Pontifícia Universidade Católica
do Rio de Janeiro

Matheus Hoffmann Brito

Predicting dry gas seals reliability with machine learning techniques developed from scarce data

Dissertação de Mestrado

Dissertation presented to the Programa de Pós–graduação em Engenharia Mecânica of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia Mecânica.

Advisor : Prof. D.Sc. Helon Vicente Hultmann Ayala
Co-advisor: D.Sc. Bruno de Barros Mendes Kassar

Rio de Janeiro
August 2022

Pontifícia Universidade Católica
do Rio de Janeiro

Matheus Hoffmann Brito

Predicting dry gas seals reliability with machine learning techniques developed from scarce data

Dissertation presented to the Programa de Pós–graduação em Engenharia Mecânica of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia Mecânica. Approved by the Examination Committee:

Prof. D.Sc. Helon Vicente Hultmann Ayala
Advisor
Departamento de Engenharia Mecânica – PUC-Rio

D.Sc. Bruno de Barros Mendes Kassar
Co-advisor
Instituto Tecgraf de Desenvolvimento de Software
Técnico-Cientifico da PUC-Rio – Tecgraf/PUC-Rio

Prof. D.Sc. Florian Alain Yannick Pradelle
Departamento de Engenharia Mecânica – PUC-Rio

Prof. D.Sc. Brunno Ferreira dos Santos
Departamento de Engenharia Química e de Materiais – PUC-Rio

Rio de Janeiro, August 19th, 2022

Matheus Hoffmann Brito

The author graduated in 2018 with a major in Mechanical Engineering from the Pontifical Catholic University of Rio de Janeiro. While undertaking his graduate studies, he worked as a researcher at Tecgraf/PUC-Rio until October 2021 and now works as a machine learning engineer at SONDA.

Bibliographic data

Brito, Matheus Hoffmann

Predicting dry gas seals reliability with machine learning techniques developed from scarce data / Matheus Hoffmann Brito; advisor: Helon Vicente Hultmann Ayala; co-advisor: Bruno de Barros Mendes Kassar. – 2022.

102 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Mecânica, 2022.

Inclui bibliografia

1. Engenharia Mecânica – Teses. 2. Selo seco a gás;. 3. Aprendizado de máquina;. 4. Dinâmica dos fluidos computacional;. 5. OpenFOAM;. 6. REFPROP.. I. Ayala, Helon Vicente Hultmann. II. Kassar, Bruno de Barros Mendes. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Mecânica. IV. Título.

CDD: 621

To my dog, Baja.

Acknowledgments

First, I would like to thank my parents, Alexandra and Zilton, and my grandmother, Lina, for all their support and encouragement throughout my life. Especially to my sister, Gabrielle, for having become a friend and companion at all times.

My sincere thanks to my advisor, Helon Ayala, for all the effort to extract my best and for all the advice. To my co-advisor, Bruno Kassar, I have no words to express my gratitude for the years of work at Tecgraf and for his trust in my work. To the teachers I had during my academic life, they were all crucial to my academic education. I would like to give a special thanks to professors Florian Pradelle, Igor Braga, and Ivan Menezes for all the encouragement and support since my graduation. You provided me with crucial tools for the development of this project.

To my colleagues in the Naval Engineering group at Tecgraf, I will be eternally grateful for all their patience and help in my professional growth, especially to Hilton Betta, for having developed a large part of this project with me. A big thank you to Petrobras engineers Renner Marques and Raphael Timbo, for all the support and insights for the project elaboration.

To my childhood friends, Mari, Tati, Vic, Ale, Boyer, Nicole, Fernando, and Vitor, my eternal thanks for being with me in all the important moments. To my teammates who became part of my family: Vivian, Nico, Pohl, Lu, Richard, Han, Henrique, Tristão, Zé, Biel, Caipira, Dedão, and Shrek, thank you for supporting me at all times. To my friends Baiano and Mathias, there are no words to express the gratitude and happiness I feel for having you as friends. You were my safe haven in times of sadness and my supporters in times of disbelief. To Bia, my most sincere thanks for all the support and motivation, I will be eternally grateful.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. This study was also financed in part by the Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) – ANP number 20932-0.

Abstract

Brito, Matheus Hoffmann; Ayala, Helon Vicente Hultmann (Advisor); Kassar, Bruno de Barros Mendes (Co-Advisor). **Predicting dry gas seals reliability with machine learning techniques developed from scarce data**. Rio de Janeiro, 2022. 102p. Dissertação de mestrado – Departamento de Engenharia Mecânica, Pontifícia Universidade Católica do Rio de Janeiro.

The correct equipment operation in the Oil and Gas industry is essential to reduce environmental, human, and financial losses. In this scenario, dry gas seals (DGS) of centrifugal compressors were studied, as they are identified as the most critical device due to the extent of the potential damage caused by their failure. In this study, 31 regression models available at Scikit-Learn were developed using machine learning (ML) techniques. They were trained with a scarce dataset, created based on a design of experiment technique, to replace numerical simulations in predicting the operational reliability of DGSs. First, a model based on Computational Fluid Dynamics (CFD) simulation was validated to represent the gas flowing between the sealing faces, to enable the calculation of the equipment's operational reliability. Thus, the open-source CFD software OpenFOAM was used together with the substance database of the software REFPROP, to allow the user to define the gas mixture and the evaluated operational conditions. Then, two case studies were carried out following a proposed generic workflow. The first comprised determining a regression model to estimate the reliability of a DGS whose mixture composition (composed of methane, ethane, and octane) is fixed but its operating conditions can vary. The second consisted of determining a more robust regressive model, where both the mixture composition and the operational conditions can vary. Finally, the feasibility of implementing both models under realistic operating conditions was evaluated, based on the infinity norm obtained for the prediction of the test set. The performances achieved were 1.872 °C and 6.951 °C for the first and second case studies, respectively.

Keywords

Dry gas seals; Machine learning; Computational fluid dynamics; OpenFOAM; REFPROP.

Resumo

Brito, Matheus Hoffmann; Ayala, Helon Vicente Hultmann; Kassar, Bruno de Barros Mendes. **Previsão de confiabilidade de selos secos a gás com técnicas de machine learning desenvolvido a partir de dados escassos**. Rio de Janeiro, 2022. 102p. Dissertação de Mestrado – Departamento de Engenharia Mecânica, Pontifícia Universidade Católica do Rio de Janeiro.

A correta operação de equipamentos na indústria de Óleo e Gás é fundamental para a reduzir perdas ambientais, humanas e financeiras. Neste cenário, foram estudados selos secos a gás (em inglês,DGS) de compressores cetrífugos, por serem identificados como os mais críticos devido à extensão dos danos potenciais causados em caso de falha. Neste estudo, foram desenvolvidos 31 modelos regressivos disponíveis no Scikit-Learn através de técnicas de aprendizado de máquina (em inglês, ML). Estes foram treinados com um conjunto de dados escassos, criado a partir de uma técnica de planejamento de experimentos, para substituir simulações numéricas na previsão de confiabilidade operacional de DGSs. Primeiramente, foi validado um modelo baseado na simulação da Dinâmica dos Fluidos Computacionais (em inglês, CFD) para representar o escoamento do gás entre as faces de selagem, a fim de possibilitar o cálculo da confiabilidade operacional do equipamento. Neste, foi utilizado o *software* de CFD de código aberto OpenFOAM em conjunto com o banco de dados de substâncias do *software* REFPROP, a fim de possibilitar ao usuário definir a mistura gasosa e as condições operacionais avaliadas. Em seguida, foram realizados dois estudos de caso seguindo um fluxograma genérico de projeto proposto. O primeiro consistiu na determinação de um modelo regressivo para estimar a confiabilidade de um DGS cuja composição gasosa (composta por metano, etano e octano) é fixa porém suas condições operacionais podem ser alteradas. Já o segundo consistiu na determinação de um modelo regressivo mais robusto, onde tanto a composição gasosa como as condições operacionais podem ser alteradas. Por fim, foi avaliada a viabilidade de implementação de ambos os modelos em condições reais de operação, baseado na norma infinita obtida para a predição do conjunto de teste. As performances atingidar foram de 1.872 °C e 6.951 °C para o primeiro e segundo estudos de caso, respectivamente.

Palavras-chave

Selo seco a gás; Aprendizado de máquina; Dinâmica dos fluidos computacional; OpenFOAM; REFPROP.

Table of contents

List of figures

List of tables

List of Abreviations

3D	Three dimensional
ABR	AdaBoostRegressor
AI	Artificial Intelligence
BgR	BaggingRegressor
ByR	BayesianRidge
CFD	Computational Fluid Dynamics
CV	Cross-validation
DGS	Dry Gas Seal
DTR	DecisionTreeRegressor
ElN	ElasticNet
ElNCV	ElasticNetCV
ETR	ExtraTreeRegressor
ETsR	ExtraTreesRegressor
FFD	Full Factorial Design
GBR	GradientBoostingRegressor
HuR	HuberRegressor
KNR	KNeighborsRegressor
Lar	Lars
LarCV	LarsCV
Las	Lasso
LasCV	LassoCV
Lin	LinearRegression
LsLr	LassoLars
LsLrCV	LassoLarsCV
LsLrIC	LassoLarsIC
LSVR	LinearSVR
ML	Machine Learning
NSVR	NuSVR
O&G	Oil and Gas
OMP	OrthogonalMatchingPursuit
OMPCV	OrthogonalMatchingPursuitCV
PAR	PassiveAggressiveRegressor

RAN	RANSACRegressor
RCV	Reduced Centralized Variable
REFPROP	Reference Fluid Thermodynamic and Transport Properties Database
RFR	RandomForestRegressor
Rid	Ridge
RidCV	RidgeCV
SGD	SGDRegressor
SM	Safety Margin
SVR	SVR
XGB	XGBRegressor

Essentially, all models are wrong, but this one is useful.

George E. P. Box, 1987 (adapted), .

1
Introduction

The Oil and Gas (O&G) industry is one of the most profitable and dangerous, responsible for providing the most used fuel in the world, petroleum [1]. Nevertheless, there may be failures in the oil extraction and refining processes, leading to environmental, human, and financial losses [2]. In order to mitigate these risks, the use of Artificial Intelligence (AI) has become increasingly widespread in the area [3]. In the study [4], three scenarios are prospected regarding adherence to the use of AI in the O&G industry in the next ten years, both to improve operational safety issues and to optimize equipment performance. Considering the worst scenario is not using any AI technique, the realistic one presented in [4] expects a cost reduction of 10 to 15 %, but in the optimistic scenario, it would be possible to reduce up to 40 or 50 %.

In several stages of the oil extraction and refining process, it is necessary to increase the pressure of the transported fluid due to the pressure drop in the lines [5, 6]. Centrifugal compressors are widely used to this end due to their size, weight, and low energy consumption [6]. Among the components of a compressor, dry gas seals (DGS) were identified as the most critical, given the extent of damage caused after the failure of the equipment as a whole and for generally failing earlier than expected, culminating in higher costs [7].

1.1
Literature review

The DGS systems, represented in Figure 1.1, are essential components for sealing rotating equipment, preventing the process gas from going into the atmosphere [2, 8]. These emerged as an alternative to mechanical seals to provide fewer contamination rates with the process gas and enable operation under high pressures, and rotations [9, 10]. The equipment is composed of two parallel faces, one of which rotates – the rotor – as it is attached to the compressor shaft, while the other remains still - the stator face. The fluid that passes between the sealing faces goes from the outer to the inner radius of the component. The rotating face has grooves that do not extend over the entire face, thus being responsible for an abrupt flow area reduction.

As a consequence, a pressure barrier appears and reduces gas leakage to the environment while ensuring the system's lubrication [5,9–11]. Another effect is the axial displacement of the stator, which compresses a set of parallel springs behind its face until an equilibrium between the inner and outer pressures and the springs' forces is reached, generating a clearance between 3 to 5 μm [10,12].

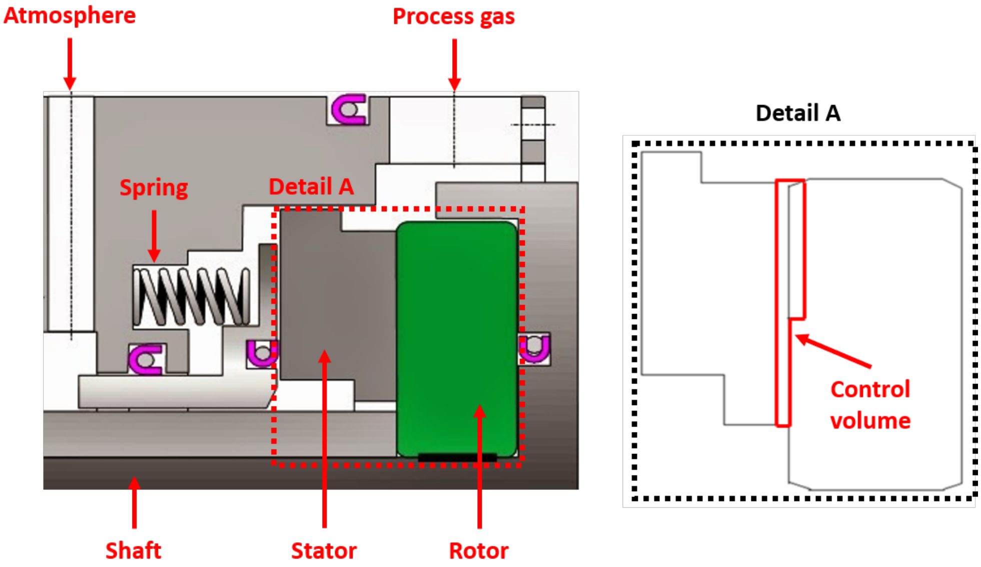


Figure 1.1: Schematic of a DGS (adapted from [5]) explaining the main components of the system (rotor, stator, shaft, and spring), the inlet (process gas) and outlet (atmosphere) of the gas in the compressor, and the control volume analyzed.

The grooves on the rotor face can have different geometries, directly influencing the fluid flow and, therefore, the system's performance [5]. The geometry types are characterized as unidirectional and bidirectional. Unidirectional geometries are only able to create the pressure barrier when the seal rotates in a particular direction, while bidirectional geometries can operate adequately when rotating in any direction [5, 10]. In the study presented in [13], it is proven that, when reversing the rotation of an unidirectional seal, the influence of the groove on the pressure distribution inside the system becomes negligible, unlike what would occur when using a bidirectional fashion.

In the study presented in [2], a set of 194 dry gas seals, unidirectional and bidirectional, were evaluated in their operating conditions until failure due to different reasons. The highest failure cause presented was supply contamination, with 43 % of the cases, and 64 % of these were due to liquid contamination. Liquid contamination consists of liquid formation between the equipment's faces, generating instability in the pressure barrier that guarantees non-contact between the faces. In order to reduce this risk, American Petroleum Institute [14] requires that the gas entering the sealing system shall be at least 20 °C above the composition's dew point. Furthermore, this temperature margin shall be satisfied inside the entire seal. Figure 1.2 shows the saturation curve of a gaseous composition and the steam curve of this gas shifted by 20 °C to the right. Thus, any point inside the DGS must be to the right of the green curve shown in the figure. In order to adequately reproduce the physical behavior between the sealing faces, the use of computational fluid dynamics (CFD) simulations is recommended [10].

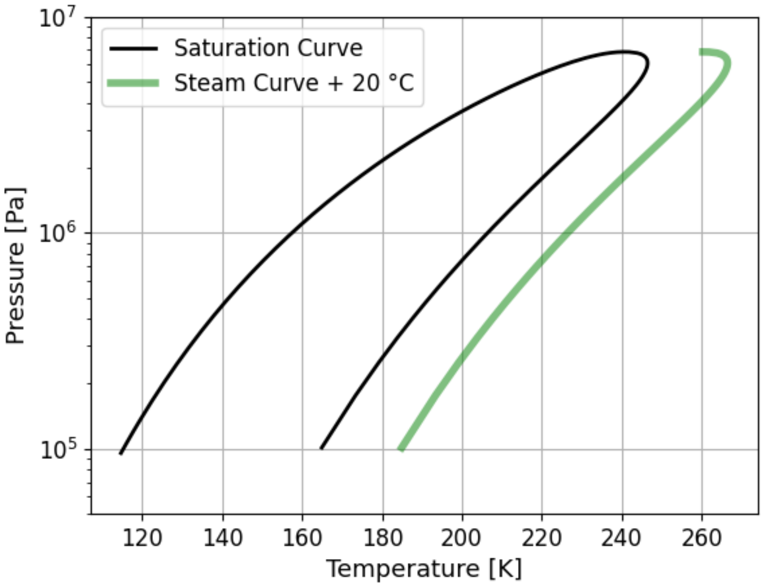


Figure 1.2: Operating limit required by the standard [14].

The fluid that passes between the sealing faces comes from the discharge or specific locations of the compressor, and its composition, mainly composed of carbon chains, may vary depending on the application of the equipment [2,15]. Before going to the DGS, the fluid passes through a device (heater) that can increase its temperature in order to guarantee the requirements set out in [14]. Figure 1.3 shows the historical record of measurements of the molar concentration of gas components, measured by chromatography, over more than two years. In this, it is possible to verify that there are variations in the concentrations, which may imply unsafe equipment operation.

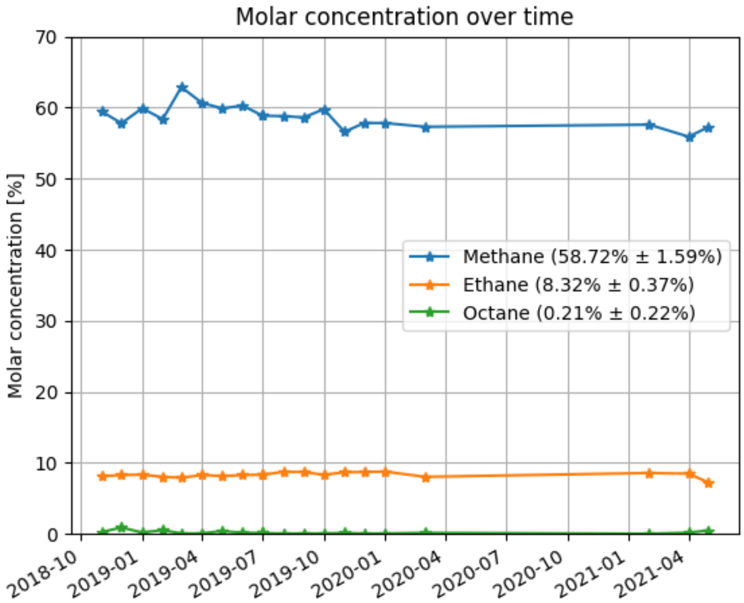


Figure 1.3: Historical composition of a standard platform. It is represented the two carbon chains with the highest molar concentration (methane and ethane) and the highest significant carbon chain (octane) observed in the chromatography.

Due to the complexity and high cost associated with carrying out experiments to obtain data regarding the behavior of DGSs following the norm [16], there are few published studies with this type of information. Table 1.1 has the authors evaluated whether their respective research had: geometric or operational optimization; development of experimental data; if, in case of numerical simulations, the data were compared with experimental data, even if not from the author himself; gas composition.

Table 1.1: Summary of contributions from studies on DGSs.

Reference	Author	Optimize parameters?	Experimental data?	Compare with experimental data?	Gas composition
[11]	Gabriel (1994)	No	Yes	No	Air
[17]	Liu et al. (2004)	Yes	No	No	-
[18]	Zheng (2005)	Yes	Yes	Yes	Air
[19]	Zhou et al. (2007)	Yes	No	No	Air
[10]	Ojile (2009)	Yes	No	Yes	Air
[20]	Jing et al. (2012)	No	No	Yes	Air
[21]	Kolomoets et al. (2012)	No	Yes	No	Air
[13]	Shahin et al. (2013)(a)	No	No	Yes	Air
[22]	Shahin et al. (2013)(b)	No	No	Yes	Air
[23]	Wang et al. (2013)	No	No	Yes	Air
[12]	Ding et al. (2016)	No	Yes	Yes	Air
[9]	Fairuz et al. (2016)	No	No	Yes	SCO_2
[24]	Ma et al. (2016)	No	No	No	N_2
[25]	Wang et al. (2017)	No	Yes	Yes	Air
[26]	Wang et al. (2018)	No	No	No	Air
[27]	Du et al. (2018)	No	No	Yes	Air and SCO_2
[15]	Kassar et al. (2020)	No	No	Yes	Hydrocarbon mixture

The study presented in [11] was an experimental research conducted to analyze the dynamic behavior of dry gas seals with spiral grooves (unidirectional). The authors evaluated the influence of different aspects of the DGS, such as the clearance between faces, the material, and possible deformations, using air as the product gas. The following information is available in the research: pressure distribution along the radius for different clearances; gas leakage from the sealing outlet; lift force on the stator face to ensure static balance. As one of the first of this field's analyses, its results are commonly used as a benchmark to validate the computational models developed.

In the research presented in [17], different layouts of DGSs with spiral grooves were evaluated and their influence on the following aspects: opening force, leakage, and energy consumption. A reduced axisymmetric domain was used to reduce the computational cost of CFD simulations, a feature used in all analyzed publications using a numerical model. A laminar flow was considered, assuming incompressible gas and a developed flow. Four different layouts were evaluated under the same operating conditions, and it was found that with increasing gas film thickness, there is a tendency to decrease the opening force and energy consumption and increase leakage. Parametric optimization of geometric factors linked to the groove design was performed using the most common layout in operations to maximize the opening force performed on the seal face, but the performance gain compared to the previous configuration is not presented.

In the article [18], the results obtained from a simplified 2D model and a complete 3D model were compared. A laminar flow was considered, assuming an incompressible gas and a developed flow. First, the two models were compared, and it was verified that the 2D representation was sufficient to reproduce the 3D model. Then, a parametric optimization was performed to define the spiral groove's best geometric characteristics that maximize the opening force. With the new geometry established, an increase of about 30 % in the performance of the equipment was achieved.

The study [19] presents the results of the parametric optimization of a spiral groove in order to maximize its efficiency, defined as the ratio between the lift force and the leakage. The airflow inside the seal was evaluated under the hypothesis of a real gas, incompressible fluid, laminar, and fully developed flow. The author considered the deformation on the seal faces due to temperature, and for that, it was used an artificial neural network (ANN) where the network input is the heat flux in the domain cells, and the output is the axial deformation. This representation proved robust to replace the Computer-Aided Engineering (CAE) model and considered more computationally efficient. Then, the parametric optimization of the seal layout was performed, obtaining an increase of 25 % in its efficiency.

The study presented in [10] is divided into two stages. In the first one, the airflow inside the seal was evaluated under the assumption of ideal gas, compressible fluid, turbulent flow following the fully developed $k - \omega$ Shear Stress Transport (SST) model. Different groove geometries were evaluated, and lift force and leakage were compared for different operating conditions. This step concluded that the spiral groove is the most efficient. The second contribution was creating an optimization model based on genetic algorithms to optimize the different variables related to sealing geometry to maximize its efficiency. In this one, the Halton [28] design of experiments (DoE) technique was used to create the dataset to train the model. As a result, it was possible to obtain a geometry 20 % more efficient, proving that the procedure followed by the author was able to generate a robust model.

In [21], experimental tests were carried out in a DGS with spiral grooves to evaluate the reliability in aeronautical use. Despite being a different application from the proposal in this research, it is possible to obtain important information about the types of sensors used in operation. Based on this information, it is possible to evaluate the measurement errors for each variable of interest according to the used sensor and stipulate threshold errors for predictive models.

In the article [20], the opening force and leakage obtained by [11] for different clearances assuming laminar and $k - \epsilon$ Re-Normalisation Group (RNG) turbulence model were compared. The airflow inside the seal was evaluated under the hypothesis of an ideal gas, compressible fluid, and fully developed flow. The author concluded that as the clearance increased, the influence of the groove was attenuated, reducing the lift force and increasing leakage, as expected. In addition, it also concluded that when using the turbulence model in the flow calculation for small clearances, there was an overestimation of the pressure field and consequently of the force on the stator, and a decrease in leakage when compared with laminar flow. However, despite being different, both results remained close to those obtained by [11], with errors lower than 5 %. Thus, the conclusion was that it is recommended to use a turbulence model when the clearance is higher than 3 μm, and to assume a laminar flow for smaller gas films.

In the research presented in [22], the behavior of pressure and temperature fields was evaluated under different operating conditions assuming different flow regimes: laminar, $k - \epsilon$ RNG and the Large Eddy Simulation (LES) turbulence models. The airflow inside the seal was evaluated under the hypothesis of an ideal gas, incompressible fluid, and fully developed flow. The result obtained was that, for the same mesh refinement, the laminar flow hypothesis is more adherent to the experimental results of [11]. Furthermore, with a decrease in the clearance between the faces, there was an increase in the pressure and temperature gradients.

In the study presented in [23], the hypothesis of turbulent flow is evaluated, using the $k-\omega$ SST model, when comparing with the experimental data presented in [11]. The airflow inside the seal was evaluated under the hypothesis of an ideal gas, incompressible fluid, and fully developed flow. First, it was found that the Reynolds number remained low even at high speeds due to the simultaneous increase in viscosity. In addition, it was found that for higher clearances, the use of turbulence models was recommended, corroborating [20].

In [12], the temperature was measured at three different points of a DGS with spiral grooves. In this study, the author aims to validate the temperature values obtained through a mathematical model with those obtained experimentally. Airflow inside the seal was evaluated under the hypothesis of ideal gas, compressible fluid, laminar, and fully developed flow. This model estimated the temperature with errors less than 0.5 °C for all three points, proving its reliability. It is important to note that the type of sensor used in this study was the same used in [21], a chromel-alumel thermocouple (type K). Thus, this study becomes an essential reference to help define temperature boundary conditions.

In the research presented in [9], the behavior of a DGS was evaluated for different temperatures under the hypothesis of ideal and real gas. The authors used the REFPROP software database, released by the NIST (National Institute of Standards and Technology) [29, 30], to calculate the real fluid's thermodynamic properties. The supercritical carbon dioxide (SCO_2) flow inside the seal was evaluated under the hypothesis of compressible fluid, $k-\omega$ SST turbulence model, and fully developed flow. The authors concluded that when comparing the pressure and temperature fields under ideal gas and real gas hypotheses for two different inlet temperatures (370 and 740 K), it can be seen that for the lowest temperature, there is a divergence between the results, which does not occur at the highest temperature. The distribution of the specific mass values under the same conditions has a disparity only when the inlet temperature is 370 K. This is due to the combination of inlet pressure and temperature in the seal being close to the phase change curve of the gas, where the ideal gas hypothesis is not valid. In conclusion, the author recommends using the real gas hypothesis to obtain more reliable results.

In the article [24], the effects of the flow of N_2 (dinitrogen) in a DGS with

T grooves (bidirectional) under different inlet pressures and rotations were evaluated. The flow inside the seal was investigated under the hypothesis of real gas, compressible fluid, laminar, and fully developed flow. An important conclusion obtained in this article was that the temperature field becomes more sensitive to operational changes when subjected to pressures below 1 MPa. Furthermore, it is possible to see from the study results that there is an abrupt drop in temperature in the region of the seal's inner radius, as observed in [13, 22]. This effect is expected due to the pressure drop to near atmospheric and, consequently, the gas temperature is also decrease due to the Joule-Thompson effect [5].

In the study presented in [25], the results obtained through experiments and computer simulations of a DGS with T grooves operating at low speeds were compared. The airflow inside the seal was evaluated under the hypothesis of ideal gas, compressible fluid, laminar, and fully developed flow. The investigated parameters were the opening force and the gas leak in the seal. For such, the proposed model has good adherence to the results. The author warns that there may be contact between the faces depending on the operating speed due to low lift force. To ensure the system's reliability, ranges of values are given for the number of grooves (10 to 16), groove depth (4 to 6 μm), and gas film thickness (2 to 4 μm). However, it is worth to mention that these recommendations are applicable in operations with rotation lower than 6000 RPM, the evaluated limit value, which may vary when subjected to higher rotations.

In [26], the orientation and magnitude of roughness on the upper wall of the T groove were evaluated. Airflow inside the seal was assessed under the hypothesis of an ideal gas, incompressible fluid, laminar, and fully developed flow. The opening force and leakage were evaluated for different clearances in grooves with and without roughness. The author concluded that roughness in the correct orientation could increase the opening force and decrease leakage under the same operating conditions. It is also evidenced that, as the clearance increases, this effect is attenuated, corroborating the conclusions presented in [11, 20].

The research presented in [27] compares the influence of gas film thickness, inlet pressure, and inlet temperature on opening force, leakage, and pressure and temperature gradients for two fluids, air and SCO_2. The flow inside the seal was evaluated under the hypothesis of a real gas, compressible fluid, laminar, and fully developed flow. First, it was concluded that the SCO_2 flow is more sensitive to temperature, pressure differential, opening force, and leakage changes than the airflow. As the inlet pressure increases, it is verified that the pressure and temperature differentials decrease, and the opening force and leakage tend to increase for both air and SCO_2. Furthermore, when evaluating the influence of the clearance for the two fluids, it is noticed that there is a tendency to obtain the same pressure distribution along the radius, corroborating with [11, 20, 26].

In the article [15], the reliability of a given operational condition is evaluated, where the evaluated fluid is a mixture composed of 13 components, mainly composed of carbon chains. In this study, the thermodynamic properties are calculated using REFPROP, similar to that performed by [9], using the $k-\omega$ SST turbulence model and assuming a developed flow. The first significant contribution was the proof that heavier carbon chains (nonane and decane, for example) tend to increase the saturation temperature of the superheated steam curve, requiring increasingly higher inlet temperatures in the sealing system to satisfy the standard [14]. In addition, unlike the other studies evaluated previously, this one quantitatively presents the system's reliability. For this, the saturation curve of the evaluated mixture, all pressure and temperature pairs of the flow solution in the evaluated domain, and the smallest convex envelope (convex hull) that encompasses these points are presented in the same graph. Thus, operational reliability (T_{margin}) was defined as the shortest distance between the convex hull and the superheated steam curve. However, the use of the convex hull may overestimate the distance to the vapor curve depending on the inlet pressure in the DGS and the saturation curve of the mixture, as exemplified in Figure 1.4.

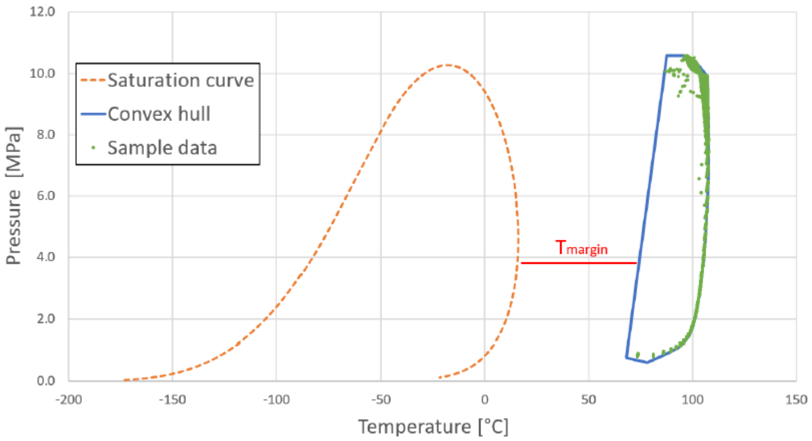


Figure 1.4: Convex hull analysis of operational reliability presented in [15].

From the studies, it is possible to notice that, with the greater availability of computational resources, the complexity of the models was increased, evaluating the hypothesis of real and compressible gas models. There is also much analysis on how the DGS' behavior is influenced by geometry and operating conditions. However, few studies aim to increase the reliability and security of the system as required by [14]. Furthermore, as they are not intended for operation, they mostly assume an airflow, which is not valid in O&G applications [2,15].

The research carried out on DGSs also highlights the scarcity of studies for creating predictive models to replace CFD simulations, which are extremely computationally expensive, to obtain reference performance parameters, such as lift force and leakage. Furthermore, the studies that were carried out for this purpose [10,17–19] aimed to improve the performance from the geometric configuration of the seal's grooves. There is also a lack of information to determine the sample conditions to be evaluated, except for [10], which uses a DoE technique to define the necessary tests. Finally, it is possible to notice that in these optimization studies, there is no consideration of variability in the composition of the gas flowing inside the system, an unlikely condition according to [2,15].

Due to the scarcity of studies, some works not linked to DGSs were evaluated to understand the feasibility of inserting such considerations in a predictive model of a sealing system. Table 1.2 has the authors evaluated and the following information regarding their respective research: source of data used to compose the dataset; how the dataset was created; whether use classic machine learning techniques or deep learning techniques were used.

Table 1.2: Summary of current contributions to the creation of predictive models.

Reference	Author	Data source	Dataset creation	Classical ML	Neural Networks
[31]	Carrillo et al. (2018)	CFD	Not specified	Yes	No
[32]	Gong et al. (2018)	Numerical	Not specified	No	Yes
[33]	Palagri et al. (2019)	Experiment	Not specified	No	Yes
[34]	Xing et al. (2019)	Experiment	Not specified	Yes	Yes
[35]	Liu et al. (2019)	Numerical	Not specified	Yes	No
[36]	Ramirez et al. (2020)	CFD	DoE	Yes	No
[37]	Elmaz et al. (2020)	Experiment	Not specified	Yes	Yes
[38]	Bakhtiari et al. (2020)	CFD	Not specified	No	Yes
[39]	Jalalifar et al. (2020)	CFD	Not specified	Yes	No
[40]	Ping et al. (2021)	Experiment	Not specified	No	Yes
[41]	Marcato et al. (2021)	CFD	DoE	No	Yes
[42]	Du et al. (2021)	Experiment	Not specified	No	Yes
[43]	Fei et al. (2021)	Experiment	Not specified	Yes	Yes
[44]	Djandja et al. (2021)	CFD	Not specified	No	Yes
[45]	Sun et al. (2021)	Experiment	Not specified	Yes	Yes
[46]	Leng et al. (2021)	Experiment	Not specified	Yes	Yes
[47]	Agbulut et al. (2021)	Experiment	Not specified	Yes	Yes
[48]	Shalaby et al. (2021)	Numerical	Not specified	No	Yes

Currently, there is a great focus on research to create representative predictive models of different types of physical phenomena, being created from experimental data [33, 34, 37, 40, 42, 43, 45–47] or from numerical simulations [31, 32, 35, 36, 38, 39, 41, 44, 48]. From the previously mentioned, the studies [38, 39] use CFD models to create datasets to be used to train the models, where the input variables are conditions that determine the operation of the evaluated equipment (*e.g.*, inlet pressure and temperature). In [32, 35], numerical data are generated to predict thermodynamic properties of gas mixtures, similar to the physical phenomenon present inside sealing systems. From the research can be verified that the predictive models whose input variables are, separately, operational conditions or mixture compositions (*e.g.*, molar fraction of methane and ethane), present satisfactory robustness and low prediction errors, obtaining correlation coefficients (R^2) of 0.99.

In [44, 46] the input variables for the predictive model are both operational conditions and mixture compositions, increasing the required complexity. The models obtained have R^2 values of 0.92 and 0.90, respectively, corroborating the hypothesis of greater complexity for the correct prediction. However, it is possible to notice that there is little information on how the experimental conditions were selected to create the datasets. In the studies [36,41], the CFD simulations to be performed are defined through DoE techniques, as presented in [10]. In these researches, it is possible to obtain greater clarity regarding the number of simulations to be carried out, the expected computational demand, and, mainly, reproducibility in the combinations of the analyzed input parameters. Furthermore, the study [49] discusses the relationship between ML and DoE, concluding that, until problems of quality and quantity of available data are resolved, such techniques can be used in a complementary way.

When evaluating the models used in the studies, it is possible to verify the use of more conventional ML techniques, such as: Polynomial [31, 36, 37], Support Vector Machine [35, 37, 39, 43, 45–47], Decision Tree [37], Elastic Net [43], Gaussian Process [43], Random Forest [34,43,46], Gradient Boosting [43,45], Extreme Gradient Boosting [45,46] and K-Nearest Neighbors [45,47]. However, in these studies, Deep Learning techniques were also evaluated, from simple ones such as Multilayer Perceptron [37, 38, 44–46], to more complex Artificial Neural Networks (ANN) [32–34, 40–43, 45, 47, 48], obtaining errors of similar magnitude to the more conventional models previously mentioned. Furthermore, it is important to realize that there is a great focus of research on the Support Vector Machines model and Deep Learning techniques, and there are not many studies with simpler models and possibly sufficient to represent the desired physical phenomena in their respective studies.

1.2 Objective and contributions

This study aims to fill the knowledge gap in predicting the reliability of dry gas seals (DGS) under realistic operational conditions, determining the safety temperature margin for a given operating setup. Furthermore, we also seek to propose a generic design workflow to create regression models from computer simulations. Firstly, a CFD model with suitable boundary conditions is developed in the open-source software OpenFOAM [50] and, thus, compare the results obtained with the reference literature [11]. The thermodynamic properties obtained from REFPROP is used, thus enabling the analysis of complex mixtures and versatility for future studies. In possession of a representative model, we use Design of Experiments (DoE) technique to define the necessary simulations to be performed to compose the dataset for the development of the model. Then, different Machine Learning models are evaluated to solve the regression problem. Finally, it is possible to define the most suitable model for the stipulated conditions.

In this work, we developed two models with different mixture composition hypothesis: one with fixed gas composition and one with variable composition. The following operating conditions are considered in both models: inlet pressure and temperature, the shaft's angular velocity, and clearance between the sealing faces. Under the hypothesis of variable mixture composition, the influence of the molar fractions of methane, ethane, and octane in a representative gas mixture is also evaluated. Thus, the contributions presented in this study are:

1. Complete generic design workflow to develop regression models based on machine learning techniques.

2. Predicting dry gas seals reliability with machine learning techniques for a fixed mixture composition.

3. Predicting dry gas seals reliability with machine learning techniques for a variable mixture composition.

1.3 Thesis outline

This study is divided into five chapters with the following topics:

- **Chapter 1**: Motivation for the study, the state-of-the-art in DGS research and predictive models developed from numerical simulations, and contributions proposed in the dissertation.

- **Chapter 2**: Theoretical conceptualization of the mathematical model, discussion of boundary conditions, and validation of the CFD model with the literature.
- **Chapter 3**: Theoretical conceptualization of the DoE and ML techniques used, creation of the dataset, and methodology for improving the regressive models.
- **Chapter 4**: Presentation of the generic design workflow proposed and used in the case studies and presentation of the results of the case studies for the two hypotheses assumed about the composition of the gas mixture.
- **Chapter 5**: Conclusions about the results obtained in the two case studies and suggestions for future work.

2
Dry gas seal

2.1
Geometry modeling

This section exposes the parametrization of a spiral groove (unidirectional) and a T groove (bidirectional) fluid domains that are simulated in this study. The equations were written in terms of easily measurable parameters and assuming circumferential periodicity according to the number of grooves in the face seal.

2.1.1
Spiral groove

The spiral groove geometry was based on the parameters proposed by [9]. The external radius (R_o), internal radius (R_i), and the radial position where the groove starts (R_g), denoted as groove radius, can be observed in Figure 2.1 and 2.2.

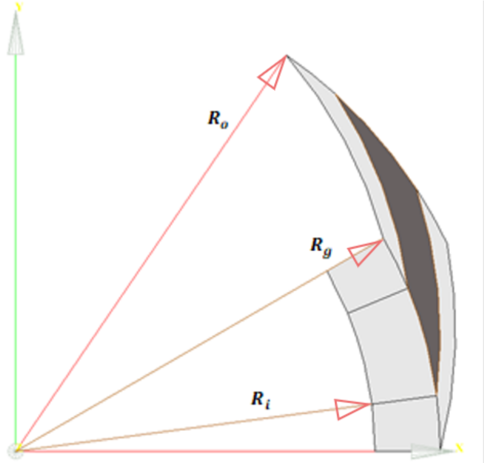


Figure 2.1: Reference radius for parametric modeling of the fluid domain of a DGS with spiral grooves.

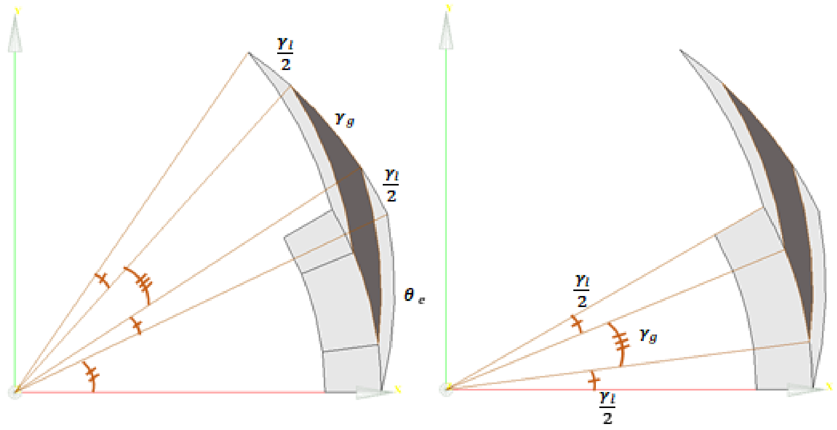


Figure 2.2: Reference angles for parametric modeling of the fluid domain of a DGS with spiral grooves.

Assuming N as the number of grooves in the face seal, the angular distance between two grooves (γ_l), the angular opening of the groove (γ_g), and the ratio between these two parameters (β), denoted as groove ratio, it is possible to write

$$\begin{cases} \beta = \frac{\gamma_g}{\gamma_l} \\ N(\gamma_g + \gamma_l) = 360 \end{cases} . \tag{2-1}$$

It is convenient to rewrite these in terms of N and β

$$\begin{cases} \gamma_l = \frac{360}{N(\beta+1)} \\ \gamma_g = \beta\gamma_l \end{cases} , \tag{2-2}$$

which are more accessible parameters to measure.

The groove shape is based on

$$r(\theta) = R_g e^{\theta tan(\alpha)}, \tag{2-3}$$

a logarithmic spiral to ensure that each point on the curve has an angle α between the moving direction and tangent of the spiral to maximize the seal performance.

In order to determine the full angle of the spiral (θ_e), it is necessary to evaluate the Equation 2-3 when r(θ) = R_o, resulting on

$$\theta_e = \frac{\ln \frac{R_o}{R_g}}{tan(\alpha)}. \tag{2-4}$$

The two parameters remaining to describe the fluid domain are the groove depth (h_g) and the clearance between the seal faces (h), that is a consequence of the dynamic behavior. They are represented in Figure 2.3.

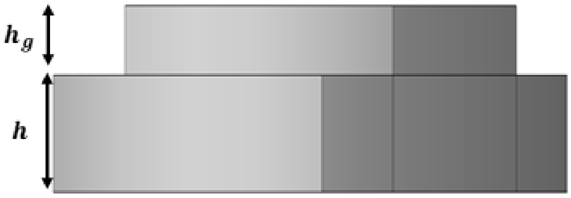


Figure 2.3: Gas film thickness between the seal faces (h) and inside the groove region (h_g).

2.1.2 T groove

The T groove geometry was based on the parameters proposed by [26] as shown in Figure 2.4 and 2.5. The parametrization is similar to the spiral groove, with some additional parameters that are shown in this section.

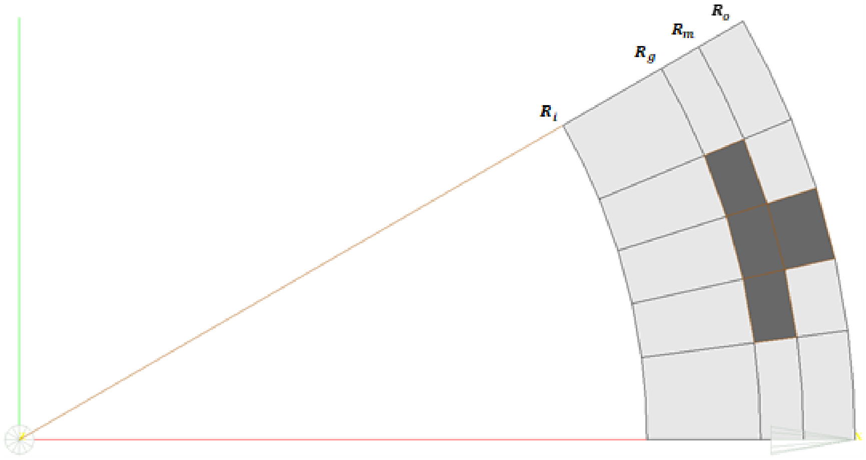


Figure 2.4: Reference radius for parametric modeling of the fluid domain of a DGS with T grooves.

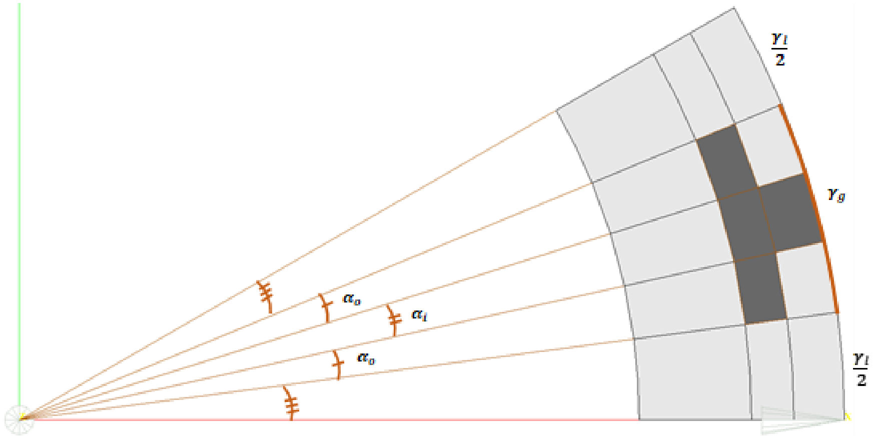


Figure 2.5: Reference angles for parametric modeling of the fluid domain of a DGS with T grooves.

The first new concept is the middle radius (R_m) that is where the groove change its cross section. Additionally, it is defined the inner angular section (α_i) and the outer angular section (α_o), and then assume that

$$\begin{cases} \beta = \beta_1 = \frac{\gamma_g}{\gamma_l} \\ \beta_2 = \frac{\alpha_i}{\alpha_o} \\ \alpha_i + 2\alpha_o = \gamma_g \\ N(\gamma_g + \gamma_l) = 360 \end{cases} \tag{2-5}$$

based on Equation 2-1.

Rewriting these in terms of N, β_1, β_2, which are more accessible parameters to measure, the relation becomes

$$\begin{cases} \gamma_l = \frac{360}{N(\beta_1+1)} \\ \gamma_g = \beta_1\gamma_l \\ \alpha_o = \frac{\gamma_g}{\beta_2+2} \\ \alpha_i = \alpha_o\beta_2 \end{cases} . \tag{2-6}$$

2.2 Mesh

A computer simulation's mesh defines how a domain is represented and the reference properties calculated. Thus, the mesh quality can directly influence the solution convergence rate and numerical errors. For this study, the resource *blockMesh* available in OpenFOAM was used, generating a hexahedral geometric mesh, seeking a mesh quality that would meet the recommended by the software (Table 2.1). It is worth mentioning that, although the software allows the execution of the simulation even with values outside the recommended range, it is essential to assess whether the results converge through a mesh convergency analysis.

Table 2.1: OpenFOAM's reference values for mesh quality.

Parameter	Minimum	Maximum
Aspect Ratio	1	1000
Non-orthogonality	0	70°
Skewness	0	4

The aspect ratio measures the proportion of deformation between a given mesh cell and an equal-sided regular cell. Due to the small clearance between the sealing faces, the proportion between the discretizations in the axial direction and the others culminates in aspect ratios above the recommended in Table 2.1. To solve this problem, increasing the number of elements in the other directions would be necessary, consequently increasing the required computational cost. Thus, it was decided to evaluate the relevance of the impact on the results through the mesh test.

The skewness and non-orthogonality parameters represent how misaligned the faces of a given cell are. Although the skewness values are within the recommended range for the two grooves types,the non-orthogonality parameter for the spiral groove is above the ideal, which may introduce numerical errors in the gradient calculations. It was expected due to the complex geometry in the groove region of the spiral seal. Gradient limiters and non-orthogonality correctors were used in the simulations to mitigate numerical errors, avoiding numerical instabilities despite compromising the convergence rate.

2.3 Mathematical modeling

This session exposes the mathematical equations solved by the used software (OpenFOAM 18.12 [50]), the boundary conditions, and the turbulence model. Furthermore, it is detailed how the thermodynamic properties are obtained with REFPROP [30] and its coupling in the numerical solver. Finally, the internal fields of velocity and pressure initializations are presented, and the adopted convergence criterion.

2.3.1 Flow governing equations

Due to the rotor movement to be represented in the model, a possible solution would be using dynamic meshes, but it would result in a substantial increase in the computational cost. The alternative was the use of a moving coordinate system methodology (Figure 2.6), thus incorporating the terms of relative velocities and centripetal and Coriolis accelerations to the conservation equations. The Multiple Rotating Frames (MRF) methodology available in the software was used in this approach. It applies the rotation terms in multiple specified zones and, unlike other available methodologies, it makes it possible to solve compressible flow simulations, a hypothesis adopted in this study. Thus, the fluid velocity in the new coordinate system can be written as

$$\bar{v} = \bar{v}_r + \bar{u}_r = \bar{v}_r + \bar{\omega} \times \bar{r}, \tag{2-7}$$

where $\bar{v}$ is the absolute velocity, $\bar{v_r}$ is the relative velocity, $\bar{u_r}$ is the moving coordinate system velocity, $\bar{\omega}$ is the angular velocity, and $\bar{r}$ is the distance to the new coordinate system's origin.

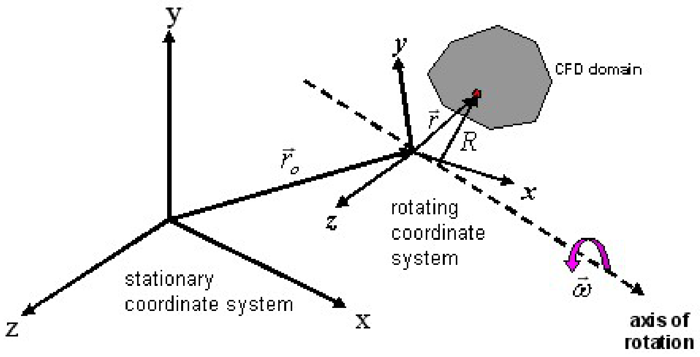


Figure 2.6: Stationary and rotating reference frames [51].

The software calculates the fluid flow using the Reynolds Averaged Navier Stokes (RANS) equations applied to a rotating frame, under the hypothesis of steady-state regime, compressible fluid, and turbulent flow. Thus, the equations that govern the flow as a function of relative velocity are [51]:

1. Conservation of mass :

$$\nabla.(\rho\bar{v_r}) = 0, \tag{2-8}$$

where ρ is the specific mass.

2. Conservation of energy:

$$\nabla.(\rho\bar{v_r}h_s + \frac{1}{2}\rho\bar{v_r}v_r^2 - \frac{1}{2}\rho\bar{v_r}u_r^2) = \nabla.(\bar{\bar{\tau}}.\bar{v}), \tag{2-9}$$

where h_s is the sensible enthalpy. The viscous stress tensor ($\bar{\bar{\tau}}$) can be written as

$$\bar{\bar{\tau}} = \mu[(\nabla\bar{v_r} + \nabla\bar{v_r}^T) - \frac{2}{3}\nabla.\bar{v_r}I], \tag{2-10}$$

where μ is the fluid dynamic viscosity and I is the unit tensor.

3. Conservation of linear momentum:

$$\nabla.(\rho\bar{v_r}\bar{v}) + \rho(2\bar{w} \times \bar{v_r} + \bar{w} \times \bar{w} \times \bar{r}) = -\nabla p + \nabla.\bar{\bar{\tau}}, \tag{2-11}$$

where p is the fluid pressure.

2.3.2 Turbulence

Turbulence is a three-dimensional and random phenomenon composed of vortices of different scales in such a way that those with higher turbulent kinetic energy transfer it to those of more minor scales until complete dissipation through viscous forces [52]. Among the different ways of evaluating turbulence, the RANS method is used due to the low computational cost compared to the others and satisfactory results. This method models the turbulence scales statistically by including terms of the average contribution of turbulence to the flow called Reynold's tensor.

There are different approaches for calculating the Reynolds tensor, but the most used ones in the studies presented in Chapter 1 are the Realizable $k - \epsilon$ and the $k - \omega$ SST models. The Realizable $k - \epsilon$ model has a lower computational cost and is indicated for high Reynolds number, overestimating the turbulence levels for lower Reynolds. The $k - \omega$ SST model combines the $Standard$ $k - \omega$ model for regions close to the wall and the $k - \epsilon$ model for regions further away from the viscous sublayer of the boundary layer. Thus, the $k - \omega$ SST model was selected due to its greater adaptability.

2.3.3 Thermodynamic and physicochemical properties

The OpenFOAM software performs the calculation of thermodynamic properties from the values of pressure (p) and temperature (T), in Pascal and Kelvin respectively. The properties of interest are:

- Dynamic viscosity (μ)
- Thermal conductivity (κ)
- Specific enthalpy (h_s)
- Specific mass (ρ)
- Specific heat at constant pressure (c_p)
- Specific heat at constant volume (c_v)
- Compressibility factor (Z)

There are different approaches to the calculation of the real thermodynamic properties of a fluid, among them: Peng-Robinson equation of state [53], JANAF tables [54], Sutherland law [55], and polynomial fits [56]. However, when it comes to gas mixtures, such correlations become more complex. Thus, such as in [9, 15], REFPROP was used to calculate the thermodynamic properties.

A general thermodynamic property Φ is described in tabular fashion (propTable) as a function of pressure and temperature (Figure 2.7). So the file consists of:

- Lowest pressure bound (p_{min}), the step between each pressure point (Δp), and the number of different pressure values (n_p).
- Lowest temperature bound (T_{min}), the step between each temperature point (ΔT), and the number of distinct temperature values (n_T).
- Table of property Φ with dimensions $n_p \times n_T$.

$$
\begin{array}{cccc}
p_{min} & \Delta\mathrm{p} & n_p & \\
T_{min} & \Delta\mathrm{T} & n_T & \\
\Phi(T_{min}, p_{min}) & \cdots & & \Phi(T_{max}, p_{min}) \\
\vdots & \ddots & & \vdots \\
\Phi(T_{min}, p_{max}) & \cdots & & \Phi(T_{max}, p_{max})
\end{array}
$$

Figure 2.7: PropTable file structure adopted to represent the fluid.

The algorithm developed uses the REFPROP database to provide the necessary information to OpenFOAM. In this algorithm, the components of the mixture and their respective molar fractions are defined, and the pressure and temperature ranges that the tables must contemplate. With this, the algorithm generates seven propTables and a file with the saturation curve of the mixture. Thus, whenever it is necessary to use some thermodynamic property in OpenFOAM, these tables are consulted through a modification in the original software code.

2.3.4 Boundary conditions

It was considered a slice of the complete fluid domain due to the periodicity of the grooves, a resource used by all the studies evaluated in Chapter 1 to reduce the computational cost. Thus, Figure 2.8 shows the reference surfaces adopted in the simulations. The side faces are defined as cyclic, the upper (rotor) and lower (stator) faces are defined as walls, and the other ones are defined as inlet or outlet according to the flow direction.

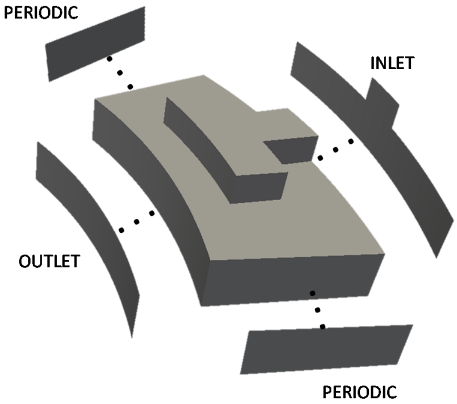


Figure 2.8: Reference faces where the boundary conditions of the fluid domain are applied.

After the research carried out in Chapter 1, it was possible to catalog in a simplified way in Table 2.2 the boundary conditions adopted in each of the studies. Based on these references, the boundary conditions of velocity, pressure, and temperature were defined for each reference face: inlet, outlet, stator, and rotor.

Table 2.2: Boundary conditions found in the literature, where: (-) means not applicable or not informed; (*defined*) the property value is prescribed; (*calculated*) the software calculates the property value to ensure consistency; the other boundary conditions are specific for each evaluated property.

	Inlet			Outlet			Stator			Rotor		
Reference	U	P	T	U	P	T	U	P	T	U	P	T
[11]	-	-	-	-	-	-	-	-	-	-	-	-
[17]	Calculated	Defined	-	Calculated	Defined	-	No Slip	Calculated	-	No Slip	Calculated	-
[18]	Calculated	Defined	Defined	Calculated	Defined	Calculated	No Slip	Calculated	Isothermal	No Slip	Calculated	Isothermal
[19]	Calculated	Defined	Defined	Calculated	Defined	Calculated	No Slip	Calculated	Heat Flux	No Slip	Calculated	Heat Flux
[10]	Calculated	Defined	Defined	Calculated	Defined	Calculated	No Slip	Calculated	Adiabatic	No Slip	Calculated	Adiabatic
[20]	Calculated	Defined	Defined	Calculated	Defined	Calculated	No Slip	Calculated	Isothermal	No Slip	Calculated	Isothermal
[21]	-	-	-	-	-	-	-	-	-	-	-	-
[13]	Calculated	Defined	Defined	Calculated	Defined	Calculated	No Slip	Calculated	Adiabatic	No Slip	Calculated	Adiabatic
[22]	Calculated	Defined	Defined	Calculated	Defined	Calculated	No Slip	Calculated	Adiabatic	No Slip	Calculated	Adiabatic
[23]	Calculated	Defined	Defined	Calculated	Defined	Calculated	No Slip	Calculated	Isothermal	No Slip	Calculated	Isothermal
[12]	Calculated	Defined	Defined	Calculated	Defined	Calculated	No Slip	Calculated	Adiabatic	Defined	Calculated	Adiabatic
[9]	Calculated	Defined	Defined	Calculated	Defined	Calculated	No Slip	Calculated	Isothermal	No Slip	Calculated	Isothermal
[24]	Calculated	Defined	Defined	Calculated	Defined	Calculated	No Slip	Calculated	Heat Flux	No Slip	Calculated	Heat Flux
[25]	Calculated	Defined	-	Calculated	Defined	-	No Slip	Calculated	-	No Slip	Calculated	-
[26]	Calculated	Defined	-	Calculated	Defined	-	No Slip	Calculated	-	No Slip	Calculated	-
[27]	Calculated	Defined	Defined	Calculated	Defined	Calculated	No Slip	Calculated	Heat Flux	No Slip	Calculated	Heat Flux
[15]	Calculated	Defined	Defined	Calculated	Defined	Calculated	No Slip	Calculated	Adiabatic	No Slip	Calculated	Adiabatic

For the sealing system's input, the velocity is defined as *zeroGradient* so that the software calculates its field. The pressure is defined as *totalPressure* so that the same static pressure is set for the entire face. Regarding the temperature field, a modified boundary condition called *tec_varyingGammaTotalTemperature* is used. This applies a reference temperature T_0 to the faces according to the formula

$$T_p = \frac{T_0}{1 + \frac{\gamma - 1}{2\gamma}\Psi|u|^2}, \tag{2-12}$$

where γ is the adiabatic expansion coefficient (ratio between c_p and c_v) calculated from the propTables.

For the domain output, the velocity is also defined as *zeroGradient*, as in the input. The temperature was also defined as *zeroGradient* so that the software could calculate its field. As for the pressure field, a modified boundary condition called *tec_subsonicSupersonicPressureOutlet* was created where, for regions where there is subsonic flow, use the *totalPressure* condition and in case of supersonic flow, use the *zeroGradient* condition. There is greater stability in the solution, mitigating incoherent oscillations in the pressure field at the outlet.

The faces representing the stator and rotor walls have the same boundary condition for pressure, *zeroGradient*, allowing the software to calculate its field. For the velocity condition, both consider no slip between the fluid and the wall. However, for the rotor, it is necessary to assign the *movingWallVelocity* condition in order to set a prescribed angular velocity on this face, even without slipping. For the temperature condition, there are different hypotheses considered: isothermal [9, 18, 20, 23], heat exchange [19, 24], and adiabatic [12, 13, 15, 22]. The adiabatic condition, defined as *zeroGradient*, was adopted since the study of [12] showed good results comparing the theoretical and experimental results.

2.3.5 Fields initialization

In order to improve the convergence rate of the solution arising from the large velocity and pressure gradients, some algorithms were developed to initialize the fields in a representative way for the studied flow. For the velocity field (Figure 2.9), a Couette flow was considered in the thickness of the gas film so that, for a given position on the z axis, we have:

- The velocity is zero on the stator wall ($z = 0$) since it is not attached to the compressor shaft.
- At the rotor wall ($z = h$), the velocity is the angular velocity of the shaft.
- Between the stator and the rotor ($0 < z < h$), a linear interpolation of the speed along the z axis is performed.
- In the groove region ($z > h$), the speed is the same as the rotor.

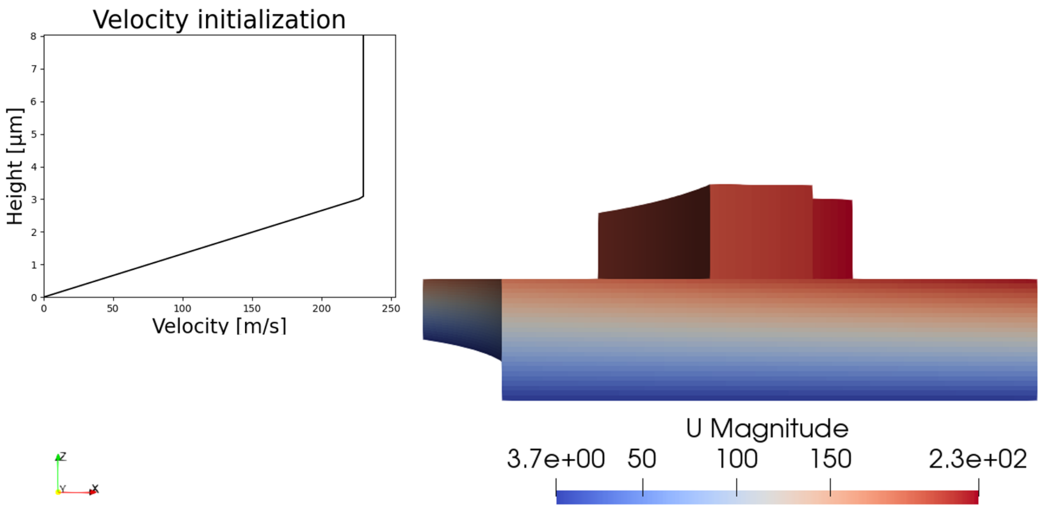


Figure 2.9: Velocity internal field initialization.

According to the literature review, it was noticed a tendency for a linear decay of the pressure field along the radius as the clearance between the faces increased. On the other hand, there is no consensus on a proportion between the pressure peak at the end of the groove region and the inlet pressure in the sealing system for smaller gaps. Thus, it was decided, for a given radial position r, to initialize the pressure field (Figure 2.10) as follows:

- In the outer radius ($r = R_o$), the pressure is the inlet pressure (P_{in}).
- In the inner radius ($r = R_i$), the pressure is the outlet pressure (P_{out}).
- Within the annular space ($R_i < z < R_o$), the linear interpolation of the pressure along the radial position is performed.

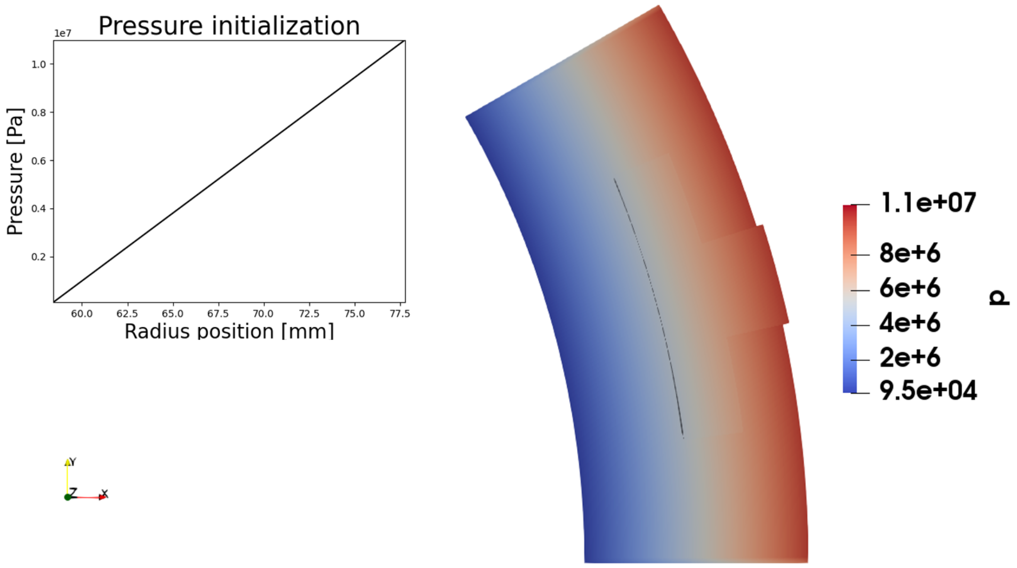


Figure 2.10: Pressure internal field initialization.

2.3.6 Solver

OpenFOAM employs the Finite Volume Method to solve the Navier-Stokes equations. For this, it is necessary to carry out the coupling between the pressure and velocity fields in order to obtain the solution. An applicable technique is the SIMPLE method (Semi Implicit Method of Linked Equations) [57], exclusively used in steady-state regime, but it showed instabilities for compressible flows. Another method is the PISO (Pressure Implicit of Split Operations) [58], which can be used for both transient and pseudo-transient regimes. In the transient case, it is recommended to limit the Courant numbers [59] (Co) below 1 for all cells in the domain, such that

$$Co = \frac{U\Delta t}{\Delta x} \leq 1, \tag{2-13}$$

where U is the magnitude of the velocity of a cell, Δt is the time step of the solver and Δx is the characteristic length of the cell.

In the pseudo-transient solution, only the steady-state result is relevant, so that the temporal progress helps in accelerating the steady-state solution calculation. Thus, as the time step does not become a limiter to satisfy the Equation 2-13, the Local Time Stepping (LTS) technique was used. This allows each cell in the domain to have its own time step, accelerating the solution.

The solver used as a base is the rhoPimpleFoam together with LTS. It uses the PIMPLE method to perform the pressure and speed coupling, applying the SIMPLE method in each PISO time step. Second order interpolation schemes (limitedLinear) were used to discretize the convective terms. As for the normal projection of the gradients and the Laplacian, non-orthogonality correctors were used for simulations with spiral grooves.

2.3.7 Convergence criteria

The commonly used way to assess whether convergence has been achieved is to verify the magnitude of the residuals of each equation. However, as presented in [60], this criterion is not suitable for most problems, and it is recommended to evaluate the stability of reference quantities of the evaluated problem. In this way, it is used:

- Mass flow balance.
- Stability of the pressure peak.
- Stability of the minimum temperature value.

The objective of calculating the mass flow balance was to assess whether the entry and exit rates at the domain boundaries are equal. For this, it calculated the mass flow at the seal's inlet ($\dot{m}_{inlet}$) and at its outlet ($\dot{m}_{outlet}$), thus making it possible to calculate the error ($Error$) from

$$Error = \left| \frac{\dot{m}_{inlet} + \dot{m}_{outlet}}{\dot{m}_{inlet}} \right| . \tag{2-14}$$

Thus, convergence is assumed after the $Error$ takes on values lower than 0.01 for 100 consecutive iterations.

To evaluate the convergence of pressure and temperature, the values in the current iteration and in the last 100 iterations are evaluated, so that convergence is reached when the difference between their maximum and minimum is less than 50 Pa and 0.5 K, respectively, acceptable errors according to professionals of the area. Therefore, when all three criteria are satisfied, convergence is assumed, and the solver is stopped. Thus, *tecProp* is defined as a set of modifications in the rhoPimpleFoam solver (coupling performed with REFPROP and convergence criteria) and boundary conditions.

2.4 Safety margin

According to the current regulations [14], the gas inside the seal must be at least 20 °C above the dew point temperature. [15] proposed the use of the convex hull technique in the pressure and temperature pairs of all cells in the domain and, from this geometry, calculate the shortest distance between one of its edges and the saturated steam curve of a gas composition, thus defining the concept of safety margin (SM). However, this method may present inconsistent results depending on the inlet pressure of the sealing gas, leading to the mistaken identification of the most critical condition. To circumvent this limitation, it evaluates the distances of all cells from the fluid domain to the saturation curve, thus redefining the concept of safety margin. This is calculated as

$$SM = T_{CFD} - T_{sat}, \tag{2-15}$$

where T_{CFD} is the temperature of a domain's cell and T_{sat} is the temperature of the steam curve for the same pressure of the evaluated cell (P_{CFD}).

To determine the T_{sat}, only the saturated vapor curve is used, so that if P_{CFD} is between the minimum and maximum values of the saturation curve, the corresponding T_{sat} is used. If the value is above or below, the value of T_{sat} is assumed as the temperature closest to the evaluated limit, as exemplified in Figure 2.11. In this one, the blue points are the pressure and temperature pairs of all cells in the fluid domain, the black curve is the complete saturation curve of the gas, the green curve is the reference steam curve for calculating the SM and the red dashed line represents the magnitude of the SM.

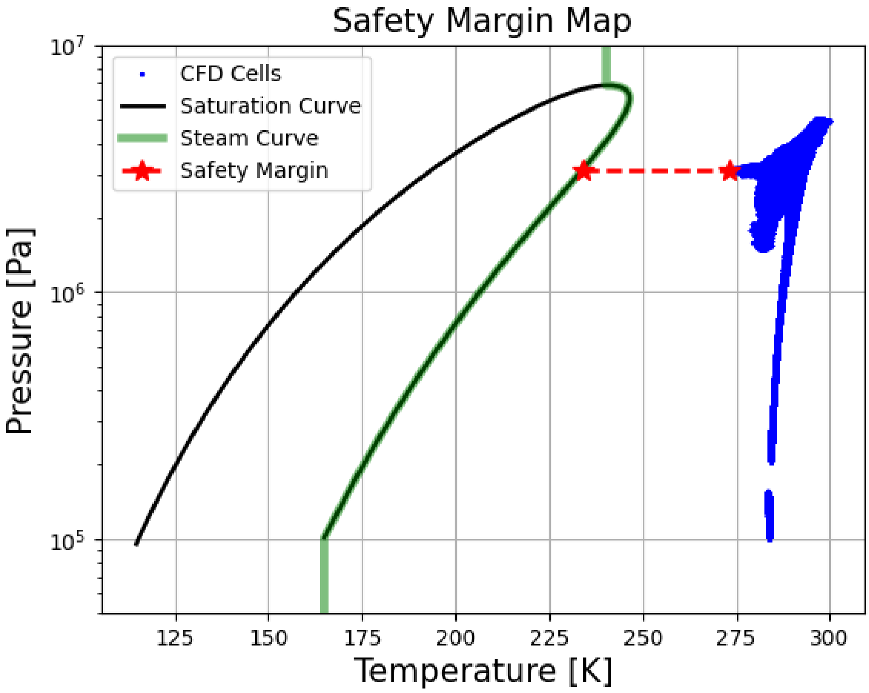


Figure 2.11: Example of a safety margin map.

Thus, it is possible to verify the intrinsic relationship between the SM and the saturation curve of the evaluated gas, confirming the importance of understanding the fluid composition to evaluate properly the reliability of the system. In order to exemplify the relationship between the saturation curve and the molar fraction, a mixture of methane, ethane and octane was created so that only the octane composition was modified (Figure 2.12). It is possible to verify that, as the concentration of the denser hydrocarbon increases, the saturation curve expands to the right of the graph, without modifying the saturated liquid curve, thus decreasing the operational safety.

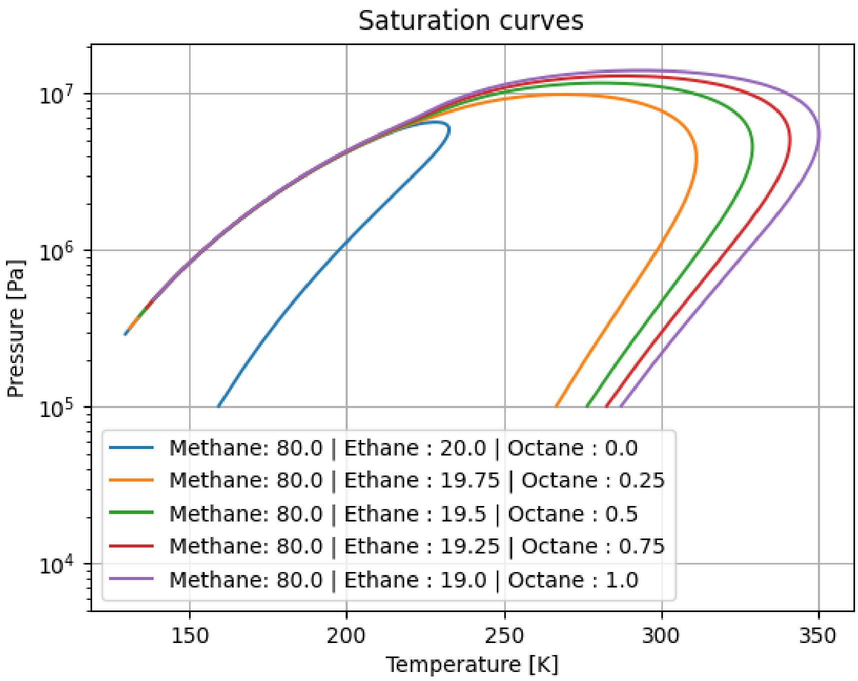


Figure 2.12: Example of saturation curves with different octane molar fraction concentration.

2.5 CFD model validation

2.5.1 Introduction

The validation of an adequate computational model is of paramount importance for its extrapolation to different geometries and operating conditions. To do so, it was used a computer with Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz and 256 GB of RAM and simulating in a parallel manner in 4 processes. The work developed by [11] was used to validate the computational model, as it provides experimental data. As it is one of the forerunners in the study of DGS, it is used as a benchmark by other authors. Thus, the results obtained through tecProp, using the previously exposed boundary conditions, are compared with the studies by [27] and [23]. In this way, comparing the congruence of results and the errors obtained among the other authors is possible with the benchmark used.

In [11], the influence of the clearance between the sealing faces on the pressure distribution along the radius and the opening force perpendicular to the faces was evaluated. These conditions were recreated following the geometric parameters and operating conditions, Tables 2.3 and 2.4 respectively, of the performed experiment.

Table 2.3: DGS geometric parameters with spiral grooves evaluated in [11].

Parameter	Symbol	Value
Inner radius	R_i	58.42 mm
Groove radius	R_g	69.00 mm
Outer radius	R_o	77.78 mm
Groove depth	h_g	5.00 μm
Ratio of groove to land	β	1.00
Spiral angle	α	15 degrees
Number of grooves	N	12

Table 2.4: Operating conditions of the DGS with spiral grooves evaluated in [11].

Parameter	Symbol	Value
Composition	-	Air
Clearance	h	2.03 | 3.05 | 5.08 μm
Inlet temperature	T_{in}	303 K
Angular velocity	ω	10380 RPM
Inlet pressure	P_{in}	4.5852 MPa
Outlet pressure	P_{out}	0.1013 MPa

The results obtained by tecProp are compared with [27] and [23] in two aspects. The first one evaluates the pressure distribution along the radius following the red line in Figure 2.13. This line consists of an 8.5 degrees straight line concerning the beginning of the domain circular section, starting at R_i to R_g at the height of half the thickness of the gas film to eliminate influences from the walls in the analysis. Then, the curve follows the profile described by the equation 2-3 from R_g to R_o at a height $h + 0.5h_g$.

The second aspect evaluated is the opening force exerted on the DGS. Since the system is in equilibrium, the force exerted perpendicularly to the parallel faces of the stator and rotor are equal. This evaluation is essential because it is complementary to the first one, considering that it consists of the sum of the product between the cell face area and its respective pressure across a face of interest. Thus, it is possible to verify if the pressure distribution in the domain as a whole is coherent.

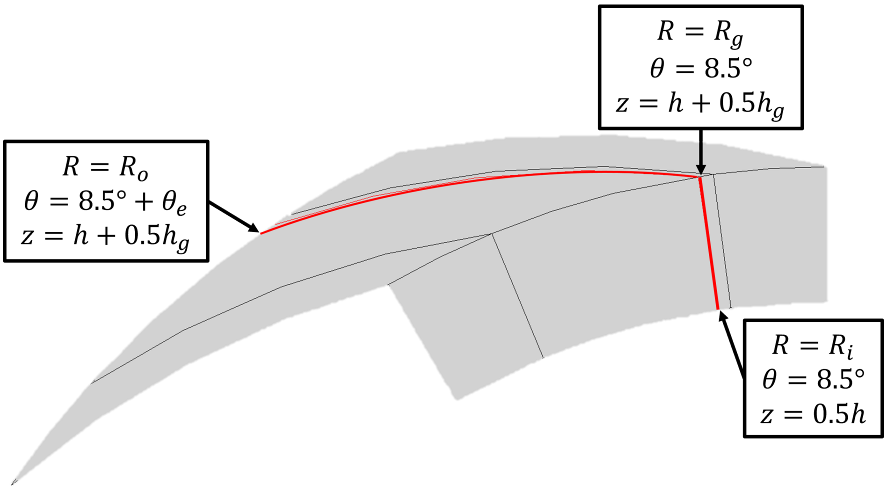


Figure 2.13: Path used for pressure measurement in model validation.

2.5.2 Spiral groove mesh convergence

Before carrying out the simulations for the different clearances, it is necessary to ensure that the appropriate mesh is used through a mesh convergence analysis. For this purpose, the 2.03 μm clearance was considered as it is the most unstable operating condition, which probably presents the highest pressure peak among the conditions to be evaluated. Thus, different mesh refinements and the highest pressure obtained in the simulated control volume were evaluated.

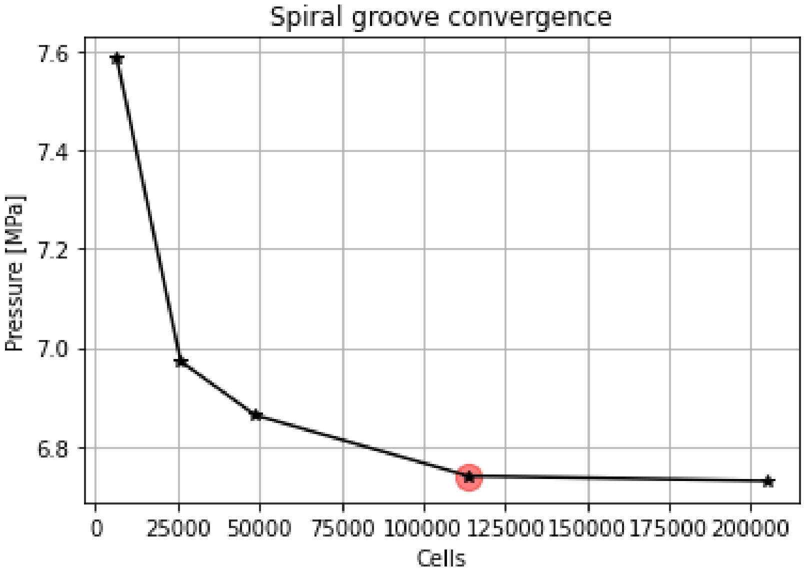


Figure 2.14: Mesh convergence analysis for a clearance of 2.03 μm with the same conditions adopted in [11].

From Figure 2.14, it is possible to verify that the maximum pressure did not show significant changes from the 113750-cell mesh, which is the refinement chosen for the other simulations. From the selected mesh, it is possible to evaluate in Figure 2.15 the pressure field in the entire domain. As the equipment rotates counterclockwise, it is possible to notice a more significant pressure on the wall on the right side of the groove than on the left side, and as one approaches the end of the groove region, this difference tends to increase. Analyzes regarding the pressure behavior radially are made separately for each of the simulated clearances.

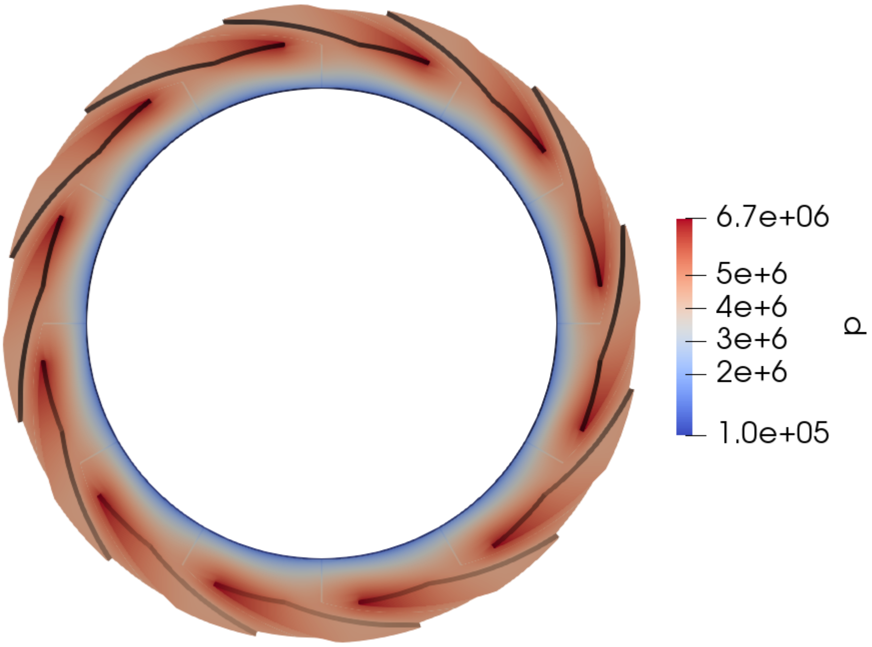
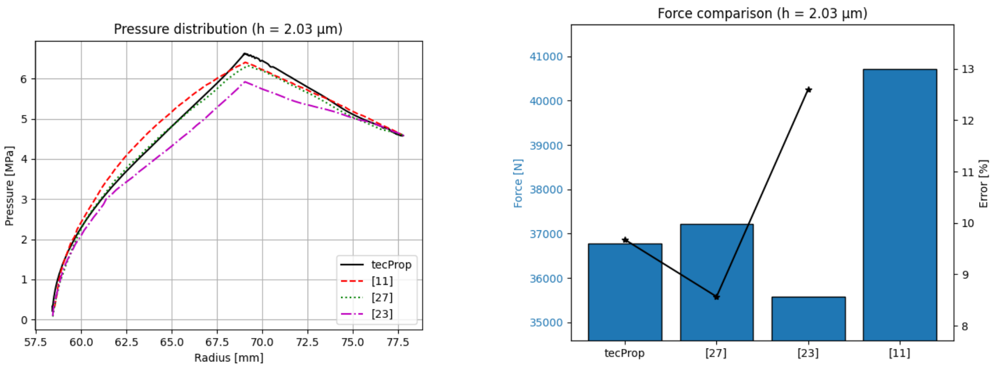


Figure 2.15: Pressure field distribution inside the complete fluid domain with a 2.03 μm clearance.

2.5.3 Results

Evaluating the pressure profile obtained with tecProp with a clearance of 2.03 μm in Figure 2.16, it is possible to verify a similar behavior to that obtained by [11], with 3.4 % average error between the two profiles. As expected, there was a pressure increase at the end of the grooved region (R_g) due to the abrupt reduction of space between the sealing faces. Furthermore, when evaluating the force performed on the stator in Figure 2.16, it is possible to see the error between the obtained result and the other references compared to [11] is similar. In this case, a more significant error was expected as it is a small clearance where turbulence effects are maximized due to high-pressure gradients between the groove region and the edges.

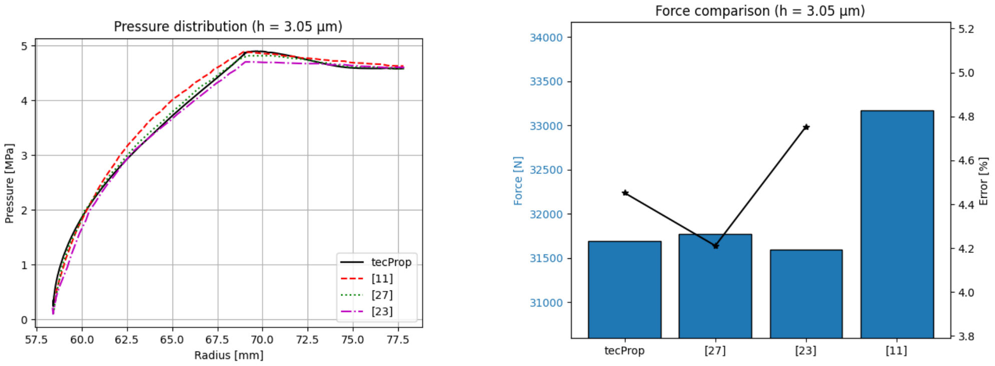
Analyzes were also carried out for the clearance of 3.05 and 5.08 μm. In these, it is verified that the influence of the groove is attenuated as expected, reducing the peak pressure in R_g as shown in Figures 2.17 and 2.18. In these, it is possible verifying similar profiles to those obtained by [11], checking for the clearance of 3.05 and 5.08 μm average errors of 2.1 % and 3.3 % respectively. It is important to note that, as the clearance increases, the peak pressure at R_g decreases. Thus, the pressure profile tends to be linear, and as a consequence, the maximum pressure is prone to be at the entrance of the system, at R_i. This behavior is also noticed in the studies developed by [27] and [23]. As a consequence, it can be seen that the values obtained for the opening force tend to be closer, as can be seen in Figures 2.17 and 2.18. In order to assess the influence of clearance on the opening force, it is verified in Table 2.5 that as the clearance increases, the force tends to decrease, as also perceived by [27] and [23]. This behavior is expected, considering that the attenuation of the groove effect directly contributes to reducing the pressure barrier in the region close to R_g. Table 2.6 evaluates the error between the case studies, including tecProp, compared to [11]. It is verified that tecProp has errors similar to other references and presents consistent pressure distributions, proving the applicability of the defined model.



(a)

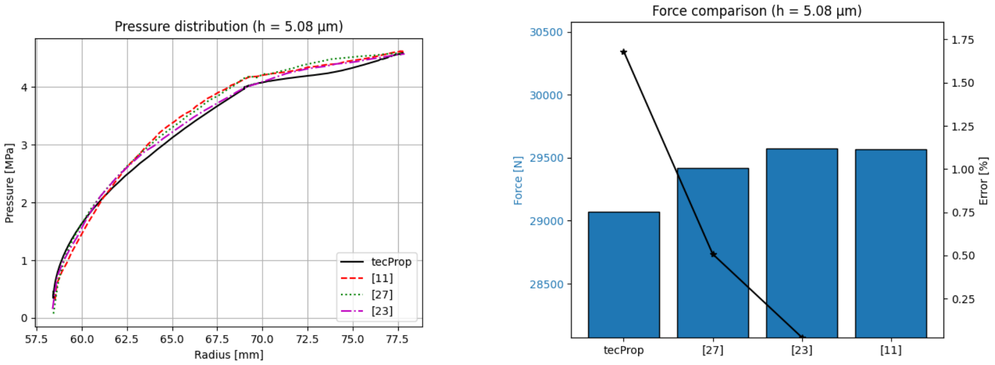
(b)

Figure 2.16: Comparison of (a) pressure distribution and (b) opening force obtained through tecProp, for a clearance of 2.03 μm, with [11], [27], and [23].



(a) (b)

Figure 2.17: Comparison of (a) pressure distribution and (b) opening force obtained through tecProp, for a clearance of 3.05 μm, with [11], [27], and [23].



(a) (b)

Figure 2.18: Comparison of (a) pressure distribution and (b) opening force obtained through tecProp, for a clearance of 5.08 μm, with [11], [27], and [23].

Table 2.5: Comparison of the opening force obtained using tecProp with [11], [27], and [23].

Clearance	tecProp	[27]	[23]	[11]
2.03 μm	36774 N	37224 N	35584 N	40712 N
3.05 μm	31692 N	31772 N	31592 N	33169 N
5.08 μm	29071 N	29418 N	29576 N	29568 N

Table 2.6: Percent error of opening force obtained through tecProp, [27], and [23] with respect to [11].

Clearance	tecProp	[27]	[23]
2.03 μm	9.67 %	8.57 %	12.60 %
3.05 μm	4.45 %	4.21 %	4.75 %
5.08 μm	1.68 %	0.51 %	0.03 %

2.5.4
T groove mesh convergence

Considering that the boundary conditions of the CFD model were validated according to the literature, a T groove geometry described in Table 2.7 was selected for the following case studies. This definition aims to decrease the execution time of each simulation, as a spiral groove takes an average of 22 hours to complete the simulation, while a T groove only takes an average of 12 hours.

Table 2.7: T groove geometry obtained from an operational DGS of a standard platform.

Parameter	Symbol	Value
Inner radius	R_i	87.50 mm
Groove radius	R_g	97.90 mm
Middle radius	R_m	103.50 mm
Outer radius	R_o	109.05 mm
Groove depth	h_g	5.00 μm
Ratio of groove to land	β_1	1.00
Ratio of groove geometry	β_2	1.00
Number of grooves	N	12

In order to guarantee the consistency of the results, a mesh convergence test with the T groove was carried out using the same operating conditions described in Table 2.4 and a 2.03 μm clearance. It is possible to verify in Figure 2.19 that from the mesh with 262500 cells, the pressure variation is not significant, which is the mesh chosen for the other simulations with this geometry. It is possible to notice an increase in the number of cells needed for simulations with the T geometry compared to the spiral one. This increment was expected due to the 77 % increase in the volume of the fluid domain.

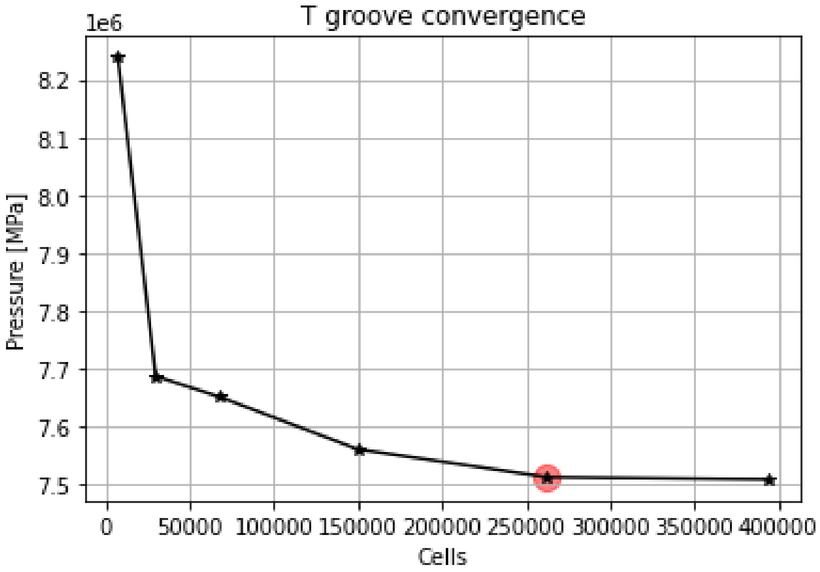


Figure 2.19: Mesh convergence analysis for a T groove DGS with a clearance of 2.03 μm.

3
Machine learning model construction

3.1
Model objective

As presented in Chapter 2, to verify the operational reliability of a DGS through its safety margin (SM), follow the steps explained in Figure 3.1:

1. Define the sealing gas composition.
2. Create propTables as described in Section 2.3.3.
3. Define the operating conditions of the DGS (inlet pressure and temperature, angular velocity and clearance between faces).
4. Use propTables and operating conditions in tecProp to calculate the pressure and temperature fields inside the seal.
5. Calculate the safety margin for these operating conditions and sealing gas composition.

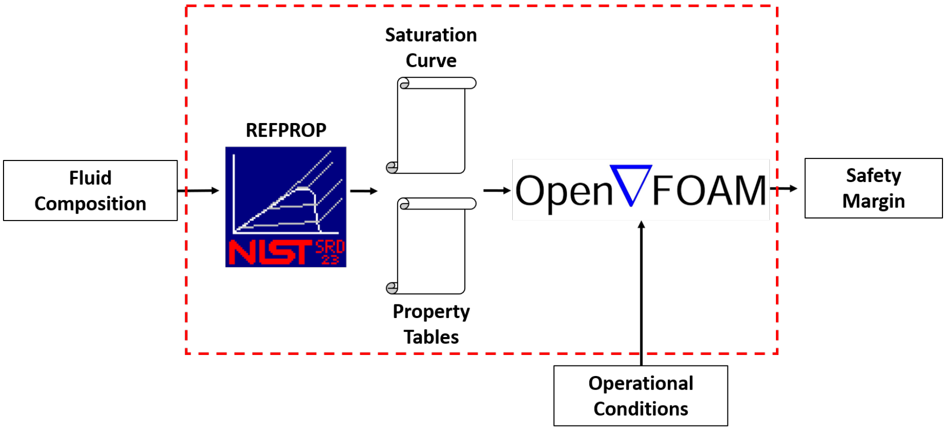


Figure 3.1: Schematic for obtaining the safety margin from a CFD simulation based on the sealing gas composition and the operating conditions of the DGS.

The aim of this project is to create a predictive model to replace the steps in the red dashed rectangle of Figure 3.1. It provides the molar concentrations of the process gas (*i.e.*, a mixture of methane, ethane and octane, where the molar concentration of ethane is defined as a linear combination of the concentrations of methane and octane) and the operational conditions as inputs, and the SM is the output of the model. In Figure 3.2, the inputs and outputs of the model are schematized, so that the inputs are on the left and the output on the right.

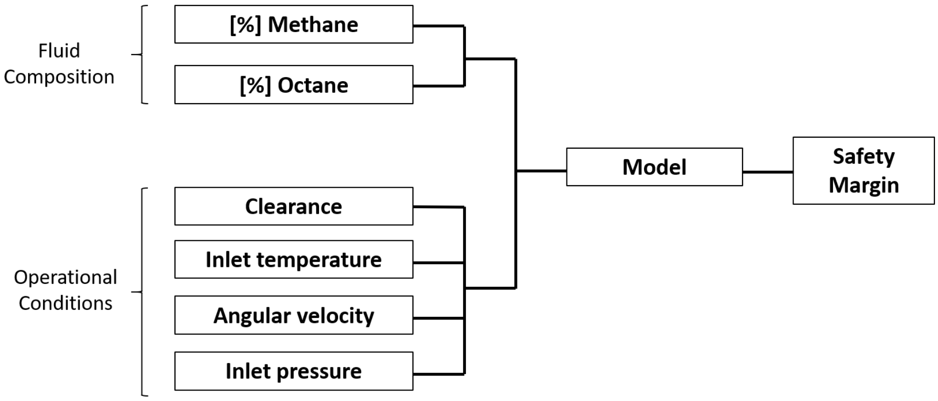


Figure 3.2: Schematic for obtaining the safety margin from the ML model based on the sealing gas composition and the operating conditions of the DGS.

3.2 Dataset

To create a robust model as described in Section 3.1, it is necessary to define firstly the ranges of values that the model should be able to cover. In Table 3.1, the upper and lower bounds of the model's input variables are exposed. The parameters referring to the operational condition of the DGS (inlet pressure and temperature, angular velocity and clearance) were defined from the ranges evaluated in the studies described in Chapter 1. The gas composition was defined based on what is shown in Figure 1.3, so that the proportions were normalized to guarantee a molar concentration sum of 100 %. Therefore, a mixture of methane, ethane, and octane was evaluated. The octane concentration in this gas mixture directly implies the position of the super-heated steam curve (Figure 2.12) and, consequently, the safe operating condition.

Table 3.1: Lower and upper bounds for each of the input variables in the machine learning model to be built.

Operational condition	lower bound	upper bound
Methane $[CH_4]$	70 %	80 %
Octane $[C_8H_{18}]$	0 %	1 %
Clearance $[h]$	3 μm	5 μm
Inlet temperature $[T_{in}]$	293 K	423 K
Angular velocity $[\omega]$	10000 RPM	20000 RPM
Inlet pressure $[P_{in}]$	1 MPa	10 MPa

To define the simulations to be performed to compose the dataset, a Design of Experiments (DoE) technique is used, as in [10, 36, 41]. The DoE technique used is the Full Factorial Design (FFD) [61], which comprises testing all combinations of inputs for the desired phenomenon. Thus, we define variables as *factors* and each value of a factor is a *level*. Thus, assuming that all factors have the same number of levels, N_{CFDs} simulations are needed to compose the dataset, according to

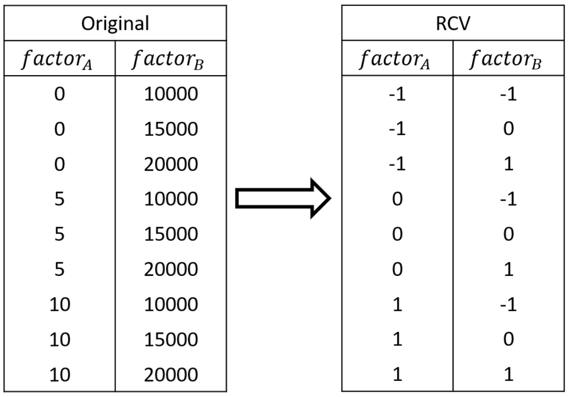
$$N_{CFDs} = (N_{levels})^{N_{factors}}, \tag{3-1}$$

where N_{levels} and $N_{factors}$ are respectively the number of levels and factors. Thus, for this study, the 3-level FFD is used so that, for a factor, the maximum, minimum and average values are evaluated.

After performing all the CFD simulations, each factor of the dataset is normalized between -1 and 1 so that all factors are on the same scale, with no distortion in the value ranges [62] (exemplified in Figure 3.3). Thus, for a level of a factor F of the dataset, its value as Reduced Centralized Variable (RCV) is defined by

$$f_i = \frac{F_i - \frac{F_{max}+F_{min}}{2}}{\frac{F_{max}-F_{min}}{2}}, \tag{3-2}$$

where F_{max} and F_{min} are, respectively, the highest and lowest levels of the factor F, and f_i is the value in RCV form of the original value F_i.



Original	
$factor_A$	$factor_B$
0	10000
0	15000
0	20000
5	10000
5	15000
5	20000
10	10000
10	15000
10	20000

RCV	
$factor_A$	$factor_B$
-1	-1
-1	0
-1	1
0	-1
0	0
0	1
1	-1
1	0
1	1

Figure 3.3: Exemple of conversion of a 3-level FFD dataset (left) to RCV form (right).

3.3 Regression models

Among the branches of AI, Machine Learning (ML) is the field of study that gives computers the ability to learn without being explicitly programmed [63]. This learning stage is called training and, later, it is possible to make predictions of new data. The training of algorithms from labeled data (i.e., the expected output for an input is known) is called *supervised learning*. On the other hand, when you only have access to the inputs, it is called *unsupervised learning*. Supervised learning methods can further subdivide into two groups, depending on the type of output desired. It is defined as a *regression* problem if the model output is continuous, and a *classification* problem if the output is categorical (e.g., true/false).

The aim of this study is to create a regressive ML model from the dataset described in Section 3.2. For this, the dataset is divided into two sets, the training and the test. The training set adjusts the model for interpreting patterns according to the errors between the true outputs and those estimated by the model. On the other hand, the test set consists of data that the model has never had access to, being used to assess the generalizability of the trained model. In this study, a proportion of 80 % of the data is used for training and the other 20 % for testing. Thus, the aim is to adjust a function that, for a set of inputs ($X_{n \times m}$), best approximates their respective outputs ($y_{n \times 1}$), where n is the number of samples and m is the number of features.

3.4 Types of regression models

In this section, the basis of the five types of regression models that are evaluated in this study are presented:

- Linear
- Neighbors
- Support vector machines
- Tree
- Ensemble

3.4.1 Linear

Linear models aim to predict the output, assuming that it can be written as a linear combination of the model inputs. The best-known model in this category is the Linear Regression [64], which seeks to calculate the values of the vector ω that minimize

$$\min_{\omega} \sum_{i=1}^{n} (X_i \omega - y_i)^2 \tag{3-3}$$

to obtain

$$y_{pred} = X\omega, \tag{3-4}$$

where y_{pred} is the predicted output.

There are also more robust linear models that also are evaluated in the present study: BayesianRidge [65], ElasticNet [66], HuberRegressor [67], Lars [68], Lasso [69], LassoLars [70], OrthogonalMatchingPursuit [71], PassiveAggressiveRegressor [72], RANSACRegressor [73], Ridge [74], and SGDRegressor [75].

3.4.2 Neighbors

Neighbor-based models estimate the output of a desired condition from the outputs of the points closest to it. The evaluated algorithm is the k-nearest Neighbors (KNN) [76], which selects the closest k points from the dataset to calculate the prediction of a condition. It is possible to use different distance metrics, but the most common is the Euclidean distance D_e, defined as

$$D_e(X_i, X_j) = \sqrt{\sum_{c=1}^{m} (X_{i,c} - X_{j,c})^2}, \tag{3-5}$$

where X_i and X_j are the points to be compared from the dataset $X_{n \times m}$, and c corresponds to the evaluated column.

3.4.3 Support vector machines

Models based on Support Vector Machines aim to determine a hyperplane that best fits the data in a larger dimension than the original $\mathbb{R}^m$. The base algorithm is the Support Vector Regressor (SVR) [77], which uses in the interpolation the dataset points that are within a margin $\pm\varepsilon$ of the desired hyperplane and penalizes, from a regularizer C, distances greater than its margin ε both above (ζ_i) and below (ζ_i^*).

The optimization problem solved is

$$\min_{\omega,b,\zeta,\zeta^*} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*) \tag{3-6}$$

subject to

$$\begin{cases} y_i - \omega^T\phi(x_i) - b \leq \varepsilon + \zeta_i \\ \omega^T\phi(x_i) + b - y_i \leq \varepsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \geq 0 \\ i = 1, \dots, n \end{cases}, \tag{3-7}$$

where $y_i \in \{-1, 1\}^n$, ω and b are the parameters that parameterize the desired hyperplane, and $\phi(x_i)$ is the Kernel function responsible for increasing the dimensionality of the problem by mapping a given x_i into a space larger than $\mathbb{R}^m$. Moreover, this study also evaluates two other methods derived from SVR: LinearSVR [77] and NuSVR [78].

3.4.4 Tree

Models based on Decision Trees aim to create hierarchical rules for separating features and their respective outputs so that, at the last node of a given tree path, it is possible to make a prediction based on the input features. The tree starts at a main node (root node), is successively divided in order to separate it into the most distinct groups possible (hidden nodes) until reaching the last node (leaf nodes), which defines the continuous value for the desired inputs.

The algorithm used is the CART (classification and regression tree) which comprises, for a feature evaluated in a node m and an associated threshold value t_m, minimizing

$$G(Q_m) = \frac{N_m^{left}}{N_m}H(Q_m^{left}) + \frac{N_m^{right}}{N_m}H(Q_m^{right}), \tag{3-8}$$

where N_m^{left} and N_m^{right} are the number of left and right samples, respectively, after the separation at node m is performed, N_m is the total of samples at node m, $H()$ is the loss function used to compute the error and Q_m^{left} and Q_m^{right} are the left and right samples, respectively. Thus, this study evaluates two tree-based methods: DecisionTreeRegressor [79] and ExtraTreeRegressor [80].

3.4.5 Ensemble

Ensemble methods comprise the use of multiple algorithms together to improve the result of the predictions of these models separately (weak learners). There are two main types of ensemble methods: bootstrap aggregating (bagging) and boosting. The bagging method comprises training each of the weak learners in parallel using random sampling of the training set. Thus, after the training stage, the output of the ensemble model is the average of the results obtained by the independent algorithms. The boosting method, on the other hand, consists of training one model at a time in sequence so that the samples with the highest prediction errors of the training set have greater weights in the training of the next model. As a result, ensemble models present more accurate results than weak learners separately. In this study, only ensemble methods based on regression trees (Section 3.4.4) are used: AdaBoostRegressor [81], BaggingRegressor [82], ExtraTreesRegressor [80], GradientBoostingRegressor [83], RandomForestRegressor [84], XGBRegressor [85].

3.5 Performance metrics

In order to compare the performance of different models, it is possible to use different metrics. First, for a given input set $X_{n \times m}$, where n is the number of simulations and m is the number of parameters. Additionally, define y_{true} and y_{pred} as the output vectors of the system, with an n dimension, of the safety margins obtained through the CFD simulations and the predictive model, respectively.

3.5.1 Coefficient of determination and adjusted coefficient of determination

To understand the concept of coefficient of determination (R^2) and adjusted coefficient of determination (R_a^2), according to [61], it is necessary to define the concept of residual (Res)

$$Res(y_{true},\ y_{pred}) = y_{true} - y_{pred}, \tag{3-9}$$

where y_{true} is the simulated value and y_{pred} is the model prediction.

Assuming that the mean of the experimental responses and those obtained by the model are equal, it is possible to decompose the quadratic sum (SQ) of the terms. Thus, the relationship between the total quadratic sum (SQ_T) and the quadratic sums of the regression (SQ_R) and the residual (SQ_r) is obtained, such that

$$\sum_{i=1}^{n}(y_{true_i}-\bar{y})^2=\sum_{i=1}^{n}(y_{pred_i}-\bar{y})^2+\sum_{i=1}^{n}Res(y_{true_i},\ y_{pred_i})^2 \leftrightarrow SQ_T=SQ_R+SQ_r, \tag{3-10}$$

where

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_{true_i}. \tag{3-11}$$

Thus, it is defined that the coefficient of determination is obtained as

$$R^2(y_{true},\ y_{pred})=\frac{SQ_R(y_{true},\ y_{pred})}{SQ_T(y_{true},\ y_{pred})}=\frac{SQ_R(y_{true},\ y_{pred})}{SQ_R(y_{true},\ y_{pred})+SQ_r(y_{true},\ y_{pred})}. \tag{3-12}$$

Thus, it is possible to verify that the model's objective must be to obtain values as close as possible to 1, as it means that the residuals tend to 0, implying a representative model.

R^2 has a significant limitation because, as the number of predictors in the model increases, it tends to model the random noise of the data, causing overfitting. In order to solve this limitation, the calculation of the adjusted correlation coefficient R_a^2 is performed, taking into account the degrees of freedom to adjust the R^2 previously calculated as

$$R_a{}^2(y_{true},\ y_{pred})=1-(1-R^2(y_{true},\ y_{pred}))\frac{n-1}{n-m}. \tag{3-13}$$

3.5.2 Infinity norm

The infinity norm ($norm_\infty$) is a particular case of the p-norm

$$norm_p(y_{true},\ y_{pred})=\|Res(y_{true},\ y_{pred})\|_p=\left(\sum_{i=1}^{n}|Res_i(y_{true},\ y_{pred})|^p\right)^{\frac{1}{p}}, \tag{3-14}$$

when $p = \infty$. In this way, the infinity norm becomes the maximum absolute value among the terms of the residual vector (Res)

$$norm_{\infty}(y_{true},\ y_{pred}) = \|Res(y_{true},\ y_{pred})\|_{p=\infty} = max\{|Res(y_{true},\ y_{pred})|\}. \tag{3-15}$$

3.5.3
Mean absolute error

The mean absolute error (MAE) consists of calculating the mean of the absolute value of the residuals, expressed as

$$MAE(y_{true},\ y_{pred}) = \frac{\sum_{i=1}^{n} |Res_i(y_{true},\ y_{pred})|}{n}. \tag{3-16}$$

It is worth pointing that, since it is the absolute value, equal residuals with opposite signs are not canceled, and do not underestimate the error.

3.5.4
Root mean square error

The mean squared error (MSE), calculated as

$$MSE(y_{true},\ y_{pred}) = \frac{\sum_{i=1}^{n} Res_i(y_{true},\ y_{pred})^2}{n}, \tag{3-17}$$

is a metric that aims to penalize more significant errors by squared each term of the residual vector.

In order to present the error value obtained by the metric in the same dimension as the analyzed variable, the root mean square error (RMSE) is calculated

$$RMSE(y_{true},\ y_{pred}) = \sqrt{MSE(y_{true},\ y_{pred})} = \sqrt{\frac{\sum_{i=1}^{n} Res_i(y_{true},\ y_{pred})^2}{n}}. \tag{3-18}$$

3.5.5
Mean absolute percentage error

The mean absolute percentage error (MAPE), calculated as

$$MAPE(y_{true},\ y_{pred}) = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{Res_i(y_{true},\ y_{pred})}{y_{true}} \right|, \tag{3-19}$$

is a metric that aims to assess point-to-point relative errors, regardless of the scale of the variable of interest.

3.5.6 Graphical and statistical analysis

For a given trained model, it is possible to calculate the error (Equation 3-9) between the true (y_{true}), and the predicted (y_{pred}) values in the test step. Three different approaches are presented. The first one is a histogram of the errors (Figure 3.4), where its x-axis represents the error and the y-axis the number of occurrences within a given error interval, defined by the scale of the graph. The objective is to obtain a greater density of occurrences at 0 and with the smallest possible scattering, resulting in a more accurate and precise model. Furthermore, the errors obtained are expected to be zero or positive for operational safety reasons, implying a model that underestimates the safety margin.

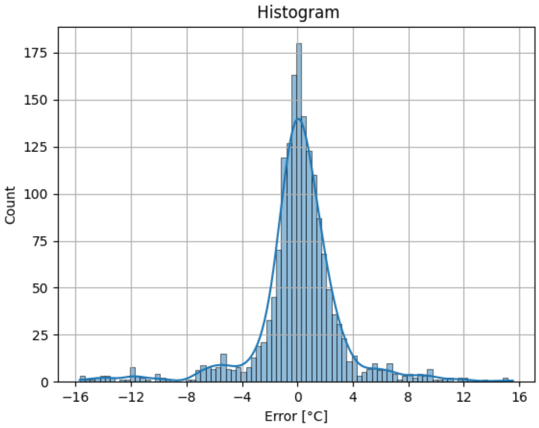


Figure 3.4: Example of histogram of residues.

The second approach is the regression diagram shown in Figure 3.5, where the x-axis has the y_{true} and the y-axis has the y_{pred}. The dotted line represents error condition 0, having a slope of 45 degrees. Points below this line correspond to an underestimated safety margin prediction, still ensuring safe operation. Furthermore, in the title of the image, three parameters of the model are presented:

- R^2: coefficient of determination of the model (Equation 3-12).
- Adjusted R^2: adjusted coefficient of determination of the model (Equation 3-13).
- Maximum Error: maximum absolute error. This is the largest absolute value between the minimum and maximum shown in Table 3.2.

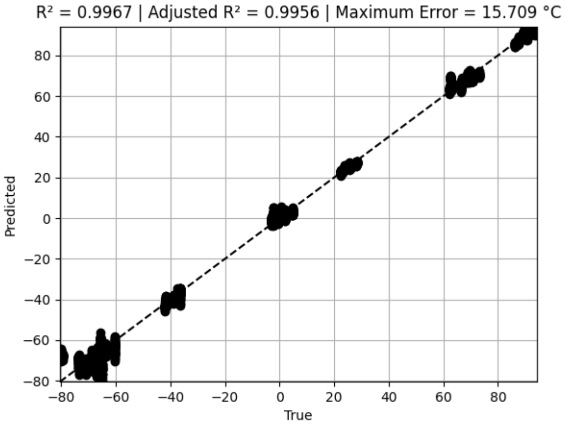


Figure 3.5: Example of regression diagram.

Finally, the Table 3.2 has some statistical characteristics of the errors obtained from a given model are evaluated, namely:

- mean: mean of the errors.
- std: standard deviation of errors.
- min: minimum error obtained.
- 25%: 25% of the errors obtained are below this value.
- 50%: median of errors.
- 75%: 75% of the errors obtained are below this value.
- max: maximum error obtained.
- time: total time to make a prediction.

Table 3.2: Example of statistic table to be evaluated

Model	mean	std	min	25%	50%	75%	max	time
Model	0.090	3.226	-15.709	-0.878	0.126	1.294	15.536	0.1 s

3.6 Model construction and validation

The first step is to choose the desired regression model. The separation of the training and test sets consists of defining a reference number, called a seed, and then performing the Mersenne Twister [86] algorithm. This algorithm aims to create, in a deterministic way, pseudo-random numbers that are used to select the samples for training and those for the test. Thus, for a given seed, the Mersenne Twister algorithm is used to select the percentage of samples chosen to compose the training set ($[\%]_{train}$) and the others to be used as a test set, a process called the holdout approach (Algorithm 2 in Appendix A).

After the split, the training data is fitted, solving the characteristic optimization problem of each regression model, considering the model's parameters as the constraints. Each model has specific and characteristic parameters for its creation, called hyper-parameters, that can be modified in order to improve its performance. With the fitted model, it is possible to predict the output for the test data inputs, and use a performance criterion, in this case, the infinity norm, where the true (y_{true}) and predicted (y_{pred}) results are compared.

However, the use of the holdout technique to create only one test group can impact the model's performance with biased results. Another approach used is the k-fold cross-validation (CV) technique, exemplified in the Algorithm 3 in Appendix A. This procedure consists of dividing the dataset into k subsets for each regression model with its respective hyper-parameters. Thus, k-1 folds are used to train the model, and the remainder is used for validation. This procedure is repeated k times until all subsets have been used for validation once. In each of the k times, the model's performance is evaluated when compared with the validation fold of the step. Thus, the model's performance for a given combination of hyper-parameters is the average of the performances of the k steps.

3.7 Tuning methods

The default configuration of the models' hyper-parameters available in Scikit-Learn [87] may not be the most suitable for creating a robust predictive model. There are different ways to improve a model's performance through modifications in the hyper-parameters values, known as tuning. The two most popular tuning techniques are Grid and Random Search. The first step for both procedures is the same, selecting the hyper-parameters to be evaluated and their possible values.

The Grid Search (GS) is the most intuitive method and evaluates all possible combinations of parameters so that the chosen one is the one that predicts the test group with the slightest error. However, this method is indicated when there are few total combinations, as it tends to become computationally expensive as the number of evaluated parameters or the discretization increases. The Random Search (RS) [88] was proposed to reduce the computational demand required by a Grid Search, evaluating only a stipulated number of random combinations. In the research, the methodology was tested and compared with the Grid Search method for different algorithms and datasets, and resulted in better models in most cases while required less computational resources.

In the Figure 3.6, it is possible to compare a case in which the objective is to improve the hyper-parameters 1 and 2. It is possible to verify that, in the Grid Search method, all possible combinations between the parameters are evaluated, which can be computationally exhaustive. On the other hand, the Random Search method uses a predefined number of samples, randomly chosen among all possible combinations, exemplified in the Algorithm 4 in Appendix A.

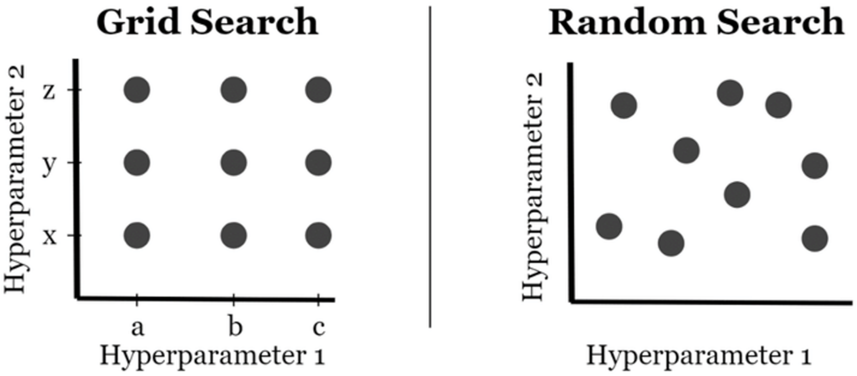


Figure 3.6: Difference between grid and random search techniques [88].

3.8 Random search k-fold cross-validation holdout

Based on the points discussed in Sections 3.6 and 3.7, a Random Search k-fold cross-validation holdout is performed to create and tune the model. For a given regression model, the detailed process for choosing the best hyper-parameters to be used and their respective performance response is presented. The performance metric used as a criterion is the infinity norm described in Subsection 3.5.2.

The first step is defining the number of seeds ($nseeds$) to be tested, so that the seed's value is incremented by 1 each iteration until $nseeds$. One hundred seeds are evaluated in order to provide stability in the model performance estimation, following the recommendation of being in the 50-200 range presented in [89]. Thus, for a given seed, the Mersenne Twister algorithm is used to select the percentage of samples chosen to compose the training set ($[\%]_{train}$) and the others to be the test set. In this study, 80 % of the data is used for training and 20 % for testing.

Then, the grid is made from the possible values that each hyper-parameter can assume ($params$). Furthermore, the number of random combinations of hyper-parameters ($ncombs$) that are tested is also defined, with 100 being the amount adopted in this study. Thus, each combination of parameters is individually tested through the k-fold CV technique. In this work, 5 folds are used for the k-folds CV step. This step consists of dividing the training set into k folds (exemplified in Figure 3.7) where the performance in the prediction of each validation group is saved in $Score_i$. Thus, the performance for a given combination of parameters ($Score_p$) is the average of the $Score_i$ of this model setup. This process is performed $ncombs$ times, and the best combination of parameters is updated at each iteration.

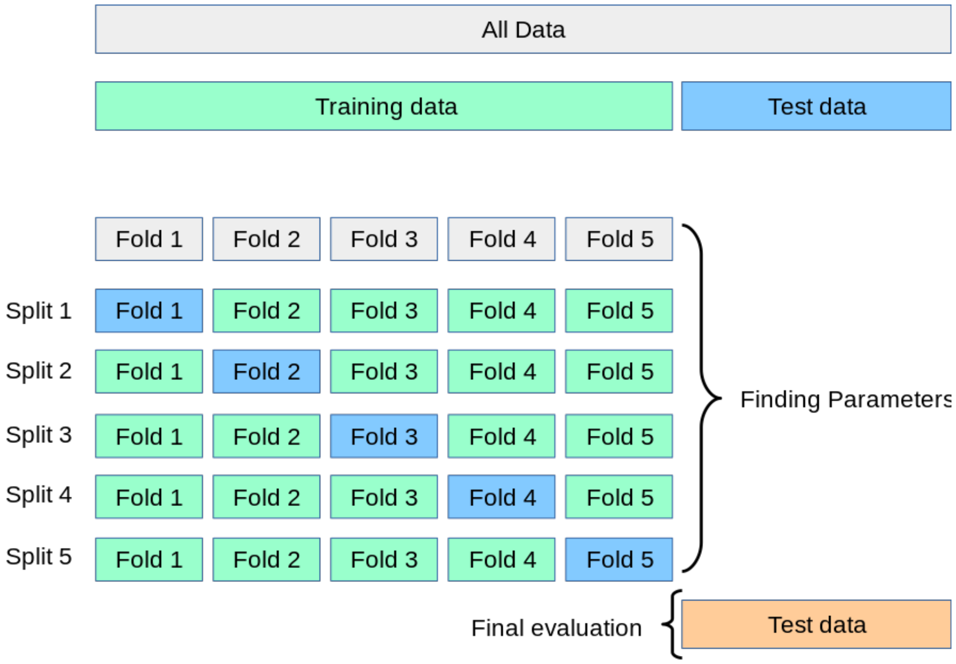


Figure 3.7: k-fold cross-validation concept [87].

After finding the best combination of hyper-parameters for a given $seed$, the model is fitted with the entire training set. Subsequently, the test set output is predicted (y_{pred}), which allows to compare it with the expected response (y_{true}). Thus, if the model's performance with its current configuration for this *seed* is better than the one previously calculated (*best_score*), it is updated the best values of error (*best_error*), hyper-parameters (*best_params*) and seed (*best_seed*). Thus, at the end of the process, the seed and its respective best combination of parameters is defined, giving the evaluated model better predicting performance. The Algorithm 1 is a pseudo-code of the presented method.

Algorithm 1 Random search k-fold cross-validation holdout.

```
$x \leftarrow input$
$y \leftarrow output$
$[\%]_{train} \leftarrow train\ percentage$
$nseeds \leftarrow number\ of\ seeds$
$model \leftarrow desired\ model$
$params \leftarrow hyper-parameters$
$ncombs \leftarrow number\ of\ combinations$
$k \leftarrow number\ of\ subsets$
procedure RS_KFOLDCV_HOLDOUT($nseeds$, $k$, $params$, $ncombs$)
    for $seed = 1 \ldots nseeds$ do
        Split the dataset into training and test set for the given $seed$.
        Create the complete combinations of parameters grid.
        Select $ncombs$ random combinations from the original grid.
        for $p = 1 \ldots ncombs$ do
            Set the p-th parameters combination to the model.
            Split the training set into k subsets.
            for $i = 1 \ldots k$ do
                Fit the $model$ to the data with the folds different than i.
                Predict the outputs for the i-th fold.
                $Score_i \leftarrow norm_\infty(y_{true}, y_{pred})$.
            end for
            $Score_p \leftarrow \frac{\sum_{i=1}^{k} Score_i}{k}$.
            if $best\_error_{seed} \geq Score_p$ then
                Update $best\_error_{seed}$.
                Update $best\_params_{seed}$.
            end if
        end for
        Set the $best\_params_{seed}$ to the model.
        Fit the $model$ to the training set.
        Predict the outputs for the test set inputs.
        if $best\_error \geq norm_\infty(y_{true},\ y_{pred})$ then
            Update best_error.
            Update best_params.
            Update best_seed.
        end if
    end for
    return $best\_params, best\_seed$
end procedure
```

4
Results and discussions

4.1
Complete generic design workflow to develop regression models based on machine learning techniques

As discussed in Chapter 1, only [10] used AI techniques to improve DGS' geometric characteristics, but there are no studies focused on using machine learning techniques to predict the reliability of DGS' systems. Moreover, the other studies that aimed to create predictive models from simulation data were unclear about the project pipeline to be followed. Therefore, a design workflow is proposed for future research whose objective is to create a regression model based on machine learning techniques where the dataset comes from computer simulations, and there is no prior knowledge about the most appropriate regression method. Thus, based on what was discussed throughout Chapter 3, the complete generic workflow (Figure 4.1) that is used in the case studies in Sessions 4.2 and 4.3 is presented.

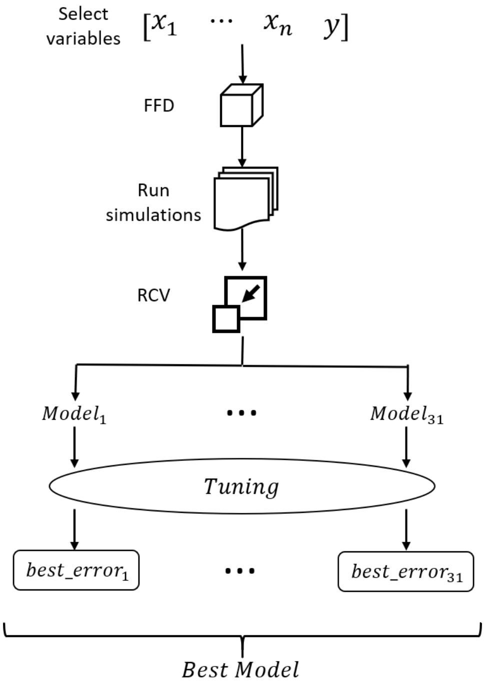


Figure 4.1: Generic design workflow for developing ML regression models.

The proposed methodology consists of:

1. Define the variables of interest used as input in the predictive model and output to be evaluated. It is important to note that this procedure considers that all variables are continuous.

2. For each input variable, the minimum and maximum boundaries are defined. Thus, this is the domain in which the regression model is applicable.

3. Perform the full factorial design (FFD) to define the combinations between the variables that are studied to compose the dataset. In this case, it is recommended to perform the 3-level FFD so that the model can better predict the physical behavior between the domain extremes.

4. Conduct the necessary experiments to obtain the required data to the stipulated conditions. In the study, CFD simulations are performed.

5. With the complete database, that is, with the values assigned to each of the variables in a simulation and their responses, it is necessary to convert the input variables into RCV.

6. Carry out the tuning process presented in Algorithm 1 for different regression models. In this study, 31 different regression models presented in Table 4.1 are tested, performing the tuning step of their hyper-parameters from the intervals presented in Appendix B. The code developed for this application is available in a repository on Github (`https://github.com/matheus-hoffmann/skl_regressor_test`).

7. With the best combination of hyper-parameters for each tested model, it is possible to evaluate their performance for different metrics presented in Chapter 3.

After performing this procedure, it is possible to compare the models based on objective information. According to experimental studies of DGSs presented in [12, 21], it is recommended to use chromel-alumel thermocouples (type K) for temperature measurement. Based on the maximum tolerance provided by this type of sensor, it is defined that the highest accepted infinity norm is 2.5 °C [90], in order to provide a safe confidence interval for the operation of the equipment. Furthermore, as the purpose of the application is real-time prediction, the threshold value to perform a prediction is defined as the human perception delay, 50 milliseconds, as presented in [91].

Table 4.1: Regression models implemented in Scikit-Learn [87] that will be tested. There is the abreviation adopted for each method, the reference where they were first presented and the type of model among those characterized in Section 3.4.

Method	Abreviation	Reference	Type
AdaBoostRegressor	ABR	[81]	Ensemble
BaggingRegressor	BgR	[82]	Ensemble
BayesianRidge	ByR	[65]	Linear
DecisionTreeRegressor	DTR	[79]	Tree
ElasticNet	ElN	[66]	Linear
ElasticNetCV	ElNCV	[66]	Linear
ExtraTreeRegressor	ETR	[80]	Tree
ExtraTreesRegressor	ETsR	[80]	Ensemble
GradientBoostingRegressor	GBR	[83]	Ensemble
HuberRegressor	HuR	[67]	Linear
KNeighborsRegressor	KNR	[76]	Neighbors
Lars	Lar	[68]	Linear
LarsCV	LarCV	[68]	Linear
Lasso	Las	[69]	Linear
LassoCV	LasCV	[69]	Linear
LassoLars	LsLr	[70]	Linear
LassoLarsCV	LsLrCV	[70]	Linear
LassoLarsIC	LsLrIC	[70]	Linear
LinearRegression	Lin	[64]	Linear
LinearSVR	LSVR	[77]	SVM
NuSVR	NSVR	[78]	SVM
OrthogonalMatchingPursuit	OMP	[71]	Linear
OrthogonalMatchingPursuitCV	OMPCV	[71]	Linear
PassiveAggressiveRegressor	PAR	[72]	Linear
RANSACRegressor	RAN	[73]	Linear
RandomForestRegressor	RFR	[84]	Ensemble
Ridge	Rid	[74]	Linear
RidgeCV	RidCV	[74]	Linear
SGDRegressor	SGD	[75]	Linear
SVR	SVR	[77]	SVM
XGBRegressor	XGB	[85]	Ensemble

4.2 Predicting dry gas seals reliability with machine learning techniques for a fixed mixture composition

4.2.1 Problem description

This first case study aims to develop a predictive machine learning model that estimates the safety margin for a T groove (geometry described at Table 2.7), a fixed composition and different operating conditions. The mixture is composed by 80 % methane, 19 % ethane, and 1 % octane. This composition is prone to be critical because, due to the large percentage of octane, the super-heated steam curve tends to have higher temperature values, requiring the model to be robust for a specific critical condition, the interior of the saturation curve. A 3-level FFD was performed to combine the different operating conditions, where the levels of each parameter are displayed in Table 4.2. Thus, to create the dataset to be used by the ML model, it is necessary to run 81 CFD simulations. It is worth highlighting that the system outlet pressure is atmospheric for all cases, approximately 101300 Pa, as exposed at [5].

Table 4.2: Levels to be evaluated in each input variable assuming a fixed gas mixture composition.

Operational condition	lower bound	midpoint	upper bound
Clearance $[h]$	3 μm	4 μm	5 μm
Inlet temperature $[T_{in}]$	293 K	358 K	423 K
Angular velocity $[\omega]$	10000 RPM	15000 RPM	20000 RPM
Inlet pressure $[P_{in}]$	1 MPa	5.5 MPa	10 MPa

4.2.2 Results

After the execution of 81 CFD simulations, the set of variables used for the elaboration of the ML model was converted into reduced centralized variables, giving all operating conditions minimum and maximum values of -1 and 1, respectively. Thus, there is an increase in the numerical stability of the model and possibly a reduction in the algorithm training time.

In the tuning step, the Algorithm 1 was performed for all models described in Table 4.1. Table 4.3 shows the results for different performance metrics of the best combination of hyper-parameters obtained for each regression model. In this one, the results are ordered according to the lowest infinity norm values, considering that this was the criterion adopted to evaluate the model's robustness. It is worth pointing that linear and neighbor-based models are not suitable for making accurate predictions. On the other hand, those based on SVM and trees, either a single one or ensembles using multiple trees, proved to be more robust and capable compared to the others. All models' R^2 and R_a^2 scores are high, demonstrating that they are insufficient criteria for choosing the most appropriate model. Furthermore, the MAE and the RMSE follow the infinity norm tendency, however, due to the high amount of low absolute residuals, these metrics present lower values than the infinity norm. Finally, MAPE is more sensitive to smaller values, being an unreliable metric for decision making. Thus, it is proved that choosing the infinity norm as the reference metric guarantees operational reliability in realistic cases.

Table 4.3: Performance metrics ordered by $norm_\infty$ for the best model achieved by each method after the tuning step for a fixed composition dataset.

Method	Type	$norm_\infty$ [°C]	R^2 [-]	R_a^2 [-]	MAE [°C]	RMSE [°C]	MAPE [%]
NuSVR	SVM	1.872	0.9997	0.9997	0.590	0.763	43.6
ExtraTreeRegressor	Tree	2.422	0.9992	0.9989	1.020	1.335	10.9
SVR	SVM	2.540	0.9993	0.9991	0.939	1.211	71.3
GradientBoostingRegressor	Ensemble	2.566	0.9995	0.9993	1.082	1.350	31.6
ExtraTreesRegressor	Ensemble	2.622	0.9994	0.9992	0.943	1.196	15.6
XGBRegressor	Ensemble	2.691	0.9994	0.9992	1.157	1.344	14.2
DecisionTreeRegressor	Tree	2.812	0.9992	0.9989	1.091	1.407	11.3
RandomForestRegressor	Ensemble	2.943	0.9989	0.9985	1.426	1.651	25.8
BaggingRegressor	Ensemble	3.144	0.9990	0.9987	1.565	1.857	9.4
AdaBoostRegressor	Ensemble	3.857	0.9985	0.9980	1.879	2.259	11.5
LassoCV	Linear	7.041	0.9901	0.9868	4.574	5.000	206.9
ElasticNet	Linear	7.066	0.9899	0.9866	4.605	5.032	211.2
Lasso	Linear	7.066	0.9899	0.9866	4.605	5.032	211.2
LassoLars	Linear	7.066	0.9899	0.9866	4.605	5.032	211.2
OrthogonalMatchingPursuitCV	Linear	7.088	0.9903	0.9871	4.456	4.931	202.8
LassoLarsIC	Linear	7.089	0.9905	0.9874	4.526	4.881	173.0
ElasticNetCV	Linear	7.244	0.9905	0.9874	4.515	4.886	198.8
SGDRegressor	Linear	7.269	0.9897	0.9862	4.673	5.105	215.5
RidgeCV	Linear	7.375	0.9897	0.9863	4.656	5.083	219.6
Ridge	Linear	7.376	0.9897	0.9863	4.656	5.083	219.6
BayesianRidge	Linear	7.405	0.9897	0.9862	4.666	5.100	220.5
LinearRegression	Linear	7.443	0.9896	0.9861	4.680	5.121	221.5
LassoLarsCV	Linear	7.443	0.9896	0.9861	4.680	5.121	221.5
LarsCV	Linear	7.443	0.9896	0.9861	4.680	5.121	221.5
RANSACRegressor	Linear	7.443	0.9896	0.9861	4.680	5.121	221.5
Lars	Linear	7.443	0.9896	0.9861	4.680	5.121	221.5
HuberRegressor	Linear	7.500	0.9896	0.9861	4.689	5.131	222.6
LinearSVR	SVM	7.531	0.9850	0.9800	5.372	5.693	139.1
PassiveAggressiveRegressor	Linear	8.549	0.9900	0.9867	5.175	5.851	146.8
OrthogonalMatchingPursuit	Linear	15.926	0.9603	0.9470	10.220	10.913	184.7
KNeighborsRegressor	Neighbors	18.872	0.9570	0.9427	9.480	11.171	80.3

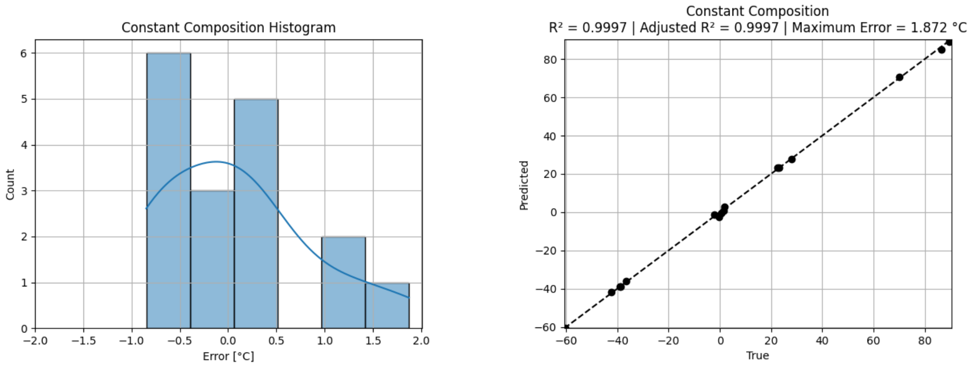
The regression model that obtained the lowest infinite norm was the NuSVR [78]. In this one, the combination of hyper-parameters that provided this performance is presented in Table 4.4, where its parameters represent:

- ν: An upper bound on the fraction of training errors and a lower bound of the fraction of support vectors.
- Kernel: Type of kernel to be used in the algorithm.
- γ: Kernel coefficient.
- C: Penalty parameter of the error term.

Table 4.4: Best NuSVR hyper-parameters for a fixed composition dataset.

Parameter	Value
ν	1
Kernel	RBF
γ	0.01
C	30000

With the model presented above, it was possible to predict the output for the test input data set (y_{pred}) and compare it with the values obtained through numerical simulation (y_{true}). In the regression diagram in Figure 4.2 (b), it is possible to verify that the points are close to the 0 error dotted line, which is the first indicator that the obtained model meets the objective. Furthermore, from the histogram in Figure 4.2 (a), it can be seen that the points with the most significant absolute errors are in conditions where the safety margin is underestimated, not compromising the integrity of the operation. In addition, it is possible to perceive a greater density of points at the origin from the histogram, representing low errors. This information is corroborated by evaluating both the mean of these residuals in Table 4.5. Also, the range of values between 25% and 75 % shows that at least 50 % of predictions have errors of magnitude below 1 °C.



(a)

(b)

Figure 4.2: NuSVR best configuration predictions for a fixed composition. (a) Histogram of the error of the test set. (b) Regression diagram of the test set.

Table 4.5: Best model prediction statistics for a fixed composition.

model	mean	std	min	25%	50%	75%	max	time
NSVR	0.069	0.783	-0.848	-0.627	0.001	0.463	1.872	70.1 μs

4.3
Predicting dry gas seals reliability with machine learning techniques for a variable mixture composition

4.3.1
Problem description

The second case study aims to develop a predictive machine learning model that estimates the safety margin for a T groove (geometry described at Table 2.7) when both the mixture composition and the operating conditions are variable. For consistency when it comes to the composition of the mixture, it is assumed that

$$[\%]_{C_2H_6} = 1 - ([\%]_{CH_4} + [\%]_{C_8H_{18}}), \tag{4-1}$$

where $[\%]_{C_2H_6}$, $[\%]_{CH_4}$ and $[\%]_{C_8H_{18}}$ correspond respectively to the molar concentrations of ethane, methane and octane. Thus, the combinations of concentrations of $[\%]_{CH_4}$ and $[\%]_{C_8H_{18}}$ in the experiments are varied, and $[\%]_{C_2H_6}$ is written as the result of Equation 4-1. The octane concentration is the parameter used to increase the super-heated steam curve temperature as exemplified in Figure 2.12, and the other operating conditions are presented in Table 4.6. It is worth highlighting that the system outlet pressure is atmospheric for all cases, approximately 101300 Pa, as exposed at [5]. Thus, a 3-level FFD was performed to create the dataset used in the ML model, requiring 729 CFD simulations.

Table 4.6: Levels to be evaluated in each input variable assuming a variable gas mixture composition.

Operational condition	lower bound	midpoint	upper bound
Methane [CH_4]	70 %	75 %	80 %
Octane [C_8H_{18}]	0 %	0.5 %	1 %
Clearance [h]	3 μm	4 μm	5 μm
Inlet temperature [T_{in}]	293 K	358 K	423 K
Angular velocity [ω]	10000 RPM	15000 RPM	20000 RPM
Inlet pressure [P_{in}]	1 MPa	5.5 MPa	10 MPa

4.3.2 Results

After performing the 729 CFD simulations, the set of variables used for the elaboration of the ML model was converted into RCV. Then, in the tuning step, the Algorithm 1 was executed for all models described in Table 4.1. Table 4.7 shows the results for different performance metrics of the best combination of hyper-parameters obtained for each regression model. The results are sorted according to the same criteria presented in Session 4.2. It is worth highlighting that considering the effects of the mixture composition on the model, greater robustness would be needed to represent these nonlinearities, as expected. Thus, it is evident that linear models continue to be the least indicated in this situation, and those based on SVM and trees showed better predictions than the other types. As in the previous case study, the MAE and the RMSE follow the infinity norm tendency, but the R^2 and R_a^2 scores have lower values as the models' infinity norm increases, evidencing the need for greater robustness in predictive models.

Table 4.7: Performance metrics ordered by $norm_\infty$ for the best model achieved by each method after the tuning step for a variable composition dataset.

Method	Type	$norm_\infty$ [°C]	R^2 [-]	R_a^2 [-]	MAE [°C]	RMSE [°C]	MAPE [%]
GradientBoostingRegressor	Ensemble	6.976	0.9995	0.9995	1.194	1.789	6.8
ExtraTreesRegressor	Ensemble	7.587	0.9989	0.9989	1.821	2.562	5.8
SVR	SVM	7.673	0.9993	0.9992	1.333	1.994	16.3
ExtraTreeRegressor	Tree	8.079	0.9991	0.9990	1.716	2.456	6.9
RandomForestRegressor	Ensemble	8.107	0.9986	0.9985	2.251	2.823	14.9
BaggingRegressor	Ensemble	8.796	0.9992	0.9991	1.496	2.159	12.6
DecisionTreeRegressor	Tree	9.216	0.9986	0.9986	1.993	2.700	6.4
XGBRegressor	Ensemble	9.279	0.9991	0.9990	1.450	2.276	9.3
NuSVR	SVM	9.941	0.9992	0.9992	1.380	2.210	4.7
AdaBoostRegressor	Ensemble	25.848	0.9714	0.9702	11.944	13.867	116.6
RANSACRegressor	Linear	32.855	0.9463	0.9440	17.298	19.256	264.4
Ridge	Linear	33.100	0.9461	0.9438	17.337	19.289	267.2
LassoCV	Linear	33.107	0.9461	0.9438	17.347	19.288	266.2
ElasticNetCV	Linear	33.107	0.9461	0.9438	17.347	19.288	266.2
LarsCV	Linear	33.108	0.9461	0.9438	17.343	19.284	267.4
LassoLarsCV	Linear	33.108	0.9461	0.9438	17.343	19.284	267.4
Lasso	Linear	33.109	0.9461	0.9438	17.338	19.279	268.8
LassoLars	Linear	33.109	0.9461	0.9438	17.338	19.279	268.8
ElasticNet	Linear	33.109	0.9462	0.9438	17.336	19.277	269.4
LassoLarsIC	Linear	33.109	0.9462	0.9438	17.335	19.277	269.6
LinearRegression	Linear	33.109	0.9462	0.9438	17.335	19.277	269.6
Lars	Linear	33.109	0.9462	0.9438	17.335	19.277	269.6
BayesianRidge	Linear	33.114	0.9461	0.9438	17.336	19.281	268.8
OrthogonalMatchingPursuitCV	Linear	33.280	0.9324	0.9295	17.458	19.582	196.9
SGDRegressor	Linear	33.390	0.9457	0.9433	17.406	19.361	264.6
LinearSVR	SVM	33.429	0.9478	0.9456	16.789	18.979	253.4
HuberRegressor	Linear	33.730	0.9480	0.9458	16.688	18.942	248.8
RidgeCV	Linear	33.785	0.9480	0.9457	16.698	18.945	250.0
KNeighborsRegressor	Neighbors	35.352	0.9548	0.9528	11.502	14.042	44.9
PassiveAggressiveRegressor	Linear	38.991	0.9467	0.9444	15.343	19.184	237.1
OrthogonalMatchingPursuit	Linear	99.681	0.5722	0.5537	46.895	51.506	602.1

The regression model that obtained the lowest infinite norm was the Gradient Boosting Regressor [83]. In this one, the combination of hyper-parameters that provided this performance is presented in Table 4.8, where its parameters represent:

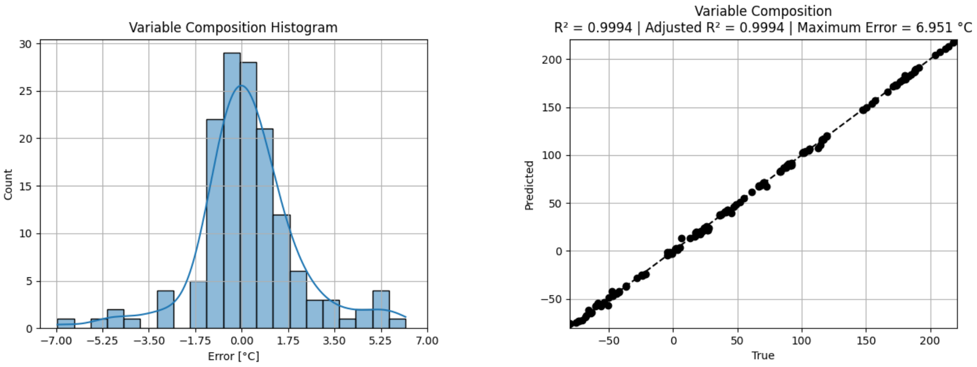
- Loss: loss function to be optimized.
- Maximum features: number of features to consider when looking for the best split.
- Learning rate: shrinks the contribution of each tree.
- Min. number of samples at a leaf: minimum number of samples required to be at a leaf node.
- Max. number of estimators: number of boosting stages to perform.

- Percentage of samples: fraction of samples to be used for fitting the individual base learners.
- Criterion: function to measure the quality of a split.
- Min. number of samples to split node: minimum number of samples required to split an internal node.

Table 4.8: Best Gradient Boosting hyper-parameters for a variable composition dataset.

Parameter	Value
Loss	'lad'
Maximum features	Auto
Learning rate	0.1
Min. number of samples at a leaf	10
Max. number of estimators	2000
Percentage of samples	1
Criterion	Friedman MSE
Min. number of samples to split node	30

From the model presented above, it was possible to compare the forecast and the actual value of the test set's outputs. In the regression diagram in Figure 4.3 (b), it is possible to verify that most of the points are close to the zero error line. However, as the values of y_{true} decrease, it is noted that some points are further away from this reference line. From the histogram in Figure 4.3 (a), it can be seen that, despite a high density of points near the origin, there is also a considerable amount further away on both sides of the histogram. This information is corroborated when evaluating the Table 4.9, where it is possible to see that although 50 % of the points have errors of magnitude below 1 °C, the minimum and maximum values obtained are considerably above the defined as a limit for safe and reliable operation.



(a)

(b)

Figure 4.3: Gradient Boosting Regressor best configuration predictions for a variable composition. (a) Histogram of the error of the test set. (b) Regression diagram of the test set.

Table 4.9: Best model prediction statistics for a variable composition.

model	mean	std	min	25%	50%	75%	max	time
GBR	0.270	1.927	-6.951	-0.610	0.122	1.062	6.180	362.8 μs

4.4 Applicability discussion

It took approximately 972 hours of simulation to create the dataset for the first case study (Section 4.2) and approximately 8748 hours for the second case study (Section 4.3), totaling approximately 41 and 365 days, respectively. However, using a developed execution queue manager, it was possible to run up to 10 simulations simultaneously, reducing the total time required to create the datasets to 23 and 207 days, respectively. Although this step of creating the dataset precedes the operation itself, where the focus is on performance, the time required to create the dataset considering a composition that varies over time is very high using the computational resources used in this research. A possible solution would be to increase the computational resources or evaluate the applicability of other DoE techniques to define a smaller number of simulations to be performed.

The tuning step took about 22 hours for the first case study and 72 hours for the second. This step considers the time required to obtain the best combination of hyper-parameters for each of the 31 regression models. As it is a step that also precedes the operation, these are added to the time needed to create the dataset, totaling 24 and 210 days, respectively, for each case study. In order to predict a safety margin for certain operational conditions, 70.1 and 362.8 μs respectively are needed for each case study. The time required for both estimates satisfies the concept of real-time defined previously, fulfilling the stipulated project requirement. Information regarding the time required in each step is presented in Table 4.10.

Table 4.10: Description of the time required to perform each step from model creation to prediction. (Section 4.2) Fixed composition dataset. (Section 4.3) Variable composition dataset.

Procedure	Section 4.2	Section 4.3
CFD simulations	23 days	207 days
Tuning	22 hours	72 hours
Prediction	70.1 μs	362.8 μs

Furthermore, in the case study presented in Section 4.2, from Table 4.5 the infinity norm of residuals is 1.872 °C, satisfying the stipulated maximum of 2.5 °C design requirement. On the other hand, in the case study presented in Section 4.3, from Table 4.9 the design requirement was not met, with a value of 6.951 °C. Thus, it is concluded that it is possible to make reliable predictions of the safety margin in operation from the fixed composition hypothesis. However, it was impossible to find a robust model when considering the variable composition hypothesis. There is the possibility of improving the model's performance that assumes a variable composition by increasing the number of CFD simulations in the dataset, but this is outside the scope proposed in the study. Figure 4.4 compares the smallest infinite norm of each regressive model, and shows the discrepancy between the performances for the two evaluated hypotheses. It is evident that incorporating the gas composition in the predictive model analysis is necessary to increase the robustness of the techniques used.

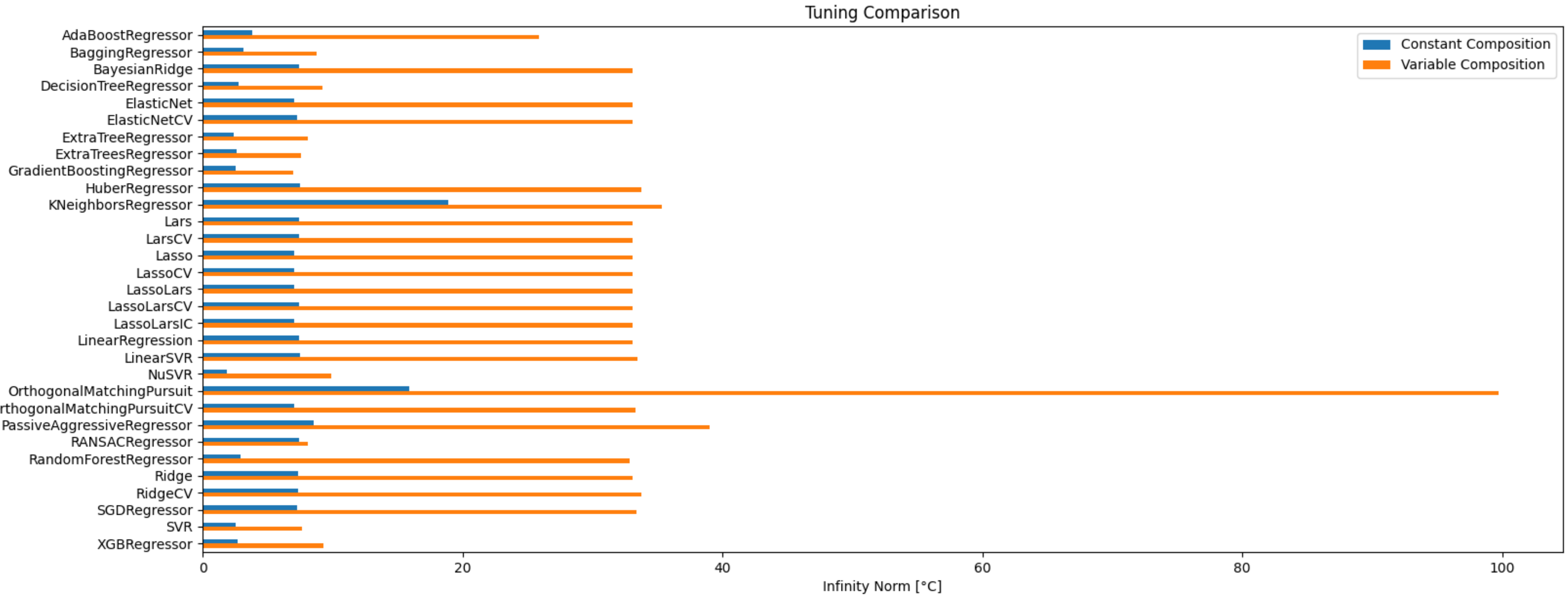


Figure 4.4: Comparison of the $norm_\infty$ obtained after the tuning process for all methods under the two hypotheses about the sealing gas composition: fixed (blue) and variable (orange).

5
Conclusion

This work evaluated the possibility of replacing CFD simulations in the operational reliability verification of dry gas seals with a machine learning regression model developed from a generic design workflow. First, a simulation solver was developed coupling OpenFOAM with thermodynamic tables generated from data available in REFPROP. In this one, the boundary conditions, initialization of internal fields of the domain, and the adequate mesh were adjusted to guarantee a CFD model that reproduced the study presented in [11]. The obtained model showed mean errors between the pressure profiles along the radius of 3.4, 2.1, and 3.3 % for clearances of 2.03, 3.05, and 5.08 μm respectively, thus proving the consistency of the solution.

Then, a generic design workflow was proposed to create regression models using ML techniques from numerical simulations. This workflow covered from the definition of the simulations to be performed based on the variables of interest to the choice of the best model to be used. This method was used in the two case studies on the reliability of dry gas seals. In the first case study, it was assumed that the composition of the gas mixture flowing through the DGS is constant. In this one, it was possible to obtain a NuSVR model with a maximum absolute error of 1.872 °C, below the maximum stipulated as a safety design requirement. The second case study assumed that the composition of the gas mixture could vary. Under these conditions, the model that best fitted was the Gradient Boosting with a maximum absolute error of 6.951 °C. However, this error is higher than the maximum permissible error, leading to the conclusion that conventional techniques were not enough to create a robust model under the assumption of variable gas composition.

Finally, the time required to create the datasets and perform the tuning of the hyper-parameters for each of the case studies was evaluated. It took 24 and 210 days to carry out these steps in the case studies 1 and 2, respectively. In addition, the forecast time for both was lower than the threshold necessary to be considered a real-time forecast, proving the applicability in operational conditions.

Some suggestions are presented for future research:

- **Test other DoE techniques for creating the dataset.** One of the most basic techniques for creating the dataset, the 3-level FFD, was used, but it is possible to define the CFD simulations to be performed using other techniques to increase the representativeness of the data and get more robust models. It is recommended to consider the central composite design [61], Box-Behnken [92] and Doehlert [93] designs. However, not getting considerable improvements in performing predictions, other DoE techniques could be used in order to reduce the number of CFD simulations needed without compromising the quality of the models. To this end, it is recommended to evaluate the D-Optimal [94] and Uniform [95] (developed for numerical simulations) designs.

- **Evaluate the use of the Leave One Out Cross-Validation (LOOCV) technique.** LOOCV [96] is a specific case of K-Fold CV (Section 3.6) when K is equal to the total number of samples n. Hence, the dataset is divided into n parts and $n - 1$ data is used for training the model and one for validation, this process being repeated n times. This procedure is recommended only for small datasets, with a maximum of 1000 samples, because of the increased computational cost required. Thus, with the use of LOOCV, it is possible to obtain a model with better precision due to the greater number of tests performed, but it can also imply a considerable increase in the execution time of the tuning step.
- **Investigate more robust regression techniques.** These techniques can improve the SM prediction of models that consider a variable mixture composition, considering that the conventional models were not enough. As a first approach, it may be interesting to create ensemble methods with combinations of weak learners not based on trees, given that these have already been evaluated in the present study. In addition, the use of neural networks may be necessary for a better representation of the phenomena, as presented in the studies [32–34, 37, 38, 40–48].
- **Optimize the operating conditions of the DGS.** It can optimize the operating conditions to reduce the operational cost of the compressor using the proposed predictive model. An example would be the decrease in the inlet temperature, still guaranteeing the 20 °C safety margin required in [14], culminating in a reduction in the energy cost of the heater that precedes the DGS.
- **Increase the number of components in the mixture composition.** This modification aims to assess the influence of other substances on the prediction of the safety margin. Including more components in the

mixture directly implies an increase in the number of simulations to be performed if the 3-level FFD is maintained, however, to circumvent this, a fractional factorial design [61] could be used to reduce the number of simulations. Including carbon dioxide (CO_2) and propane (C_3H_8) is recommended, as their concentrations tend to increase over time.

Appendices

A
Model construction algorithms

The algorithms presented here are based on the Scikit-Learn documentation [87] and were described in Chapter 3. The Algorithm 1 used in the study is based on the following procedures:

- **Algorithm 2**: Holdout algorithm described in Section 3.6.
- **Algorithm 3**: K-fold cross-validation algorithm described in Section 3.6.
- **Algorithm 4**: Random-search algorithm described in Section 3.7.

Algorithm 2 Holdout technique.

```
$x \leftarrow input$
$y \leftarrow output$
$[\%]_{train} \leftarrow train\ percentage$
$seed \leftarrow seed$
$model \leftarrow desired\ model$
procedure HOLDOUT($seed$)
    Split the dataset into training and test set for the given seed.
    Fit the model to the training set.
    Predict the outputs for the test set inputs.
    return $norm_{\infty}(y_{true},\ y_{pred})$
end procedure
```

Algorithm 3 K-fold cross-validation

```
$x \leftarrow input$
$y \leftarrow output$
$model \leftarrow regression\ model$
$k \leftarrow number\ of\ subsets$
procedure KFOLDCV($k$)
    Split the dataset into k subsets.
    for $i = 1 \dots k$ do
        Fit the model to the data with the folds different than i.
        Predict the outputs for the i-th fold.
        $Score_i \leftarrow norm_\infty(y_{true}, y_{pred})$.
    end for
    return $\frac{\sum_{i=1}^{k} Score_i}{k}$
end procedure
```

Algorithm 4 Random search

```
$x \leftarrow input$
$y \leftarrow output$
$model \leftarrow regression\ model$
$params \leftarrow hyper - parameters$
$ncombs \leftarrow number\ of\ combinations$
procedure RANDOM_SEARCH($params$, $ncombs$)
    Split the dataset into training and test set.
    Create the complete combinations of parameters grid.
    Randomly select ncombs parameters combinations from the original grid.
    for $p = 1 \dots ncombs$ do
        Set the p-th parameters combination to the model.
        Fit the model to the training set.
        Predict the outputs for the test set inputs.
        if $best\_error \geq norm_\infty(y_{true},\ y_{pred})$ then
            Update best_error.
            Update best_params.
        end if
    end for
    return best_params
end procedure
```

B
Models valid hyper-parameters

Table B.1 presents, for each regression model to be evaluated, the ranges of values of possible hyper-parameters. From these, the combinations will be performed as presented in Section 3.8 to tune the hyper-parameters. The columns in this table correspond to:

- Model: Regression model.
- Hyper-parameter: model parameter to be evaluated.
- Distribution/option: possible values for the evaluated hyper-parameter.

Model	Hyper-parameter	Distribution/option
AdaBoostRegressor	Learning rate	Log-uniform $[10^{-1}, 10^{3}]$
	Loss function	Linear, square, exponential
	Max. number of estimators	Uniform integer $\{50, 100, ..., 1000\}$
BaggingRegressor	Percentage of samples	Log-uniform $[10^{-1}, 10^{0}]$
	Max. number of estimators	Uniform integer $\{50, 100, ..., 1000\}$
BayesianRidge	α_1	Log-uniform $[10^{-9}, 10^{2}]$
	α_2	Log-uniform $[10^{-9}, 10^{2}]$
	λ_1	Log-uniform $[10^{-9}, 10^{2}]$
	λ_2	Log-uniform $[10^{-9}, 10^{2}]$
	Max. number of iterations	Uniform integer $\{50, 100, ..., 10^{5}\}$
	Compute intercept?	True, False
	Normalize data?	True, False
DecisionTreeRegressor	Criterion	MSE, Friedman MSE, MAE
	Max. number of features	Auto, sqrt, log2
	Min. number of samples at a leaf	[1, 5, 10, 15, 20]
	Min. number of samples to split node	[2, 11, 21, 30, 40]
ElasticNet	α	Log-uniform $[10^{-3}, 10^{3}]$
	Compute intercept?	True, False
	Penalty combination of L1 and L2	[0, 0.25, 0.5, 0.75, 1]
	Normalize data?	True, False
ElasticNetCV	Compute intercept?	True, False

Model	Hyper-parameter	Distribution/option
	Penalty combination of L1 and L2	[0, 0.25, 0.5, 0.75, 1]
	Normalize data?	True, False
ExtraTreeRegressor	Criterion	MSE, Friedman MSE, MAE
	Maximum features	Auto, sqrt, log2
	Min. number of samples at a leaf	[1, 5, 10, 15, 20]
	Min. number of samples to split node	[2, 11, 21, 30, 40]
ExtraTreesRegressor	Criterion	MSE, MAE
	Maximum features	Auto, sqrt, log2
	Min. number of samples at a leaf	[1, 5, 10, 15, 20]
	Min. number of samples to split node	[2, 11, 21, 30, 40]
	Max. number of estimators	Uniform integer {50, 100, ..., 1000}
GradientBoostingRegressor	Criterion	MSE, Friedman MSE
	Learning rate	Log-uniform $[10^{-3}, 10^{0}]$
	Loss function	Huber, ls, lad, quantile
	Maximum features	Auto, sqrt, log2
	Min. number of samples at a leaf	[1, 5, 10, 15, 20]
	Min. number of samples to split node	[2, 11, 21, 30, 40]
	Max. number of estimators	Uniform integer {50, 100, ..., 1000}
	Percentage of samples	Log-uniform $[10^{-3}, 10^{0}]$
HuberRegressor	α	Log-uniform $[10^{-4}, 10^{1}]$
	ϵ	Log-uniform $[10^{0}, 10^{1}]$

Model	Hyper-parameter	Distribution/option
	Max. number of iteration	[10, 100, 500, 1000]
KNeighborsRegressor	Algorithm	Auto, Ball Tree, KD Tree, Brute force
	Distance metric	Euclidean, Manhattan, Chebyshev
	Number of neighbors	[2, 3, 4, 5, 6]
	Weights	Uniform, distance
Lars	Compute intercept?	True, False
	Normalize data?	True, False
LarsCV	Compute intercept?	True, False
	Normalize data?	True, False
Lasso	α	Log-uniform $[10^{-3}, 10^{3}]$
	Compute intercept?	True, False
	Normalize data?	True, False
LassoCV	Compute intercept?	True, False
	Normalize data?	True, False
LassoLars	α	Log-uniform $[10^{-3}, 10^{3}]$
	Compute intercept?	True, False
	Normalize data?	True, False
LassoLarsCV	Compute intercept?	True, False
	Normalize data?	True, False
LassoLarsIC	Criterion	bic, aic
	Compute intercept?	True, False

Model	Hyper-parameter	Distribution/option
	Normalize data?	True, False
LinearRegression	Compute intercept?	True, False
	Normalize data?	True, False
	Force positive coefficients?	True, False
LinearSVR	C	Log-uniform $[10^{-2}, 10^{5}]$
	ϵ	Log-uniform $[10^{-2}, 10^{0}]$
	Loss function	ϵ-insensitive, ϵ^2-insensitive
NuSVR	C	Log-uniform $[10^{-2}, 10^{5}]$
	γ	Log-uniform $[10^{-5}, 10^{4}]$
	Kernel	Linear, RBF, Sigmoid
	ν	Log-uniform $[10^{-2}, 10^{0}]$
OrthogonalMatchingPursuit	Compute intercept?	True, False
	Normalize data?	True, False
OrthogonalMatchingPursuitCV	Compute intercept?	True, False
	Normalize data?	True, False
PassiveAggressiveRegressor	C	Log-uniform $[10^{-2}, 10^{5}]$
	ϵ	Log-uniform $[10^{-2}, 10^{0}]$
	Compute intercept?	True, False
RandomForestRegressor	Criterion	MSE, MAE
	Maximum features	Auto, sqrt, log2
	Min. number of samples at a leaf	[1, 5, 10, 15, 20]

Model	Hyper-parameter	Distribution/option
	Min. number of samples to split node	[2, 11, 21, 30, 40]
	Max. number of estimators	Uniform integer {50, 100, ..., 1000}
RANSACRegressor	Loss function	absolute_loss, squared_loss
	Percentage of samples	[0.1, 0.5, 0.9]
Ridge	α	Log-uniform $[10^{-1}, 10^{3}]$
	Compute intercept?	True, False
	Normalize data?	True, False
	Solver	Auto, SVD, Cholesky, Least-squares, Sparse CG, Stochastic Average Gradient Descent
RidgeCV	Compute intercept?	True, False
	LOOCV strategy	Auto, SVD, Eigen
	Normalize data?	True, False
SGDRegressor	α	Log-uniform $[10^{-3}, 10^{3}]$
	Compute intercept?	True, False
	Penalty combination of L1 and L2	[0, 0.25, 0.5, 0.75, 1]
	Learning rate	Constant, Optimal, Inverse Scaling, Adaptive
	Loss function	Squared Loss, Huber, ϵ-insensitive, ϵ^2-insensitive

Model	Hyper-parameter	Distribution/option
	Penalty	L1, L2, ElasticNET
SVR	C	Log-uniform $[10^{-2}, 10^{5}]$
	ϵ	Log-uniform $[10^{-2}, 10^{0}]$
	γ	Log-uniform $[10^{-5}, 10^{4}]$
	Kernel	Linear, RBF, Sigmoid
XGBRegressor	Booster technique	gbtree, gblinear, dart
	Learning rate	Log-uniform $[10^{-3}, 10^{0}]$
	Maximum tree depth	[9, 10, 11, 12]
	Min. sum of weights in a node	[5, 6, 7, 8]
	Percentage of samples	[0.7, 0.8, 0.9, 1.0]

Table B.1: Hyper-parameter settings for tuning the models in the present work using random search k-fold cross-validation holdout.

Bibliography

[1] INKPENY, A. C.; MOFFETT, M. H.. **The Global Oil Gas Industry: Management, Strategy and Finance**. PennWell Books, Tulsa, Oklahoma, 1 edition, 2011.

[2] DAY, M.; ALLISON, T.. **Analysis of historical dry gas seal failure data**. Volume 9: Oil and Gas Applications; Supercritical CO2 Power Cycles; Wind Energy, 06 2016.

[3] BRAVO, C.; SAPUTELLI, L.; RIVAS, F.; PÉREZ, A. G.; NIKOLAOU, M.; ZANGL, G.; DE GUZMÁN, N.; MOHAGHEGH, S. ; NUNEZ, G.. **State of the art of artificial intelligence and predictive analytics in the e&p industry: A technology survey**. SPE Journal, 19(04):547–563, 06 2013.

[4] KOROTEEV, D.; TEKIC, Z.. **Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future**. Energy and AI, 3:100041, 2021.

[5] STAHLEY, J.. **Dry Gas Seals Handbook**. PennWell Corporation, Tulsa, Oklahoma, 1 edition, 2005.

[6] BROWN, R.. **Compressors: Selection and Sizing**. Ann Arbor, Michigan, 2 edition, 1997.

[7] MASSALA, D. C.. **Ram analysis applied to centrifugal gas compressors: "case study of an oil and gas company"**. Masters thesis, Universidade de Lisboa, Lisboa, Portugal, 2018.

[8] SAHADEVAN, R.; MENHALI, B. A.. **Enhancement of Dry Gas Seal Reliability**. Day 3 Wed, November 14, 2018, 11 2018. D032S177R002.

[9] FAIRUZ, Z. M.; JAHN, I.. **The influence of real gas effects on the performance of supercritical co2 dry gas seals**. Tribology International, 102:333–347, 2016.

[10] OJILE, J. O.. **Numerical Modelling of Bidirectional Dry Gas Face Seals**. Phd thesis, Cranfield University, 2009.

[11] GABRIEL, R. P.. **Fundamentals of spiral groove noncontacting face seals**. Lubrication Engineering, 50(3):215–224, 1994.

[12] DING, X.; LU, J.. **Theoretical analysis and experiment on gas film temperature in a spiral groove dry gas seal under high speed and pressure**. International Journal of Heat and Mass Transfer, 96:438–450, 2016.

[13] SHAHIN, I.; GADALA, M.; ALQARADAWI, M. ; BADR, O.. **Centrifugal compressor spiral dry gas seal simulation working at reverse rotation**. Procedia Engineering, 68:285–292, 2013. INTERNATIONAL TRIBOLOGY CONFERENCE MALAYSIA 2013.

[14] American Petroleum Institute. **API Standard 614: Lubrication, Shaft-Sealing, and Control-Oil Systems and Auxiliaries for Petroleum, Chemical, and Gas Industry Services**, 5 edition, 2007.

[15] KASSAR, B.; MARQUES, R.; JUNIOR, H. ; BRITO, M.. **Improving operational equipment reliability with cfd analysis: case study of dry gas seal**. Rio Oil & Gas Expo and Conference, (187), 2020.

[16] American Petroleum Institute. **API Standard 617: Axial and Centrifugal Compressors and Expander-compressors**, 8 edition, 2014.

[17] LIU, Y.; SHEN, X.; XU, W. ; WANG, Z.. **Performance comparison and parametric study on spiral groove gas film face seals**. Science in China Series G: Physics, Mechanics and Astronomy, 2004.

[18] ZHENG, X.. **Parametrical study of hydrodynamic seal using a 2d design code and comparing with a 3d cfd model**. Volume 3: Turbo Expo 2005, Parts A and B:1173–1180, 06 2005.

[19] ZHOU, J.; GU, B. ; CHEN, Y.. **An improved design of spiral groove mechanical seal**. Chinese Journal of Chemical Engineering, 15(4):499–506, 2007.

[20] JING, X.; XUDONG, P.; SHAOXIAN, B. ; XIANGKAI, M.. **Cfd simulation of microscale flow field in spiral groove dry gas seal**. In: PROCEEDINGS OF 2012 IEEE/ASME 8TH IEEE/ASME INTERNATIONAL CONFERENCE ON MECHATRONIC AND EMBEDDED SYSTEMS AND APPLICATIONS, p. 211–217, 2012.

[21] KOLOMOETS, A.; DOTSENKO, V.. **Experimental investigation dry gas-dynamic seals used for gas-compressor unit**. Procedia Engineering, 39:379–386, 2012.

[22] SHAHIN, I.; GADALA, M.; ALQARADAWI, M. ; BADR, O.. **Three dimensional computational study for spiral dry gas seal with constant groove depth and different tapered grooves.** Procedia Engineering, 68:205–212, 2013. INTERNATIONAL TRIBOLOGY CONFERENCE MALAYSIA 2013.

[23] HONG, W.; JIANSHU, L. ; CHANGLIU, Y.. **A thermohydrodynamic analysis of dry gas seals for high-temperature gas-cooled reactor.** Journal of Tribology, 135, 04 2013.

[24] MA, C.; BAI, S. ; PENG, X.. **Thermoelastohydrodynamic characteristics of t-grooves gas face seals.** International Journal of Heat and Mass Transfer, 102:277–286, 2016.

[25] WANG, Y.; LU, L.; ZHANG, H. ; LYU, S.. **A simulation analysis and experimental research on t groove end face seal under mid-and-low speed.** International Journal of Precision Engineering and Manufacturing, 18, 04 2017.

[26] YAN, W.; QIONG, H.; JIANJUN, S.; DA, W. ; XIAOQING, Z.. **Numerical analysis of t-groove dry gas seal with orientation texture at the groove bottom.** Advances in Mechanical Engineering, 11(1):1687814018821775, 2019.

[27] DU, Q.; GAO, K.; ZHANG, D. ; XIE, Y.. **Effects of grooved ring rotation and working fluid on the performance of dry gas seal.** International Journal of Heat and Mass Transfer, 126:1323–1332, 2018.

[28] HALTON, J. H.. **On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals.** Numerische Mathematik, 2, 1960.

[29] LINSTROM, P.; MALLARD, W.. **NIST Chemistry WebBook, NIST Standard Reference Database Number 69.** National Institute of Standards and Technology, 2021.

[30] LEMMON, E. W.; BELL, I.; HUBER, M. L. ; MCLINDEN, M. O.. **NIST Standard Reference Database 23: Reference Fluid Thermodynamic and Transport Properties-REFPROP, Version 10.0, National Institute of Standards and Technology.** National Institute of Standards and Technology, 2018.

[31] CARRILLO, J. A. E.; DE LA FLOR, F. J. S. ; LISSÉN, J. M. S.. **Single-phase ejector geometry optimisation by means of a multi-objective evolutionary algorithm and a surrogate cfd model**. Energy, 164:46–64, 2018.

[32] GONG, Z.; WU, Y.; WU, L. ; SUN, H.. **Predicting thermodynamic properties of alkanes by high-throughput force field simulation and machine learning**. Journal of Chemical Information and Modeling, 58(12):2502–2516, 2018. PMID: 30205676.

[33] PALAGI, L.; PESYRIDIS, A.; SCIUBBA, E. ; TOCCI, L.. **Machine learning for the prediction of the dynamic behavior of a small scale orc system**. Energy, 166:72–82, 2019.

[34] XING, J.; WANG, H.; LUO, K.; WANG, S.; BAI, Y. ; FAN, J.. **Predictive single-step kinetic model of biomass devolatilization for cfd applications: A comparison study of empirical correlations (ec), artificial neural networks (ann) and random forest (rf)**. Renewable Energy, 136:104–114, 2019.

[35] LIU, Y.; HONG, W. ; CAO, B.. **Machine learning for predicting thermodynamic properties of pure fluids and their mixtures**. Energy, 188:116091, 2019.

[36] RAMIREZ, R.; AVILA, E.; LOPEZ, L.; BULA, A. ; DUARTE FORERO, J.. **Cfd characterization and optimization of the cavitation phenomenon in dredging centrifugal pumps**. Alexandria Engineering Journal, 59(1):291–309, 2020.

[37] ELMAZ, F.; ÖZGÜN YÜCEL ; MUTLU, A. Y.. **Predictive modeling of biomass gasification with machine learning-based regression methods**. Energy, 191:116541, 2020.

[38] BAKHTIARI, M.; GHASSEMI, H.. **Cfd data based neural network functions for predicting hydrodynamic performance of a low-pitch marine cycloidal propeller**. Applied Ocean Research, 94:101981, 2020.

[39] JALALIFAR, S.; MASOUDI, M.; ABBASSI, R.; GARANIYA, V.; GHIJI, M. ; SALEHI, F.. **A hybrid svr-pso model to predict a cfd-based optimised bubbling fluidised bed pyrolysis reactor**. Energy, 191:116414, 2020.

[40] PING, X.; YANG, F.; ZHANG, H.; ZHANG, J.; ZHANG, W. ; SONG, G.. **Introducing machine learning and hybrid algorithm for prediction and optimization of multistage centrifugal pump in an orc system**. Energy, 222:120007, 2021.

[41] MARCATO, A.; BOCCARDO, G. ; MARCHISIO, D.. **A computational workflow to study particle transport and filtration in porous media: Coupling cfd and deep learning**. Chemical Engineering Journal, 417:128936, 2021.

[42] DU, B.; LUND, P. D. ; WANG, J.. **Combining cfd and artificial neural network techniques to predict the thermal performance of all-glass straight evacuated tube solar collector**. Energy, 220:119713, 2021.

[43] FEI, Z.; YANG, F.; TSUI, K.-L.; LI, L. ; ZHANG, Z.. **Early prediction of battery lifetime via a machine learning based framework**. Energy, 225:120205, 2021.

[44] DJANDJA, O. S.; DUAN, P.-G.; YIN, L.-X.; WANG, Z.-C. ; DUO, J.. **A novel machine learning-based approach for prediction of nitrogen content in hydrochar from hydrothermal carbonization of sewage sludge**. Energy, 232:121010, 2021.

[45] SUN, L.; LIU, T.; XIE, Y.; ZHANG, D. ; XIA, X.. **Real-time power prediction approach for turbine using deep learning techniques**. Energy, 233:121130, 2021.

[46] LENG, E.; HE, B.; CHEN, J.; LIAO, G.; MA, Y.; ZHANG, F.; LIU, S. ; E, J.. **Prediction of three-phase product distribution and bio-oil heating value of biomass fast pyrolysis based on machine learning**. Energy, 236:121401, 2021.

[47] AGBULUT, U.; GÜREL, A. E. ; SARIDEMIR, S.. **Experimental investigation and prediction of performance and emission responses of a ci engine fuelled with different metal-oxide based nanoparticles–diesel blends using different machine learning algorithms**. Energy, 215:119076, 2021.

[48] SHALABY, A.; ELKAMEL, A.; DOUGLAS, P. L.; ZHU, Q. ; ZHENG, Q. P.. **A machine learning approach for modeling and optimization of a co2 post-combustion capture unit**. Energy, 215:119113, 2021.

[49] FREIESLEBEN, J.; KEIM, J. ; GRUTSCH, M.. **Machine learning and design of experiments: Alternative approaches or complementary methodologies for quality improvement?** Quality and Reliability Engineering International, 36(6):1837–1848, 2020.

[50] WELLER, H. G.; TABOR, G.; JASAK, H. ; FUREBY, C.. **A tensorial approach to computational continuum mechanics using object-oriented techniques.** Computers in Physics, 12:620, 1998.

[51] Ansys Inc. **Ansys Fluent Theory Guide**, 12 edition, 2015.

[52] POPE, S. B.. **Turbulent Flows.** Cambridge University Press, Cambridge, United Kingdom, 10 edition, 2000.

[53] PENG, D.-Y.; ROBINSON, D. B.. **A new two-constant equation of state.** American Chemical Society, 15, 1976.

[54] CHASE, M. W.; CURNUTT, J. L.; HU, A. T.; PROPHET, H.; SYVERUD, A. N. ; WALKER, L. C.. **Janaf thermochemical tables, 1974 supplement.** Journal of Physical and Chemical Reference Data, 3(2):311–480, 1974.

[55] SUTHERLAND, W.. **Lii. the viscosity of gases and molecular force.** The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 36(223):507–531, 1893.

[56] MCBRIDE, B. J.; GORDON, S. ; RENO, M. A.. **Coefficients for calculating thermodynamic and transport properties of individual species.** NASA Techinical Memorandum, 1993.

[57] PATANKAR, S. V.. **Numerical heat transfer and fluid flow.** Series on Computational Methods in Mechanics and Thermal Science. Hemisphere Publishing Corporation (CRC Press, Taylor & Francis Group), 1 edition, 1980.

[58] ISSA, R.. **Solution of the implicitly discretised fluid flow equations by operator-splitting.** Journal of Computational Physics, 62(1):40–65, 1986.

[59] COURANT, R.; FRIEDRICHS, K. ; LEWY, H.. **Über die partiellen differenzengleichungen der mathematischen physik.** Mathematische Annalen, 100, 1928.

[60] COLEMAN, H.; MEMBERS, C.. **ASME V&V 20-2009 Standard for Verification and Validation in Computational Fluid Dynamics**

and Heat Transfer (V&V20 Committee Chair and principal author). 01 2009.

[61] GOUPY, J.; CREIGHTON, L.. **Introduction to Design of Experiments with JMP Examples**. SAS Press, Stockholm, Sweden, 3 edition, 2007.

[62] GÉRON, A.. **Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent System**. O'Reilly Media, Inc, Sebastopol, California, 5 edition, 2019.

[63] SAMUEL, A. L.. **Some studies in machine learning using the game of checkers**. IBM JOURNAL OF RESEARCH AND DEVELOPMENT, p. 71–105, 1959.

[64] GALTON, S. F.. **Natural inheritance**. London :Macmillan,, 1 edition, 1889. https://www.biodiversitylibrary.org/bibliography/32181.

[65] MACKAY, D. J. C.. **Bayesian interpolation**. Neural Computation, 4(3):415–447, 05 1992.

[66] ZOU, H.; HASTIE, T.. **Regularization and variable selection via the elastic net**. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67(2):301–320, 2005.

[67] HUBER, P. J.. **Robust Statistics**. Wiley, 1 edition, 1981.

[68] EFRON, B.; HASTIE, T.; JOHNSTONE, I. ; TIBSHIRANI, R.. **Least angle regression**. The Annals of Statistics, 32(2):407 – 499, 2004.

[69] TIBSHIRANI, R.. **Regression shrinkage and selection via the lasso**. Journal of the Royal Statistical Society. Series B (Methodological), 58(1):267–288, 1996.

[70] JANUAVIANI, T. M. A.; GUSRIANI, N.; JOEBAEDI, K.; SUPIAN, S. ; SUBIYANTO, S.. **The best model of lasso with the lars (least angle regression and shrinkage) algorithm using mallow's cp**. World Scientific News, 116:245–252, 2019.

[71] PATI, Y. C.; REZAIIFAR, R. ; KRISHNAPRASAD, P. S.. **Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition**. p. 1–3, 1993.

[72] SHALEV-SHWARTZ, S.; CRAMMER, K.; DEKEL, O. ; SINGER, Y.. **Online passive-aggressive algorithms**. Journal of Machine Learning Research, 16, 2004.

[73] FISCHLER, M. A.; BOLLES, R. C.. **Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography**. Commun. ACM, 24(6):381–395, jun 1981.

[74] HOERL, A. E.; KENNARD, R. W.. **Ridge regression: Biased estimation for nonorthogonal problems**. Technometrics, 42(1):80–86, 2000.

[75] KIEFER, J.; WOLFOWITZ, J.. **Stochastic Estimation of the Maximum of a Regression Function**. The Annals of Mathematical Statistics, 23(3):462 – 466, 1952.

[76] FIX, E.; HODGES, J. L.. **Discriminatory analysis. nonparametric discrimination: Consistency properties**. International Statistical Review / Revue Internationale de Statistique, 57(3):238–247, 1989.

[77] DRUCKER, H.; BURGES, C. J. C.; KAUFMAN, L.; SMOLA, A. ; VAPNIK, V.. **Support vector regression machines**. In: Mozer, M. C.; Jordan, M. ; Petsche, T., editors, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, volume 9. MIT Press, 1997.

[78] SCHÖLKOPF, B.; SMOLA, A. J.; WILLIAMSON, R. C. ; BARTLETT, P. L.. **New support vector algorithms**. Neural Computation, 12(5):1207–1245, 05 2000.

[79] BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A. ; STONE, C. J.. **Classification And Regression Trees**. Routledge, 1 edition, 1984.

[80] GEURTS, P.; ERNST, D. ; WEHENKEL, L.. **Extremely randomized trees**. Machine Learning, 63, 2006.

[81] FREUND, Y.; SCHAPIRE, R. E.. **A decision-theoretic generalization of on-line learning and an application to boosting**. Journal of Computer and System Sciences, 55(1):119–139, 1997.

[82] BREIMAN, L.. **Bagging predictors**. Machine Learning, 24, 1996.

[83] FRIEDMAN, J. H.. **Greedy function approximation: A gradient boosting machine.** The Annals of Statistics, 29(5):1189 – 1232, 2001.

[84] BREIMAN, L.. **Random forests**. Machine Learning, 45, 2001.

[85] CHEN, T.; GUESTRIN, C.. **Xgboost: A scalable tree boosting system**. p. 785–794, 2016.

[86] MATSUMOTO, M.; NISHIMURA, T.. **Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator**. ACM Trans. Model. Comput. Simul., 8(1):3–30, jan 1998.

[87] PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M. ; DUCHESNAY, E.. **Scikit-learn: Machine learning in Python**. Journal of Machine Learning Research, 12:2825–2830, 2011.

[88] BERGSTRA, J.; BENGIO, Y.. **Random search for hyper-parameter optimization**. Journal of Machine Learning Research, 13(10):281–305, 2012.

[89] KUHN, M.; JOHNSON, K.. **Applied predictive modeling**. Springer, 5 edition, 2018.

[90] International Electrotechnical Committee. **Thermocouples – Part 3: Extension and compensating cables – Tolerances and identification system**, 2 edition, 2008.

[91] KOPETZ, H.. **Real-Time Systems: Design Principles for Distributed Embedded Applications**. Springer, Wien, Austria, 2 edition, 2011.

[92] BOX, G. E. P.; BEHNKEN, D. W.. **Some new three level designs for the study of quantitative variables**. Technometrics, 2(4):455–475, 1960.

[93] DOEHLERT, D. H.. **Uniform shell designs**. Journal of the Royal Statistical Society. Series C (Applied Statistics), 19(3):231–239, 1970.

[94] HOTELLING, H.. **Some improvements in weighing and other experimental techniques**. The Annals of Mathematical Statistics, 15(3):297–306, 1944.

[95] FANG, K.-T.; HICKERNELL, F.. **Uniform Experimental Design**. 1 edition, 03 2008.

[96] EFRON, B.. **The jackknife, the bootstrap and other resampling plans**. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1 edition, 1982.