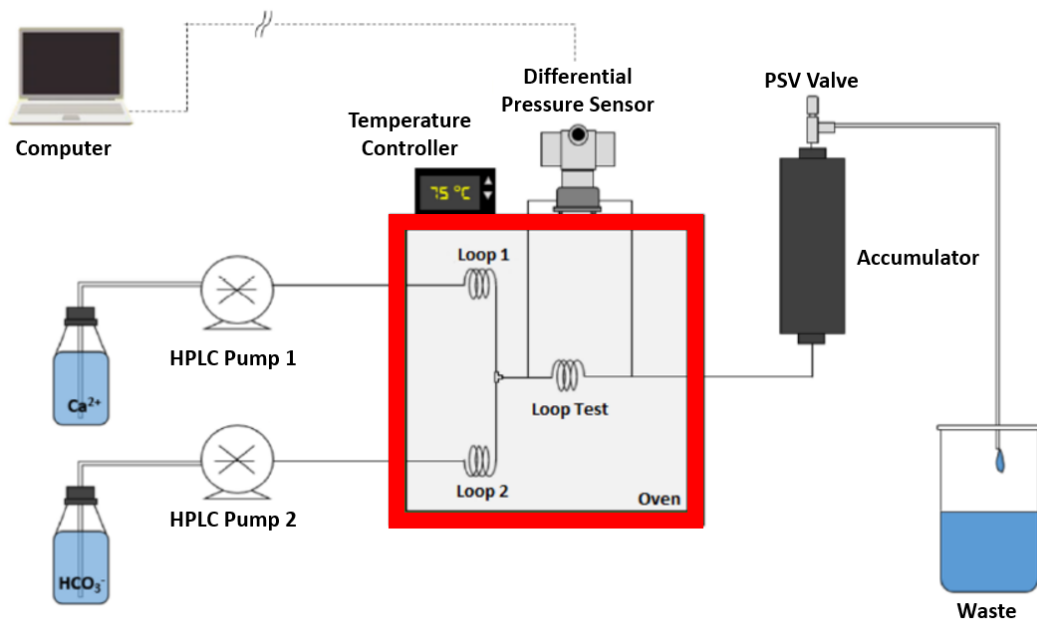


Figure 2.3: Scheme of a Dynamic Scale Loop (DSL) system used in a TBT experiment (adapted from KARTNALLER *et al.*, 2018).

Organic fouling is another serious flow assurance problem. Paraffin wax deposition tends to occur peripherally of the flow, like the walls of the pipelines, progressively decreasing the sectional area and blocking the pipelines completely. This question is more preoccupation in subsea pipelines due to the low-temperature conditions once the wax formation happens when the

operational conditions are below the wax appearance temperature (WAT) (MAHIR *et al.*, 2019; De SOUZA *et al.*, 2019; ZHENG *et al.*, 2017). To remediate wax fouling, several kinds of treatments can be adopted, such as thermal, mechanical, chemical, and biological methods (ALADE *et al.*, 2020). An example of a chemical treatment used is the Nitrogen Generated Systems (NGSs) application, which can also be applied to solve gas hydrates problems (DE OLIVEIRA, 2019).

NGS consists of a highly exothermic reaction with nitrogen gas ( $N_2$ ) and water ( $H_2O$ ) as products. A reaction that can be classified as NGS happens between ammonium chloride ( $NH_4Cl$ ) and the sodium nitrite ( $NaNO_2$ ) ( $\Delta H_{Rx} = -79.95 \text{ kcal}\cdot\text{mol}^{-1}$ ), Eq. 2.5. This kind of reaction has its kinetics strongly influenced by the pH value, as shown in Figure 2.4, in which the rate constant ( $k$ ) has an exponential increase for pH values below 4 (NGUYEN *et al.*, 2001; NGUYEN *et al.*, 2003; and DE OLIVEIRA, 2019).

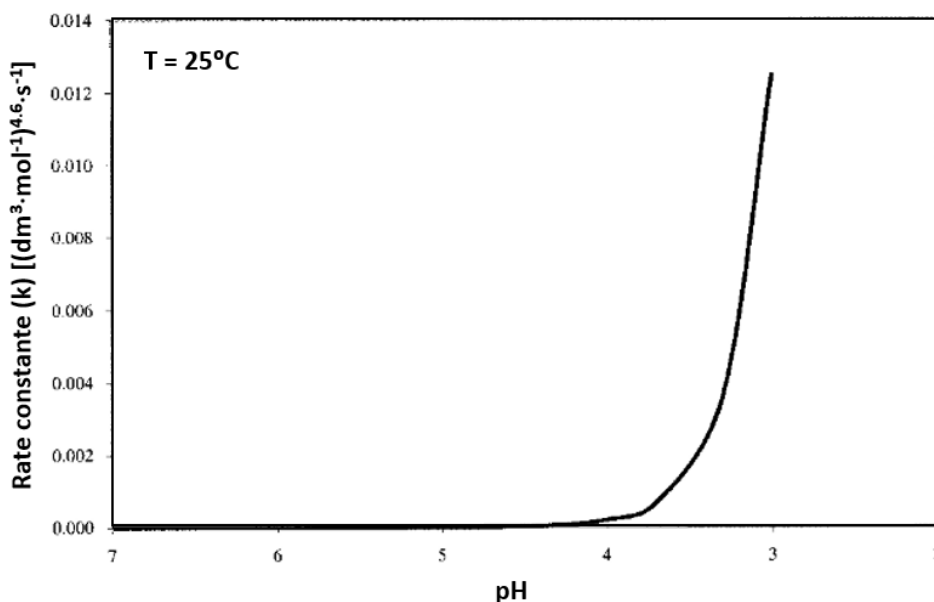
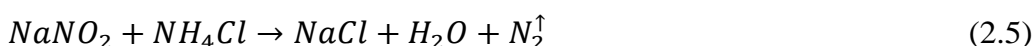


Figure 2.4: Reaction rate constant as a function of pH (adapted from NGUYEN *et al.*, 2001).

## 2.2

### Modeling in the oil and gas industry

The day-by-day monitoring and controlling of tests and experiments to solve flow assurance issues or other sources of problems have been generating a massive amount of datasets that only tends to rise with the introduction of the new technologies in oil and gas production (MOHAMMADPPOR and TORABI, 2020). However, the digitalization of most companies and their deepening entry into the "Oil and Gas 4.0" phase has been slow (LU *et al.*, 2019). This digitalization process demands a rigorous use of the Big Data (BD) analysis, but that could lead to an improvement in operational efficiency (NGUYEN *et al.*, 2020). That represents a promising field for applying different types of Artificial Intelligence (AI) strategies to solve some of the problems and challenges of this industry sector, Figure 2.5.

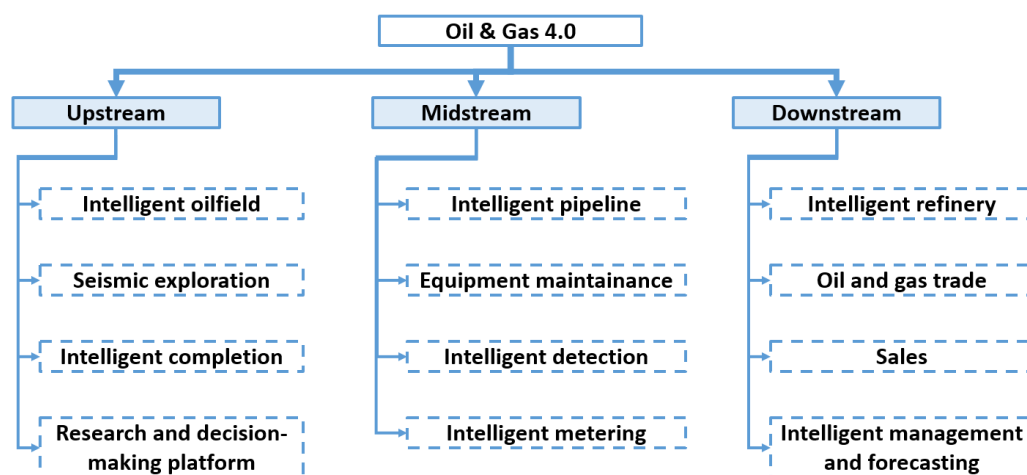


Figure 2.5: Scenarios with good application potential in the context of the "Oil and Gas 4.0" era (adapted from LU *et al.*, 2020).

Following this necessity of the industry, many works have been published applying different kinds of AI to use those datasets and develop models that help to solve some issues or to facilitate the integration of the equipment and the control strategies. For example, creating digital twins, Figure 2.6, and soft sensors to be applied in all three big areas of the oil and gas industry (WANASINGHE *et al.*, 2020; LU *et al.*, 2019; RAHMANIFARD and PLAKSINA, 2018).

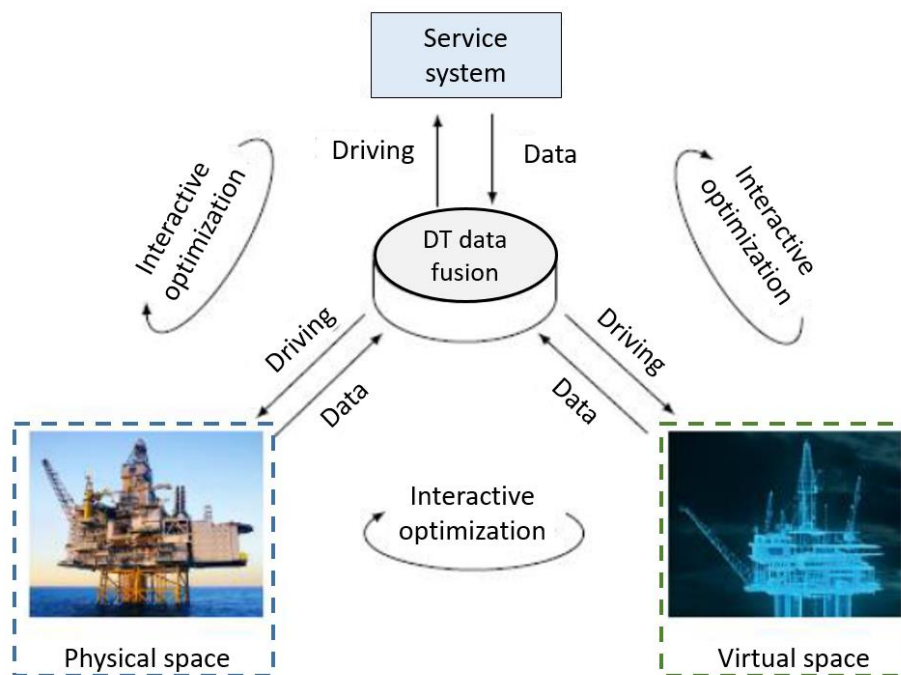


Figure 2.6: Scheme of physical and virtual space using a digital twin framework with five components (physical space, virtual space, connection between them, data, and service) (adapted from WANASINGHE *et al.*, 2020).

Some examples of the application of AI in the oil and gas industry are Multilayer Perceptron (MLP) for wax deposition (AMAR *et al.*, 2021), MLP for prediction of volume fraction in a three-phase flow meter (ISLAMI RAD and PEYVANDI, 2019), MLP to model the asphaltene precipitation (ZAREI and BAGHBAN, 2017), CNN to predict oil and gas flow rate of a two-phase flow (XU *et al.*, 2020), CNN to predict the volume flow rates of the individual phases in a three-phase mixture (LI *et al.*, 2021), LS-SVM to determine the stability region in crude oil (CHAMKALANI *et al.*, 2012).

## 2.3

### pH meter techniques




The pH measurement is essential to be monitored in several kinds of processes associated with the chemical industry. Its control can be used to regulate the solubility of chemicals or biomolecules, avoid undesired side reactions or promote the mechanism for the desired product and influence the kinetics of the chemical reactions. There are several kinds of techniques to gauge the pH value, using the classical chemical indicators, glass electrodes to



optical fiber, fluorometric, and Ion sensitive FET (field effect transistor) pH sensors (KHAN *et al.*, 2017). The pH glass electrodes are the most commonly used type of pH measurement. Still, they usually do not perform well under extreme conditions like extremes of the pH scales, high pressure, or high temperature. (GOTOR *et al.*, 2017 and BYCHKOV *et al.*, 2020).

The *in situ* measurement for this kind of environmental condition does not have a wide range of equipment available in the market. Some of that, for the high pressurized system, are shown in Table 1, although it has been studied through the last decades using different approaches (BYCHKOV *et al.*, 2020 CROLET and BONIS, 1983). For example, Samaranayake and Sastry (2013) used a high pressure pH sensor based on electrical signals to measure the properties under hydrostatic pressure up to 800 MPa. This study also reported the use of different methods to develop high pressure pH sensors, such as glass electrodes, electrical conductivity, reaction volume, and spectrophotometry in the period between 1959 and 2010.

Table 1: pH meter electrodes for pressurized system (Hanna Instruments, 2021; Ato, 2021; Winn-Marion Companies, 2021).

Electrodes	Companies	Maximum pressure (bar)	Price (Dollars)
	Hanna Instruments	6	261.16
	Ato	10	351.75
	Winn-Marion Companies	13.8	890.41

### 2.3.1

#### pH sensors using image processing

As an exciting alternative to be applied in extreme conditions, the optical pH sensors, using fluorescence or absorbance, or image-based pH sensors, have been developed for *in situ* and laboratory analysis. These techniques have the disadvantage of the use of optically active molecules to act as indicators. The fluorescence methods have been used to determine the pH values in the extremes of the traditional pH scale, using ANN models to interpret the spectrophotometer signal (SAFAVI and BAGHERI, 2003), smartphone apps to predict the pH value through a picture of the sensor under a UV light (GOTOR *et al.*, 2017) or picture of a pH sensor stripes using Least Squares-Support Vector Machine (LS-SVM) to classify the pH values (MUTLU *et al.*, 2017).

De Oliveira *et al.* (2019) proposed a method to measure the pH values in a pressurized system for real-time application through image analysis, Figure 2.7. In the study, a webcam was used to collect an image of the pressurized reactor in the RGB color system, where software pre-developed by the research group processed the RGB values. Then the RGB was converted to the HSV system to be applied in the proposed equation that correlated the hue value with the pH. The method was developed to work on the pH range of 2 to 10, using as an indicator a mix of buffer solution known as the Korthoff indicator, being tested on the pressure range of 0 to 6.0 MPa.

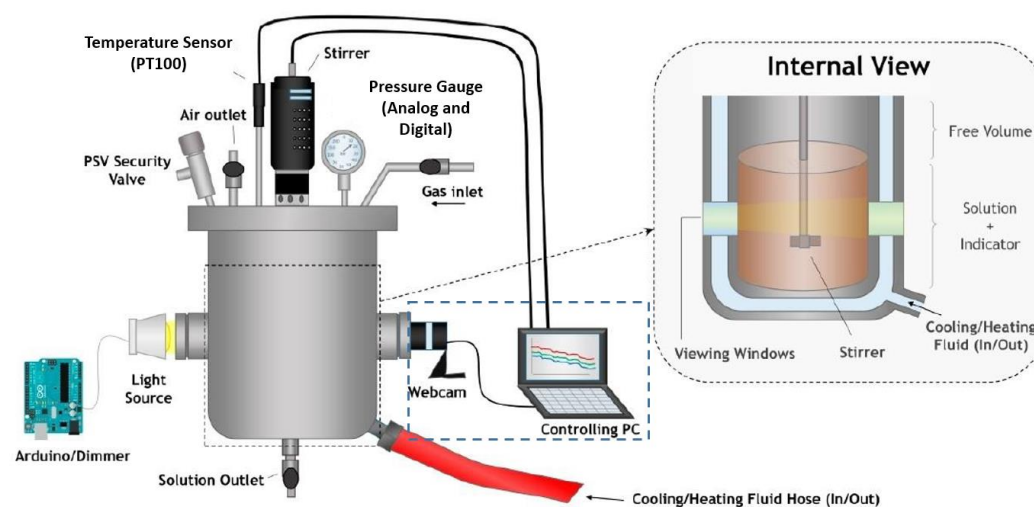


Figure 2.7: Experimental setup scheme (adapted from DE OLIVEIRA *et al.*, 2019)

Although the good performance obtained in that work, a solution using only one software, with the possibility of a future wireless application developed in open-source software are interesting reasons to develop studies in this area. For that, different deep learning strategies can be used to create models to classify and predict the pH values through image analysis.

## 2.4

### Soft sensors

Soft sensors are usually predictive models for a variable of interest, using the information of other available variables and process parameters. This characteristic allows the estimation and monitoring, in real-time, of operational parameters that before needed to be sent to the laboratory to be analyzed. Also, software tools are not subject to mechanical problems and have easier to maintain than a conventional sensor, given an economy for the process manager (KADLEC *et al.*, 2009; POERIO and BROWN, 2018). Some of the challenges that could be found during their development are presented in Figure 2.8.

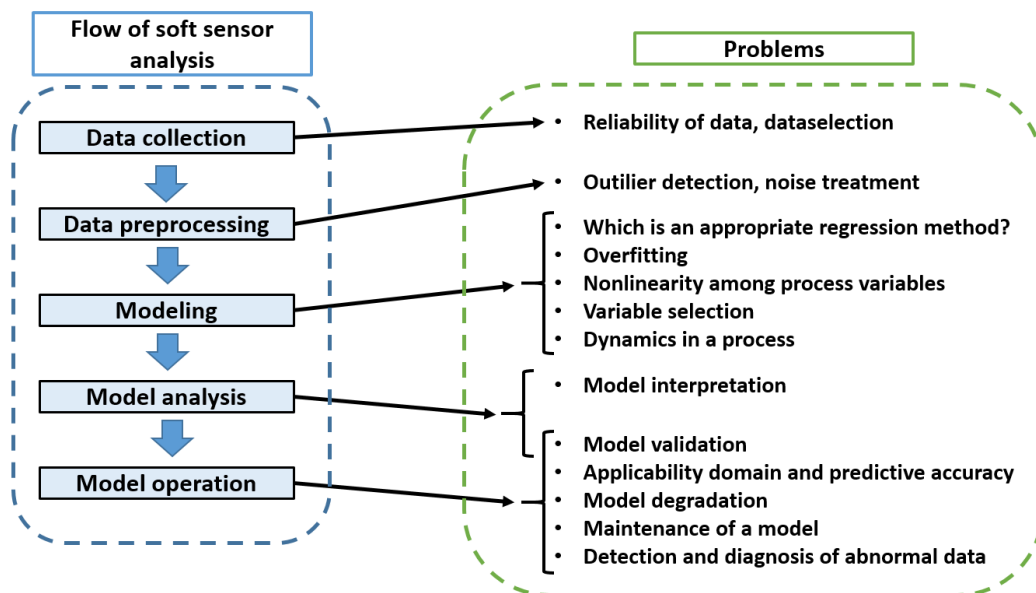


Figure 2.8: Flow of soft sensor analysis and problems involved at each stage (adapted from FUNATSU, 2018).

They can be divided into three main types: First Principles Models (FPM),

data-driven models, and mixed models. The FPM models (white-box models) are built mainly from the mechanical knowledge of the process, which can be hard or complex to determine and demand a significant computation effort and time. In turn, the data-driven models (black-box models) are created using only the process dataset, making them a more popular strategy for developing soft sensors. For that, several kinds of AI methods can be used, such as Artificial Neural Network (ANN), fuzzy logic, Support Vector Machine (SVM), Decision Tree (DT), Principal Component Analysis (PCA), hybrid methods, and others (YAN *et al.*, 2017; SHANG *et al.*, 2014).

ANN are mathematical models developed based on the biological neural systems, initially presented in the work of MacCulloh and Pitts (1943) (KUMAR *et al.*, 2013). ANN models are one of the most common strategies explored due to their advantages as training and adaptive structure (LI *et al.*, 2017). ANNs represent a large class of model structures, and one of the most popular ones is the Multi-Layer Perceptrons (MLP). However, the MLP could present some optimization problems with more deep structures with more than two hidden layers (SHANG *et al.*, 2014).

In this case, the model needs a bigger number of hidden layers or even more complex structures. They are known as Deep Learning (DL) techniques (SUN and GE, 2021) and are very present in the chemical engineering field, Figure 2.9. Among the ANN techniques that could be classified as DL, one can be pointed out the Convolutional Neural Network (CNN), traditionally used in image classification (MADHAN *et al.*, 2021).

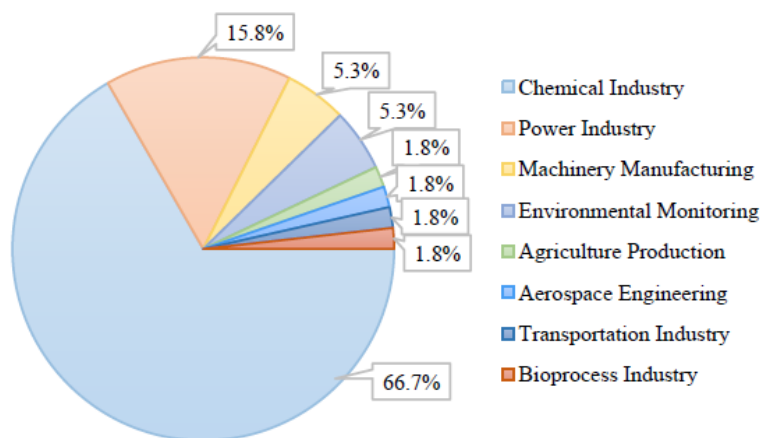


Figure 2.9: Statistics on exiting relevant work applications in different fields (adapted from SUN and GE, 2021).

### 2.4.1

#### Multi-Layer Perceptron (MLP)

MLPs are usually formed by three layers: input layer, hidden layer, and output layer, Figure 2.10. In some cases, more than one hidden layer can be used. The input layer has the same number of neurons as the model's input variables, and the output one has the number of neurons equal to the number of target variables. The number of neurons in the hidden layers is one of the parameters adapted during the development of the model (CHOJACZYK *et al.*, 2015; LI *et al.*, 2017).

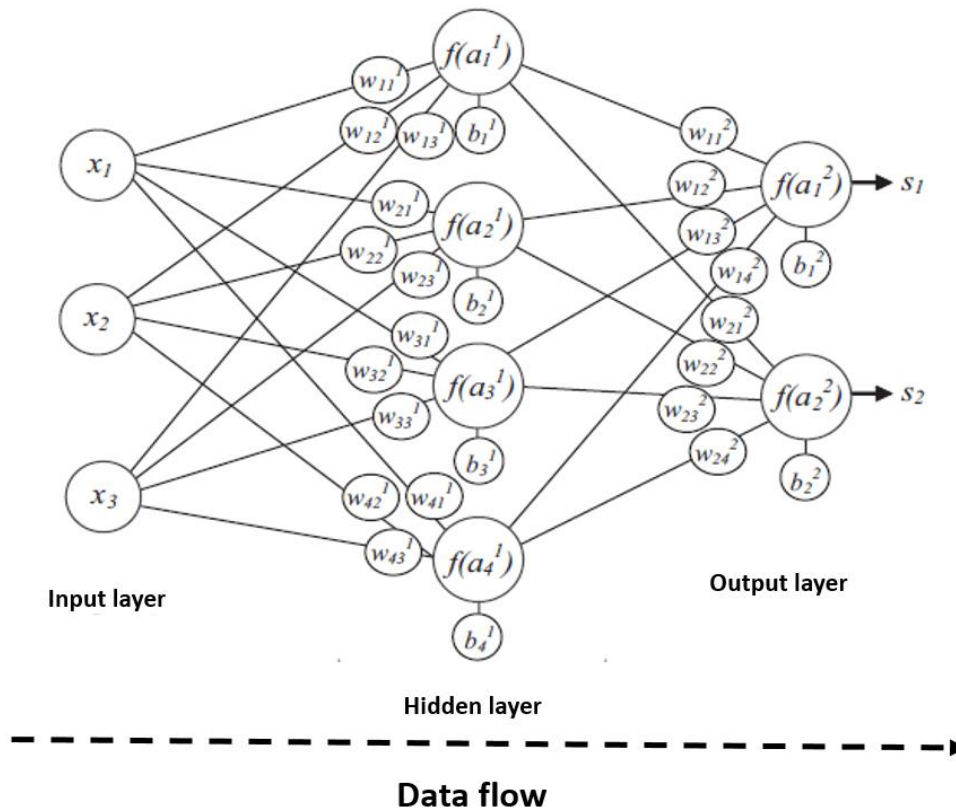


Figure 2.10: Multi-Layer Neural Network scheme (adapted from CHOJACZYK *et al.*, 2015).

The information flow through the interconnected neurons from the input to the output layer. When the neuron receives the information, input ( $x_i$ ), the information is processed according to Eq. 2.6, in which the input is multiplied by a factor called weight ( $w_{ij}$ ), related to the importance of the variable, and is added to a constant named bias ( $b_j$ ). This constant is responsible for avoiding

the resultant value ( $a_j$ ) from assuming negative values before it was passed to the activation function ( $f_{(a_j)}$ ) (LI *et al.*, 2017; HAMMOUDI *et al.*, 2019).

$$a_j = \sum_{i=1}^n w_{ij} \cdot x_i + b_j \quad (2.6)$$

where  $j$  and  $i$ , respectively, represent the identification of the origin and the destination neuron.

The activation functions' process is responsible for calculating the information that leaves the neuron to the next layer or as the model's output. The most common types of activation functions used on MLP are the sigmoid functions and the linear function (Eq. 2.7), Figure 2.11. The first type is usually represented by the logsigmoid function and the hyperbolic tangent, Eqs. 2.8-9 (SOLEIMANI *et al.*, 2013; CHOJACZYK *et al.*, 2015 and VALIM *et al.*, 2017).

$$f_{(a_m)} = x \quad (2.7)$$

$$f_{(a_m)} = \frac{1}{(1+e^{-x})} \quad (2.8)$$

$$f_{(a_m)} = \frac{2}{(1+e^{-2x})} - 1 \quad (2.9)$$

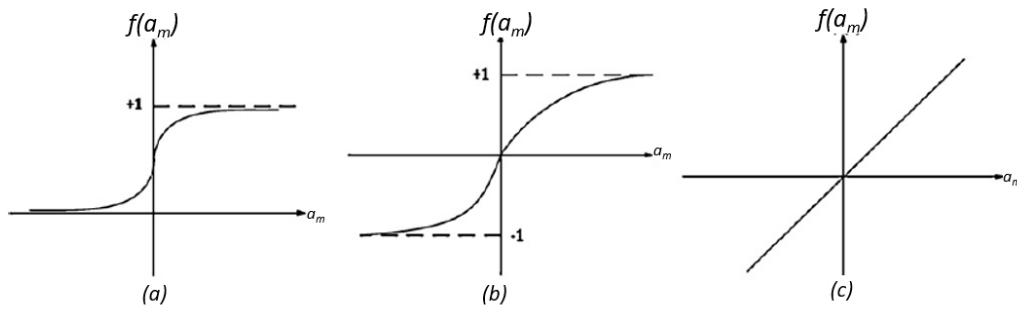


Figure 2.11: Activation functions: (a) logsigmoid, (b) hyperbolic tangent, (c) linear (adapted from SOLEIMANI *et al.*, 2013).

An essential part of the application of the MLP models is the training algorithm, being the backpropagation (BP) is one of the most common kinds applied to the MLPs. They belong to the supervised classification of the training algorithms, in which the model outputs are compared to a corresponding target

data using an error function, such as mean-squared error (MSE), and the weights and bias of the parameters are modified to minimize the evaluation parameter (SOLEIMANI *et al.*, 2013; HAYKIN, 2001).

Modifications of the BP algorithm have been proposed aiming to improve the method. The Bayesian Regularization BP algorithm uses a Bayesian regulation to enhance efficiency. The Levenberg-Marquardt BP and gradient descent with momentum and adaptive learning rate BP use a quasi-Newton method to make the convergence faster and with a smaller computation effort due to the use of an approximation of the Hessian matrix (PLUMB *et al.*, 2005; FORESEE and HAGAN, 1997; HAGAN and MEHAJ, 1994; HAYKIN, 2001).

## 2.4.2

### Convolutional Neural Network (CNN)

CNNs are a type of DL architecture based on the animal visual cortex, and it has been successfully used to extract features through image analysis. The first ones were proposed by LeCun *et al.* (1989), but they became more popular after overcoming some technological challenges at the beginning of the last decade (MARQUES, 2018; BOUWMANS, 2019; YUAN *et al.*, 2020). As a DL method, the CNN models have some advantages compared to more traditional AI strategies (ZAN *et al.*, 2020):

- Use of the raw data directly in the training and test in many cases, avoiding the pre-processing data necessity;
- Being able to be applied in more complex tasks;
- Learn the most appropriate features from the classification problems.

The structure of the CNN, Figure 2.12, similarly to the MLP, can be described in three parts: input layer, hidden layers, and output layer. However, in these cases, there are multiple hidden layers, which can be split into three classes: convolutional, pooling (or subsampling), and fully connected (YAO *et al.*, 2019).

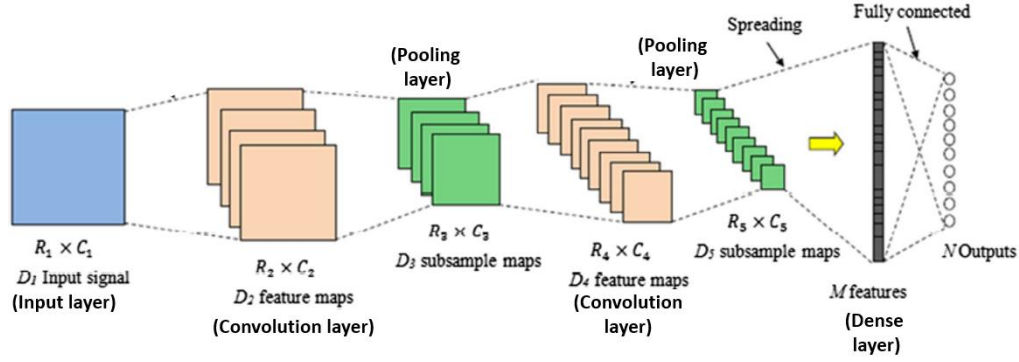


Figure 2.12: Typical CNN structure (adapted from ZAN *et al.*, 2020).

The most significant component of a CNN is the convolutional layers. The weights and biases are organized in a series of convolutional filters (or kernels). The filter coefficients are optimized during the model training, where each filter learns to extract specific features or patterns from its input layer. As the convolution process, Figure 2.13, is a linear operation, this non-linearity in the signal is granted by the activation function, which in the case of the CNN, the ReLU function (Rectified Linear Unit), Eq. 2.10, is usually chosen to be used (SHEN *et al.*, 2021; ZAN *et al.*, 2020; CASTANEDA, 2017).

$$ReLU(x) = \max(0, x) \quad (2.10)$$

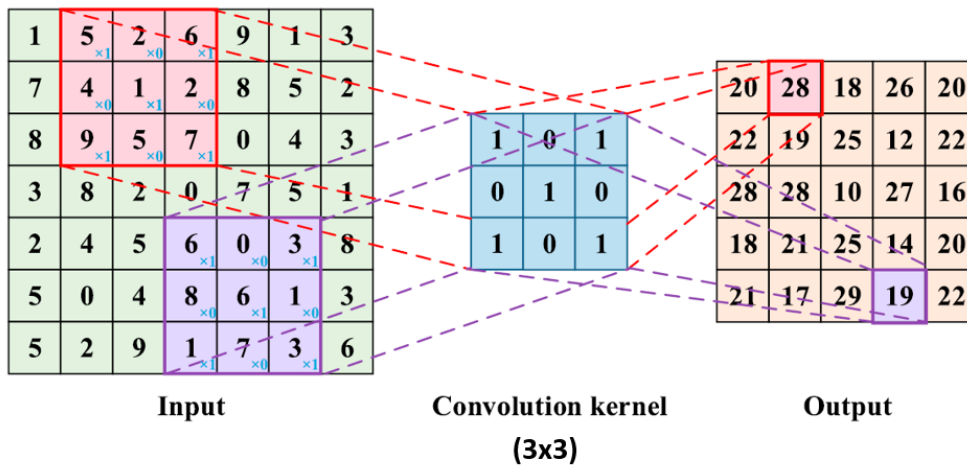


Figure 2.13: Convolutional operation with a 3 x 3 convolutional kernel (adapted from YUAN *et al.*, 2020).



The pooling layer, Figure 2.14, is completed by downsampling the feature maps resulting from the convolutional operations. Unlike the convolutional layer, the subsampling layer has no specific value or parameters to represent the compositive features of the receptive field. It can quickly reduce the scale of the feature maps but also the sensitivity of similar light features. The most common kinds of strategies applied are max pooling and average pooling (YUAN *et al.*, 2020; ZAN *et al.*, 2020; MARQUES, 2018).

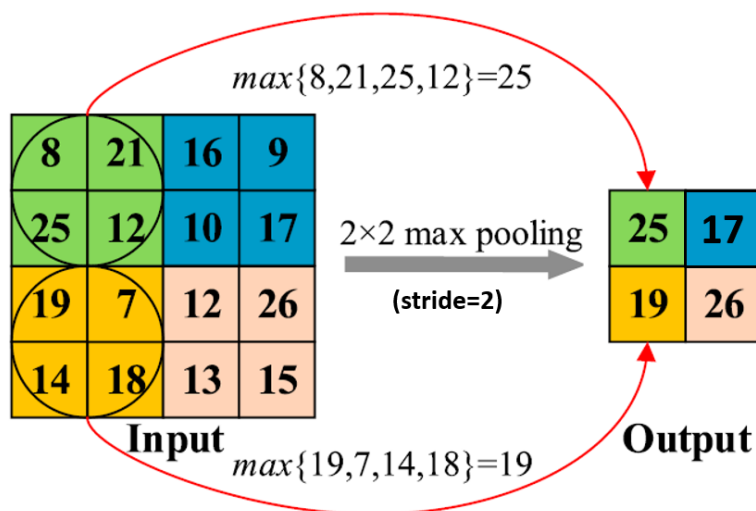


Figure 2.14: Max pooling operation with 2 x 2 size (adapted from YUAN *et al.*, 2020).

Fully connected layers (or dense layers) are the last ones of the hidden layers. They have all their neurons fully connected with the previous layer, such as the hidden layer in the MLP structure (ZAN *et al.*, 2020). The last layer of the CNN, the output layer, is also a dense layer, which in the case of a regression model, has the number of the output variables. For classification models, the number of neurons is equal to the number of classes, having as output in each neuron the probability of the belongs to each class. Usually, the function *softmax* is used for this last case (MARQUES, 2018; ZAN *et al.*, 2020).

### 2.4.3

#### Support Vector Machine (SVM)

SVM is a Machine Learning technique to create supervised learning models, which could be used for classification and regression problems. It was presented in the work of Cortes and Vapnik (1995), based on a statistical

learning theory and structural risk minimization, for binary classification (WU *et al.*, 2018). SVM is based on the geometric distance class interval maximization strategy, giving the model a strong generalization ability (PENG *et al.*, 2020).

The first applications of the SVM were developed for binary classification with linearly separable data, in which the classes can be split by a straight line, as shown in Figure 2.15. (NOGUEIRA, 2021).

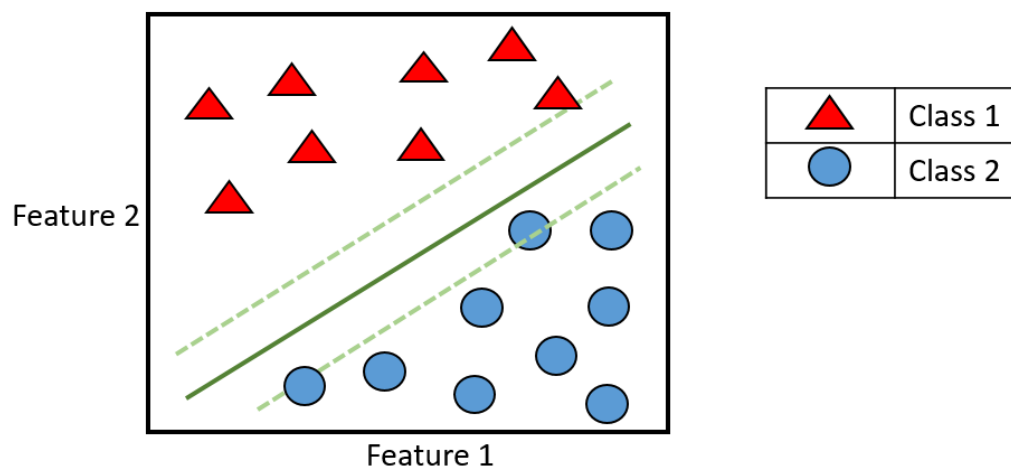


Figure 2.15: Linearly separable data with two dimensions and two classes (solid line – hyperplane separating the classes; dashed lines – margins of the hyperplane) (adapted from NOGUEIRA, 2021).

In the following years, the SVM models were evolving, allowing them to be used for non-linearly separable data problems and multi-class problems. The first challenge was solved using kernels, which are mathematical functions ( $\phi$ ) that transform the data from a given space (Input space) to a new high-dimensional one (Feature Space), where the classes can be separated by a linear surface (hyperplane), Figure 2.16 (CHAUHAN *et al.*, 2019).

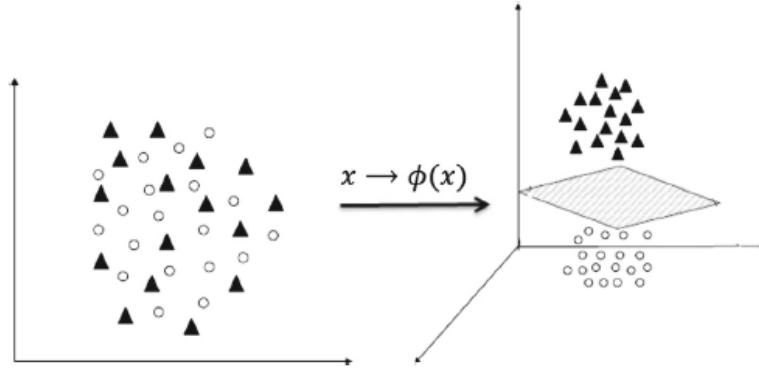


Figure 2.16: Representation of the input transformation from the Input space (right) to the Feature space (left) by the use of the kernels (adapted from CHAUHAN *et al.*, 2019).

The most often used kernel functions, Figure 2.17, are: linear (Eq. 2.11), polynomial (Eq. 2.12), RBF (Radial-Basis Function, Eq. 2.13), and sigmoid (Eq. 2.14) (CHAUHAN *et al.*, 2019; GONG *et al.*, 2019).

$$\varphi(x) = \text{Kernel}_{linear}(x_i, x_j) = (\text{gamma}(x_i, x_j) + \text{coef}) \quad (2.11)$$

$$\varphi(x) = \text{Kernel}_{poly}(x_i, x_j) = (\text{gamma}(x_i, x_j) + \text{coef})^{\text{degreee}} \quad (2.12)$$

$$\varphi(x) = \text{Kernel}_{RBF}(x_i, x_j) = \exp(-\text{gamma}\|x_i - x_j\|^2) \quad (2.13)$$

$$\varphi(x) = \text{Kernel}_{sigmoid}(x_i, x_j) = \tanh(\text{gamma}(x_i, x_j) + \text{coef}) \quad (2.14)$$

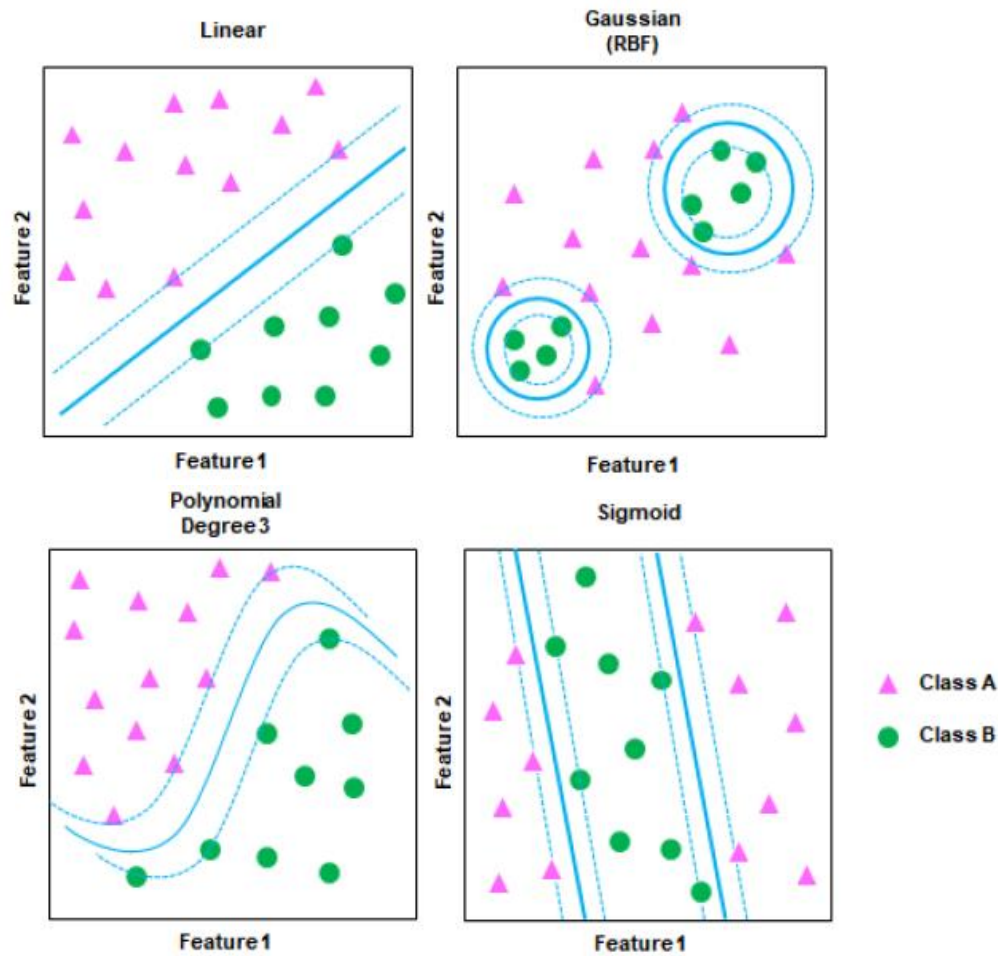


Figure 2.17: Kernel functions behavior in classification with SVM (adapted from NOGUEIRA, 2021).

The *gamma* and *degree* are hyperparameters that could be modified to adjust the model fit. The *gamma* represents the influence of each data in training in general and in the surface position. In turn, the *degree* is the parameter associated with the polynomial level. The *coef* is an independent term of each function (RHYS *et al.*, 2020; SCIKIT-LEARN, 2022 and NOGUEIRA, 2021).

The application of the SVM models to the multi-class problem was allowed by the implementation of strategies such as One-versus-One (OvO) and One-versus-Rest (OvR or One-versus-All (OvA)). OvR is probably one of the first techniques applied for the multi-class classification problem. Its class is separated from the others by a hyperplane, reducing the situation to a group of binary classification problems. For the OvO technique, the classification is realized between each pair of classes, usually resulting in a higher number of hyperplanes, although it could demand less from the computer (CHAUHAN *et*

*al.*, 2019; DING *et al.*, 2019; RHYS *et al.*, 2020 and NOGUEIRA, 2021). Figure 2.18 shows an example of OvR and OvO approaches with a three-class situation.

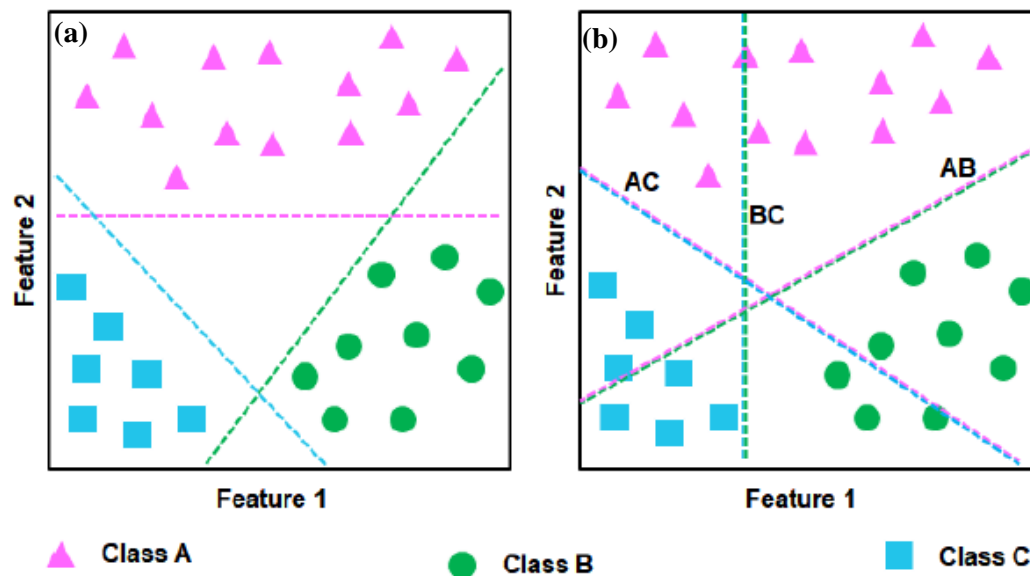


Figure 2.18: Multi-class SVM approaches (a) OvR (One-versus-Rest) and (b) OvO (One-versus-One) (dashed lines – hyperplanes) (adapted from NOGUEIRA, 2021).

#### 2.4.4

##### Decision tree (DT)

The decision tree is another AI technique with supervised learning algorithms and is commonly used for classification problems. DT has a simple form that combines several binary tests in its structure (GEURTS *et al.*, 2009 TANGIRALA, 2020). It is structured as a tree, Figure 2.19, hierarchically structured with a group of interconnected nodes. The process starts in the root node where the input is inserted, then it and each internal node of the tree are responsible for a test, and each terminal node (or leaf node) is labeled with a class (PRIYAM *et al.*, 2013; GEURTS *et al.*, 2009; ARAUJO, 2017).

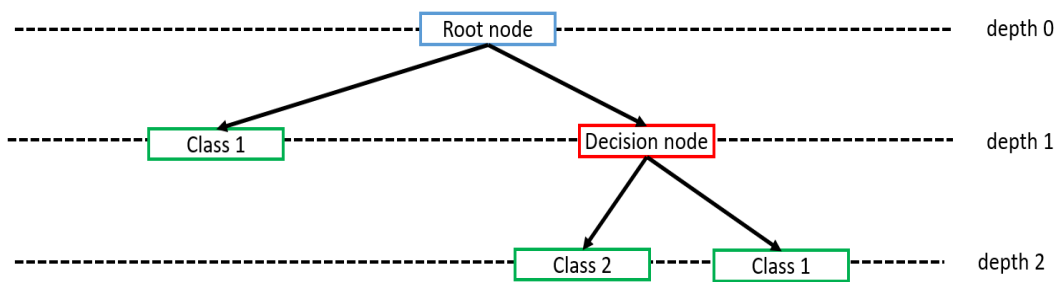


Figure 2.19: Example of a general decision tree for classification (adapted from BARROS, 2014).

Other important concepts are *depth* and *breadth*. The first one is related to the number of levels (layers) that DT has from the root node to the terminal node. The *breadth* refers to the number of internal nodes in each level of the tree (BARROS, 2014).

For the training process of the DT, models have used some functions to measure the impurity level of a node, in which the lower this parameter, the better the prediction. This helps to decide the necessity to split the node. In the case of classification problems, the most common functions are the Gini impurity (*gini*) and the cross-entropy (*entropy*) (HASTIE et. al., 2009; BARROS, 2014).

## 2.5

## References

- AALSALEM, M. Y., KHAN, W. Z., GHARIBI, W., *et al.* "Wireless Sensor Networks in oil and gas industry: Recent advances, taxonomy, requirements, and open challenges", **Journal of Network and Computer Applications**, v. 113, n. October 2017, p. 87–97, 2018. DOI: 10.1016/j.jnca.2018.04.004.
- ALADE, O. S., HASSAN, A., MAHMOUD, M., *et al.* "Novel Approach for Improving the Flow of Waxy Crude Oil Using Thermochemical Fluids: Experimental and Simulation Study", **ACS Omega**, v. 5, n. 8, p. 4313–4321, 2020. DOI: 10.1021/acsomega.9b04268.
- AMAR, M. N., JAHANBANI GHAFAROKHI, A., SHANG WUI NG, C. "Predicting wax deposition using robust machine learning techniques", **Petroleum**, 2021. DOI: 10.1016/j.petlm.2021.07.005.
- ARAUJO, M. A. de. **Metodologia Baseada em Medidas Dispersas de Tensão e Árvores de Decisão para Localização de Faltas em Sistemas de Distribuição Modernos**. 2017. 208 f. Universidade de São Paulo, 2017.
- ATO. **pH Electrode for High Pressure**. [S.d.]. Disponível em: <https://www.ato.com/ph-electrode-for-high-pressure>. Acesso em: 14 jun. 2021.
- BARROS, R. C. **On the automatic design of decision-tree induction algorithms**. 2014. 202 f. Universidade de São Paulo - Campus São Carlos, 2014.
- BARUPAL, D. K., FIEHN, O. "Generating the blood exposome database using a comprehensive text mining and database fusion approach", **Environmental Health Perspectives**, v. 127, n. 9, p. 2825–2830, 2019. DOI: 10.1289/EHP4713.
- BELL, E., LU, Y., DARABOINA, N., *et al.* "Thermal methods in flow assurance: A review", **Journal of Natural Gas Science and Engineering**, v. 88, n. July 2020, p. 103798, 2021. DOI: 10.1016/j.jngse.2021.103798.
- BOUWMANS, T., JAVED, S., SULTANA, M., *et al.* "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation", **Neural Networks**, v. 117, p. 8–66, 2019. DOI: 10.1016/j.neunet.2019.04.024.
- BYCHKOV, A. Y., BÉNÉZETH, P., POKROVSKY, O. S., *et al.* "Experimental determination of calcite solubility and the stability of aqueous Ca- and Na-carbonate and -bicarbonate complexes at 100–160 °C and 1–50 bar pCO<sub>2</sub> using in situ pH measurements", **Geochimica et Cosmochimica Acta**, v. 290, p. 352–365, 2020. DOI: 10.1016/j.gca.2020.09.004.
- CHAMKALANI, A., MOHAMMADI, A. H., ESLAMIMANESH, A., *et al.* "Diagnosis of asphaltene stability in crude oil through “two parameters” SVM model", **Chemical Engineering Science**, v. 81, p. 202–208, 2012. DOI: 10.1016/j.ces.2012.06.060.

CHAO, J., ZHANG, L., FENG, R., *et al.* "Experimental study on the compatibility of scale inhibitors with Mono Ethylene Glycol", **Petroleum Research**, v. 5, n. 4, p. 315–325, 2020. DOI: 10.1016/j.ptlrs.2020.07.003.

CHAUHAN, V. K., DAHIYA, K., SHARMA, A. "Problem formulations and solvers in linear SVM: a review", **Artificial Intelligence Review**, v. 52, n. 2, p. 803–855, 2019. DOI: 10.1007/s10462-018-9614-6.

CHEMICALS, D. **Triethyle Glycol - Material Safety Data Sheet 2007**. [S.d.]. Disponível em: [http://msdssearch.dow.com/PublishedLiteratureDOWCOM/dh\\_0952/0901b80380952386.pdf?filepa%0Ath=ethy](http://msdssearch.dow.com/PublishedLiteratureDOWCOM/dh_0952/0901b80380952386.pdf?filepa%0Ath=ethy). Acesso em: 12 jan. 2018.

CHOJACZYK, A. A., TEIXEIRA, A. P., NEVES, L. C., *et al.* "Review and application of Artificial Neural Networks models in reliability analysis of steel structures", **Structural Safety**, v. 52, n. PA, p. 78–89, 2015. DOI: 10.1016/j.strusafe.2014.09.002.

COMPANIES, W.-M. **ABB TB567.2.3.1.0.8.T.25, TB567 High-Pressure pH Sensor, Glass**. [S.d.]. Disponível em: <https://www.winn-marion.com/itemdetail/?itemCode=TB567.2.3.1.0.8.T.25>. Acesso em: 12 jun. 2021.

CORTES, C., VAPNIK, V. "Support-Vector Networks", **Machine Learning**, v. 20, p. 273–297, 1995.

CROLET, J. L., BONIS, M. R. "pH MEASUREMENTS IN AQUEOUS CO<sub>2</sub> SOLUTIONS UNDER HIGH PRESSURE AND TEMPERATURE.", **Corrosion**, v. 39, n. 2, p. 39–46, 1983. DOI: 10.5006/1.3580813.

DE OLIVEIRA, A. V. B. **SISTEMA DE GERAÇÃO DE NITROGÊNIO E CALOR IN-SITU, PARA APLICAÇÃO EM AMBIENTES SUBMARINHOS: UM ESTUDO EM MEIO PRESSURIZADO COM GERAÇÃO DE CATALISADOR IN-SITU**. 2019. Universidade Federal do Rio de Janeiro, 2019.

DE OLIVEIRA, A. V. B., ORTIZ, R. W. P., KARTNALLER, V., *et al.* "Real-Time Measurement of pH in Atmospheric and Pressurized Systems Using a Low-Cost Image Analysis Method", **IEEE Sensors Journal**, v. 19, n. 23, p. 10991–10998, 2019. DOI: 10.1109/JSEN.2019.2936442.

DE OLIVEIRA, M. C. K., GONÇALVES, M. A. "An effort to establish correlations between brazilian crude oils properties and flow assurance related issues", **Energy and Fuels**, v. 26, n. 9, p. 5689–5701, 2012. DOI: 10.1021/ef300650k.

DE SOUZA, A. V. A., ROSÁRIO, F., CAJAIBA, J. "Evaluation of calcium carbonate inhibitors using sintered metal filter in a pressurized dynamic system", **Materials**, v. 12, n. 11, p. 1–13, 2019. DOI: 10.3390/ma12111849. .

DEVOLD, H. **OIL AND GAS PRODUCTION HANDBOOK: An introduction to oil and gas production**. [S.l.], ABB ATPA Oil and Gas, 2006.

DING, S., ZHAO, X., ZHANG, J., *et al.* "A review on multi-class TWSVM", **Artificial Intelligence Review**, v. 52, n. 2, p. 775–801, 2019. DOI:



10.1007/s10462-017-9586-y. .

DYER, S. J., GRAHAM, G. M. "The effect of temperature and pressure on oilfield scale formation", **Journal of Petroleum Science and Engineering**, v. 35, n. 1–2, p. 95–107, 2002. DOI: 10.1016/S0920-4105(02)00217-6.

EDGAR EDUARDO MEDINA. **DEEP CNN AND MLP-BASED VISION SYSTEMS FOR ALGAE DETECTION IN AUTOMATIC INSPECTION OF UNDERWATER PIPELINES**. 2017. 88 f. UNIVERSIDADE FEDERAL DO RIO DE JANEIRO (UFRJ), 2017.

ENGINEERING., H. I. O. P. **Why are Gas Hydrates Important?** [S.d.]. Disponível em: [http://www.pet.hw.ac.uk/research/hydrate/hydrates\\_why.cfm](http://www.pet.hw.ac.uk/research/hydrate/hydrates_why.cfm). Acesso em: 27 dez. 2017.

FORESEE, F. D., HAGAN, M. T. "GAUSS-NEWTON APPROXIMATION TO BAYESIAN LEARNING". 3, 1997. **Anais [...]** [S.l: s.n.], 1997. p. 1930–1935. DOI: 10.1109/ICNN.1997.614194.

FUNATSU, K. "Process Control and Soft Sensors", **Applied Chemoinformatics**, p. 571–584, 2018. DOI: 10.1002/9783527806539.ch13.

GEURTS, P., IRRTHUM, A., WEHENKEL, L. "Supervised learning with decision tree-based methods in computational and systems biology", **Molecular BioSystems**, v. 5, n. 12, p. 1593–1605, 2009. DOI: 10.1039/b907946g.

GONG, W., CHEN, H., ZHANG, Z., *et al.* "A Novel Deep Learning Method for Intelligent Fault Diagnosis of Rotating Machinery Based on Improved CNN-SVM and Multichannel Data Fusion", **Sensors**, v. 19, 2019. DOI: 10.3390/s19071693.

GOTOR, R., ASHOKKUMAR, P., HECHT, M., *et al.* "Optical pH Sensor Covering the Range from pH 0-14 Compatible with Mobile-Device Readout and Based on a Set of Rationally Designed Indicator Dyes", **Analytical Chemistry**, v. 89, n. 16, p. 8437–8444, 2017. DOI: 10.1021/acs.analchem.7b01903.

HAGAN, M. T., MENHAJ, M. B. "Training Feedforward Networks with the Marquardt Algorithm", **IEEE Transactions on Neural Networks**, v. 5, n. 6, p. 989–993, 1994. DOI: 10.1109/72.329697.

HAMMOUDI, A., MOUSSACEB, K., BELEBCHOUHE, C., *et al.* "Comparison of artificial neural network (ANN) and response surface methodology (RSM) prediction in compressive strength of recycled concrete aggregates", **Construction and Building Materials**, v. 209, p. 425–436, 2019. DOI: 10.1016/j.conbuildmat.2019.03.119.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Second ed. New York, Springer, 2009.

HAYKIN, S. **Redes Neurais: Princípio e prática**. 2ed. ed. Porto Alegre - RS - Brazil, Bookman, 2001.

INSTRUMENTS, H. **HI-1003/3 pH Electrode for High Pressure**. [S.d.]. Disponível em: <https://www.hannainstruments.co.uk/home/21-ph-electrode-for-high-pressure>. Acesso em: 14 jun. 2021.

IRMANN-JACOBSEN, T. B. HÆGLAND, B. **Flow Assurance & Operability: A System Perspective**. 2014. Disponível em: [http://www.uio.no/studier/emner/matnat/math/MEK4450/h14/undervisningsmateriale/module-%0A4/mek4450\\_flowassurance\\_pensum.pdf](http://www.uio.no/studier/emner/matnat/math/MEK4450/h14/undervisningsmateriale/module-%0A4/mek4450_flowassurance_pensum.pdf). Acesso em: 27 dez. 2017.

ISLAMI RAD, S. Z., GHOLIPOUR PEYVANDI, R. "A simple and inexpensive design for volume fraction prediction in three-phase flow meter: Single source-single detector", **Flow Measurement and Instrumentation**, v. 69, n. June, p. 101587, 2019. DOI: 10.1016/j.flowmeasinst.2019.101587.

KADLEC, P., GABRYS, B., STRANDT, S. "Data-driven Soft Sensors in the process industry", **Computers and Chemical Engineering**, v. 33, n. 4, p. 795–814, 2009. DOI: 10.1016/j.compchemeng.2008.12.012.

KARTNALLER, V. **AVALIAÇÃO DA INFLUÊNCIA DO USO DE INIBIDORES DE HIDRATOS NO PROCESSO DE INCRUSTAÇÃO DE CARBONATO DE CÁLCIO EM SISTEMA DINÂMICO PRESSURIZADO**. 1–167 f. Universidade Federal do Rio de Janeiro (UFRJ), 2018.

KARTNALLER, V., VENÂNCIO, F., DO ROSÁRIO, F. F., *et al.* "Application of multiple regression and design of experiments for modelling the effect of monoethylene glycol in the calcium carbonate scaling process", **Molecules**, v. 23, n. 4, p. 1–12, 2018. DOI: 10.3390/molecules23040860.

KEIFER, G., EFFENBERGER, F. **Angewandte Chemie International Edition**, v. 6, n. 11, p. 951–952, 1967.

KHAN, M. I., MUKHERJEE, K., SHOUKAT, R., *et al.* "A review on pH sensitive materials for sensors and detection methods", **Microsystem Technologies**, v. 23, n. 10, p. 4391–4404, 2017. DOI: 10.1007/s00542-017-3495-5.

KUMAR, S., NAIYA, T. K., KUMAR, T. "Developments in oilfield scale handling towards green technology-A review", **Journal of Petroleum Science and Engineering**, v. 169, n. May, p. 428–444, 2018. DOI: 10.1016/j.petrol.2018.05.068.

LECUN, Y., BOSER, B., DENKER, J. S., *et al.* "Backpropagation Applied to handwritten Zip Code Recognition", **Neural Computation**, v. 1, p. 541–551, 1989. DOI: 10.1162/neco.1989.1.4.541.

LI, H., ZHANG, Z., LIU, Z. "Application of artificial neural networks for catalysis: A review", **Catalysts**, v. 7, n. 10, 2017. DOI: 10.3390/catal7100306.

LI, J., HU, D., CHEN, W., *et al.* "Cnn-based volume flow rate prediction of oil–gas–water three-phase intermittent flow from multiple sensors", **Sensors (Switzerland)**, v. 21, n. 4, p. 1–20, 2021. DOI: 10.3390/s21041245.

LIM, V. W. S., METAXAS, P. J., STANWIX, P. L., *et al.* "Gas hydrate formation probability and growth rate as a function of kinetic hydrate inhibitor (KHI) concentration", **Chemical Engineering Journal**, v. 388, n. January, p. 124177, 2020. DOI: 10.1016/j.cej.2020.124177. Disponível em: <https://doi.org/10.1016/j.cej.2020.124177>.

LU, H., GUO, L., AZIMI, M., *et al.* "Oil and Gas 4.0 era: A systematic review and

outlook", **Computers in Industry**, v. 111, p. 68–90, 2019. DOI: 10.1016/j.compind.2019.06.007.

MADHAN, E. S., KANNAN, K. S., RANI, P. S., *et al.* "A distributed submerged object detection and classification enhancement with deep learning", **Distributed and Parallel Databases**, n. 0123456789, 2021. DOI: 10.1007/s10619-021-07342-1.

MAHIR, L. H. A., VILAS BÔAS FÁVERO, C., KETJUNTIWA, T., *et al.* "Mechanism of Wax Deposition on Cold Surfaces: Gelation and Deposit Aging", **Energy and Fuels**, v. 33, n. 5, p. 3776–3786, 2019. DOI: 10.1021/acs.energyfuels.8b03139.

MARQUES, A. C. R. **Contribuição a Abordagem de Problemas de Classificação por Redes Convolucionais Profundas**. 2018. 121 f. Universidade Estadual de Campinas, 2018.

MELCHUNA, A., ZHANG, X., SA, J. H., *et al.* "Flow Risk Index: A New Metric for Solid Precipitation Assessment in Flow Assurance Management Applied to Gas Hydrate Transportability", **Energy and Fuels**, v. 34, n. 8, p. 9371–9378, 2020. DOI: 10.1021/acs.energyfuels.0c01203.

MOHAMMADPOOR, M., TORABI, F. "Big Data analytics in oil and gas industry: An emerging trend", **Petroleum**, v. 6, n. 4, p. 321–328, 2020. DOI: 10.1016/j.petlm.2018.11.001.

MUTLU, A. Y., KILIÇ, V., ÖZDEMİR, G. K., *et al.* "Smartphone-based colorimetric detection: Via machine learning", **Analyst**, v. 142, n. 13, p. 2434–2441, 2017. DOI: 10.1039/c7an00741h.

NASIR, Q., SULEMAN, H., ELSHEIKH, Y. A. "A review on the role and impact of various additives as promoters/ inhibitors for gas hydrate formation", **Journal of Natural Gas Science and Engineering**, v. 76, n. December 2019, p. 103211, 2020. DOI: 10.1016/j.jngse.2020.103211.

NGUYEN, D. A., FOGLER, H. S., CHAVADEJ, S. "Fused chemical reactions. 2. Encapsulation: Application to remediation of paraffin plugged pipelines", **Industrial and Engineering Chemistry Research**, v. 40, n. 23, p. 5058–5065, 2001. DOI: 10.1021/ie0009886. .

NGUYEN, D. A., IWANIW, M. A., FOGLER, H. S. "Kinetics and mechanism of the reaction between ammonium and nitrite ions: Experimental and theoretical studies", **Chemical Engineering Science**, v. 58, n. 19, p. 4351–4362, 2003. DOI: 10.1016/S0009-2509(03)00317-8.

NGUYEN, T. B., NGUYEN, T. H., CHUNG, W. Y. "Battery-free and noninvasive estimation of food ph and co2 concentration for food monitoring based on pressure measurement", **Sensors (Switzerland)**, v. 20, n. 20, p. 1–12, 2020. DOI: 10.3390/s20205853. .

NOGUEIRA, J. do N. P. **FAULT DETECTION AND DIAGNOSIS OF A TWO-COLUMN SOUR WATER TREATMENT UNIT BASED ON ARTIFICIAL INTELLIGENCE ALGORITHMS**. 2021. 162 f. Federal University of Rio de Janeiro (UFRJ), 2021.

OLAJIRE, A. A. "A review of oilfield scale management technology for oil and gas production", **Journal of Petroleum Science and Engineering**, v. 135, p. 723–737, 2015. DOI: 10.1016/j.petrol.2015.09.011.

OLAJIRE, A. A. "Flow assurance issues in deep-water gas well testing and mitigation strategies with respect to gas hydrates deposition in flowlines—A review", **Journal of Molecular Liquids**, v. 318, p. 114203, 2020. DOI: 10.1016/j.molliq.2020.114203.

PAZ, P. A., CAPRACE, J.-D., CAJAIBA, J. F., *et al.* "Prediction of Calcium Carbonate Scaling in Pipes Using Artificial Neural Networks". 2017. **Anais [...]** Trondheim, Norway, [s.n.], 2017. p. 1–10.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., *et al.* "Scikit-learn: Machine Learning in Python", **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PENG, Y., LIAO, M., DENG, H., *et al.* "CNN-SVM: A classification method for fruit fL image with the complex background", **IET Cyber-Physical Systems: Theory and Applications**, v. 5, n. 2, p. 181–185, 2020. DOI: 10.1049/iet-cps.2019.0069. .

PLUMB, A. P., ROWE, R. C., YORK, P., *et al.* "Optimisation of the predictive ability of artificial neural network (ANN) models: A comparison of three ANN programs and four classes of training algorithm", **European Journal of Pharmaceutical Sciences**, v. 25, n. 4–5, p. 395–405, 2005. DOI: 10.1016/j.ejps.2005.04.010. .

POERIO, D. V., BROWN, S. D. "Highly-overlapped, recursive partial least squares soft sensor with state partitioning via local variable selection", **Chemometrics and Intelligent Laboratory Systems**, v. 175, n. December 2017, p. 104–115, 2018. DOI: 10.1016/j.chemolab.2018.02.006.

PRIYAM, A., ABHIJEET, GUPTA, R., *et al.* "Comparative Analysis of Decision Tree Classification Algorithms", **International Journal of Current Engineering and Technology**, v. 3, p. 334–337, 2013.

QASIM, A., KHAN, M. S., LAL, B., *et al.* "A perspective on dual purpose gas hydrate and corrosion inhibitors for flow assurance", **Journal of Petroleum Science and Engineering**, v. 183, n. August, p. 106418, 2019. DOI: 10.1016/j.petrol.2019.106418.

RAHMANIFARD, H., PLAKSINA, T. "Application of artificial intelligence techniques in the petroleum industry: a review", **Artificial Intelligence Review**, v. 52, n. 4, p. 2295–2318, 2019. DOI: 10.1007/s10462-018-9612-8.

RAMZI, M., HOSNY, R., EL-SAYED, M., *et al.* "Evaluation of scale inhibitors performance under simulated flowing field conditions using dynamic tube blocking test", **International Journal of Chemical Sciences**, v. 14, n. 1, p. 16–28, 2016. .

RHYS, H. I. **Machine Learning with R, the tidyverse, and mlr**. [S.l.], Manning Publications, 2020.

SAFAVI, A., BAGHERI, M. "Novel optical pH sensor for high and low pH values", **Sensors and Actuators, B: Chemical**, v. 90, n. 1–3, p. 143–150, 2003. DOI:

10.1016/S0925-4005(03)00039-X. .

SAMARANAYAKE, C. P., SASTRY, S. K. "In-situ pH measurement of selected liquid foods under high pressure", **Innovative Food Science and Emerging Technologies**, v. 17, p. 22–26, 2013. DOI: 10.1016/j.ifset.2012.09.006. .

SANTOS, H. F. L., CASTRO, B. B., BLOCH, M., *et al.* "A physical model for scale growth during the dynamic tube blocking test", **OTC Brasil 2017**, n. Figure 1, p. 161–180, 2017. DOI: 10.4043/27956-ms. .

SCIKIT-LEARN. **1.4. Support Vector Machines**. [S.d.]. Disponível em: <https://scikit-learn.org/stable/modules/svm.html>. Acesso em: 10 jul. 2022.

SHANG, C., YANG, F., HUANG, D., *et al.* "Data-driven soft sensor development based on deep learning technique", **Journal of Process Control**, v. 24, n. 3, p. 223–233, 2014. DOI: 10.1016/j.jprocont.2014.01.012.

SHEN, S., LU, H., SADOUGHI, M., *et al.* "A physics-informed deep learning approach for bearing fault detection", **Engineering Applications of Artificial Intelligence**, v. 103, n. May, p. 104295, 2021. DOI: 10.1016/j.engappai.2021.104295.

SOLEIMANI, R., SHOUSHARI, N. A., MIRZA, B., *et al.* "Experimental investigation, modeling and optimization of membrane separation using artificial neural network and multi-objective optimization using genetic algorithm", **Chemical Engineering Research and Design**, v. 91, n. 5, p. 883–903, 2013. DOI: 10.1016/j.cherd.2012.08.004.

SUN, Q., GE, Z. "A Survey on Deep Learning for Data-driven Soft Sensors", **IEEE Transactions on Industrial Informatics**, v. 3203, n. c, 2021. DOI: 10.1109/TII.2021.3053128. .

TANGIRALA, S. "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm", **International Journal of Advanced Computer Science and Applications**, v. 11, n. 2, p. 612–619, 2020. DOI: 10.14569/ijacsa.2020.0110277.

THEYAB, M. A. "Fluid Flow Assurance Issues: Literature Review", **SciFed Journal of Petroleum**, v. 2, n. 1, p. 1–11, 2018.

VALIM, I. C., FIDALGO, J. L. G., REGO, A. S. C., *et al.* "Neural network modeling to support an experimental study of the delignification process of sugarcane bagasse after alkaline hydrogen peroxide pre-treatment", **Bioresource Technology**, v. 243, p. 760–770, 2017. DOI: 10.1016/j.biortech.2017.06.029.

WANASINGHE, T. R., WROBLEWSKI, L., PETERSEN, B. K., *et al.* "Digital Twin for the Oil and Gas Industry: Overview, Research Trends, Opportunities, and Challenges", **IEEE Access**, v. 8, p. 104175–104197, 2020. DOI: 10.1109/ACCESS.2020.2998723.

WU, H., HUANG, Q., WANG, D., *et al.* "A CNN-SVM combined model for pattern recognition of knee motion using mechanomyography signals", **Journal of Electromyography and Kinesiology**, v. 42, n. December 2017, p. 136–142, 2018. DOI: 10.1016/j.jelekin.2018.07.005.

XU, Z., WU, F., YANG, X., *et al.* "Measurement of gas-oil two-phase flow patterns by using CNN algorithm based on dual ECT sensors with venturi tube", **Sensors (Switzerland)**, v. 20, n. 4, 2020. DOI: 10.3390/s20041200.

YAN, W., TANG, D., LIN, Y. "A data-driven soft sensor modeling method based on deep learning and its application", **IEEE Transactions on Industrial Electronics**, v. 64, n. 5, p. 4237–4245, 2017. DOI: 10.1109/TIE.2016.2622668.

YAO, G., LEI, T., ZHONG, J. "A review of Convolutional-Neural-Network-based action recognition", **Pattern Recognition Letters**, v. 118, p. 14–22, 2019. DOI: 10.1016/j.patrec.2018.05.018.

YUAN, X., QI, S., SHARDT, Y., *et al.* "Soft sensor model for dynamic processes based on multichannel convolutional neural network", **Chemometrics and Intelligent Laboratory Systems**, v. 203, n. January, p. 104050, 2020. DOI: 10.1016/j.chemolab.2020.104050.

ZAN, T., LIU, Z., WANG, H., *et al.* "Control chart pattern recognition using the convolutional neural network", **Journal of Intelligent Manufacturing**, v. 31, n. 3, p. 703–716, 2020. DOI: 10.1007/s10845-019-01473-0.

ZAREI, F., BAGHBAN, A. "Phase behavior modelling of asphaltene precipitation utilizing MLP-ANN approach", **Petroleum Science and Technology**, v. 35, n. 20, p. 2009–2015, 2017. DOI: 10.1080/10916466.2017.1377233.

ZHENG, S., FOGLER, H. S., HAJI-AKBARI, A. "A Fundamental Wax Deposition Model for Water-in-Oil Dispersed Flows in Subsea Pipelines", **AIChE Journal**, v. 63, p. 4201–4213, 2017. DOI: 10.1002/aic.15750.

## **Development of Artificial Neural Networks models for the simulation of $\text{CaCO}_3$ scale formation process in the presence of monoethylene glycol (MEG) in Dynamic Tube Blocking Test equipment**

In the first part of this study, the target goal was to model the scale formation process caused by inorganic salts (scale). As presented in the previous section, they are formed by the deposition of salts in the pipeline wall caused by the variation of operation conditions (temperature, pressure, pH, and others) encountered during the production process. Monitoring this deposition process and being able to understand and predict how the other chemicals used during the process, such as inhibitors for other sources of incrustation, affects these dynamics is very important to the operation of the exploration and to avoid product losses.

In this scenario, the current section presents the development of models for the simulation of the scale formation in the presence of MEG, a hydrate inhibitor, using data from the Dynamic Tube Blocking Test (TBT). The results presented here were already published at Energy & Fuels (<https://doi.org/10.1021/acs.energyfuels.1c03364>), presented in Appendix A. The appendix published together with the main file of the article is presented in Appendix B.

This section contains the full version of the article (as found online), and the Supporting Information available with the published article is presented in Appendix C.

## 3.1

## Development of artificial neural network models for the simulation of $\text{CaCO}_3$ scale formation process in the presence of monoethylene glycol (MEG) in a dynamic tube blocking test (TBT) equipment

*AUTHOR NAMES: Bruno X. Ferreira<sup>1</sup>, Carlos R. Hall Barbosa<sup>3</sup>, João Cajaíba<sup>2</sup>, Vinicius Kartnaller<sup>2</sup>, Brunno F. Santos<sup>1\*</sup>.*

*\*Email corresponding author: bsantos@puc-rio.br*

### AUTHOR ADDRESS:

*1 Department of Chemical and Materials Engineering (DEQM), Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rua Marquês de São Vicente, 225 – Gávea, Rio de Janeiro, RJ 22430-060, Brazil.*

*2 Instituto de Química, Pólo de Xistoquímica, Universidade Federal do Rio de Janeiro (UFRJ), Rua Hélio de Almeida 40, Cidade Universitária, Rio de Janeiro 21941-614, Brazil.*

*3 Postgraduate Program in Metrology, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rua Marquês de São Vicente, 225 – Gávea, Rio de Janeiro, RJ 22430-060, Brazil.*

**KEYWORDS:** Artificial Neural Network, Calcium carbonate, Monoethylene glycol, Multi-Layer Perceptron, Scale, Fouling.

**ABSTRACT:** The precipitation of gas hydrate and inorganic salts (scale) during oil and gas production represents a significant flow assurance hindrance for the industry. Chemical inhibitors can prevent the fouling process, but specific inhibitors to address a problem could result in synergistic or adverse effects. Simulations in tubes and pipelines are necessary to understand these behaviors by assessing the scaling tendency of the water. The primary objective of this study was to create models using an Artificial Neural Network (ANN) of the Multi-Layer Perceptron (MLP) type for the simulation of the calcium carbonate scaling formation process



in the presence of monoethylene glycol (MEG), a typical gas hydrate inhibitor. A database was obtained from 38 Tube Blocking Test (TBT) experiments with different conditions. The models were developed using MATLAB R2020a, splitting the database into two groups on the ratio of 70:30, respectively, train and test ones, preserving the time-dependency of the differential pressure ( $\Delta P$ ) data. The ANNs were created using six inputs (temperature, pressure, calcium and bicarbonate concentration, MEG concentration, and the  $\Delta P$  measured at a selected time) and one output (the  $\Delta P$  measured at a later time). The goal was to explore how monitoring the conditions in a pipeline can predict the evolution of the scaling process. We investigated two scenarios for the  $\Delta P$  prediction: a near future (1 step ahead) and a far future (5 steps ahead). The MLP models demonstrated high performance, with an  $R^2$  higher than 92.9% for both training and test groups for both prediction horizons. Then the models were tested with a second data group to evaluate their applicability to control systems. The best models showed good scaling prediction, with  $R^2$  ranging from 80.0 to 99.9%. The results represent a promising step towards applying machine learning techniques to simulate and predict scaling tendencies in controlled pipelines.

### 3.1.1

#### Introduction

Flow assurance is a significant concern during oil and gas production and is achieved by guaranteeing that hydrocarbon production from wells is maintained without loss over time due to flow restrictions. During production, the oil–gas–water mixture undergoes drastic variations in operating conditions, such as temperature and pressure, so the solubility of certain compounds can decrease, leading to the formation of deposits (fouling). Fouling may occur in pipelines and equipment and is generally caused by the formation of wax, gas hydrate, and scale (inorganic salts). This scenario can require expensive and complex remediation processes and, in severe cases, production stoppage and well shutdown (SOUZA *et al.*, 2019; KHORMALI *et al.*, 2018; NGUYEN *et al.*, 2003). This problem is of great concern, especially for wells in the Brazilian pre-salt region located in ultradeep waters with mainly carbonaceous reservoir rocks, and can result in potential issues such as calcium carbonate and gas hydrates fouling (DE OLIVEIRA and GONÇALVES, 2012).

Gas hydrate originates from the crystallization of water molecules encapsulating small and light gas molecules (e.g., CO<sub>2</sub>, methane, and propane) under operating conditions with high pressure and low temperature, such as those found in deep and ultradeep water (KIM *et al.*, 2020; NASIR *et al.*, 2020). The most practical and economical method for preventing hydrate formation or others kinds of obstructions in lines (e.g., scales) is using chemical inhibitors (KUMAR *et al.*, 2018; DE ROSA *et al.*, 2019; AHMED *et al.*, 2020). Thermodynamic Hydrate Inhibitors (THIs) are typically injected into the production line to prevent the formation of gas hydrates. THIs consist of alcohols or glycols, such as methanol, triethylene glycol (TEG), and monoethylene glycol (MEG), and function by moving the equilibrium curve envelope towards lower temperature and higher pressure (KAN *et al.*, 2002; LIM *et al.*, 2002).

Scale forms as a result of the deposition of inorganic salts precipitating from the supersaturated water. Their formation depends on several factors such as temperature, pressure, ion concentration, pH, and others (OLAJIRE, 2015). Barium sulfate, strontium sulfate, and calcium carbonate are the most common types of scale found during oil and gas production (DYER and GRAHAM, 2002; ODDO and TOMSON, 1994). However, calcium carbonate (CaCO<sub>3</sub>) formation is of greater concern since the water may be in equilibrium with carbonaceous rocks in the reservoir, leading to a significant number of bicarbonate ions dissolved in the water phase (Eq. B1–B3, in the Appendix B). The precipitation of CaCO<sub>3</sub> occurs as this fluid is produced and faces a pressure drop, which decreases the CO<sub>2</sub> solubility and increases pH, leading to precipitation (Eq. B4, in the Appendix B).

There are dozens of different inhibitor types used for typical inorganic scale. There are three main classes of inhibitors: phosphate esters, phosphonates, and polymers. The first two classes act as chelators, sequestering the metals from solution, while the polymeric class achieves scale control through crystal distortion. In 2002, the average cost due to scale formation was more than 1.4 billion dollars (FRENIER and ZIAUDDIN, 2008). As a result, the market for scale inhibitors for the oil and gas industry continues to grow and currently represents millions of dollars annually. Market analyses predict further increases in the expenditures with a CAGRs (Compound Annual Growth Rates) of 5.5% and 6.9% for the scale and hydrate inhibitors markets, respectively (Global Oilfield Scale Inhibitor Market, 2021; Hydrate Inhibitors Market Analysis, 2021; Oilfield Scale Inhibitor Market,

2021).

A concern with the use of inhibitors for production is the compatibility between the different inhibitors and other chemicals. Several studies have investigated these compatibilities, including the effects of the Enhanced Oil Recovery (EOR) chemicals on scale inhibitors (WANG *et al.*, 2018) and the interaction between scale inhibitors and hydrate inhibitors (CHAO *et al.*, 2020). For example, Seiersten and Kundu (2018) and Kartnaller *et al.* (2018) studied the impact of MEG as a gas hydrate scale inhibitor, concluding that MEG serves as an inhibitor by increasing the scaling time. This result was unexpected because the presence of MEG in water increases ion activities. That behavior has been proposed to be connected to the high-energy bond between -OH groups and the  $\text{CaCO}_3$  surface; this indicates that thermodynamic hydrate inhibitors can also benefit wells experiencing calcium carbonate scale formation.

Understanding the interactions between inhibitors, water, and ions is essential for predicting the phase behavior during production and estimating the solid accumulation tendency in production lines. A common and well-known methodology to evaluate inhibitor efficiency is the Dynamic Tube Blocking Test (TBT). It is usually applied to verify a product's performance and Minimum Inhibitor Concentration (MIC), allowing comparison with other commercially available products (RAMZI *et al.*, 2016; MACEDO *et al.*, 2019; FERNANDES *et al.*, 2021). TBT experiments are also used to study inorganic salt morphologies (DE MORAIS *et al.*, 2020; SANNI *et al.*, 2019) and develop scale formation models. However, it is difficult to predict how the scaling process will develop using flow and phase behavior models due to the system's complexity, the large number of variables, and some stochastic behavior. A previous work has attempted to model the scale formation in pipelines, specifically in TBT experiments, but only using physical models (SANTOS *et al.*, 2017). These models, based on Darcy Weisbach equation for pressure loss in pipes and on a growth rate scale formation model, were successful in fitting the TBT experiments curves, enabling an estimation on how fast the process was happening. However, the model was learning only the information regarding that specific experiment and not acquiring information for predicting the behavior of the system.

Other studies have explored the use of Artificial Neural Networks (ANNs) and other machine learning algorithms to create new models since they do not

demand an understanding of the scale formation mechanism, only requiring a “black-box” model. These models were able to predict the thermodynamics related to the calcium carbonate precipitation (saturation ratio of the solution) and its dissolution capacity (PAZ *et al.*, 2017; AHMADI *et al.*, 2015). However, literature still lacks kinetic modeling related to the scale formation process. Recently, Wang *et al.* (2019) have developed an Elman Neural Network (ENN) with genetic algorithm (GA) to predict calcium carbonate scale formation in shell and tube heat exchangers over time. They were able to successfully predict the fouling resistance as a function of conductivity, pH and dissolved oxygen. Still, as far as the author’s knowledge, no study relating scale formation and variables to simulate conditions during oil and gas production has been previously assessed.

In recent decades, different types of Artificial Intelligence (AI), such as ANN, GA, Support Vector Machines (SVMs), Adaptive Neuro-Fuzzy Inference System (ANFIS), Least Square Support Vector Machine (LSSVM), Principal Component Analysis (PCA), Committee Machine Intelligent System (CMIS) have been applied to solve problems and challenges in several fields like nanofluids properties (BAGHBAN *et al.*, 2018a; BAGHBAN *et al.*, 2018b; BAGHBAN *et al.*, 2019), systems efficiency (AHMADI *et al.*, 2020; ZAMEN *et al.*, 2019) and in the oil and gas industry, from the reservoir to production (ALKINANI *et al.*, 2019; OTCHERE *et al.*, 2021; RAHMANIFARD and PLAKSINA, 2019). The ANN was inspired by the neural arrangement of the human brain. It is easy to train and has tunable parameters and an adaptive structure, making it one of the most widely used machine learning techniques (LI *et al.*, 2017). One of the most common classes of the ANN is the Feedforward Neural Network (FFNN) with MLP (Multi-Layer Perceptron) topologies, which can model complex systems (HEIDARI *et al.*, 2020). The usual structure of the MLP consists of an input layer, where the number of neurons is equal to the number of model inputs, and an output layer. In addition, there is at least one hidden layer between them with several neurons to be selected by the user (LI *et al.*, 2021; HAMMOUDI *et al.*, 2019). This structure has been used to predict different parameters for the oil and gas industry, such as the gas-oil ratio (SEFIDI and AJORKARAN, 2019), the volume fraction percentage in three-phase systems (ISLAMI RAD and PEYVANDI, 2019), and the deposition process of asphaltene (ZAREI *et al.*, 2017) and wax (AMAR *et al.*, 2021).

Knowing the importance of digital transformation, AI, and process

monitoring in the oil and gas industries, this work intends to model the scale formation process using the MLP to predict the differential pressure ( $\Delta P$ ) one and five steps ahead in time. The goal is to explore how monitoring the conditions in a pipeline (i.e., temperature, pressure, ion concentrations, and differential pressure) can predict the evolution of the scaling process. This study may lead to deeper investigations into applications in monitoring systems and fault detection. TBT differential pressures were monitored over time for different temperatures, pressures, calcium and bicarbonate concentrations, and MEG concentrations. MEG concentration was used as a variable since many scale inhibitor products are solutions of the active molecule in a mixture of water and MEG. Also, MEG can be directly injected in high amounts as thermodynamic gas hydrate inhibitors. Even further, MEG can change the viscosity of the solution and can influence the crystallization of calcium carbonate, which would lead to different effects to be modeled in order to best simulate the scale formation process. Two scenarios were considered: a near future time (differential pressure measured 1 step ahead) and a far future time (differential pressure measured 5 steps ahead). The models showed good scaling prediction for both time horizons, showing a promising step towards simulating and predicting scaling tendencies in controlled pipes in production lines.

### **3.1.2**

#### **Methodology**

### **3.2.1**

#### **Experimental details**

Experiments were performed in a TBT equipment, in which two solutions containing incompatible cations and anions are pumped into tubes inside an oven, conditioned to the test temperature, mixed in a micro-chamber, and then flown into a capillary tube called loop test. The apparatus consisted of two high performance liquid chromatography (HPLC) pumps pushing newly prepared calcium chloride and sodium bicarbonate solutions, with pH ranging from 7.0-7.5 depending on the salts concentration, into a thermostat-regulated oven through 1.8 m long stainless-steel tubes with 1 mm inner diameters (i.e., two conditioning loops, one for each solution).

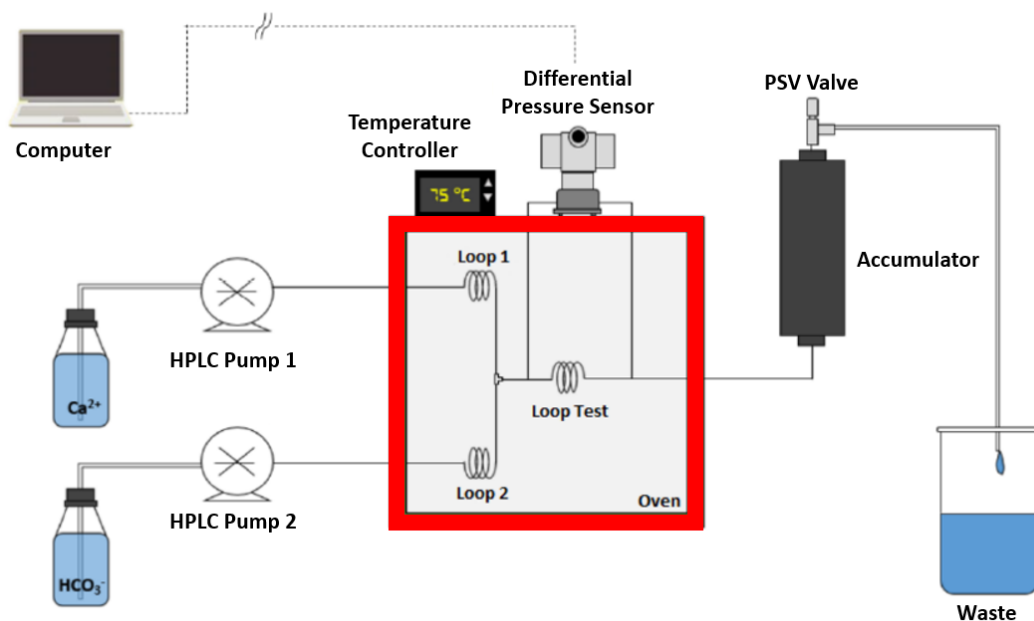


Figure 3.1: Scheme of a Dynamic Scale Loop (DSL) system used in a TBT experiment (adapted from KARTNALLER *et al.*, 2018).

These loops ensured the solutions reached the mixture chamber at the correct temperature for the experiments. After mixing, the combined solution flowed through a third tube (loop test) with the same dimensions as the other tubes. This process resulted in a supersaturated solution leading to calcium carbonate formation and deposition. When deposition occurred, the inlet pressure became higher than the outlet pressure, generating a differential pressure. This differential pressure was measured using a model EJA 130A high-static differential pressure transmitter (Yokogawa, Musashino, Tokyo, Japan). The data were acquired at 1 s intervals using a LabView-based software program. The injection flow rate was  $10.0 \text{ mL min}^{-1}$  ( $5.00 \text{ mL min}^{-1}$  for each solution, leading to a 1:1 mixture ratio of the two solutions). The pressure of the system was regulated using a PSV valve connected outside the oven.

### 3.1.2.2

#### ANN database preparation

The experimental data used in this study are the results from 38 TBT experiments previously presented in Kartnaller *et al.* (2018), which used a modeling approach with experiments from a central composite design of experiment and multivariate linear regression (MLR). In the previous work, MLR was applied to

model the scaling time to reach several differential pressure levels (1 to 25 psi, in intervals of 1 psi). For each pressure, a different model had to be made, totaling 25 different models to predict a single scaling tendency. These experiments varied the pressure, temperature, concentration of MEG ( $C_{\text{MEG}}$ ) (v/v %), and concentration of the carbonate ( $\text{CHCO}_3^-$ ) (ppm) and calcium ( $\text{C}_{\text{Ca}^{2+}}$ ) (ppm) ions over the operating ranges shown in Table 3.1. The experiments measure the  $\Delta P$  every second as the monitored variable.

Table 3.1: Range of the experimental variables.

Variable	Unit	Minimum value	Maximum value
Pressure	bar	0	170
Temperature	°C	40	110
$C_{\text{MEG}}$	v/v %	0	80
$\text{C}_{\text{Ca}^{2+}}$	ppm	1000	6000
$\text{CHCO}_3^-$	ppm	1000	6000

The goal for the ANN modeling in the present work was to improve the prediction of the scale formation process, in which the differential pressure was also an input for the modeling. The measurement of the differential pressure at a moment in time, plus the experimental variables, was used to estimate the differential pressure in a later time. Hence, experimental data were first preprocessed to adjust the signal baseline and create the differential pressure variables one step ahead, ( $\Delta P_{(t+1)}$ ) and five steps ahead ( $\Delta P_{(t+5)}$ ) (one second and five seconds ahead respectively) to be used in the prediction models. The database was then split into two parts. The first database consisted of 32 experiments, totaling 46,698 data points. This database was separated into two groups, train (70%) and test (30%), and was used to train the MLP models. To preserve the time information about the scale formation associated with the pressure differential, this division was accomplished by selecting 7 data points for the train group and 3 for the test group from every 10 data points.

The second database consisted of 6 experiments, totaling 7,705 data points. Those experiments were conducted with fixed values of pressure, temperature,  $\text{CHCO}_3^-$  and  $\text{C}_{\text{Ca}^{2+}}$  in their central values of the design of experiments, and varying

$C_{MEG}$  (10, 20, 30, 50, 60, and 70 v/v %). This database was used to separately validate the models constructed by the ANN for each experiment.

### 3.1.2.3

#### Artificial Neural Network optimization

For this study, MLP type ANN models with one output neuron were developed using Matlab R2020a (developed by Mathworks, Inc) to predict  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$ . Six input variables were chosen for the input layer: the five independent variables (pressure, temperature,  $CHCO_3^-$  and  $CCa^{2+}$ , and  $C_{MEG}$ ) and the differential pressure at the selected time  $t$  ( $\Delta P_{(t)}$ ). The proposed MLP structure had one hidden layer, in which the number of neurons is one of the hyperparameters to be optimized. The search was started with the same number of neurons as the input layer.

The activation function, applied to the connection between the input and hidden layers, was the second hyperparameter studied, and the hyperbolic tangent (*tansig*) and log sigmoid (*logsig*) functions were used. Both functions are commonly used due to their sigmoidal form. The linear activation function (*purelin*) was used between the hidden layer and the output layer (CHOJACZYK *et al.*, 2015; SOLEIMANI *et. al.*, 2013; HAYKIN, 2001).

The last hyperparameter optimized for the MLP models was the training algorithm. The Gradient Descent with Momentum and Adaptive Learning Rate Backpropagation (*traingdx*), Levenberg-Marquardt Backpropagation (*trainlm*), and Bayesian Regularization Backpropagation (*trainbr*) functions were selected for testing. The first of these algorithms improves upon traditional backpropagation with a combination of an adaptive learning rate and momentum training, while the others apply a quasi-Newton method for faster convergence (MATHWORKS, 2020 a; MATHWORKS, 2020b; MATHWORKS, 2020c).

### 3.1.2.4

#### Statistical performance evaluation

To evaluate the performance of the ANN models, the coefficient of determination parameter ( $R^2$ , Eq. B5), Sum of Squared Errors (SSE, Eq. B1), Mean Squared Error (MSE, Eq. B2), and Root Mean Squared Error (RMSE, Eq. B3), were



chosen. For  $R^2$ , the goal was to achieve a value close to one, while the goal for the others was to achieve the lowest value possible, indicating the best fit between the experimental data and the predicted data from the ANN models. To calculate  $R^2$ , it is also necessary to calculate the Total Sum of Squares (TSS, Eq. B4). The equations are available in the Appendix B.

Figure 3.2 shows a schematic for the process adopted in this study, from the data acquisition on the experiments to the determination of the best MLP model.

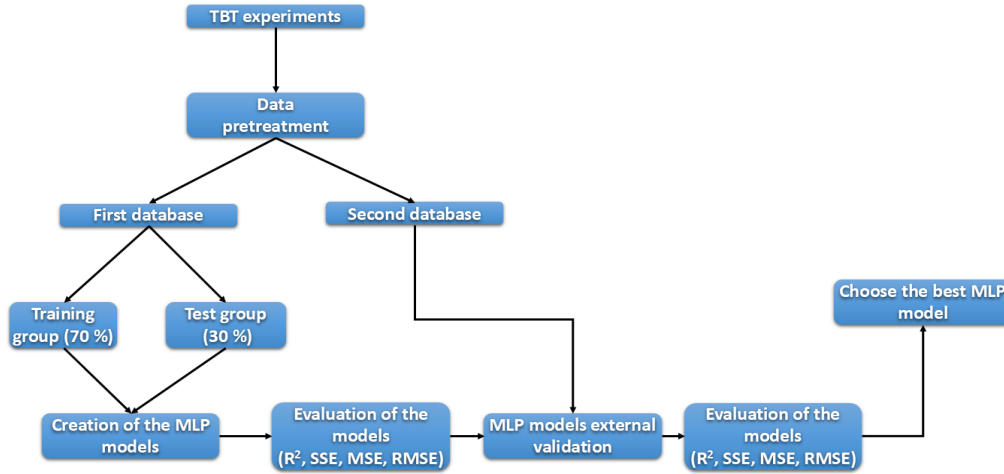


Figure 3.2: Flowchart of the methodology.

### 3.1.2.5

#### Sensitivity analysis

The “black-box” group of models, in which the ANN models are often included, present some difficult to extract information about the process from their parameters. However, the evaluation of the input variables effects over the output variable can be determined by a sensitivity analysis.

For that, in this study two approaches were explored. First, it was used the relevancy factor ( $r$ , Eq. 3.1), which can be applied to quantify these effects, with values on the range from -1 to +1. The highest absolute value of  $r$  indicates the variables that most affect the target variable, in which the positive values indicate an elevation on the output variable whereas the negative ones designate a decrease on the target variable (BAGHBAN, 2019b; AHMADI *et. al.*, 2020b).

$$r = \frac{\sum_{i=1}^N (X_{k,i} - \bar{X}_k)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (X_{k,i} - \bar{X}_k)^2 \sum_{i=1}^N (y_i - \bar{y})^2}} \quad (3.1)$$

where  $N$  is the total number of data points,  $X_{k,i}$  is the  $i$ th input value of the  $k$ th parameter,  $y_i$  is  $i$ th output value,  $\overline{X_k}$  is the average value of the  $k$ th input parameter and  $\bar{y}$  is the mean value of the output parameter.

The second parameter adopted was the Relative Importance (RI), in which the methodology proposed by Garson (1991) (Eq. 3.2) was chosen to obtain the RI values, varying between 0 and 1, which is based on the connection weights between the ANN layers (DE ONA and GARRIDO, 2014; XU *et al.*, 2013; PENTÓS, 2016).

$$RI_{ij} = \frac{\sum_{j=1}^P \frac{|w_{ij}| \cdot |w_{jk}|}{\sum_{i=1}^N |w_{ij}|}}{\sum_{i=1}^N \sum_{j=1}^P \frac{|w_{ij}| \cdot |w_{jk}|}{\sum_{i=1}^N |w_{ij}|}} \quad (3.2)$$

where  $RI_{ij}$  is the parameters RI of the variable  $x_i$  concerning the output neuron  $j$ ,  $w_{ij}$  is the weight parameter of the connection between the input  $x_i$  and the  $j$ th hidden neuron,  $w_{jk}$  is the weight parameter of the connection between the  $j$ th hidden neuron and the  $k$ th output variable.

### 3.1.3

#### Results and Discussion

The data were selected, processed, and separated into two groups for training and testing to optimize the ANN model. The training data were used to construct the model and calculate the estimated parameters. Once the model was constructed, it was applied to the testing data to predict the output and compare it to the known values. Different types of models were tested by changing the hyperparameters of the ANN and were compared to indicate the best ones.

#### 3.1.3.1

##### Evaluation of the ANN models

MLP topologies developed to predict  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$  are shown in Table B1 in the Appendix B, along with the optimized hyperparameters of the trained models and the performance parameters from the train and test groups, for models having 6-8 neurons in the hidden layer. These results show that the best performance for the  $\Delta P_{(t+1)}$  was achieved with seven neurons in the hidden layer using the *tansig*

activation function and the *trainlm* training algorithm. This topology had an  $R^2$  equal to 99.88% for the test set and the lowest values for error. However, only three trained topologies had an  $R^2$  lower than 99%, showing that the models have very similar accuracy.

For the topologies built to predict  $\Delta P_{(t+5)}$ , the model with the best results had the same hidden layer configuration as the best model for  $\Delta P_{(t+1)}$  but used the *trainbr* as the training algorithm. Its performance had an  $R^2$  equal to 98.93% and the lowest values for the other error parameters as well. However, as observed in the predictions for the  $\Delta P_{(t+1)}$  case, most of the models had very similar figures of merit, indicating that the accuracy was largely independent of the activation function and training algorithm used (*trainlm* and *trainbr*). It is also interesting to point out that the worst results, in both cases, were obtained when using the *traingdx* training algorithm.

This investigation optimizing the hyperparameters of the MLP model for each output, primarily the number of neurons and the transfer function on the hidden layer, is an important step toward achieving the best models. Another essential phase in the model development is to validate them with new experimental data, verifying the model's prediction capability before using it in real applications.

### 3.1.3.2

#### Validation of the MLP models

Since the MLP models demonstrated similar accuracy for both time horizons, all were used in this validation phase. This evaluation used the second database in which the MEG concentration was changed from 10% to 70%, while all other variables were unchanged. This series of experiments tested the behavior of the scaling process in the presence of the glycol molecule. In a previously published article (KARTNALLER *et al.*, 2018), our research group has shown that MEG can act as a calcium carbonate inhibitor at concentrations above 30%.

The correct mechanism to explain how MEG acts in the calcium carbonate crystallization is still not completely known. The interaction of alcohols (and therefore polyols) have been studied by several works in the past years, and simulations have shown that the -OH group can bind to specific faces of the calcite polymorph, which can lead to control of the crystal growth (SAND *et al.*, 2010; BOVET *et al.*, 2015; ZHANG *et al.*, 2008). Okhrimenko *et al.* (2013) showed that

this adsorption could also happen for aragonite and vaterite (other calcium carbonate polymorphs), although the binding energy in these cases is lower than for calcite. This adsorption comes from the fact that the Ca–CO<sub>3</sub> ion pair (note that this is just a representation of pairs, not chemical bond) delocalizes charges by ordering the –OH group of the organic molecules. Thus, the O of this group is associated with Ca, while the H is associated with CO<sub>3</sub> (Okhrimenko *et al.*, 2013). This causes a highly organized monolayer structure to form on the surface of the crystal, in which the hydrophobic part of the chains face away from the surface. Many other types of organic molecules have also been studied on the calcium carbonate crystallization, specifically related to biomineralization.

Biomineralization is the process in which living organisms produce hard minerals that act as support, protection or nourishment structures. A wide variety of minerals can be synthesized by these organisms, such as silica, calcium phosphate and calcium carbonate. The calcite polymorph synthesized in pure solution in a laboratory has a large crystalline difference from that synthesized by mineralization (YANG *et al.*, 2008). This control of crystal growth is generally attributed to complex organic molecules known as coccolith-associated polysaccharides (CAPs). These are large polymeric carbohydrate molecules containing a variety of functional groups, such as –COOH and –OH. Hence, since MEG contains 3 hydroxyl groups in its structure, it is possible to suppose an association that there is an interaction of this molecule with the surface of the particles being formed, controlling crystal growth, which would also explain how it controls inhibition. Also, changing its concentration changes the viscosity of the solution (affecting the flow dynamics inside the tube).

The performance parameters for all MLP models for each new experiment are presented in the Appendix C in Tables C1-6. The models are validated by observing how they predict the scaling process under conditions different from the training or testing. Although the models showed very high accuracy for both training and test sets, their application to the new data was not completely successful. Some of the models' prediction of the scaling process over time was unsatisfactory for a few experiments, which showed that certain regions in the modeled response did not fit the actual expected experimental values. For the  $\Delta P_{(t+1)}$  scenario, the logsig\_7\_purelin\_1\_trainbr model (values of the weights and bias are available in the Appendix C, Table C7) was the best with an  $R^2$  over 99.3% for all

new experiments. Figure 3.3 shows the predicted differential pressure from this MLP model and the experimental data for all six experiments. In addition, four other topologies had an  $R^2$  higher than 97% showing that they are also very accurate models.

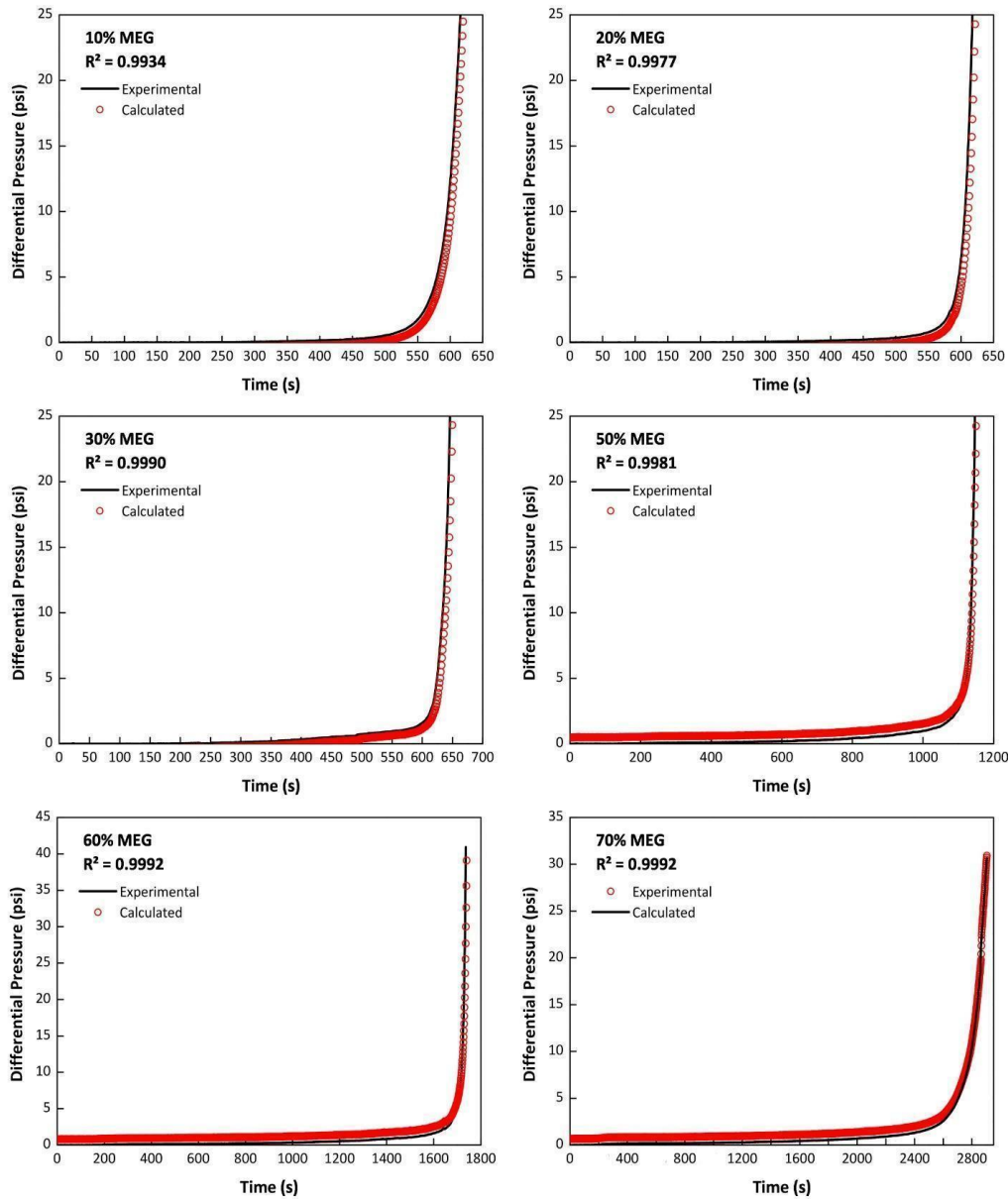


Figure 3.3: Representation of the behavior of the experimental data of the six experiments of the second database and the respective predicted data for the output  $\Delta P_{(t+1)}$  by the MLP model `logsig_7_purelin_1_trainbr`.

The lack of fit of parts of the predicted region was mainly observed for the  $\Delta P_{(t+5)}$  case. For example, the best model for this case could not predict the scaling tendency for MEG concentrations between 20 – 50%. For some of the experiments,

the  $R^2$  of the fit was actually negative, indicating that the scaling process was not being accurately modeled (or that the residues of the regression in that region did not follow a normal distribution with a mean equal to zero).

While most models did not present a good prediction performance for the new experiments, some were still very accurate. For the  $\Delta P_{(t+5)}$  time horizon, the `logsig_6_purelin_1_trainlm` model (values of the weights and bias are available in the Appendix C, Table C8) was the most accurate, with an  $R^2$  ranging from 79.7% to 96.4%. Figure 3.4 shows the predicted differential pressure from this MLP model and the experimental data for all six experiments. These results are important because they show that even though accurate predictions can be made for some regions of the studied response, continuous validation of the best models is necessary as new data is obtained.

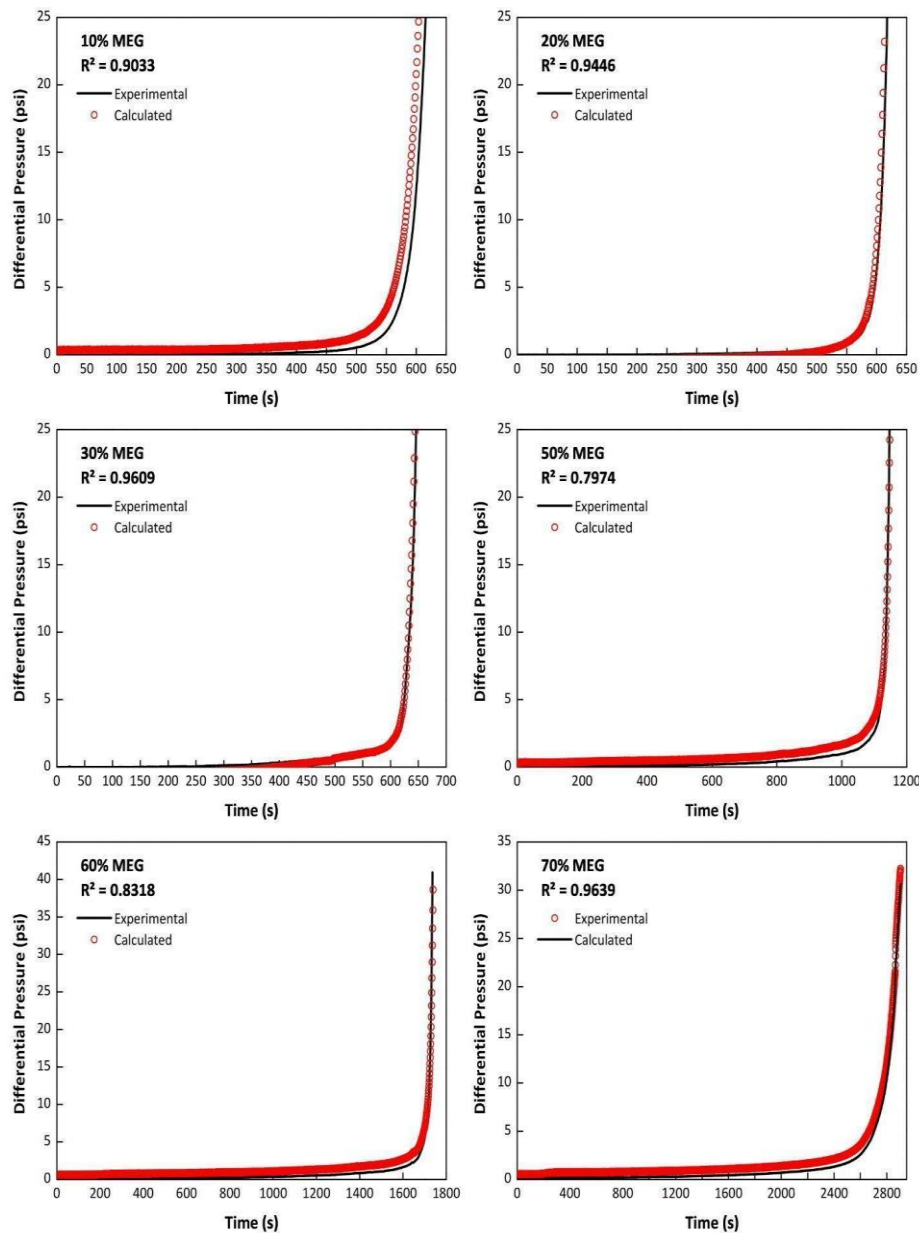


Figure 3.4. Representation of the behavior of the experimental data of the six experiments of the second database and the respective predicted data for the output  $\Delta P_{(t+5)}$  by the MLP model `logsig_6_purelin_1_trainlm`.

For the best models chosen for each output variable,  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$ , a deeper evaluation was performed, starting for a comparison between the experimental and predicted values for the training and test datasets, shown on Figure S1A-B respectively for the variables  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$ . These results also show that the model chosen to predict the  $\Delta P_{(t+1)}$  has the best prediction power.

Another investigation adopted was to evaluate the behavior of the normalized residuals according to the  $\Delta P$  values, comparing the response for the both output

variables  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$  for the training and test datasets, respectively, Figure S2A-B. From that could be extract that the MLP model for the  $\Delta P_{(t+5)}$  variable has a tendency to predict higher values than the experimental measures, what is worsen in higher values of  $\Delta P$ . However, it is important to highlight that the amount of data points with absolute normalized residuals higher than 0.1 is less than 1 % for the analyzed datasets for both output variables.

### 3.1.3.3

#### Sensitivity analysis

For the sensitivity analysis, the best models for each output variable,  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$ , were chosen, which had the topologies `logsig_7_purelin_1_trainbr` and `logsig_6_purelin_1_trainlm`. The first sensitivity evaluation was made for the relevancy factor ( $r$ ), Figures 3.5A-B show the values of  $r$  of each input variable for both target variables, respectively,  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$ . They indicate that the  $\Delta P_{(t)}$  are by far the most influential parameter for the two prediction horizons with a  $r$  close to 1, indicating expected strong correlation between the measure of the  $\Delta P$  and its prediction for future horizons.

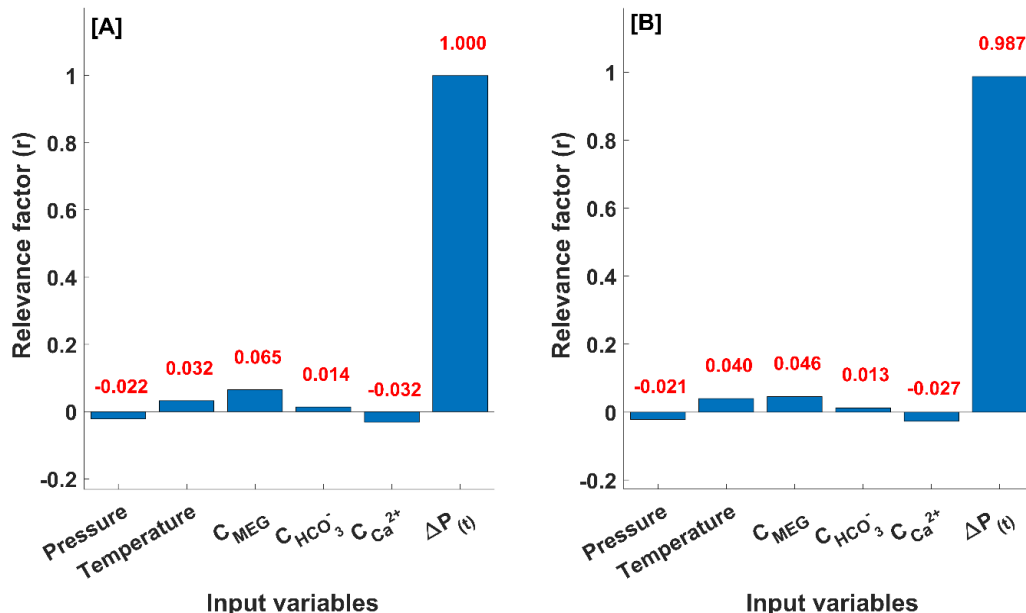


Figure 3.5. Relevancy factor of both output variables  $\Delta P_{(t+1)}$  (A) and  $\Delta P_{(t+5)}$  (B).

Then, these MLP models were analyzed for the Relative Importance (RI) parameter, which the values are presented on Figures 3.6A-B for the output variables  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$  respectively. For the best  $\Delta P_{(t+1)}$  model, the inputs



pressure, temperature,  $C_{\text{MEG}}$  and  $C_{\text{HCO}_3^-}$  presented an RI varying between 14 % and 19 %, and the input variable  $C_{\text{Ca}^{2+}}$  was the most relevant one for the  $\Delta P_{(t+1)}$  prediction. In turn, the input with less impact was the  $\Delta P_{(t)}$ .

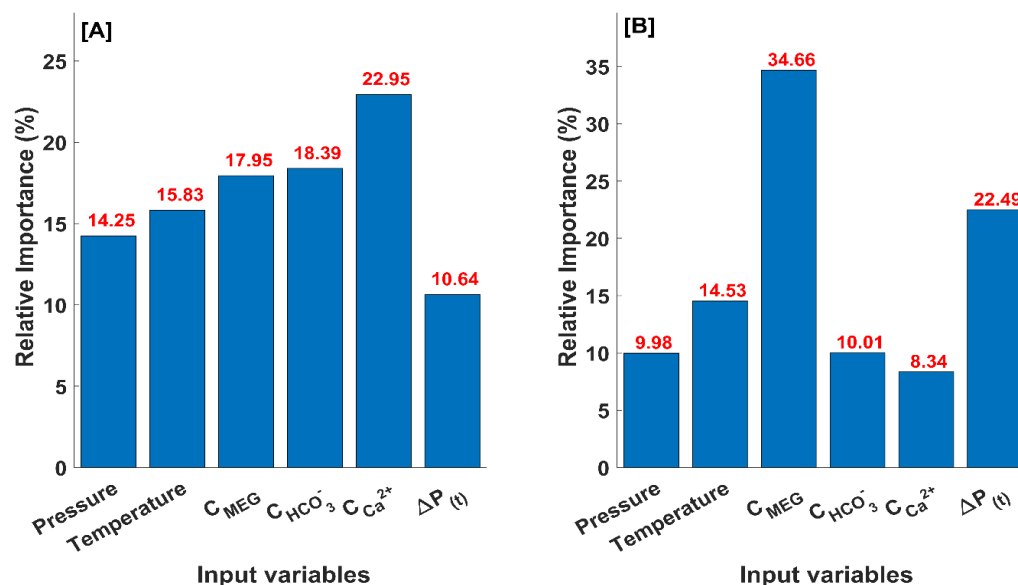


Figure 3.6. Relative Importance (RI) of both output variables  $\Delta P_{(t+1)}$ (A) and  $\Delta P_{(t+5)}$ (B) calculated by the Garson method (GARSON, 1991).

Conversely, for the best  $\Delta P_{(t+5)}$  model the most significant variables were the  $C_{\text{MEG}}$  followed for the  $\Delta P_{(t)}$ , respectively with the values of 34.7 % and 22.5 %, while the other inputs variables presented RI values lower than 15 %. This difference on the influence hierarchy of the input variables is interesting, since it shows an increase on the importance of the  $\Delta P_{(t)}$  for the prediction of the future. Also, for the  $\Delta P_{(t+5)}$  model, the high RI value of the variable  $C_{\text{MEG}}$  indicates a reason for this MLP model presenting the best performance against the validation data group. This may indicate a strong implication that MEG has in impacting the development of the scale formation process due to its inhibitor effect.

The two analyzed parameters,  $r$  and RI, led to different levels of influence for each input in the target variables. While the parameter  $r$  indicates the effect of the input values on the target variable, the RI parameter shows how the model attributes the importance for these inputs. Although, the  $\Delta P_{(t)}$  variable has a huge absolute value for the parameter  $r$ , a model that only uses this variable as input probably could predict the tendency of the  $\Delta P$  curve but it would not be able to distinguish between the different scenarios. That way, the combination of these

results indicates that maybe a hybrid model could be a better approach for this problem, applying the MLP to lead with the  $\Delta P$  curve behavior and another kind of model to handle the environment conditions information. However, this premise is outside of the scope of this work.

Finally, the modeling results indicated that ANN could be applied to predict the differential pressure and to understand the evolution of the scaling process at earlier as well as later times. For process monitoring, this appears to be a promising tool for transforming digital data acquired during production to establish the scaling tendency of a well over time, by relating the scale formation process with operational variables as a start to develop a model that could simulate the conditions during oil and gas production.

### 3.1.4

#### Conclusions

This study showed that using an MLP-type ANN enabled the modeling of the scaling process in a tube with a dynamic flow containing precipitated calcium carbonate. Even though the scaling process is a very complex system with stochastic behavior, this machine learning technique permitted its prediction over different time horizons: a “near future”, or one step ahead ( $\Delta P_{(t+1)}$ ), and a “far future”, or five steps ahead ( $\Delta P_{(t+5)}$ ). The generated models were highly accurate for both training and test data sets and for both time horizons, regardless of the activation function and the training algorithm used (*trainlm* and *trainbr*). However, using *traingdx* as a training algorithm gave poorer results. When using the models to predict a different series of experiments that simulated various viscosities with calcium carbonate inhibition, most models did not show the same initial high accuracy. In fact, only a few models were very accurate for all the experiments. Overall, for the  $\Delta P_{(t+1)}$  time horizon, the *logsig\_7\_purelin\_1\_trainbr* was the best model, with an  $R^2$  over 99.3% for the additional experiments. The *logsig\_6\_purelin\_1\_trainlm* model was the best model for the  $\Delta P_{(t+5)}$  time horizon, with an  $R^2$  ranging from 79.7% to 96.4%. These results show that ANN can predict the differential pressure in a tube to understand the evolution of the scaling process in the near time as well as its development in the future. This strategy represents an important application of digital transformation to oil and gas production to establish the scaling tendency during the lifetime of a well based on differential pressure

process monitoring.

## ASSOCIATED CONTENT

### Supporting Information

The following files are available free of charge in the Appendix B and Appendix C: The equilibrium equations of the calcium carbonate scale formation (Equations B1-4), Regression plot between experimental versus the predicted values (Figure B1A-B), Comparison between normalized residuals of the prediction of the  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$  variables for the training and test datasets (Figure B2A-B), Performance values for all topologies for the validation experiments (Tables C1-6) and Optimized parameters of the best MLP topologies (Tables C7-8). (PDF)

## AUTHOR INFORMATION

### Author Contributions

All authors have contributed to the writing of this manuscript and have approved its final version.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENT

The authors acknowledge the financial support by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Petrobras in the development of this work. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

### 3.5

#### References

- AHMADI, M. A., BAHADORI, A., SHADIZADEH, S. R. "A rigorous model to predict the amount of dissolved calcium carbonate concentration throughout oil field brines: Side effect of pressure and temperature", **Fuel**, v. 139, n. 1, p. 154–159, 2015. DOI: 10.1016/j.fuel.2014.08.044.
- AHMADI, M. H., BAGHBAN, A., GHAZVINI, M., *et al.* "An insight into the prediction of TiO<sub>2</sub>/water nanofluid viscosity through intelligence schemes", **Journal of Thermal Analysis and Calorimetry**, v. 139, n. 3, p. 2381–2394, 2020. DOI: 10.1007/s10973-019-08636-4.
- AHMADI, M. H., BAGHBAN, A., SADEGHZADEH, M., *et al.* "Evaluation of electrical efficiency of photovoltaic thermal solar collector", **Engineering Applications of Computational Fluid Mechanics**, v. 14, n. 1, p. 545–565, 2020. DOI: 10.1080/19942060.2020.1734094.
- AHMED, M., HUSSEIN, I. A., ONAWOLE, A. T., *et al.* "Dissolution kinetics of different inorganic oilfield scales in green formulations", **ACS Omega**, v. 5, n. 46, p. 29963–29970, 2020. DOI: 10.1021/acsomega.0c04357.
- ALKINANI, H. H., AL-HAMEEDI, A. T. T., DUNN-NORMAN, S., *et al.* "Applications of artificial neural networks in the petroleum industry: A review". 2019-March, 2019. **Anais [...]** [S.l: s.n.], 2019. DOI: 10.2118/195072-ms.
- BAGHBAN, A., JALALI, A., SHAFIEE, M., *et al.* "Developing an ANFIS-based swarm concept model for estimating the relative viscosity of nanofluids", **Engineering Applications of Computational Fluid Mechanics**, v. 13, n. 1, p. 26–39, 2019. DOI: 10.1080/19942060.2018.1542345.
- BAGHBAN, A., KAHANI, M., NAZARI, M. A., *et al.* "Sensitivity analysis and application of machine learning methods to predict the heat transfer performance of CNT/water nanofluid flows through coils", **International Journal of Heat and Mass Transfer**, v. 128, p. 825–835, 2019. DOI: 10.1016/j.ijheatmasstransfer.2018.09.041.
- BAGHBAN, A., POURFAYAZ, F., AHMADI, M. H., *et al.* "Connectionist intelligent model estimates of convective heat transfer coefficient of nanofluids in circular cross-sectional channels", **Journal of Thermal Analysis and Calorimetry**, v. 132, n. 2, p. 1213–1239, 2018. DOI: 10.1007/s10973-017-6886-z.
- BAGHBAN, A., SASANIPOUR, J., POURFAYAZ, F., *et al.* "Towards experimental and modeling study of heat transfer performance of water- SiO<sub>2</sub> nanofluid in quadrangular cross-section channels", **Engineering Applications of Computational Fluid Mechanics**, v. 13, n. 1, p. 453–469, 2019. DOI: 10.1080/19942060.2019.1599428.
- BOVET, N., YANG, M., JAVADI, M. S., *et al.* "Interaction of alcohols with the calcite surface", **Physical Chemistry Chemical Physics**, v. 17, n. 5, p. 3490–3496, 2015. DOI: 10.1039/c4cp05235h.
- CHAO, J., ZHANG, L., FENG, R., *et al.* "Experimental study on the compatibility of scale inhibitors with Mono Ethylene Glycol", **Petroleum Research**, v. 5, n. 4, p.

315–325, 2020. DOI: 10.1016/j.ptlrs.2020.07.003.

CHESHMEH SEFIDI, A., AJORKARAN, F. "A novel MLP-ANN approach to predict solution gas-oil ratio", **Petroleum Science and Technology**, v. 37, n. 23, p. 2302–2308, 2019. DOI: 10.1080/10916466.2018.1490759.

CHOJACZYK, A. A., TEIXEIRA, A. P., NEVES, L. C., *et al.* "Review and application of Artificial Neural Networks models in reliability analysis of steel structures", **Structural Safety**, v. 52, n. PA, p. 78–89, 2015. DOI: 10.1016/j.strusafe.2014.09.002.

DA ROSA, K. R. S. A., FONTES, R. A., DO ROSÁRIO, F. F., *et al.* "Improved protocol for scale inhibitor evaluation: A meaningful step on scale management", **Offshore Technology Conference Brasil 2019, OTCB 2019**, 2020. DOI: 10.4043/29683-ms.

DATA BRIGDE - MARKET RESEARCH. **Global Oilfield Scale Inhibitor Market – Industry Trends and Forecast to 2027**. 2020. Disponível em: <https://www.databridgemarketresearch.com/reports/global-oilfield-scale-inhibitor-market>.

DE MORAIS, S. C., DE LIMA, D. F., FERREIRA, T. M., *et al.* "Effect of pH on the efficiency of sodium hexametaphosphate as calcium carbonate scale inhibitor at high temperature and high pressure", **Desalination**, v. 491, n. May, p. 114548, 2020. DOI: 10.1016/j.desal.2020.114548.

DE OÑA, J., GARRIDO, C. "Extracting the contribution of independent variables in neural network models: A new approach to handle instability", **Neural Computing and Applications**, v. 25, n. 3–4, p. 859–869, 2014. DOI: 10.1007/s00521-014-1573-5.

DE SOUZA, A. V. A., ROSÁRIO, F., CAJAIBA, J. "Evaluation of calcium carbonate inhibitors using sintered metal filter in a pressurized dynamic system", **Materials**, v. 12, n. 11, p. 1–13, 2019. DOI: 10.3390/ma12111849.

DYER, S. J., GRAHAM, G. M. "The effect of temperature and pressure on oilfield scale formation", **Journal of Petroleum Science and Engineering**, v. 35, n. 1–2, p. 95–107, 2002. DOI: 10.1016/S0920-4105(02)00217-6.

FERNANDES, R. S., SANTOS, W. D. L., DE LIMA, D. F., *et al.* "Application of water-soluble polymers as calcium carbonate scale inhibitors in petroleum wells: A uni- and multivariate approach", **Desalination**, v. 515, n. June, p. 115201, 2021. DOI: 10.1016/j.desal.2021.115201.

FRENIER, W. W. ., ZIAUDDIN, M. **Formation, Removal, and Inhibition of Inorganic Scale in the Oilfield Environment**. [S.l.], Society of Petroleum Engineers, 2008.

GARSON, G. D. "Interpreting neural-network connection weights", **AI Expert**, v. 6, n. 4, p. 47–51, 1991.

GHRITLAHRE, H. K., PRASAD, R. K. "Application of ANN technique to predict the performance of solar collector systems - A review", **Renewable and Sustainable Energy Reviews**, v. 84, n. September 2017, p. 75–88, 2018. DOI: 10.1016/j.rser.2018.01.001.

HAMMOUDI, A., MOUSSACEB, K., BELEBCHOUHE, C., *et al.* "Comparison of artificial neural network (ANN) and response surface methodology (RSM) prediction in compressive strength of recycled concrete aggregates", **Construction and Building Materials**, v. 209, p. 425–436, 2019. DOI: 10.1016/j.conbuildmat.2019.03.119.

HAYKIN, S. **Redes Neurais - Principios prática**. 2ed. ed. Porto Alegre - RS - Brazil, Bookman, 2001.

HEIDARI, A. A., FARIS, H., MIRJALILI, S., *et al.* **Ant lion optimizer: Theory, literature review, and application in multi-layer perceptron neural networks**. [S.l.], Springer International Publishing, 2020. v. 811. Disponível em: [http://dx.doi.org/10.1007/978-3-030-12127-3\\_3](http://dx.doi.org/10.1007/978-3-030-12127-3_3).

INSIGHTS, C. M. **Hydrate Inhibitors Market Analysis (Hydrate Inhibitors Market Report, by Inhibitor Type (Thermodynamic Inhibitors, Low Dosage Hydrate Inhibitors (LDHI's), and Hybrid) - Size, Share, trends, and Forecast to 2025)**. 2018. Disponível em: <https://www.coherentmarketinsights.com/market-insight/hydrate-inhibitors-market-555>.

ISLAMI RAD, S. Z., GHOLIPOUR PEYVANDI, R. "A simple and inexpensive design for volume fraction prediction in three-phase flow meter: Single source-single detector", **Flow Measurement and Instrumentation**, v. 69, n. June, p. 101587, 2019. DOI: 10.1016/j.flowmeasinst.2019.101587.

KAN, A. T., FU, G., WATSON, M. A., *et al.* "Effect of hydrate inhibitors on oilfield scale formation and inhibition", **SPE Oilfield Scale Symposium**, p. 83–94, 2002. DOI: 10.2118/74657-ms.

KARTNALLER, V. **AVALIAÇÃO DA INFLUÊNCIA DO USO DE INIBIDORES DE HIDRATOS NO PROCESSO DE INCRUSTAÇÃO DE CARBONATO DE CÁLCIO EM SISTEMA DINÂMICO PRESSURIZADO**. **Energies**. 1–167 f. Universidade Federal do Rio de Janeiro (UFRJ), 2018.

KARTNALLER, V., VENÂNCIO, F., DO ROSÁRIO, F. F., *et al.* "Application of multiple regression and design of experiments for modelling the effect of monoethylene glycol in the calcium carbonate scaling process", **Molecules**, v. 23, n. 4, p. 1–12, 2018. DOI: 10.3390/molecules23040860.

KHALIL DE OLIVEIRA, M. C., GONÇALVES, M. A. "An effort to establish correlations between brazilian crude oils properties and flow assurance related issues", **Energy and Fuels**, v. 26, n. 9, p. 5689–5701, 2012. DOI: 10.1021/ef300650k.

KHORMALI, A., SHARIFOV, A. R., TORBA, D. I. "Increasing efficiency of calcium sulfate scale prevention using a new mixture of phosphonate scale inhibitors during waterflooding", **Journal of Petroleum Science and Engineering**, v. 164, n. February, p. 245–258, 2018. DOI: 10.1016/j.petrol.2018.01.055.

KIM, H., YOO, W., LIM, Y., *et al.* "Economic evaluation of MEG injection and regeneration process for oil FPSO", **Journal of Petroleum Science and Engineering**, v. 164, n. February, p. 417–426, 2018. DOI: 10.1016/j.petrol.2018.01.071.

KUMAR, S., NAIYA, T. K., KUMAR, T. "Developments in oilfield scale handling

towards green technology-A review", **Journal of Petroleum Science and Engineering**, v. 169, n. May, p. 428–444, 2018. DOI: 10.1016/j.petrol.2018.05.068.

LI, Hong, YU, H., CAO, N., *et al.* "Applications of Artificial Intelligence in Oil and Gas Development", **Archives of Computational Methods in Engineering**, v. 28, n. 3, p. 937–949, 2021. DOI: 10.1007/s11831-020-09402-8.

LI, Hao, ZHANG, Z., LIU, Z. "Application of artificial neural networks for catalysis: A review", **Catalysts**, v. 7, n. 10, 2017. DOI: 10.3390/catal7100306.

LIM, V. W. S., METAXAS, P. J., STANWIX, P. L., *et al.* "Gas hydrate formation probability and growth rate as a function of kinetic hydrate inhibitor (KHI) concentration", **Chemical Engineering Journal**, v. 388, n. January, p. 124177, 2020. DOI: 10.1016/j.cej.2020.124177.

MACEDO, R. G. M. d. A., MARQUES, N. do N., PAULUCCI, L. C. S., *et al.* "Water-soluble carboxymethylchitosan as green scale inhibitor in oil wells", **Carbohydrate Polymers**, v. 215, n. February, p. 137–142, 2019. DOI: 10.1016/j.carbpol.2019.03.082.

MARKETS, M. and. **Oilfield Scale Inhibitor Market by Type (Phosphonates, Carboxylate/Acrylic, Sulfonates, and Others), and by Region (North America, Europe, Asia-Pacific, Middle East, Africa, and South America) - Global Forecast to 2020.** 2016. Disponível em: <https://www.marketsandmarkets.com/Market-Reports/oilfield-scale-inhibitor-market-268225660.html>.

MATHWORKS. **trainbr - Bayesian regularization backpropagation.** 2021a. Disponível em: <https://www.mathworks.com/help/deeplearning/ref/trainbr.html>. Acesso em: 27 set. 2020.

MATHWORKS. **traindxd - Gradient descent with momentum and adaptive learning rate backpropagation.** 2021b. Disponível em: <https://www.mathworks.com/help/deeplearning/ref/traindxd.html>. Acesso em: 27 set. 2020.

MATHWORKS. **trainlm - Levenberg-Marquardt backpropagation.** 2021c. Disponível em: <https://www.mathworks.com/help/deeplearning/ref/trainlm.html>. Acesso em: 27 set. 2020.

MELCHUNA, A., ZHANG, X., SA, J. H., *et al.* "Flow Risk Index: A New Metric for Solid Precipitation Assessment in Flow Assurance Management Applied to Gas Hydrate Transportability", **Energy and Fuels**, v. 34, n. 8, p. 9371–9378, 2020. DOI: 10.1021/acs.energyfuels.0c01203.

NAIT AMAR, M., JAHANBANI GHAFAROKHI, A., SHANG WU NG, C. "Predicting wax deposition using robust machine learning techniques", **Petroleum**, 2021. DOI: 10.1016/j.petlm.2021.07.005.

NASIR, Q., SULEMAN, H., ELSHEIKH, Y. A. "A review on the role and impact of various additives as promoters/ inhibitors for gas hydrate formation", **Journal of Natural Gas Science and Engineering**, v. 76, n. December 2019, p. 103211, 2020. DOI: 10.1016/j.jngse.2020.103211.

NGUYEN, D. A., IWANIW, M. A., FOGLER, H. S. "Kinetics and mechanism of the reaction between ammonium and nitrite ions: Experimental and theoretical studies", **Chemical Engineering Science**, v. 58, n. 19, p. 4351–4362, 2003. DOI: 10.1016/S0009-2509(03)00317-8. .

ODDO, J. E., TOMSON, M. B. "Why Scale Forms in the Oil Field and Methods To Predict It", **SPE Production & Facilities**, v. 9, n. 01, p. 47–54, 1 fev. 1994. DOI: 10.2118/21710-PA.

OKHRIMENKO, D. V., NISSENBAUM, J., ANDERSSON, M. P., *et al.* "Energies of the adsorption of functional groups to calcium carbonate polymorphs: The importance of -OH and -COOH groups", **Langmuir**, v. 29, n. 35, p. 11062–11073, 2013. DOI: 10.1021/la402305x. .

OLAJIRE, A. A. "A review of oilfield scale management technology for oil and gas production", **Journal of Petroleum Science and Engineering**, v. 135, p. 723–737, 2015. DOI: 10.1016/j.petrol.2015.09.011.

OTCHERE, D. A., ARBI GANAT, T. O., GHOLAMI, R., *et al.* "Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models", **Journal of Petroleum Science and Engineering**, v. 200, n. August 2020, p. 108182, 2021. DOI: 10.1016/j.petrol.2020.108182.

PAZ, P. A., CAPRACE, J.-D., CAJAIBA, J. F., *et al.* "Prediction of Calcium Carbonate Scaling in Pipes Using Artificial Neural Networks". 25 jun. 2017. **Anais [...]** Trondheim, Norway, ASME, 25 jun. 2017. p. 1–10. DOI: 10.1115/OMAE2017-61233.

PENTOS, K. "The methods of extracting the contribution of variables in artificial neural network models - Comparison of inherent instability", **Computers and Electronics in Agriculture**, v. 127, p. 141–146, 2016. DOI: 10.1016/j.compag.2016.06.010.

RAHMANIFARD, H., PLAKSINA, T. "Application of artificial intelligence techniques in the petroleum industry: a review", **Artificial Intelligence Review**, v. 52, n. 4, p. 2295–2318, 2019. DOI: 10.1007/s10462-018-9612-8.

RAMZI, M., HOSNY, R., EL-SAYED, M., *et al.* "Evaluation of scale inhibitors performance under simulated flowing field conditions using dynamic tube blocking test", **International Journal of Chemical Sciences**, v. 14, n. 1, p. 16–28, 2016. .

SAND, K. K., YANG, M., MAKOVICKY, E., *et al.* "Binding of ethanol on calcite: The role of the OH bond and its relevance to biomineralization", **Langmuir**, v. 26, n. 19, p. 15239–15247, 2010. DOI: 10.1021/la101136j.

SANNI, O. S., BUKUAGHANGIN, O., CHARPENTIER, T. V. J., *et al.* "Evaluation of laboratory techniques for assessing scale inhibition efficiency", **Journal of Petroleum Science and Engineering**, v. 182, n. July, p. 106347, 2019. DOI: 10.1016/j.petrol.2019.106347.

SANTOS, H. F. L., CASTRO, B. B., BLOCH, M., *et al.* "A physical model for scale growth during the dynamic tube blocking test", **OTC Brasil 2017**, n. Figure 1, p. 161–180, 2017. DOI: 10.4043/27956-ms.



SEIERSTEN, M., KUNDU, S. S. "Scale management in monoethylene glycol MEG systems - A review", **Society of Petroleum Engineers - SPE International Oilfield Scale Conference and Exhibition 2018**, n. June, p. 20–21, 2018. DOI: 10.2118/190738-ms.

SOLEIMANI, R., SHOUSHARI, N. A., MIRZA, B., *et al.* "Experimental investigation, modeling and optimization of membrane separation using artificial neural network and multi-objective optimization using genetic algorithm", **Chemical Engineering Research and Design**, v. 91, n. 5, p. 883–903, 2013. DOI: 10.1016/j.cherd.2012.08.004.

WANG, J., LV, Z., LIANG, Y., *et al.* "Fouling resistance prediction based on GA–Elman neural network for circulating cooling water with electromagnetic anti-fouling treatment", **Journal of the Energy Institute**, v. 92, n. 5, p. 1519–1526, 2019. DOI: 10.1016/j.joei.2018.07.022.

WANG, Q., AL-NASSER, W., CHEN, T., *et al.* "Calcium Carbonate Scale Inhibition: Effects of EOR Chemicals". 2018. **Anais [...]** Phoenix, Arizona, USA, NACE International, 2018. p. 1–12.

XU, M., WONG, T. C., CHIN, K. S. "Modeling daily patient arrivals at Emergency Department and quantifying the relative importance of contributing variables using artificial neural network", **Decision Support Systems**, v. 54, n. 3, p. 1488–1498, 2013. DOI: 10.1016/j.dss.2012.12.019. .

YANG, M., STIPP, S. L. S., HARDING, J. "Biological control on calcite crystallization by polysaccharides", **Crystal Growth and Design**, v. 8, n. 11, p. 4066–4074, 2008. DOI: 10.1021/cg800508t.

ZAMEN, M., BAGHBAN, A., POURKIAEI, S. M., *et al.* "Optimization methods using artificial intelligence algorithms to estimate thermal efficiency of PV/T system", **Energy Science and Engineering**, v. 7, n. 3, p. 821–834, 2019. DOI: 10.1002/ese3.312.

ZAREI, F., BAGHBAN, A. "Phase behavior modelling of asphaltene precipitation utilizing MLP-ANN approach", **Petroleum Science and Technology**, v. 35, n. 20, p. 2009–2015, 2017. DOI: 10.1080/10916466.2017.1377233.

ZHANG, L., YUE, L. H., WANG, F., *et al.* "Divisive effect of alcohol-water mixed solvents on growth morphology of calcium carbonate crystals", **Journal of Physical Chemistry B**, v. 112, n. 34, p. 10668–10674, 2008. DOI: 10.1021/jp8034659.

## 4

### **Machine learning models for measurement of pH using a low-cost image analysis strategy**

The second part of this work is indirectly connected with the other type of fouling that may occur during production and has its origin in the organic components from the oil and gas. They are also caused by the different conditions (mainly pressure and temperature) that the three-phase mixture (oil-gas-water) is exposed to during the extraction process. This problem can be avoided using inhibitors, but once the fouling is formed, there are some strategies that can be used to unplug the pipes and valves. One of these alternatives has been used extensively in the last decades, known as Nitrogen Generating System (NGS). This system releases a great amount of heat and  $N_2$  gas that act to redissolve the wax precipitates and gas hydrates.

As presented in Section 2, this system has its kinetics very dependent on the pH conditions, which is a complex parameter to measure and monitor under high-pressure conditions. That motivates the development of a model to determine the pH in a pressurized system that could be applied to monitor the NGS in future applications.

This section contains the manuscript version of the article that presents the results of the development of the model. Supporting Information, which will be available with the manuscript during the submission process, is presented in Appendix B.

## 4.1

## Article Manuscript

# Machine learning models for measurement of pH using a low-cost image analysis strategy

*Bruno X. Ferreira<sup>a</sup>, Alline V. B. de Oliveira<sup>b</sup>, João Cajaiba<sup>b</sup>, Vinicius Kartnaller<sup>b</sup>,  
Brunno F. Santos<sup>a\*</sup>.*

## AUTHOR ADDRESS:

*a* - Department of Chemical and Materials Engineering (DEQM), Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rua Marquês de São Vicente, 225 – Gávea, Rio de Janeiro, RJ 22430-060, Brazil.

*b* - Instituto de Química, Núcleo de Desenvolvimento de Processos e Análises Químicas em Tempo Real (NQTR), Universidade Federal do Rio de Janeiro, Rua Hélio de Almeida 40, Cidade Universitária, Rio de Janeiro 21941-614, Brazil.

*\*bsantos@puc-rio.br*

## ABSTRACT:

One difficult measurement to be performed is the pH values in a pressurized system, requiring specialized equipment. With this problem as a goal, this work aims to develop models to determine the pH in pressurized systems (up to 6 MPa) as an initial step to create an applicable soft sensor. For that, classification and prediction models were created using image analysis and different Machine Learning techniques: Convolution Neural Networks (CNN), Support Vector Machines, and Decision Trees. All of them were explored in the classification models, but CNN was the only used for the regression ones. The best models for each technique were tested in two study cases: titration curve and CO<sub>2</sub>-H<sub>2</sub>O equilibrium systems. The best classification models were the CNN ones, but the model with the best performance was the predictive CNN using the reduce RGB images with 30 neurons in the last dense layer, which presents R<sup>2</sup> values higher than 80%.

**KEYWORDS:** Convolutional Neural Networks; Image processing; pH meter models; Flow assurance.

#### 4.1.1

##### Introduction

Industrial processes demand the use of a large number of sensors to control and monitor the operational conditions for several different variables, such as temperature, pressure, liquid level, and concentration. However, some of these parameters are hard to measure in real time because of technical difficulties, elevated costs, and other factors (KADLEC *et al.*, 2009; FUNATSU, 2018). One special process variable is the pH, which is present in several types of chemical industries, from the control of the kinetics of a reaction to the monitoring of the quality of the product or reagents (KHAN *et al.*, 2017).

Inside the pH measurement context, monitoring the pH values in the high-pressure system is challenging due to the difficulty of producing the equipment, even though there have been published studies in this field since the middle of the last century. Usually, the standard pH sensors, such as glass electrode ones, are available for pressures up to 16 bar. Making that suitable equipment available in the market has high prices. That makes the development of indirect methods to measure or predict pH-value in high pressure conditions an interest subject to be explored, as highlighted by Lemmer *et al.* (2017) (DE OLIVEIRA *et al.*, 2019; BYCHKOV *et al.*, 2020; CROLET and BONI, 1983; SAMARAYAKE and SASTRY, 2013).

As an alternative to physical sensors, the development and application of the called soft sensors are becoming more common in the industrial scenario, with an emphasis on the chemical industry (POERIO and BROWN, 2018; Sun and GE, 2021; YAN *et al.*, 2017). Soft sensors are predictive models, which are usually created using two main strategies: using first-principal models (white-box models) or using the available database store from the past measurements (data-driven or black-box models) (SHANG *et al.*, 2014; KADLEC *et al.*, 2009). Data-driven models are a very popular strategy adopted to develop soft sensors since they do not require extensive knowledge about the system but a sufficient amount of information with enough quality to estimate the process's properties properly. That

makes this strategy very attractive to complex processes found in the industry. For that, conventionally, the modeling process uses a variety of statistical inference and Machine Learning (ML) techniques (SUN and GE, 2021; SANSANA *et al.*, 2021).

Due to the huge importance of pH for the industry, many soft sensors were created to measure the pH values in different specific industrial processes, applying different modeling strategies. For example, Dixit *et al.* (2021) have used Convolution Neural Network (CNN) models to predict the pH in red meat using hyperspectral images intended to monitor this important quality parameter. Also, the work of Capel-cuevas *et al.* (2011) developed a Multi-layer Perceptron (MLP) model to predict the pH value in a solution through image analysis for that using the hue value ( $h$ ) of a picture with 11 immobilized sensing elements, covering the pH values on the range 0-14.

This study aimed to develop classification and prediction models to determine the pH in pressurized systems, using image analysis and different ML strategies. The models will be developed using known buffer solutions. Then the best model will be tested in two other scenarios (study cases) to evaluate their performance.

#### 4.1.1.1

#### Modeling Strategies

##### 4.1.1.1.1

#### Convolution Neural Networks (CNNs)

Considered a subtype of deep discriminative architecture, the CNN is inspired by the animal visual cortex organization and has its concept based on a Time-delay Neural Network (TDNN). In the CNN, the convolution process replaced the general matrix multiplication presented in others Artificial Neural Networks (ANN). CNN has been demonstrated to be suitable for processing two-dimensional data with grid-like topologies, like images and videos. Additionally, the use of CNN requires minimal pre-processing, allowing end-to-end solutions. With the rapid development of computation, the use of GPU-accelerated computing has improved the CNNs train efficiency (BOUWMANS *et al.*, 2019; LIU *et al.*, 2017). CNN has been applied in several fields, such as sea surface temperature prediction (HAGHBIN *et al.*, 2021), detection of fracture in coal (KARIMPOULI *et al.*, 2020), identification of superheat situations (LEI *et al.*, 2020), and to predict

soil properties (WADOUX *et al.*, 2019).

CNN structure, Figure 4.1, consists of an input layer, an output layer, and multiple hidden layers. The hidden layers are divided into three classes: convolutional, pooling, and fully connected. In the convolutional layers, the most important part of the CNNs, are applied the convolution operations, the addition of the bias as the input data, and the transference of the results to the activation functions, so the result can be directed to the next layer. The weights and biases of this layer are organized into a series of kernels (or filters) responsible for the local feature extractions (YAO *et al.*, 2019; ZAN *et al.*, 2020; SHEN *et al.*, 2021). The most common type of activation functions used for the CNN is the sigmoid function (*sigmoid*) and rectified linear units (ReLU) (ZAN *et al.*, 2020), but other kinds, such as hyperbolic tangents (*tanh*), can be used (RIZKIN *et al.*, 2019).

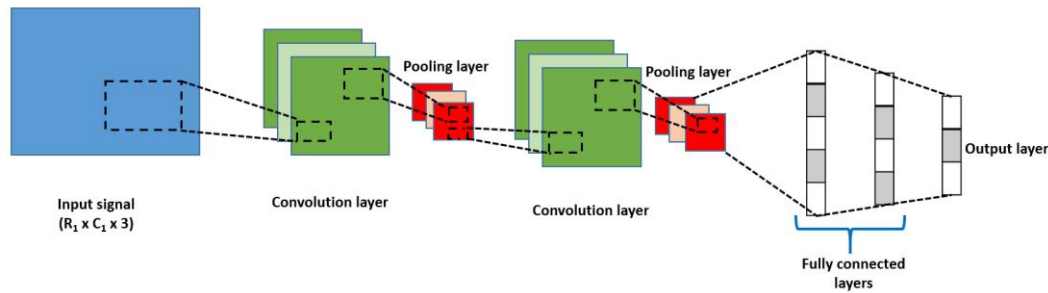


Figure 4.1: CNN schematic representation.

In the pooling layers (or subsampling layer), the downsampling is completed, reducing the dimension of feature maps. Commonly, the strategies used in this layer are maximum pooling (*max pooling*) (used in this study) and average pooling. These layers are used after one or two convolutional layers. Finally, the last hidden layer type is the fully connected layer (or dense layer), where all the neurons are connected with active ones from the previous layer. Then the last dense layers are connected with the output layer that aims to integrate the highly abstract features for classification or regression tasks. In this type, all the neurons are connected with active ones from the previous layer (YAO *et al.*, 2019; ZAN *et al.*, 2020; YUAN *et al.*, 2020).

#### 4.1.1.1.2

### Support Vector Machines (SVMs)

Support vector machine, Figure 4.2, is a ML method that was first developed for binary linear classification problems proposed by Cortes and Vapnik (1995). The technique separates the classes with the largest gap (optimal margin) between the borderline instances (Support Vectors), which leads to the method being known as an optimal margin classifier. SVM is widely used in classification problems due to its simplicity, strong generalization ability, and computational efficiency (ASGHER *et al.*, 2020; PENG *et al.*, 2020). The method evolution allows it to be applied to multi-class problems, using techniques like One-versus-One (OvO) and One-versus-Rest (OvR), and to be used for non-linearly separable data using kernels (CHAUHAN *et al.*, 2019; DING *et al.*, 2019).

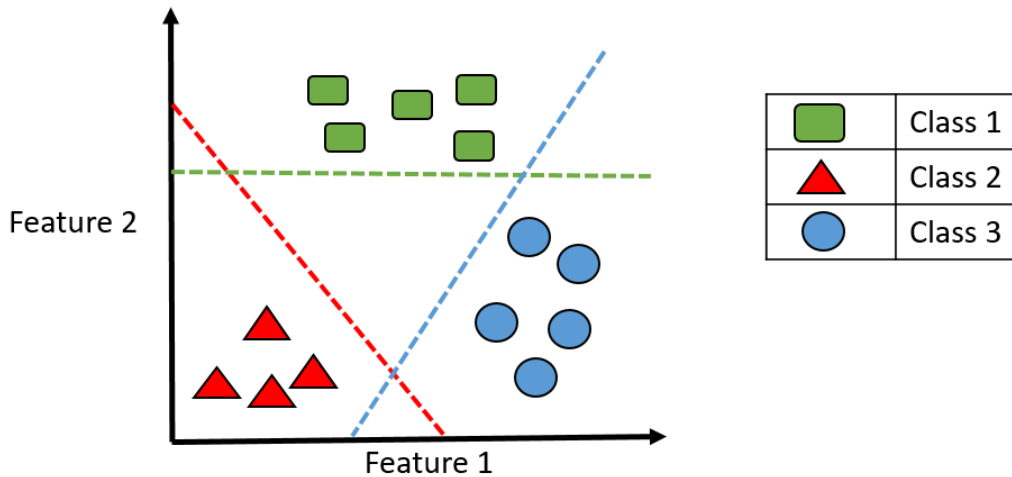


Figure 4.2: SVM schematic representation, with a three classes problem using linear kernels (dashed lines)

Kernels are mathematical functions that transform the data from a given space (input space) to a new one with more dimensions (feature space), where this data can be separated with the linear surfaces (hyperplanes) (CHAUHAN *et al.*, 2019). The most common kinds of a kernel are linear (Eq. 4.1), polynomial (Eq. 4.2), RBF (Radial-Basis Function, Eq. 4.3), and sigmoid (Eq. 4.4).

$$Kernel_{linear}(x_i, x_j) = (gamma(x_i, x_j) + coef) \quad (4.1)$$

$$Kernel_{poly}(x_i, x_j) = (gamma(x_i, x_j) + coef)^{degree} \quad (4.2)$$

$$Kernel_{RBF}(x_i, x_j) = \exp(-gamma\|x_i - x_j\|^2) \quad (4.3)$$

$$Kernel_{sigmoid}(x_i, x_j) = \tanh(gamma(x_i, x_j) + coef) \quad (4.4)$$

They depend on the hyperparameters *degree* and *gamma*, which can be optimized. The first one is related to the degree of the polynomial function, being present only in Eq. 4.2. The hyperparameter *gamma* represents the influence of each data in the training database in the optimal decision surface position, which then can be a function of only the numbers of the variable or also the variance of the normalized dataset matrix. Another important hyperparameter is the *C*, which represents a regularization cost of the misclassification and the influence on the margin width and hardness of the SVM models (LORENA *et al.*, 2007; RHYS *et al.*, 2020 and SCIKIT-LEARN, 2022a).

#### 4.1.1.3

##### Decision Trees (DTs)

Decision Tree (DT) is another very widely used ML technique for classification problems due to the easy implementation and understanding of its step. DT structure, Figure 4.3, is composed of several binary tests along the tree, where the tests start in the root node of the DT and progress through the different nodes, still reaching one of the leaf nodes the determine the class of the data (GEURTS *et al.*, 2009; TANGIRALA, 2020; PRIYAM *et al.*, 2013 and SCIKIT-LEARN, 2022b).

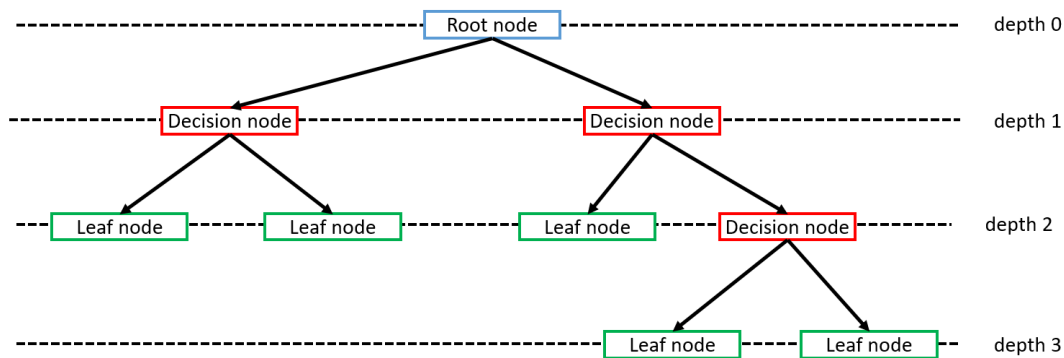


Figure 4.3: DT schematic representation.

Some of the hyperparameters which can be explored during the development of the models are the depth of the tree that is determined by the



number of layers from the root node to the leaf node, the maximum number of leaves in the last layer of DT, and the criteria applied to evaluate the quality of the split during the training process. To evaluate the quality of the split can be used different metrics parameters the measure the purity of the nodes along the DT. In this study, the two ones tested were the Gini impurity (*gini*) and the cross-entropy (*entropy*) (HASTIE *et al.*, 2009 and SCIKIT-LEARN, 2022b).

#### 4.1.2.

#### Methodology

##### 4.1.2.1

##### Case study: Pressurized reactor

The experiments were conducted on a midiclave reactor apparatus (*Büchiglasuster*, *Gschwaderstrasse* 12, Uster, Switzerland) with a double-walled reaction vessel constructed on AISI 316 stainless steel, Figure 4.4. It had two borosilicate windows disposed at a 180° angle horizontally, temperature and pressure transducers, mechanical stirring, and data acquisition controlled by Büchi software bls2 2.7e. In one window, a light source was adapted, provided by a LED lamp dimmer GU 10 5 W, dual voltage, with a luminous flux of 280 lm, coupled with a polymer circular polarizer filter (with a diameter of 40 mm) and with the light intensity controlled using a dimmer shield together with an Arduino Uno. On the second window was connected to a Microsoft LifeCam Cinema HD. The image acquisition occurred at the rate of 20 images of 32 bits per second with a resolution of 1280 x 800 pixels.

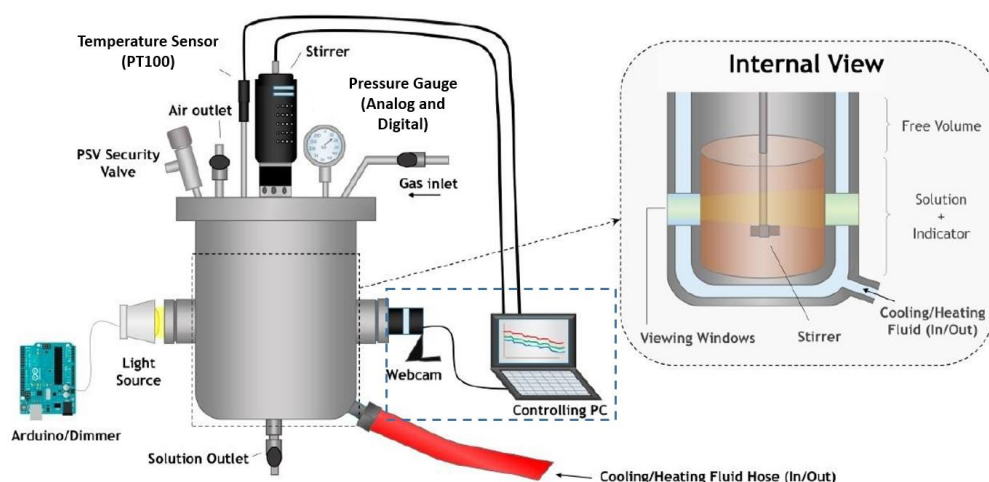


Figure 4.4: Experimental setup scheme (adapted from DE OLIVEIRA *et al.*, 2019)

The Korthoff indicator was prepared by mixing different pH indicators, previously diluted to 0.1 %, and more details can be found in De Oliveira *et al.* (2019). It was added to the proportion of 1 % v/v of the liquid phase for all experiments. The nine buffer solutions in the pH range 2-10 were prepared in a concentration of 0.1 mol L<sup>-1</sup> with deionized water in the final volume of 500 mL, according to the procedure described in De Oliveira *et al.* (2019). All the buffer solutions had their exact pH value at atmospheric pressure determined using a previously calibrated pH meter (*Mettler-Toledo SevenMulti<sup>TM</sup> S47*, Columbus, USA).

To construct the calibration curve, the batch reactor was loaded with 200 mL of the buffer solution containing the Korthoff indicator for each pH value. The experiments were maintained at 298.15 K, 200 rpm stirring rate and using different working pressure (0.0, 0.5, 1.0, 2.0, 4.0, 6.0 MPa with N<sub>2</sub>), waiting around 10 minutes to stabilize the signal at each pressure.

Aiming to test the developed models with the different strategies, they were applied to predict the pH values in two situations. First, an acid-base titration curve was performed using a strong base (NaOH 0.04724 mol L<sup>-1</sup>) and a strong acid (HCl 0.01789 mol L<sup>-1</sup>). The reactor, at 298.15 K, 200 rpm, and atmospheric pressure, was filled with 200 mL of HCl solution with the Korthoff indicator, and the NaOH solution was added to 150 mL at the flow rate of 2.00 mL min<sup>-1</sup> using an HPLC pump (Shimadzu model LC-20AR, Kyoto, Japan). In some experiments, the system was pressurized with N<sub>2</sub> at 6 MPa. The pH values obtained through the models were

compared with the values calculated using the concentrations of the solutions for the theoretical titration curve. The second case was the pH measurement in the pressurized CO<sub>2</sub>-H<sub>2</sub>O systems. For that, first, the system was pressurized with CO<sub>2</sub> at the desired pressures (0.1, 0.3, 0.5, 1.0, 2.0, and 5.0 MPa). Then, 200 mL of distilled water containing the Korthoff indicator was added to the reactor using an HPLC pump, waiting around 2 h for the stabilization of the system. The system pressure raised with the addition of water, but then it decreased due to the dissolution of the CO<sub>2</sub> in the water until it reached the system equilibrium. The pH values measured in the equilibrium condition were compared with obtained data in the literature.

#### 4.1.2.2

##### Database preparation

The experimental data used in this work were previously presented by De Oliveira *et al.* (2019). For the development of the model, a dataset with 386 images on the RGB color system with 1280 x 800 pixels of resolution was selected, composed of images for all the nine pH categories in the amount shown in Table 4.1. Figure 4.5 shows examples of images for each pH category. The pH values were verified using the equation presented on the based work and using, when necessary, the function *round* (NUMPY, 2022). This database was split into three groups, train (70%), validation (15%), and test (15%) to develop the models.

Table 4.1. Number of images for each pH category in the training database

pH categories	Number of images
2	40
3	49
4	45
5	45
6	39
7	24
8	50
9	43
10	51

Total	386
-------	-----

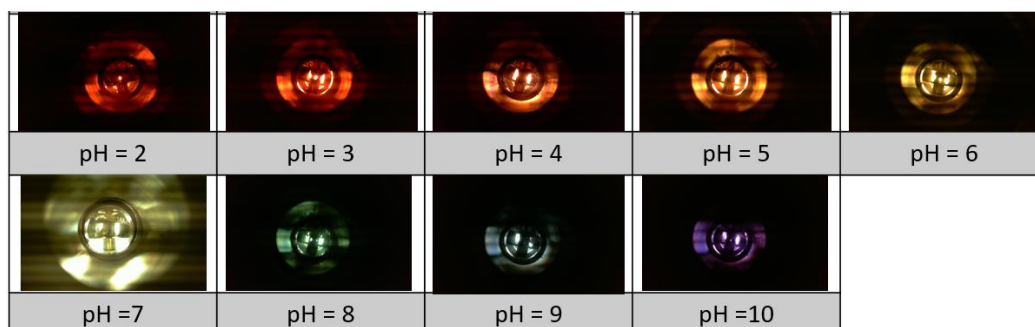


Figure 4.5: Examples of the images presented in the dataset for each pH category.

To build the models, different kinds of input were tested. For that, the reactor images pass for some simple pre-processing steps, such as changing the color system from RGB to HSV and reducing their resolution by cropping the center region of the images. Thus, resulting in the input types presented in Table 4.2.

Table 4.2. Types of input tested in the different models developed.

Code	Color system	Input information	Resolution
Input 1	RGB	RGB components	1280 x 800
Input 2	RGB	RGB components	280 x 280
Input 3	HSV	HSV components	280 x 280
Input 4	HSV	<i>hue</i>	280 x 280
Input 5	HSV	<i>saturation</i>	280 x 280
Input 6	HSV	<i>value</i>	280 x 280

A second image dataset was also obtained from the application test of the models developed on the acid-base titration curve and the pressurized CO<sub>2</sub>-H<sub>2</sub>O systems cases for those following the same procedures to determine the pH values or classes applied for the first database.

#### 4.1.2.3

##### Modeling strategies

This work explored three methodologies to build a soft sensor to determine the pH value as a class: CNN, SVM, and DT. The CNN models also developed a sensor to predict the pH value with one decimal case of accuracy. All the models were

built and tested using Python v3.6 as a programming language on the Google Colaboratory Pro environment.

#### 4.1.2.3.1

##### CNN

The proposed architecture for the CNNs had two dense layers using Rectified Linear Units (ReLU) as activation functions. For the classifier models, the output layer had nine neurons, one for each class that gives the probability of the image belong to each class, using the softmax as an activation function. The regression models had only one neuron in this layer that gives a value for the pH in the range of 2-10, in which linear activation function was used. The other parameters of the architecture were explored as hyperparameters and summarized with their options or search regions on Tab. 4.3. For the CNNs models, all six types of inputs were tested.

Table 4.3: Hyperparameters tested in the CNN models

Hyperparameters	Search region or Options
Number epochs for training	[80 2000]
Batch size	4, 8, 16, 32, 64, 128
Number of convolution layer	[2 8]
Activation functions	'linear', 'ReLU', ' <i>sigmoid</i> ', ' <i>tanh</i> '
Filter size	8, 16, 32, 64
filter kernel	1, 3, 5
Optimizer	'adam', 'SGD', 'Adadelata'
Dropout layer percentage	0.05, 0.1, 0.15, 0.2, 0.25, 0.3
Learning rate	0.01, 0.005, 0.001, 0.0005, 0.0001
Number of neurons - first dense layer	40, 50, 60, 70, 80, 90, 100, 120
Number of neurons - second dense layer	10, 20, 30, 40, 50, 60

The results of the training and initial tests of both types of CNN models were compiled in the Weight and Biases (wandb) platform (Weight and Bias, 2022). This platform developed in Python gives an alternative to organize the machine learning results using different kinds of frameworks and libraries (such as PyTorch, Keras, and Scikit-learn) on a regular computer or using cloud-hosted ones (such as

Azure, Google Cloud, and AWS). It also gives an iterative way to compare the performance of the models and the influence of each hyperparameter, using, for example, the parallel coordinate plot.

#### 4.1.2.3.2

##### SVM

The SVM classifier models were developed using the Input 2 kind. The scikit-learn library 1.0.2 (Scikit-learn, 2022a; Pedregosa *et al.*, 2011) is used to create, train and test the models, and permits the implementation of two different approaches for the multi-class cases. The first one, known as OvR, creates a binary classification for each class versus the rest of the dataset, and the second one is called OvO, which also builds a binary classification for each class but against every other class. This was explored as a hyperparameter (Dec\_func\_shape). The other four hyperparameters investigated, Tab. 4.4, are related to the type of kernel applied and their parameters.

Table 4.4. Hyperparameters tested in SVM

Hyperparameters	Search region or options
Dec_func_shape	'OvO', 'OvR'
C	[0.001; 0.1; 0.5; 1; 2; 5]
kernel	'linear', 'poly', 'rbf', 'sigmoid'
degree	1, 2, 3, 4, 5, 6
gamma	'scale', 'auto'

#### 4.2.3.1

##### Decision Tree (DT)

The classifier models were developed using Input 2 and using the scikit-learn library 1.0.2 (Scikit-learn, 2022b; Pedregosa *et al.*, 2011) to train and test them. Table 4.5 shows the four hyperparameters explored during the development of the models.

Table 4.5: Hyperparameters tested in DT

Hyperparameters	Search region or options
Crit	'gini', 'entropy'
max_depth	5, 7, 9, 11, None

max_leaf_nodes	10, 20, None
min_samples_leaf	1, 5, 7, 10

#### 4.1.2.4

##### Statistical Performance Evaluation

To evaluate the performance of the classifiers of the models, with the different hyperparameters and architecture, the metrics were obtained using the python library sklearn.metrics 1.0.2 (SCIKIT-LEARN, 2022c; PEDREGOSA *et al.*, 2011). A very common metric is accuracy (ACC, Eq. 4.6), which is calculated by the ratio of the number of correct predictions to the total number of them (NAMUDURI *et al.*, 2020). Another two parameters used were the precision (PR, Eq. 4.7), a measure of the quantity of the prediction for a class is correct, and the recall (RC, Eq. 4.8) (or sensitivity), which represents the models' ability to correctly detect the objects that belong to the class. For the case of these three parameters, the results are in the range [0 1], being the best values closer to 1.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (4.6)$$

$$PR = \frac{TP}{TP+FP} \quad (4.7)$$

$$RC = \frac{TP}{TP+FN} \quad (4.8)$$

where TP was the number of true positives, TN was the number of true negatives, FP was the number of false positives, and FN was the number of false negatives.

Another common technique to evaluate the performance of a classifier is the Confusion Matrix (CM), which allows visualization of the classification results, and was also applied in this study. In a binary case, it is a square matrix, a 2x2 matrix (Figure 4.6), where the number of rows and columns is equal to the number of classes. CM presents information about how often a certain behavior is detected correctly or not, in which the values for parameters TP, TN, FP, and FN are reported (CAELEN, 2017; RUUSKA *et al.*, 2018; HASNAIN *et al.*, 2020).

		Predicted Class	
		Class A	Class B
True Class	Class A	TP	FN
	Class B	FP	TN

Figure 4.6: Scheme of a confusion matrix (2x2) in a binary case

For the regression CNN models, the evaluation parameters chosen were Sum of Squared Errors (SSE, Eq. 4.9), Root Mean Squared Error (RMSE, Eq. 4.10), and coefficient of determination parameter ( $R^2$ , Eq. 4.12). To calculate  $R^2$ , it was also necessary to calculate the Total Sum of Squares (TSS, Eq. 4.11). For the errors criteria adopted, the goal was to achieve the lowest values, and for the  $R^2$  the best results were indicated for values closer to 1.

$$SSE = \sum_{i=1}^n (x_i - \hat{x}_i)^2 \quad (4.9)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (x_i - \hat{x}_i)^2}{n}} \quad (4.10)$$

$$TSS = \sum_{i=1}^n (x_i - \bar{x})^2 \quad (4.11)$$

$$R^2 = 1 - \frac{SSE}{TSS} \quad (4.12)$$

where variables  $n$ ,  $x_i$ ,  $\hat{x}_i$  and  $\bar{x}$  represent the total number of data points, the observed value, the predicted value, and the mean value of the samples, respectively.

Figure 4.7 shows a schematic representation of the proposed methodology, from the dataset pre-processing up to the choice of the best models, which are then tested with the data from the system pressurized with  $\text{CO}_2$  and the acid-base titration curve.



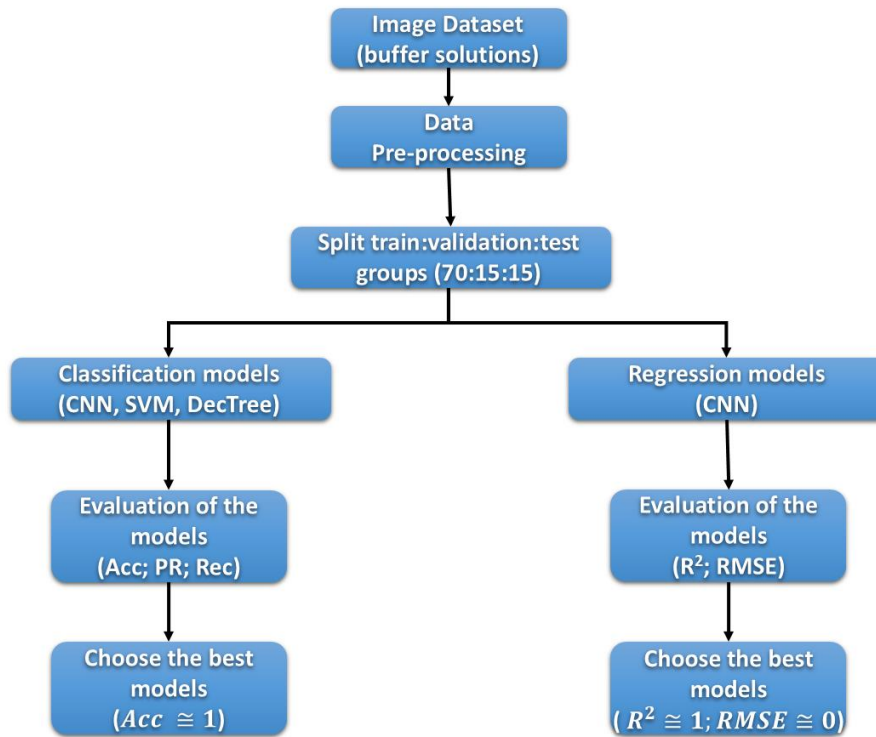


Figure 4.7: Flowchart of the methodology

### 4.1.3

#### Results and Discussion

Once the results were determined and processed into the database, they were split into training, validation, and test groups to build and optimize the models. The training data of the first two groups were directly used in the development of the models, and the test group was used as the first validation step. The different classificatory and prediction models were explored with their respective hyperparameters to compare them and find the best ones to be tested in specific applied situations.

#### 4.1.3.1

##### Evaluation of the classification models

#### 4.1.3.1.1

##### CNN classification models

Table 4.6 shows the hypermeters and the performance parameters for the best five models for Inputs 1, 2, and 3, in which the activation function used was ReLU for all models. Table 4.7 also show the same kind of information but referent

to Inputs 4, 5, and 6, which had the different types of activation function used. The results indicated that the best option for input was the image with 280 x 280 pixels using the RGB color system (Input 2) since these models presented the highest ACC values, greater than 96% for the validation and test groups, with the highest ACC value equal to 97.87% for the test group, and they also had a small number of neurons on the dense layers, making the models lighter. The better performance of the models using Input 2 is probably due to the section of the image selected containing the most important part of the information.

Table 4.6: CNN classification models topologies for the best models – Part I

models ID	Input	Hyperparameters <sup>a</sup>											training time (s)	Validation group ACC	Test group ACC
		1	2	3	4	5	6	7	8	9	10	11			
CNN_class_1	Input 1	12	139	8	8	1	3	0.001	4	0.2	120	30	578	0.9772	0.9575
CNN_class_2		12	157	4	8	3	5	0.0005	3	0.15	100	40	688	0.9635	0.95745
CNN_class_3		8	156	8	8	3	5	0.0005	3	0.2	120	20	724	0.9635	0.95745
CNN_class_4		12	137	4	8	3	5	0.0005	3	0.25	120	60	610	0.9543	0.95745
CNN_class_5		12	137	8	8	3	1	0.0005	3	0.2	100	50	580	0.9543	0.95745
CNN_class_6	Input 2	4	128	8	8	3	5	0.0001	4	0.05	100	60	98	0.9817	0.97872
CNN_class_7		12	140	8	8	3	3	0.0005	3	0.1	70	30	84	0.9817	0.97872
CNN_class_8		8	132	8	8	5	5	0.0001	4	0.05	90	60	143	0.9772	0.97872
CNN_class_9		12	123	8	8	5	5	0.0005	4	0.15	80	10	95	0.9635	0.97872
CNN_class_10		4	118	8	8	3	5	0.0001	4	0.1	120	60	99	0.9635	0.97872
CNN_class_11	Input 3	12	145	8	8	5	5	0.005	4	0.1	70	60	112	0.9178	0.85106
CNN_class_12		8	153	8	8	5	3	0.005	4	0.05	40	50	129	0.8950	0.85106
CNN_class_13		12	152	8	8	3	5	0.005	4	0.05	50	40	105	0.9452	0.80851
CNN_class_14		12	148	8	8	5	3	0.005	4	0.05	40	30	113	0.8356	0.78723
CNN_class_15		8	152	4	8	3	5	0.01	2	0.05	120	40	100	0.8721	0.78723

a - hyperparameters: 1 – batch\_size; 2- epochs; 3 – filter\_size\_1; 4 – filter\_size\_2; 5 - kernel\_size\_1; 6 - kernel\_size\_2; 7 - learning rate, 8 - n\_layers; 9 – p\_dropout; 10 - size\_dense\_1; 11 – size\_sense\_2

Table 4.7: CNN classification models topologies for the best models – Part II

models ID	Input	Hyperparameters <sup>a</sup>													training time (s)	Validation group	Test group
		1	2	3	4	5	6	7	8	9	10	11	12	13		ACC	ACC
CNN_class_16	Input 4	8	261	4	8	1	3	0.0001	2	0.25	90	40	linear	sigmoid	154	0.1598	0.1064
CNN_class_17		12	146	8	8	5	5	0.0001	4	0.2	40	60	sigmoid	linear	98	0.1507	0.1489
CNN_class_18		12	130	8	4	5	3	0.01	3	0.3	40	30	sigmoid	linear	83	0.1461	0.1489
CNN_class_19		12	163	8	4	5	5	0.005	4	0.05	50	40	relu	relu	104	0.1461	0.1489
CNN_class_20		4	134	4	8	5	5	0.01	4	0.15	50	30	relu	linear	100	0.1461	0.1489
CNN_class_21	Input 5	4	261	4	8	5	5	0.0005	4	0.2	80	60	tanh	relu	180	0.6347	0.6596
CNN_class_22		4	615	4	8	3	3	0.001	3	0.15	100	20	relu	sigmoid	345	0.5890	0.6596
CNN_class_23		4	257	8	8	5	5	0.001	3	0.15	40	50	tanh	relu	246	0.7580	0.6596
CNN_class_24		8	455	4	8	3	1	0.01	4	0.1	100	50	tanh	relu	196	0.6621	0.6596
CNN_class_25		12	683	8	4	5	5	0.005	2	0.2	50	20	relu	sigmoid	221	0.6667	0.6596
CNN_class_26	Input 6	8	222	4	4	5	3	0.005	4	0.05	40	30	linear	relu	146	0.9224	0.8723
CNN_class_27		12	416	8	8	3	5	0.005	4	0.3	120	60	relu	relu	247	0.9909	0.8723
CNN_class_28		12	447	8	8	3	3	0.005	4	0.25	90	60	linear	relu	253	0.9909	0.8723
CNN_class_29		8	446	8	8	5	5	0.001	4	0.2	90	30	relu	relu	437	0.8904	0.8085
CNN_class_30		8	549	8	8	5	5	0.001	3	0.15	120	30	linear	tanh	355	0.9315	0.7872

a - hyperparameters: 1 – batch\_size; 2- epochs; 3 – filter\_size\_1; 4 – filter\_size\_2; 5 - kernel\_size\_1; 6 - kernel\_size\_2; 7 - learning rate, 8 - n\_layers; 9 – p\_dropout; 10 - size\_dense\_1; 11 – size\_sense\_2;

12 – activation\_function\_1; 13 – activation\_function\_2

The performance of the models can also be analyzed using the CM, which allows an examination of the models' performance for each class. To exemplify that, Figure 4.8A-B shows the CMs from the CNN\_class\_10 for the training and test datasets, respectively. The images confirm the good classification results presented in Table 6 but also show that the misclassification happens between the pH classes 2, 3, and 4, presenting the class that the models could find more difficult to classify in other tests.

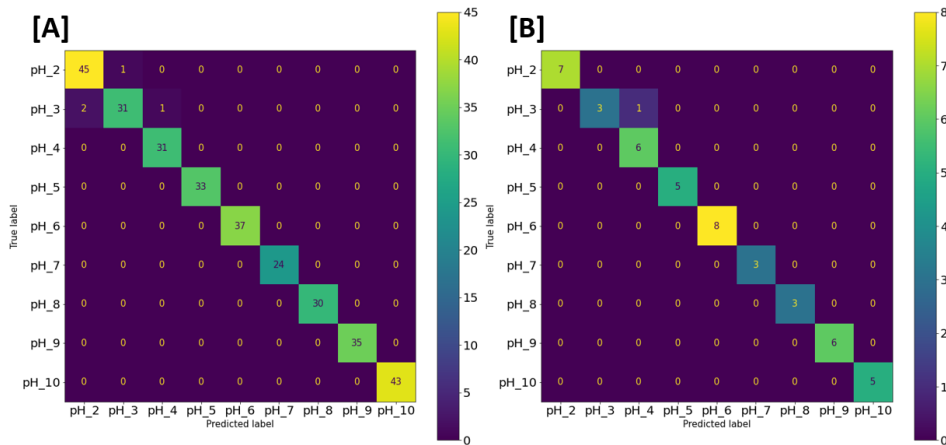


Figure 4.8: Confusion matrix (CM) of the CNN\_class\_10 for the training (A) and test (B) datasets.

The results also showed that the models using the RGB values had a better performance compared to the ones using the HSV information. This behavior was not expected since prediction models for the pH value presented in the works of Capel-Cuevas *et al.* (2011), and De Oliveira *et al.* (2019) had good results using the component *hue* of the HSV system to predict the pH value. Thus, indicating that the convolution process could extract the representative information without the need to swap the color system.

For the models using as input the RGB information and all the components of the HSV, the hyperparameters of the activation\_functions were optimized in an initial exploratory step, in which it was found that the best activation function for the convolution layer was the ReLU, being applied in all the models shown on Table 6. Another hyperparameter optimized in the initial search was the optimization algorithm, in which the “adam” had the best results and was the one used to build all the CNN models.

Since the CNN models created using the individual components of the HSV did not present the expected performance, other kinds of activation functions were explored for the convolution layers. Still, no significant improvements were noticed, being the ReLU type the most common one applied in the best models, as presented in Table 4.7.

The hyperparameter `filtr_size` for these best models was commonly equal to eight, one of the lowest values tested, reducing the number of parameters of the models. Another interesting result was the number of layers of the best models (*n\_layers*) indicates that the best ones are formed by three or four convolution layers. The fact that the CNNs were not too deep also gives a small number of parameters for the models.

#### 4.1.3.1.2

##### SVM models

The hyperparameters of the five best models are shown in Table 4.8, along with the performance parameters for the validation and test groups, since all of them presented accuracy equal to 1 for the training group. The results for all trained SVM models are available in Appendix D in Table D1. It was observed that the best performance was obtained using the kernel of the *polynomial* form with the lowest degrees. The best way to determine the parameter gamma was by using their dependency on the number of classes. Also, the best methodology to approach this multiclass problem was the OvO, present in the three best models.

All five models presented high values for all the performance parameters, higher than 90% for all groups. However, the values for the test group were higher than the ones for the validation group, which could indicate that models could have a problem with overfitting or that the division of the unbalanced dataset could result in a problem during the training process.

Table 4.8. SVM topologies for the best models

Model ID	Hyperparameters						Validation group			Test group		
	Dec_func_shape	C	Kernel	degree	gamma	training time (s)	PR	RC	ACC	PR	RC	ACC
SVM_1	OvO	0.01	poly	1	Auto	73.63	0.9434	0.9471	0.9362	0.9815	0.9630	0.9787
SVM_2	OvO	0.02	poly	2	Auto	67.54	0.9352	0.9378	0.9362	0.9815	0.9630	0.9787
SVM_3	OvO	0.03	poly	3	Auto	65.69	0.9352	0.9378	0.9362	0.9815	0.9630	0.9787
SVM_4	OvR	0.04	poly	1	Auto	78.13	0.9434	0.9471	0.9362	0.9815	0.9630	0.9787
SVM_5	OvR	0.05	poly	2	Auto	68.44	0.9352	0.9378	0.9362	0.9815	0.9630	0.9787

#### 4.1.3.1.3

##### DT models

Table 4.9 shows the results of the five best models according to the performance parameters and their respective hyperparameters. The results for all trained DT models are available in Appendix D in Table D2. From the results, it was possible to observe that the “*entropy*” were the best criteria to measure the split quality, being applied in all five models. The other hyperparameter that can be noticed is the *max\_leaf\_nodes* equal to 10, which appears in three of the five models shown that is concordant with the number of pH classes, which are equal to 9. Another interesting result is that the DT models demanded a longer training time than the SVM ones, even though they were simple models. Analyzing the performance parameters can be observed the same problems that were pointed out for the SVM models, and the ACC values are lower than the ones obtained for those models.



Table 4.9: DT topologies for the best models

Model ID	Hyperparameters					Validation group			Test group		
	Crit	max_depth	max_leaf_nodes	min_samples_leaf	training time (s)	PR	RC	ACC	PR	RC	ACC
DT_1	<i>entropy</i>	None	10	7	198.70	0.8977	0.8479	0.8723	0.9397	0.9434	0.9362
DT_2	<i>entropy</i>	7	None	1	224.67	0.8825	0.8405	0.8511	0.9139	0.9249	0.9149
DT_3	<i>entropy</i>	None	20	1	222.43	0.8726	0.8442	0.8511	0.9212	0.9063	0.9149
DT_4	<i>entropy</i>	11	10	10	188.88	0.8636	0.8479	0.8511	0.9212	0.9063	0.9149
DT_5	<i>entropy</i>	None	10	5	206.46	0.8852	0.8442	0.8511	0.9119	0.9249	0.9149

## 4.3.2

### Evaluation of the prediction models

#### 4.3.2.1

##### CNN prediction models

Table 4.10 shows the performance values, and the hyperparameters of the best five models of each kind of input tested. These models had  $R^2$  and  $RMSE$  as the performance parameters evaluated, where the lowest value indicated the best response for all of them. As observed in the classification models, the best results were obtained with the Input 2 type. The results also showed that these models also had a low number of convolutional layers and neurons in the dense layers.

Different from the strategy for the classification of CNN models, the prediction models have only one neuron in the output layer, given the predicted pH value. In the hyperparameters, exploration for the prediction models tested different types of activation functions for all input types. Although the ReLU function type was present in several topologies, it did not have the same predominance observed in the classificatory models. The batch size use to train the all best models of the Input 1 was equal to 12, while for the models using Input 2 and Input 3 the best results were obtain with batch sizes of 8 and 4. For all the prediction models, “adam” was also applied as the optimizer algorithm.

Also, in the prediction models case, the results for the ones using RGB values had a better performance than those using the HSV information, and again Input 6 showed the worst results.

For the prediction models, the best performance results were often obtained when the CNN had three or four convolution layers in its topology. However, some models with only two layers appeared among the best ones.

Table 4.10: CNN prediction models topologies for the best models

models ID	Input variables	Hyperparameters													Training time (s)	RMSE (test)	R <sup>2</sup> (test)
		1	2	3	4	5	6	7	8	9	10	11	12	13			
CNN_pred_1	Input 1	12	99	8	8	1	5	0.005	3	0.25	100	60	relu	relu	277	0.316	0.982
CNN_pred_2		12	106	8	8	1	3	0.0005	3	0.3	50	50	linear	linear	251	0.438	0.978
CNN_pred_3		12	104	8	8	3	5	0.0005	3	0.25	90	50	tanh	relu	359	0.481	0.971
CNN_pred_4		12	101	8	8	3	3	0.001	4	0.3	120	50	relu	relu	334	0.636	0.942
CNN_pred_5		12	90	4	4	1	1	0.01	4	0.25	40	20	linear	relu	199	0.643	0.940
CNN_pred_6	Input 2	8	86	8	8	3	3	0.0005	4	0.3	60	40	tanh	tanh	46	0.150	0.993
CNN_pred_7		8	102	8	8	3	1	0.0005	3	0.15	60	50	sigmoid	relu	53	0.184	0.994
CNN_pred_8		4	113	8	4	3	1	0.005	2	0.25	70	40	relu	linear	58	0.200	0.990
CNN_pred_9		4	118	8	8	3	1	0.001	3	0.15	100	30	relu	linear	67	0.219	0.995
CNN_pred_10		8	104	8	8	1	5	0.0005	4	0.2	100	60	relu	sigmoid	56	0.191	0.994
CNN_pred_11	Input 3	4	118	8	8	1	3	0.005	2	0.2	100	40	linear	relu	78	0.480	0.979
CNN_pred_12		8	120	8	8	3	3	0.005	2	0.05	120	20	relu	linear	58	0.486	0.971
CNN_pred_13		8	119	8	8	5	5	0.005	2	0.05	120	40	linear	relu	68	0.463	0.971
CNN_pred_14		4	118	8	8	3	3	0.01	2	0.1	100	40	relu	relu	77	0.499	0.966
CNN_pred_15		12	106	8	8	1	3	0.01	2	0.15	100	20	linear	relu	48	0.487	0.967
CNN_pred_16	Input 4	12	102	8	8	5	1	0.01	2	0.05	100	40	tanh	tanh	41	2.740	-0.005
CNN_pred_17		12	116	8	8	5	1	0.01	2	0.2	70	60	sigmoid	sigmoid	35	2.740	-0.004
CNN_pred_18		12	93	4	8	5	1	0.0001	2	0.05	70	30	linear	sigmoid	36	2.740	-0.005
CNN_pred_19		12	114	8	8	3	5	0.01	2	0.25	70	60	sigmoid	sigmoid	43	2.740	-0.005
CNN_pred_20		12	113	8	8	3	3	0.01	2	0.25	120	60	sigmoid	tanh	42	2.740	-0.005
CNN_pred_21	Input 5	12	111	8	8	5	5	0.01	3	0.2	70	60	relu	linear	62	0.854	0.921

CNN_pred_22		12	98	4	8	5	3	0.005	2	0.3	90	10	linear	relu	38	0.876	0.920
CNN_pred_23		8	92	4	8	5	5	0.01	2	0.15	120	60	relu	linear	43	0.881	0.919
CNN_pred_24		8	96	4	4	5	1	0.005	2	0.3	90	60	linear	linear	42	0.901	0.863
CNN_pred_25		12	111	4	8	5	3	0.01	2	0.2	70	60	linear	relu	41	0.904	0.975
CNN_pred_26	Input 6	8	120	8	8	5	5	0.01	3	0.1	80	50	tanh	relu	93	0.446	0.972
CNN_pred_27		4	93	8	8	3	3	0.005	3	0.2	90	60	relu	relu	63	0.815	0.908
CNN_pred_28		4	105	4	8	3	3	0.01	2	0.15	90	20	linear	relu	59	0.830	0.917
CNN_pred_29		8	104	4	8	1	3	0.01	3	0.1	40	60	tanh	relu	63	0.965	0.885
CNN_pred_30		4	112	8	8	3	3	0.01	4	0.1	90	50	sigmoid	relu	94	0.976	0.879

a - hyperparameters: 1 – batch\_size; 2- epochs; 3 – filter\_size 1; 4 – filter\_size\_2; 5 - kernel\_size\_1; 6 - kernel\_size\_2; 7 - learning rate, 8 - n\_layers; 9 – p\_dropout; 10 - size\_dense\_1; 11 – size\_sense\_2; 12 – activation\_function\_1; 13 – activation\_function\_2

### 4.1.3.3

#### Validation of the models

Once the best models were selected, a natural next step was to test their applicability to determine the pH value, using the two scenarios in which they could be exposed during their use. For that, the classification and prediction CNN models with Input 2 images were chosen since they presented the best performance results, respectively models CNN\_class\_6-10 and CNN\_pred\_6-10. DT and SVM models were also tested to compare their efficiency with the classification CNN ones.

#### 4.1.3.3.1

##### Case study: Titration curve of strong acid with a strong base

The first case studied was the already known strong acid – strong base titration. For that, the images of six experiments were used (four at atmospheric pressure and two pressurized at 6 MPa) to test the best five classification models of each technique CNN, SVM, and DT, and the five CNN prediction models.

Figure 4.9 shows a comparison between the ACC values of each CNN classificatory model tested, presenting the mean, low and high values of the performance parameter. The average ACC values are very similar for all the models, ranging from 88% to 92%. In general, the performance of the CNN models was lower than expected, although all the models had no ACC values lower than 80%, except for the CNN\_class\_7.

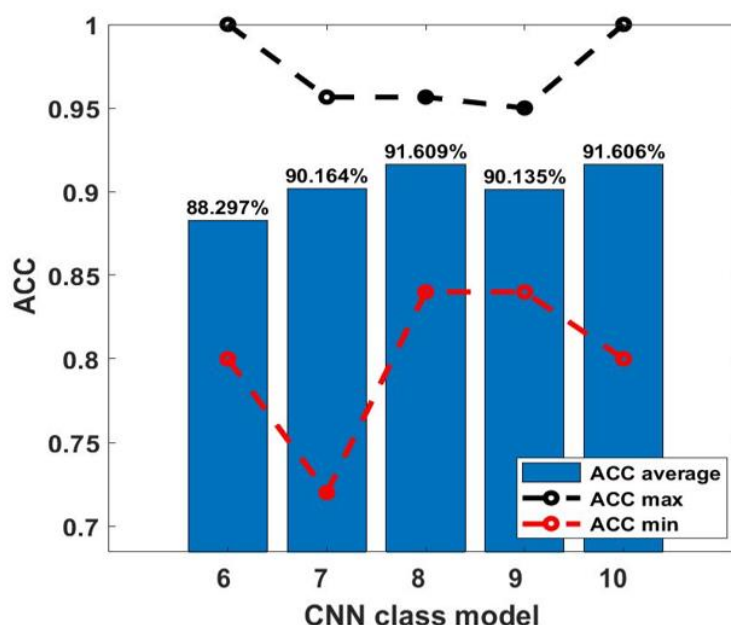


Figure 4.9: Accuracy values for the CNN classification models in the

### neutralization curve scenario

Figure 4.10A shows the ACC values for the SVM model tested, and Figure 4.10B shows the values for the DT models, presenting the mean, low and high values of the performance parameter. Both strategies present a worst performance than the CNN classification models, with average ACC values, lowers than 75%. SVM models show a better performance than the DT models, but both strategies had models with a big range of ACC values.

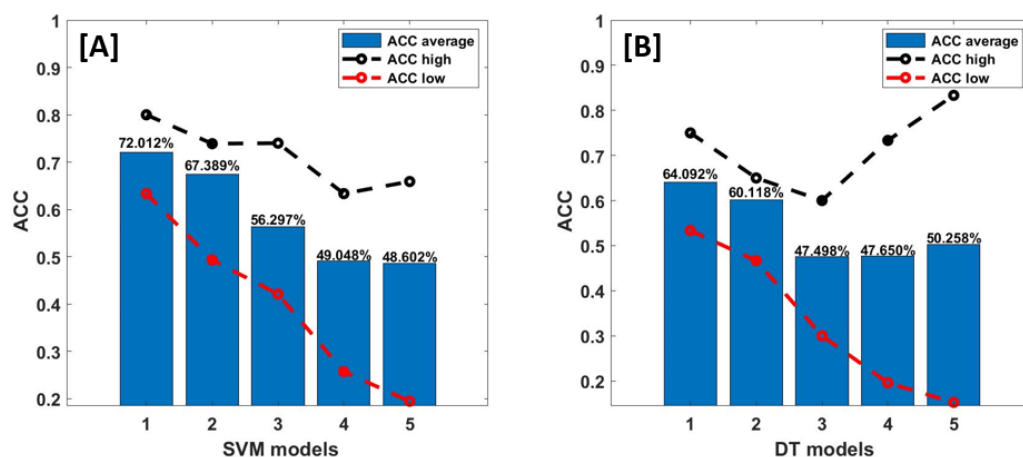


Figure 4.10: Accuracy values in the neutralization curve scenario for the classification models: SVM (A) and DT (B).

The performance parameters of the prediction CNN models are presented in Figures 4.11A-B, respectively, the parameters RMSE and  $R^2$ . Comparing the parameters' results, it is observed that the five models showed a good fit with the experimental values, presenting  $R^2$  values high than 90 %, in which the best model was the CNN\_pred\_6 with the average  $R^2$  and RMSE, respectively, equal to 94.96% and 0.8198. The result indicates that the prediction models had a better performance than the classification ones.

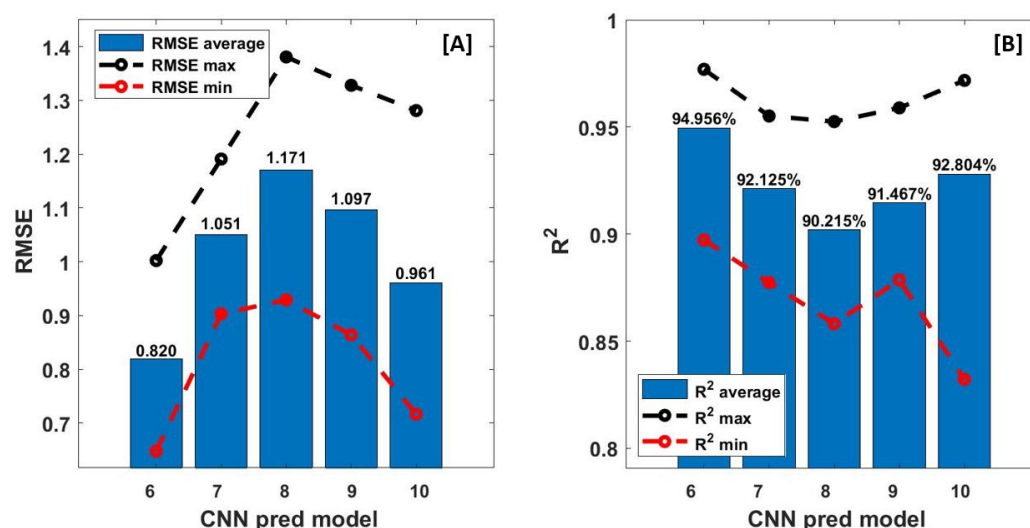


Figure 4.11: RMSE (A) and  $R^2$  (B) values for the CNN predict models in the neutralization curve scenario.

#### 4.1.3.3.2

##### Case study: CO<sub>2</sub>-H<sub>2</sub>O equilibrium systems

In this second study case, the analyzed scenario was the change in the pH of the aqueous solution due to the dissolution of the CO<sub>2</sub> in the solution due to the pressure applied in the reactor. The ten CNN models, five classificatory and five predictions, and the DT and SVM models were evaluated using data obtained from eight experiments, in which the CO<sub>2</sub> pressure in the system varied between 0.1 MPa and 5 MPa.

The performance parameter analyzed for the CNN classification models was the ACC, shown in Figure 4.12. As observed in the first case study, the models presented similar average ACC values in the range of 81% to 85%. Although these values were not distant from those found in the previous case, when the lowest values obtained were analyzed, ACC values were lower than 71% for all models, indicating that they could classify with a lower precision in some of the experimental situations. This worsening in the results was expected since the CO<sub>2</sub>-H<sub>2</sub>O equilibrium leads to low pH values with the increase of the pressure, and the differentiation between the pH classes 2, 3, and 4 was one of the challenges during the models' development.

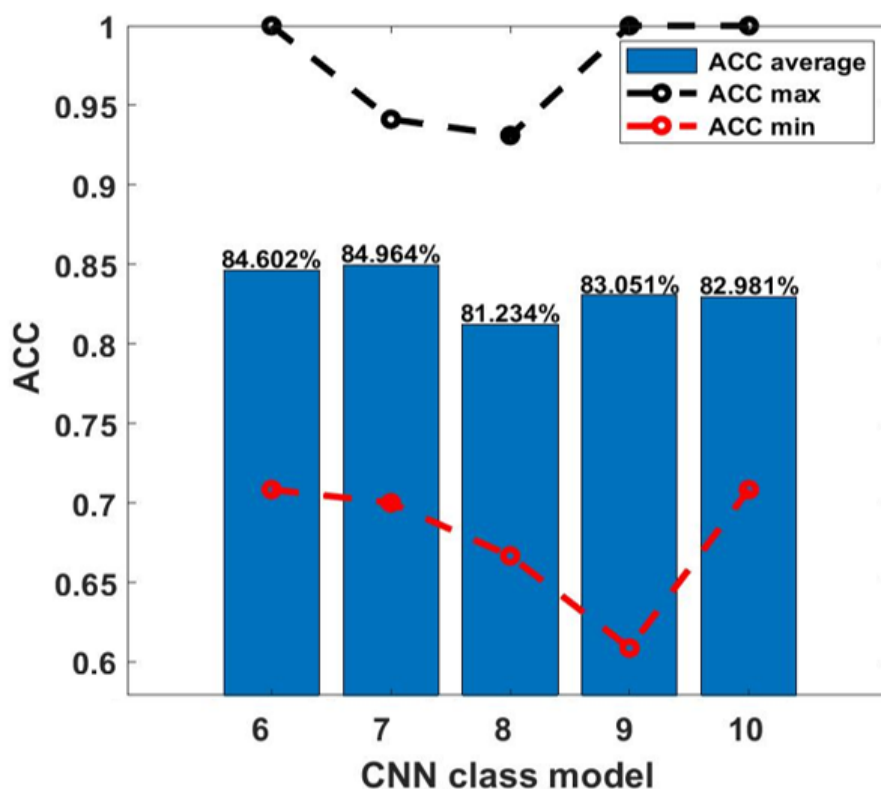


Figure 4.12: Accuracy values for the CNN classification models in the equilibrium CO<sub>2</sub>-H<sub>2</sub>O system scenario

The ACC values for the SVM and DT models are shown in Figure 4.13A-B, presenting the mean, low and high values of the performance parameter. Again, the models of both strategies had a worse performance than the CNN models. However, for this case study, the average ACC values were lower than 50%, indicating that these models are not to be applied in this experimental condition.

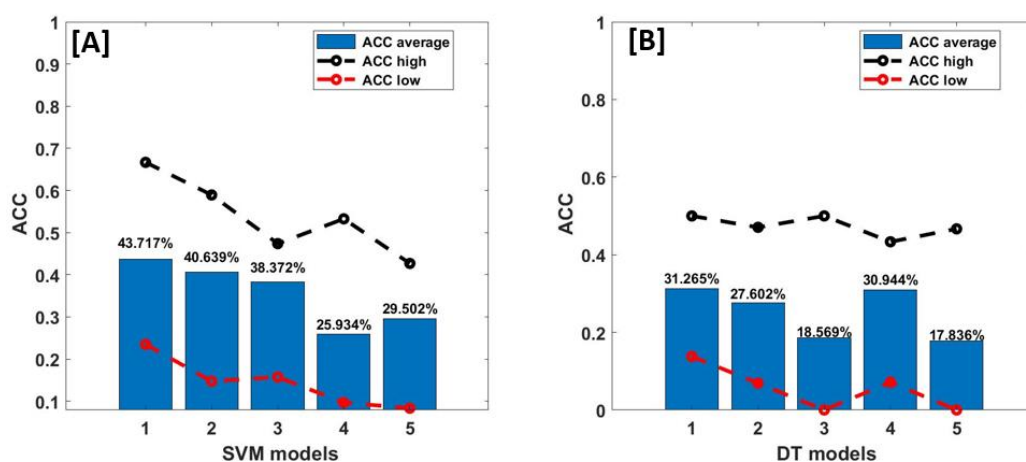


Figure 4.13: Accuracy values in the equilibrium CO<sub>2</sub>-H<sub>2</sub>O system scenario for the classification models: SVM (A) and DT (B).



Figure 14A-B shows the results of the performance parameters RMSE and  $R^2$ , respectively, for the prediction CNN models. As observed with the classification models, the prediction models also presented a significant worsening in their performance. However, the models had different behaviors, with the average  $R^2$  values varying between 63% and 87%. The model CNN\_pred\_8 presented the worst performance with a low  $R^2$  value equal to 37.87%. Otherwise, the model CNN\_pred\_9 showed the most promising one, with its lowest  $R^2$  value equaling 80.13%.

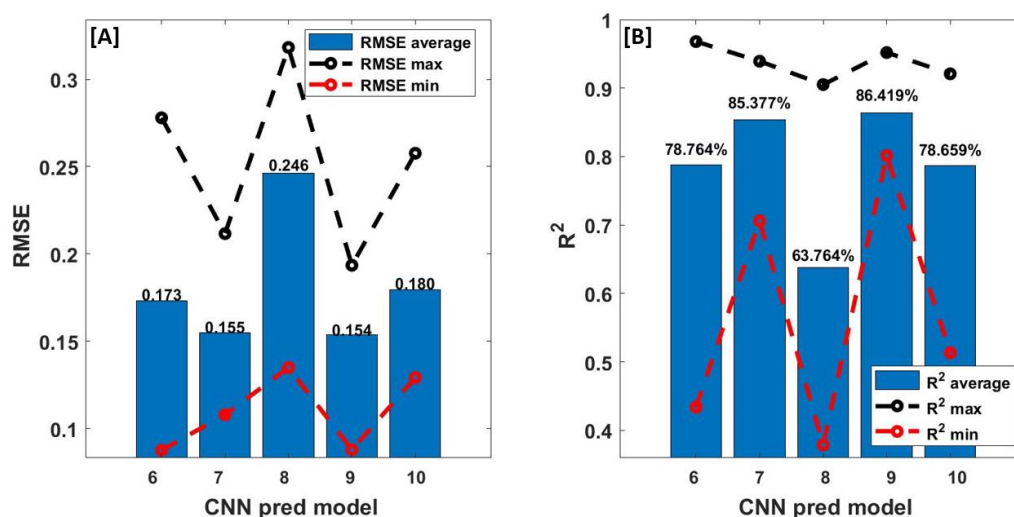


Figure 4.14: RMSE (A) and  $R^2$  (B) values for the CNN predict models in the equilibrium  $\text{CO}_2\text{-H}_2\text{O}$  system scenario

In this work was developed different types of classification and regression models to determine the pH values in the range of 2-10 using different ML techniques. It is the first step in developing a soft sensor to be applied for real-time monitoring situation with pressurized system, such as the NGS (Nitrogen generation system) process using a submersion probe to acquire the images.

#### 4.1.4.

#### Conclusions

This study developed models to determine the pH values in atmospheric and

pressurized systems (up to 6 MPa) using images from the reactor vessel acquired using low-cost methods. For that purpose, it was explored three different types of ML strategies (CNN, SVM, and DT) for the development of the classification models, which classify the aqueous solution pH into one of the nine classes. Also, regression models using the CNN strategy were developed to predict the pH values in the range of 2-10. The best five models of explored strategies were tested in two scenarios to verify their application in other operational situations. The best classification model was the CNN one, with both the buffer solutions and the cases of study datasets. Although, the best performance was obtained by the prediction CNN models, highlighting the model CNN\_pred\_9, which presents  $R^2$  values higher than 80% for all tested datasets. Thus, the regression CNN models are the most interesting strategy to continue developing the soft sensor to determine the pH values in high pressure systems.

## ASSOCIATED CONTENT

### Supplementary materials

The performance results of all DT and SVM classification models are available in the supplementary material in Appendix D, respectively in Table D1 and Table D2.

## AUTHOR INFORMATION

### Corresponding Author

\*Email corresponding author: bsantos@puc-rio.br

ORCID: 0000-0001-8755-7749

### Author Contributions

All authors have contributed to the writing of this manuscript and have approved its final version.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENT

The authors acknowledge the financial support by Conselho Nacional de

Desenvolvimento Científico e Tecnológico (CNPq) and Petrobras in the development of this work. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

## 4.5

### References

- ASGHER, U., KHALIL, K., AYAZ, Y., *et al.* "Classification of Mental Workload (MWL) using Support Vector Machines (SVM) and Convolutional Neural Networks (CNN)", **2020 3rd International Conference on Computing, Mathematics and Engineering Technologies: Idea to Innovation for Building the Knowledge Economy, iCoMET 2020**, p. 1–6, 2020. DOI: 10.1109/iCoMET48670.2020.9073799.
- BIASES, W. and. **Weights & Biases**. Disponível em: <https://docs.wandb.ai/>. Acesso em: 10 jul. 2022.
- BIEWALD, L. **Experiment Tracking with Weights and Biases**. Disponível em: <https://www.wandb.com/>. Acessado em: 10 jul. 2022.
- BOUWMANS, T., JAVED, S., SULTANA, M., *et al.* "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation", **Neural Networks**, v. 117, p. 8–66, 2019. DOI: 10.1016/j.neunet.2019.04.024.
- BYCHKOV, A. Y., BÉNÉZETH, P., POKROVSKY, O. S., *et al.* "Experimental determination of calcite solubility and the stability of aqueous Ca– and Na–carbonate and –bicarbonate complexes at 100–160 °C and 1–50 bar pCO<sub>2</sub> using in situ pH measurements", **Geochimica et Cosmochimica Acta**, v. 290, p. 352–365, 2020. DOI: 10.1016/j.gca.2020.09.004.
- CAELEN, O. "A Bayesian interpretation of the confusion matrix", **Annals of Mathematics and Artificial Intelligence**, v. 81, n. 3–4, p. 429–450, 2017. DOI: 10.1007/s10472-017-9564-8.
- CAPEL-CUEVAS, S., CUÉLLAR, M. P., DE ORBE-PAYÁ, I., *et al.* "Full-range optical pH sensor array based on neural networks", **Microchemical Journal**, v. 97, n. 2, p. 225–233, 2011. DOI: 10.1016/j.microc.2010.09.008.
- CHAUHAN, V. K., DAHIYA, K., SHARMA, A. "Problem formulations and solvers in linear SVM: a review", **Artificial Intelligence Review**, v. 52, n. 2, p. 803–855, 2019. DOI: 10.1007/s10462-018-9614-6. Disponível em: <https://doi.org/10.1007/s10462-018-9614-6>.
- CORTES, C., VAPNIK, V. "Support-Vector Networks", **Machine Learning**, v. 20, p. 273–297, 1995.
- CROLET, J. L., BONIS, M. R. "pH MEASUREMENTS IN AQUEOUS CO<sub>2</sub> SOLUTIONS UNDER HIGH PRESSURE AND TEMPERATURE.", **Corrosion**, v. 39, n. 2, p. 39–46, 1983. DOI: 10.5006/1.3580813.
- DE OLIVEIRA, A. V. B., ORTIZ, R. W. P., KARTNALLER, V., *et al.* "Real-Time Measurement of pH in Atmospheric and Pressurized Systems Using a Low-Cost Image Analysis Method", **IEEE Sensors Journal**, v. 19, n. 23, p. 10991–10998, 2019. DOI: 10.1109/JSEN.2019.2936442.
- DING, S., ZHAO, X., ZHANG, J., *et al.* "A review on multi-class TWSVM", **Artificial Intelligence Review**, v. 52, n. 2, p. 775–801, 2019. DOI: 10.1007/s10462-017-9586-y.

DIXIT, Y., AL-SARAYREH, M., CRAIGIE, C. R., *et al.* "A global calibration model for prediction of intramuscular fat and pH in red meat using hyperspectral imaging", **Meat Science**, v. 181, n. November, p. 108405, 2021. DOI: 10.1016/j.meatsci.2020.108405.

FUNATSU, K. "Process Control and Soft Sensors", **Applied Chemoinformatics**, p. 571–584, 2018. DOI: 10.1002/9783527806539.ch13.

GEURTS, P., IRRTHUM, A., WEHENKEL, L. "Supervised learning with decision tree-based methods in computational and systems biology", **Molecular BioSystems**, v. 5, n. 12, p. 1593–1605, 2009. DOI: 10.1039/b907946g.

HAGHBIN, M., SHARAFATI, A., MOTTA, D., *et al.* "Applications of soft computing models for predicting sea surface temperature: a comprehensive review and assessment", **Progress in Earth and Planetary Science**, v. 8, n. 1, 2021. DOI: 10.1186/s40645-020-00400-9.

HASNAIN, M., PASHA, M. F., GHANI, I., *et al.* "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking", **IEEE Access**, v. 8, p. 90847–90861, 2020. DOI: 10.1109/ACCESS.2020.2994222.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Second ed. New York, Springer, 2009.

KADLEC, P., GABRYS, B., STRANDT, S. "Data-driven Soft Sensors in the process industry", **Computers and Chemical Engineering**, v. 33, n. 4, p. 795–814, 2009. DOI: 10.1016/j.compchemeng.2008.12.012.

KARIMPOULI, S., TAHMASEBI, P., RAMANDI, H. L. "A review of experimental and numerical modeling of digital coalbed methane: Imaging, segmentation, fracture modeling and permeability prediction", **International Journal of Coal Geology**, v. 228, n. March, p. 103552, 2020. DOI: 10.1016/j.coal.2020.103552.

KHAN, M. I., MUKHERJEE, K., SHOUKAT, R., *et al.* "A review on pH sensitive materials for sensors and detection methods", **Microsystem Technologies**, v. 23, n. 10, p. 4391–4404, 2017. DOI: 10.1007/s00542-017-3495-5.

LEI, Y., CHEN, X., MIN, M., *et al.* "A semi-supervised Laplacian extreme learning machine and feature fusion with CNN for industrial superheat identification", **Neurocomputing**, v. 381, p. 186–195, 2020. DOI: 10.1016/j.neucom.2019.11.012. Disponível em: <https://doi.org/10.1016/j.neucom.2019.11.012>.

LEMMER, A., MERKLE, W., BAER, K., *et al.* "Effects of high-pressure anaerobic digestion up to 30 bar on pH-value, production kinetics and specific methane yield", **Energy**, v. 138, p. 659–667, 2017. DOI: 10.1016/j.energy.2017.07.095.

LIU, W., WANG, Z., LIU, X., *et al.* "A survey of deep neural network architectures and their applications", **Neurocomputing**, v. 234, n. October 2016, p. 11–26, 2017. DOI: 10.1016/j.neucom.2016.12.038.

LORENA, A. C., DE CARVALHO, A. C. P. L. F. "An Introduction to Support Vector Machines", **Revista de Informática Teórica e Aplicada**, v. 14, p. 43–67, 2007. DOI: 10.22456/2175-2745.5690.

NAMUDURI, S., NARAYANAN, B. N., DAVULURU, V. S. P., *et al.* "Review—Deep Learning Methods for Sensor Based Predictive Maintenance and Future Perspectives for Electrochemical Sensors", **Journal of The Electrochemical Society**, v. 167, n. 3, p. 037552, 2020. DOI: 10.1149/1945-7111/ab67a8. .

NUMPY. **numpy.round**. . Disponível em: [https://numpy.org/doc/stable/reference/generated/numpy.round\\_.html](https://numpy.org/doc/stable/reference/generated/numpy.round_.html). Acesso em: 30 mar. 2022.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., *et al.* "Scikit-learn: Machine Learning in Python", **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PENG, Y., LIAO, M., DENG, H., *et al.* "CNN-SVM: A classification method for fruit fL image with the complex background", **IET Cyber-Physical Systems: Theory and Applications**, v. 5, n. 2, p. 181–185, 2020. DOI: 10.1049/iet-cps.2019.0069.

POERIO, D. V., BROWN, S. D. "Highly-overlapped, recursive partial least squares soft sensor with state partitioning via local variable selection", **Chemometrics and Intelligent Laboratory Systems**, v. 175, n. December 2017, p. 104–115, 2018. DOI: 10.1016/j.chemolab.2018.02.006.

PRIYAM, A., ABHIJEET, GUPTA, R., *et al.* "Comparative Analysis of Decision Tree Classification Algorithms", **International Journal of Current Engineering and Technology**, v. 3, p. 334–337, 2013. .

RHYS, H. I. **Machine Learning with R, the tidyverse, and mlr**. [S.l.], Manning Publications, 2020.

RIZKIN, B. A., POPOVICH, K., HARTMAN, R. L. "Artificial Neural Network control of thermoelectrically-cooled microfluidics using computer vision based on IR thermography", **Computers and Chemical Engineering**, v. 121, p. 584–593, 2019. DOI: 10.1016/j.compchemeng.2018.11.016.

RUUSKA, S., HÄMÄLÄINEN, W., KAJAVA, S., *et al.* "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle", **Behavioural Processes**, v. 148, n. January, p. 56–62, 2018. DOI: 10.1016/j.beproc.2018.01.004.

SAMARANAYAKE, C. P., SASTRY, S. K. "In-situ pH measurement of selected liquid foods under high pressure", **Innovative Food Science and Emerging Technologies**, v. 17, p. 22–26, 2013. DOI: 10.1016/j.ifset.2012.09.006. .

SANSANA, J., JOSWIAK, M. N., CASTILLO, I., *et al.* "Recent trends on hybrid modeling for Industry 4.0", **Computers and Chemical Engineering**, v. 151, p. 107365, 2021. DOI: 10.1016/j.compchemeng.2021.107365.

SCIKIT-LEARN. **1.10. Decision Trees**. [S.d.]. Disponível em: <https://scikit-learn.org/stable/modules/tree.html>. Acesso em: 10 jul. 2022a.

SCIKIT-LEARN. **1.4. Support Vector Machines**. [S.d.]. Disponível em: <https://scikit-learn.org/stable/modules/svm.html>. Acesso em: 10 jul. 2022b.

SCIKIT-LEARN. **3.3. Metrics and scoring: quantifying the quality of predictions**. [S.d.]. Disponível em: <https://scikit-learn.org/stable/modules/metrics.html>.

learn.org/stable/modules/model\_evaluation.html. Acesso em: 10 jul. 2022c.

SHANG, C., YANG, F., HUANG, D., *et al.* "Data-driven soft sensor development based on deep learning technique", **Journal of Process Control**, v. 24, n. 3, p. 223–233, 2014. DOI: 10.1016/j.jprocont.2014.01.012.

SHEN, S., LU, H., SADOUGHI, M., *et al.* "A physics-informed deep learning approach for bearing fault detection", **Engineering Applications of Artificial Intelligence**, v. 103, n. May, p. 104295, 2021. DOI: 10.1016/j.engappai.2021.104295.

SUN, Q., GE, Z. "A Survey on Deep Learning for Data-driven Soft Sensors", **IEEE Transactions on Industrial Informatics**, v. 3203, n. c, 2021. DOI: 10.1109/TII.2021.3053128. .

TANGIRALA, S. "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm", **International Journal of Advanced Computer Science and Applications**, v. 11, n. 2, p. 612–619, 2020. DOI: 10.14569/ijacsa.2020.0110277.

WADOUX, A. M. J. C. "Using deep learning for multivariate mapping of soil with quantified uncertainty", **Geoderma**, v. 351, n. November 2018, p. 59–70, 2019. DOI: 10.1016/j.geoderma.2019.05.012.

YAN, W., TANG, D., LIN, Y. "A data-driven soft sensor modeling method based on deep learning and its application", **IEEE Transactions on Industrial Electronics**, v. 64, n. 5, p. 4237–4245, 2017. DOI: 10.1109/TIE.2016.2622668. .

YAO, G., LEI, T., ZHONG, J. "A review of Convolutional-Neural-Network-based action recognition", **Pattern Recognition Letters**, v. 118, p. 14–22, 2019. DOI: 10.1016/j.patrec.2018.05.018.

YUAN, X., QI, S., SHARDT, Y., *et al.* "Soft sensor model for dynamic processes based on multichannel convolutional neural network", **Chemometrics and Intelligent Laboratory Systems**, v. 203, n. January, p. 104050, 2020. DOI: 10.1016/j.chemolab.2020.104050.

ZAN, T., LIU, Z., WANG, H., *et al.* "Control chart pattern recognition using the convolutional neural network", **Journal of Intelligent Manufacturing**, v. 31, n. 3, p. 703–716, 2020. DOI: 10.1007/s10845-019-01473-0.

## 5 Conclusions

In this work, several models were developed using AI, intended to be applied directly or indirectly to oil and gas production problems related to flow assurance. In the first part of the study, it was possible to develop MLP models to predict the scaling process in a tube with a dynamic flow using the differential pressure ( $\Delta P$ ) to monitor this process. The models were created using the six process variables as inputs. The prediction of the  $\Delta P$  in two-time horizons (one step ahead ( $\Delta P_{(t+1)}$ ) and five steps ahead ( $\Delta P_{(t+5)}$ )) was explored as output variable individually. The best model for variable  $\Delta P_{(t+1)}$  was the one with the topology `logsig_7_purelin_1_trainbr`, with  $R^2$  over 99.3%. Otherwise, for the  $\Delta P_{(t+5)}$  the best model with the best overall performance has the topology `logsig_6_purelin_1_trainlm`, presenting an  $R^2$  between 79.7% and 96.4%.

In the second part, the creation of classification and prediction models, using different AI techniques (CNN, SVM, and DT), to determine the pH values in the atmospheric and pressurized system from image analysis was accomplished. The best classification model was `CNN_clas_RGB_crop_model_4` presenting accuracy values equal to 97.87 % for the test group. The best prediction model was `CNN_pred_RGB_crop_model_4`, which also uses Input 2, having an  $R^2$  value higher than 80% in all tested scenarios.

In conclusion, the models developed during this study presented high levels in their respective performance parameters, indicating that they are exciting candidates that keep being studied and developed to be applied for the tasks of monitoring and controlling in the oil and gas industry.



## 6

### Suggestions for future works

As a first suggestion, it would be interesting to start by the suggestion that the MLP models developed to predict  $\Delta P$  of the tubes in both time horizons ( $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$ ) were tested on real-time experiments as validation of their performance and their applicability. Also another interesting path is to use this concept and create new models to predict the  $\Delta P$  during the scale formation using a more extensive database that could amplify the application range or include more variables to create a more robust model that could be applied in several scenarios.

Regarding the models developed to determine the pH base in imaging analysis, the first suggestion is to test the performance in real-time experiments. These tests could also be used to evaluate the full time to determine the pH value, from the image capture to the models' response, which could be an important parameter in case this was used in a controlling strategy in future works. Also, it would be possible to develop a control loop based on this soft sensor. Another point to be explored is to test the viability of using this form of detection on the NGS due to the bubbles obtained during the process that could disturb the results.

**A**

**Published article: Development of MLP artificial neural network models for the simulation of CaCO<sub>3</sub> scale formation process in the presence of monoethylene glycol (MEG) in a dynamic tube blocking test (TBT) equipment**

# Development of Artificial Neural Network Models for the Simulation of a $\text{CaCO}_3$ Scale Formation Process in the Presence of Monoethylene Glycol (MEG) in Dynamic Tube Blocking Test Equipment

Bruno X. Ferreira, Carlos R. Hall Barbosa, João Cajaiba, Vinicius Kartnaller, and Brunno F. Santos\*



Cite This: *Energy Fuels* 2022, 36, 2288–2299



Read Online

ACCESS |



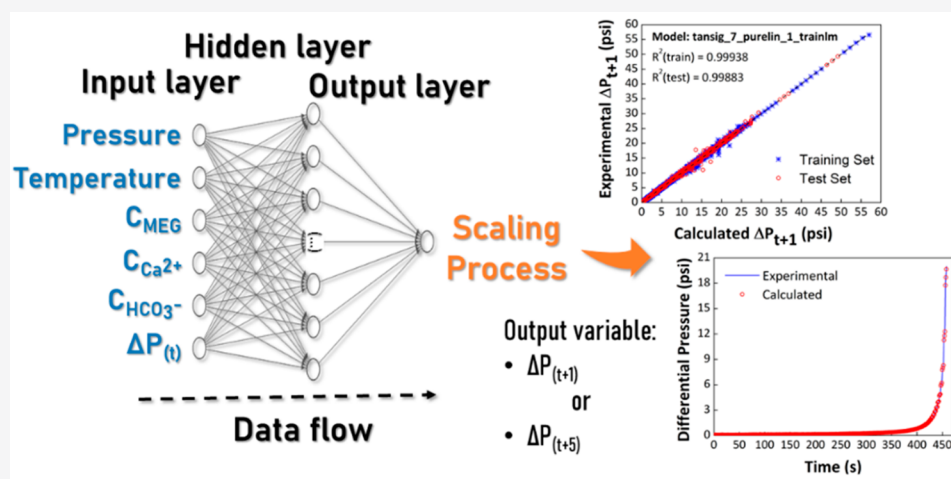
Metrics & More



Article Recommendations



Supporting Information



**ABSTRACT:** The precipitation of gas hydrate and inorganic salts (scale) during oil and gas production represents a significant flow assurance hindrance for the industry. Chemical inhibitors can prevent the fouling process, but specific inhibitors to address a problem could result in synergistic or adverse effects. Simulations in tubes and pipelines are necessary to understand these behaviors by assessing the scaling tendency of the water. The primary objective of this study was to create models using an artificial neural network (ANN) of the multilayer perceptron (MLP) type for the simulation of the calcium carbonate scaling formation process in the presence of monoethylene glycol (MEG), a typical gas hydrate inhibitor. A database was obtained from 38 tube blocking test (TBT) experiments with different conditions. The models were developed using MATLAB R2020a, splitting the database into two groups on the ratio of 70:30, respectively, train and test ones, preserving the time dependency of the differential pressure ( $\Delta P$ ) data. The ANNs were created using six inputs (temperature, pressure, calcium and bicarbonate concentrations, MEG concentration, and  $\Delta P$  measured at a selected time) and one output ( $\Delta P$  measured at a later time). The goal was to explore how monitoring the conditions in a pipeline can predict the evolution of the scaling process. We investigated two scenarios for the  $\Delta P$  prediction: a near future (one step ahead) and a far future (five steps ahead). The MLP models demonstrated high performance, with an  $R^2$  higher than 92.9% for both training and test groups for both prediction horizons. Then, these models were tested with a second data group to evaluate their applicability to control the systems. The best models showed good scaling prediction, with  $R^2$  ranging from 80.0% to 99.9%. These results represent a promising step toward applying machine learning techniques to simulate and predict scaling tendencies in controlled pipelines.

## 1. INTRODUCTION

Flow assurance is a significant concern during oil and gas production and is achieved by guaranteeing that hydrocarbon production from wells is maintained without loss over time due to flow restrictions. During production, the oil–gas–water mixture undergoes drastic variations in operating conditions, such as temperature and pressure, so that the solubility of

**Received:** September 30, 2021

**Revised:** January 16, 2022

**Published:** January 31, 2022



ACS Publications

© 2022 American Chemical Society

2288

<https://doi.org/10.1021/acs.energyfuels.1c03364>  
Energy Fuels 2022, 36, 2288–2299

certain compounds can decrease, leading to the formation of deposits (fouling). This fouling may occur in pipelines and equipment and is generally caused by the formation of wax, gas hydrate, and scale (inorganic salts). These can require expensive and complex remediation processes and, in severe cases, production stoppage and well shutdown.<sup>1–3</sup> This problem is of great concern, especially for wells in the Brazilian presalt region located in ultradeep waters with mainly carbonaceous reservoir rocks, and can result in potential issues such as calcium carbonate and gas hydrate fouling.<sup>4,5</sup>

Gas hydrate originates from the crystallization of water molecules encapsulating small and light gas molecules (e.g., CO<sub>2</sub>, methane, and propane) under operating conditions with high pressure and low temperature, such as those found in deep and ultradeep water.<sup>6,7</sup> The most practical and economical method for preventing hydrate formation or others kinds of obstructions in lines (e.g., scales) is using chemical inhibitors.<sup>8–10</sup> Thermodynamic hydrate inhibitors (THIs) are typically injected into the production line to prevent the formation of gas hydrates. THIs consist of alcohols or glycols, such as methanol, triethylene glycol (TEG), and monoethylene glycol (MEG), and function by moving the equilibrium curve envelope toward lower temperature and higher pressure.<sup>11,12</sup>

Scale forms as a result of the deposition of inorganic salts precipitating from the supersaturated water. Their formation depends on several factors such as temperature, pressure, ion concentration, pH, and others.<sup>13</sup> Barium sulfate, strontium sulfate, and calcium carbonate are the most common types of scale found during oil and gas production.<sup>14,15</sup> However, calcium carbonate (CaCO<sub>3</sub>) formation is of greater concern since the water may be in equilibrium with carbonaceous rocks in the reservoir, leading to a significant number of bicarbonate ions dissolved in the water phase (eqs S1–S3, [Supporting Information](#)). The precipitation of CaCO<sub>3</sub> occurs as this fluid is produced and faces a pressure drop, which decreases the CO<sub>2</sub> solubility and increases pH, leading to precipitation (eq 4, [Supporting Information](#)).

There are dozens of different inhibitor types used for typical organic scale. There are three main classes of inhibitors: phosphate esters, phosphonates, and polymers. The first two classes act as chelators, sequestering the metals from solution, while the polymeric class achieves scale control through crystal distortion.

In 2002, the average cost due to scale formation was more than 1.4 billion dollars.<sup>16</sup> As a result, the market for scale inhibitors for the oil and gas industry continues to grow and currently represents millions of dollars annually. Market analyses predict further increases in these expenditures with a CAGRs (compound annual growth rates) of 5.5% and 6.9% for the scale and hydrate inhibitors markets, respectively.<sup>17–19</sup>

A concern with the use of inhibitors for production is the compatibility between the different inhibitors and other chemicals. Several studies have investigated these compatibilities, including the effects of the enhanced oil recovery (EOR) chemicals on scale inhibitors<sup>20</sup> and the interaction between scale inhibitors and hydrate inhibitors.<sup>21</sup> For example, Seiersten and Kundu<sup>22</sup> and Kartnaller et al.<sup>23</sup> studied the impact of MEG as a gas hydrate scale inhibitor, concluding that MEG serves as an inhibitor by increasing the scaling time. This result was unexpected because the presence of MEG in water increases ion activities. That behavior has been proposed to be connected to the high-energy bond between –OH groups and the CaCO<sub>3</sub> surface; this indicates that thermody-

namic hydrate inhibitors can also benefit wells experiencing calcium carbonate scale formation.

Understanding the interactions between inhibitors, water, and ions is essential for predicting the phase behavior during production and estimating the solid accumulation tendency in production lines. A common and well-known methodology to evaluate inhibitor efficiency is the dynamic tube blocking test (TBT). It is usually applied to verify a product's performance and minimum inhibitor concentration (MIC), allowing comparison with other commercially available products.<sup>24–26</sup> TBT experiments are also used to study inorganic salt morphologies<sup>27,28</sup> and develop scale formation models. However, it is difficult to predict how the scaling process will develop using flow and phase behavior models due to the system's complexity, the large number of variables, and some stochastic behavior. A previous work has attempted to model the scale formation in pipelines, specifically in TBT experiments, but only using physical models.<sup>29</sup> These models, based on the Darcy Weisbach equation for pressure loss in pipes and on a growth rate scale formation model, were successful in fitting the TBT experiments curves, enabling an estimation on how fast the process was happening. However, the model was learning only the information regarding that specific experiment and not acquiring information for predicting the behavior of the system.

Other studies have explored the use of artificial neural networks (ANNs) and other machine learning algorithms to create new models since they do not demand an understanding of the scale formation mechanism, only requiring a "black-box" model. These models were able to predict the thermodynamics related to the calcium carbonate precipitation (saturation ratio of the solution) and its dissolution capacity.<sup>30,31</sup> However, literature still lacks kinetic modeling related to the scale formation process. Recently, Wang et al.<sup>32</sup> have developed an Elman neural network (ENN) with a genetic algorithm (GA) to predict calcium carbonate scale formation in shell and tube heat exchangers over time. They were able to successfully predict the fouling resistance as a function of conductivity, pH, and dissolved oxygen. Still, as far as the author's knowledge, no study relating scale formation and variables to simulate conditions during oil and gas production has been previously assessed.

In recent decades, different types of artificial intelligence (AI), such as ANN, GA, support vector machines (SVMs), the adaptive neuro-fuzzy inference system (ANFIS), least square support vector machine (LSSVM), principal component analysis (PCA), and the committee machine intelligent system (CMIS) have been applied to solve problems and challenges in several fields like nanofluids properties<sup>33–35</sup> and systems efficiency<sup>36,37</sup> and in the oil and gas industry, from the reservoir to production.<sup>38–40</sup> ANN was inspired by the neural arrangement of the human brain. It is easy to train and has tunable parameters and an adaptive structure, making it one of the most widely used machine learning techniques.<sup>41</sup> One of the most common classes of ANN is the feedforward neural network (FFNN) with MLP (multilayer perceptron) topologies, which can model complex systems.<sup>42</sup> The usual structure of MLP consists of an input layer, where the number of neurons is equal to the number of model inputs, and an output layer. In addition, there is at least one hidden layer between them with several neurons to be selected by the user.<sup>43,44</sup> This structure has been used to predict different parameters for the oil and gas industry, such as the gas–oil

ratio,<sup>45</sup> volume fraction percentage in three-phase systems,<sup>46</sup> and deposition process of asphaltene<sup>47</sup> and wax.<sup>48</sup>

Knowing the importance of digital transformation, AI, and process monitoring in the oil and gas industries, this work intends to model the scale formation process using MLP to predict the differential pressure ( $\Delta P$ ) one and five steps ahead in time. The goal is to explore how monitoring the conditions in a pipeline (i.e., temperature, pressure, ion concentrations, and differential pressure) can predict the evolution of the scaling process. This study may lead to deeper investigations into applications in monitoring systems and fault detection. TBT differential pressures were monitored over time for different temperatures, pressures, calcium and bicarbonate concentrations, and MEG concentrations. MEG concentration was used as a variable since many scale inhibitor products are solutions of the active molecule in a mixture of water and MEG. Also, MEG can be directly injected in high amounts as thermodynamic gas hydrate inhibitors. Even further, MEG can change the viscosity of the solution and can influence the crystallization of calcium carbonate, which would lead to different effects to be modeled in order to best simulate the scale formation process. Two scenarios were considered: a near future time (differential pressure measured one step ahead) and a far future time (differential pressure measured five steps ahead). The models showed good scaling prediction for both time horizons, showing a promising step toward simulating and predicting scaling tendencies in controlled pipes in production lines.

## METHODOLOGY

**2.1. Experimental Details.** Experiments were performed in TBT equipment, in which two solutions containing compatible cations and anions are pumped into tubes inside an oven, conditioned to the test temperature, mixed in a microchamber, and then flown into a capillary tube called a loop test. The apparatus consisted of two high performance liquid chromatography (HPLC) pumps pushing newly prepared calcium chloride and sodium bicarbonate solutions, with pH ranging from 7.0 to 7.5 depending on the salts concentration, into a thermostat-regulated oven through 1.8 m long stainless-steel tubes with 1 mm inner diameters (i.e., two conditioning loops, one for each solution). These loops ensured that the solutions reached the mixture chamber at the correct temperature for the experiments. After mixing, the combined solution flowed through a third tube (loop test) with the same dimensions as the other tubes. This process resulted in a supersaturated solution leading to calcium carbonate formation and deposition. When deposition occurred, the inlet pressure became higher than the outlet pressure, generating a differential pressure. This differential pressure was measured using a model EJA 130A high-static differential pressure transmitter (Yokogawa, Musashino, Tokyo, Japan). The data were acquired at 1 s intervals using a LabView-based software program. The injection flow rate was 10.0 mL min<sup>-1</sup> (5.00 mL min<sup>-1</sup> for each solution, leading to a 1:1 mixture ratio of the two solutions). The pressure of the system was regulated using a PSV valve connected outside the oven.

**2.2. ANN Database Preparation.** The experimental data used in this study are the results from 38 TBT experiments previously presented in Kartnaller et al.,<sup>23</sup> which used a modeling approach with experiments from a central composite design of the experiment and multivariate linear regression

(MLR). In the previous work, MLR was applied to model the scaling time to reach several differential pressure levels (1–25 psi, in intervals of 1 psi). For each pressure, a different model had to be made, totaling 25 different models to predict a single scaling tendency. These experiments varied the pressure, temperature, concentration of MEG ( $C_{\text{MEG}}$ ) (v/v %), and concentration of the carbonate ( $C_{\text{HCO}_3^-}$ ) (ppm) and calcium ( $C_{\text{Ca}^{2+}}$ ) (ppm) ions over the operating ranges shown in Table 1. The experiments measure the  $\Delta P$  every second as the monitored variable.

**Table 1. Range of Experimental Variables**

Variable	Unit	Minimum value	Maximum value
Pressure	bar	0	170
Temperature	°C	40	110
$C_{\text{MEG}}$	v/v %	0	80
$C_{\text{Ca}^{2+}}$	ppm	1000	6000
$C_{\text{HCO}_3^-}$	ppm	1000	6000

The goal for the ANN modeling in the present work was to improve the prediction of the scale formation process, in which the differential pressure was also an input for the modeling. The measurement of the differential pressure at a moment in time, plus the experimental variables, was used to estimate the differential pressure in a later time. Hence, experimental data were first preprocessed to adjust the signal baseline and create the differential pressure variables one step ahead ( $\Delta P_{(t+1)}$ ) and five steps ahead ( $\Delta P_{(t+5)}$ ) to be used in the prediction models. The database was then split into two parts. The first database consisted of 32 experiments, totaling 46,698 data points. This database was separated into two groups, train (70%) and test (30%), and was used to train the MLP models. To preserve the time information about the scale formation associated with the pressure differential, this division was accomplished by selecting seven data points for the train group and three for the test group from every 10 data points.

The second database consisted of six experiments, totaling 7705 data points. Those experiments were conducted with fixed values of pressure, temperature,  $C_{\text{HCO}_3^-}$  and  $C_{\text{Ca}^{2+}}$ , and varying  $C_{\text{MEG}}$  (10, 20, 30, 50, 60, and 70 v/v %). This database was used to separately validate the models constructed by the ANN for each experiment.

**2.3. Artificial Neural Network Optimization.** For this study, MLP type ANN models with one output neuron were developed using Matlab R2020a (developed by Mathworks, Inc.) to predict  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$ . The inputs chosen were the five independent variables (pressure, temperature,  $C_{\text{HCO}_3^-}$  and  $C_{\text{Ca}^{2+}}$ , and  $C_{\text{MEG}}$ ) plus the differential pressure at the selected time  $t$  ( $\Delta P_{(t)}$ ), resulting in six neurons on the input layer. The proposed MLP structure had one hidden layer, in which the number of neurons is one of the hyperparameters to be optimized. The search was started with the same number of neurons as the input layer.

The activation function, applied to the connection between the input and hidden layers, was the second hyperparameter studied, and the hyperbolic tangent (*tansig*) and log sigmoid (*logsig*) functions were used. Both functions are commonly used due to their sigmoidal form. The linear activation function (*purelin*) was used between the hidden layer and the output layer.<sup>49–51</sup>



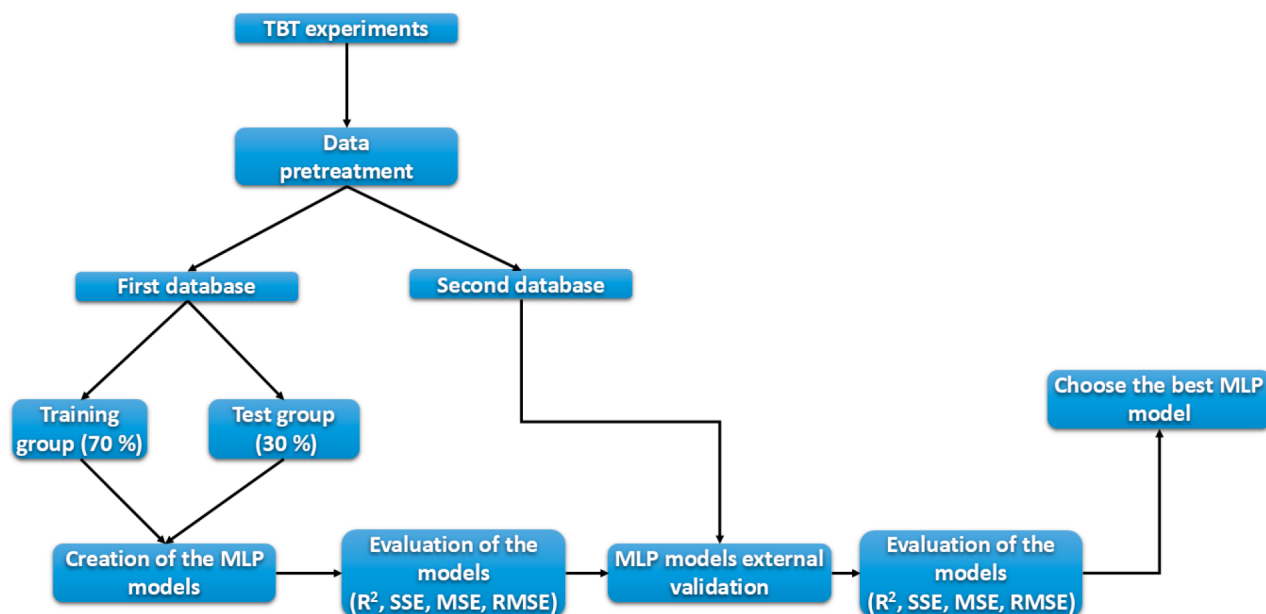


Figure 1. Flowchart of the methodology.

The last hyperparameter optimized for the MLP models was the training algorithm. The gradient descent with momentum and adaptive learning rate backpropagation (*traindx*), Levenberg–Marquardt backpropagation (*trainlm*), and Bayesian regularization backpropagation (*trainbr*) functions were selected for testing. The first of these algorithms improves upon traditional backpropagation with a combination of an adaptive learning rate and momentum training, while the others apply a quasi-Newton method for faster convergence.<sup>52–54</sup>

**2.4. Statistical Performance Evaluation.** To evaluate the performance of the ANN models, the coefficient of determination parameter ( $R^2$ , eq A5), sum of squared errors (SSE, eq A1), mean squared error (MSE, eq A2), and root mean squared error (RMSE, eq A3) were chosen. For  $R^2$ , the goal was to achieve a value close to one, while the goal for the others was to achieve the lowest value possible, indicating the best fit between the experimental data and the predicted data from the ANN models. To calculate  $R^2$ , it is also necessary to calculate the total sum of squares (TSS, eq A4). The equations are available in Appendix A.

Figure 1 shows a schematic for the process adopted in this study, from the data acquisition on the experiments to the determination of the best MLP model.

**2.5. Sensitivity Analysis.** The “black-box” group of models, in which the ANN models are often included, present some difficulty to extract information about the process from their parameters. However, the evaluation of the input variables effects over the output variable can be determined by a sensitivity analysis.

For that, in this study two approaches were explored. First, it was used the relevancy factor ( $r$ , eq 1), which can be applied to quantify these effects, with values on the range from  $-1$  to  $+1$ . The highest absolute value of  $r$  indicates the variables that most affect the target variable, in which the positive values indicate an elevation on the output variable, whereas the negative ones designate a decrease on the target variable.<sup>55,56</sup>

$$r = \frac{\sum_{i=1}^N (X_{k,i} - \bar{X}_k)(y_i - \bar{y})}{\sum_{i=1}^N (X_{k,i} - \bar{X}_k)^2 \sum_{i=1}^N (y_i - \bar{y})^2} \quad (1)$$

where  $N$  is the total number of data points,  $X_{k,i}$  the  $i$ th input value of the  $k$ th parameter,  $y_i$  the  $i$ th output value,  $\bar{X}_k$  the average value of the  $k$ th input parameter, and  $\bar{y}$  the mean value of the output parameter.

The second parameter adopted was the relative importance (RI), in which the methodology proposed by Garson<sup>57</sup> (eq 2) was chosen to obtain the RI values, varying between 0 and 1, which are based on the connection weights between the ANN layers.<sup>58–60</sup>

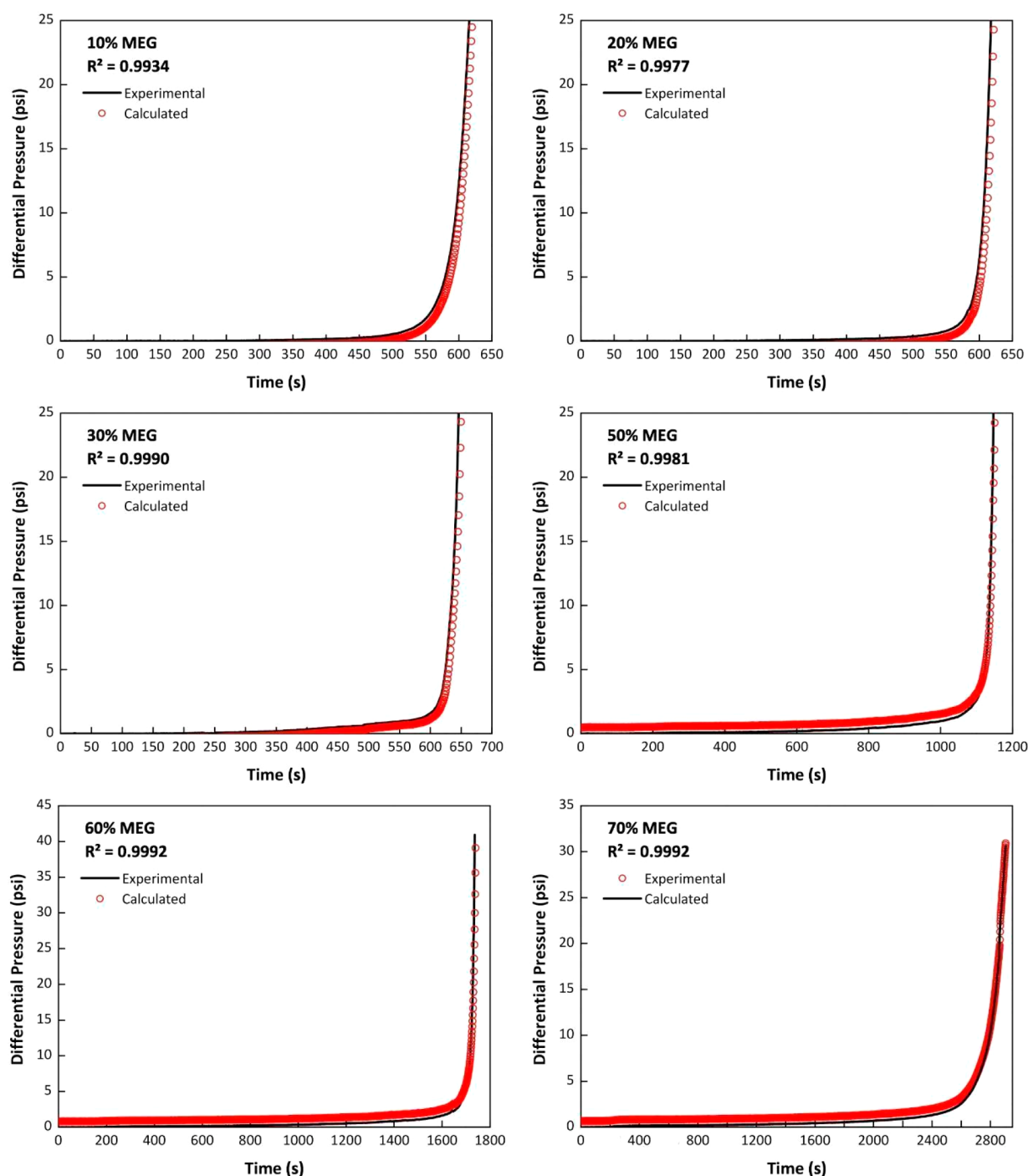
$$RI_{ij} = \frac{\sum_{j=1}^P \frac{|w_{ij}| \cdot |w_{jk}|}{\sum_{i=1}^N |w_{ij}|}}{\sum_{i=1}^N \sum_{j=1}^P \frac{|w_{ij}| \cdot |w_{jk}|}{\sum_{i=1}^N |w_{ij}|}} \quad (2)$$

where  $RI_{ij}$  is the parameters RI of the variable  $x_i$  concerning the output neuron  $j$ ,  $w_{ij}$  the weight parameter of the connection between the input  $x_i$  and the  $j$ th hidden neuron, and  $w_{jk}$  the weight parameter of the connection between the  $j$ th hidden neuron and the  $k$ th output variable.

### 3. RESULTS AND DISCUSSION

The data were selected, processed, and separated into two groups for training and testing to optimize the ANN model. The training data were used to construct the model and calculate the estimated parameters. Once the model was constructed, it was applied to the testing data to predict the output and compare it to the known values. Different types of models were tested by changing the hyperparameters of ANN and were compared to indicate the best ones.

**3.1. Evaluation of ANN Models.** MLP topologies developed to predict  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$  are shown in Table B1 in Appendix B, along with the optimized hyperparameters of the trained models and the performance parameters from the train and test groups, for models having six–eight neurons in the hidden layer. These results show that the best



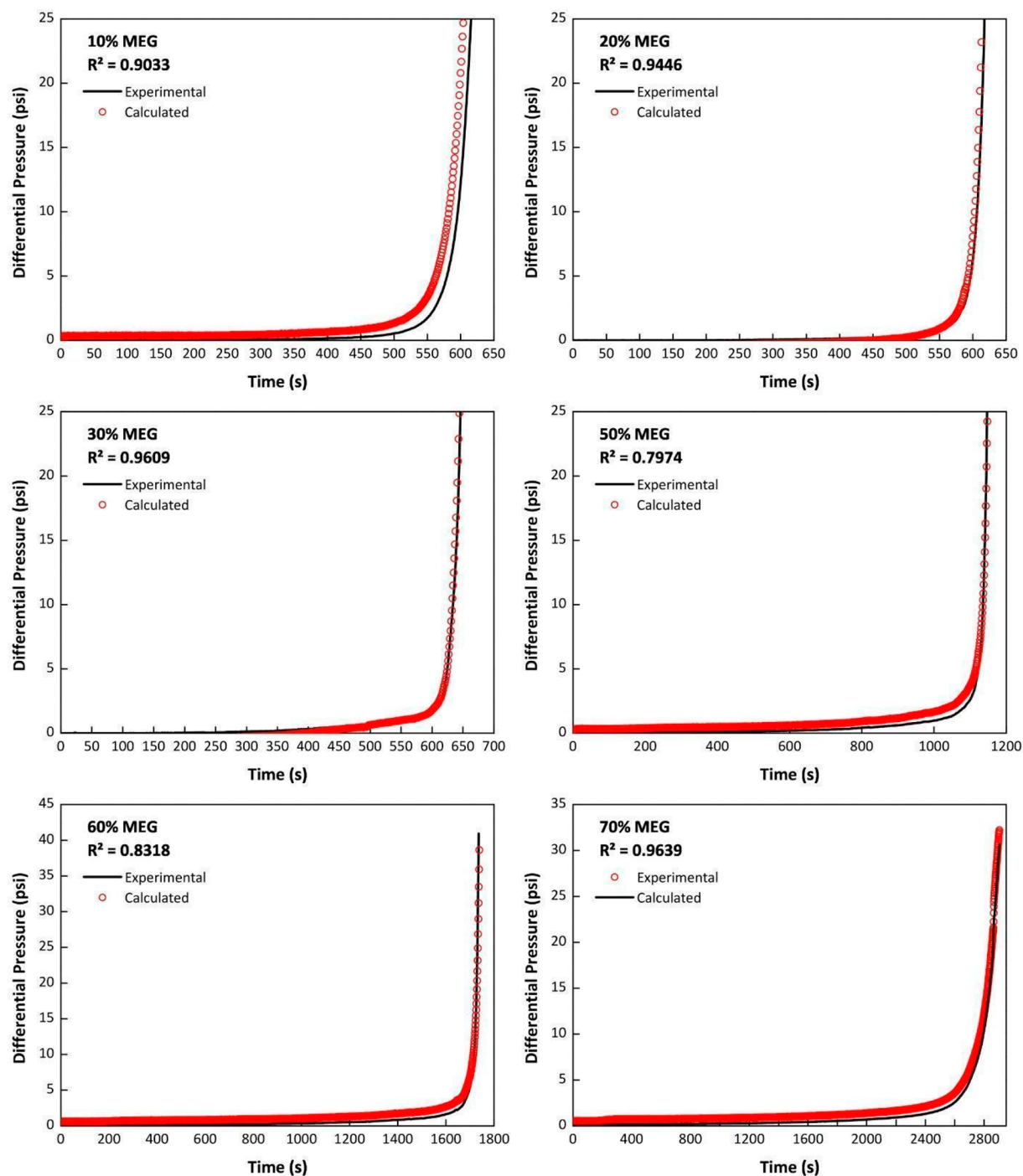
**Figure 2.** Representation of the behavior of the experimental data of the six experiments of the second database and the respective predicted data for the output  $\Delta P_{(t+1)}$  by the MLP model `logsig_7_purelin_1_trainbr`.

performance for  $\Delta P_{(t+1)}$  was achieved with seven neurons in the hidden layer using the *tansig* activation function and the *trainlm* training algorithm. This topology had an  $R^2$  equal to 99.88% for the test set and the lowest values for error. However, only three trained topologies had an  $R^2$  lower than 99%, showing that the models have very similar accuracy.

For the topologies built to predict  $\Delta P_{(t+5)}$ , the model with the best results had the same hidden layer configuration as the best model for  $\Delta P_{(t+1)}$  but used the *trainbr* as the training algorithm. Its performance had an  $R^2$  equal to 98.93% and the lowest values for the other error parameters as well. However,

as observed in the predictions for the  $\Delta P_{(t+1)}$  case, most of the models had very similar figures of merit, indicating that the accuracy was largely independent of the activation function and training algorithm used (*trainlm* and *trainbr*). It is also interesting to point out that the worst results, in both cases, were obtained when using the *traingdx* training algorithm.

This investigation optimizing the hyperparameters of the MLP model for each output, primarily the number of neurons and the transfer function on the hidden layer, is an important step toward achieving the best models. Another essential phase in the model development is to validate them with new



**Figure 3.** Representation of the behavior of the experimental data of the six experiments of the second database and the respective predicted data for the output  $\Delta P_{(t+s)}$  by the MLP model `logsig_6_purelin_1_trainlm`.

experimental data, verifying the model's prediction capability before using it in real applications.

**3.2. Validation of MLP Models.** Since the MLP models demonstrated similar accuracy for both time horizons, all were used in this validation phase. This evaluation used the second database in which the MEG concentration was changed from 10% to 70%, while all other variables were unchanged. This series of experiments tested the behavior of the scaling process in the presence of the glycol molecule. In a previously published article,<sup>23</sup> our research group has shown that MEG

can act as a calcium carbonate inhibitor at concentrations above 30%.

The correct mechanism to explain how MEG acts in the calcium carbonate crystallization is still not completely known. The interactions of alcohols (and therefore polyols) have been studied by several works in the past years, and simulations have shown that the  $-OH$  group can bind to specific faces of the calcite polymorph, which can lead to control of crystal growth.<sup>61–63</sup> Okhrimenko et al.<sup>64</sup> showed that this adsorption could also happen for aragonite and vaterite (other calcium carbonate polymorphs), although the binding energy in these



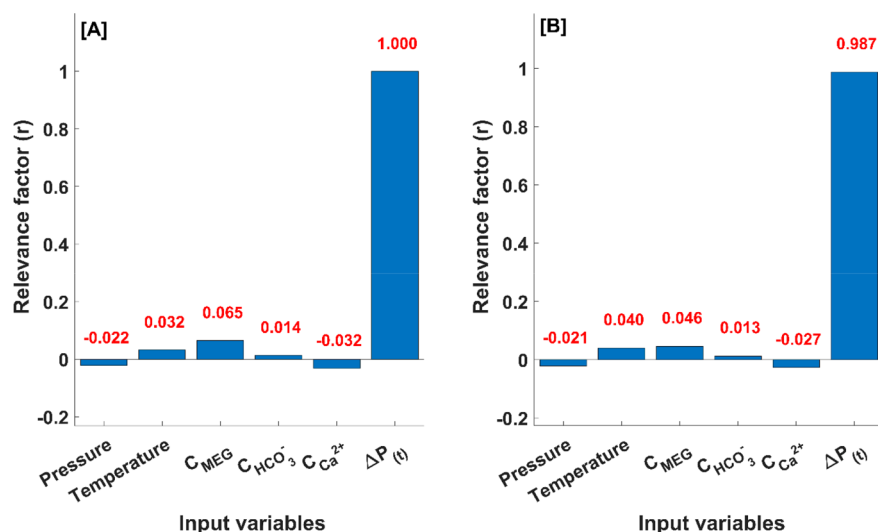


Figure 4. Relevancy factor of both output variables  $\Delta P_{(t+1)}$  (A) and  $\Delta P_{(t+5)}$  (B).

cases is lower than for calcite. This adsorption comes from the fact that the Ca–CO<sub>3</sub> ion pair (note that this is just a representation of pairs, not chemical bond) delocalizes charges by ordering the –OH group of the organic molecules. Thus, the O of this group is associated with Ca, while the H is associated with CO<sub>3</sub>.<sup>64</sup> This causes a highly organized monolayer structure to form on the surface of the crystal, in which the hydrophobic parts of the chains face away from the surface. Many other types of organic molecules have also been studied on the calcium carbonate crystallization, specifically related to biomineralization.

Biomineralization is the process in which living organisms produce hard minerals that act as support, protection, or nourishment structures. A wide variety of minerals can be synthesized by these organisms, such as silica, calcium phosphate, and calcium carbonate. The calcite polymorph synthesized in pure solution in a laboratory has a large crystalline difference from that synthesized by mineralization.<sup>65</sup> This control of crystal growth is generally attributed to complex organic molecules known as coccolith-associated polysaccharides (CAPs). These are large polymeric carbohydrate molecules containing a variety of functional groups, such as –COOH and –OH. Hence, since MEG contains three hydroxyl groups in its structure, it is possible to suppose an association that there is an interaction of this molecule with the surface of the particles being formed, controlling crystal growth, which would also explain how it controls inhibition. Also, changing its concentration changes the viscosity of the solution (affecting the flow dynamics inside the tube).

The performance parameters for all MLP models for each new experiment are presented in the Supporting Information in Tables S3–S8. The models are validated by observing how they predict the scaling process under conditions different from the training or testing. Although the models showed very high accuracy for both training and test sets, their application to the new data was not completely successful. Some of the models' predictions of the scaling process over time were unsatisfactory for a few experiments, which showed that certain regions in the modeled response did not fit the actual expected experimental values. For the  $\Delta P_{(t+1)}$  scenario, the logsig\_7\_purelin\_1\_trainbr model (values of the weights and bias are available in the Supporting Information, Table S1) was the

best with an R<sup>2</sup> over 99.3% for all new experiments. Figure 2 shows the predicted differential pressure from this MLP model and the experimental data for all six experiments. In addition, four other topologies had an R<sup>2</sup> higher than 97% showing that they are also very accurate models.

The lack of fit of parts of the predicted region was mainly observed for the  $\Delta P_{(t+5)}$  case. For example, the best model for this case could not predict the scaling tendency for MEG concentrations between 20%–50%. For some of the experiments, the R<sup>2</sup> of the fit was actually negative, indicating that the scaling process was not being accurately modeled (or that the residues of the regression in that region did not follow a normal distribution with a mean equal to zero).

While most models did not present a good prediction performance for the new experiments, some were still very accurate. For the  $\Delta P_{(t+5)}$  time horizon, the logsig\_6\_purelin\_1\_trainlm model (values of the weights and bias are available in the Supporting Information, Table S2) was the most accurate, with an R<sup>2</sup> ranging from 79.7% to 96.4%. Figure 3 shows the predicted differential pressure from this MLP model and the experimental data for all six experiments. These results are important because they show that even though accurate predictions can be made for some regions of the studied response continuous validation of the best models is necessary as new data are obtained.

For the best models chosen for each output variable,  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$ , a deeper evaluation was performed, starting for a comparison between the experimental and predicted values for the training and test data sets, shown on Figure S1A and B, respectively, for the variables  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$ . These results also show that the model chosen to predict the  $\Delta P_{(t+1)}$  has the best prediction power.

Another investigation adopted was to evaluate the behavior of the normalized residuals according to the  $\Delta P$  values, comparing the response for the both output variables  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$  for the training and test data sets, respectively (Figure S2A, B). From that could be extract that the MLP model for the  $\Delta P_{(t+5)}$  variable has a tendency to predict higher values than the experimental measures, which is worse in higher values of  $\Delta P$ . However, it is important to highlight that the amount of data points with absolute normalized residuals

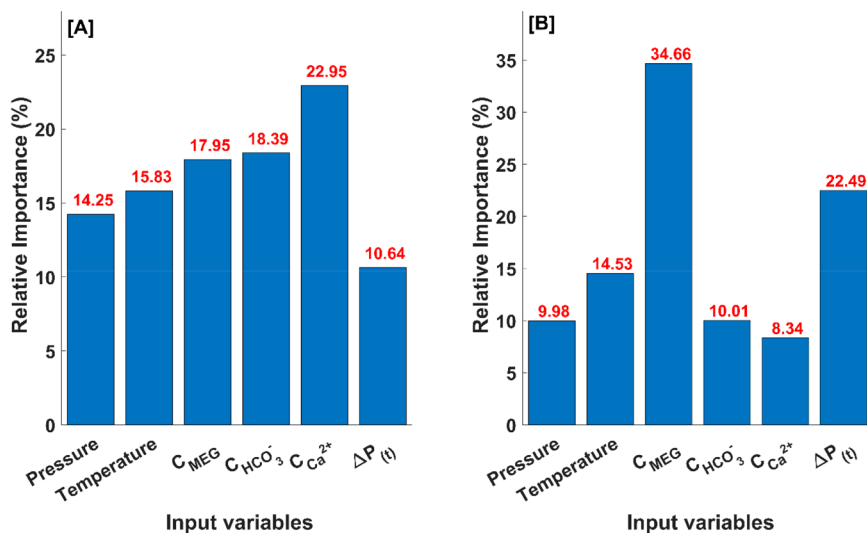


Figure 5. Relative importance (RI) of both output variables  $\Delta P_{(t+1)}$  (A) and  $\Delta P_{(t+5)}$  (B) calculated by the Garson method.<sup>57</sup>

higher than 0.1 is less than 1% for the analyzed data sets for both output variables.

**3.3. Sensitivity Analysis.** For the sensitivity analysis, the best models for each output variable,  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$ , were chosen, which had the topologies `logsig_7_purelin_1_trainbr` and `logsig_6_purelin_1_trainlm`. The first sensitivity evaluation was made for the relevancy factor ( $r$ ); Figure 4A and B shows the values of  $r$  of each input variable for both target variables, respectively,  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$ . They indicate that  $\Delta P(t)$  is by far the most influential parameter for the two prediction horizons with a  $r$  close to 1, indicating expected strong correlation between the measure of the  $\Delta P$  and its prediction for future horizons.

Then, these MLP models were analyzed for the relative importance (RI) parameter, where the values are presented in Figure 5A and B for the output variables  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$ , respectively. For the best  $\Delta P_{(t+1)}$  model, the inputs pressure, temperature,  $C_{MEG}$ , and  $C_{HCO_3^-}$  presented an RI varying between 14% and 19%, and the input variable  $C_{Ca^{2+}}$  was the most relevant one for the  $\Delta P_{(t+1)}$  prediction. In turn, the input with the less impact was  $\Delta P(t)$ .

Conversely, for the best  $\Delta P_{(t+5)}$  model, the most significant variables were  $C_{MEG}$  followed for  $\Delta P(t)$ , respectively with the values of 34.7% and 22.5%, while the other inputs variables presented RI values lower than 15%. This difference on the influence hierarchy of the input variables is interesting, since it shows an increase on the importance of  $\Delta P(t)$  for the prediction of the future. Also, for the  $\Delta P_{(t+5)}$  model, the high RI value of the variable  $C_{MEG}$  indicates a reason for this MLP model presenting the best performance against the validation data group. This may indicate a strong implication that MEG has in impacting the development of the scale formation process due to its inhibitor effect.

The two analyzed parameters,  $r$  and RI, led to different levels of influence for each input in the target variables. While the parameter  $r$  indicates the effect of the input values on the target variable, the RI parameter shows how the model attributes the importance for these inputs. Although, the  $\Delta P(t)$  variable has a huge absolute value for the parameter  $r$ , a model that only uses this variable as input probably could predict the tendency of the  $\Delta P$  curve, but it would not be able to distinguish between the different scenarios. That way, the combination of these

results indicates that maybe a hybrid model could be a better approach for this problem, applying the MLP to lead with the  $\Delta P$  curve behavior and another kind of model to handle the environment conditions information. However, this premise is outside of the scope of this work.

Finally, the modeling results indicated that ANN could be applied to predict the differential pressure and to understand the evolution of the scaling process at earlier as well as later times. For process monitoring, this appears to be a promising tool for transforming digital data acquired during production to establish the scaling tendency of a well over time, by relating the scale formation process with operational variables as a start to develop a model that could simulate the conditions during oil and gas production.

## 4. CONCLUSIONS

This study showed that using an MLP-type ANN enabled modeling of the scaling process in a tube with a dynamic flow containing precipitated calcium carbonate. Even though the scaling process is a very complex system with stochastic behavior, this machine learning technique permitted its prediction over different time horizons: a “near future” or one step ahead ( $\Delta P_{(t+1)}$ ) and a “far future” or five steps ahead ( $\Delta P_{(t+5)}$ ). The generated models were highly accurate for both training and test data sets and for both time horizons, regardless of the activation function and the training algorithm used (*trainlm* and *trainbr*). However, using *traingdx* as a training algorithm gave poorer results. When using the models to predict a different series of experiments that simulated various viscosities with calcium carbonate inhibition, most models did not show the same initial high accuracy. In fact, only a few models were very accurate for all the experiments. Overall, for the  $\Delta P_{(t+1)}$  time horizon, the `logsig_7_purelin_1_trainbr` was the best model, with an  $R^2$  over 99.3% for the additional experiments. The `logsig_6_purelin_1_trainlm` model was the best model for the  $\Delta P_{(t+5)}$  time horizon, with an  $R^2$  ranging from 79.7% to 96.4%. These results show that ANN can predict the differential pressure in a tube to understand the evolution of the scaling process in the near time as well as its development in the future. This strategy represents an important application of digital transformation to oil and gas

Table B1. MLP Topology Models for the Variables  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$

Variables	Hidden Layer		Training algorithm	$R^2$ (train)	$R^2$ (test)	SSE (train) <sup>a</sup>	SSE (test) <sup>a</sup>	MSE (train) <sup>a</sup>	MSE (test) <sup>a</sup>	RMSE (train) <sup>a</sup>	RMSE (test) <sup>a</sup>
	Number of neurons	Activation function									
$\Delta P_{(t+1)}$	7	tansig	trainlm	0.99938	0.99883	0.1830	0.1532	0.0045	0.0087	0.0669	0.0935
	8	tansig	trainbr	0.99941	0.99879	0.1754	0.1575	0.0043	0.0090	0.0655	0.0948
	7	logsig	trainbr	0.99937	0.99877	0.1862	0.1606	0.0046	0.0092	0.0675	0.0958
	7	tansig	trainbr	0.99939	0.99877	0.1810	0.1616	0.0044	0.0092	0.0665	0.0960
	6	logsig	trainbr	0.99934	0.99875	0.1954	0.1638	0.0048	0.0093	0.0691	0.0967
	6	tansig	trainbr	0.99920	0.99869	0.2373	0.1706	0.0058	0.0097	0.0762	0.0987
	7	logsig	trainlm	0.99876	0.99861	0.3679	0.1809	0.0090	0.0103	0.0949	0.1016
	6	tansig	trainlm	0.99921	0.99859	0.2332	0.1843	0.0057	0.0105	0.0755	0.1026
	8	tansig	trainlm	0.99934	0.99848	0.1950	0.1983	0.0048	0.0113	0.0691	0.1064
	8	tansig	traingdx	0.97068	0.97227	8.4319	3.5668	0.2062	0.2036	0.4541	0.4512
$\Delta P_{(t+5)}$	8	logsig	traingdx	0.93977	0.94472	16.9728	6.9699	0.4151	0.3978	0.6443	0.6307
	7	tansig	traingdx	0.93472	0.93567	18.4571	7.9237	0.4514	0.4522	0.6719	0.6725
	7	tansig	trainbr	0.99049	0.98927	3.6088	1.8554	0.0883	0.1059	0.2971	0.3254
	8	logsig	trainbr	0.99087	0.98886	3.4662	1.9151	0.0848	0.1093	0.2912	0.3306
	7	logsig	trainlm	0.99105	0.98884	3.3971	1.9314	0.0831	0.1102	0.2882	0.3320
	8	tansig	trainbr	0.99099	0.98846	3.4214	1.9797	0.0837	0.1130	0.2893	0.3361
	7	tansig	trainlm	0.98958	0.98834	3.9501	2.0475	0.0966	0.1169	0.3108	0.3418
	7	logsig	trainbr	0.98940	0.98584	4.0165	2.4110	0.0982	0.1376	0.3134	0.3709
	6	logsig	trainbr	0.98860	0.98570	4.3171	2.4271	0.1056	0.1385	0.3249	0.3722
	6	tansig	trainbr	0.98828	0.98411	4.4384	2.7042	0.1085	0.1543	0.3295	0.3929
$\Delta P_{(t+5)}$	6	tansig	trainlm	0.98192	0.97995	6.8011	3.4036	0.1663	0.1943	0.4078	0.4407
	6	logsig	trainlm	0.98435	0.97816	5.9041	3.6896	0.1444	0.2106	0.3800	0.4589
	6	logsig	traingdx	0.94447	0.93134	16.4750	8.9705	0.4029	0.5120	0.6348	0.7155
	6	tansig	traingdx	0.93674	0.92913	22.7857	11.3463	0.5573	0.6476	0.7465	0.8047

Using normalized data.

roduction to establish the scaling tendency during the lifetime  
f a well based on differential pressure process monitoring.

## APPENDIX A

### Performance Evaluation Equations

$$SSE = \sum_{i=1}^n (x_i - \hat{x}_i)^2 \quad (A1)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2 \quad (A2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (x_i - \hat{x}_i)^2}{n}} \quad (A3)$$

$$TSS = \sum_{i=1}^n (x_i - \bar{x})^2 \quad (A4)$$

$$R^2 = 1 - \frac{SSE}{TSS} \quad (A5)$$

In the above equations, variables  $n$ ,  $x_i$ ,  $\hat{x}_i$ , and  $\bar{x}$  represent the total number of data points, the observed value, the predicted value, and the mean value of the samples, respectively.

## APPENDIX B

### MLP Topologies Performances

MLP topologies performances are presented in Table B1.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.energyfuels.1c03364>.

Equilibrium equations of calcium carbonate scale formation (eqs S1–S4), regression plot between experimental versus predicted values (Figure S1A, B), comparison between normalized residuals of prediction of  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$  variables for training and test data sets (Figure S2A, B), optimized parameters of best MLP topologies (Tables S1 and S2), and performance values for all topologies for the validation experiments (Tables S3–S8)(PDF)

## AUTHOR INFORMATION

### Corresponding Author

Brunno F. Santos – Department of Chemical and Materials Engineering (DEQM), Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, RJ 22430-060, Brazil; [orcid.org/0000-0001-8755-7749](https://orcid.org/0000-0001-8755-7749); Email: [bsantos@puc-rio.br](mailto:bsantos@puc-rio.br)

### Authors

Bruno X. Ferreira – Department of Chemical and Materials Engineering (DEQM), Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, RJ 22430-060, Brazil; [orcid.org/0000-0001-8378-1102](https://orcid.org/0000-0001-8378-1102)

Carlos R. Hall Barbosa – Postgraduate Program in Metrology, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, RJ 22430-060, Brazil

João Cajaiba – Instituto de Química, Pólo de Xistoquímica, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ 21941-614, Brazil; [orcid.org/0000-0001-7552-7614](https://orcid.org/0000-0001-7552-7614)

Vinicius Kartnaller – Instituto de Química, Pólo de Xistoquímica, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ 21941-614, Brazil

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.energyfuels.1c03364>

## Author Contributions

All authors have contributed to the writing of this manuscript and have approved its final version.

## Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

The authors acknowledge the financial support by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Petrobras in the development of this work. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001 and Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

## ABBREVIATIONS

I, artificial intelligence; ANN, artificial neural networks; NFIS, adaptive neuro-fuzzy inference system;  $\text{CaCO}_3$ , calcium carbonate; CAGR, compound annual growth rate;  $[\text{Ca}^{2+}]$ , concentration of carbonate ions; CAP, coccolith-associated polysaccharide;  $[\text{CHCO}_3^-]$ , concentration of calcium ions;  $C_{\text{MEG}}$ , concentration of MEG; CMIS, committee machine intelligent system;  $\text{CO}_2$ , carbon dioxide;  $\Delta P$ , differential pressure;  $\Delta P_{(t+1)}$ , differential pressure prediction one step ahead;  $\Delta P_{(t+5)}$ , differential pressure prediction five steps ahead; EOR, enhanced oil recovery; ENN, Elman neural network; FFNN, feedforward neural networks; GA, genetic algorithm;  $-\text{OH}$ , hydroxyl group; HPLC, high performance liquid chromatography; *logsig*, *logsigmoid* (activation function); LSSVM, least square support vector machine; MEG, monoethylene glycol; MIC, minimum inhibitor concentration; MLP, multilayer perceptron; MLR, multivariate linear regression; MSE, mean squared error; PCA, principal component analysis; *purelin*, linear (activation function); RI, relative importance;  $r$ , relevancy factor;  $R^2$ , coefficient of determination; RMSE, root mean squared error; SSE, sum of squared errors; SVM, support vector machine; *tansig*, hyperbolic tangent (activation function); *trainbr*, training algorithm Bayesian regularization backpropagation; *traingdx*, training algorithm gradient descent with momentum and adaptive learning rate backpropagation; *trainlm*, training algorithm Levenberg–Marquardt backpropagation; TBT, tube blocking test; TEG, triethylene glycol; THI, thermodynamic hydrate inhibitors; TSS, total sum of squares

## REFERENCES

- (1) de Souza, A. V. A.; Rosário, F.; Cajaiba, J. Evaluation of Calcium Carbonate Inhibitors Using Sintered Metal Filter in a Pressurized Dynamic System. *Materials (Basel)*. **2019**, *12* (11), 1849.
- (2) Khormali, A.; Sharifov, A. R.; Torba, D. I. Increasing Efficiency of Calcium Sulfate Scale Prevention Using a New Mixture of Phosphonate Scale Inhibitors during Waterflooding. *J. Pet. Sci. Eng.* **2018**, *164*, 245–258.

- (3) Nguyen, D. A.; Iwaniw, M. A.; Fogler, H. S. Kinetics and Mechanism of the Reaction between Ammonium and Nitrite Ions: Experimental and Theoretical Studies. *Chem. Eng. Sci.* **2003**, *58* (19), 4351–4362.
- (4) Khalil De Oliveira, M. C.; Gonçalves, M. A. An Effort to Establish Correlations between Brazilian Crude Oils Properties and Flow Assurance Related Issues. *Energy Fuels* **2012**, *26* (9), 5689–5701.
- (5) Melchuna, A.; Zhang, X.; Sa, J. H.; Abadie, E.; Glénat, P.; Sum, A. K. Flow Risk Index: A New Metric for Solid Precipitation Assessment in Flow Assurance Management Applied to Gas Hydrate Transportability. *Energy Fuels* **2020**, *34* (8), 9371–9378.
- (6) Kim, H.; Yoo, W.; Lim, Y.; Seo, Y. Economic Evaluation of MEG Injection and Regeneration Process for Oil FPSO. *J. Pet. Sci. Eng.* **2018**, *164*, 417–426.
- (7) Nasir, Q.; Suleman, H.; Elsheikh, Y. A. A Review on the Role and Impact of Various Additives as Promoters/ Inhibitors for Gas Hydrate Formation. *J. Nat. Gas Sci. Eng.* **2020**, *76*, 103211.
- (8) Kumar, S.; Naiya, T. K.; Kumar, T. Developments in Oilfield Scale Handling towards Green Technology-A Review. *J. Pet. Sci. Eng.* **2018**, *169* (May), 428–444.
- (9) da Rosa, K. R. S. A.; Fontes, R. A.; do Rosário, F. F.; Freitas, T. C.; de Oliveira Penna, M.; Castro, B. B.; da Silva, M. G. F.; da Silva, G. M. L. L.; Silva, J. M.; Figueiredo, M. R. Improved Protocol for Scale Inhibitor Evaluation: A Meaningful Step on Scale Management. In *Offshore Technology Conference Brasil*, Rio de Janeiro, Brazil, October 2019. DOI: 10.4043/29683-ms.
- (10) Ahmed, M.; Hussein, I. A.; Onawole, A. T.; Saad, M. A.; Mahmoud, M. Dissolution Kinetics of Different Inorganic Oilfield Scales in Green Formulations. *ACS Omega* **2020**, *5* (46), 29963–29970.
- (11) Kan, A. T.; Fu, G.; Watson, M. A.; Tomson, M. B. Effect of Hydrate Inhibitors on Oilfield Scale Formation and Inhibition. *SPE Oilf. Scale Symp.* **2002**, 83–94.
- (12) Lim, V. W. S.; Metaxas, P. J.; Stanwix, P. L.; Johns, M. L.; Haandrikman, G.; Crosby, D.; Aman, Z. M.; May, E. F. Gas Hydrate Formation Probability and Growth Rate as a Function of Kinetic Hydrate Inhibitor (KHI) Concentration. *Chem. Eng. J.* **2020**, *388*, 124177.
- (13) Olajire, A. A. A Review of Oilfield Scale Management Technology for Oil and Gas Production. *J. Pet. Sci. Eng.* **2015**, *135*, 723–737.
- (14) Dyer, S. J.; Graham, G. M. The Effect of Temperature and Pressure on Oilfield Scale Formation. *J. Pet. Sci. Eng.* **2002**, *35* (1–2), 95–107.
- (15) Oddo, J. E.; Tomson, M. B. Why Scale Forms in the Oil Field and Methods To Predict It. *SPE Prod. Facil.* **1994**, *9* (01), 47–54.
- (16) Frenier, W. W.; Ziauddin, M. *Formation, Removal, and Inhibition of Inorganic Scale in the Oilfield Environment*; Society of Petroleum Engineers, 2008.
- (17) Global Oilfield Scale Inhibitor Market – Industry Trends and Forecast to 2027. *Data Bridge Market Research*. <https://www.databridgemarketresearch.com/reports/global-oilfield-scale-inhibitor-market> (accessed Aug 26, 2021).
- (18) Hydrate Inhibitors Market Analysis. *Coherent Market Insights*. <https://www.coherentmarketinsights.com/market-insight/hydrate-inhibitors-market-555> (accessed Aug 27, 2021).
- (19) Oilfield Scale Inhibitor Market. *Markets and Markets*. <https://www.marketsandmarkets.com/Market-Reports/oilfield-scale-inhibitor-market-268225660.html> (accessed Aug 26, 2021).
- (20) Wang, Q.; Al-nasser, W.; Chen, T.; Aramco, S.; Liang, F. Calcium Carbonate Scale Inhibition: Effects of EOR Chemicals. In *Corrosion Conference & Expo 2018*; NACE International: Phoenix, Arizona, USA, 2018; pp 1–12.
- (21) Chao, J.; Zhang, L.; Feng, R.; Wang, Z.; Xu, S.; Zhang, C.; Ren, S. Experimental Study on the Compatibility of Scale Inhibitors with Mono Ethylene Glycol. *Pet. Res.* **2020**, *5* (4), 315–325.



- (22) Seiersten, M.; Kundu, S. S. Scale Management in Monoethylene Glycol MEG Systems - A Review. *Soc. Pet. Eng. - SPE Int. Oilf. Scale Conf. Exhib.* **2018**, *2018*, No. June, 20–21.
- (23) Kartnaller, V.; Venâncio, F.; do Rosário, F. F.; Cajaiba, J. Application of Multiple Regression and Design of Experiments for Modelling the Effect of Monoethylene Glycol in the Calcium Carbonate Scaling Process. *Molecules* **2018**, *23* (4), 860–12.
- (24) Ramzi, M.; Hosny, R.; El-Sayed, M.; Fathy, M.; Moghny, T. A. Evaluation of Scale Inhibitors Performance under Simulated Flowing Field Conditions Using Dynamic Tube Blocking Test. *Int. J. Chem. Sci.* **2016**, *14* (1), 16–28.
- (25) Macedo, R. G. M. d. A.; Marques, N. do N.; Paulucci, L. C. S.; Cunha, J. V. M.; Villetti, M. A.; Castro, B. B.; Balaban, R. de C. Water-Soluble Carboxymethylchitosan as Green Scale Inhibitor in Oil Wells. *Carbohydr. Polym.* **2019**, *215*, 137–142.
- (26) Fernandes, R. S.; Santos, W. D. L.; de Lima, D. F.; de Souza, M. A. F.; Castro, B. B.; Balaban, R. C. Application of Water-Soluble Polymers as Calcium Carbonate Scale Inhibitors in Petroleum Wells: A Uni- and Multivariate Approach. *Desalination* **2021**, *515* (June), 115201.
- (27) de Moraes, S. C.; de Lima, D. F.; Ferreira, T. M.; Domingos, J. B.; de Souza, M. A. F.; Castro, B. B.; Balaban, R. de C. Effect of PH on the Efficiency of Sodium Hexametaphosphate as Calcium Carbonate Scale Inhibitor at High Temperature and High Pressure. *Desalination* **2020**, *491* (May), 114548.
- (28) Sanni, O. S.; Bukuaghangin, O.; Charpentier, T. V. J.; Neville, J. Evaluation of Laboratory Techniques for Assessing Scale Inhibition Efficiency. *J. Pet. Sci. Eng.* **2019**, *182* (July), 106347.
- (29) Santos, H. F. L.; Castro, B. B.; Bloch, M.; Martins, A. L.;chlüter, H. E. P.; Júnior, M. F. S.; Jacinto, C. M. C.; Rosário, F. F. A Physical Model for Scale Growth during the Dynamic Tube Blocking Test. *OTC Bras. 2017* **2017**, 161–180.
- (30) Paz, P. A.; Caprace, J.-D.; Cajaiba, J. F.; Netto, T. A. Prediction of Calcium Carbonate Scaling in Pipes Using Artificial Neural Networks. In *ASME 2017 36th International Conference on Ocean, Offshore and Arctic Engineering*; ASME: Trondheim, Norway, 2017; pp 9–10. DOI: 10.1115/OMAE2017-61233.
- (31) Ahmadi, M. A.; Bahadori, A.; Shadizadeh, S. R. A Rigorous Model to Predict the Amount of Dissolved Calcium Carbonate Concentration throughout Oil Field Brines: Side Effect of Pressure and Temperature. *Fuel* **2015**, *139* (1), 154–159.
- (32) Wang, J.; Lv, Z.; Liang, Y.; Deng, L.; Li, Z. Fouling Resistance Prediction Based on GA–Elman Neural Network for Circulating Cooling Water with Electromagnetic Anti-Fouling Treatment. *J. Energy Inst.* **2019**, *92* (5), 1519–1526.
- (33) Baghban, A.; Sasanipour, J.; Pourfayaz, F.; Ahmadi, M. H.; Kasaeian, A.; Chamkha, A. J.; Oztog, H. F.; Chau, K.-w. Towards Experimental and Modeling Study of Heat Transfer Performance of Water–SiO<sub>2</sub> Nanofluid in Quadrangular Cross-Section Channels. *Eng. Appl. Comput. Fluid Mech.* **2019**, *13* (1), 453–469.
- (34) Baghban, A.; Pourfayaz, F.; Ahmadi, M. H.; Kasaeian, A.; Pourkiaei, S. M.; Lorenzini, G. Connectionist Intelligent Model Estimates of Convective Heat Transfer Coefficient of Nanofluids in Circular Cross-Sectional Channels. *J. Therm. Anal. Calorim.* **2018**, *132* (2), 1213–1239.
- (35) Baghban, A.; Kahani, M.; Nazari, M. A.; Ahmadi, M. H.; Yan, W. M. Sensitivity Analysis and Application of Machine Learning Methods to Predict the Heat Transfer Performance of CNT/Water Nanofluid Flows through Coils. *Int. J. Heat Mass Transfer* **2019**, *128*, 825–835.
- (36) Ahmadi, M. H.; Baghban, A.; Sadeghzadeh, M.; Zamen, M.; Mosavi, A.; Shamshirband, S.; Kumar, R.; Mohammadi-Khanaposhtani, M. Evaluation of Electrical Efficiency of Photovoltaic Thermal Solar Collector. *Eng. Appl. Comput. Fluid Mech.* **2020**, *14* (1), 545–565.
- (37) Zamen, M.; Baghban, A.; Pourkiaei, S. M.; Ahmadi, M. H. Optimization Methods Using Artificial Intelligence Algorithms to Estimate Thermal Efficiency of PV/T System. *Energy Sci. Eng.* **2019**, *7* (3), 821–834.
- (38) Alkinani, H. H.; Al-Hameedi, A. T. T.; Dunn-Norman, S.; Flori, R. E.; Alsaba, M. T.; Amer, A. S. Applications of Artificial Neural Networks in the Petroleum Industry: A Review. In *SPE Middle East Oil and Gas Show and Conference, MEOS, Proceedings*; March 2019. DOI: 10.2118/195072-ms.
- (39) Otchere, D. A.; Arbi Ganat, T. O.; Gholami, R.; Ridha, S. Application of Supervised Machine Learning Paradigms in the Prediction of Petroleum Reservoir Properties: Comparative Analysis of ANN and SVM Models. *J. Pet. Sci. Eng.* **2021**, *200*, 108182.
- (40) Rahmanifard, H.; Plaksina, T. Application of Artificial Intelligence Techniques in the Petroleum Industry: A Review. *Artif. Intell. Rev.* **2019**, *52* (4), 2295–2318.
- (41) Li, H.; Zhang, Z.; Liu, Z. Application of Artificial Neural Networks for Catalysis: A Review. *Catalysts* **2017**, *7* (10), 306.
- (42) Heidari, A. A.; Faris, H.; Mirjalili, S.; Aljarah, I.; Mafarja, M. *Ant Lion Optimizer: Theory, Literature Review, and Application in Multi-Layer Perceptron Neural Networks*; Springer International Publishing, 2020; Vol. 811. DOI: 10.1007/978-3-030-12127-3\_3.
- (43) Li, H.; Yu, H.; Cao, N.; Tian, H.; Cheng, S. Applications of Artificial Intelligence in Oil and Gas Development. *Arch. Comput. Methods Eng.* **2021**, *28* (3), 937–949.
- (44) Hammoudi, A.; Moussaceb, K.; Belebchouche, C.; Dahmoune, F. Comparison of Artificial Neural Network (ANN) and Response Surface Methodology (RSM) Prediction in Compressive Strength of Recycled Concrete Aggregates. *Constr. Build. Mater.* **2019**, *209*, 425–436.
- (45) Cheshmeh Sefidi, A.; Ajorkaran, F. A Novel MLP-ANN Approach to Predict Solution Gas-Oil Ratio. *Pet. Sci. Technol.* **2019**, *37* (23), 2302–2308.
- (46) Islami rad, S. Z.; Gholipour Peyvandi, R. A Simple and Inexpensive Design for Volume Fraction Prediction in Three-Phase Flow Meter: Single Source-Single Detector. *Flow Meas. Instrum.* **2019**, *69* (June), 101587.
- (47) Zarei, F.; Baghban, A. Phase Behavior Modelling of Asphaltene Precipitation Utilizing MLP-ANN Approach. *Pet. Sci. Technol.* **2017**, *35* (20), 2009–2015.
- (48) Nait Amar, M.; Jahanbani Ghahfarokhi, A.; Shang Wui Ng, C. Predicting Wax Deposition Using Robust Machine Learning Techniques. *Petroleum* **2021**, DOI: 10.1016/j.petlm.2021.07.005.
- (49) Chojaczyk, A. A.; Teixeira, A. P.; Neves, L. C.; Cardoso, J. B.; Guedes Soares, C. Review and Application of Artificial Neural Networks Models in Reliability Analysis of Steel Structures. *Struct. Saf.* **2015**, *52* (PA), 78–89.
- (50) Soleimani, R.; Shoushtari, N. A.; Mirza, B.; Salahi, A. Experimental Investigation, Modeling and Optimization of Membrane Separation Using Artificial Neural Network and Multi-Objective Optimization Using Genetic Algorithm. *Chem. Eng. Res. Des.* **2013**, *91* (5), 883–903.
- (51) Haykin, S. *Redes Neurais – Princípios Prática*, 2ed.; Bookman: Porto Alegre, RS, Brazil, 2001.
- (52) traingdx: Gradient descent with momentum and adaptive learning rate backpropagation. *MathWorks*. <https://www.mathworks.com/help/deeplearning/ref/traingdx.html> (accessed Sep 27, 2020).
- (53) trainlm: Levenberg–Marquardt backpropagation. *MathWorks*. <https://www.mathworks.com/help/deeplearning/ref/trainlm.html> (accessed Sep 27, 2020).
- (54) trainbr: Bayesian regularization backpropagation. *MathWorks*. <https://www.mathworks.com/help/deeplearning/ref/trainbr.html> (accessed Sep 27, 2020).
- (55) Baghban, A.; Jalali, A.; Shafiee, M.; Ahmadi, M. H.; Chau, K.-w. Developing an ANFIS-Based Swarm Concept Model for Estimating the Relative Viscosity of Nanofluids. *Eng. Appl. Comput. Fluid Mech.* **2019**, *13* (1), 26–39.
- (56) Ahmadi, M. H.; Baghban, A.; Ghazvini, M.; Hadipoor, M.; Ghasempour, R.; Nazemzadegan, M. R. An Insight into the Prediction of TiO<sub>2</sub>/Water Nanofluid Viscosity through Intelligence Schemes. *J. Therm. Anal. Calorim.* **2020**, *139* (3), 2381–2394.
- (57) Garson, G. D. Interpreting Neural-Network Connection Weights. *AI Expert* **1991**, *6* (4), 47–51.

(58) de Oña, J.; Garrido, C. Extracting the Contribution of Independent Variables in Neural Network Models: A New Approach to Handle Instability. *Neural Comput. Appl.* **2014**, *25* (3–4), 859–869.

(59) Xu, M.; Wong, T. C.; Chin, K. S. Modeling Daily Patient Arrivals at Emergency Department and Quantifying the Relative Importance of Contributing Variables Using Artificial Neural Network. *Decis. Support Syst.* **2013**, *54* (3), 1488–1498.

(60) Pentoś, K. The Methods of Extracting the Contribution of Variables in Artificial Neural Network Models - Comparison of Inherent Instability. *Comput. Electron. Agric.* **2016**, *127*, 141–146.

(61) Sand, K. K.; Yang, M.; Makovicky, E.; Cooke, D. J.; Hassenkam, T.; Bechgaard, K.; Stipp, S. L. S. Binding of Ethanol on Calcite: The Role of the OH Bond and Its Relevance to Biomineralization. *Langmuir* **2010**, *26* (19), 15239–15247.

(62) Bovet, N.; Yang, M.; Javadi, M. S.; Stipp, S. L. S. Interaction of Alcohols with the Calcite Surface. *Phys. Chem. Chem. Phys.* **2015**, *17* (5), 3490–3496.

(63) Zhang, L.; Yue, L. H.; Wang, F.; Wang, Q. Dissive Effect of Alcohol-Water Mixed Solvents on Growth Morphology of Calcium Carbonate Crystals. *J. Phys. Chem. B* **2008**, *112* (34), 10668–10674.

(64) Okhrimenko, D. V.; Nissenbaum, J.; Andersson, M. P.; Olsson, M. H. M.; Stipp, S. L. S. Energies of the Adsorption of Functional Groups to Calcium Carbonate Polymorphs: The Importance of -OH and -COOH Groups. *Langmuir* **2013**, *29* (35), 11062–11073.

(65) Yang, M.; Stipp, S. L. S.; Harding, J. Biological Control on calcite Crystallization by Polysaccharides. *Cryst. Growth Des.* **2008**, *8* (11), 4066–4074.

## Recommended by ACS

### Evaluation of Silica and Related Matrix Ion Effects on Common Scale Inhibitors

Yue Zhao, Mason Tomson, *et al.*

JANUARY 17, 2021  
ENERGY & FUELS

READ 

### Machine-Learning Approach for Forecasting Steam-Assisted Gravity-Drainage Performance in the Presence of Noncondensable Gases

Serhat Canbolat and Emre Artun

JUNE 07, 2022  
ACS OMEGA

READ 

### Application of Machine Learning Methods in Modeling the Loss of Circulation Rate while Drilling Operation

Ahmed Alsaihati, Dhafer Al Shehri, *et al.*

JUNE 08, 2022  
ACS OMEGA

READ 

### Experimental Evaluation of Common Sulfate Mineral Scale Coprecipitation Kinetics in Oilfield Operating Conditions

Ping Zhang, Mason B. Tomson, *et al.*

JUNE 18, 2019  
ENERGY & FUELS

READ 

Get More Suggestions >

## B

### Appendix of the article: Development of MLP artificial neural network models for the simulation of CaCO<sub>3</sub> scale formation process in the presence of monoethylene glycol (MEG) in a dynamic tube blocking test (TBT) equipment

#### PERFORMANCE EVALUATION EQUATIONS

$$SSE = \sum_{i=1}^n (x_i - \hat{x}_i)^2 \quad (B1)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2 \quad (B2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (x_i - \hat{x}_i)^2}{n}} \quad (B3)$$

$$TSS = \sum_{i=1}^n (x_i - \bar{x})^2 \quad (B4)$$

$$R^2 = 1 - \frac{SSE}{TSS} \quad (B5)$$

In the above equations, variables  $n$ ,  $x_i$ ,  $\hat{x}_i$  and  $\bar{x}$  represent the total number of data points, the observed value, the predicted value, and the mean value of the samples, respectively.

Table B1. MLP topology models for the variables  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$ .

Variables	Hidden Layer		Training algorithm	$R^2$ (train)	$R^2$ (test)	SSE (train) <sup>a</sup>	SSE (test) <sup>a</sup>	MSE (train) <sup>a</sup>	MSE (test) <sup>a</sup>	RMSE (train) <sup>a</sup>	RMSE (test) <sup>a</sup>
	Number of neurons	Activation function									
$\Delta P_{(t+1)}$	7	<i>tansig</i>	<i>trainlm</i>	0.99938	0.99883	0.1830	0.1532	0.0045	0.0087	0.0669	0.0935
	8	<i>tansig</i>	<i>trainbr</i>	0.99941	0.99879	0.1754	0.1575	0.0043	0.0090	0.0655	0.0948
	7	<i>logsig</i>	<i>trainbr</i>	0.99937	0.99877	0.1862	0.1606	0.0046	0.0092	0.0675	0.0958
	7	<i>tansig</i>	<i>trainbr</i>	0.99939	0.99877	0.1810	0.1616	0.0044	0.0092	0.0665	0.0960
	6	<i>logsig</i>	<i>trainbr</i>	0.99934	0.99875	0.1954	0.1638	0.0048	0.0093	0.0691	0.0967
	6	<i>tansig</i>	<i>trainbr</i>	0.99920	0.99869	0.2373	0.1706	0.0058	0.0097	0.0762	0.0987
	7	<i>logsig</i>	<i>trainlm</i>	0.99876	0.99861	0.3679	0.1809	0.0090	0.0103	0.0949	0.1016
	6	<i>tansig</i>	<i>trainlm</i>	0.99921	0.99859	0.2332	0.1843	0.0057	0.0105	0.0755	0.1026
	8	<i>tansig</i>	<i>trainlm</i>	0.99934	0.99848	0.1950	0.1983	0.0048	0.0113	0.0691	0.1064
	8	<i>tansig</i>	<i>traingdx</i>	0.97068	0.97227	8.4319	3.5668	0.2062	0.2036	0.4541	0.4512
	8	<i>logsig</i>	<i>traingdx</i>	0.93977	0.94472	16.9728	6.9699	0.4151	0.3978	0.6443	0.6307
	7	<i>tansig</i>	<i>traingdx</i>	0.93472	0.93567	18.4571	7.9237	0.4514	0.4522	0.6719	0.6725
$\Delta P_{(t+5)}$	7	<i>tansig</i>	<i>trainbr</i>	0.99049	0.98927	3.6088	1.8554	0.0883	0.1059	0.2971	0.3254
	8	<i>logsig</i>	<i>trainbr</i>	0.99087	0.98886	3.4662	1.9151	0.0848	0.1093	0.2912	0.3306
	7	<i>logsig</i>	<i>trainlm</i>	0.99105	0.98884	3.3971	1.9314	0.0831	0.1102	0.2882	0.3320
	8	<i>tansig</i>	<i>trainbr</i>	0.99099	0.98846	3.4214	1.9797	0.0837	0.1130	0.2893	0.3361
	7	<i>tansig</i>	<i>trainlm</i>	0.98958	0.98834	3.9501	2.0475	0.0966	0.1169	0.3108	0.3418
	7	<i>logsig</i>	<i>trainbr</i>	0.98940	0.98584	4.0165	2.4110	0.0982	0.1376	0.3134	0.3709
	6	<i>logsig</i>	<i>trainbr</i>	0.98860	0.98570	4.3171	2.4271	0.1056	0.1385	0.3249	0.3722
	6	<i>tansig</i>	<i>trainbr</i>	0.98828	0.98411	4.4384	2.7042	0.1085	0.1543	0.3295	0.3929
	6	<i>tansig</i>	<i>trainlm</i>	0.98192	0.97995	6.8011	3.4036	0.1663	0.1943	0.4078	0.4407



6	<i>logsig</i>	<i>trainlm</i>	0.98435	0.97816	5.9041	3.6896	0.1444	0.2106	0.3800	0.4589
6	<i>logsig</i>	<i>traingdx</i>	0.94447	0.93134	16.4750	8.9705	0.4029	0.5120	0.6348	0.7155
6	<i>tansig</i>	<i>traingdx</i>	0.93674	0.92913	22.7857	11.3463	0.5573	0.6476	0.7465	0.8047

---

a - Using normalized data

**C**

**Supporting information of the article: Development of MLP artificial neural network models for the simulation of CaCO<sub>3</sub> scale formation process in the presence of monoethylene glycol (MEG) in a dynamic tube blocking test (TBT) equipment**

# Development of MLP artificial neural network models for the simulation of CaCO<sub>3</sub> scale formation process in the presence of monoethylene glycol (MEG) in a dynamic tube blocking test (TBT) equipment

*AUTHOR NAMES : Bruno X. Ferreira<sup>1</sup>, Vinicius Kartnaller<sup>2</sup>, Carlos R. Hall Barbosa<sup>3</sup>, João Cajaiba<sup>2</sup>, Brunno F. Santos<sup>1\*</sup>.*

AUTHOR ADDRESS :

1 Department of Chemical and Materials Engineering (DEQM), Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rua Marquês de São Vicente, 225 – Gávea, Rio de Janeiro, RJ 22430-060, Brazil.

2 Instituto de Química, Pólo de Xistoquímica, Universidade Federal do Rio de Janeiro

(UFRJ), Rua H lio de Almeida 40, Cidade Universit ria, Rio de Janeiro 21941-614, Brazil.

3 Postgraduate Program in Metrology, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rua Marqu s de S o Vicente, 225 – G vea, Rio de Janeiro, RJ 22430-060, Brazil.

### Equilibrium equations of the calcium carbonate scale formation

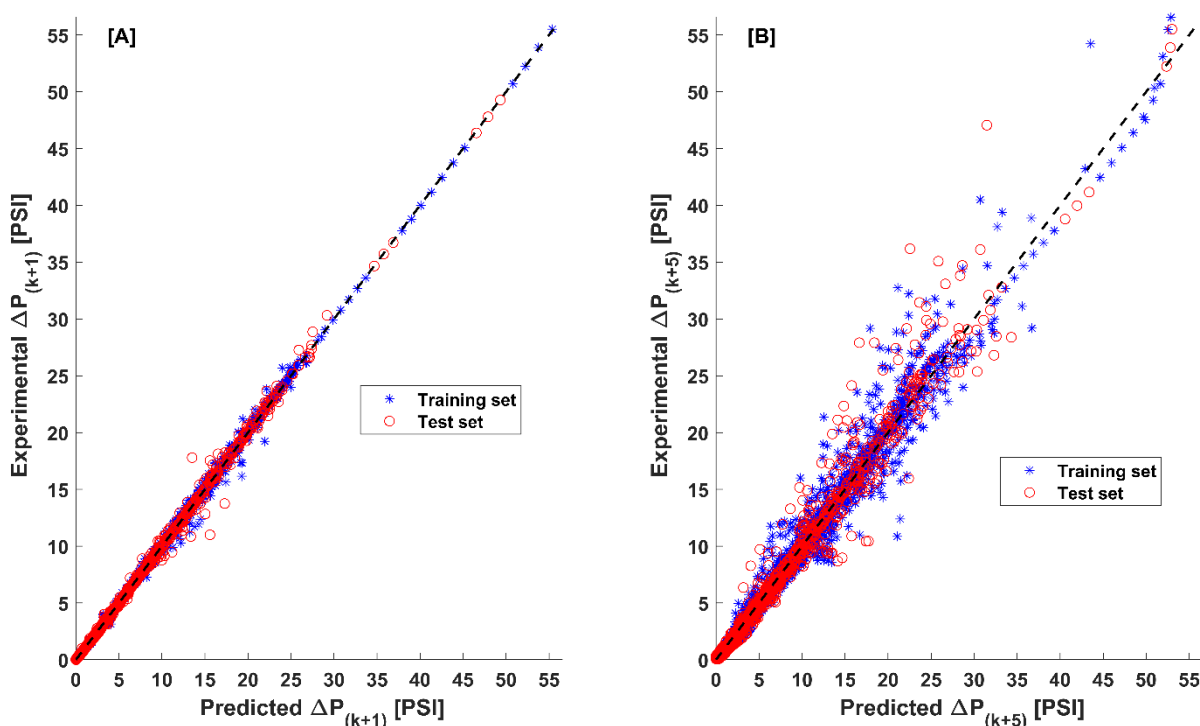
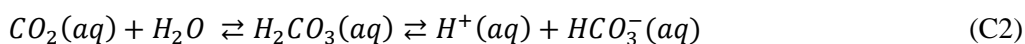


Figure C1. Regression plot between experimental versus the predicted values for the variables  $\Delta P_{(t+1)}$ (A) and  $\Delta P_{(t+5)}$ (B).

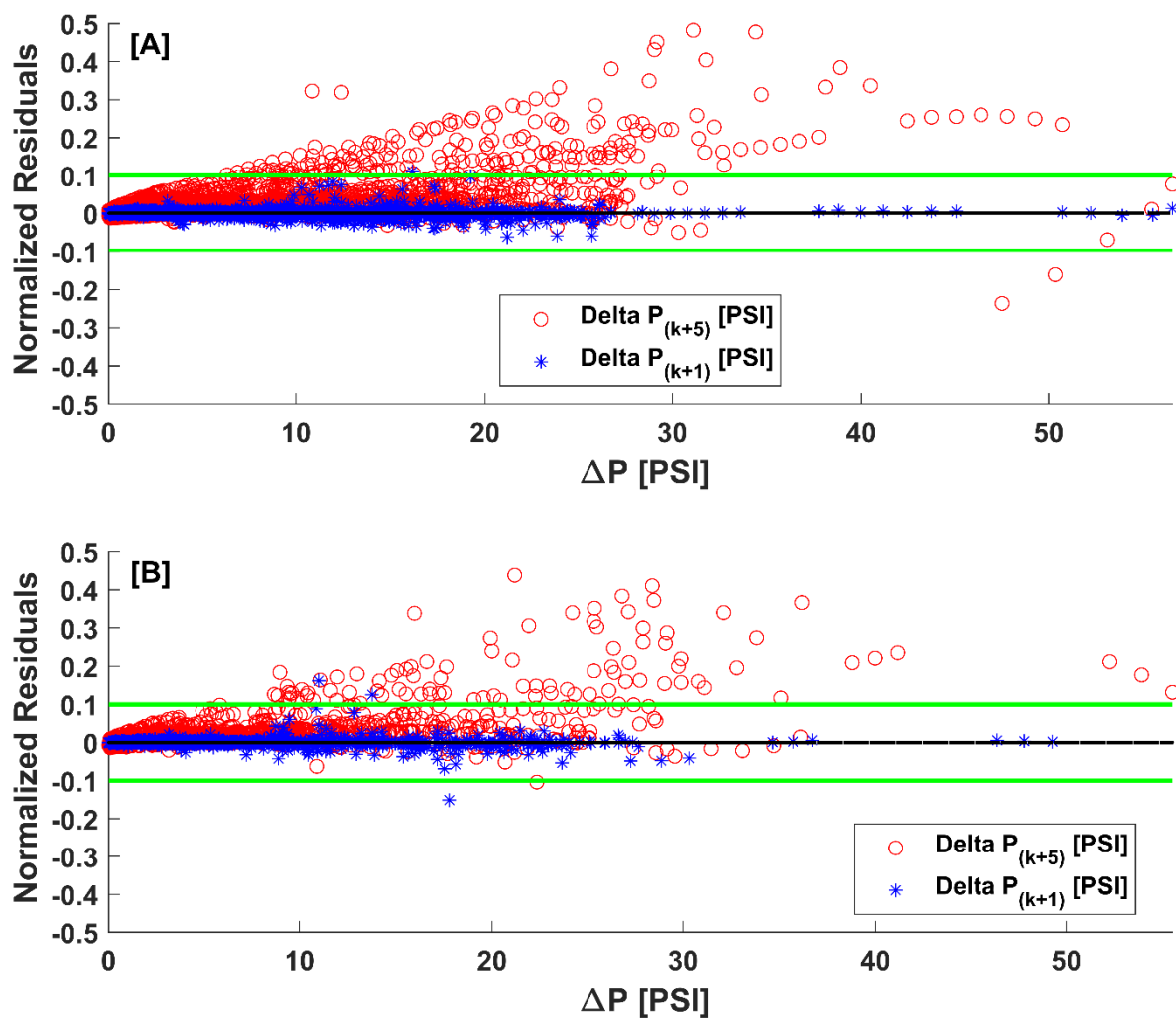


Figure C2. Comparison between normalized residuals of the prediction of the  $\Delta P_{(t+1)}$  and  $\Delta P_{(t+5)}$  variables for the training dataset (A) and test dataset (B).

Table C1: Performance values for each MLP topology for the experiment with 10 v/v% MEG.

Hidden Layer				C <sub>MEG</sub>			
Variables	Number of neurons	Activation function	Training algorithm	10 v/v %			
				R <sup>2</sup>	SSE <sup>a</sup>	MSE <sup>a</sup>	RMSE <sup>a</sup>
ΔP (k+1)	7	<i>tansig</i>	<i>trainlm</i>	0.9404	1.7051	2.1569	1.4686
	8	<i>tansig</i>	<i>trainbr</i>	0.7485	7.9869	10.1033	3.1786
	7	<i>logsig</i>	<i>trainbr</i>	0.9934	0.1745	0.2208	0.4699
	7	<i>tansig</i>	<i>trainbr</i>	0.9932	0.1527	0.1931	0.4395
	6	<i>logsig</i>	<i>trainbr</i>	0.9660	1.0771	1.3625	1.1673
	6	<i>tansig</i>	<i>trainbr</i>	0.9877	0.3445	0.4358	0.6602
	7	<i>logsig</i>	<i>trainlm</i>	0.9972	0.0617	0.0780	0.2793
	6	<i>tansig</i>	<i>trainlm</i>	0.9829	0.4988	0.6309	0.7943
	8	<i>tansig</i>	<i>trainlm</i>	0.9917	0.2201	0.2784	0.5276
	8	<i>tansig</i>	<i>traingdx</i>	0.9712	0.4982	0.6302	0.7939
	8	<i>logsig</i>	<i>traingdx</i>	0.9742	0.6964	0.8810	0.9386
	7	<i>tansig</i>	<i>traingdx</i>	0.5967	3.5043	4.4329	2.1054
ΔP (k+5)	7	<i>tansig</i>	<i>trainbr</i>	0.6573	41.4766	52.4490	7.2422
	8	<i>logsig</i>	<i>trainbr</i>	-27.4508	2490.9010	3149.8582	56.1236
	7	<i>logsig</i>	<i>trainlm</i>	-10.9546	963.4933	1218.3813	34.9053
	8	<i>tansig</i>	<i>trainbr</i>	-11.0535	2101.0571	2656.8828	51.5450
	7	<i>tansig</i>	<i>trainlm</i>	0.5905	24.1539	30.5438	5.5266
	7	<i>logsig</i>	<i>trainbr</i>	0.5555	125.0289	158.1048	12.5740
	6	<i>logsig</i>	<i>trainbr</i>	0.6414	36.7254	46.4410	6.8148
	6	<i>tansig</i>	<i>trainbr</i>	-0.0358	69.5040	87.8910	9.3750

6	<i>tansig</i>	<i>trainlm</i>	0.1875	61.7697	78.1105	8.8380
6	<i>logsig</i>	<i>trainlm</i>	0.9033	5.0087	6.3337	2.5167
6	<i>logsig</i>	<i>traingdx</i>	-0.1372	32.4990	41.0964	6.4106
6	<i>tansig</i>	<i>traingdx</i>	0.9678	0.7685	0.9717	0.9858

a - Using normalized data

Table C2: Performance values for each MLP topology for the experiment with 20 v/v% MEG.

Hidden Layer				$C_{MEG}$			
Variables	Number of neurons	Activation function	Training algorithm	20 v/v %	SSE <sup>a</sup>	MSE <sup>a</sup>	RMSE <sup>a</sup>
				$R^2$			
$\Delta P$ (k+1)	7	<i>tansig</i>	<i>trainlm</i>	0.8438	1.1654	1.4931	1.2219
	8	<i>tansig</i>	<i>trainbr</i>	0.5093	3.1002	3.9720	1.9930
	7	<i>logsig</i>	<i>trainbr</i>	0.9977	0.0156	0.0200	0.1415
	7	<i>tansig</i>	<i>trainbr</i>	0.9675	0.1834	0.2350	0.4848
	6	<i>logsig</i>	<i>trainbr</i>	0.9535	0.4052	0.5192	0.7205
	6	<i>tansig</i>	<i>trainbr</i>	0.9849	0.1033	0.1323	0.3638
	7	<i>logsig</i>	<i>trainlm</i>	0.9878	0.0694	0.0889	0.2982
	6	<i>tansig</i>	<i>trainlm</i>	0.9993	0.0044	0.0057	0.0755
	8	<i>tansig</i>	<i>trainlm</i>	0.9836	0.1103	0.1413	0.3759
	8	<i>tansig</i>	<i>traingdx</i>	0.9942	0.0380	0.0487	0.2208
	8	<i>logsig</i>	<i>traingdx</i>	0.9311	0.5920	0.7585	0.8709
	7	<i>tansig</i>	<i>traingdx</i>	0.6628	0.8787	1.1258	1.0610
$\Delta P$ (k+5)	7	<i>tansig</i>	<i>trainbr</i>	-0.0338	72.5838	92.9622	9.6417

8	<i>logsig</i>	<i>trainbr</i>	-18.7222	1210.2739	1550.0678	39.3709
7	<i>logsig</i>	<i>trainlm</i>	-8.9682	309.1142	395.9004	19.8972
8	<i>tansig</i>	<i>trainbr</i>	-61.5184	4641.2074	5944.2628	77.0990
7	<i>tansig</i>	<i>trainlm</i>	-8.2588	75.1440	96.2413	9.8103
7	<i>logsig</i>	<i>trainbr</i>	0.5621	47.5270	60.8705	7.8020
6	<i>logsig</i>	<i>trainbr</i>	-3.9839	205.6587	263.3990	16.2296
6	<i>tansig</i>	<i>trainbr</i>	0.5886	10.2781	13.1638	3.6282
6	<i>tansig</i>	<i>trainlm</i>	-6.5523	167.3544	214.3404	14.6404
6	<i>logsig</i>	<i>trainlm</i>	0.9446	1.2849	1.6456	1.2828
6	<i>logsig</i>	<i>traingdx</i>	-10.5975	87.5857	112.1761	10.5913
6	<i>tansig</i>	<i>traingdx</i>	0.8979	0.8438	1.0808	1.0396

a - Using normalized data

Table C3: Performance values for each MLP topology for the experiment with 30 v/v% MEG.

Hidden Layer				$C_{MEG}$			
				30 v/v %			
Variables	Number of neurons	Activation function	Training algorithm	$R^2$	SSE <sup>a</sup>	MSE <sup>a</sup>	RMSE <sup>a</sup>
$\Delta P$ (k+1)	7	<i>tansig</i>	<i>trainlm</i>	0.9460	0.4234	0.5183	0.7200
	8	<i>tansig</i>	<i>trainbr</i>	0.9744	0.1796	0.2199	0.4689
	7	<i>logsig</i>	<i>trainbr</i>	0.9990	0.0075	0.0091	0.0956
	7	<i>tansig</i>	<i>trainbr</i>	0.9843	0.1061	0.1299	0.3604
	6	<i>logsig</i>	<i>trainbr</i>	0.9414	0.5028	0.6156	0.7846
	6	<i>tansig</i>	<i>trainbr</i>	0.9958	0.0305	0.0373	0.1931

$\Delta P(k+5)$	7	<i>logsig</i>	<i>trainlm</i>	0.9900	0.0632	0.0774	0.2782
	6	<i>tansig</i>	<i>trainlm</i>	0.9921	0.0588	0.0720	0.2683
	8	<i>tansig</i>	<i>trainlm</i>	0.9996	0.0029	0.0036	0.0598
	8	<i>tansig</i>	<i>traingdx</i>	0.9877	0.1025	0.1255	0.3543
	8	<i>logsig</i>	<i>traingdx</i>	0.9582	0.3593	0.4399	0.6633
	7	<i>tansig</i>	<i>traingdx</i>	0.9093	0.3842	0.4703	0.6858
	7	<i>tansig</i>	<i>trainbr</i>	-571.6200	1366.2200	1672.0900	40.8900
	8	<i>logsig</i>	<i>trainbr</i>	-0.5428	59.3964	72.6940	8.5261
	7	<i>logsig</i>	<i>trainlm</i>	-8.1633	153.1347	187.4183	13.6901
	8	<i>tansig</i>	<i>trainbr</i>	-21.5072	1175.7158	1438.9338	37.9333
	7	<i>tansig</i>	<i>trainlm</i>	-14.3092	122.3726	149.7692	12.2380
	7	<i>logsig</i>	<i>trainbr</i>	0.3275	25.2963	30.9597	5.5641
	6	<i>logsig</i>	<i>trainbr</i>	-5.7828	213.8681	261.7487	16.1786
	6	<i>tansig</i>	<i>trainbr</i>	0.9467	1.1251	1.3770	1.1735
	6	<i>tansig</i>	<i>trainlm</i>	-15.0054	357.4552	437.4818	20.9161
	6	<i>logsig</i>	<i>trainlm</i>	0.9609	0.8400	1.0280	1.0139
	6	<i>logsig</i>	<i>traingdx</i>	-5.8235	54.0330	66.1299	8.1320
	6	<i>tansig</i>	<i>traingdx</i>	0.9450	0.5433	0.6650	0.8155

---

a - Using normalized data



Table C4: Performance values for each MLP topology for the experiment with 50 v/v% MEG.

Hidden Layer				C <sub>MEG</sub>			
Variables	Number of neurons	Activation function	Training algorithm	50 v/v %			
				R <sup>2</sup>	SSE <sup>a</sup>	MSE <sup>a</sup>	RMSE <sup>a</sup>
ΔP (k+1)	7	<i>tansig</i>	<i>trainlm</i>	0.9376	0.5845	0.4042	0.6358
	8	<i>tansig</i>	<i>trainbr</i>	0.9715	0.2954	0.2043	0.4520
	7	<i>logsig</i>	<i>trainbr</i>	0.9981	0.0192	0.0133	0.1152
	7	<i>tansig</i>	<i>trainbr</i>	0.9586	0.4245	0.2936	0.5418
	6	<i>logsig</i>	<i>trainbr</i>	0.9671	0.2896	0.2003	0.4475
	6	<i>tansig</i>	<i>trainbr</i>	0.9908	0.0901	0.0623	0.2496
	7	<i>logsig</i>	<i>trainlm</i>	0.9885	0.1086	0.0751	0.2740
	6	<i>tansig</i>	<i>trainlm</i>	0.9768	0.2241	0.1550	0.3937
	8	<i>tansig</i>	<i>trainlm</i>	0.9950	0.0498	0.0345	0.1856
	8	<i>tansig</i>	<i>traingdx</i>	0.9747	0.3215	0.2223	0.4715
	8	<i>logsig</i>	<i>traingdx</i>	0.9776	0.1919	0.1327	0.3643
	7	<i>tansig</i>	<i>traingdx</i>	0.9625	0.3940	0.2725	0.5220
ΔP (k+5)	7	<i>tansig</i>	<i>trainbr</i>	0.0298	9.4548	6.5365	2.5567
	8	<i>logsig</i>	<i>trainbr</i>	-3.9217	43.3274	29.9541	5.4730
	7	<i>logsig</i>	<i>trainlm</i>	-18.7737	320.6140	221.6544	14.8881
	8	<i>tansig</i>	<i>trainbr</i>	-484.8908	287.0268	198.4341	14.0867
	7	<i>tansig</i>	<i>trainlm</i>	-28.9126	267.9047	185.2141	13.6093
	7	<i>logsig</i>	<i>trainbr</i>	-1.1301	20.5066	14.1771	3.7652
	6	<i>logsig</i>	<i>trainbr</i>	-17.8273	27985.8641	19347.8407	139.0965

6	<i>tansig</i>	<i>trainbr</i>	0.8653	2.5848	1.7870	1.3368
6	<i>tansig</i>	<i>trainlm</i>	-1.1671	45.6665	31.5712	5.6188
6	<i>logsig</i>	<i>trainlm</i>	0.7974	3.3790	2.3360	1.5284
6	<i>logsig</i>	<i>traingdx</i>	-9.0137	93.6943	64.7749	8.0483
6	<i>tansig</i>	<i>traingdx</i>	0.5214	6.2932	4.3507	2.0858

a - Using normalized data

Table C5. Performance values for each MLP topology for the experiment with 60 v/v% MEG.

Hidden Layer				$C_{MEG}$			
				60 v/v %			
Variables	Number of neurons	Activation function	Training algorithm	$R^2$	SSE <sup>a</sup>	MSE <sup>a</sup>	RMSE <sup>a</sup>
$\Delta P$ (k+1)	7	<i>tansig</i>	<i>trainlm</i>	0.8635	1.5935	0.7318	0.8554
	8	<i>tansig</i>	<i>trainbr</i>	0.8520	1.9841	0.9111	0.9545
	7	<i>logsig</i>	<i>trainbr</i>	0.9992	0.0109	0.0050	0.0707
	7	<i>tansig</i>	<i>trainbr</i>	0.8801	1.5705	0.7212	0.8492
	6	<i>logsig</i>	<i>trainbr</i>	0.9716	0.3115	0.1431	0.3782
	6	<i>tansig</i>	<i>trainbr</i>	0.9894	0.1334	0.0612	0.2475
	7	<i>logsig</i>	<i>trainlm</i>	0.9931	0.0859	0.0394	0.1986
	6	<i>tansig</i>	<i>trainlm</i>	0.9945	0.0672	0.0309	0.1757
	8	<i>tansig</i>	<i>trainlm</i>	0.9298	0.8811	0.4046	0.6361
	8	<i>tansig</i>	<i>traingdx</i>	0.9799	0.3414	0.1568	0.3959
	8	<i>logsig</i>	<i>traingdx</i>	0.9442	0.5271	0.2421	0.4920
	7	<i>tansig</i>	<i>traingdx</i>	0.9346	1.0160	0.4665	0.6830

$\Delta P$ (k+5)	7	<i>tansig</i>	<i>trainbr</i>	0.7816	3.2599	1.4964	1.2233
	8	<i>logsig</i>	<i>trainbr</i>	-62.3342	229.2742	105.2466	10.2590
	7	<i>logsig</i>	<i>trainlm</i>	-20.8211	326.8558	150.0407	12.2491
	8	<i>tansig</i>	<i>trainbr</i>	-577.0620	506.9002	232.6887	15.2541
	7	<i>tansig</i>	<i>trainlm</i>	-17.8622	218.2728	100.1965	10.0098
	7	<i>logsig</i>	<i>trainbr</i>	-3.1168	48.8241	22.4123	4.7342
	6	<i>logsig</i>	<i>trainbr</i>	-12.0190	362993.7519	166629.5401	408.2028
	6	<i>tansig</i>	<i>trainbr</i>	0.8718	2.8607	1.3132	1.1459
	6	<i>tansig</i>	<i>trainlm</i>	-1.3260	39.5380	18.1496	4.2602
	6	<i>logsig</i>	<i>trainlm</i>	0.8318	2.8749	1.3197	1.1488
	6	<i>logsig</i>	<i>traingdx</i>	-17.4929	219.2948	100.6656	10.0332
	6	<i>tansig</i>	<i>traingdx</i>	0.8104	3.0840	1.4157	1.1898

a - Using normalized data

Table C6: Performance values for each MLP topology for the experiment with 70 v/v% MEG.

Hidden Layer				$C_{MEG}$			
				70 v/v %			
Variables	Number of neurons	Activation function	Training algorithm	$R^2$	SSE <sup>a</sup>	MSE <sup>a</sup>	RMSE <sup>a</sup>
$\Delta P$ (k+1)	7	<i>tansig</i>	<i>trainlm</i>	0.9579	2.0775	0.5717	0.7561
	8	<i>tansig</i>	<i>trainbr</i>	0.9490	2.7186	0.7482	0.8650
	7	<i>logsig</i>	<i>trainbr</i>	0.9992	0.0435	0.0120	0.1094
	7	<i>tansig</i>	<i>trainbr</i>	0.9662	1.7686	0.4867	0.6977
	6	<i>logsig</i>	<i>trainbr</i>	0.9498	2.3628	0.6503	0.8064

$\Delta P(k+5)$	6	<i>tansig</i>	<i>trainbr</i>	0.9980	0.1014	0.0279	0.1670
	7	<i>logsig</i>	<i>trainlm</i>	0.9992	0.0433	0.0119	0.1092
	6	<i>tansig</i>	<i>trainlm</i>	0.9993	0.0340	0.0093	0.0967
	8	<i>tansig</i>	<i>trainlm</i>	0.8533	7.2194	1.9868	1.4095
	8	<i>tansig</i>	<i>traingdx</i>	0.9746	1.8028	0.4962	0.7044
	8	<i>logsig</i>	<i>traingdx</i>	0.9434	1.8831	0.5182	0.7199
	7	<i>tansig</i>	<i>traingdx</i>	0.9554	2.8289	0.7785	0.8823
	7	<i>tansig</i>	<i>trainbr</i>	0.9884	0.7300	0.2008	0.4482
	8	<i>logsig</i>	<i>trainbr</i>	-10.5339	161.0831	44.3155	6.6570
	7	<i>logsig</i>	<i>trainlm</i>	0.8767	5.8786	1.6173	1.2717
	8	<i>tansig</i>	<i>trainbr</i>	-15.9223	45.2201	12.4405	3.5271
	7	<i>tansig</i>	<i>trainlm</i>	-0.7421	89.2030	24.5406	4.9538
	7	<i>logsig</i>	<i>trainbr</i>	0.5290	25.7058	7.0719	2.6593
	6	<i>logsig</i>	<i>trainbr</i>	-401.5163	40425.2866	11121.3776	105.4579
	6	<i>tansig</i>	<i>trainbr</i>	0.9326	5.1796	1.4250	1.1937
	6	<i>tansig</i>	<i>trainlm</i>	-1.1537	150.2429	41.3332	6.4291
	6	<i>logsig</i>	<i>trainlm</i>	0.9639	2.2455	0.6178	0.7860
	6	<i>logsig</i>	<i>traingdx</i>	-2.5397	175.3307	48.2351	6.9452
	6	<i>tansig</i>	<i>traingdx</i>	0.9867	0.8371	0.2303	0.4799

---

a - Using normalized data

Table C7: Optimized parameters (weight and bias) of the MLP logsig\_7\_purelin\_1\_trainbr used to predict the  $\Delta P_{(t+1)}$ .

Parameters connecting the input and hidden neurons								Parameters connecting the hidden and output neurons		
	wj1 (i=1)	wj2 (i=2)	wj3 (i=3)	wj4 (i=4)	wj5 (i=5)	wj6 (i=6)	bj1 (i=1)		wj1 (k=1)	bk (k=1)
j=1	-0.0719	0.5510	-0.8545	0.9714	1.6225	0.0680	0.27177	j=1	6.3578	-0.1811
j=2	0.6330	-2.0230	2.8739	-0.5121	-0.6166	-1.6893	1.421	j=2	-0.4174	
j=3	1.4327	-0.9769	-0.7022	-0.7209	0.7137	-0.3642	-0.3152	j=3	-4.2264	
j=4	0.8456	0.7333	0.3548	0.2277	0.2538	-0.1296	-0.7738	j=4	-7.4539	
j=5	8.2765	-5.4457	-0.6467	-3.5122	5.3282	-0.9758	-1.5888	j=5	0.68132952	
j=6	0.0007	-0.5577	-0.4325	0.7543	-0.8844	-0.1982	-0.263	j=6	-8.9164202	
j=7	1.0517	-0.6232	0.1544	0.0566	-1.4625	0.0042	-0.0744	j=7	10.1681218	

Table C8. Optimized parameters (weight and bias) of the MLP logsig\_6\_purelin\_1\_trainlm used to predict the  $\Delta P_{(t+5)}$ .

Parameters connecting the input and hidden neurons								Parameters connecting the hidden and output neurons		
	wj1 (i=1)	wj2 (i=2)	wj3 (i=3)	wj4 (i=4)	wj5 (i=5)	wj6 (i=6)	bj1 (i=1)		wj1 (k=1)	bk (k=1)
j=1	0.1772	-0.0609	4.2388	-0.0234	0.0847	-2.8046	1.66666	j=1	50.6089	-52.5521
j=2	-0.3512	-0.3111	0.1776	0.2761	-0.3175	1.6985	1.42789	j=2	2.6236	
j=3	-5.4440	-1.9806	-0.1860	8.7545	-6.9442	-3.0622	-3.1773	j=3	-0.4000	
j=4	-1.8506	1.6964	1.6203	-3.4523	1.8862	3.7338	-0.0657	j=4	27.3521	
j=5	-0.1484	0.1629	-4.2434	0.0656	-0.0813	2.8580	-1.5943	j=5	50.2158673	
j=6	-1.8321	1.6810	1.6508	-3.4241	1.8643	3.6932	-0.0901	j=6	-27.555778	

**D**

**Supporting information of the article: Machine learning models for measurement of pH using a low-cost image analysis strategy**

## **Supplementary material**

# Machine learning models for measurement of pH using a low-cost image analysis strategy: Study case in incrustation

*Bruno X. Ferreira<sup>1</sup>, Alline V. B. de Oliveira<sup>2</sup>, João Cajaíba<sup>2</sup>, Vinicius Kartnaller<sup>2</sup>,  
Brunno F. Santos<sup>1\*</sup>.*

AUTHOR ADDRESS:

*a* - Department of Chemical and Materials Engineering (DEQM), Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rua Marquês de São Vicente, 225 – Gávea, Rio de Janeiro, RJ 22430-060, Brazil.

*b* - Instituto de Química, Núcleo de Desenvolvimento de Processos e Análises Químicas em Tempo Real (NQTR), Universidade Federal do Rio de Janeiro, Rua Hélio de Almeida 40, Cidade Universitária, Rio de Janeiro 21941-614, Brazil.

[\\*bsantos@puc-rio.br](mailto:bsantos@puc-rio.br)

Table D1: Performance values for all SVM models (PR = Precision, RC = Recall, ACC = Accuracy)

Model ID	Hyperparameters					Training time (s)	Training group			Validation group			Test group		
	Dec_func_shape	C	Kernel	degree	gamma		PR	RC	ACC	PR	RC	ACC	PR	RC	ACC
SVM_1	OvO	0.01	poly	1	auto	73.6327	1.000	1.000	1.000	0.943	0.947	0.936	0.981	0.963	0.979
SVM_2	OvO	0.01	poly	2	auto	67.5487	1.000	1.000	1.000	0.935	0.938	0.936	0.981	0.963	0.979
SVM_3	OvO	0.01	poly	3	auto	65.6985	1.000	1.000	1.000	0.935	0.938	0.936	0.981	0.963	0.979
SVM_4	OvR	0.01	poly	1	auto	78.1301	1.000	1.000	1.000	0.943	0.947	0.936	0.981	0.963	0.979
SVM_5	OvR	0.01	poly	2	auto	68.4474	1.000	1.000	1.000	0.935	0.938	0.936	0.981	0.963	0.979
SVM_6	OvR	0.01	poly	3	auto	66.1859	1.000	1.000	1.000	0.935	0.938	0.936	0.981	0.963	0.979
SVM_7	OvO	0.01	linear	-	-	73.9518	1.000	1.000	1.000	0.943	0.947	0.936	0.944	0.947	0.957
SVM_8	OvR	0.01	linear	-	-	72.3457	1.000	1.000	1.000	0.943	0.947	0.936	0.944	0.947	0.957
SVM_9	OvO	0.01	poly	4	auto	66.4678	1.000	1.000	1.000	0.916	0.916	0.915	0.944	0.947	0.957
SVM_10	OvR	0.01	poly	4	auto	66.5141	1.000	1.000	1.000	0.916	0.916	0.915	0.944	0.947	0.957
SVM_11	OvO	0.1	linear	-	-	71.9084	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_12	OvO	0.1	poly	1	auto	72.1042	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_13	OvR	0.1	linear	-	-	73.1657	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_14	OvR	0.1	poly	1	auto	71.7229	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_15	OvO	0.5	linear	-	-	20.4319	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_16	OvO	0.5	poly	1	auto	20.3909	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_17	OvR	0.5	linear	-	-	19.9021	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_18	OvR	0.5	poly	1	auto	20.2430	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_19	OvO	1	linear	-	-	19.5853	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_20	OvO	1	poly	1	auto	19.7388	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957

SVM_21	OvR	1	linear	-	-	20.7276	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_22	OvR	1	poly	1	auto	20.2096	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_23	OvO	2	linear	-	-	20.4769	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_24	OvO	2	poly	1	auto	20.1835	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_25	OvR	2	linear	-	-	20.1731	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_26	OvR	2	poly	1	auto	20.2983	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_27	OvO	5	linear	-	-	19.9585	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_28	OvO	5	poly	1	auto	20.7580	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_29	OvR	5	linear	-	-	20.2323	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_30	OvR	5	poly	1	auto	20.0705	1.000	1.000	1.000	0.932	0.899	0.894	0.956	0.956	0.957
SVM_31	OvO	0.1	poly	2	auto	68.8614	1.000	1.000	1.000	0.914	0.881	0.872	0.944	0.940	0.936
SVM_32	OvR	0.1	poly	2	auto	67.6416	1.000	1.000	1.000	0.914	0.881	0.872	0.944	0.940	0.936
SVM_33	OvO	0.5	poly	2	auto	19.0878	1.000	1.000	1.000	0.914	0.881	0.872	0.944	0.940	0.936
SVM_34	OvR	0.5	poly	2	auto	18.8967	1.000	1.000	1.000	0.914	0.881	0.872	0.944	0.940	0.936
SVM_35	OvO	1	poly	2	auto	18.9032	1.000	1.000	1.000	0.914	0.881	0.872	0.944	0.940	0.936
SVM_36	OvR	1	poly	2	auto	18.7249	1.000	1.000	1.000	0.914	0.881	0.872	0.944	0.940	0.936
SVM_37	OvO	2	poly	2	auto	18.8447	1.000	1.000	1.000	0.914	0.881	0.872	0.944	0.940	0.936
SVM_38	OvR	2	poly	2	auto	20.1455	1.000	1.000	1.000	0.914	0.881	0.872	0.944	0.940	0.936
SVM_39	OvO	5	poly	2	auto	18.9318	1.000	1.000	1.000	0.914	0.881	0.872	0.944	0.940	0.936
SVM_40	OvO	5	rbf	-	scale	83.6070	0.992	0.992	0.991	0.932	0.913	0.915	0.944	0.940	0.936
SVM_41	OvR	5	poly	2	auto	18.7070	1.000	1.000	1.000	0.914	0.881	0.872	0.944	0.940	0.936
SVM_42	OvR	5	rbf	-	scale	57.6971	0.992	0.992	0.991	0.932	0.913	0.915	0.944	0.940	0.936
SVM_43	OvO	0.1	poly	3	auto	67.6992	1.000	1.000	1.000	0.895	0.881	0.872	0.937	0.924	0.915
SVM_44	OvO	0.1	poly	4	auto	66.4869	1.000	1.000	1.000	0.860	0.853	0.851	0.944	0.940	0.936
SVM_45	OvR	0.1	poly	3	auto	66.4615	1.000	1.000	1.000	0.895	0.881	0.872	0.937	0.924	0.915



SVM_46	OvR	0.1	poly	4	auto	65.4905	1.000	1.000	1.000	0.860	0.853	0.851	0.944	0.940	0.936
SVM_47	OvO	0.5	poly	3	auto	18.8342	1.000	1.000	1.000	0.895	0.881	0.872	0.937	0.924	0.915
SVM_48	OvO	0.5	poly	4	auto	17.9543	1.000	1.000	1.000	0.860	0.853	0.851	0.944	0.940	0.936
SVM_49	OvR	0.5	poly	3	auto	18.4414	1.000	1.000	1.000	0.895	0.881	0.872	0.937	0.924	0.915
SVM_50	OvR	0.5	poly	4	auto	17.8248	1.000	1.000	1.000	0.860	0.853	0.851	0.944	0.940	0.936
SVM_51	OvO	1	poly	3	auto	18.3908	1.000	1.000	1.000	0.895	0.881	0.872	0.937	0.924	0.915
SVM_52	OvO	1	poly	4	auto	18.0865	1.000	1.000	1.000	0.860	0.853	0.851	0.944	0.940	0.936
SVM_53	OvR	1	poly	3	auto	18.4174	1.000	1.000	1.000	0.895	0.881	0.872	0.937	0.924	0.915
SVM_54	OvR	1	poly	4	auto	17.7328	1.000	1.000	1.000	0.860	0.853	0.851	0.944	0.940	0.936
SVM_55	OvO	2	poly	3	auto	19.3454	1.000	1.000	1.000	0.895	0.881	0.872	0.937	0.924	0.915
SVM_56	OvO	2	poly	4	auto	18.4938	1.000	1.000	1.000	0.860	0.853	0.851	0.944	0.940	0.936
SVM_57	OvR	2	poly	3	auto	18.6079	1.000	1.000	1.000	0.895	0.881	0.872	0.937	0.924	0.915
SVM_58	OvR	2	poly	4	auto	18.1854	1.000	1.000	1.000	0.860	0.853	0.851	0.944	0.940	0.936
SVM_59	OvO	5	poly	3	auto	18.3817	1.000	1.000	1.000	0.895	0.881	0.872	0.937	0.924	0.915
SVM_60	OvO	5	poly	4	auto	20.2391	1.000	1.000	1.000	0.860	0.853	0.851	0.944	0.940	0.936
SVM_61	OvR	5	poly	3	auto	19.2050	1.000	1.000	1.000	0.895	0.881	0.872	0.937	0.924	0.915
SVM_62	OvR	5	poly	4	auto	18.3742	1.000	1.000	1.000	0.860	0.853	0.851	0.944	0.940	0.936
SVM_63	OvO	0.01	poly	5	auto	65.9490	1.000	1.000	1.000	0.888	0.842	0.872	0.878	0.888	0.894
SVM_64	OvO	0.01	poly	6	auto	65.6963	1.000	1.000	1.000	0.885	0.842	0.872	0.878	0.888	0.894
SVM_65	OvR	0.01	poly	5	auto	66.2891	1.000	1.000	1.000	0.888	0.842	0.872	0.878	0.888	0.894
SVM_66	OvR	0.01	poly	6	auto	67.2551	1.000	1.000	1.000	0.885	0.842	0.872	0.878	0.888	0.894
SVM_67	OvO	0.1	poly	5	auto	67.5335	1.000	1.000	1.000	0.860	0.853	0.851	0.883	0.903	0.894
SVM_68	OvR	0.1	poly	5	auto	65.7002	1.000	1.000	1.000	0.860	0.853	0.851	0.883	0.903	0.894
SVM_69	OvO	0.5	poly	5	auto	18.4619	1.000	1.000	1.000	0.860	0.853	0.851	0.883	0.903	0.894
SVM_70	OvR	0.5	poly	5	auto	18.2668	1.000	1.000	1.000	0.860	0.853	0.851	0.883	0.903	0.894

SVM_71	OvO	1	poly	5	auto	18.2470	1.000	1.000	1.000	0.860	0.853	0.851	0.883	0.903	0.894
SVM_72	OvR	1	poly	5	auto	18.4467	1.000	1.000	1.000	0.860	0.853	0.851	0.883	0.903	0.894
SVM_73	OvO	2	poly	5	auto	18.6633	1.000	1.000	1.000	0.860	0.853	0.851	0.883	0.903	0.894
SVM_74	OvR	2	poly	5	auto	18.5536	1.000	1.000	1.000	0.860	0.853	0.851	0.883	0.903	0.894
SVM_75	OvO	5	poly	5	auto	18.4825	1.000	1.000	1.000	0.860	0.853	0.851	0.883	0.903	0.894
SVM_76	OvR	5	poly	5	auto	18.5648	1.000	1.000	1.000	0.860	0.853	0.851	0.883	0.903	0.894
SVM_77	OvO	5	poly	2	scale	19.5816	0.989	0.985	0.986	0.922	0.897	0.894	0.900	0.903	0.894
SVM_78	OvR	5	poly	2	scale	19.5647	0.989	0.985	0.986	0.922	0.897	0.894	0.900	0.903	0.894
SVM_79	OvO	0.1	poly	6	auto	66.8405	1.000	1.000	1.000	0.815	0.757	0.787	0.863	0.880	0.872
SVM_80	OvR	0.1	poly	6	auto	65.7677	1.000	1.000	1.000	0.815	0.757	0.787	0.863	0.880	0.872
SVM_81	OvO	0.5	poly	6	auto	23.6465	1.000	1.000	1.000	0.815	0.757	0.787	0.863	0.880	0.872
SVM_82	OvR	0.5	poly	6	auto	17.9289	1.000	1.000	1.000	0.815	0.757	0.787	0.863	0.880	0.872
SVM_83	OvO	1	poly	6	auto	18.1661	1.000	1.000	1.000	0.815	0.757	0.787	0.863	0.880	0.872
SVM_84	OvR	1	poly	6	auto	18.1014	1.000	1.000	1.000	0.815	0.757	0.787	0.863	0.880	0.872
SVM_85	OvO	2	poly	6	auto	18.2913	1.000	1.000	1.000	0.815	0.757	0.787	0.863	0.880	0.872
SVM_86	OvR	2	poly	6	auto	18.0889	1.000	1.000	1.000	0.815	0.757	0.787	0.863	0.880	0.872
SVM_87	OvO	5	poly	3	scale	19.8430	0.985	0.980	0.982	0.875	0.832	0.851	0.889	0.909	0.894
SVM_88	OvO	5	poly	6	auto	18.3918	1.000	1.000	1.000	0.815	0.757	0.787	0.863	0.880	0.872
SVM_89	OvR	5	poly	3	scale	19.4896	0.985	0.980	0.982	0.875	0.832	0.851	0.889	0.909	0.894
SVM_90	OvR	5	poly	6	auto	18.4311	1.000	1.000	1.000	0.815	0.757	0.787	0.863	0.880	0.872
SVM_91	OvO	5	poly	1	scale	20.5908	0.959	0.954	0.954	0.922	0.897	0.894	0.929	0.917	0.915
SVM_92	OvR	5	poly	1	scale	20.3333	0.959	0.954	0.954	0.922	0.897	0.894	0.929	0.917	0.915
SVM_93	OvO	2	poly	2	scale	20.8323	0.959	0.949	0.950	0.886	0.823	0.851	0.900	0.903	0.894
SVM_94	OvR	2	poly	2	scale	20.1677	0.959	0.949	0.950	0.886	0.823	0.851	0.900	0.903	0.894
SVM_95	OvO	2	rbf	-	scale	82.8456	0.947	0.938	0.941	0.916	0.876	0.894	0.884	0.880	0.872

SVM_96	OvR	2	rbf	-	scale	84.4934	0.947	0.938	0.941	0.916	0.876	0.894	0.884	0.880	0.872
SVM_97	OvO	5	poly	4	scale	20.0378	0.960	0.949	0.950	0.831	0.758	0.787	0.908	0.925	0.915
SVM_98	OvR	5	poly	4	scale	19.5585	0.960	0.949	0.950	0.831	0.758	0.787	0.908	0.925	0.915
SVM_99	OvO	2	poly	1	scale	24.6015	0.944	0.932	0.936	0.784	0.802	0.851	0.879	0.880	0.872
SVM_100	OvR	2	poly	1	scale	21.8771	0.944	0.932	0.936	0.784	0.802	0.851	0.879	0.880	0.872
SVM_101	OvO	2	poly	3	scale	20.7346	0.950	0.935	0.936	0.853	0.786	0.809	0.895	0.887	0.872
SVM_102	OvR	2	poly	3	scale	20.1369	0.950	0.935	0.936	0.853	0.786	0.809	0.895	0.887	0.872
SVM_103	OvO	1	poly	2	scale	21.5510	0.935	0.911	0.918	0.748	0.765	0.809	0.860	0.843	0.830
SVM_104	OvR	1	poly	2	scale	21.5702	0.935	0.911	0.918	0.748	0.765	0.809	0.860	0.843	0.830
SVM_105	OvO	2	poly	4	scale	20.2783	0.937	0.909	0.909	0.807	0.721	0.745	0.939	0.925	0.915
SVM_106	OvR	2	poly	4	scale	20.2763	0.937	0.909	0.909	0.807	0.721	0.745	0.939	0.925	0.915
SVM_107	OvO	1	rbf	-	scale	81.6530	0.917	0.896	0.904	0.794	0.806	0.851	0.849	0.825	0.809
SVM_108	OvR	1	rbf	-	scale	85.2323	0.917	0.896	0.904	0.794	0.806	0.851	0.849	0.825	0.809
SVM_109	OvO	5	poly	5	scale	20.1014	0.940	0.913	0.913	0.760	0.696	0.723	0.878	0.847	0.851
SVM_110	OvR	5	poly	5	scale	20.1531	0.940	0.913	0.913	0.760	0.696	0.723	0.878	0.847	0.851
SVM_111	OvO	1	poly	3	scale	20.9656	0.925	0.891	0.890	0.710	0.712	0.745	0.867	0.850	0.830
SVM_112	OvR	1	poly	3	scale	21.4649	0.925	0.891	0.890	0.710	0.712	0.745	0.867	0.850	0.830
SVM_113	OvO	2	poly	5	scale	21.3406	0.943	0.884	0.886	0.848	0.702	0.723	0.830	0.769	0.787

SVM_11 4	OvR	2	poly	5	scale	20.2931	0.943	0.884	0.886	0.848	0.702	0.723	0.830	0.769	0.787
SVM_11 5	OvO	5	poly	6	scale	20.2130	0.953	0.894	0.895	0.774	0.637	0.660	0.835	0.792	0.809
SVM_11 6	OvR	5	poly	6	scale	20.4282	0.953	0.894	0.895	0.774	0.637	0.660	0.835	0.792	0.809
SVM_11 7	OvO	1	poly	1	scale	23.5627	0.880	0.853	0.863	0.731	0.727	0.766	0.835	0.809	0.787
SVM_11 8	OvR	1	poly	1	scale	24.9392	0.880	0.853	0.863	0.731	0.727	0.766	0.835	0.809	0.787
SVM_11 9	OvO	0.5	poly	2	scale	24.1426	0.868	0.833	0.840	0.671	0.660	0.702	0.857	0.841	0.830
SVM_12 0	OvR	0.5	poly	2	scale	24.0889	0.868	0.833	0.840	0.671	0.660	0.702	0.857	0.841	0.830
SVM_12 1	OvO	1	poly	4	scale	21.0243	0.922	0.856	0.858	0.702	0.647	0.681	0.821	0.757	0.766
SVM_12 2	OvR	1	poly	4	scale	21.2123	0.922	0.856	0.858	0.702	0.647	0.681	0.821	0.757	0.766
SVM_12 3	OvO	1	poly	5	scale	21.2567	0.925	0.855	0.858	0.698	0.644	0.681	0.789	0.714	0.745
SVM_12 4	OvR	1	poly	5	scale	21.1990	0.925	0.855	0.858	0.698	0.644	0.681	0.789	0.714	0.745
SVM_12 5	OvO	2	poly	6	scale	21.2306	0.930	0.864	0.868	0.617	0.578	0.617	0.778	0.692	0.723
SVM_12 6	OvR	2	poly	6	scale	31.5717	0.930	0.864	0.868	0.617	0.578	0.617	0.778	0.692	0.723
SVM_12 7	OvO	1	poly	6	scale	25.5722	0.922	0.845	0.849	0.698	0.644	0.681	0.778	0.692	0.723
SVM_12 8	OvR	1	poly	6	scale	21.3149	0.922	0.845	0.849	0.698	0.644	0.681	0.778	0.692	0.723
SVM_12 9	OvO	0.5	poly	4	scale	22.9189	0.901	0.826	0.831	0.667	0.622	0.660	0.694	0.717	0.745
SVM_13 0	OvR	0.5	poly	4	scale	25.3530	0.901	0.826	0.831	0.667	0.622	0.660	0.694	0.717	0.745

SVM_13 1	OvO	0.5	poly	3	scale	23.0260	0.887	0.810	0.817	0.727	0.690	0.723	0.675	0.716	0.723
SVM_13 2	OvR	0.5	poly	3	scale	22.5973	0.887	0.810	0.817	0.727	0.690	0.723	0.675	0.716	0.723
SVM_13 3	OvO	0.5	poly	5	scale	22.7091	0.904	0.819	0.822	0.667	0.622	0.660	0.671	0.695	0.723
SVM_13 4	OvR	0.5	poly	5	scale	22.4033	0.904	0.819	0.822	0.667	0.622	0.660	0.671	0.695	0.723
SVM_13 5	OvO	0.5	rbf	-	scale	59.6960	0.859	0.802	0.822	0.682	0.648	0.660	0.685	0.734	0.702
SVM_13 6	OvR	0.5	rbf	-	scale	59.7815	0.859	0.802	0.822	0.682	0.648	0.660	0.685	0.734	0.702
SVM_13 7	OvO	0.5	poly	6	scale	22.7071	0.926	0.803	0.804	0.665	0.594	0.638	0.657	0.673	0.702
SVM_13 8	OvR	0.5	poly	6	scale	22.4208	0.926	0.803	0.804	0.665	0.594	0.638	0.657	0.673	0.702
SVM_13 9	OvO	1	rbf	-	auto	118.8883	1.000	1.000	1.000	0.017	0.111	0.149	0.017	0.111	0.149
SVM_14 0	OvR	1	rbf	-	auto	124.4400	1.000	1.000	1.000	0.017	0.111	0.149	0.017	0.111	0.149
SVM_14 1	OvO	2	rbf	-	auto	124.8738	1.000	1.000	1.000	0.017	0.111	0.149	0.017	0.111	0.149
SVM_14 2	OvR	2	rbf	-	auto	125.8524	1.000	1.000	1.000	0.017	0.111	0.149	0.017	0.111	0.149
SVM_14 3	OvO	5	rbf	-	auto	124.1344	1.000	1.000	1.000	0.017	0.111	0.149	0.017	0.111	0.149
SVM_14 4	OvR	5	rbf	-	auto	82.9192	1.000	1.000	1.000	0.017	0.111	0.149	0.017	0.111	0.149
SVM_14 5	OvO	0.1	poly	6	scale	96.6185	0.794	0.734	0.735	0.607	0.511	0.532	0.678	0.677	0.702
SVM_14 6	OvR	0.1	poly	6	scale	97.3522	0.794	0.734	0.735	0.607	0.511	0.532	0.678	0.677	0.702
SVM_14 7	OvO	0.1	poly	5	scale	99.1677	0.779	0.723	0.726	0.581	0.503	0.532	0.678	0.677	0.702

SVM_14 8	OvR	0.1	poly	5	scale	99.9254	0.779	0.723	0.726	0.581	0.503	0.532	0.678	0.677	0.702
SVM_14 9	OvO	0.5	poly	1	scale	28.0745	0.689	0.684	0.717	0.622	0.593	0.596	0.563	0.712	0.681
SVM_15 0	OvR	0.5	poly	1	scale	27.4439	0.689	0.684	0.717	0.622	0.593	0.596	0.563	0.712	0.681
SVM_15 1	OvO	0.1	poly	4	scale	103.2796	0.748	0.698	0.703	0.502	0.489	0.532	0.650	0.654	0.681
SVM_15 2	OvR	0.1	poly	4	scale	102.9692	0.748	0.698	0.703	0.502	0.489	0.532	0.650	0.654	0.681
SVM_15 3	OvO	0.1	poly	3	scale	110.1472	0.719	0.660	0.671	0.523	0.461	0.489	0.646	0.661	0.681
SVM_15 4	OvR	0.1	poly	3	scale	110.0066	0.719	0.660	0.671	0.523	0.461	0.489	0.646	0.661	0.681
SVM_15 5	OvO	2	sigmoid	-	scale	30.3691	0.558	0.568	0.612	0.517	0.380	0.426	0.478	0.481	0.489
SVM_15 6	OvR	2	sigmoid	-	scale	29.9800	0.558	0.568	0.612	0.517	0.380	0.426	0.478	0.481	0.489
SVM_15 7	OvO	0.1	poly	2	scale	121.9873	0.444	0.548	0.575	0.355	0.430	0.426	0.469	0.601	0.617
SVM_15 8	OvR	0.1	poly	2	scale	121.8429	0.444	0.548	0.575	0.355	0.430	0.426	0.469	0.601	0.617
SVM_15 9	OvO	1	sigmoid	-	scale	33.8316	0.374	0.490	0.562	0.388	0.412	0.468	0.360	0.524	0.532
SVM_16 0	OvR	1	sigmoid	-	scale	34.1463	0.374	0.490	0.562	0.388	0.412	0.468	0.360	0.524	0.532
SVM_16 1	OvO	0.5	sigmoid	-	scale	37.9933	0.292	0.360	0.425	0.262	0.296	0.340	0.371	0.437	0.468
SVM_16 2	OvR	0.5	sigmoid	-	scale	38.5091	0.292	0.360	0.425	0.262	0.296	0.340	0.371	0.437	0.468
SVM_16 3	OvO	5	sigmoid	-	scale	25.5321	0.661	0.416	0.457	0.500	0.298	0.319	0.407	0.331	0.319
SVM_16 4	OvR	5	sigmoid	-	scale	25.2560	0.661	0.416	0.457	0.500	0.298	0.319	0.407	0.331	0.319

5	SVM_16	OvO	0.01	poly	6	scale	124.0644	0.518	0.402	0.402	0.561	0.449	0.447	0.474	0.398	0.404
6	SVM_16	OvR	0.01	poly	6	scale	123.6579	0.518	0.402	0.402	0.561	0.449	0.447	0.474	0.398	0.404
7	SVM_16	OvO	0.1	rbf	-	scale	163.5958	0.227	0.367	0.406	0.203	0.324	0.319	0.213	0.370	0.383
8	SVM_16	OvR	0.1	rbf	-	scale	163.2050	0.227	0.367	0.406	0.203	0.324	0.319	0.213	0.370	0.383
9	SVM_16	OvO	0.01	poly	5	scale	128.9090	0.534	0.387	0.388	0.376	0.394	0.383	0.317	0.325	0.319
0	SVM_17	OvR	0.01	poly	5	scale	130.9366	0.534	0.387	0.388	0.376	0.394	0.383	0.317	0.325	0.319
1	SVM_17	OvO	0.1	poly	1	scale	146.0755	0.210	0.350	0.388	0.192	0.296	0.298	0.199	0.333	0.362
2	SVM_17	OvR	0.1	poly	1	scale	145.9695	0.210	0.350	0.388	0.192	0.296	0.298	0.199	0.333	0.362
3	SVM_17	OvO	0.01	poly	4	scale	135.6832	0.413	0.358	0.361	0.376	0.394	0.383	0.347	0.344	0.340
4	SVM_17	OvR	0.01	poly	4	scale	138.2390	0.413	0.358	0.361	0.376	0.394	0.383	0.347	0.344	0.340
5	SVM_17	OvO	0.1	sigmoid	-	scale	155.6791	0.069	0.222	0.288	0.061	0.222	0.255	0.074	0.222	0.298
6	SVM_17	OvR	0.1	sigmoid	-	scale	155.4850	0.069	0.222	0.288	0.061	0.222	0.255	0.074	0.222	0.298
7	SVM_17	OvO	0.01	poly	2	scale	149.1872	0.129	0.216	0.224	0.129	0.222	0.213	0.129	0.222	0.213
8	SVM_17	OvO	0.01	poly	3	scale	138.5773	0.129	0.216	0.224	0.129	0.222	0.213	0.129	0.222	0.213
9	SVM_17	OvR	0.01	poly	2	scale	149.3683	0.129	0.216	0.224	0.129	0.222	0.213	0.129	0.222	0.213
0	SVM_18	OvR	0.01	poly	3	scale	140.1225	0.129	0.216	0.224	0.129	0.222	0.213	0.129	0.222	0.213
1	SVM_18	OvO	0.01	poly	1	scale	154.0070	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149

2	SVM_18	OvO	0.01	rbf	-	scale	163.4359	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
3	SVM_18	OvO	0.01	rbf	-	auto	180.1035	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
4	SVM_18	OvO	0.01	sigmoid	-	scale	155.7240	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
5	SVM_18	OvO	0.01	sigmoid	-	auto	151.5917	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
6	SVM_18	OvR	0.01	poly	1	scale	155.6617	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
7	SVM_18	OvR	0.01	rbf	-	scale	163.1617	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
8	SVM_18	OvR	0.01	rbf	-	auto	176.2421	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
9	SVM_18	OvR	0.01	sigmoid	-	scale	159.7846	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
0	SVM_19	OvR	0.01	sigmoid	-	auto	150.0988	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
1	SVM_19	OvO	0.1	rbf	-	auto	175.2291	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
2	SVM_19	OvO	0.1	sigmoid	-	auto	150.8025	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
3	SVM_19	OvR	0.1	rbf	-	auto	175.3881	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
4	SVM_19	OvR	0.1	sigmoid	-	auto	151.5459	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
5	SVM_19	OvO	0.5	rbf	-	auto	81.0795	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
6	SVM_19	OvO	0.5	sigmoid	-	auto	42.2812	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
7	SVM_19	OvR	0.5	rbf	-	auto	80.6303	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
8	SVM_19	OvR	0.5	sigmoid	-	auto	41.7899	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149



SVM_19 9	OvO	1	sigmoid	-	auto	42.1566	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
SVM_20 0	OvR	1	sigmoid	-	auto	42.9797	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
SVM_20 1	OvO	2	sigmoid	-	auto	43.2634	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
SVM_20 2	OvR	2	sigmoid	-	auto	42.5592	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
SVM_20 3	OvO	5	sigmoid	-	auto	43.1205	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149
SVM_20 4	OvR	5	sigmoid	-	auto	43.3328	0.016	0.111	0.146	0.017	0.111	0.149	0.017	0.111	0.149

---

Table D2: Performance values for all DT models (PR = Precision, RC = Recall, ACC = Accuracy)

Model ID	Hyperparameters				Training time (s)	Training group			Validation group			Test group		
	Crit	max_depth	max_leaf_nodes	min_samples_leaf		PR	RC	ACC	PR	RC	ACC	PR	RC	ACC
DT_1	entropy	None	7	10	198.6981	0.989	0.987	0.986	0.898	0.848	0.872	0.940	0.943	0.936
DT_2	entropy	7	1	None	224.6720	1.000	1.000	1.000	0.883	0.840	0.851	0.914	0.925	0.915
DT_3	entropy	None	1	20	222.4325	1.000	1.000	1.000	0.873	0.844	0.851	0.921	0.906	0.915
DT_4	entropy	11	10	10	188.8843	0.989	0.987	0.986	0.864	0.848	0.851	0.921	0.906	0.915
DT_5	entropy	None	5	10	206.4564	0.989	0.987	0.986	0.885	0.844	0.851	0.912	0.925	0.915
DT_6	entropy	5	1	20	228.2701	1.000	1.000	1.000	0.813	0.728	0.745	0.915	0.925	0.915
DT_7	entropy	5	5	10	209.3055	0.989	0.987	0.986	0.824	0.790	0.787	0.900	0.906	0.915
DT_8	entropy	9	7	None	199.6455	0.989	0.987	0.986	0.808	0.757	0.766	0.924	0.928	0.915
DT_9	entropy	7	1	20	225.0789	1.000	1.000	1.000	0.939	0.880	0.915	0.899	0.890	0.894
DT_10	gini	11	1	None	193.8204	1.000	1.000	1.000	0.887	0.880	0.894	0.909	0.905	0.894
DT_11	entropy	5	7	20	201.3851	0.989	0.987	0.986	0.928	0.894	0.915	0.899	0.890	0.894
DT_12	entropy	5	10	10	190.0051	0.989	0.987	0.986	0.911	0.872	0.894	0.887	0.890	0.894
DT_13	entropy	7	1	10	223.8978	0.993	0.990	0.991	0.896	0.854	0.872	0.884	0.888	0.894
DT_14	entropy	7	5	10	207.9183	0.989	0.987	0.986	0.914	0.885	0.894	0.895	0.906	0.894
DT_15	entropy	11	5	20	206.9108	0.989	0.987	0.986	0.912	0.857	0.894	0.899	0.890	0.894
DT_16	entropy	9	5	None	208.1672	0.989	0.987	0.986	0.888	0.858	0.872	0.899	0.893	0.894
DT_17	entropy	None	1	10	222.4532	0.993	0.990	0.991	0.869	0.848	0.851	0.880	0.888	0.894
DT_18	entropy	None	7	None	198.4000	0.989	0.987	0.986	0.900	0.863	0.872	0.887	0.906	0.894
DT_19	entropy	5	10	20	191.8055	0.989	0.987	0.986	0.873	0.844	0.851	0.887	0.906	0.894
DT_20	entropy	None	10	None	187.5435	0.989	0.987	0.986	0.878	0.844	0.851	0.878	0.888	0.894

DT_21	entropy	7	7	None	201.2883	0.989	0.987	0.986	0.840	0.813	0.830	0.899	0.872	0.894
DT_22	entropy	11	7	None	198.7149	0.989	0.987	0.986	0.875	0.803	0.830	0.909	0.906	0.894
DT_23	entropy	9	5	20	209.3711	0.989	0.987	0.986	0.863	0.761	0.809	0.899	0.890	0.894
DT_24	entropy	11	1	20	224.6336	1.000	1.000	1.000	0.886	0.844	0.872	0.865	0.872	0.872
DT_25	entropy	5	7	10	201.0931	0.989	0.987	0.986	0.912	0.861	0.894	0.874	0.872	0.872
DT_26	entropy	None	10	10	188.0587	0.989	0.987	0.986	0.900	0.863	0.872	0.856	0.869	0.872
DT_27	entropy	None	10	20	188.4272	0.989	0.987	0.986	0.875	0.829	0.851	0.881	0.884	0.872
DT_28	gini	11	10	10	180.7146	0.972	0.973	0.973	0.887	0.880	0.894	0.870	0.868	0.872
DT_29	gini	11	10	20	181.1326	0.972	0.973	0.973	0.887	0.880	0.894	0.870	0.868	0.872
DT_30	gini	None	5	20	188.3494	0.972	0.973	0.973	0.887	0.880	0.894	0.870	0.868	0.872
DT_31	gini	None	10	10	181.5692	0.972	0.973	0.973	0.897	0.880	0.894	0.870	0.868	0.872
DT_32	gini	None	10	20	180.2820	0.972	0.973	0.973	0.887	0.880	0.894	0.870	0.868	0.872
DT_33	entropy	9	10	10	191.0532	0.989	0.987	0.986	0.857	0.822	0.830	0.850	0.869	0.872
DT_34	gini	None	7	20	184.3694	0.972	0.973	0.973	0.875	0.866	0.872	0.870	0.868	0.872
DT_35	gini	9	10	10	181.1932	0.972	0.973	0.973	0.861	0.829	0.851	0.870	0.868	0.872
DT_36	gini	None	7	None	185.7719	0.972	0.973	0.973	0.855	0.829	0.851	0.870	0.868	0.872
DT_37	entropy	9	1	10	225.9190	0.993	0.990	0.991	0.935	0.917	0.936	0.846	0.853	0.851
DT_38	gini	9	1	None	193.5661	0.997	0.996	0.995	0.887	0.880	0.894	0.826	0.831	0.851
DT_39	gini	None	1	None	194.0073	1.000	1.000	1.000	0.856	0.848	0.851	0.853	0.852	0.851
DT_40	entropy	None	1	None	222.0445	1.000	1.000	1.000	0.870	0.835	0.851	0.837	0.835	0.851
DT_41	entropy	7	10	20	190.5960	0.989	0.987	0.986	0.911	0.872	0.894	0.837	0.835	0.851
DT_42	entropy	7	10	10	190.9484	0.989	0.987	0.986	0.877	0.862	0.872	0.869	0.838	0.851
DT_43	gini	9	1	10	195.5095	0.980	0.977	0.977	0.891	0.880	0.894	0.857	0.846	0.851
DT_44	gini	11	1	10	193.6495	0.980	0.977	0.977	0.895	0.880	0.894	0.859	0.852	0.851
DT_45	gini	9	7	20	184.7079	0.972	0.973	0.973	0.891	0.880	0.894	0.853	0.852	0.851

DT_46	gini	11	10	None	181.4457	0.972	0.973	0.973	0.887	0.880	0.894	0.853	0.852	0.851
DT_47	gini	None	10	None	180.1495	0.972	0.973	0.973	0.891	0.880	0.894	0.853	0.852	0.851
DT_48	gini	9	7	None	184.3993	0.972	0.973	0.973	0.869	0.861	0.872	0.852	0.846	0.851
DT_49	gini	11	7	None	184.6349	0.972	0.973	0.973	0.875	0.866	0.872	0.859	0.852	0.851
DT_50	gini	None	5	10	186.8215	0.972	0.973	0.973	0.873	0.864	0.872	0.852	0.846	0.851
DT_51	gini	None	5	None	186.6592	0.972	0.973	0.973	0.873	0.864	0.872	0.853	0.852	0.851
DT_52	entropy	7	7	20	203.1044	0.989	0.987	0.986	0.836	0.785	0.809	0.865	0.859	0.851
DT_53	entropy	11	10	20	189.5346	0.989	0.987	0.986	0.833	0.782	0.787	0.850	0.872	0.851
DT_54	gini	11	5	None	187.0591	0.972	0.973	0.973	0.831	0.799	0.830	0.853	0.852	0.851
DT_55	gini	None	1	20	193.8636	1.000	1.000	1.000	0.944	0.922	0.936	0.854	0.830	0.830
DT_56	entropy	9	1	20	227.2644	1.000	1.000	1.000	0.869	0.839	0.851	0.829	0.819	0.830
DT_57	entropy	11	1	None	223.5777	1.000	1.000	1.000	0.863	0.844	0.851	0.829	0.835	0.830
DT_58	entropy	9	1	None	226.3461	1.000	1.000	1.000	0.860	0.784	0.830	0.870	0.859	0.830
DT_59	entropy	11	1	10	224.5103	0.993	0.990	0.991	0.888	0.862	0.872	0.840	0.835	0.830
DT_60	entropy	5	5	None	208.4174	0.989	0.987	0.986	0.910	0.850	0.872	0.830	0.835	0.830
DT_61	entropy	7	5	20	208.5141	0.989	0.987	0.986	0.877	0.821	0.851	0.850	0.838	0.830
DT_62	entropy	9	7	20	199.7175	0.989	0.987	0.986	0.865	0.839	0.851	0.839	0.813	0.830
DT_63	gini	9	5	20	187.0887	0.972	0.973	0.973	0.891	0.880	0.894	0.826	0.831	0.830
DT_64	gini	11	5	20	186.9737	0.972	0.973	0.973	0.891	0.880	0.894	0.826	0.831	0.830
DT_65	gini	9	5	None	189.6498	0.972	0.973	0.973	0.869	0.857	0.872	0.826	0.831	0.830
DT_66	gini	None	1	10	194.4044	0.980	0.977	0.977	0.856	0.847	0.851	0.847	0.824	0.830
DT_67	entropy	9	10	None	189.9840	0.989	0.987	0.986	0.844	0.790	0.809	0.862	0.856	0.830
DT_68	entropy	None	5	None	206.1365	0.989	0.987	0.986	0.832	0.816	0.809	0.851	0.856	0.830
DT_69	entropy	5	7	None	200.5122	0.989	0.987	0.986	0.829	0.786	0.787	0.828	0.835	0.830
DT_70	entropy	9	10	20	191.1951	0.989	0.987	0.986	0.821	0.786	0.787	0.871	0.851	0.830

DT_71	entropy	11	7	20	199.7647	0.989	0.987	0.986	0.821	0.786	0.787	0.835	0.832	0.830
DT_72	gini	9	1	20	193.1450	0.997	0.996	0.995	0.857	0.840	0.851	0.804	0.809	0.809
DT_73	entropy	5	1	None	227.1413	1.000	1.000	1.000	0.853	0.820	0.830	0.835	0.822	0.809
DT_74	entropy	7	10	None	190.5470	0.989	0.987	0.986	0.907	0.881	0.894	0.827	0.822	0.809
DT_75	entropy	11	5	10	207.0994	0.989	0.987	0.986	0.904	0.863	0.872	0.834	0.832	0.809
DT_76	gini	9	7	10	184.7154	0.972	0.973	0.973	0.900	0.880	0.894	0.808	0.815	0.809
DT_77	gini	11	7	10	186.1690	0.972	0.973	0.973	0.861	0.840	0.851	0.804	0.794	0.809
DT_78	gini	9	5	10	189.7158	0.972	0.973	0.973	0.833	0.822	0.830	0.808	0.815	0.809
DT_79	entropy	9	5	10	209.4935	0.989	0.987	0.986	0.807	0.764	0.766	0.813	0.838	0.809
DT_80	gini	11	7	20	184.4725	0.972	0.973	0.973	0.848	0.811	0.809	0.796	0.809	0.809
DT_81	gini	7	10	10	168.8146	0.814	0.862	0.890	0.751	0.806	0.851	0.704	0.757	0.809
DT_82	gini	7	7	10	171.5456	0.814	0.862	0.890	0.730	0.783	0.830	0.704	0.757	0.809
DT_83	gini	7	7	20	171.5379	0.814	0.862	0.890	0.727	0.769	0.830	0.704	0.757	0.809
DT_84	gini	7	7	None	171.6062	0.814	0.862	0.890	0.741	0.790	0.830	0.704	0.757	0.809
DT_85	gini	7	10	None	169.2891	0.814	0.862	0.890	0.721	0.773	0.809	0.704	0.757	0.809
DT_86	entropy	5	1	10	225.8063	0.993	0.990	0.991	0.914	0.898	0.915	0.794	0.779	0.787
DT_87	entropy	5	5	20	208.9354	0.989	0.987	0.986	0.888	0.858	0.872	0.806	0.801	0.787
DT_88	entropy	5	10	None	189.6989	0.989	0.987	0.986	0.872	0.825	0.851	0.805	0.803	0.787
DT_89	entropy	7	5	None	207.7634	0.989	0.987	0.986	0.834	0.811	0.809	0.790	0.785	0.787
DT_90	entropy	11	5	None	208.7633	0.989	0.987	0.986	0.824	0.804	0.809	0.787	0.785	0.787
DT_91	entropy	None	5	20	206.5249	0.989	0.987	0.986	0.831	0.799	0.809	0.794	0.782	0.787
DT_92	gini	9	10	None	180.6487	0.972	0.973	0.973	0.868	0.835	0.851	0.774	0.772	0.787
DT_93	gini	11	5	10	187.4583	0.972	0.973	0.973	0.868	0.835	0.851	0.789	0.776	0.787
DT_94	gini	None	7	10	184.0691	0.972	0.973	0.973	0.868	0.838	0.851	0.786	0.778	0.787
DT_95	gini	7	1	20	176.0546	0.827	0.876	0.904	0.751	0.806	0.851	0.698	0.760	0.787

DT_96	entropy	7	7	10	202.9389	0.989	0.987	0.986	0.842	0.826	0.830	0.778	0.785	0.766
DT_97	entropy	9	7	10	200.4623	0.989	0.987	0.986	0.842	0.818	0.830	0.772	0.782	0.766
DT_98	entropy	11	7	10	201.7677	0.989	0.987	0.986	0.848	0.818	0.830	0.770	0.761	0.766
DT_99	entropy	11	10	None	188.2425	0.989	0.987	0.986	0.861	0.821	0.830	0.771	0.761	0.766
DT_100	gini	9	10	20	180.8291	0.972	0.973	0.973	0.883	0.857	0.872	0.784	0.781	0.766
DT_101	entropy	None	7	20	199.6641	0.989	0.987	0.986	0.832	0.808	0.809	0.763	0.782	0.766
DT_102	gini	7	1	None	175.9758	0.827	0.876	0.904	0.744	0.769	0.830	0.691	0.725	0.766
DT_103	gini	7	5	10	171.5165	0.814	0.862	0.890	0.773	0.780	0.830	0.674	0.720	0.766
DT_104	gini	7	5	20	171.0032	0.814	0.862	0.890	0.729	0.783	0.830	0.693	0.720	0.766
DT_105	gini	7	5	None	173.3215	0.814	0.862	0.890	0.765	0.780	0.830	0.674	0.720	0.766
DT_106	gini	11	1	20	194.8051	1.000	1.000	1.000	0.773	0.751	0.766	0.751	0.761	0.745
DT_107	gini	7	10	20	168.5786	0.814	0.862	0.890	0.765	0.780	0.830	0.662	0.704	0.745
DT_108	gini	7	1	10	175.7481	0.821	0.870	0.900	0.694	0.661	0.745	0.637	0.688	0.723
DT_109	gini	5	1	20	146.8047	0.696	0.665	0.721	0.491	0.593	0.574	0.618	0.639	0.638
DT_110	gini	5	1	10	147.4505	0.696	0.665	0.721	0.525	0.620	0.596	0.489	0.598	0.596
DT_111	gini	5	5	None	143.2902	0.573	0.651	0.708	0.510	0.620	0.596	0.464	0.579	0.596
DT_112	gini	5	7	20	142.3548	0.573	0.651	0.708	0.510	0.620	0.596	0.444	0.579	0.596
DT_113	gini	5	7	None	141.8879	0.573	0.651	0.708	0.510	0.620	0.596	0.444	0.579	0.596
DT_114	gini	5	10	10	139.6752	0.573	0.651	0.708	0.510	0.620	0.596	0.461	0.579	0.596
DT_115	gini	5	10	20	140.1616	0.573	0.651	0.708	0.510	0.620	0.596	0.461	0.579	0.596
DT_116	gini	5	10	None	139.9196	0.573	0.651	0.708	0.525	0.620	0.596	0.461	0.579	0.596
DT_117	gini	5	5	10	145.0393	0.573	0.651	0.708	0.515	0.583	0.574	0.461	0.579	0.596
DT_118	gini	5	5	20	143.8020	0.573	0.651	0.708	0.483	0.598	0.574	0.444	0.579	0.596
DT_119	gini	5	7	10	142.1834	0.573	0.651	0.708	0.488	0.583	0.574	0.461	0.579	0.596
DT_120	gini	5	1	None	146.8746	0.696	0.665	0.721	0.506	0.552	0.532	0.443	0.563	0.574

## E

### Graphical Abstracts from the articles

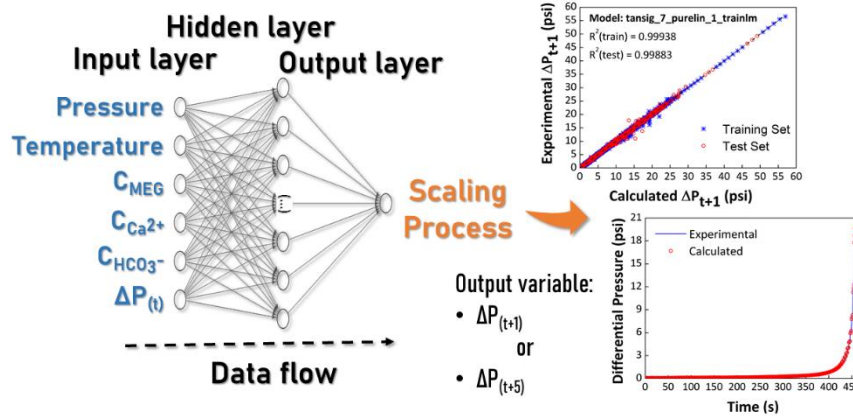


Figure E.1: Graphical Abstract of the article: Development of artificial neural network models for the simulation of  $\text{CaCO}_3$  scale formation process in the presence of monoethylene glycol (MEG) in a dynamic tube blocking test (TBT) equipment

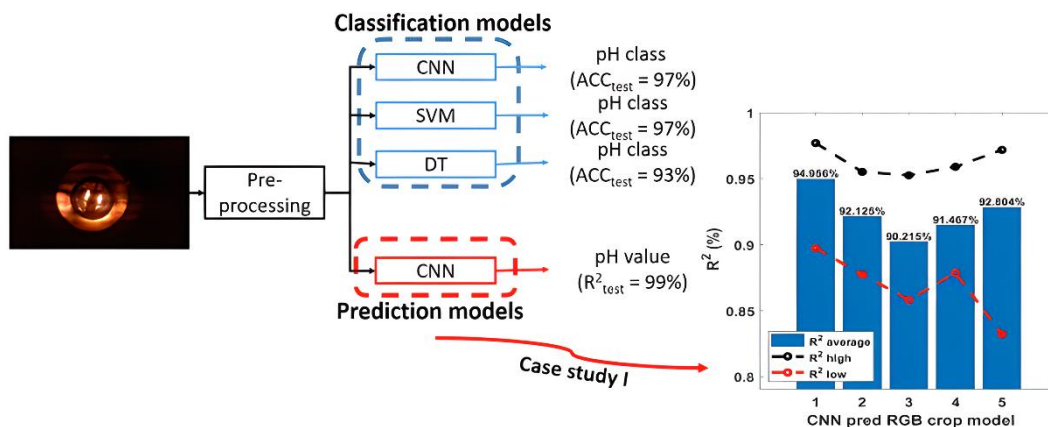


Figure E.2: Graphical Abstract of the article: Machine learning models for measurement of pH using a low-cost image analysis strategy

## **F**

### **Databases and codes**

The databases and the codes developed in this study are available at:  
<https://github.com/FerreiraBX95/Master-Thesis---Bruno-Xavier-Ferreira>.